Examples of High-Dimensional Data

Genevera I. Allen

Statistical Learning: High-Dimensional Data

January 10, 2011

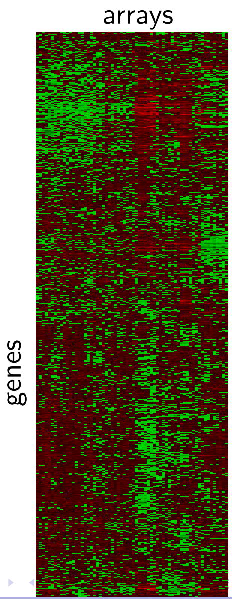
1. Large p, Small n: Biological Data

2. Random Fields Data

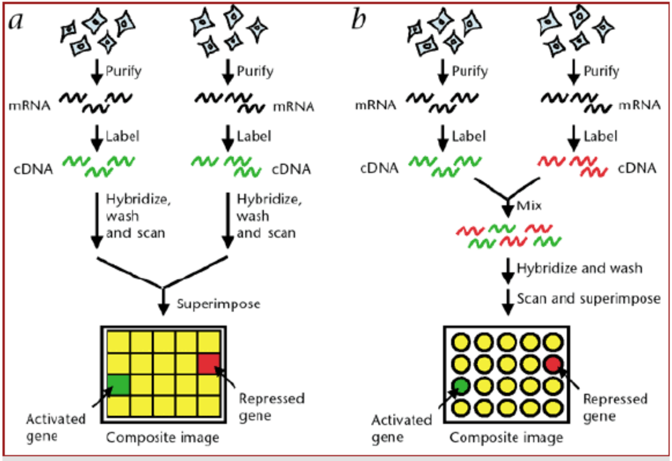
3. Collaborative Filtering Data

Example: Microarrays

- Measure gene expression.
- Often tens of thousands of genes (features).
- Only tens of hundreds of samples.



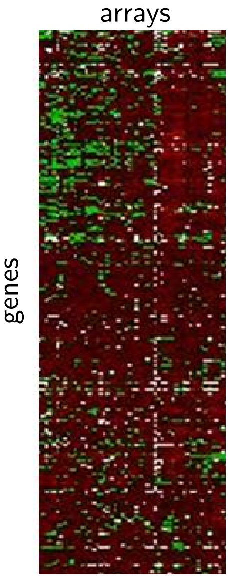
Review: Microarrays



(Stears, R. L. *et. al*, 2003)

Statistical Questions: Microarrays

- Data pre-processing:
  - Normalization.
  - Missing data imputation.
- Inference:
  - Which genes are significant?
- Clustering:
  - Groups of genes, groups of samples.
- Model Building:
  - Small n, large p.



Other Types of Biological Data

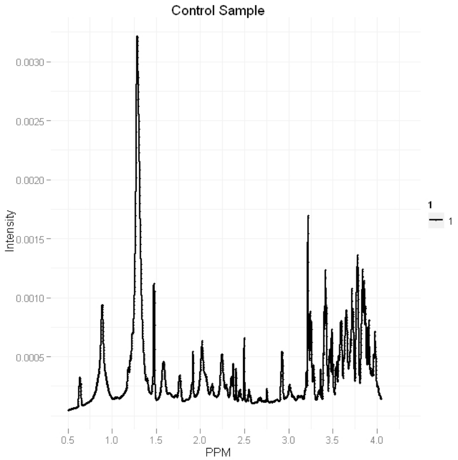
Genetics:

- Deep Sequencing - Counts.
- Micro RNA Expression - Continuous.
- CGH (Copy Number Variation) - Continuous / Categorical.
- SNPs (Single Nucleotide Polymorphisms) - Binary / Categorical.
- Methalaytion - Continuous.

Other Types of Biological Data

Proteomics / Metabolomics (Chemometrics):

- (H-NMR) Measures the chemical shift associated with various metabolites.



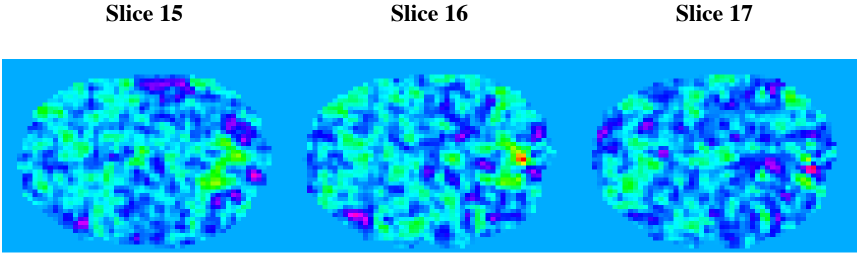
1. Large p, Small n: Biological Data

2. Random Fields Data

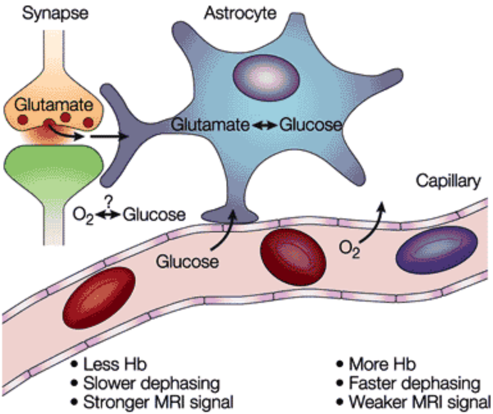
3. Collaborative Filtering Data

Example: Functional MRIs (fMRI)

- Rows: Voxels.
- Columns: Subjects (And/or replicates and times).
- Measurement: Hemodynamic response (change in blood flow).



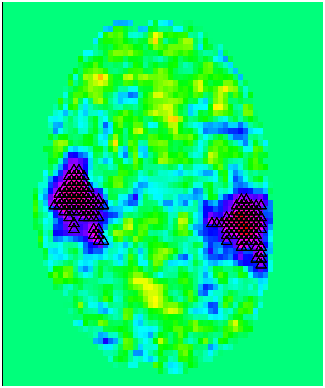
Review: fMRIs



(Heeger & Ress, 2002)

Statistical Questions: fMRIs

- Inference:
  - Which voxels are significant?
  - Which groups of voxels (regions of interest) are significant?
- Clustering:
  - Groups of voxels that behave similarly - finding regions of interest.
- Networks (Functional Connectivity):
  - How are voxels or groups of voxels related to each other?
  - How are voxels or groups of voxels related through time?



Others

- Finance.
  - Time Series Data.
- Climate Data.
  - Spatial Data.
  - Spatio-temporal Data.
- Neuroimaging.
  - DTI - Diffusion Tensor Imaging.
  - Calcium-Florescence Imaging.
  - EEG & MEG.

1. Large p, Small n: Biological Data

2. Random Fields Data

3. Collaborative Filtering Data

Example: Netflix Movie Rating Data

- Rows: Movies.
- Columns: Customers.
- Measurement: Movie ratings (scale of 1 - 5).

	Anne	Ben	Charlie	Doug	Eve	...
Star Wars	2	5	4	4	3	...
Harry Potter	3	4	5	3	?	...
Pretty Woman	4	?	2	?	5	...
Titanic	5	?	2	1	3	...
Lord of the Rings	?	5	5	4	4	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

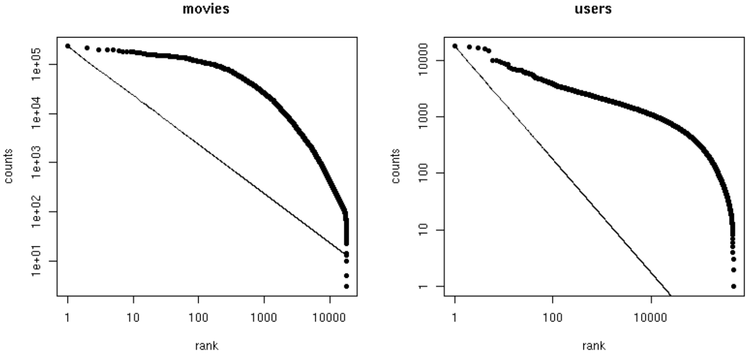
Netflix Prize

- Challenge: Predict un-rated movies with 10% improvement over Cinematch.
- Training Set: 480,000 customer ratings on 18,000 movies.
- Around 98.7% missing ratings!
- $1,000,000 prize!

- Contest: October 2006 - August 2009.
- Winners: Team led by Robert Bell and Yehuda Koren.
- Methods: Variations on the SVD and *k*-nearest neighbors (Bell & Koren, 2008).
- Fields: Recommender systems & Collaborative filtering.

Visualizing Netflix Data

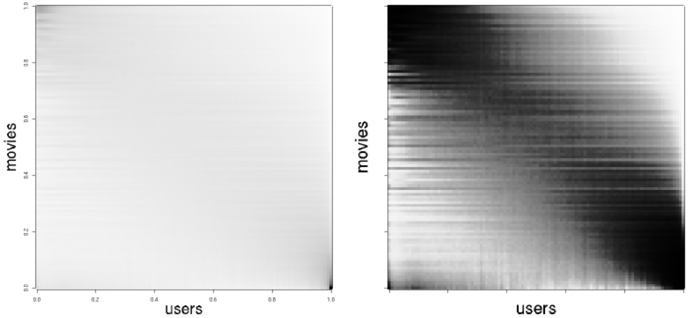
Zipf



(Justin S. Dyer & Art B. Owen, 2010)

Visualizing Netflix Data

Copulas



(Justin S. Dyer & Art B. Owen, 2010)

Other Examples

- Amazon
- Facebook
- Yahoo!
- Twitter