

## Penalized Clustering of Large Scale Functional Data with Multiple Covariates

Ping Ma

University of Illinois at Urbana-Champaign

**Abstract:** With the rapid advancement in high throughput technology, extensive repeated measurements have been taken to monitor the system-wide dynamics in many scientific investigations. A typical example is temporal gene expression studies, in which a series of micorarray experiments are conducted sequentially during a biological process, e.g., cell cycle microarray experiments. At each time point, mRNA expression levels of thousands of genes are measured simultaneously. Collected over time, a gene's "temporal expression profile" gives the scientist some clues on what role this gene might play during the process. A group of genes with similar profiles are often "co-regulated" or participants of a common and important biological function. Thus clustering genes into homogeneous groups is a crucial first step to decipher the underlying mechanism. In addition to the time factor, such repeated measurements often contain other covariates, e.g., replicates at each time point, species in comparative genomics studies, and treatment groups in case-control studies, as well as many factors in a factorial designed experiment. Incorporation of multiple covariates adds another layer of complexity. Clustering methods taking all these factors into account are still lacking. In this talk, I will present a penalized clustering method for large scale data with multiple covariates through a functional data approach. Simulation studies and real-data examples are presented to investigate the empirical performance of the proposed method. Open-source code is available in the R package MFDA.