Stat 315c: Introduction

Art B. Owen

Stanford Statistics

Stat 315c "Analysis of Transposable Data"

Usual Statistics Setup

- there's Y (we'll predict it)
- and there's $X_1, \ldots, X_d$ (to predict from)
- and n IID copies of (X, Y) to infer with

Stat 315c "Analysis of Transposable Data"

Usual Statistics Setup

- there's Y (we'll predict it)
- and there's $X_1, \ldots, X_d$ (to predict from)
- and n IID copies of (X, Y) to infer with

Data matrix (X, Y) is n by $d+1$

- $d+1$ named columns (variables)
- and n anonymous exchangeable rows
- with $n \to \infty$ and d fixed

Stat 315c "Analysis of Transposable Data"

Usual Statistics Setup

- there's Y (we'll predict it)
- and there's $X_1, \ldots, X_d$ (to predict from)
- and n IID copies of (X, Y) to infer with

Data matrix (X, Y) is n by $d+1$

- $d+1$ named columns (variables)
- and n anonymous exchangeable rows
- with $n \to \infty$ and d fixed

This course

- Both rows and columns are named
- We learn about the cols using rows as obs, and conversely
- n and d may both be large

Problem domains

Two mode examples

- Movies $\times$ Raters $\longrightarrow$ Ratings
- Terms $\times$ Documents $\longrightarrow$ Counts
- Genes $\times$ Experiments $\longrightarrow$ Expression level
- IP-address $\times$ Books $\longrightarrow$ Purchases
- Questions $\times$ Test takers $\longrightarrow$ Grade

Problem domains

Two mode examples

- Movies $\times$ Raters $\longrightarrow$ Ratings
- Terms $\times$ Documents $\longrightarrow$ Counts
- Genes $\times$ Experiments $\longrightarrow$ Expression level
- IP-address $\times$ Books $\longrightarrow$ Purchases
- Questions $\times$ Test takers $\longrightarrow$ Grade

Single mode examples (Rows $\times$ Cols are the same entities)

- Actors $\times$ Actors $\longrightarrow$ co-appearance
- Articles $\times$ Articles $\longrightarrow$ co-citation
- Web pages $\times$ Web pages $\longrightarrow$ hyper-links

Problem domains

Two mode examples

- Movies $\times$ Raters $\longrightarrow$ Ratings
- Terms $\times$ Documents $\longrightarrow$ Counts
- Genes $\times$ Experiments $\longrightarrow$ Expression level
- IP-address $\times$ Books $\longrightarrow$ Purchases
- Questions $\times$ Test takers $\longrightarrow$ Grade

Single mode examples (Rows $\times$ Cols are the same entities)

- Actors $\times$ Actors $\longrightarrow$ co-appearance
- Articles $\times$ Articles $\longrightarrow$ co-citation
- Web pages $\times$ Web pages $\longrightarrow$ hyper-links

Higher dimensional layouts

- genes $\times$ conditions $\times$ tissues

About the course

History

- Response to common thread in lots of problems
- Began as seminar in spring 2000
- Guest speakers from Netflix
- and biomedical informatics
- and statistics

Goals

- Look at existing methods
- Look for remaining holes

Materials

- Articles online
- There's no book

Data types

Dyadic data

For $X \in \mathcal{X}$ (eg actors) and $Y \in \mathcal{Y}$ (eg movies)
Record pairs

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots (X_N, Y_N)$$

Actor X_i was in movie Y_i
$N \ll |\mathcal{X}| \times |\mathcal{Y}|$ so the full matrix would be very sparse

So it is "variables and cases as usual", after all

- Variable 1 = actor, Variable 2 = movie
- maybe Variable 3 = box office
- N = # pairs $\to \infty$, with d = 2 or 3

(Well almost)

Back to anonymous rows, but with

A special kind of random variable

- Categorical with many levels, e.g.:
  1. Phone number
  2. IP address
  3. Actor
  4. Query string
- Number of levels grows with N
- There may be many unseen levels

Back to anonymous rows, but with

A special kind of random variable

- Categorical with many levels, e.g.:
  1. Phone number
  2. IP address
  3. Actor
  4. Query string
- Number of levels grows with N
- There may be many unseen levels

Different from classical categorical variables, e.g.:

- Binary variables, or,
- Setosa vs Verginica vs Versicolor, etc.

Non-dyadic examples

Dense data

- for microarrays we have all genes in all experiments apart from missing values
- for dyadic case, it's mostly missing apart from a few observed values

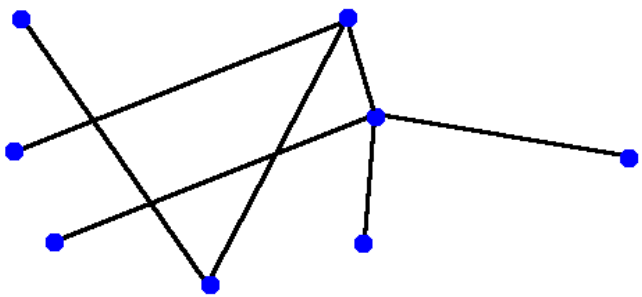
Triadic data

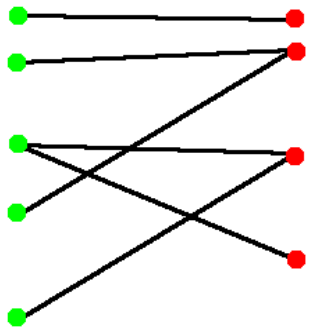
- Actors, Directors, and Year
- Genes, Conditions, Tissues

Graphs

Transposable data often have a graph representation.

- Edges can be directed or undirected
- Bipartite graphs for data of two modes (eg actors and films)

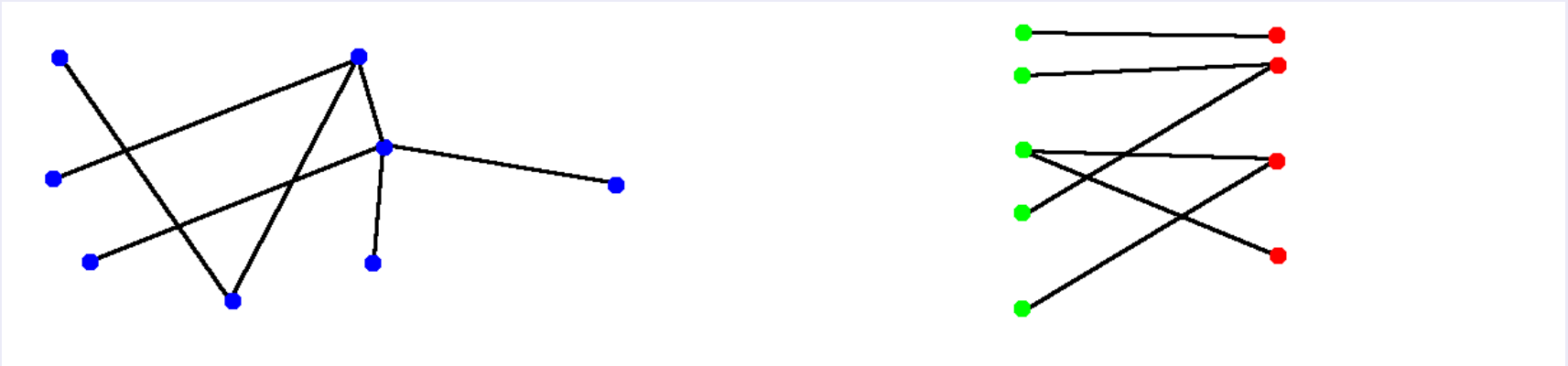




Graphs

Transposable data often have a graph representation.

- Edges can be directed or undirected
- Bipartite graphs for data of two modes (eg actors and films)



- Edges can have weights (more generally feature vectors)
- Nodes can have features
- Hypergraphs for triadic data
- Generalize ad infinitum (but then we break the graph paradigm)

Methods

Methods for these problems are of several (overlapping) types

- Classical
  1. ANOVA
  2. Correspondence analysis
  3. Rasch model

Methods

Methods for these problems are of several (overlapping) types

- Classical
  1. ANOVA
  2. Correspondence analysis
  3. Rasch model
- Unsupervised learning
  1. Clustering (group the rows or the columns)
  2. Biclustering (jointly group the rows and the columns)
  3. Spectral clustering
  4. Independent components analysis

Methods

Methods for these problems are of several (overlapping) types

- Classical
  1. ANOVA
  2. Correspondence analysis
  3. Rasch model
- Unsupervised learning
  1. Clustering (group the rows or the columns)
  2. Biclustering (jointly group the rows and the columns)
  3. Spectral clustering
  4. Independent components analysis
- Matrix approximation
  1. Singular value decomposition
  2. Nonnegative decomposition
  3. Semi-Discrete decomposition

But wait there's more

Some more ideas, not yet forced into a category

- PageRank, TrustRank, Hubs and Authorities
- Smoothing on graphs
- Subsampling matrices
- Recommender engines
- Archetypal analysis
- Latent Dirichlet Allocation
- Compositional data
- Canonical correlation and generalizations
- Head versus long tail

Problems and tasks

We'd like to

- Predict missing labels (eg spam)
- Find anomalies (eg unusual credit card patterns)
- Decide where to get labels
- Group rows/columns/both
- Reduce dimension
- Predict missing links

Goals

- Find common structures in these problem
- Learn some specific methods
- Learn to compare|mix|hybridize methods
- Move from "could to" to "should do"
- Spot research opportunities

High level view

Approaches include

- Principled Bayesian methods
- Ad hoc but very fast algorithms
- Moments
- Maximum likelihood

High level view

Approaches include

- Principled Bayesian methods
- Ad hoc but very fast algorithms
- Moments
- Maximum likelihood

Persistent issues

- What happens to the bootstrap and cross-validation?
- How should we window data arriving in time?
- Does anything go to ∞?
- Do we model missingness?

High level view

Approaches include

- Principled Bayesian methods
- Ad hoc but very fast algorithms
- Moments
- Maximum likelihood

Persistent issues

- What happens to the bootstrap and cross-validation?
- How should we window data arriving in time?
- Does anything go to ∞?
- Do we model missingness?

When we're done

- There will be lots of holes in the material

High level view

Approaches include

- Principled Bayesian methods
- Ad hoc but very fast algorithms
- Moments
- Maximum likelihood

Persistent issues

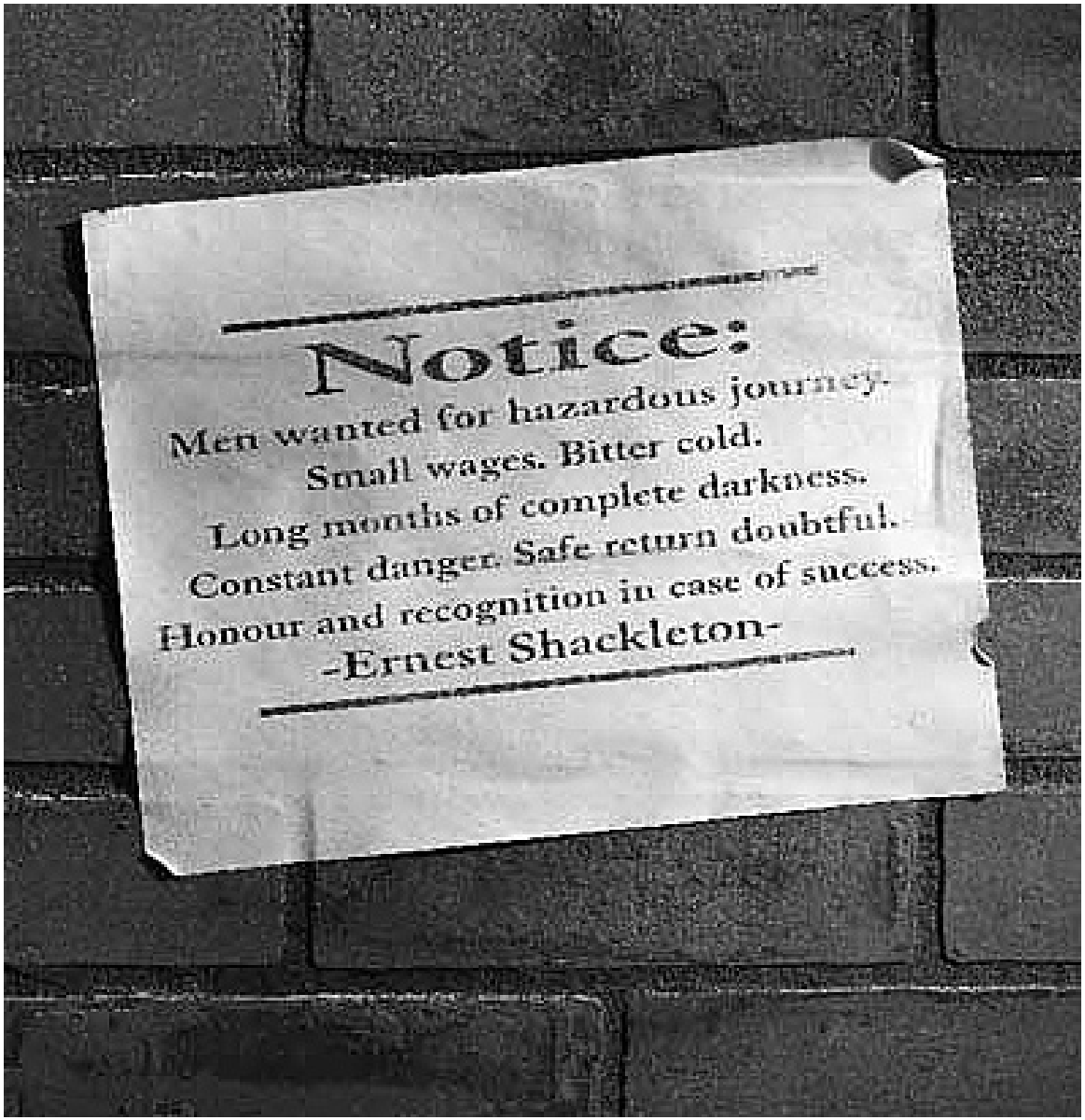
- What happens to the bootstrap and cross-validation?
- How should we window data arriving in time?
- Does anything go to ∞?
- Do we model missingness?

When we're done

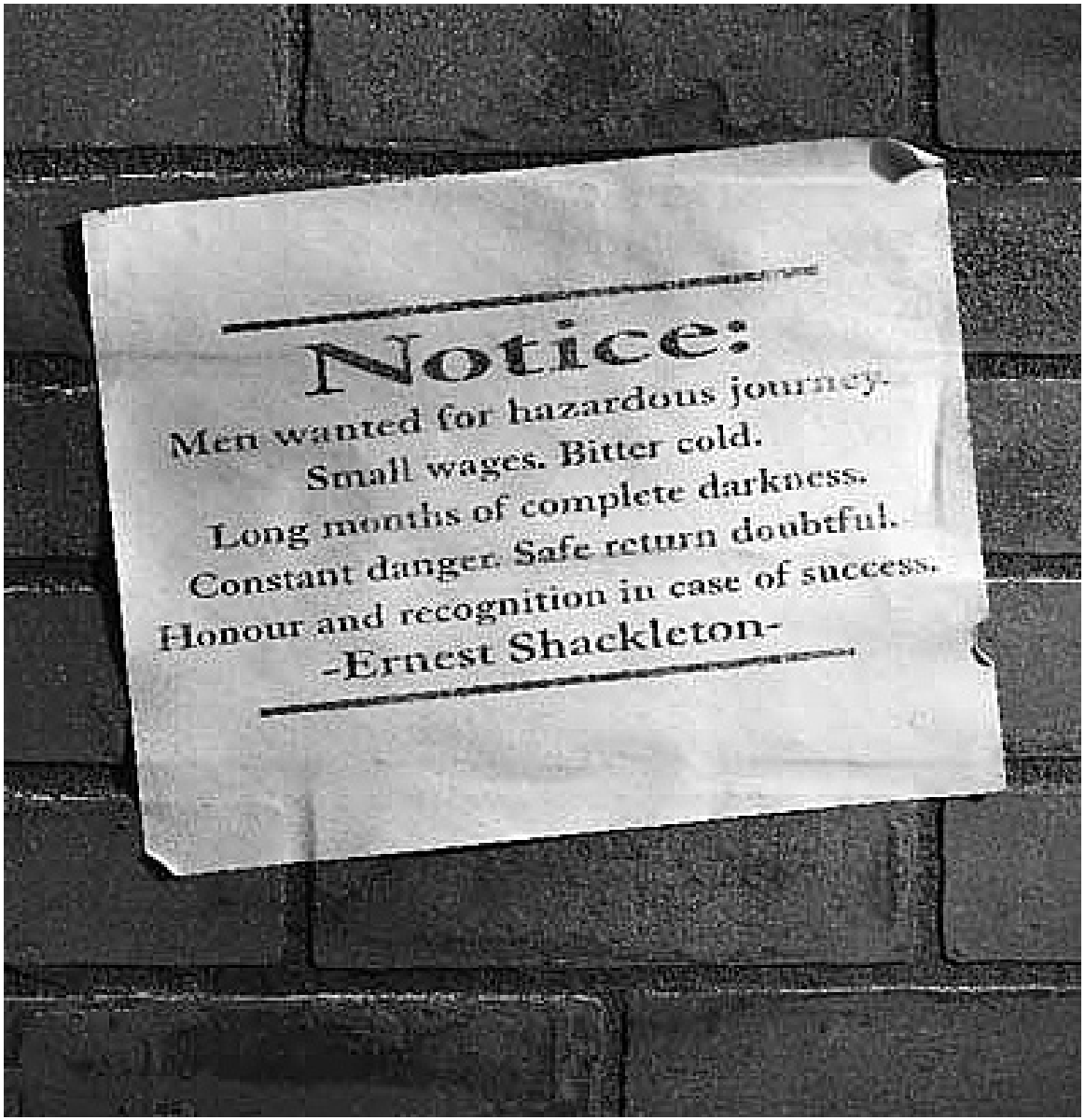
- There will be lots of holes in the material
- Right now there are disconnected islands

Not a usual course

Not a usual course



Not a usual course



Then

- It was 1914
- 5000 people applied

Now

- Men and women wanted
- It won't be cold

Some results

From 07/08

- New cross-validation method for (Perry and O.)
- New bootstrap for non-IID data (O.)
- Two papers on spectral clustering (Salzman)