

Statistical Analysis of Text

- Statistical text analysis has a long history in literary analysis and in solving disputed authorship problems
- First (?) is Thomas C. Mendenhall in 1887

SCIENCE.

FRIDAY, MARCH 11, 1887.

*THE CHARACTERISTIC CURVES OF COM-
POSITION.*

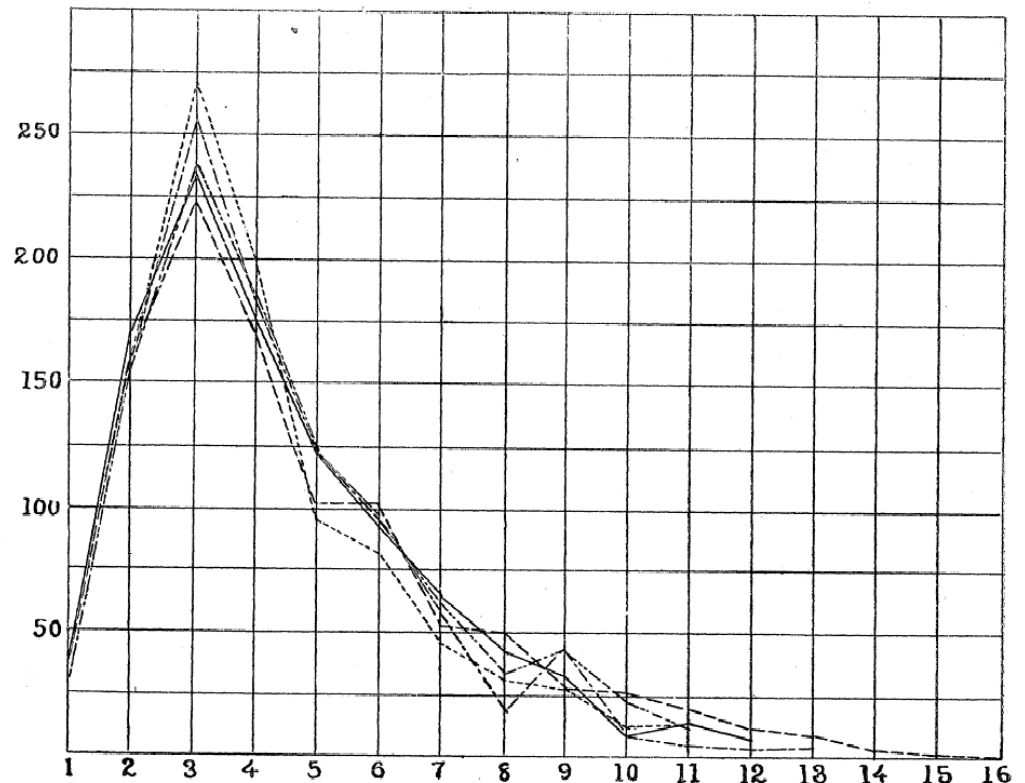


FIG. 2.—SHOWING FIVE GROUPS, OF ONE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'

Mendenhall

- Mendenhall was Professor of Physics at Ohio State and at University of Tokyo, Superintendent of the USA Coast and Geodetic Survey, and later, President of Worcester Polytechnic Institute

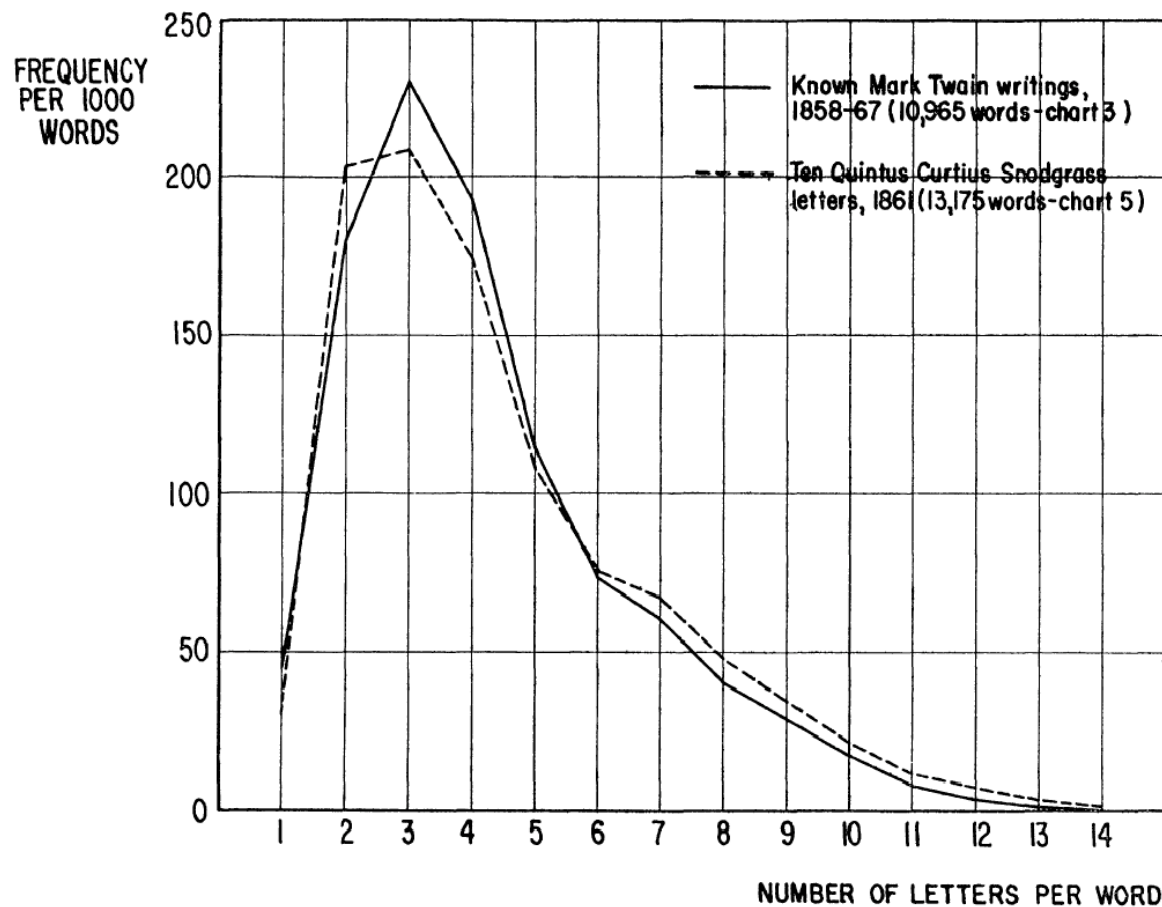


Mendenhall Glacier,
Juneau, Alaska

AMERICAN STATISTICAL ASSOCIATION JOURNAL, MARCH 1963

MARK TWAIN AND THE QUINTUS CURTIUS SNODGRASS LETTERS: A STATISTICAL TEST OF AUTHORSHIP

CLAUDE S. BRINEGAR



$$X^2 = 127.2, df=12$$



**ON SENTENCE-LENGTH AS A STATISTICAL CHARACTER-
ISTIC OF STYLE IN PROSE: WITH APPLICATION TO TWO
CASES OF DISPUTED AUTHORSHIP**

By G. UDN YULE

, Vol. 30, No. 3/4 (Jan., 1939), 363-390.

**A NOTE ON THE STATISTICAL ANALYSIS OF SENTENCE
LENGTH AS A CRITERION OF LITERARY STYLE**

By C. B. WILLIAMS, Sc.D.

Department of Entomology, Rothamsted Experimental Station

Vol. 31, No. 3/4 (Mar., 1940), 356-361.

Did Shakespeare write a newly-discovered poem?

By RONALD THISTED

Department of Statistics, University of Chicago, Chicago, Illinois 60637, U.S.A.

AND BRADLEY EFRON

1, Vol. 74, No. 3 (Sep., 1987), 445-455.



On a Distribution Law for Word Frequencies

H. S. SICHEL*

, Vol. 70, No. 351 (Sep., 1975), 542-547.

**MARK TWAIN AND THE QUINTUS CURTIUS SNODGRASS
LETTERS: A STATISTICAL TEST OF AUTHORSHIP**

CLAUDE S. BRINEGAR

AMERICAN STATISTICAL ASSOCIATION JOURNAL, MARCH 1963

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 302

JUNE, 1963

Volume 58

INFERENCE IN AN AUTHORSHIP PROBLEM^{1,2}

A comparative study of discrimination methods applied
to the authorship of the disputed *Federalist* papers

FREDERICK MOSTELLER

Harvard University

and

Center for Advanced Study in the Behavioral Sciences

AND

DAVID L. WALLACE

University of Chicago

- Used Naïve Bayes with Poisson and Negative Binomial model
- Out-of-sample predictive performance

Today

- Statistical methods routinely used for textual analyses of all kinds
- Machine translation, part-of-speech tagging, information extraction, question-answering, text categorization, etc.
- Not reported in the statistical literature

Text Categorization

- Automatic assignment of documents with respect to manually defined set of categories
- Applications automated indexing, spam filtering, content filters, medical coding, CRM, essay grading
- Dominant technology is supervised machine learning:

Manually classify some documents, then learn a classification rule from them (possibly with manual intervention)

Document Representation

- Documents usually represented as “bag of words:”

$$\mathbf{x}_i = \{x_{i1}, \dots, x_{id}\}$$

- x_i 's might be 0/1, counts, or weights (e.g. tf/idf, LSI)
- Many text processing choices: stopwords, stemming, phrases, synonyms, NLP, etc.

Naïve Bayes via a Toy Spam Filter

Example

- Naïve Bayes is a generative model that makes drastic simplifying assumptions
- Consider a small training data set for spam along with a bag of words representation

#	Message	Spam
1	the quick brown fox	no
2	the quick rabbit ran and ran	yes
3	rabbit run run run	no
4	rabbit at rest	yes

Training data comprising four labeled e-mail messages.

#	and	at	brown	fox	quick	rabbit	ran	rest	run	the
1	0	0	1	1	1	0	0	0	0	1
2	1	0	0	0	1	1	2	0	0	1
3	0	0	0	0	0	1	0	0	3	0
4	0	1	0	0	0	1	0	1	0	0

Term vectors corresponding to the training data.

	X_1	X_2	X_3	X_4	X_5	X_6	Y
#	brown	fox	quick	rabbit	rest	run	Spam
1	1	1	1	0	0	0	0
2	0	0	1	1	0	2	1
3	0	0	0	1	0	3	0
4	0	0	1	1	1	0	1

Term vectors after stemming and stopword removal with the Spam label, coded as 0=no, 1=yes.

Naïve Bayes Machinery

- We need a way to estimate:

$$Pr(Y = 1|X_1 = x_1, \dots, X_d = x_d)$$

- Via Bayes theorem we have:

$$= \frac{Pr(Y = 1) \times Pr(X_1 = x_1, \dots, X_d = x_d|Y = 1)}{Pr(X_1 = x_1, \dots, X_d = x_d)}$$

or, on the log-odds scale:

$$\begin{aligned} \log \frac{Pr(Y = 1|X_1 = x_1, \dots, X_d = x_d)}{Pr(Y = 0|X_1 = x_1, \dots, X_d = x_d)} \\ = \log \frac{Pr(Y = 1)}{Pr(Y = 0)} + \log \frac{Pr(X_1 = x_1, \dots, X_d = x_d|Y = 1)}{Pr(X_1 = x_1, \dots, X_d = x_d|Y = 0)} \end{aligned}$$

Naïve Bayes Machinery

- Naïve Bayes assumes:

$$Pr(X_1 = x_1, \dots, X_d = x_d | Y = 1) = \prod_{i=1}^d Pr(X_i = x_i | Y = 1)$$

and

$$Pr(X_1 = x_1, \dots, X_d = x_d | Y = 0) = \prod_{i=1}^d Pr(X_i = x_i | Y = 0)$$

leading to:

$$\begin{aligned} \log \frac{Pr(Y = 1 | X_1 = x_1, \dots, X_d = x_d)}{Pr(Y = 0 | X_1 = x_1, \dots, X_d = x_d)} \\ = \log \frac{Pr(Y = 1)}{Pr(Y = 0)} + \sum_{i=1}^d \log \frac{Pr(X_i = x_i | Y = 1)}{Pr(X_i = x_i | Y = 0)} \end{aligned}$$

Maximum Likelihood Estimation

$$\log \frac{\widehat{Pr}(Y = 1)}{\widehat{Pr}(Y = 0)} = \log \frac{2/4}{2/4} = 0$$

weights
of
evidence

$$\log \frac{\widehat{Pr}(X_3 = 1|Y = 1)}{\widehat{Pr}(X_3 = 1|Y = 0)} = \log \frac{2/2}{1/2} = \log 2$$

	X_1	X_2	X_3	X_4	X_5	X_6	Y
#	brown	fox	quick	rabbit	rest	run	Spam
1	1	1	1	0	0	0	0
2	0	0	1	1	0	2	1
3	0	0	0	1	0	3	0
4	0	0	1	1	1	0	1

Naïve Bayes Prediction

- Usually add a small constant (e.g. 0.5) to avoid divide by zero problems and to reduce bias

	X_1	X_2	X_3	X_4	X_5	X_6
	brown	fox	quick	rabbit	rest	run
Term Present	-1.10	-1.10	0.51	0.51	1.10	0
Term Absent	0.51	0.51	-1.10	-1.10	-0.51	0

Estimated Weights of evidence for the example.

- New message: “the quick rabbit rests”

- New message: “the quick rabbit rests”

	X_1	X_2	X_3	X_4	X_5	X_6
	brown	fox	quick	rabbit	rest	run
Term Vector	0	0	1	1	1	0
Weight of Evidence	0.51	0.51	0.51	0.51	1.10	0

- Predicted log odds:

$$0.51 + 0.51 + 0.51 + 0.51 + 1.10 + 0 = 3.04$$

- Corresponds to a spam probability of 0.95

Logistic Regression Model

- Linear model for log odds of category membership:

$$\log \frac{p(y=1 | \mathbf{x}_i)}{p(y=-1 | \mathbf{x}_i)} = \sum \beta_j x_{ij} = \boldsymbol{\beta} \mathbf{x}_i$$

- Conditional probability model

Maximum Likelihood Training

- Choose parameters (β_j 's) that maximize probability (likelihood) of class labels (y_i 's) given documents (\mathbf{x}_i 's)

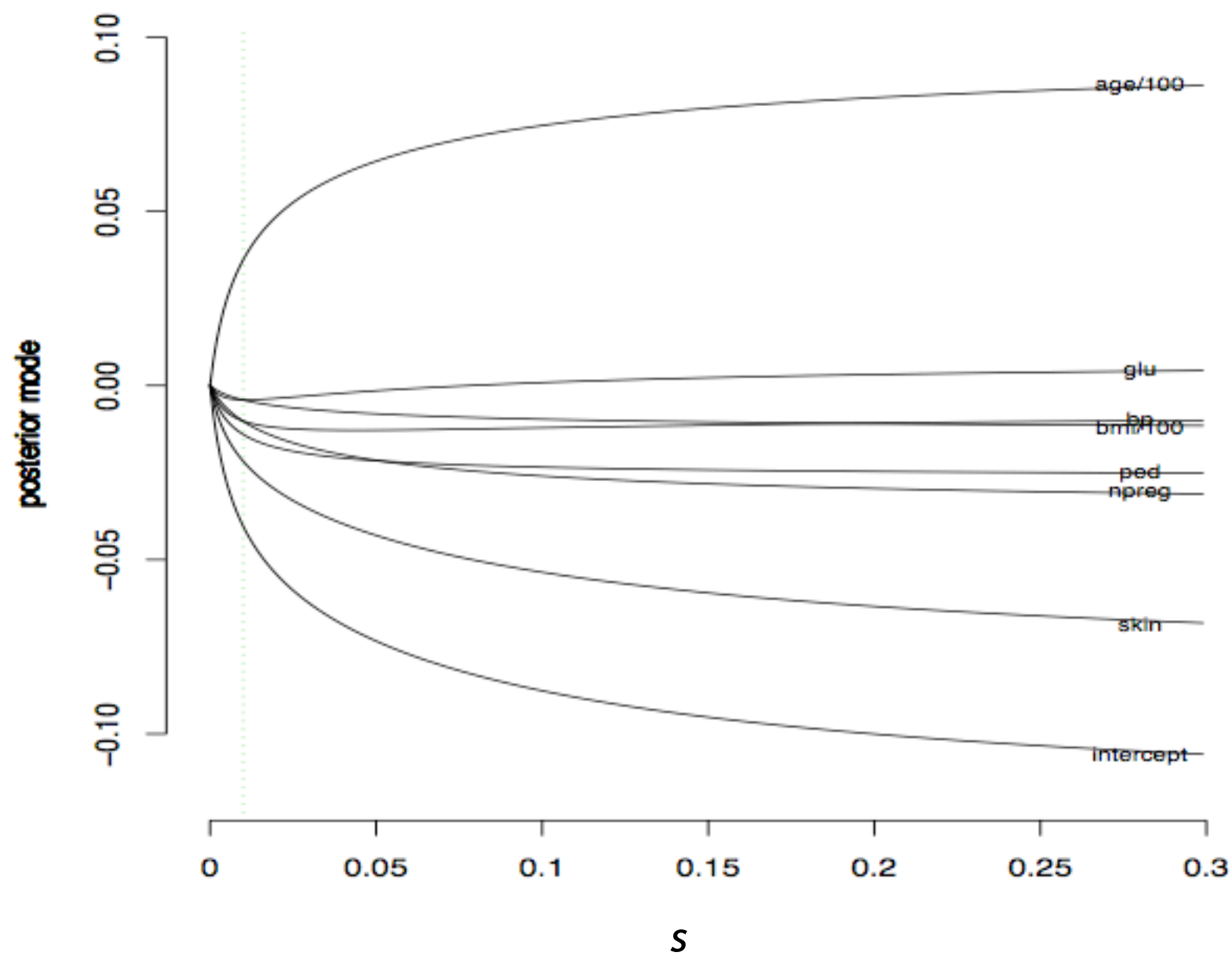
$$L(\boldsymbol{\beta}) = p(\boldsymbol{\beta}|D) = \left(\prod_{i=1}^n \frac{1}{1 + \exp(-\boldsymbol{\beta}^T \mathbf{x}_i y_i)} \right)$$

- Tends to overfit
- Not defined if $d > n$
- Feature selection

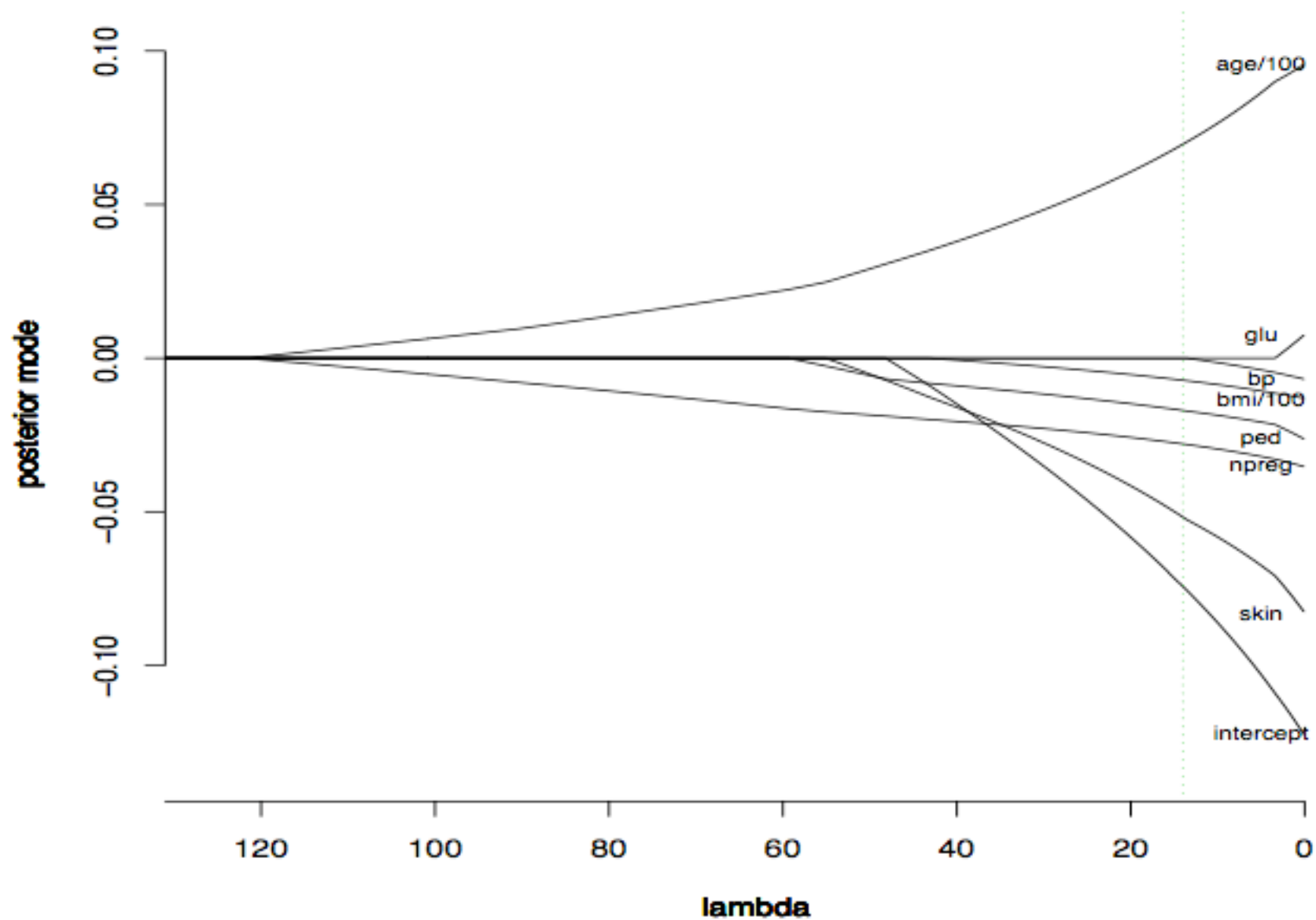
Shrinkage Methods

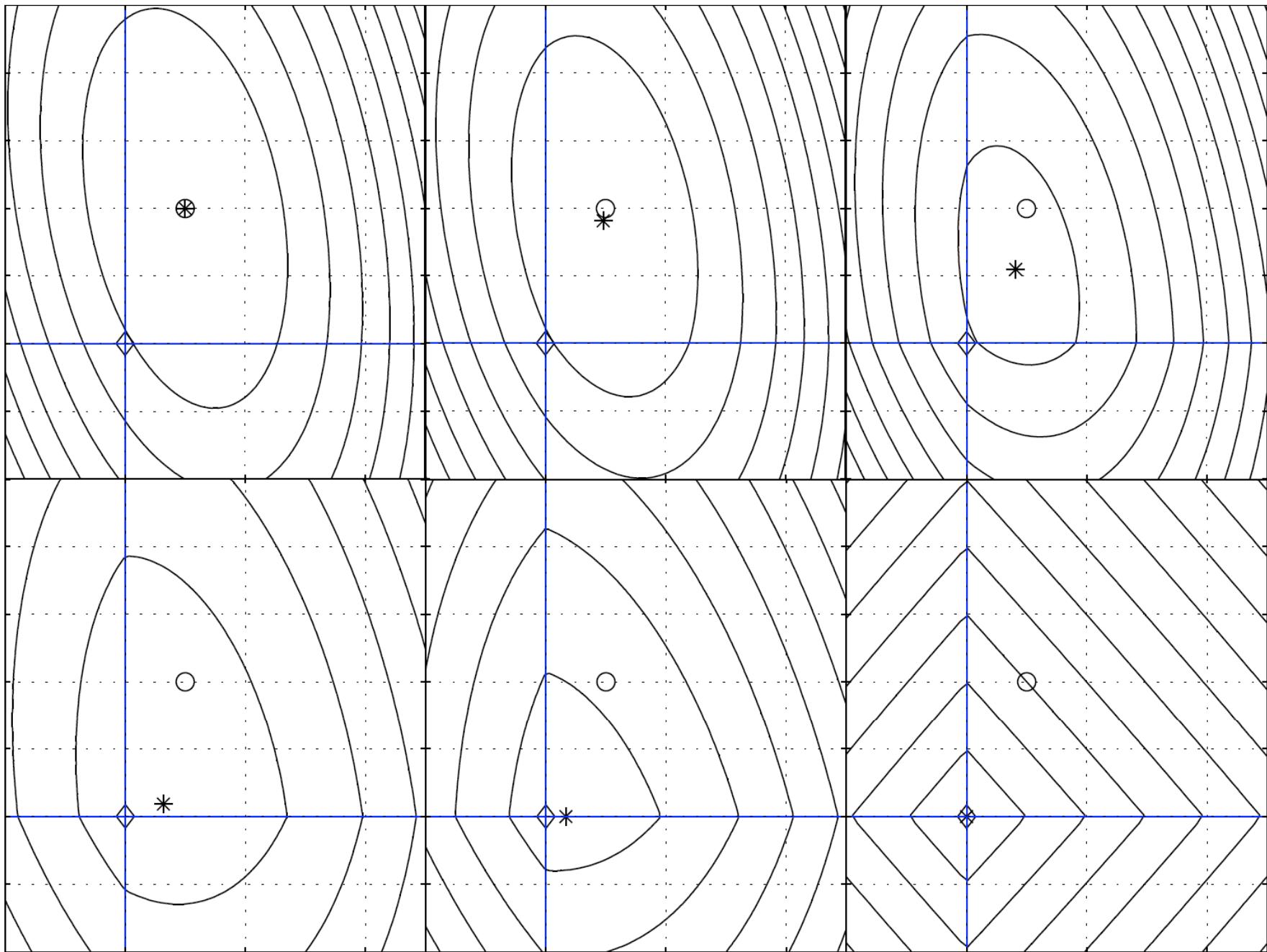
- Feature selection is a discrete process - individual variables are either in or out. Combinatorial nightmare.
- This method can have high variance - a different dataset from the same source can result in a totally different model
- Shrinkage methods allow a variable to be partly included in the model. That is, the variable is included but with a shrunk co-efficient
- Elegant way to tackle over-fitting

Posterior Modes with Varying Hyperparameter – Gaussian

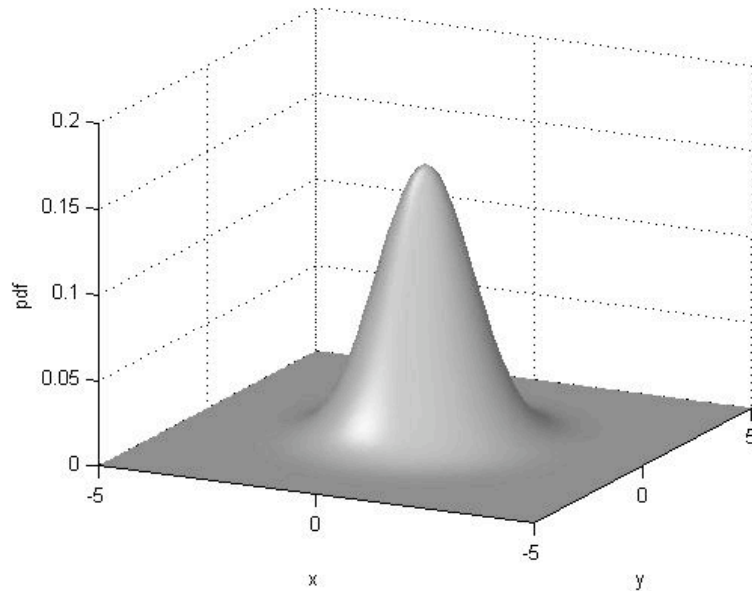


Posterior Modes with Varying Hyperparameter – Laplace

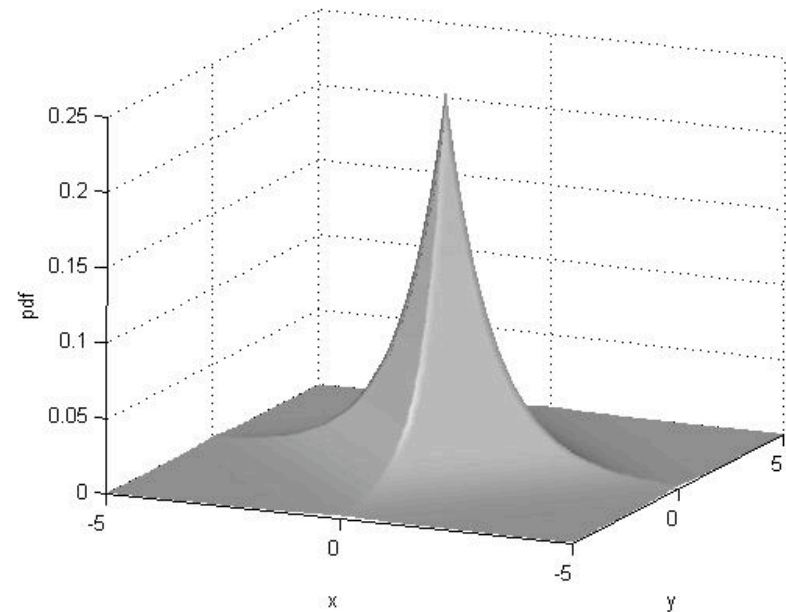




Bayesian Perspective



$$\beta_j \sim N(0, \tau^2)$$



$$\beta_j \sim N(0, \tau_j^2)$$

$$\tau_j^2 \sim \exp(\gamma)$$

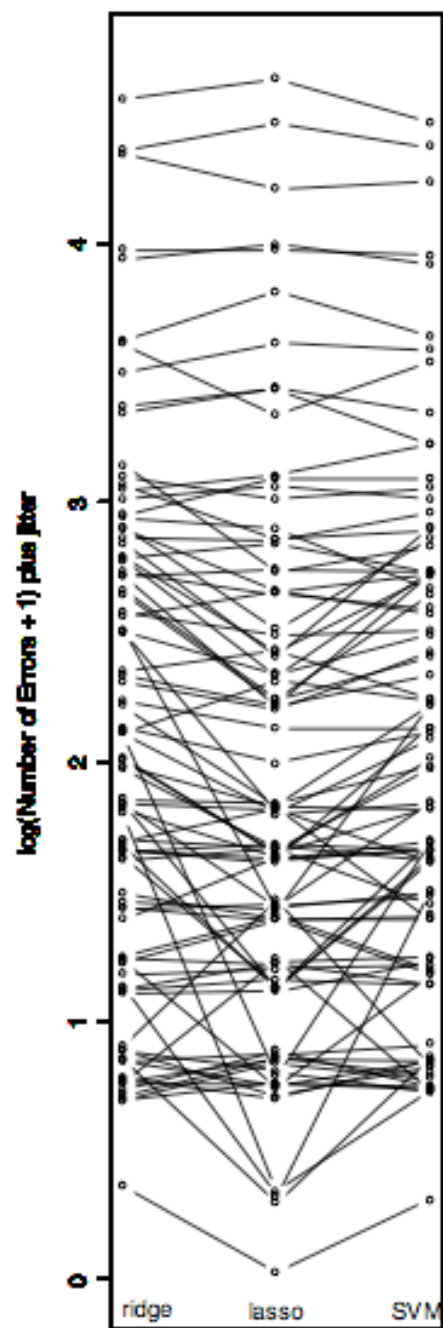
Data Sets

- ModApte subset of Reuters-21578
 - 90 categories; 9603 training docs; 18978 features
- Reuters RCV1-v2
 - 103 cats; 23149 training docs; 47152 features
- OHSUMED heart disease categories
 - 77 cats; 83944 training docs; 122076 features
- Cosine normalized TFxIDF weights

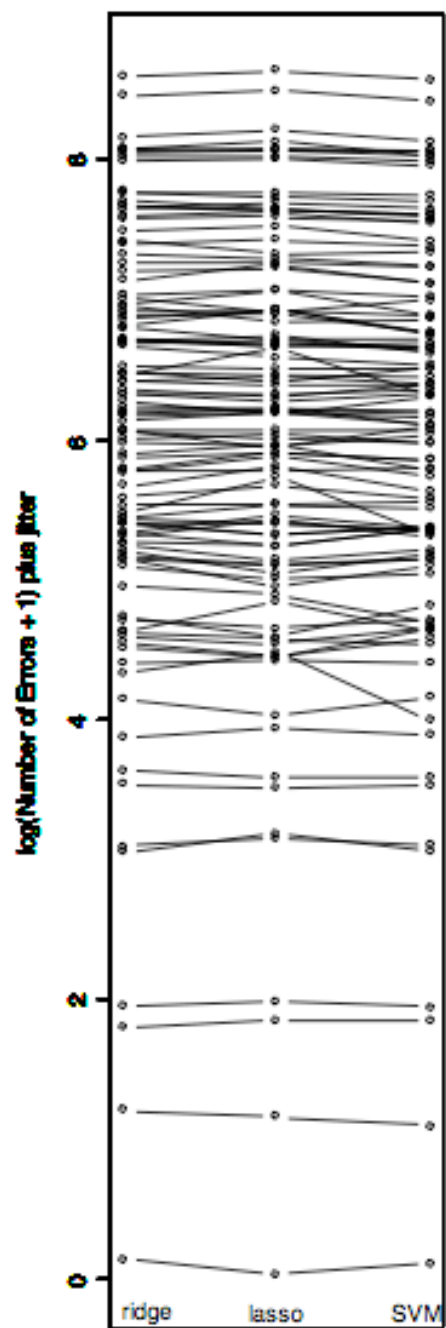
Dense vs. Sparse Models (Macroaveraged F1)

	ModApte	RCV1-v2	OHSUMED
Lasso	52.03	56.54	51.30
Ridge	39.71	51.40	42.99
Ridge/500	38.82	46.27	36.93
Ridge/50	45.80	41.61	42.59
Ridge/5	46.20	28.54	41.33
SVM	53.75	57.23	50.58

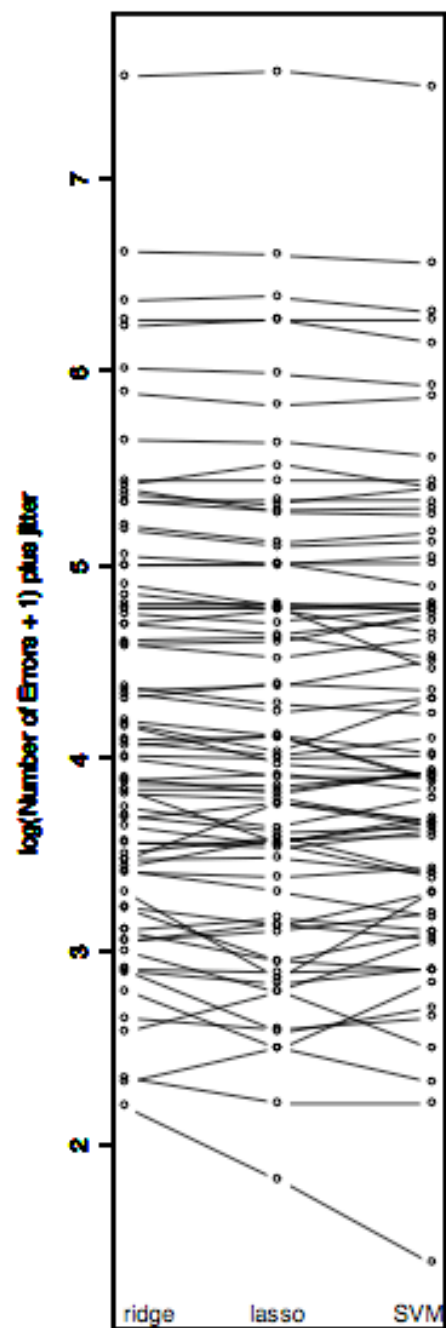
ModApte (90 categories)



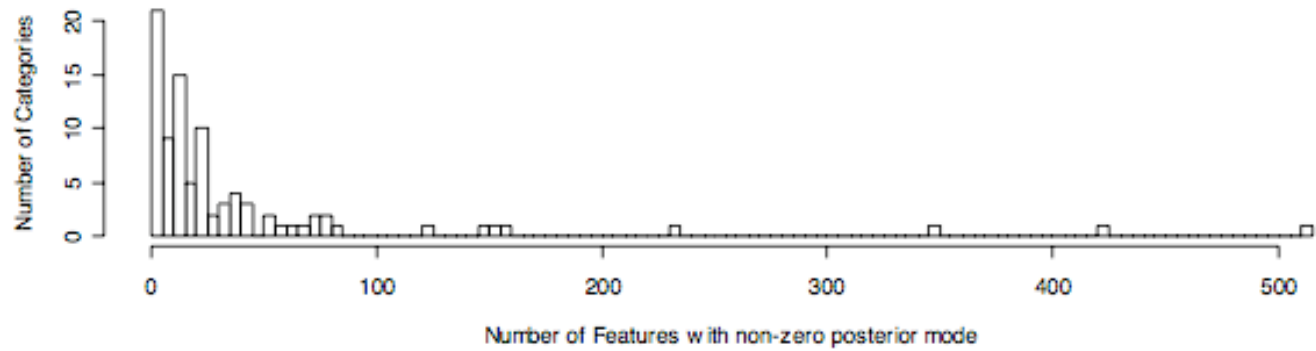
RCV1-v2 (103 categories)



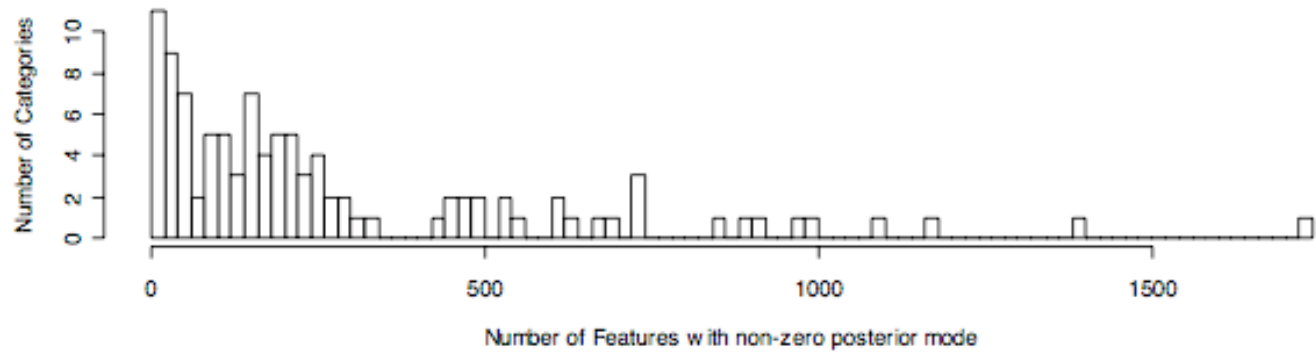
OHSUMED (77 categories)



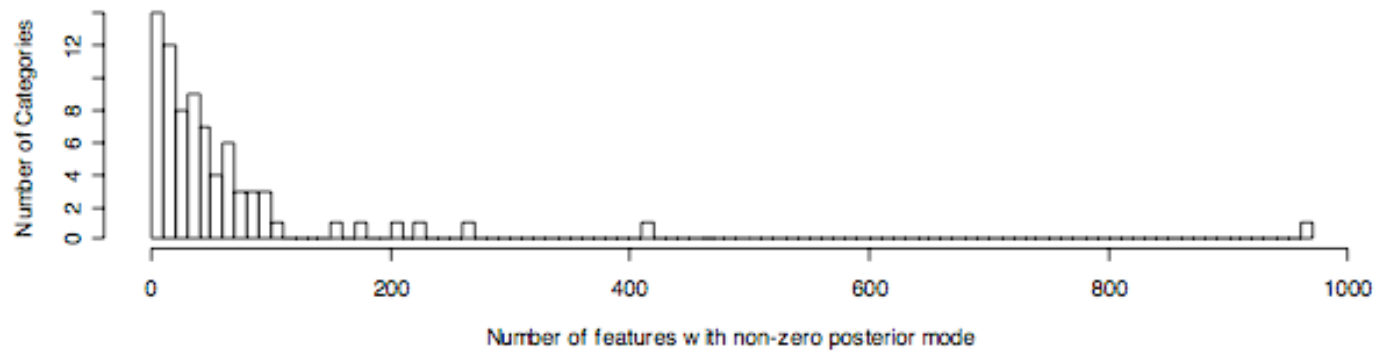
ModApte - 21,989 features



RCV1 - 47,152 features

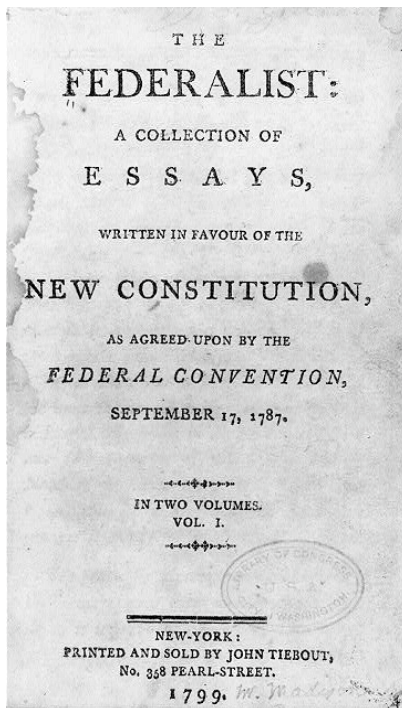


OHSUMED - 122,076 features



The Federalist

- “The authorship of certain numbers of the ‘Federalist’ has fairly reached the dignity of a well-established historical controversy.” (Henry Cabot Lodge, 1886)
- Historical evidence is muddled



Paper Number	Author
1	Hamilton
2-5	Jay
6-9	Hamilton
10	Madison
11-13	Hamilton
14	Madison
15-17	Hamilton
18-20	Joint: Hamilton and Madison
21-36	Hamilton
37-48	Madison
49-58	Disputed
59-61	Hamilton
62-63	Disputed
64	Jay
65-85	Hamilton



JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 302

JUNE, 1963

Volume 58

INFERENCE IN AN AUTHORSHIP PROBLEM^{1,2}

A comparative study of discrimination methods applied
to the authorship of the disputed *Federalist* papers

FREDERICK MOSTELLER

Harvard University

and

Center for Advanced Study in the Behavioral Sciences

AND

DAVID L. WALLACE

University of Chicago


- Used function words with Naïve Bayes with Poisson and Negative Binomial model
- Out-of-sample predictive performance

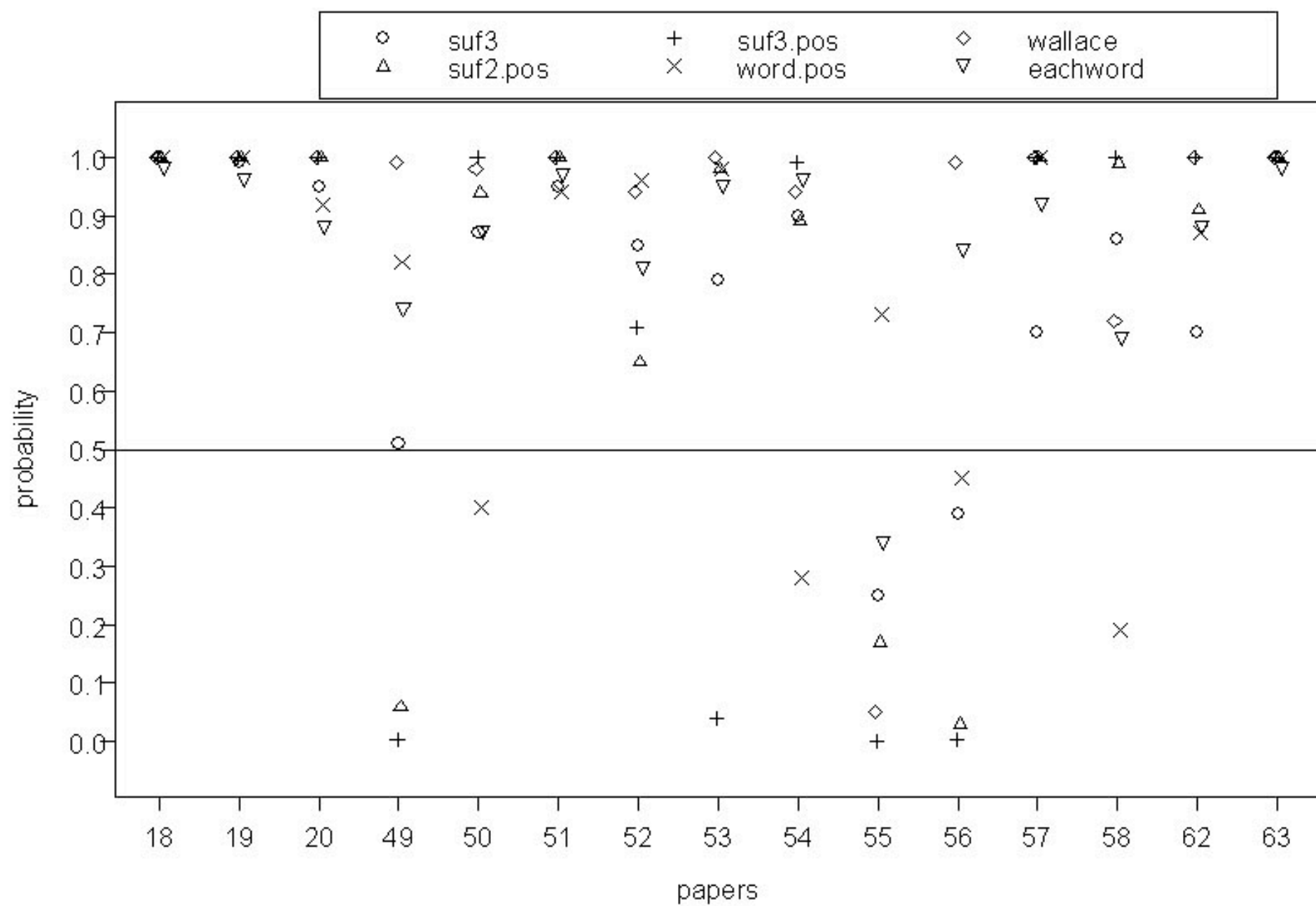
F. Summing up

In summary, the following points are clear:

1) Madison is the principal author. These data make it possible to say far more than ever before that the odds are enormously high that Madison wrote the 12 disputed papers. Weakest support is given for No. 55. Support for Nos. 62 and 63, most in doubt by current historians, is tremendous.

Feature Set	10-fold Error Rate
Charcount	0.21
POS	0.19
Suffix2	0.12
Suffix3	0.09
Words	0.10
Charcount+POS	0.12
Suffix2+POS	0.08
Suffix3+POS	0.04
Words+POS	0.08
484 features	0.05
Wallace features	0.05
Words (≥ 2)	0.05
Each Word	0.05





Polytomous Logistic Regression (PLR)

$$P(y_i = k \mid \mathbf{x}_i) = \frac{\exp(\vec{\beta}_k \mathbf{x}_i)}{\sum_{k'} \exp(\vec{\beta}_{k'} \mathbf{x}_i)}$$

- Elegant approach to multiclass problems
- Also known as *polychotomous LR*, *multinomial LR*, and, ambiguously, *multiple LR* and *multivariate LR*

1-of-K Sample Results: brittany-l

Feature Set	% errors	Number of Features
“Argamon” function words, raw tf	74.8	380
POS	75.1	44
1suff	64.2	121
1suff*POS	50.9	554
2suff	40.6	1849
2suff*POS	34.9	3655
3suff	28.7	8676
3suff*POS	27.9	12976
3suff+POS+3suff*POS+Argamon	27.6	22057
All words	23.9	52492

4.6 million parameters



89 authors with at least 50 postings. 10,076 training documents, 3,322 test documents.

BMR-Laplace classification, default hyperparameter

Term Weighting

- How strongly does a particular word indicate the content of a document?
- Some clues:
 - Number of times word occurs in this document
 - Number of times word occurs in other documents
 - Length of document

TF (term frequency)

IDF (inverse document frequency)

$$w_{ij}^{\text{raw}} = \begin{cases} (1 + \ln f_{ij}) \ln \frac{N}{n_j}, & \text{if } t_j \text{ present in } d_i \\ 0, & \text{otherwise} \end{cases}$$

$$w_{ij} = \frac{w_{ij}^{\text{raw}}}{\sqrt{\sum_{j'=1}^d w_{ij'}^{\text{raw}} \times w_{ij'}^{\text{raw}}}}$$

Set L2-norm to 1.0

- “Cosine-normalized TFIDF weighting”
 - Many minor variants on this theme

Variants on Term Weighting

- Explicit models of word frequency (e.g. Poisson mixtures, multinomial,...)
- Smoothing
 - Across similar documents
 - Across similar terms
- Trainable term weighting for complex features (e.g. Darmstadt indexing approach)

Domain Knowledge in Text Classification

- Certain words are positively or negatively associated with category
- **Domain Knowledge:** textual descriptions for categories
- Prior mean quantifies the strength of positive or negative association
- Prior variance quantifies our confidence in the domain knowledge

An Example Model

(category “grain”)

Word	Beta		Word	Beta
corn	29.78		formal	-1.15
wheat	20.56		holder	-1.43
rice	11.33		hungarian	-6.15
sindt	10.56		rubber	-7.12
madagascar	6.83		special	-7.25
import	6.79	
grain	6.77		beet	-13.24
contract	3.08		rockwood	-13.61

Using Domain Knowledge (DK)

- Give domain words higher mean or variance
- **Two methods:** For each DK term t and category q , and manually chosen C ,
 - First method sets **DK-based variance:**

$$\text{variance}(t, q) = C \cdot \text{significance}(t, q) \cdot \sigma^2$$

- Second method sets **DK-based mode:**

$$\text{mode}(t, q) = C \cdot \text{significance}(t, q) \cdot \sigma$$

Here σ^2 is variance for all other words chosen by 5-fold CV on training data

- Used TFxIDF weighting on the prior knowledge documents to compute $\text{significance}(t, q)$

Experiments

- Data sets
 - 1) TREC 2004 Genomics data:
 - **Categories:** 32 MeSH categories under “Cells” hierarchy
 - **Documents:** 3742 training and 4175 test
 - **Prior Knowledge:** MeSH category descriptions
 - 2) ModApte subset of Reuters-21578
 - **Categories:** 10 most frequent categories
 - **Documents:** 9603 training and 3299 test
 - **Prior Knowledge:** keywords selected by hand (Wu & Srihari, 2004)
- Big (all training examples) and small size training data
- Limited, biased data often the case in applications

MeSH Prior Knowledge Example

- **MeSH Heading:** Neurons
- **Scope Note:** The basic cellular units of nervous tissue. Each neuron consists of a body, an axon, and dendrites. Their purpose is to receive, conduct, and transmit impulses in the nervous system.
- **Entry Term:** Nerve Cells
- **See Also:** Neural Conduction

MeSH Results (Big training data)

	Macro F1	ROC
Laplace	50.2	88.7
Laplace & DK-based variance	53.7	89.2
Laplace & DK-based mode	52.8	89.4

MeSH Results

(training: 500 random examples)

	Macro F1	ROC
Laplace	35.1	78.3
Laplace & DK-based variance	49.7	83.8
Laplace & DK-based mode	44.4	84.2

MeSH Results

(training: 5 positive and 5 random examples for each category)

	Macro F1	ROC
Laplace	29.3	65.9
Laplace & DK-based variance	43.7	77.6
Laplace & DK-based mode	35.8	83.3

Prior Knowledge for ModApte

Category	Prior Knowledge
earn	cents cts net profit quarter qtr revenue rev share shr
acq	acquire acquisition company merger stake
money-fx	bank currency dollar money
grain	agriculture corn crop grain wheat usda
crude	barrel crude oil opec petroleum
trade	deficit import surplus tariff trade
interest	bank money lend rate
wheat	wheat
ship	port ship tanker vessel warship
corn	corn

ModApte Results

(training: 100 random samples)

	Macro F1	ROC
Laplace	37.2	76.2
Laplace & DK-based variance	65.3	87.1
Laplace & DK-based mode	72.0	93.5

ModApte Results

(training: 5 positive + 5 random samples for each category)

	Macro F1	ROC
Laplace	42.7	77.8
Laplace & DK-based variance	63.8	88.1
Laplace & DK-based mode	66.5	94.4