Recommenders, Topics, and Text

Susan Athey and Guido Imbens, Stanford University
NBER Methods Lectures, 2015

Introduction

- **Netflix Prize:**
  Beat Netflix recommender system, using Netflix data →
  Win $1 million
- **Data:**
  480,000 users
  18,000 movies
  100 million observed ratings = only 1.1% of ratings observed

Recommendation Systems

- Predict whether a user will like an item (or how much: here binary choice for simplicity)
- Scenarios: what is observed?
  - Lots of users and lots of items
  - Observable item characteristics
  - Matrix of users x choices with 1's (or ratings) for items consumer likes/chose in past

- Economics Literature
  - Obvious connection to discrete choice modeling
  - Multiple choices per user
  - Small literature incorporating multiple choices: voting records, Neilsen TV viewing, school choice, small supermarket bundles

- ML Literature
  - Massive literature with lots of disconnected subliteratures: too many to cite
  - Netflix challenge results (and almost all contests): average of hundreds of models
  - Big strength: incorporating multiple sparse choice data at scale
  - Focus on prediction, not interpretability or counterfactual prediction, but some approaches deliver interpretability and parameters that look like primitives

- For economic applications:
  - ML methods have a lot of promise for large scale choice modeling, but more work to tailor them to economist's needs/tastes. One strand of literature builds models with latent variables that connect closely to structural economic models.

Strands of Recommendation Literature

"Unsupervised": Item Similarity

- Suppose observe item characteristics
- Recommend similar items to what the user likes
- Use unsupervised learning techniques to reduce dimensionality (see earlier lectures)

"Supervised": Collaborative Filtering

- Suppose observe choice data but limited item characteristics
- Find other users with similar tastes
- Recommend items liked by similar users
- E.g. Matrix decomposition

Hierarchical/Graphical Models

Bringing in User Choice Data

- It can be useful to model item cluster/topics SEPARATELY from user data if:
  - You want to make predictions when the context changes
  - Website may change what types of articles it displays together; journal may do special issues or article groupings

- It can be useful to model topics together with user data if:
  - You want to make the best possible recommendations in a stable context

Collaborative Filtering

Items

Users

0	0	0	0	1	0	0	1
1	0	0	1	1	1	0	0
0	1	0	0	1	1	0	1
0	0	0	1	1	1	0	?
0	0	1	1	1	0	0	?
0	1	0	0	1	0	1	?
0	1	0	0	0	1	0	?

Goal: predict probabilities that users choose item

Note: in Netflix challenge data, have ratings 1-5 but many items not rated at all.

Collaborative Filtering: Classification problem (e.g. logit, CART)

- Simple, familiar approach:
  - Use a logit or other classifier item by item
- Model: $\Pr(u_{ij}=1) = f(\text{indicators for other items user likes})$
- Approach doesn't work well with lots of items and few items chosen per user
  - Need to combine with other methods to reduce dimensionality of feature space
  - Item by item prediction is costly

K-nearest neighbor (KNN) Prediction

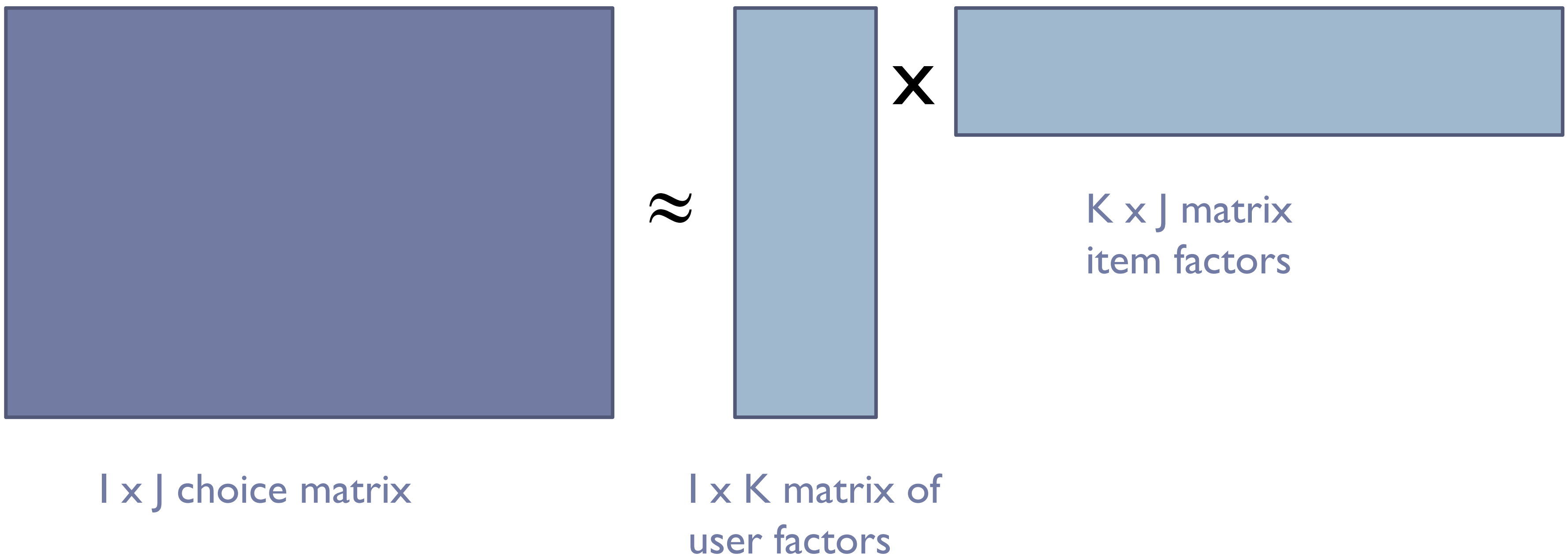
- To predict $\Pr(u_{ij}=1)$:
  - Compute similarity between item *j* and all other items
  - Find the *k* closest by similarity metric
  - Calculate weighted average of user choices of the neighbor items

- Similarity:
  - Measured in terms of user preference vectors
  - Cosine similarity: the angle between the vectors

- Very popular method! Easy and fast.

- Cons
  - With sparse matrix, may predict a lot of zeros
  - May not work very well with simple weights
  - More complex versions "learn"=estimate weights, but if you want to estimate a model, other models may work better

Collaborative Filtering: Matrix Factorization



Analyst picks K. Use various methods (that work at large scale) to find factor matrices that minimize mean squared error of predicts versus true choice matrix.

PROS: Easy and fast. Transparent. Results can be inspected and maybe interpreted.
CONS: Harder to link to a model and extend in various directions. No inference/hypothesis testing.

Other Approaches

- Iterative clustering
  - Can use any of a variety of unsupervised learning methods to find groups of similar users or items
  - Iterate:
    - Group items with similar user (category) vectors
    - Then group users with similar item (category) vectors

- Structural Models/Hierarchical Models
  - Specify latent factors and distribution they are drawn from
  - Use Bayesian methods or EM algorithm to estimate the latent factors
  - More principled way to do decomposition and incorporate user and item characteristics
  - Will return to this after brief digression…

Recommendations and Topic Modeling for Documents and Text

Collaborative Filtering versus Topic Models

Items
Words/Phrases

Users

Documents

0	0	0	0	1	0	0	1
1	0	0	1	1	1	0	0
0	1	0	0	1	1	0	1
0	0	0	1	1	1	0	?
0	0	1	1	1	0	0	?
0	1	0	0	1	0	1	?
0	1	0	0	0	1	0	?

Goal: predict probabilities that users choose item *by finding latent factors*
Goal: put documents into topics with similar collections of words

Finding topics of articles: Methods for clustering described above can be used as quick and dirty ways to put documents into clusters. Hierarchical models are more sophisticated.
Classifying articles (good/bad, left/right): Supervised learning, and particular classification methods, can be applied if some documents are labelled, e.g using human judges. Can combine unsupervised learning techniques for dimensionality reduction with classification techniques.

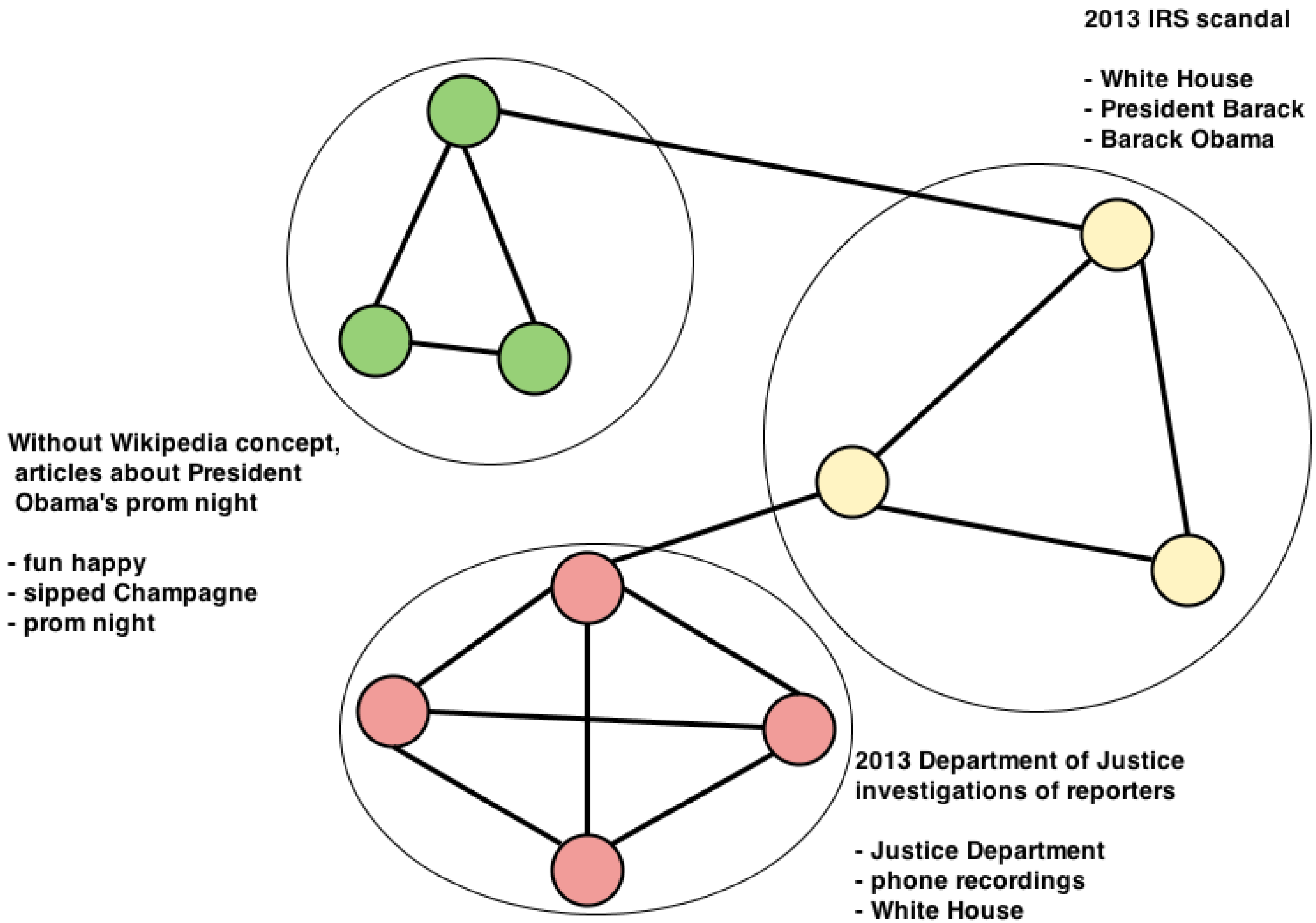
Classifying Large Number of Articles

- Take articles, have humans label as many as you can afford
- Build a classifier to classify unlabeled articles
  - See, e.g., Vowpal Wabbit
  - Use unsupervised learning for feature reduction
  - And/or, use method such as LASSO or other regularized ML methods with cross-validation to select model complexity, specification of features
  - Warning: this doesn't work as well as you might think w/ limited training data and many features. Needs tweaking in feature reduction.
- Expect to see this used widely in economics for text.
  - A number of applications already (e.g. political bias, review sentiment, etc. See Gentzkow/Shapiro, Luca, etc.)
- Possible alternatives/improvements
  - Athey, Mobius, and Pal (in progress)
  - Use unsupervised learning PRIOR to human labeling
  - Put articles into topics, and then crowd source features of topics, also of articles within each topic (stratified)
  - For article-specific characteristics that don't apply to topic as a whole, train classifier (e.g. for left/right political bias)
    - Classifier works much better within a topic (since significance of words different by topic)
  - Can avoid asking about political bias of article about a sports event

Quick and Dirty Clustering for Topics: Network Methods

- Another approach to clustering is to interpret the document – word matrix as a network
- There are network methods for clustering or "community detection," some of which are different from matrix factorization
  - Approach based on "modularity" (see e.g. Newman) – greedy, scalable algorithm to maximize links within communities while penalizing links outside
- Athey, Mobius, and Pal (in progress) apply this to news articles
  - Strength of link between articles is determined by common words, as well as links to common Wikipedia articles

Network Classification Example



Topic Examples in April 2013

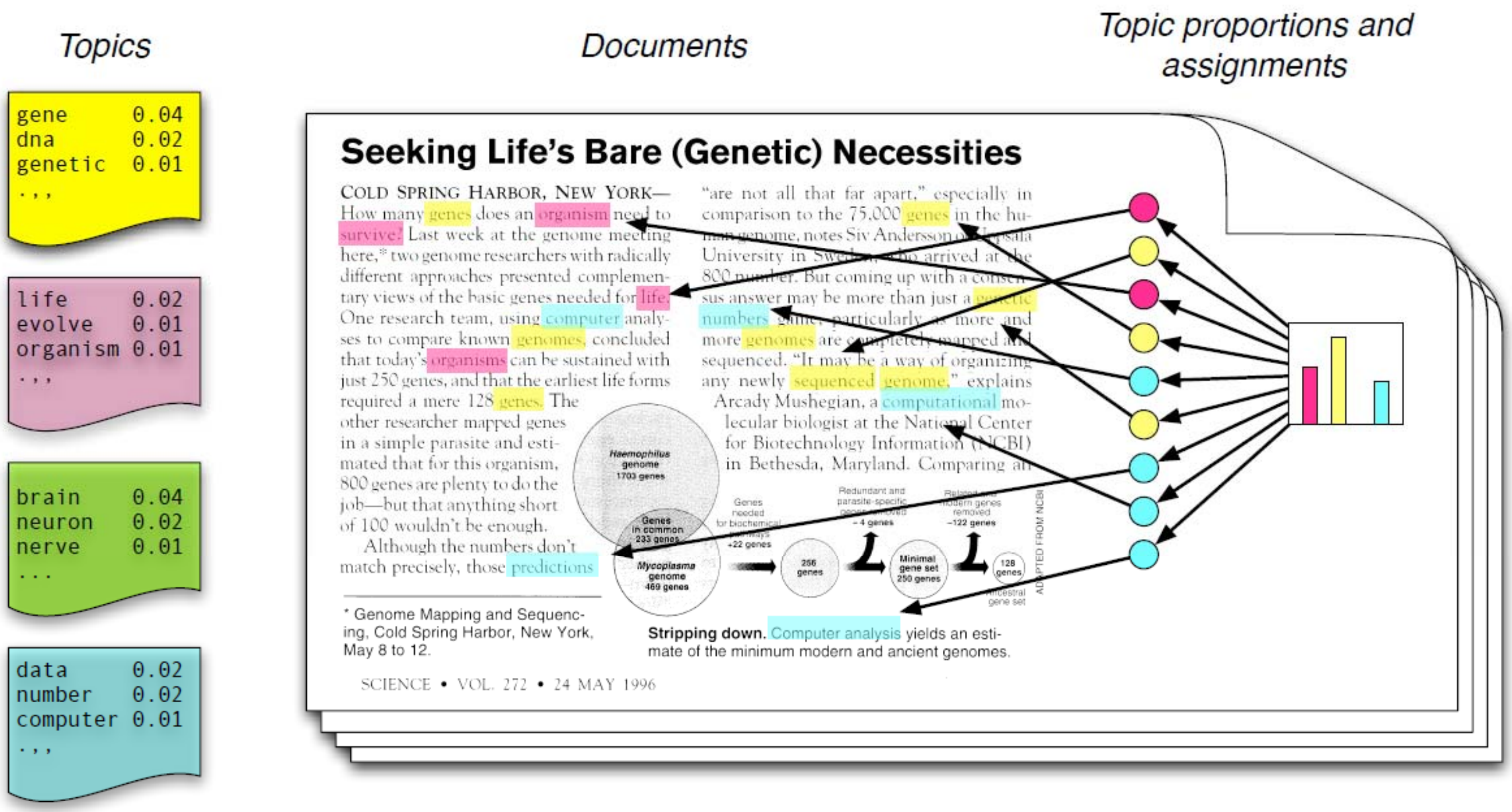
Boston Marathon bombings
2013 Korean crisis
2012–13 in English football
Ariel Castro kidnappings
2013 NFL season
2013 in baseball
2013 in American television
2013 NBA Finals
2013 in film
2012–13 NHL season
2013 Masters Tournament
2013 NASCAR Sprint Cup Series

2013 Moore tornado
Death of Lee Rigby
2012 Benghazi attack
Murder of Travis Alexander
List of school shootings in the United States
Death and funeral of Margaret Thatcher
Timeline of the Syrian civil war (May–August 2012)
2013 IRS scandal
NCAA Men's Division I Basketball Championship
Malaysian general election, 2013
Shooting of Trayvon Martin
Phil Mickelson

Hierarchical Models (Closer to Structural Models)

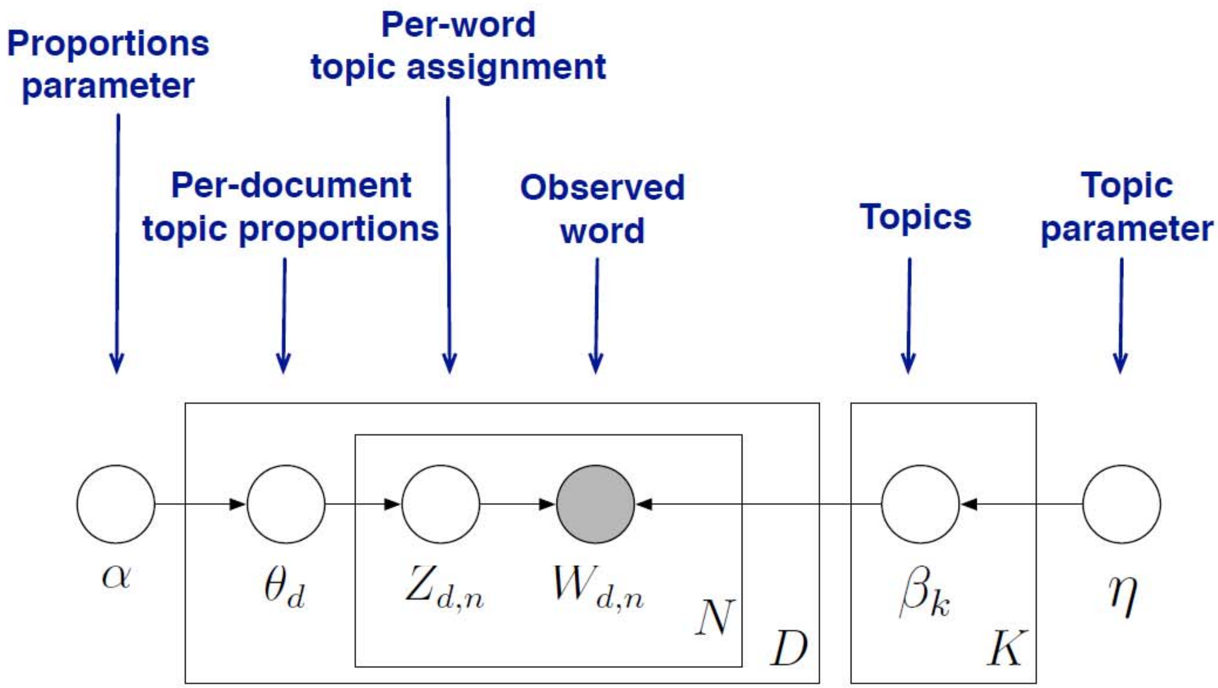
Many of these slides based on a tutorial by David Blei;
See David Blei's web page at Columbia for more details
http://www.cs.columbia.edu/~blei/

Latent Dirichlet Allocation (LDA)



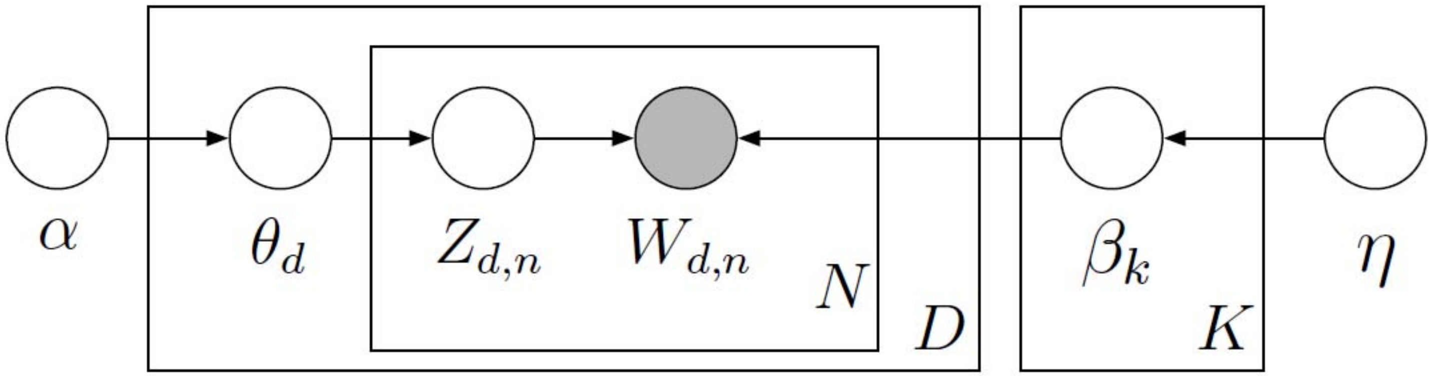
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

LDA As a Graphical Model



$$p(\beta, \theta, \mathbf{z}, \mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i \mid \eta) \right) \left(\prod_{d=1}^{D} p(\theta_d \mid \alpha) \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

Estimation and Inference of LDA



Approximate posterior inference algorithms

- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Distributed sampling (Newman et al., 2008; Ahmed et al., 2012)
- Collapsed variational inference (Teh et al., 2006)
- Online variational inference (Hoffman et al., 2010)
- Factorization based inference (Arora et al., 2012; Anandkumar et al., 2012)

Topics Estimated from Articles in *Science*

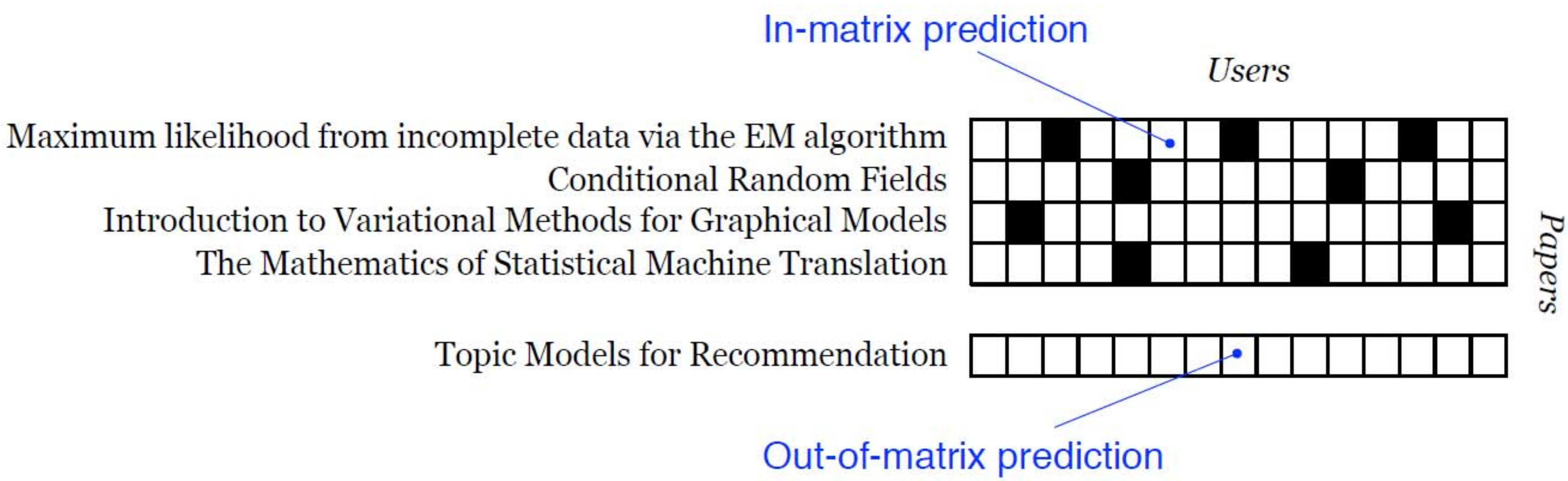
1	2	3	4	5
dna	protein	water	says	mantle
gene	cell	climate	researchers	high
sequence	cells	atmospheric	new	earth
genes	proteins	temperature	university	pressure
sequences	receptor	global	just	seismic
human	fig	surface	science	crust
genome	binding	ocean	like	temperature
genetic	activity	carbon	work	earths
analysis	activation	atmosphere	first	lower
two	kinase	changes	years	earthquakes

6	7	8	9	10
end	time	materials	dna	disease
article	data	surface	rna	cancer
start	two	high	transcription	patients
science	model	structure	protein	human
readers	fig	temperature	site	gene
service	system	molecules	binding	medical
news	number	chemical	sequence	studies
card	different	molecular	proteins	drug
circle	results	fig	specific	normal
letters	rate	university	sequences	drugs

11	12	13	14	15
years	species	protein	cells	space
million	evolution	structure	cell	solar
ago	population	proteins	virus	observations
age	evolutionary	two	hiv	earth
university	university	amino	infection	stars
north	populations	binding	immune	university
early	natural	acid	human	mass
fig	studies	residues	antigen	sun
evidence	genetic	molecular	infected	astronomers
record	biology	structural	viral	telescope

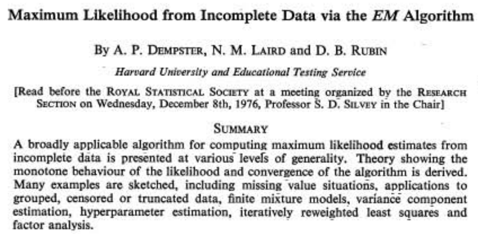
16	17	18	19	20
fax	cells	energy	research	neurons
manager	cell	electron	science	brain
science	gene	state	national	cells
aaas	genes	light	scientific	activity
advertising	expression	quantum	scientists	fig
sales	development	physics	new	channels
member	mutant	electrons	states	university
recruitment	mice	high	university	cortex
associate	fig	laser	united	neuronal
washington	biology	magnetic	health	visual

Making Recommendations with User Choice Data and Content Data



User and Content Data

- Consider EM (Dempster et al., 1977). The text lets us estimate its topics:



Maximum Likelihood from Incomplete Data via the *EM* Algorithm

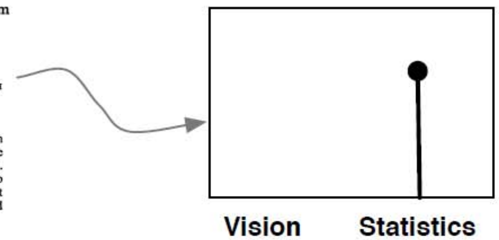
By A. P. Dempster, N. M. Laird and D. B. Rubin

Harvard University and Educational Testing Service

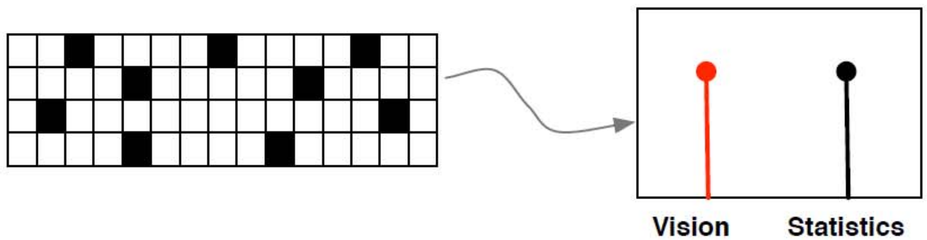
[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 8th, 1976, Professor S. D. Silvey in the Chair]

Summary

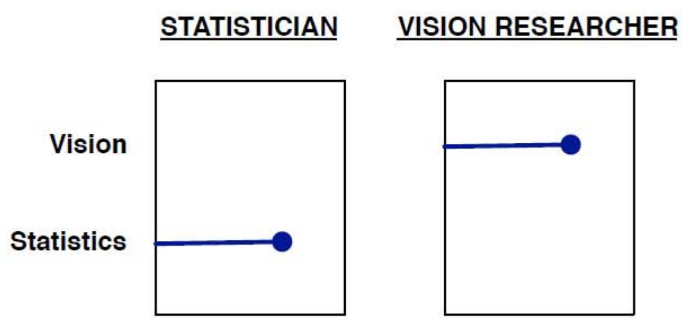
A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.



- With user data, we adjust the topics to account for who liked it:

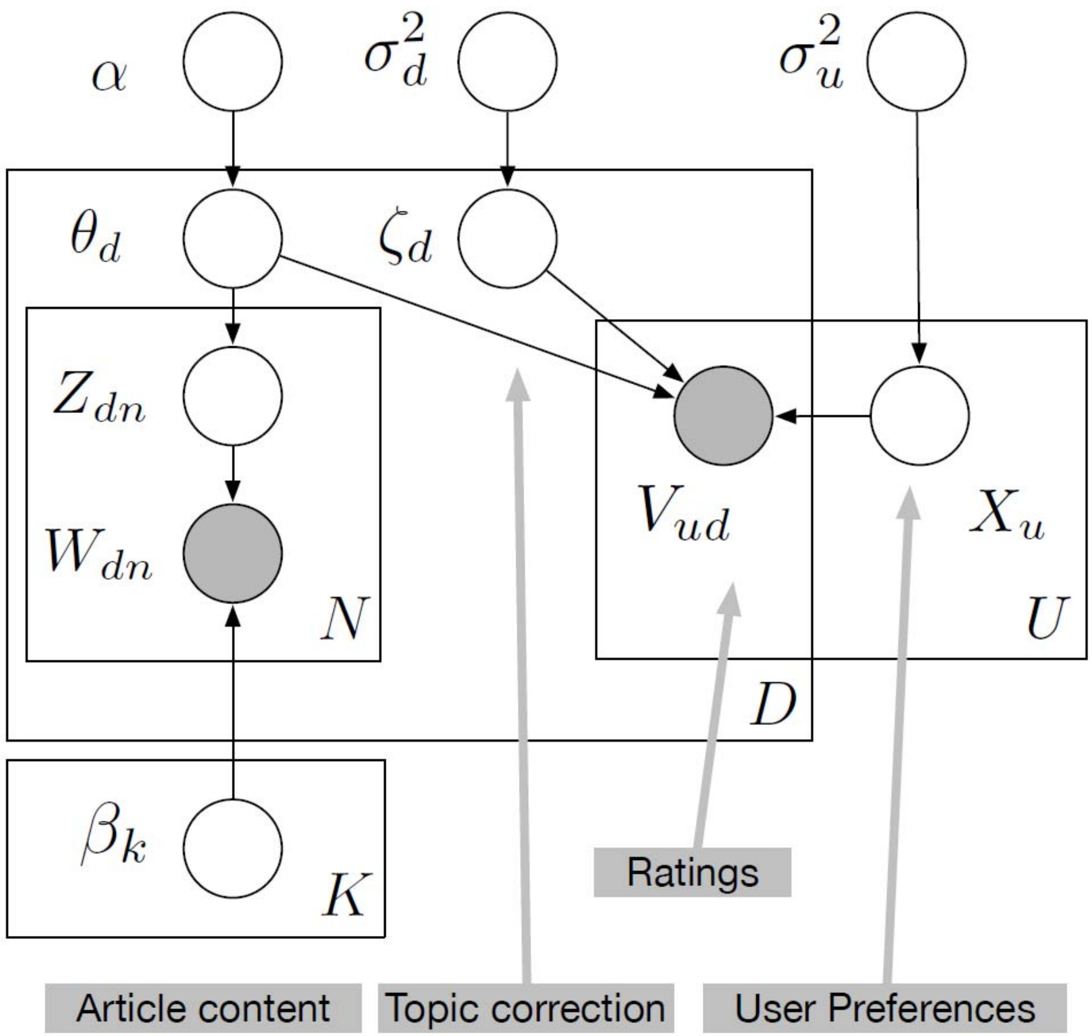


- We can then recommend to users:



Complex Model with Latent Variables for Topics and User Preferences

Grey circles: observables



Conclusions

- Recommendation systems literature is close to demand literature in goals
  - ML Literature is massive, disconnected, ad hoc, and hard to absorb
- Quick and dirty techniques can be a substitute for complex models for both recommendation systems and for topic modeling
  - Clustering methods
  - Matrix decomposition techniques give a sense of what would come out of a latent variable model with much less work
  - Network literature also has some fast and easy methods
- One branch of ML (graphical models) is very close to structural models with unobserved product characteristics
  - Topic models help understand structure underlying documents, and can be used as part of user choice models
  - Open questions for economists include identification and mapping onto utility framework

- Applications/directions
  - My work with Jeno Pal and Markus Mobius looks empirically at role of aggregators and intermediaries on what articles people choose
  - With David Blei, Jake Hofman, and Markus Mobius, looks methodologically at how to link topic models to structural IO models and incorporate role of availability and prominence of what is shown

Machine Learning and Causal Inference

Susan Athey and Guido Imbens, Stanford University

NBER Lectures, 2015

Introduction

Supervised Machine Learning v. Econometrics/Statistics Lit. on Causality

Supervised ML

- Well-developed and widely used nonparametric prediction methods that work well with big data
  - Used in technology companies, computer science, statistics, genomics, neuroscience, etc.
  - Rapidly growing in influence
- Cross-validation for model selection
- Focus on prediction and applications of prediction
- Weaknesses
  - Causality (with notable exceptions, e.g. Pearl, but not much on data analysis)

Econometrics/Soc Sci/Statistics

- Formal theory of causality
  - Potential outcomes method (Rubin) maps onto economic approaches
- "Structural models" that predict what happens when world changes
  - Used for auctions, anti-trust (e.g. mergers) and business decision-making (e.g. pricing)
- Well-developed and widely used tools for estimation and inference of causal effects in exp. and observational studies
  - Used by social science, policy-makers, development organizations, medicine, business, experimentation
- Weaknesses
  - Non-parametric approaches fail with many covariates
  - Model selection unprincipled

Lessons for Economists

Engineering approach

- Methods that scale
- Asymptotic normality of estimates or predictions for hypothesis testing not important goal
- Lots of incremental improvements in algorithms, judged by performance at prediction
- Formal theory and perfect answers not required: "it works"

More systematic in key respects

- Cross-validation for model selection

Low hanging fruit

- Model selection/variable selection for exogenous covariates, prediction component of model
- Heterogeneity
  - Heterogeneous treatment effects/elasticities
  - Personalized recommendations based on estimates
- Some specific areas
  - Recommendation systems
  - Topic modeling
  - Text analysis/classifiers

Causal Inference

A Research Agenda on Causal Inference

Problems

- Many problems in social sciences entail a combination of prediction and causal inference
- Existing ML approaches to estimation, model selection and robustness do not directly apply to the problem of estimating causal parameters
- Inference more challenging for some ML methods

Proposals

- Formally model the distinction between causal and predictive parts of the model and treat them differently for both estimation and inference
  - Abadie, Athey, Imbens and Wooldridge (2014, under review; also work in progress)
- Develop new estimation methods that combine ML approaches for prediction component of models with causal approaches
  - Athey-Imbens (2015, work in progress)
- Develop new approaches to cross-validation optimized for causal inference and optimal policy estimation
  - Athey-Imbens (2015, work in progress)
- Develop robustness measures for causal parameters inspired by ML
  - Athey-Imbens (*AER P&P* 2015; work in progress)
- Develop methods for causal inference for network analysis drawing on CS tools for networks
  - Athey-Eckles-Imbens (2015)
- Large scale structural models with latent variables
  - Athey-Nekipelov (2012, 2015); Athey, Blei, Hofman, Mobius (in progress)

Model for Causal Inference

- For causal questions, we wish to know what would happen if a policy-maker changes a policy
  - Potential outcomes notation:
    - $Y_i(w)$ is the outcome unit i would have if assigned treatment w
    - For binary treatment, treatment effect is $\tau_i = Y_i(1) - Y_i(0)$
  - Administer a drug, change minimum wage law, raise a price
  - Function of interest: mapping from alt. CF policies to outcomes
  - Holland: Fundamental Problem of Causal Inference
    - We do not see the same units at the same time with alt. CF policies

- Units of study typically have fixed attributes x_i
  - These would not change with alternative policies
  - E.g. we don't contemplate moving coastal states inland when we change minimum wage policy

Causal Inference Versus Prediction

When is Prediction Primary Focus?

- Economics: "allocation of scarce resources"
- An allocation is a decision.
  - Generally, optimizing decisions requires knowing the counterfactual payoffs from alternative decisions.
- Hence: intense focus on causal inference in applied economics

- Examples where prediction plays the dominant role in causal inference
  - Decision is obvious given an unknown state
  - Many decisions hinge on a prediction of a future state
  - Prediction dominant for a component of causal inference
    - Propensity score estimation
    - First stage of IV/2SLS
    - Predicting the baseline in difference in difference settings
    - Predicting the baseline in time series settings

Prediction and Decision-Making: Predicting a State Variable

Kleinberg, Ludwig, Mullainathan, and Obermeyer (2015)

- Motivating examples:
  - Will it rain? (Should I take an umbrella?)
  - Which teacher is best? (Hiring, promotion)
  - Unemployment spell length? (Savings)
  - Risk of violation of regulation (Health inspections)
  - Riskiest youth (Targeting interventions)
  - Creditworthiness (Granting loans)
- Empirical applications:
  - Will defendant show up for court? (Should we grant bail?)
  - Will patient die within the year? (Should we replace joints?)

A formal model

- Payoff Y_i is, for all i, known function of policy (W_i) and state of the world (S)
$$Y_i = \pi(W_i, S)$$
- State of the world may depend on policy choice
- Then, the impact of changing policy is
$$\frac{\partial}{\partial W_i} Y_i = \frac{\partial}{\partial W_i}\pi(W_i, S) + \frac{\partial}{\partial S}\pi(W_i, S) \cdot \frac{\partial S}{\partial W_i}$$
  - Paper refers to second term as "causal component"
    - Argue that taking an umbrella doesn't effect rain, so the main problem is predicting rain
  - But in general $\frac{\partial}{\partial W_i}\pi$ is unknown/heterogeneous, as is π – can also think of that as the causal effect
  - But idea still carries over if knowing S tells you the sign of $\frac{\partial}{\partial W_i}\pi$

Application: Joint Replacements

- Methods:
  - Regularized logistic regression, choosing penalty parameter for number of covariates using 10-fold c-v
- Data
  - 65K Medicare patients
  - 3305 variables and 51 state dummies

TABLE 1—RISKIEST JOINT REPLACEMENTS

Predicted mortality percentile	Observed mortality rate	Futile procedures averted	Futile spending ($ mill.)
1	0.435 (0.028)	1,984	30
2	0.422 (0.028)	3,844	58
5	0.358 (0.027)	8,061	121
10	0.242 (0.024)	10,512	158
20	0.152 (0.020)	12,317	185
30	0.136 (0.019)	16,151	242

Columns(3) and (4) show results of a simulation exercise: we identify a population of eligibles (using published Medicare guidelines: those who had multiple visits to physicians for osteoarthritis and multiple claims for physical therapy or therapeutic joint injections) who did not receive replacement and assign them a predicted risk. We then substitute the high risk surgeries in each row with patients from this eligible distribution for replacement, starting at median predicted risk. Column (3) counts the futile procedures averted (i.e., replaced with non-futile procedures) and (4) quantifies the dollars saved in millions by this substitution.