UCLA

BIOSTATISTICS SEMINAR

SPRING 2017

Correlation and Mixture in High Dimensional Data: Should the Distribution Look Normal?

Armin Schwartzman, PhD

Associate Professor of Biostatistics

University of California, San Diego

Wednesday, May 24, 2017

3:30pm - 4:30pm, CHS 33-105

Refreshments served at 3:00 PM in room 51-254 CHS

ABSTRACT:

Large scale multiple testing problems, such as in brain imaging and genomics, base their inference on a large number of z-scores. If most effects are null, it seems natural that the empirical distribution of z-scores should follow a standard normal distribution. But should it? In this talk I show two ways in which the empirical distribution of z-scores can be deceiving, because of correlation and mixture. First, following Efron's (2007) conjecture, I show that even if the z-scores are standard normal, the empirical distribution may depart from it, due to strong correlation caused by hidden random effects. Instead, it may be approximated by a Gaussian mixture that generalizes Efron's empirical null distribution. Second, I show that if the original data is a Gaussian mixture, then within-class standardization using a template-based EM algorithm produces z-scores whose empirical distribution looks standard normal. However, their true distribution has in fact lighter tails.