Filling in Missing Data: Elections, ______, Healthcare

Madeleine Udell

Assistant Professor
Operations Research and Information Engineering
Cornell University

Cornell WAM, March 16 2019

Definitions

what is operations research?

Definitions

what is operations research?

my answer: using data to make decisions.

- using data to make predictions
- using predictions to make decisions

Outline

Elections

Data Tables

Generalized Low Rank Models

Applications

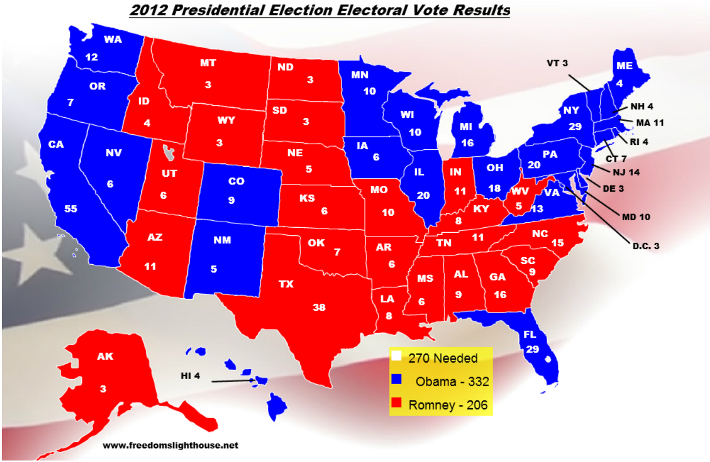
Bias

Analytics for political campaigns

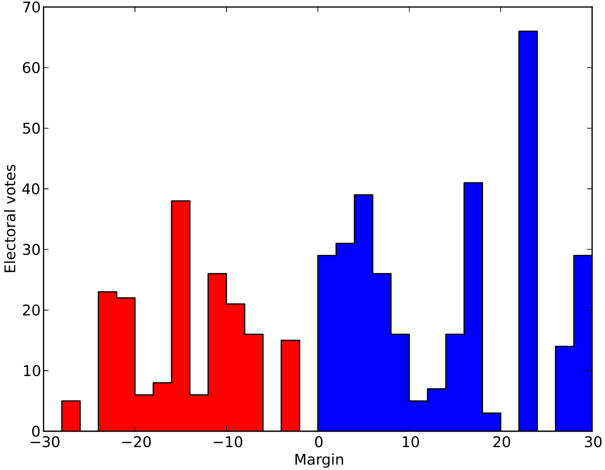
goal: allocate limited resources to optimize electoral vote

2012

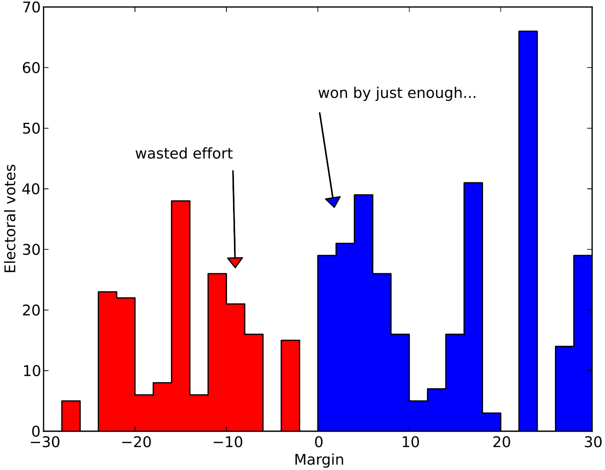
2012 electoral map



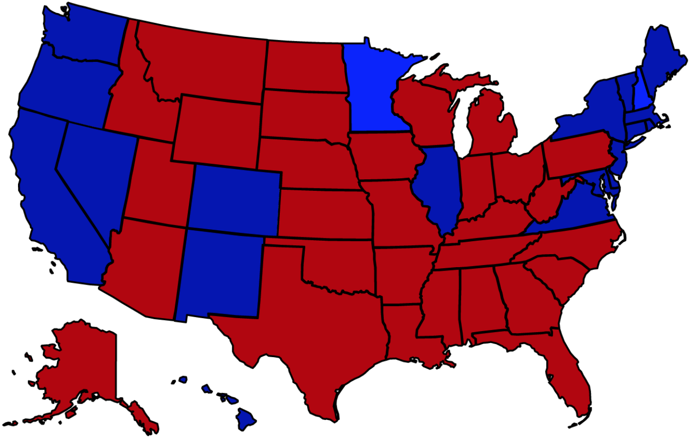
Optimization on the Obama campaign



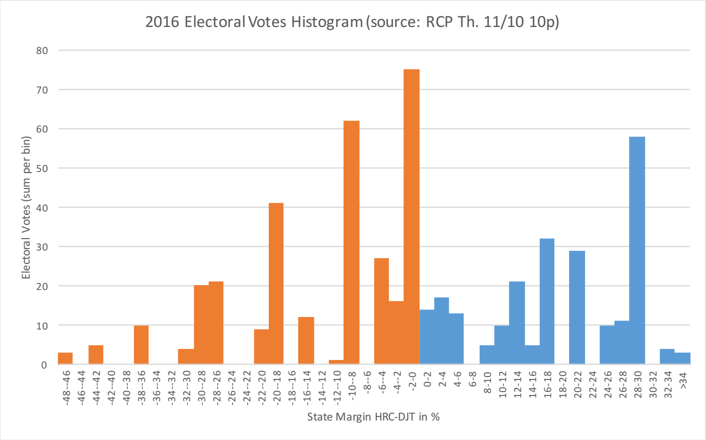
Optimization on the Obama campaign



2016 electoral map



Not much (effective) optimization on the Clinton campaign



Predictions and decisions in electoral campaigns

predictions

- who will vote?
- who will they vote for?
- how effective are interventions?

decisions

- volunteers: who should they talk to?
- money: what ads to display on what platforms?
- candidate's time: where to travel?
- policy positions: what to emphasize?

to maximize probability of electoral win

Outline

Data table: politics

age	gender	state	income	education	voted?	support	⋯
29	F	CT	\$53,000	college	yes	Clinton	⋯
57	?	NY	\$19,000	high school	yes	?	⋯
?	M	CA	\$102,000	masters	no	Trump	⋯
41	F	NV	\$23,000	?	yes	Trump	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Data table: politics

age	gender	state	income	education	voted?	support	$\cdots$
29	F	CT	$53,000	college	yes	Clinton	$\cdots$
57	?	NY	$19,000	high school	yes	?	$\cdots$
?	M	CA	$102,000	masters	no	Trump	$\cdots$
41	F	NV	$23,000	?	yes	Trump	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

goals:

- detect demographic groups?
- find typical responses?
- identify related features?
- impute missing entries?

Data table

m examples (patients, respondents, assets)
n features (tests, questions, performance indicators)

$$\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}$$

- ith row of A is feature vector for ith example
- jth column of A gives values for jth feature across all examples

Outline

Low rank model

given: $m \times n$ data table A, $k \ll m, n$
find: $X \in \mathbf{R}^{m \times k}$, $Y \in \mathbf{R}^{k \times n}$ for which

$$\begin{bmatrix} \\ X \\ \\ \end{bmatrix} \begin{bmatrix} & Y & \end{bmatrix} \approx \begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix}$$

i.e., $x_i y_j \approx A_{ij}$, where

$$\begin{bmatrix} \\ X \\ \\ \end{bmatrix} = \begin{bmatrix} \text{—} x_1 \text{—} \\ \vdots \\ \text{—} x_m \text{—} \end{bmatrix} \qquad \begin{bmatrix} & Y & \end{bmatrix} = \begin{bmatrix} | & & | \\ y_1 & \cdots & y_n \\ | & & | \end{bmatrix}$$

Low rank model

given: $m \times n$ data table A, $k \ll m, n$
find: $X \in \mathbf{R}^{m \times k}$, $Y \in \mathbf{R}^{k \times n}$ for which

$$\begin{bmatrix} \\ X \\ \\ \end{bmatrix} \begin{bmatrix} \quad & Y & \quad \end{bmatrix} \approx \begin{bmatrix} & & \\ \quad & A & \quad \\ & & \end{bmatrix}$$

i.e., $x_i y_j \approx A_{ij}$, where

$$\begin{bmatrix} \\ X \\ \\ \end{bmatrix} = \begin{bmatrix} \text{—} x_1 \text{—} \\ \vdots \\ \text{—} x_m \text{—} \end{bmatrix} \qquad \begin{bmatrix} \quad & Y & \quad \end{bmatrix} = \begin{bmatrix} | & & | \\ y_1 & \cdots & y_n \\ | & & | \end{bmatrix}$$

interpretation:

- X and Y are (compressed) representation of A

Low rank model

given: $m \times n$ data table A, $k \ll m, n$
find: $X \in \mathbf{R}^{m\times k}$, $Y \in \mathbf{R}^{k\times n}$ for which

$$\begin{bmatrix} \\ X \\ \\ \end{bmatrix} \begin{bmatrix} & Y & \end{bmatrix} \approx \begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix}$$

i.e., $x_i y_j \approx A_{ij}$, where

$$\begin{bmatrix} \\ X \\ \\ \end{bmatrix} = \begin{bmatrix} -x_1- \\ \vdots \\ -x_m- \end{bmatrix} \qquad \begin{bmatrix} & Y & \end{bmatrix} = \begin{bmatrix} | & & | \\ y_1 & \cdots & y_n \\ | & & | \end{bmatrix}$$

interpretation:

- X and Y are (compressed) representation of A
- $x_i^T \in \mathbf{R}^k$ is a point associated with example i
- $y_j \in \mathbf{R}^k$ is a point associated with feature j

Low rank model

given: $m \times n$ data table A, $k \ll m, n$
find: $X \in \mathbf{R}^{m \times k}$, $Y \in \mathbf{R}^{k \times n}$ for which

$$\begin{bmatrix} \\ X \\ \\ \end{bmatrix} \begin{bmatrix} & Y & \end{bmatrix} \approx \begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix}$$

i.e., $x_i y_j \approx A_{ij}$, where

$$\begin{bmatrix} \\ X \\ \\ \end{bmatrix} = \begin{bmatrix} \text{—}x_1\text{—} \\ \vdots \\ \text{—}x_m\text{—} \end{bmatrix} \qquad \begin{bmatrix} & Y & \end{bmatrix} = \begin{bmatrix} | & & | \\ y_1 & \cdots & y_n \\ | & & | \end{bmatrix}$$

interpretation:

- X and Y are (compressed) representation of A
- $x_i^T \in \mathbf{R}^k$ is a point associated with example i
- $y_j \in \mathbf{R}^k$ is a point associated with feature j
- inner product $x_i y_j$ approximates A_{ij}

Why?

- reduce storage; speed transmission
- understand (visualize, cluster)
- remove noise
- infer missing data
- simplify data processing

Principal components analysis

PCA: for $A \in \mathbf{R}^{m \times n}$,

$$\text{minimize} \quad \|A - XY\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n (A_{ij} - x_i y_j)^2$$

with variables $X \in \mathbf{R}^{m \times k}$, $Y \in \mathbf{R}^{k \times n}$

- old roots (Pearson 1901, Hotelling 1933)
- least squares low rank fitting
- (analytical) solution via SVD of $A = U \Sigma V^T$
- (numerical) solution via alternating minimization

Generalized low rank model

$$\text{minimize} \quad \sum_{(i,j)\in\Omega} L_j(x_i y_j, A_{ij})$$

- loss functions L_j for each column
  - *e.g.*, different losses for reals, booleans, categoricals, ordinals, ...
- observe only $(i, j) \in \Omega$ (other entries are missing)

Note:

- can be (NP-)hard to optimize exactly
- alternating minimiziation still works well

Matrix completion

observe A_{ij} only for $(i,j) \in \Omega \subset \{1,\ldots,m\} \times \{1,\ldots,n\}$

$$\text{minimize} \quad \sum_{(i,j)\in\Omega}(A_{ij} - x_i y_j)^2 + \gamma\|X\|_F^2 + \gamma\|Y\|_F^2$$

two regimes:

- **some entries missing:** don't waste data; "borrow strength" from entries that are *not* missing
- **most entries missing:** matrix completion still works!

Losses

$$\text{minimize} \quad \sum_{(i,j)\in\Omega} L_j(x_i y_j, A_{ij}) + \sum_{i=1}^m r_i(x_i) + \sum_{j=1}^n \tilde{r}_j(y_j)$$

choose loss $L(u, a)$ adapted to data type:

data type	**loss**	$L(u, a)$
real	quadratic	$(u-a)^2$
real	absolute value	$\|u-a\|$
real	huber	$\mathbf{huber}(u-a)$
boolean	hinge	$(1-ua)_+$
boolean	logistic	$\log(1+\exp(-au))$
integer	poisson	$\exp(u) - au + a\log a - a$
ordinal	ordinal hinge	$\sum_{a'=1}^{a-1}(1-u+a')_+ + \sum_{a'=a+1}^{d}(1+u-a')_+$
categorical	one-vs-all	$(1-u_a)_+ + \sum_{a'\neq a}(1+u_{a'})_+$
categorical	multinomial logit	$\frac{\exp(u_a)}{\sum_{a'=1}^d \exp(u_{a'})}$

Implementations

find code to fit GLRMs in

- Python (serial)
- Julia (shared memory parallel)
- Spark (parallel distributed)
- H20 (parallel distributed)

Implementations

find code to fit GLRMs in

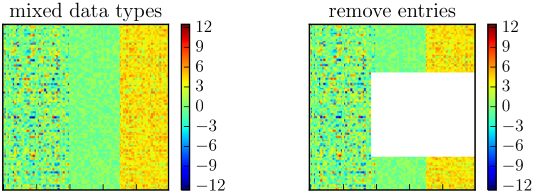
- Python (serial)
- Julia (shared memory parallel)
- Spark (parallel distributed)
- H20 (parallel distributed)

example: (Julia) fit rank 5 GLRM in 2 lines of code:

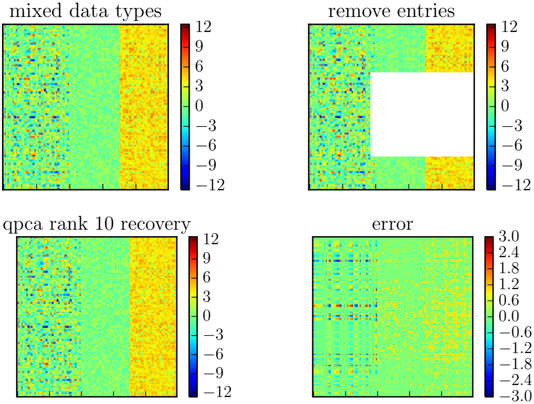
```
glrm, labels = GLRM(A, 5)
X,Y = fit!(glrm)
```

Outline

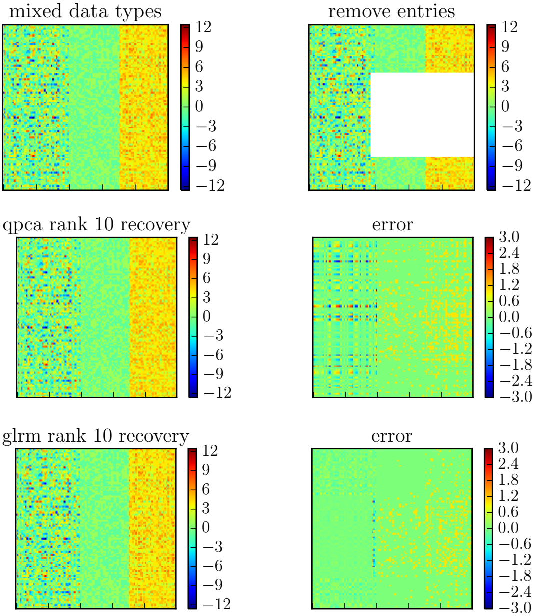
Impute heterogeneous data



Impute heterogeneous data

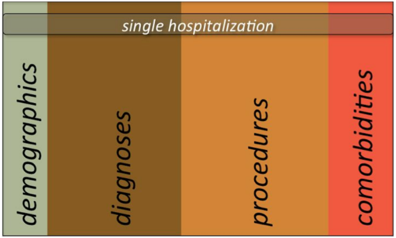


Impute heterogeneous data

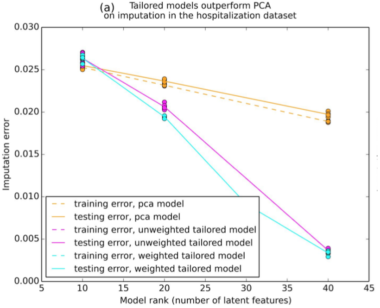


Hospitalizations are low rank

hospitalization data set



GLRM outperforms PCA



(Schuler Liu, Wan, Callahan, U, Stark, Shah 2016)

American community survey

2013 ACS:

- 3M respondents, 87 economic/demographic survey questions
  - income
  - cost of utilities (water, gas, electric)
  - weeks worked per year
  - hours worked per week
  - home ownership
  - looking for work
  - use foodstamps
  - education level
  - state of residence
  - ...
- 1/3 of responses missing

American community survey

most similar features (in *demography space*):

- Alaska: Montana, North Dakota
- California: Illinois, cost of water
- Colorado: Oregon, Idaho
- Ohio: Indiana, Michigan
- Pennsylvania: Massachusetts, New Jersey
- Virginia: Maryland, Connecticut
- Hours worked: weeks worked, education

Low rank models for dimensionality reduction[1]

U.S. Wage & Hour Division (WHD) compliance actions:

company	zip	violations	⋯
Holiday Inn	14850	109	⋯
Moosewood Restaurant	14850	0	⋯
Cornell Orchards	14850	0	⋯
Lakeside Nursing Home	14850	53	⋯
⋮	⋮	⋮	

- 208,806 rows (cases) $\times$ 252 columns (violation info)
- 32,989 zip codes...

[1] labor law violation demo: `https://github.com/h2oai/h2o-3/blob/master/h2o-r/demos/rdemo.census.labor.violations.large.R`

Low rank models for dimensionality reduction

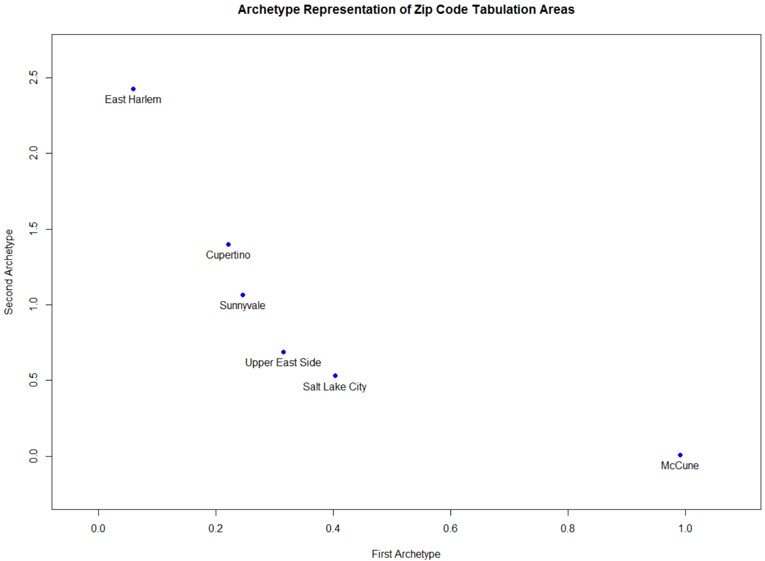
ACS demographic data (US census bureau):

zip	unemployment	mean income	$\cdots$
94305	12%	\$47,000	$\cdots$
06511	19%	\$32,000	$\cdots$
60647	23%	\$23,000	$\cdots$
94121	4%	\$178,000	$\cdots$
$\vdots$	$\vdots$	$\vdots$	

- 32,989 rows (zip codes) $\times$ 150 columns (demographic info)
- GLRM embeds zip codes into (low dimensional) *demography space*

Low rank models for dimensionality reduction

Zip code features:



Low rank models for dimensionality reduction

build 3 sets of features to predict violations:

- categorical: expand zip code to categorical variable
- concatenate: join tables on zip
- GLRM: replace zip code by low dimensional zip code features

fit a supervised (deep learning) model:

method	train error	test error	runtime
categorical	0.2091690	0.2173612	23.7600000
concatenate	0.2258872	0.2515906	4.4700000
GLRM	0.1790884	0.1933637	4.3600000

Outline

The trouble with polls

Q: are people who respond to polls like people who don't?

The trouble with polls

Q: are people who respond to polls like people who don't?
A: no:

> *There is a 19-year-old black man in Illinois who has no idea of the role he is playing in this election.*
>
> *He is sure he is going to vote for Donald J. Trump.*
>
> *In some polls, he's weighted as much as 30 times more than the average respondent, and as much as 300 times more than the least-weighted respondent.*

`http://www.nytimes.com/2016/10/13/upshot/how-one-19-year-old-illinois-man-is-distorting-national-p`
`html`

Correct biased sample

two types of people

- type A always fill out all questions
- type B leave question 3 blank half the time

question 1	question 2	question 3	question 4	$\cdots$
2.7	yes	4	yes	$\cdots$
9.2	no	?	no	$\cdots$
2.7	yes	4	yes	$\cdots$
9.2	no	1	no	$\cdots$
9.2	no	1	no	$\cdots$
9.2	no	?	no	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	

estimate population mean of question 3

- excluding missing entries: 2.5
- imputing missing entries: 2

Correct biased sample

two types of people

- type A always fill out all questions
- type B leave question 3 blank half the time

question 1	question 2	question 3	question 4	$\cdots$
2.7	yes	4	yes	$\cdots$
9.2	no	?	no	$\cdots$
2.7	yes	4	yes	$\cdots$
9.2	no	1	no	$\cdots$
9.2	no	1	no	$\cdots$
9.2	no	?	no	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	

estimate population mean of question 3 if
type B people have two subtypes:

- one that answers "1" to question 3
- another that doesn't answer, but whose true answer is "27"

How does this apply to election models?

simple model: suppose that in each demographic group,

- there are some Trump and some Clinton supporters
- the Trump supporters respond to pollsters at lower rates (or lie about their support)

there is *no way* to detect this from polling data!

How does this apply to election models?

simple model: suppose that in each demographic group,

- there are some Trump and some Clinton supporters
- the Trump supporters respond to pollsters at lower rates (or lie about their support)

there is *no way* to detect this from polling data!

n.b. confidence intervals (as usually computed)

- account for *statistical* error
- do not account for *systematic* error

Dealing with systematic bias

for correct estimation, need *outcome* to be independent of *missingness* conditional on *covariates*

$$\text{support for Trump} \perp \text{non-response} \mid \text{demographics}$$

Dealing with systematic bias

for correct estimation, need *outcome* to be independent of *missingness* conditional on *covariates*

$$\text{support for Trump} \perp \text{non-response} \mid \text{demographics}$$

problem with systematic bias:

- even if you know it *exists*, you don't know *how much*!

Dealing with systematic bias

for correct estimation, need *outcome* to be independent of *missingness* conditional on *covariates*

$$\text{support for Trump} \perp \text{non-response} \mid \text{demographics}$$

problem with systematic bias:

- even if you know it *exists*, you don't know *how much*!
- modeling systemic bias?

Summary

generalized low rank models

- fill in missing data
- handle huge, heterogeneous data coherently
- transform big messy data into small clean data

paper:
`http://arxiv.org/abs/1410.0342`

code:
`https://github.com/madeleineudell/LowRankModels.jl`