TASC Texas A&M at Qatar Advanced Scientific Computing Center

Clustering high-dimensional data using summary statistics

23 September, 1 p.m., LH 238

Dr. Valen Johnson
Dean of Science
University Distinguished Professor
Texas A&M University

The ability to cluster high dimensional data, in which large numbers of features are measured on a comparatively small number of objects, has become increasingly important as new measurement and data acquisition technologies have become prevalent. We propose a model-based clustering algorithm to address this challenge. The algorithm is insensitive to the distributions of features that are measured on each object and does not require the specification of tuning parameters. Assuming that P feature measurements for N objects are accumulated in an N x P matrix X}, where N<<P, the method is based on exploiting the cluster-dependent structure of the Nx N matrix XX'. Computational burden thus depends primarily on N, the number of objects to be clustered, rather than P, the number of features that are measured. This makes the method particularly useful in high dimensional settings, where it is substantially faster than a number of other popular clustering algorithms. The method is best applied to data in which the structural dependence between features need not be explicitly modeled, for instance when analyzing many types of genomic data. To illustrate the method, we compared it to 17 other clustering algorithms applied to 32 genomic data sets in which gold standards were available. We found that it provides the most accurate cluster configuration more than twice as often as its closest competitors. Detailed examples involving specific genomic data sets are provided.

Dr. Valen Johnson

Dr. Valen Johnson is dean of the College of Science at Texas A&M University. He joined the Texas A&M faculty in 2012 as professor of statistics and has served as head of the department. He previously was a professor of biostatistics at the University of Texas M.D. Anderson Cancer Center (2004-2012) and the University Michigan (2002-2004) and professor of statistics at Duke University (1989-2001). He also worked for one year at Los Alamos National Laboratory (2001-2002). He received his Ph.D. in statistics from The University of Chicago in 1989. His applied research interests include educational assessment, ordinal data and rank data analysis, clinical trial design, image analysis, and reliability analysis. His current methodological interests focus on Bayesian hypothesis testing and its connections to classical testing procedures; Bayesian variable selection; Markov chain Monte Carlo model diagnostics; and latent variable modeling.

For more information or to RSVP, contact Dr. Othmane Bouhali, **othmane.bouhali@qatar.tamu.edu**