
High-Dimensional Matrix Models for Biological Data

Genevera I. Allen

Advances in high-throughput and imaging technologies have created massive biological data sets that present a challenge for classical statistical methodology. Often these data sets can be arranged in the form of a large matrix: tens of thousands of genes or isoforms with gene expression measured on a few samples in microarrays and RNA-Seq data, or hundreds of thousands of voxels from functional magnetic resonance images (fMRIs) measured on a few subjects.

In this talk, we will first review standard matrix-models based on the singular value decomposition (SVD) and the multivariate normal, giving an overview of recent statistical learning developments based on these two models. In the later half of this talk, we will introduce a new matrix-model, the transposable regularized covariance model, based on the matrix-variate normal distribution. This matrix-model allows one to directly capture the relationships between both the row and column variables. Applications to missing data imputation, multiple testing, graphical models and clustering will be presented with examples on microarrays, flow cytometry data, functional MRIs and RNA-Seq data.