SCIENCE AT THE EDGE

2018 SEMINAR SERIES

Quantitative Biology Graduate Program | Gene Expression in Development and Disease

Aaditya Rangan

Courant Institute of Mathematical Sciences
New York University

"*Covariate-Corrected Low-Rank Biclustering Methods for Gene-Expression and GWAS Data*"

A common goal in data-analysis is to sift through a large matrix and detect any significant submatrices (i.e., biclusters) that have a low numerical rank. To give an example from genomics, one might imagine a data-matrix involving several genetic-measurements taken across many patients. In this context a 'bicluster' would correspond to a subset of genetic-measurents that are correlated across a subset of the patients. While some biclusters might extend across most (or all) of the patients, it is also possible for biclusters to involve only a small subset of patients. Detecting biclusters such as these provides a first step towards unraveling the physiological mechanisms underlying the heterogeneity within a patient population.

In this talk I'll describe a simple algorithm for tackling this biclustering problem – i.e., for detecting low-rank submatrices from within a larger data-matrix. The basic method itself is very straightforward (c.f. Rangan 2012), and involves examining the 'loops' (i.e., 2-x-2 submatrices) within the data-set. Importantly, this method can easily be modified to account for many considerations which commonly arise in practice. For example, by counting loops slightly differently, we can correct for controls: finding biclusters that manifest only within a 'case'-population without manifesting within a 'control'-population. Moreover, this methodology can also correct for categorical- and continuous-covariates, as well as sparsity within the data. I'll illustrate these practical features with two examples; the first drawn from gene-expression analysis and the second drawn from a much larger genome-wide-association-study (GWAS).

In addition to being quite practical, this loop-counting method exhibits a few interesting mathematical features, and much can be proven about this approach. For example, we can show that this method is `close to optimal' within a certain subset of local algorithms, and that this method comes close to achieving the (conjectured) computational phase-transition for the planted-bicluster problem. We can also explain why loop-counting outperforms more traditional spectral-methods. I'll discuss these mathematical details towards the end of the talk if I have time.

References:

Rangan A, McGrouther C, Kelsoe J, Schork N, Stahl E, Zhu Q, Krishnan A, Yao V, Troyanskaya, Bilaloglu S, Raghavan P, Bergen S, Jureus A, Landen M, Bipolar Disorders Working Group of the Psychiatric Genomics Consortium (2018) Supplementary Information for: A loop-counting method for covariate-corrected low-rank biclustering of gene-expression and genome-wide association study data.

Rangan A, McGrouther C, Kelsoe J, Schork N, Stahl E, Zhu Q, Krishnan A, Yao V, Troyanskaya, Bilaloglu S, Raghavan P, Bergen S, Jureus A, Landen M, Group of the Psychiatric Genomics Consortium (2018) A loop-counting method for covariate-corrected low-rank biclustering of gene-expression and genome-wide association study data

FRIDAY, FEBRUARY 16, 2018

11:30 AM, ROOM 1400 BPS

Refreshments at 11:15

MICHIGAN STATE UNIVERSITY