Numerical Analysis and Scientific Computing Seminar

Imputing Missing Data with the Gaussian Copula

Madeleine Udell
Cornell University

Abstract: Missing data imputation forms the first critical step of many data analysis pipelines. The challenge is greatest for mixed data sets, including real, Boolean, and ordinal data, where standard techniques for imputation fail basic sanity checks: for example, the imputed values may not follow the same distributions as the data. This talk introduces a new semiparametric algorithm to impute missing values, with no tuning parameters. The algorithm models mixed data as a Gaussian copula. This model can fit arbitrary marginals for continuous variables and can handle ordinal variables with many levels, including Boolean variables as a special case. We develop an efficient approximate EM algorithm to estimate copula parameters from incomplete mixed data, and low rank and online extensions of the method that can handle extremely large datasets. The resulting model reveals the statistical associations among variables. Experimental results on several synthetic and real datasets show the superiority of the proposed algorithm to state-of-the-art imputation algorithms for mixed data.

Friday, October 2, 2020, 2:40 pm
https://emory.zoom.us/j/95900585494

Mathematics
Emory University