Data Science Seminar

Hosted by the Department of Mathematical Sciences

- Date: Tuesday, October 25, 2022
- Time: 1:15pm – 2:15pm
- Room: Whitney Hall 100E
- Speaker: Dr. John Stufken (George Mason University)
- Title: Musings on Subdata Selection

Abstract

Data reduction or summarization methods for large datasets (full data) aim at making inferences by replacing the full data by the reduced or summarized data. Data storage and computational costs are among the primary motivations for this. In this presentation, data reduction will mean the selection of a subset (subdata) of the observations in the full data. While data reduction has been around for decades, its impact continues to grow with approximately 2.5 exabytes (2.5 x 10 18 bytes) of data collected per day. We will begin by discussing an information-based method for subdata selection under the assumption that a linear regression model is adequate . A strength of this method, which is inspired by ideas from optimal design of experiments, is that it is superior to competing methods in terms of statistical performance and computational cost when the model is correct. A weakness of the method, shared with other model-based methods, is that it can give poor results if the model is incorrect. We will therefore conclude with a discussion of a model-free method.

The work discussed here has been with various collaborators, including Rakhi Singh (Binghamton U), HaiYing Wang (U of Connecticut), and Min Yang (U of Illinois at Chicago).

Biography of the speaker: Since 2022, Dr. John Stufken is Professor of Statistics at George Mason University. Before this he held positions at the University of Georgia (1986-1990 and 2003-2014), Iowa State University (1988-2002), Arizona State University (2014-2019) and the University of North Carolina at Greensboro (2019-2022). He was Head of the Department of Statistics at UGA (2003-2014), Coordinator for Statistics in ASU's School of Mathematical and Statistical Sciences (2014-2019), where he was the Charles Wexler endowed Professor of Statistics, and Director for Informatics and Analytics at UNCG, where he was Bank of America Excellence Professor. In addition to these academic positions, he was Program Director for Statistics at the National Science Foundation (2000-2003). Stufken's research interests are primarily in design of experiments and, more recently, big data analysis. He is author of multiple articles in top statistics journals, and of the Springer Verlag book Orthogonal Arrays: Theory and Applications. He is co-editor of the Handbook of Design and Analysis of Experiments, by CRC Press. He has been editor for the Journal of Statistical Planning and Inference and The American Statistician, and is currently associate editor for the Journal of the American Statistical Association, Statistica Sinica, International Statistical Review, and Journal of Statistical Theory and Practice. He is an Elected Fellow of both the American Statistical Association and the Institute of Mathematical Statistics and is an elected member of the International Statistical Institute. He held the title of Rothschild Distinguished Visiting Fellow at the Isaac Newton Institute of Mathematical Sciences in Cambridge, UK.