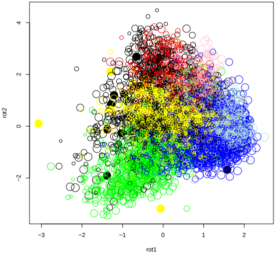


Fast Linear methods for linguistics (Eigenword-based language models from large corpora)

Dean P. Foster

Probably better title for talk:

> **"Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions"**

- Unfortunately it was written by me.
- by Halko, Martinsson, and Tropp.
- But it is currently my favorite paper.

Mathematical summary

problem	Find a low rank approximation to a matrix M.
solution	Multiply a random matrix times M and "clean it up."

Mathematical summary

problem Find a low rank approximation to a matrix M.

solution Multiply a random matrix times M and "clean it up."

outline Applying this method to words:

- bilinear: defining eigenwords
- trilinear: applying eigenwords to HMM
- tetralinear: clustering

BI-LINEAR: CCA for Semi-supervised data

CCA: Usual data table for data mining

$$\begin{bmatrix} Y \\ \\ (n \times 1) \\ \\ \end{bmatrix} \begin{bmatrix} X \\ \\ (n \times p) \\ \\ \end{bmatrix}$$

with $p \gg n$

With unlabeled data

m rows of unlabeled data:

$$\begin{bmatrix} Y \\ n \times 1 \\ \end{bmatrix} \quad \begin{bmatrix} X \\ (n+m) \times p \\ \\ \\ \end{bmatrix}$$

With alternative X's

m rows of unlabeled data, and two sets of equally useful X's:

$$\begin{bmatrix} Y \\ \\ n \times 1 \\ \\ \end{bmatrix} \quad \begin{bmatrix} X \\ \\ (n+m) \times p \\ \\ \\ \\ \\ \\ \end{bmatrix} \quad \begin{bmatrix} Z \\ \\ (n+m) \times p \\ \\ \\ \\ \\ \\ \end{bmatrix}$$

With: $m \gg n$

Examples

- Named entity recognition
  - Y = person / place
  - X = name itself
  - Z = words before target
- Topic identification (medline)
  - Y = topic
  - X = abstract
  - Z = text
- Speaker identification:
  - Y = which character is speaking in a sitcom
  - X = sound track
  - Z = video

What if we run CCA on X and Z?

CCA = canonical correlation analysis

- Find the directions that are most highly correlated
- Close to PCA (principal components analysis)

What if we run CCA on X and Z?

CCA = canonical correlation analysis

- Find the directions that are most highly correlated
- Close to PCA (principal components analysis)
- The Main Result

Theorem (Foster and Kakade, '06)

Let $\hat{\beta}$ be the Ridge regression estimator with weights induced by the CCA. Then

$$Risk(\hat{\beta}) \leq \left(5\alpha + \frac{\sum \lambda_i^2}{n}\right)\sigma^2$$

What if we run CCA on X and Z?

CCA = canonical correlation analysis

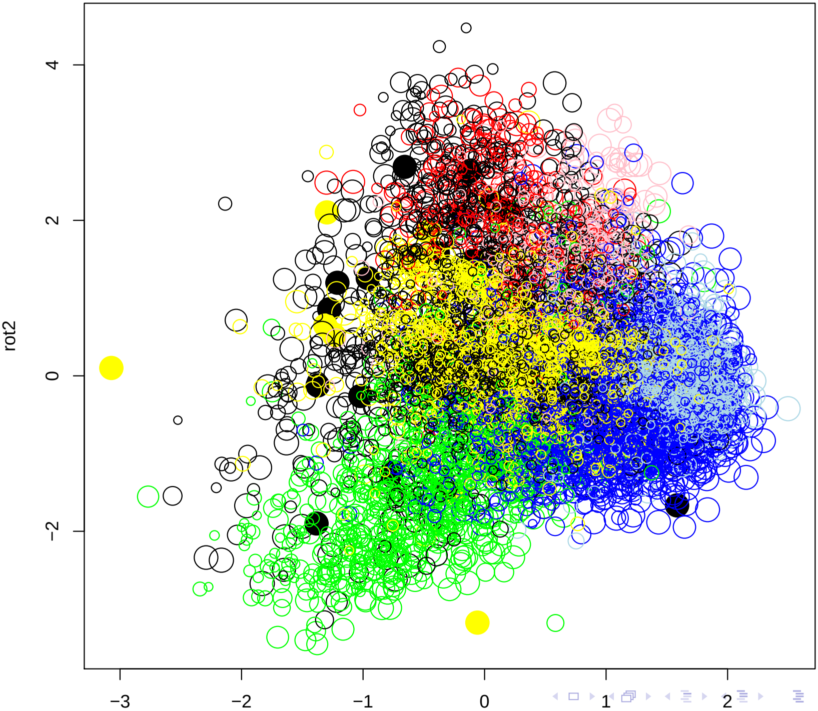
- Find the directions that are most highly correlated
- Close to PCA (principal components analysis)
- The Main Result

Theorem (Foster and Kakade, '06)

Let $\hat{\beta}$ be the Ridge regression estimator with weights induced by the CCA. Then

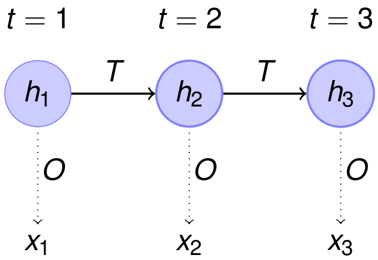
$$Risk(\hat{\beta}) \leq \left(5\alpha + \frac{\sum \lambda_i^2}{n}\right)\sigma^2$$

- Let's take a quick look at such CCA variables


rot2
4
2
0
−2
−3
−2
−1
0
1
2

TRI-LINEAR: HMMs

The Hidden Markov Model



Reduced Dimension Hidden Markov Model

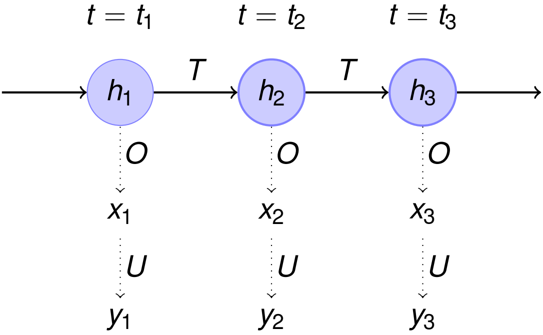


Figure: The Y's are our eigenfeatures.

Reduced Dimension Hidden Markov Model

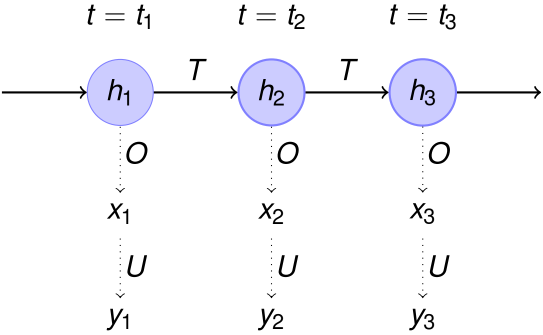


Figure: The Y's are our eigenfeatures.

$$\Pr(x_t, \ldots, x_1) = 1^T T \mathrm{diag}(OU^\top y_t) \cdots T \mathrm{diag}(OU^\top y_1) \pi$$

HMM theorem

Theorem (with Rodu, Ungar)

Let X_t be generated by an $m \geq 2$ state HMM. Suppose we are given a U which has the property that range(O) $\subset$ range(U) and $|U_{ij}| \leq 1$. Using N independent triples, we have

$$N \geq \frac{128 m^2 (2t+3)^2}{\epsilon^2 \; \Lambda^2 \sigma_m^4} \log\left(\frac{2m}{\delta}\right) \cdot \overbrace{\frac{\epsilon^2/(2t+3)^2}{\left(\sqrt[2t+3]{1+\epsilon} - 1\right)^2}}^{\approx 1}$$

implies that

$$1 - \epsilon \leq \left| \frac{\widehat{\Pr}(x_1, \ldots, x_t)}{\Pr(x_1, \ldots, x_t)} \right| \leq 1 + \epsilon$$

holds with probability at least $1 - \delta$.

Results on ConLL task

- Results on 2 NLP sequence labeling problems: NER (CoNLL '03 shared task) and Chunking (CoNLL '00 shared task).
- Trained on $\sim$ 65 million tokens of unlabeled text in a few hours!

Relative reduction in error over state-of-the-art:

Embedding/Model	NER	Chunking
C&W	15.0%	18.8%
HLBL	19.5%	20.2%
Brown	12.1%	18.7%
Ando+Zhang	5.6%	14.6%

Extending to trees

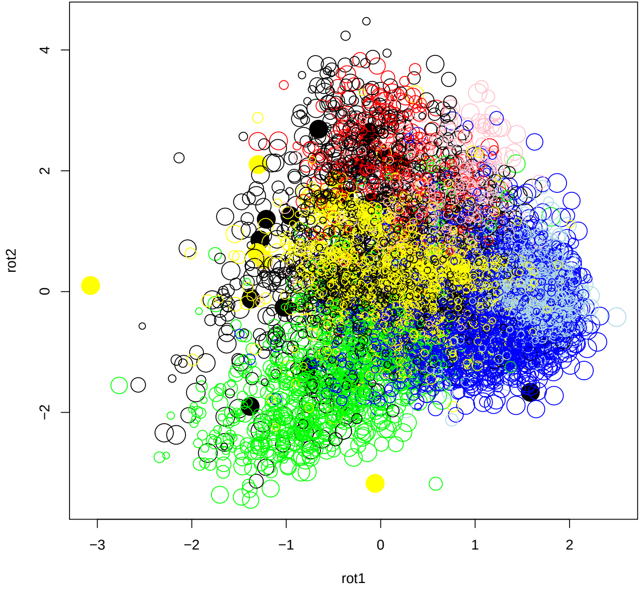
Theorem (with Rodu, Ungar, Dhillon, Collins)

Same as before–but for dependency parse trees.

Ran Dependency parsing on Penn Treebank

- Raw MST Parser is 91.8% accurate
- Adding eigenwords: 2.6% error reduction
- eigenwords plus Re-ranking: 7.3% error reduction

TETRA-LINEAR: Clustering



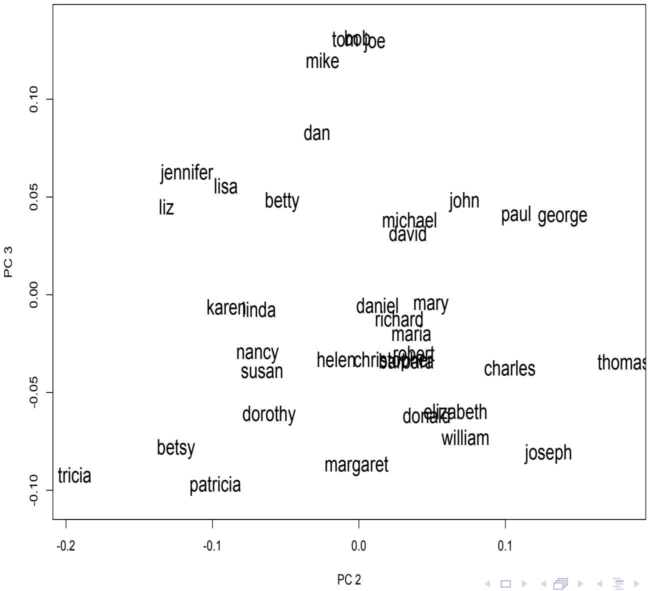
If you rotate this, you will see there are “pointy” directions

Clustering theorem

Theorem (with Hsu, Kakade, Liu, Anima, NIPS 2012)

Maximizing $E(\mu^\top X)^4$ will find the natural coordinate system for LDA.

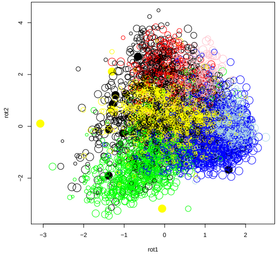
Sample direction sticking out of the main body of data



Conclusions

Linear methods:

- have nice math.
- are useful for real questions.
- are fast on computers.



Key

Colors:

- nouns = Blue (dark = NN1, light = NN2)
- verbs = red (dark = VV1, light = VV2)
- adj = green
- unk = yellow
- black = all else

Size = 1/Ziff order, top 50 are solid, rest are open.