Nathanael Kazmierczak
Dr. Douglas A. Vander Griend
07/28/2016

Does My Data Mean Anything? – Studying the Methodology of Factor Analysis

Numbers are ubiquitous in the modern world. The rise of information technology has made access to data easier than ever before, and the multiplicity of readily available data analysis programs makes it tempting to process data without understanding how data analysis techniques actually work. This "black-box" approach can lead to serious mistakes in scientific research, so it is important to research and fully document the techniques being used.

My research this summer has focused on understanding the methodology and limitations of factor analysis, a powerful data analysis technique used in chemistry to obtain information about multiple chemicals when they are mixed together and cannot be separated. Factor analysis was originally used in the social sciences to analyze surveys where multiple 'factors' may contribute to a person's responses. Starting with the publication of the seminal monograph *Factor Analysis in Chemistry* by Edmund Malinowski, this technique began to be used for chemistry research. However, factor analysis has multiple variants, and few chemists have attempted to define the strengths and weaknesses of the different methods. Additionally, the theory of error propagation is much less straightforward in factor analysis than in most scientific study. This makes it easy to obtain results that appear to be important, but in reality are so filled with error as to be practically meaningless.

To determine when factor analysis produces chemically meaningful results, I applied artificial data and Monte Carlo simulations. In artificial data, I generated data sets in a computer (using the programming language MATLAB) that represent what a "true" data set would look like according to the pertinent laws of chemistry, such as chemical equilibrium, without any interfering error from instruments or human mistakes in the lab. I then added an error pattern that simulates different types of possible error, and analyzed the data using factor analysis. Because the 'true' answers from which the data set was generated were known, I could compare the calculated values with the true values and quantify how much the added error affected the answer. If the error pattern involved randomness (as with instrument noise), then I repeated this process for 30 – 100 data sets in a Monte Carlo simulation, which produces a distribution of the possible calculated values for the given amount of random error added.

Many results have come from my research this summer. Using Monte Carlo simulations, we have concluded that factor analysis begins to yield meaningless results when applied to reactions stronger than a ΔG of -70. This work will help scientists have a quantitative measure of confidence when using the factor analysis technique. We have also quantified the effects of instrument noise, mistakes in solution concentration, insufficient mathematical constraints, and a variety of other topics that will provide new guidelines for acceptable and inacceptable uses of factor analysis in chemistry.

As a result of my research this summer, I have made huge gains in my ability to solve problems creatively and think independently. I have gained data analysis and computer science skills that are very useful in chemistry, and I have also gained confidence that I will be well prepared to complete graduates studies. Perhaps most importantly, though, my work this summer showed me how much I enjoy scientific research. I am now convinced that a scientific career represents the vocation through which I will serve God in the years to come.