SURG Application (Summer 2018): Statistics; Computer Science; Mathematical Methods in the Social Sciences; Psychology

Statistical Learning Theory Methods applied to Psychological MMCAR Data: An Investigation

SURG | Social Sciences and Journalism (SSJ) | *Tags: Design/Build; Quantitative Data Analysis; Lab-based*

This cover page is meant to focus your reading of the sample proposal, summarizing important aspects of proposal writing that the author did well or could have improved. ***Review the following sections before reading the sample****. The proposal is also annotated throughout to highlight key elements of the proposal's structure and content.*

Proposal Strengths	Areas for Improvement
While the methodology and subject of this proposal crosses disciplines, the researcher does a great job of writing for a general audience, and reviewing literature from relevant fields.	The student uses rhetorical questions in some places of the proposal. This is not a technique commonly used in grant writing.
The background section goes beyond summarizing past literature, and instead presents an argument exposing a gap in knowledge in the current literature, and why it should be filled.	The student does not include 1-2 sentences describing how this project relates to the student's future goals. This information is useful since the committee is funding both the project, and the person.
The specific aims of the research are clearly laid out and well organized.	
The output of the research is clear and explicitly connected to the research aims.	

Other Key Features to Take Note Of
Since this is a "build" based project, the student identifies a clear plan to validate the functionality of the proposed output/tool relative to what is accepted in the field.
Since this project was conducted within a psychology lab with human subject data, the researcher submitted IRB and was ruled exempt due to the nature of this project. You may have to do this if anything you are doing in your research relates to human subjects. The researcher also submitted proof of CITI Training Certificate in Social and Behavioral Research in the appendix, which was removed from the sample grant for anonymity.

SURG Application (Summer 2018): Statistics; Computer Science; Mathematical Methods in the Social Sciences; Psychology

Throughout psychology's history, the randomized controlled experiment has been regarded by many as the "gold standard" for furthering knowledge about how humans think, act, feel and behave. However, in recent decades the advent of rapid computation and the Internet have opened a multitude of possibilities in psychological research that previously were practically impossible. Two such tools have sprung forth that hold enormous potential understanding and predictive value for the field of personality psychology: large-scale statistical learning theory (SLT) methods, and a type of data structure called Massively Missing Completely-At-Random (MMCAR) data. SLT methods, from the field of statistics, are used to explore datasets with the aims of *prediction* and *inference*—either predicting outcomes using data, or trying to learn more about a subject using data. MMCAR data collection is a relatively easy, feasible way to collect personality data from hundreds of thousands of subjects on dozens of psychometric scales, without tiring out survey respondents with too many questions. Both SLT methods and MMCAR data are capable of revealing deep insights about personality, but currently there is no technology or procedure to interface between these two tools. Without these tools in existence, analysis will otherwise require tedious, repetitive, data cleaning that is a barrier to entry for many researchers. With my research, I will fill this gap by writing a code package in the open-source computer language R to implement SLT analysis methods designed to work with MMCAR data, and will test this code against equivalent psychology standard methods as a control. Long term, these code tools will allow for more psychological studies on large MMCAR datasets with newfound ease, reduction of error and speed.

Intro moves from broad topic to specific issue

Clear research/project statement occurs in 1st paragraph

While statistical learning theory could not have been feasibly applied to large datasets before the era of high-speed computation, that technology is now available—and though it has been utilized in many other fields, SLT methods are not yet widespread in psychology. However, recent literature (Yarkoni & Westfall, 2017) suggests a wide spread of gains through which SLT could benefit psychology. These researchers suggest that the field has a "near-total focus on *explaining* the causes of behavior" (Yarkoni & Westfall, 2017) which has resulted in psychology in general having "little (or unknown) ability to *predict* future behaviors with any appreciable accuracy" (Yarkoni & Westfall, 2017). This new pattern has begun to take hold among some researchers; for example, one study applied statistical learning theory methods to personality data and predicted mortality-rate (Chapman et al., 2016).

At the same time as machine learning has bloomed into an accessible, powerful analytic tool, the rise of the internet and online surveys make huge personality datasets possible (Revelle et al., 2013), and a type of data called MMCAR has emerged as a result. MMCAR and is based on a type of data long studied in statistics called MCAR, or "Missing Completely-At-Random" (Little, 1988). The huge advantage of MMCAR data is that, due to its sheer size, MMCAR makes it possible to survey participants using a small set of questions randomly chosen from a larger pool of questions, and still obtain all the information needed for personality scales. Because of the sheer scope of the sample size, all the questions are asked, but no one participant has to answer all or even a majority of the questions. Thus, a mass of data is aggregated without tiring out participants (Revelle et al., 2013). In the past, MCAR could not have been integrated into studies on purpose because of constrained sample size, which is now possible through web/smartphone surveys (Revelle et al., 2013).

Background goes beyond summarizing relevant literature. It sets up a justification for the specific research aims

Identifies gap in knowledge

Where does this research project find its place? MMCAR is an emerging strategy of data collection and could hugely benefit from having more tools at its disposal. At the same time, SLT has huge potential value to the field of psychology. Therefore, my project will further *both* of these new approaches, SLT and MMCAR, by creating an open-source, publicly accessible code package filled with tools that allow easier analysis of MMCAR data by SLT methods.

Since this project's goal is to create tools for the methodology of future psychological experiments—that is, for the interface between MMCAR data and statistical learning theory (SLT) methods—there will be two main parts to the project: the tool creation and the validation of the tools by comparing against a control analysis. Ideally, the code functions created would

Justifies why gap should be filled

Background leads to specific goals of 8 week project

cover a wide range of types of statistical learning theory methods, because researchers looking to examine MMCAR data likely have diverse interests. Therefore, I plan to build in this project at least *five* SLT methods that can interface with MMCAR data. For researchers interested in making predictions about their MMCAR datasets, I will create two *predictive methods*, one for regression (quantitative) variable prediction (A) and one for classification (categorical) variable use (B). For researchers who are focusing on understanding the relationships between variables in their datasets, an *inference method* (C) is essential. All researchers would need to validate their models in question, so I will create at least one SLT *validation method* function (D). Finally, the fifth minimum method that I will create is a *high-dimensionality method* (E), because MMCAR datasets by nature tend to have a large number of variables, which requires analysis that accounts for this increased dimension. In general, the purpose of the first three types of methods is to address researchers' many different potential interests; validation is crucial for any model, and the fifth method is necessary for many MMCAR datasets.

Great organization of specific aims of the project.

Methods are defined and justified

In my research, I will choose Multiple Linear Regression (A) and Logistic Regression (B), two reliable and commonly-used methods, to fill the need for a regression predictive method and a classification predictive method, respectively. K-Nearest Neighbors (C) is an SLT method that would fulfil the need for an inference method of looking at MMCAR data, and is convenient because it can be used on either quantitative or categorical variables. For validation, I will implement K-fold Cross-Validation (D), which provides a useful measure of the effectiveness of a model, and for a dimensionality-addressing method I would implement Ridge Regression (E), a type of method that can eliminate variables that do not contribute to the model. When implementing SLT methods, I will research and improve upon a variety of techniques for overcoming the unique missing-data challenge of MMCAR data, and I will start with *multiple imputation*, a data analysis technique that fills in missing data entries with values that do not affect the data distribution but allows other analysis to be performed on the data. For all SLT methods, I will create functions written in R; this approach will easily accept MMCAR data as an input and smoothly perform SLT methods on this type of data.

Specific description of the analysis process

Clear plan to validate functionality of tool within field of study

Finally, after writing the code functions, I will test them by running simulated MMCAR datasets and comparing them against current lab procedures used at the Personality, Motivation and Cognition Lab, or, in the case of K-fold Cross-Validation, against an equivalent standard called a Best Scales analysis (Elleman et al., under review). Because a MMCAR dataset might have hundreds of thousands of subjects, *p-values* are not a good enough measure of model effectiveness. I will examine p-values below 0.01 as an extra check after correcting for multiple comparisons, but to determine the effectiveness of a tool, I will compare *effect* sizes of each created code function's output versus its control's. I will use a multiple-correlation "Cohen's statistic to measure effect size, and as a measure of significant impact, I will look for an effect size difference of at least 0.15 to determine whether the new model is effective.

Output and connection to research aims are incredibly clear

Throughout this research project, I will need to write functions in R, understand the basic theory of SLT methods and how to apply them, and understand the principles of MMCAR data. I am well-qualified to write functions in R due to both my past experience creating code functions in other languages—I have taken two computer science fundamentals courses, EECS 111 and 211, at Northwestern, and also have AP Computer Science experience—as well as my experience working with R. I have used R regularly in Data Science I (Stat 301-1) and Data Science II (Stat 301-2). In these classes, I learned about and regularly analyzed datasets using statistical functions and SLT methods, which I will apply in this research. In addition, my faculty advisor is highly competent in R and is a resource I can consult. In order to prepare for research, I have already begun studying MMCAR datasets and will continue to do so through the spring. I will draw on my faculty advisor's expertise with MMCAR data as I create code functions; however, although this project builds on work that the PMC Lab has done in the past, my project is independent from the lab's regular routine. I hope to create greater awareness and accessibility of SLT methods in psychology with the creation of this code package.

Could have included 1-2 sentences showing why the research relates to future goals

---

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science, 12*(6), 1100-1122. doi:10.1177/1745691617693393

Little, R. J. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association, 83*(404), 1198. doi:10.2307/2290157

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2013). Web- and Phone-based Data Collection using Planned Missing Designs. *The SAGE Handbook of Online Research Methods,* 578-594. doi:10.4135/9781473957992.n33

Chapman, B. P., Weiss, A., & Duberstein, P. (2016). Statistical Learning Theory for High Dimensional Prediction: Application to Criterion-Keyed Scale Development. *Psychological Methods, 21*(4), 603-620. doi:10.1037/met0000088.supp

Elleman, L. G., Condon, D. M., Holtzman, N. S., Allen, V., & Revelle, W. (under review). Personality is associated with U.S. neighborhood demographics: Analysis at the factor, facet, and item levels.