Featurizing Text

Bob Stine
Dept of Statistics, Wharton School
University of Pennsylvania

Slides and draft manuscript available at
www-stat.wharton.upenn.edu/~stine

Thank you, NSF! (#1106743)
Thanks also to Dean Foster, Mark Liberman, and Trulia.

Wharton
Department of Statistics

Challenges in Data Analysis

- Election survey (ANES)
  - Predict voting behavior
  - Open ended responses to questions
- Medical outcomes
  - Predict health outcome based on
    Biometric data (weight, height, age, BP)
    Physician descriptions
  - Biometrics are 'easy' to use, but text?
- Real estate listings
  - Predict price from text in 7,400 listings
  - Suggest over-priced listings, identify comps

small dataset
for linguists

Wharton
Department of Statistics

Methodology

- Regression analysis
  - Flexible, familiar, well-understood
- Text is not well matched to regression
  - Regression designed for an ‘Excel table’
  - Columns of numbers
- Featurizing
  - Create the numerical Excel table
  - Emphasize ease-of-use rather than finding the best possible, domain-specific strategy
  - Three related methods that can be combined

Plan

- Regression models
- Featurizing for regression
  - New spin on existing methods
  - Novel aspects of empirical results
- Real estate example in detail
  - Cross validation
- Probability models
  - Topic models as explanation for success
- Discussion and plans

Wharton
Department of Statistics

Interpretation?

- Personal interest in prediction
    - Can use statistical tests to measure how well a model predicts, and to determine whether 'improvements' produce a better model.
    - What would it mean to find the right interpretation?
- By-and-large leave interpretation to others

Regression Analysis

Regression Model

- Typical data
  - Start with representative sample
  - Numerical data (category encoded as number)
- Build equation
  - Relate response to regressors, (y_i, X_i)

    regressor = predictor, explanatory variable, independent variable
  - Use differences in regressors to ‘explain’ simultaneous differences in response
  - Find weighted sum of regressors that is most correlated with response
- Prediction
  - Weighted average of regressors for new case

Wharton
Department of Statistics

Issues in Regression

- Which characteristics to use?
  - Substantive insight
  - Automated search
  - Everything
- How to separate wheat from chaff?
  - Statistical significance
    Everything passes this test with large samples
- Goodness of fit
  - R^2 is the percentage of 'explained' variation
    Adjusted R^2 compensates for size of model
  - Not appropriate for automated searches
    Over-fitting inflates R^2
  - Cross-validation: predict data you have not used

Wharton
Department of Statistics

Evaluating Coefficients

- Classical models
  - Handful of estimated coefficients
  - t-statistic compares observed statistic to 'null model' in which regressor has no effect
- Two issues in large models with big samples
  - t-statistic proportional to $\sqrt{\text{sample size}}$
    Regressors with tiny impact on predictions (small effect size) are 'statistically significant'
  - Multiplicity produces many apparently significant effects (statistics rewards persistence)
    Bonferroni threshold at $\approx\sqrt{2 \log (\#\text{regressors})}$

Wharton
Department of Statistics

Summarizing Model Estimates

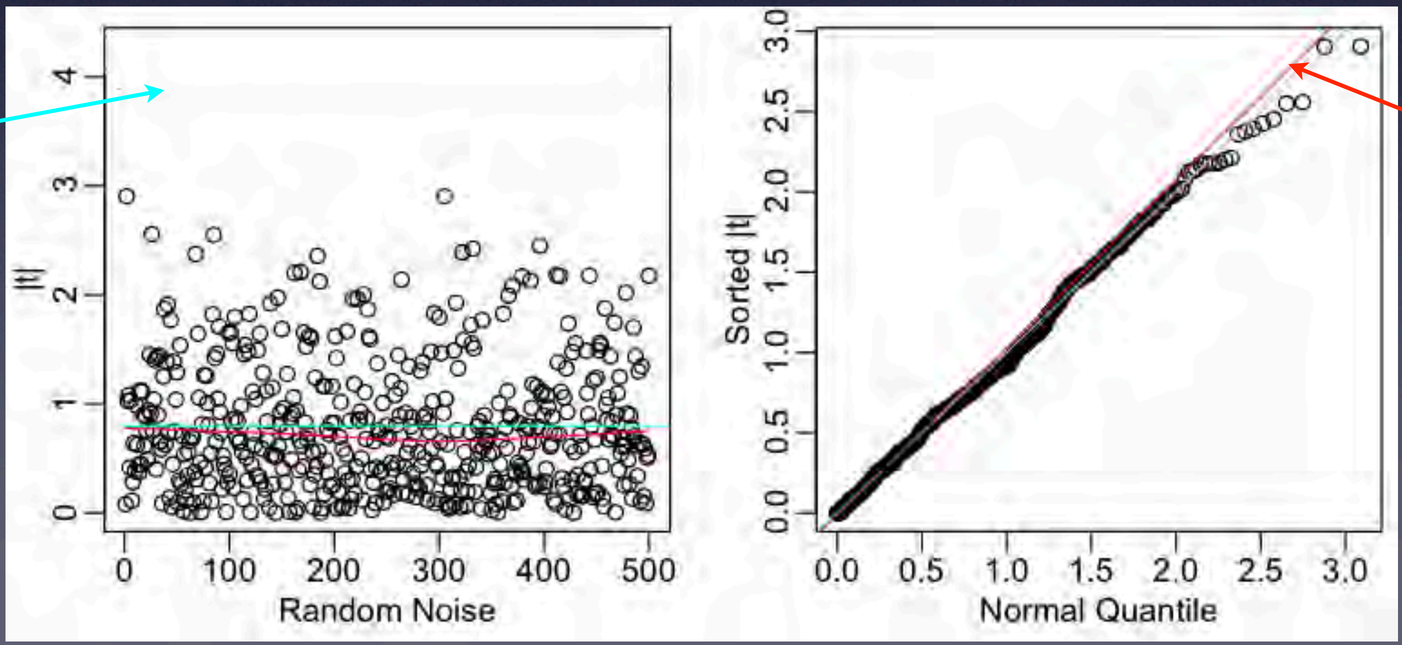
- Models have 100s of regressors, 1000s cases
- Graphical summary of coefficients
  - Absolute size of t-statistic
  - Half-normal plot of t-statistics

Plots for Random Gaussian Noise (null model)

Bonferroni threshold

Local average

sqrt(2/π)



Fit to estimates with small |t|

Wharton
Department of Statistics

Featurizing Text

Wharton
Department of Statistics

Methods

- All convert text into numerical variables
  - Text must be tokenized first
- Three direct, unsupervised approaches
  - Counts of words in documents
  - Apply principal components analysis to the counts of the different words
  - Form eigenwords from the sequence of words and build numerical variables from these

Less and less obvious

- Terminology
  - Principal components = latent semantic analysis
  - So called spectral methods

Tokenization

- What's a word?
  - Word versus word type
- Simple
  - White-space delimited sequence of characters
  - Alphabetic characters in lower case
  - Distinguish punctuation.

    Yes, . is a word type.
- Nothing fancy, such as
  - Stemming
  - Tagging with part of speech (parsing)
  - Correcting spelling errors
  - Encoding phone numbers, e-mail addresses

Sample after Tokenizing

Data from trulia.com for Chicago in June 2013.

```
1125000 recently gut rehabbed bucktown beauty on quiet , treelined street - extraordinary high end finis\
hes and quality workmanship throughout . fabulous chef 's kitchen , 4 bedrooms , office , two family roo\
ms , custom closets , 3 beautiful stone bathrooms , master with rainshower and body sprays , dual zone h\
vac , two laundry rooms , speaker system , wonderful outdoor deck and patio . two car garage . welcome h\
ome !
15000 nice lot to build your dream home . lot is fenced . close to expressway and shopping . this is s s\
hort sale , seller will look at all offers ! !
1250000 come see this 5 bedroom , 3.5 bath home of upgraded comfort , perfect location & easily compared\
 to new const . gourmet kitchen with custom banquet , ss appliances with double viking oven , gran count\
ers & large island / breakfast bar perfect for entertain . spacious great room steps from large deck ove\
r 2 car garage . great size beds with master bath with separate vanities whirlpool bath & huge steam sho\
wer . hardwood floors throughout , high ceil , dual zoned , 2 laundry , state of art elect system , stor\
age , steps to rest , transportation & in coveted coonley school district . for more information on this\
 property , contact the listing agent , mary gott at ( 312 ) 475-7772 or mgott@koenigstrey . com .
99000 location ! location ! great opportunity ! vacant lot in one of the fastest appreciating neighborho\
ods in chicago . commercial residential building all around . ideal for new construction . close to expr\
ess ways , downtown chicago , public transportation & shopping centers . many opportunities here , zoned\
 b3-1 . call on this one ! !
84000 10521 s kedzie ave , is located in chicago , il 60655 . it is currently listed for $AMOUNT$ . for \
more information , contact us at expert@govlisted . com . &lt ; br / &gt ; &lt ; br / &gt ; 10521 s kedz\
ie ave is a single family home and was built in 1948 . it has 3 bedrooms and 1.00 baths . 10521 s kedzie\
 ave was listed on 06 / 08 / 2013 . &lt ; br / &gt ; &lt ; br / &gt ; 10521 s kedzie ave , chicago , il \
```

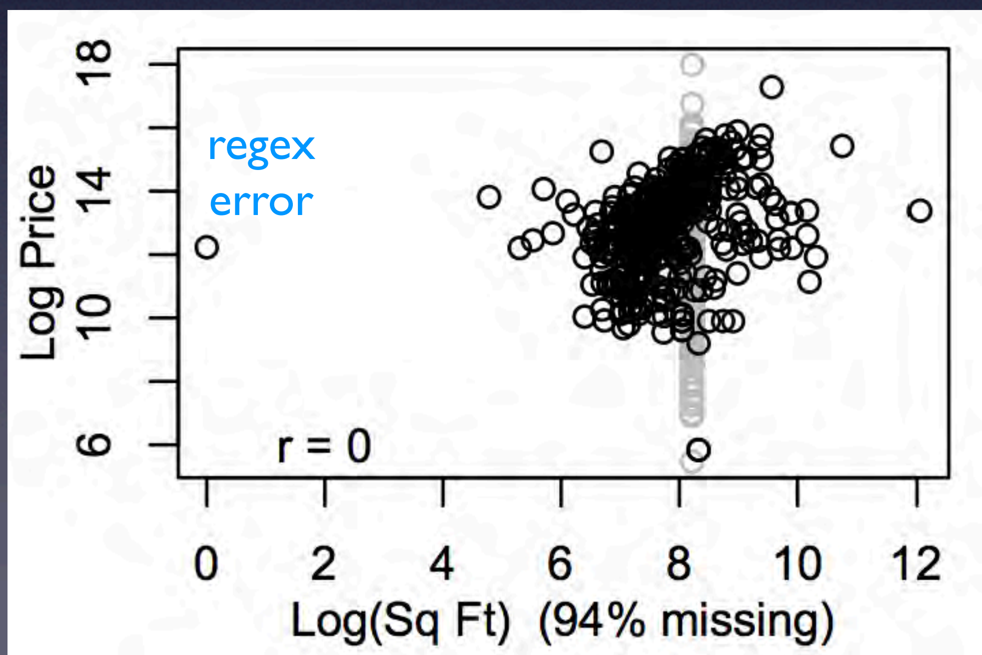
Each listing defines a document.

Not exactly proper English grammar!

Wharton
Department of Statistics

Parsing Is Hard

- Create regressors by matching text to regular expressions
- Example: square footage
  - Most listings do not show this: 94% missing
  - Weak correlation with log prices in the observed cases



Wharton
Department of Statistics

Featurizing Method 1

Word Counts

Wharton
Department of Statistics

Document/Word Matrix

- Sparse matrix W counts how many times each word type appears within each document

$W =$ 7400 x 6000

	w_1	w_2	w_3	w_4	w_5	...
d_1	0	0	0	2	0	
d_2	3	4	0	0	1	
d_3	0	2	4	0	0	
d_4	0	0	0	0	0	
d_5	0	2	0	0	0	
d_6	0	3	0	0	0	
d_7	0	0	0	1	1	
...						

Documents

Words

Wharton
Department of Statistics

Document/Word Matrix

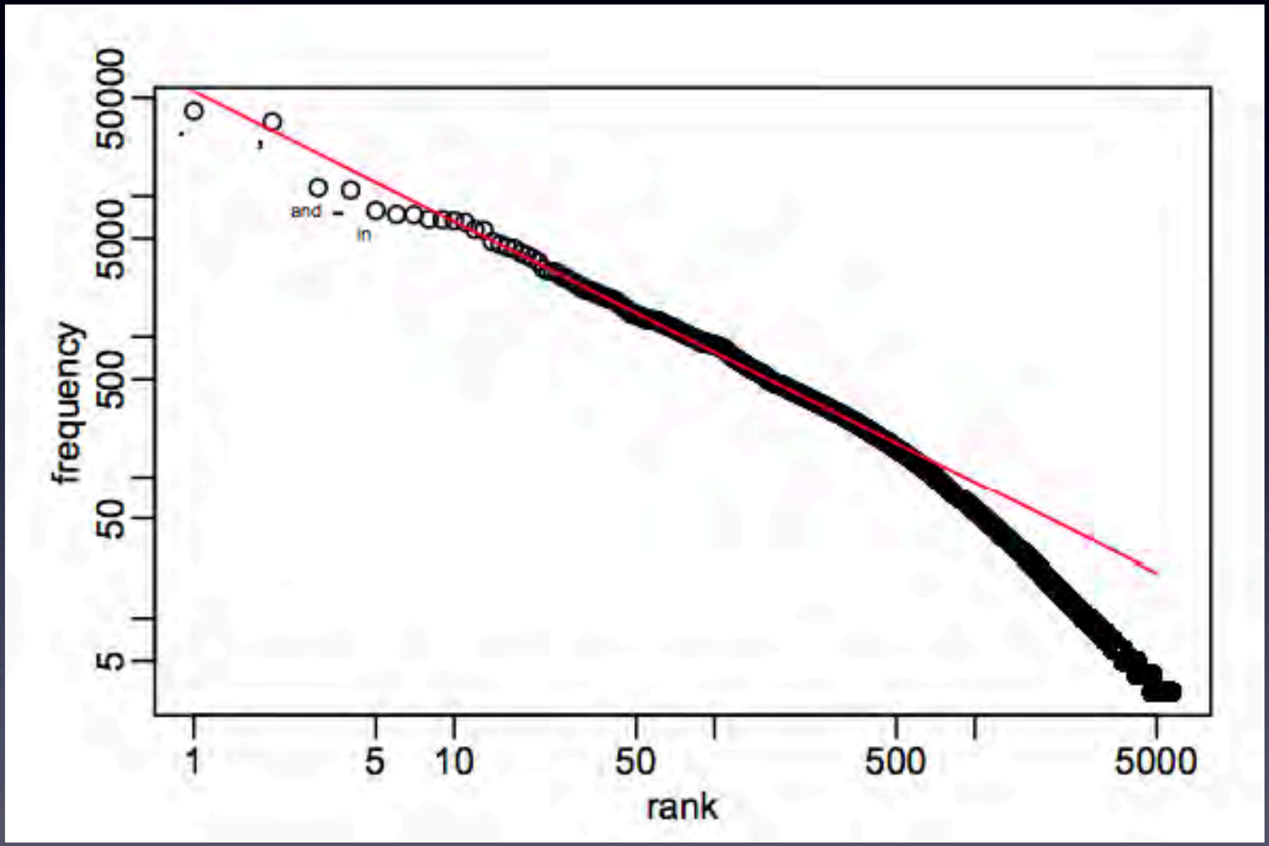
- Combine rare words
  - Most words types have small counts (Zipf dist.)
  - Combine those seen only once or twice throughout the corpus into type 'OOV'

    Reduces vocabulary from 15,000 to 6,000 for real estate
- Columns of W define regressors
  - Regressor = count of specific words
  - Fit a regression with several thousand columns
  - No variable selection – just use them all
- Benchmark
  - Can one predict as well (better?) with fewer

Shown on next slide

Zipf Distribution

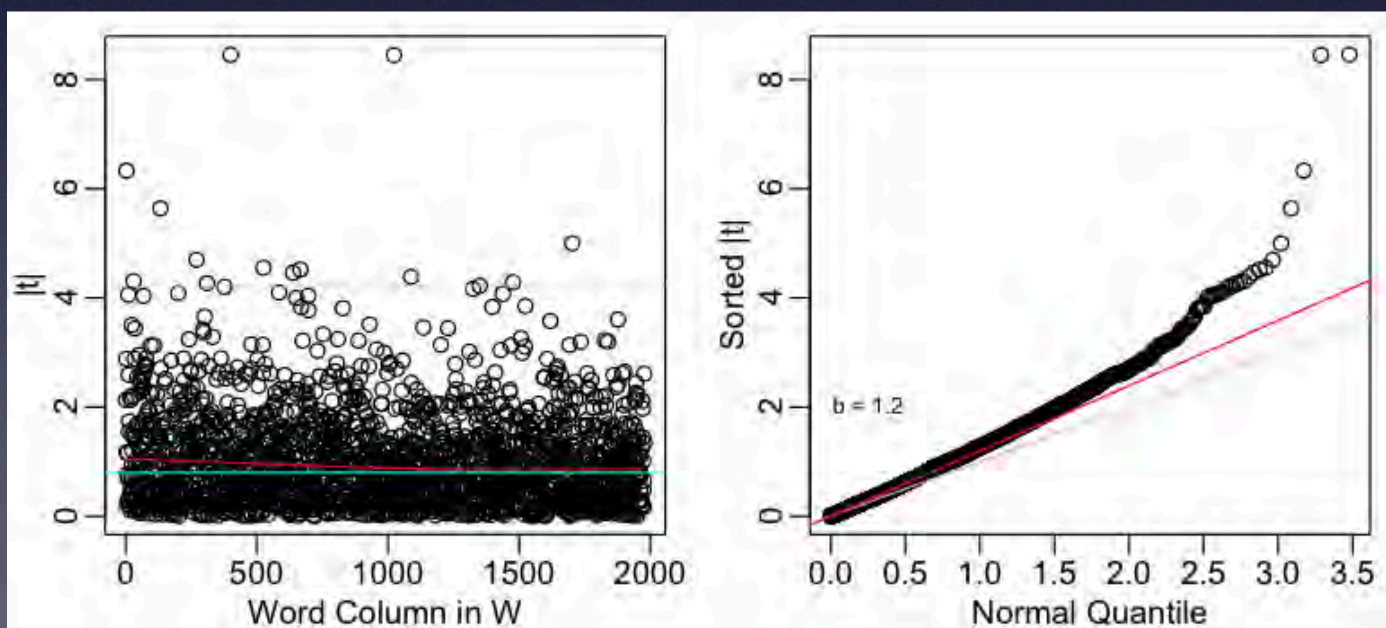
- Counts of word types in real estate listings



After merging OOV

Results for Real Estate

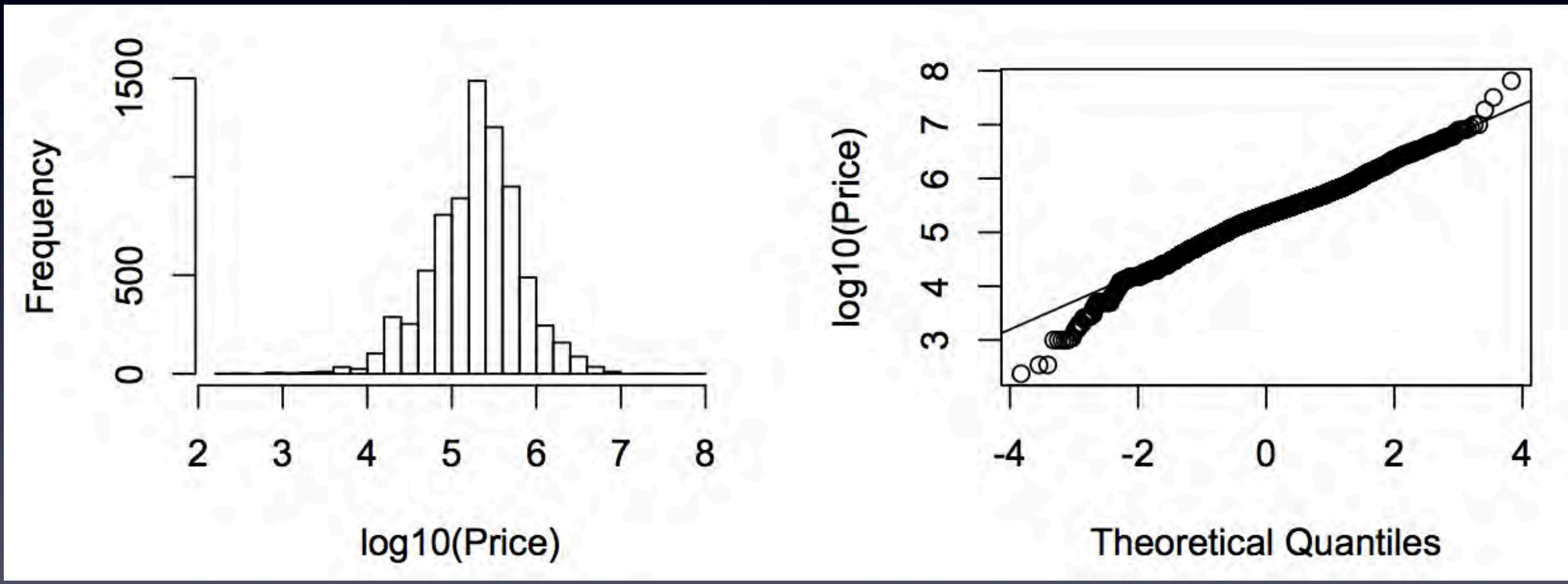
- W generates surprisingly good fit…
  - Regress log price on counts of 2,000 most common word types
- Performance
  - Adjusted R^2 = 68%
  - Diffuse statistically significant coefficients



Wharton
Department of Statistics

Why Logs?

- Prices for real estate in Chicago follow roughly log normal distribution



Featurizing Method 2

Principal Components

Wharton
Department of Statistics

Concentrate Signal

- Regression on words
  - Explains substantial of variation among prices
  - Cannot limit attention to the big ones
    If retain only those coefficients that pass the Bonferroni threshold, then adj R^2 drops to 19%.
- Heuristic model
  - Response lives in low-dimension + noise
    $$y = g(\mu) + \text{random noise}$$
  - Each regressor is μ plus random noise
    $$x_j = \mu + \text{more random error}$$
  - Get a better regressor by averaging the x_js
    $$\overline{x} = (x_1 + x_2 + \ldots + x_p)/p$$

Wharton
Department of Statistics

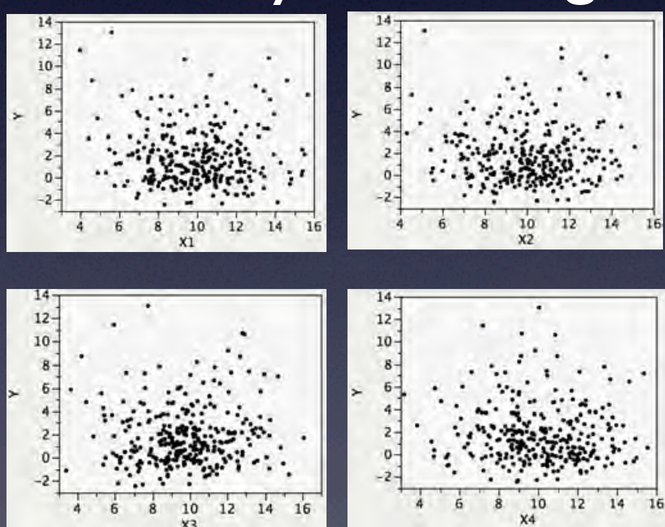
Better Averaging

- How would you know that you should just average the x_js to recover μ?
- Search for interesting directions

unsupervised

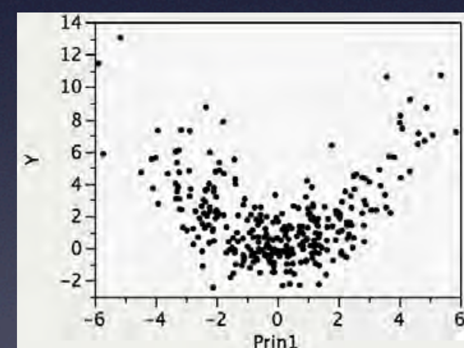
  - Find weighted sums of the x_js. y is not used.
  - Rely on enough variation among elements of μ

simulated example



Association with observed

	Prin1
X1	0.23774
X2	0.16353
X3	0.21825
X4	0.21200
X5	0.21186
X6	0.23490
X7	0.19348
X8	0.18074
X9	0.29655
X10	0.21732
X11	0.23696
X12	0.23441
X13	0.25799
X14	0.22608
X15	0.22582
X16	0.21216
X17	0.21205
X18	0.23009
X19	0.23701
X20	0.19993



$g(\mu) = (\mu-10)^2$

Association with PC

- Generalization: Principal components analysis

Wharton
Department of Statistics

Latent Semantic Analysis

- Idea
  - Replace columns of W by matrix M with fewer columns
  - New columns (principal components) are weighed sums of the original columns

    Chosen to have maximal variance and be uncorrelated
  - Fewer dimensions while preserving document separation

Embarrassed to admit how long before I realized this!

- Classical eigenvalue problem
  - Albeit applied to much larger matrix than usual
  - W in real estate has

    7,400 rows and 5,700 columns

Clustering

- LSA also used for clustering documents (LSI)
  - W with 1,000s of columns replaced by PC matrix M with fewer columns
- New coordinates
  - W
    Each document represented by long, sparse vector of word counts.
  - M
    Each document represented by point in lower dimensional space
- Cluster the documents in this new space
  - Early approach to document retrieval

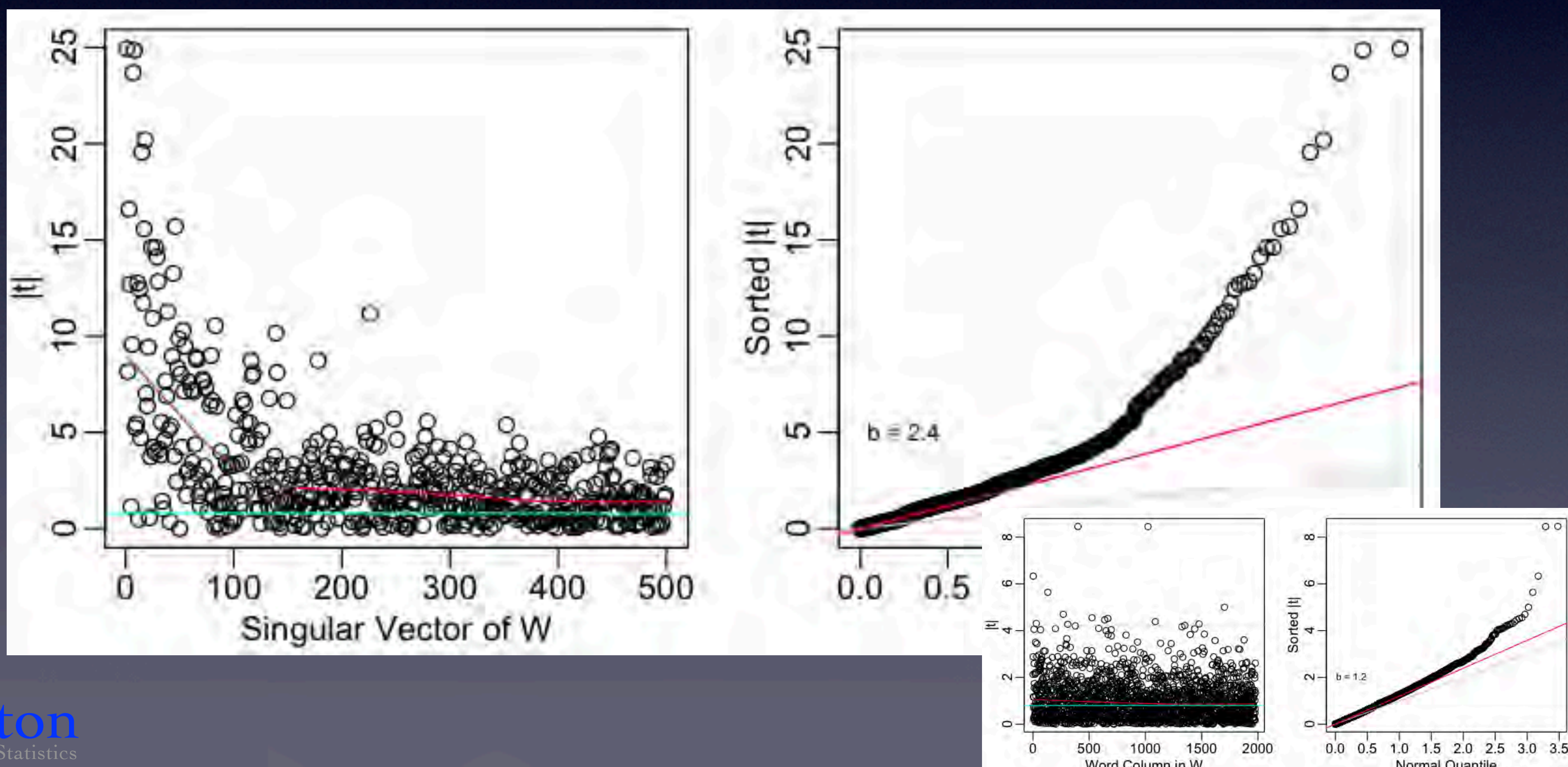
Computation?

- W is a very big matrix, but...
- W is sparse
  - Most elements of W are zero so don't have to reserve space or manipulate
    $$7{,}400 \times 5{,}700 \approx 42{,}000{,}000$$
    elements
- Random algorithms
  - Computers are pretty fast, and
  - Modern algorithms based on random projection make this a fast calculation.

Results for Real Estate

- Retaining 500 PCs produces nearly as good a fit as words, but more concentrated
  - Adjusted R^2 = 61%
  - High variance components also more predictive

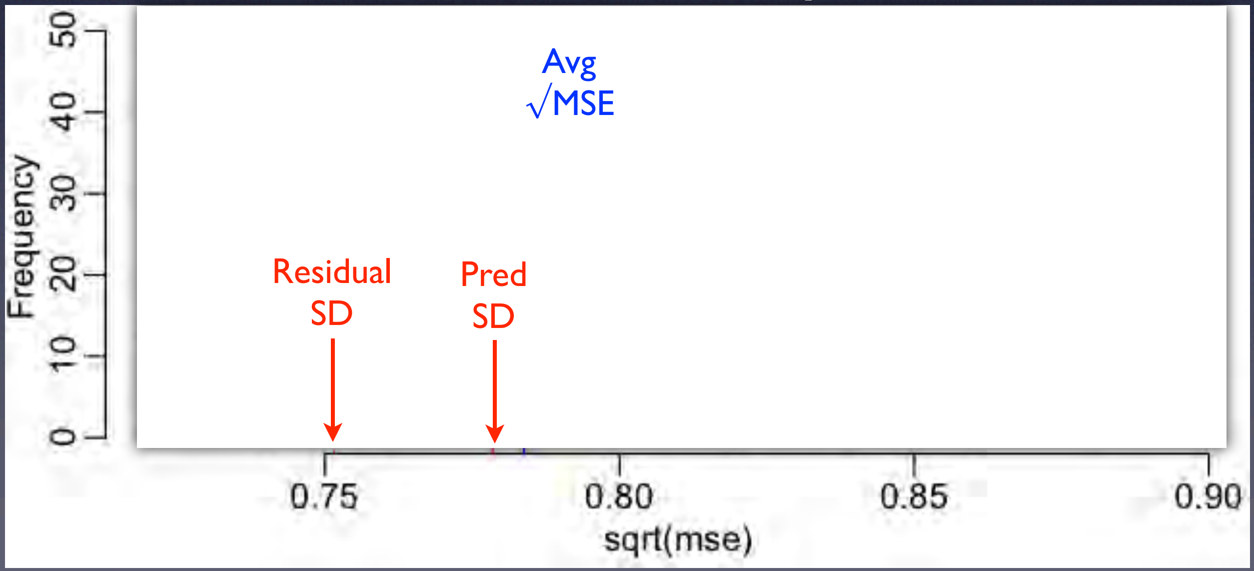
2000 words has adj R^2 68%



Wharton
Department of Statistics

Cross Validation

- Model predicts as well as it claims
- Validate using out-of-sample test cases
  - Transductive case
    Regressors for test cases are available when building model (ie, used in PCA)
- Model prediction
  - 10-fold cross-validation, repeated 20 times



Featurizing Method 3

Bigram Components

Wharton
Department of Statistics

Bigram Analysis

- W = bag-of-words
  - W defines word space based on co-occurrence within a document
  - Treats document as a multiset, losing information related to order
- Bigram matrix counts adjacent word pairs

$$
B = \begin{array}{c|c}
\begin{bmatrix} 0 & 0 & 2 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 1 \end{bmatrix} & \begin{matrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{matrix} \\
\begin{matrix} w_1 & w_2 & w_3 & w_4 & w_5 \end{matrix} &
\end{array}
$$

6000 x 6000

Wharton
Department of Statistics

Singular Value Decomposition

- Represent a matrix as a weighted sum of simpler matrices

$$B = d_1\, u_1\, v_1^t + d_2\, u_2\, v_2^t + \ldots$$

  - $d_1 \geq d_2 \geq \ldots$ are constants (singular values)
  - u_j and v_j are vectors (left and right singular vec)

- Truncated sum = 'low rank' approximation
  - Heuristic: remaining terms random noise

- Alternative expression, as a product

$$B = U\, D\, V^t$$

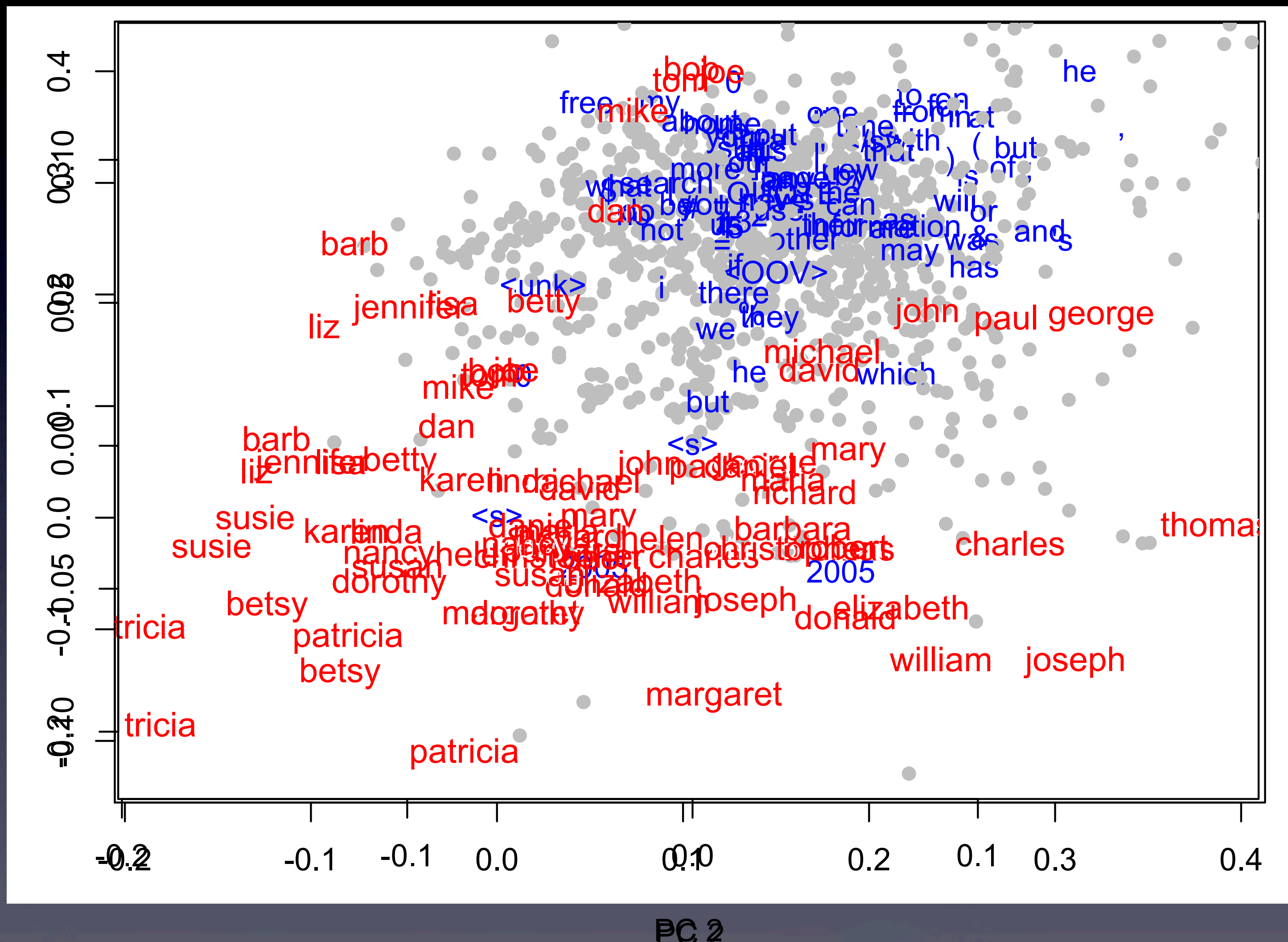
  - D is diagonal with elements dj
  - u_j and v_j are columns of U and V

Truncation retains only leading columns of U,V

Building Regressors

- Singular vectors identify new coordinates for words based on adjacency
  - Such coordinates called ‘eigenwords’ by Ungar and colleagues
  - Can be constructed from counts of other n-grams (three, four, or more consecutive words)
- Examples from Google n-grams provide some sense of what these measure
  - Vocabulary of 50,000 words with Internet as source text
  - A word is a point in a space of lower dimension
  - Labelling selected words provides intuition

Example of Eigenwords



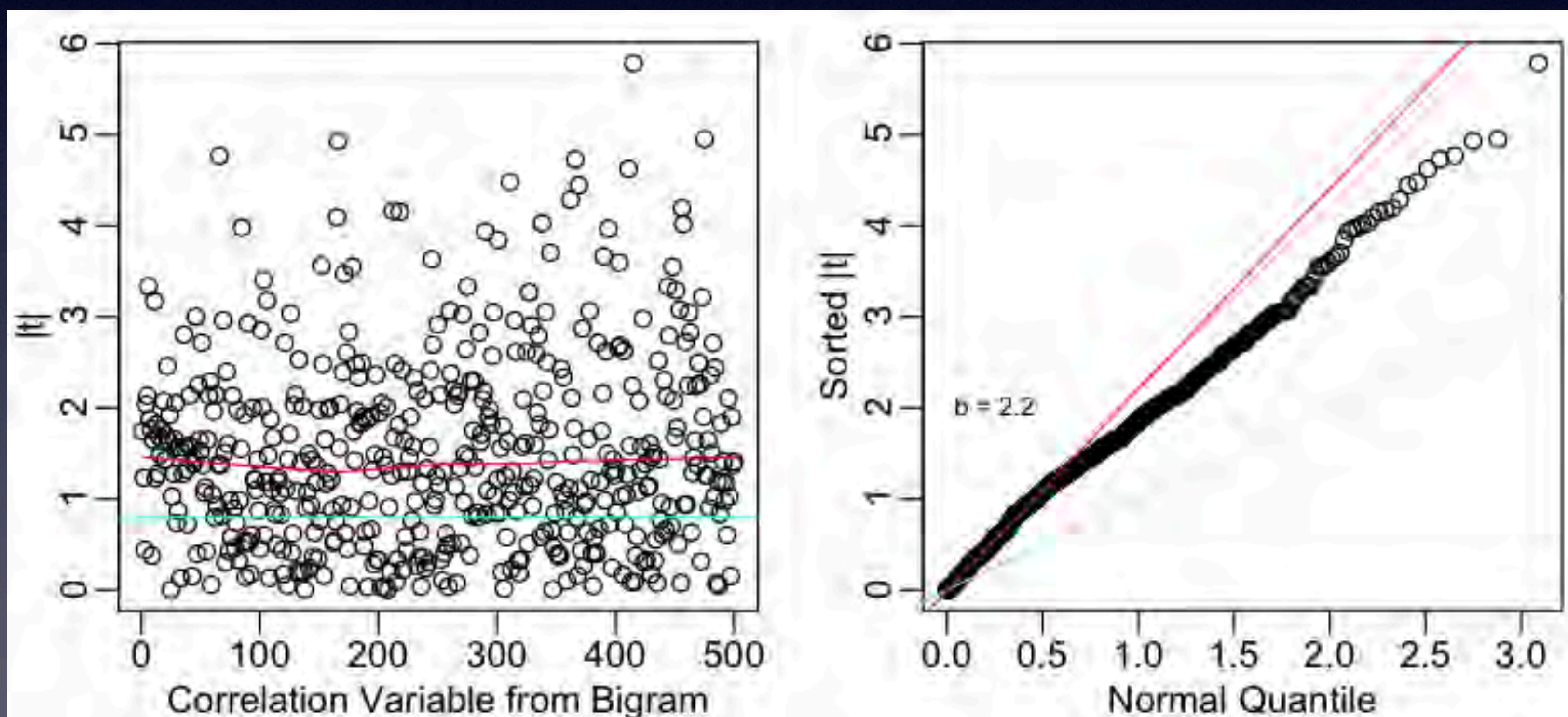
Getting Regressors

- Eigenwords define locations of words in a lower-dimension space, say C
- To represent documents in this new space, compute the average position of its words
  - Each word in a document is point in C
  - Represent document as the average position of its words (centroid)
- 'Equivalent' to correlation between word mix of document (row of W) and singular vectors

Wharton
Department of Statistics

Results for Real Estate

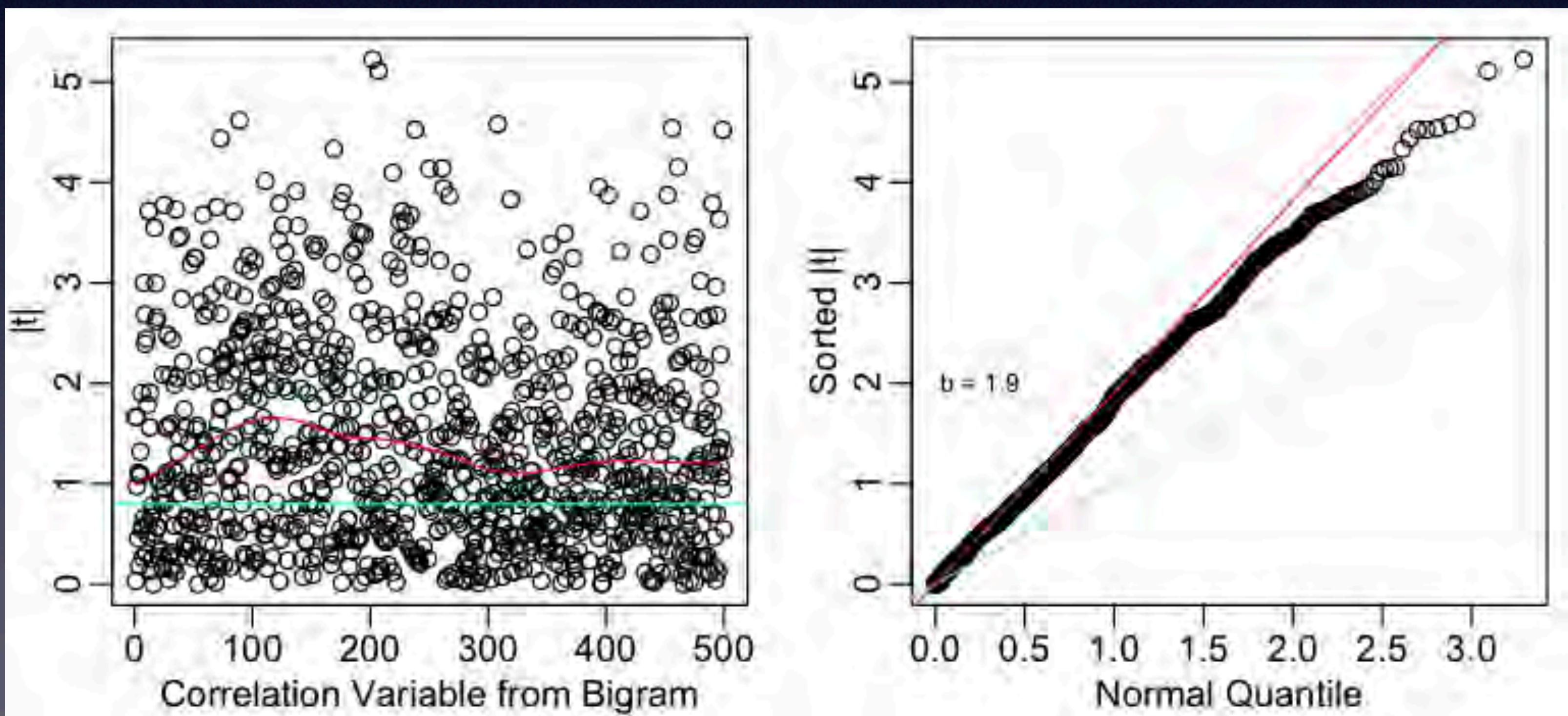
- Leading 500 left singular vectors explain similar variation to LSA (61%), but
- Lose the concentration of signal



Wharton
Department of Statistics

Results for Real Estate

- Adding the right singular vectors lifts adjusted R^2 to 66%, but without concentrating signal
  - Collinearity between left/right singular vectors



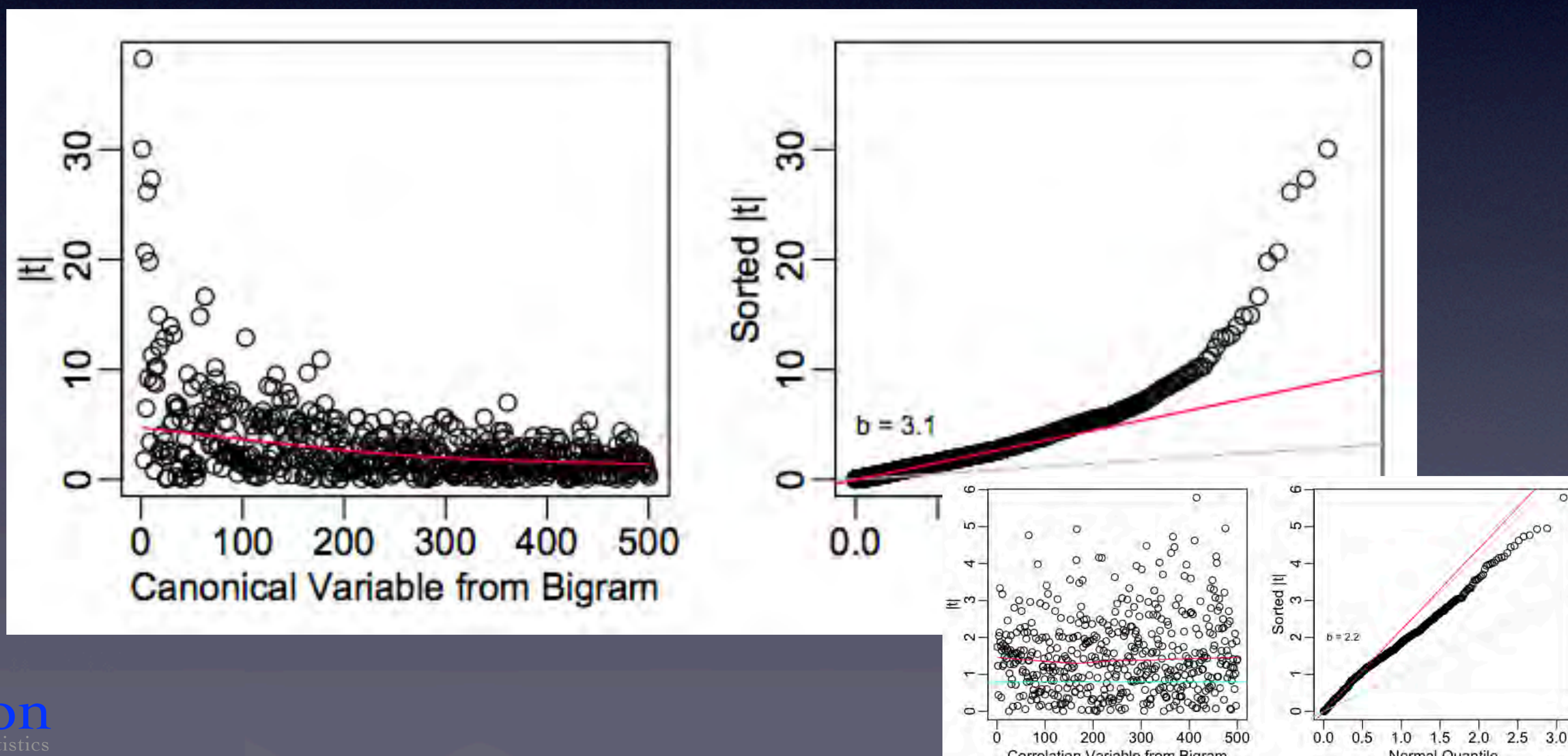
Better than raw words, but we can do better...

Nicer Regressors

- Left and right singular vectors of B are
  - Correlated
  - Defined by adjacent co-occurrence
- Use common information to form better regressors
  - More power in fewer coordinates
- Technique: canonical correlation analysis
  - Find weighted sum of one collection of variables that is most related to a weighted sum of a second collection
  - Weighted sums are known as canonical variables

Results for Real Estate

- Canonical variables formed from the CCA of the bigram singular vectors again concentrate signal
  - Same fit, just rearranged into fewer components



Adding more?

- Combine regressors
  - For instance

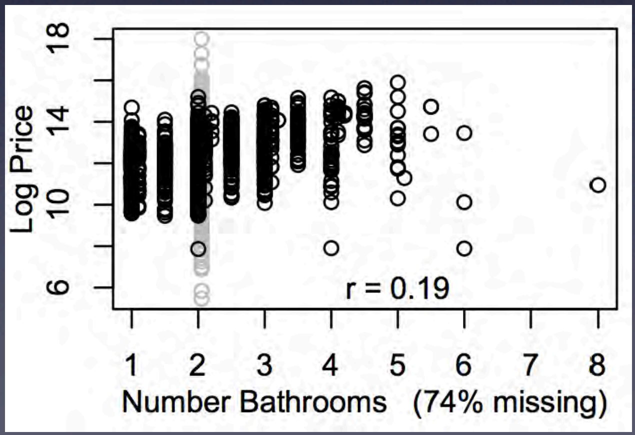
Regress on LSA variables	61%
Add bigram l.h.s. variables	68%

- Substantive parsing
  - Tried originally to use regular expressions
  - Parse for #bathrooms, bedrooms and sq. ft.
  - Adds 0.3% to R^2…
    Statistically significant but not noticeable.

Interpretation

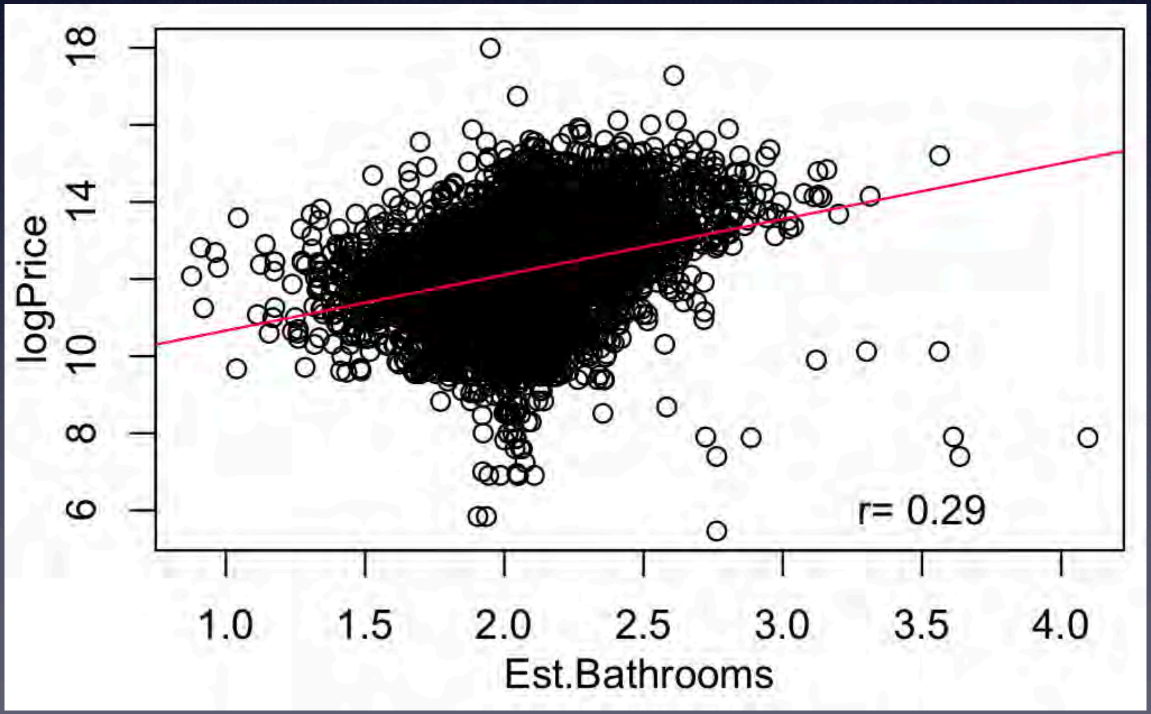
Predictive but Unattractive

- Offer some interpretable variables
- Lighthouse variables
  - Create substantively oriented variable, perhaps from partial information
  - Use substantive variable to form interpretable combinations of PCA or singular vectors
- Example: number bathrooms
  - Partially observed
    3/4 missing
  - Correlation r = 0.4 when limited to observed cases



Guided PCA

- Form combination of PCA variables that is most correlated available parsed count
- Use this new variable as regressor in place of bathrooms



Wharton
Department of Statistics

Topic Models

Wharton
Department of Statistics

Topic Model

- Data generating process, a probability model
  - Cluster documents based on common ‘topics’
  - Bag-of-words model
- Typical analysis
  - Unsupervised (no response to predict)
  - Specify priors for Bayesian model
  - Given model, use Markov Chain Monte Carlo (MCMC) to find distribution of latent topics
- Example
  - Cluster articles that appear in *Science* magazine
  - Explore how topics evolve

You get to play ‘name that topic’ as in factor analysis.

Basic Model

- Each document mixes words from collection of topics
  - topic = probability distribution over words
  - Details: Blei, Ng, and Jordan 2003

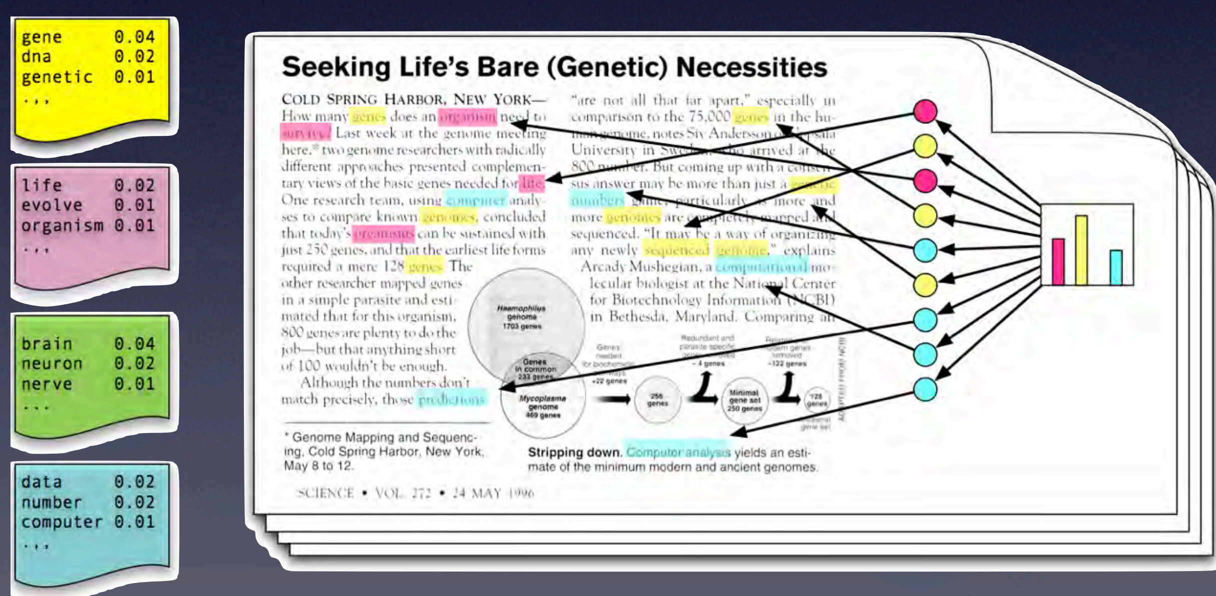


Figure from Blei, Intro to Topic Models

Probability Model

Beta:Binomial
as
Dirichlet:Multinomial

- Latent Dirichlet allocation (LDA)
- Define K topics
  - Discrete distributions over vocabulary

    $P_k \sim$ Dirichlet, $k = 1,\ldots, K$
- Each document covers a mixture of topics
  - Random distribution

    $Z_i \sim$ Dirichlet, $i = 1,\ldots, n$
- Topic mixture
  - Determines words that appear

    $P(\text{word } w \text{ in doc } i) = P_{kw}$ $\quad k \sim \text{Multi}(Z_i)$
  - Defines the response

    $y_i = Z_i'\beta + \text{noise}$

Wharton
Department of Statistics

Simulate Topic Data

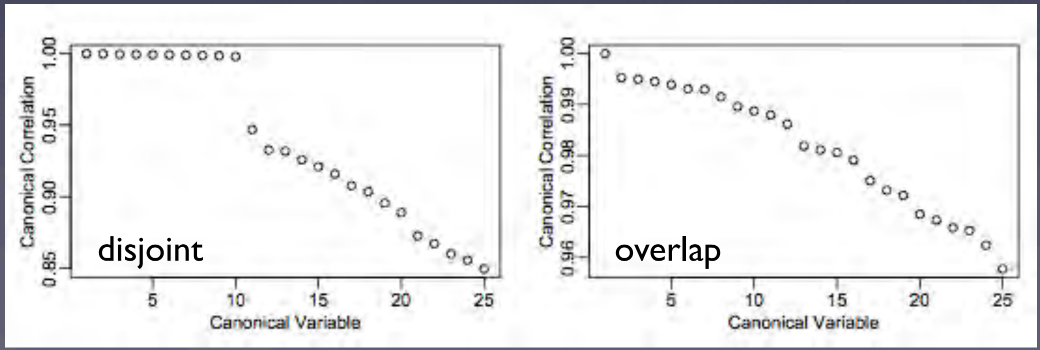
- Suppose data were generated in this fashion
- Simulation
  - 10 topics (K=10, hidden in analysis)
  - 2000 word types
  - 4000 documents
- Nature of the topics
  - Disjoint… few words in common
  - Overlapping… many words in common
- Response
  - Weighted sum of topic shares, R^2=0.92

Results for Topics

- Modeling
  - 100 PCs of W, 100 left and 100 right from B
  - Predicts well
- Impact of topic overlap
  - Better fitting model with distinct topics

Topic Structure	Num Regressors	Origin	$\overline{R}^2$
Disjoint	100	W	0.746
	200	B	0.788
Overlapping	100	W	0.516
	200	B	0.626

  - CCA reveals K if disjoint



Wharton
Department of Statistics

Comments on LDA

- Nice to have probability model that 'explains' why
  - direct methods work
  - results from W and B are similar
- Not perfect
  - Need to enrich with some sequential dependence to mimic text
  - Insert Markov chain into sequence of topics that generate words within document

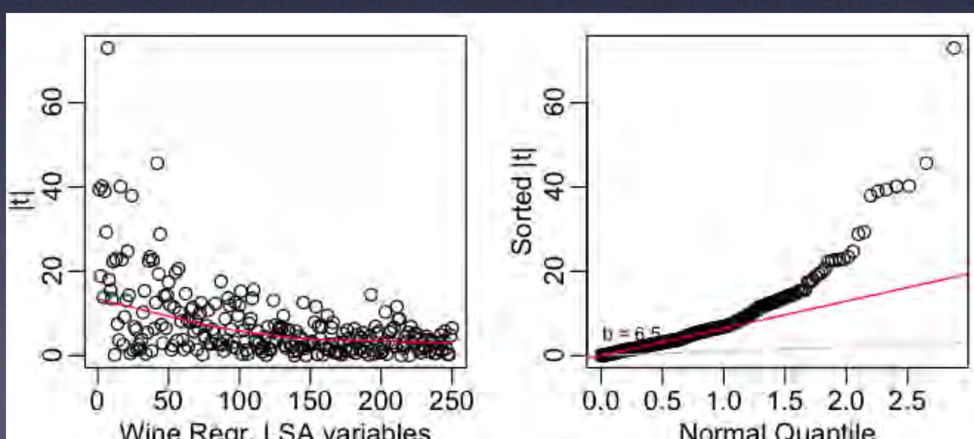
Wrap-Up

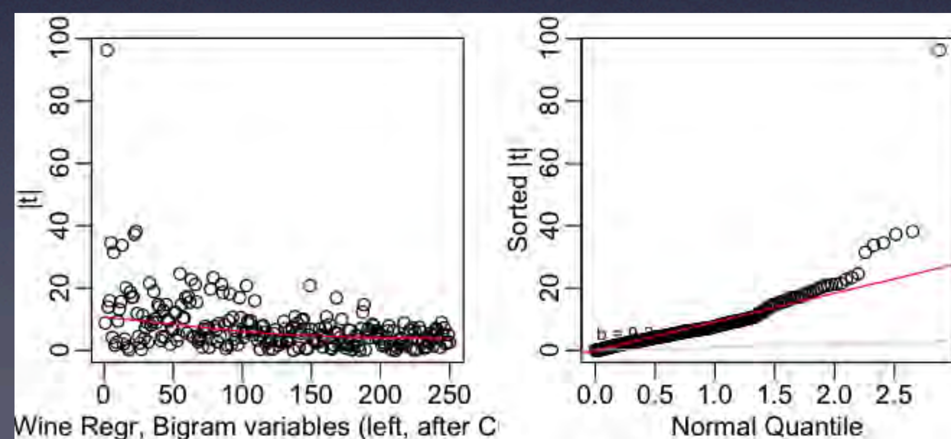
Take-Aways

- Direct conversions of text to numerical variables allow one to easily exploit unstructured text in regression models
  - Exploit conventional statistical routines in a different context
  - The analysis is fast to run
- Related to probability models for documents (LDA, topic models)
- The results illustrated for real estate seem representative rather than exceptional
  - It works in other problems too...

Wine Ratings

- Data
  - 22,000 wine tasting notes (Thank you, Mark)
  - Response is rating of wine
- Results
  - 250 PCs of W: adj R^2 = 67%
  - 500 SVs of B : adj R^2 = 68%
- Similar qualitative concentration of signal





Next Steps

- Transfer learning
  - Chicago real estate next year
  - Miami real estate
- More elaborate tokenization
  - Stemming, parsing/tagging
- Exploiting other word counts and sources
  - Trigrams
  - Merging with other quantitative data
- Statistics: variable selection
  - Outside the 'nearly black' context of theory
  - Capturing nonlinearities, word synergies

Thanks for coming!