Stat 399 Research Seminar: Transposable Data

Art Owen, owen@stat

In most statistical work we arrange data in a table where each row is an observation and each column describes a measured variable. But in some settings the columns can equally well be viewed as observations with the rows being the variables. Examples of such transposable data arise when:

- people say which movies they saw, and perhaps how much they liked them
- for a set of words (or other features) and documents the number of appearances of each word in each document is counted
- expression data are gathered on a large number of genes under many different conditions.

This sort of data raises interesting questions and has been studied in many disjoint literatures. It is now prominent in applications of recent intense interest, including: analyzing genetic expression data, modeling users arriving at a website, retrieving documents. Methods proposed include: singular value decomposition, fuzzy clustering, semi-discrete decomposition, non-negative decomposition, Rasch models, correspondence analysis, gene shaving, plaid models.

Research challenges

Interesting issues include: how to cluster rows or columns or both at the same time, how to exploit covariates that arise just for rows or just for columns, how to handle data that is almost all missing, how to bootstrap permute or cross-validate, what to do when the rows and columns are the same individuals, extensions to three way and higher tables.

Activities

1. Survey and present research papers.
2. Have informal talks from domain experts.
3. Experiment with some real data sets.

Administrative

1. You can combine this with Stat 319 (take both, split the units)
2. Organizational meeting: Friday March 31 at noon in Girshick library, Sequoia Hall 105.