How to Read 100 Million Blogs (& Classify Deaths Without Physicians)

Gary King
Harvard University

August 23, 2007

References

- Daniel Hopkins and Gary King. "Extracting Systematic Social Science Meaning from Text"
- Gary King and Ying Lu. "Verbal Autopsy Methods with Multiple Causes of Death," tentatively to appear, *Statistical Science*
- Copies at `http://gking.harvard.edu`

Inputs and Target Quantities of Interest

- Input Data:
  - Large set of text documents
  - A set of (mutually exclusive and exhaustive) categories
  - A small set of documents hand-coded into the categories
- Quantities of interest
  - individual document classifications (spam filters)
  - proportion in each category (proportion email which is spam)
- Estimation
  - *Can* get the 2nd by counting the 1st (turns out not to be necessary!)
  - High classification accuracy $\not\Rightarrow$ unbiased category proportions

Blogs as a Running Example

- Blogs (web logs): web version of a daily diary, with posts listed in reverse chronological order.
- 8% of U.S. Internet users (12 million) have blogs
- Growth: ≈ 0 in 2000; 44–100 million worldwide now.
- A democratic technology: 6 million in China and 700,000 in Iran
- "We are living through the largest expansion of expressive capability in the history of the human race"

One specific quantity of interest

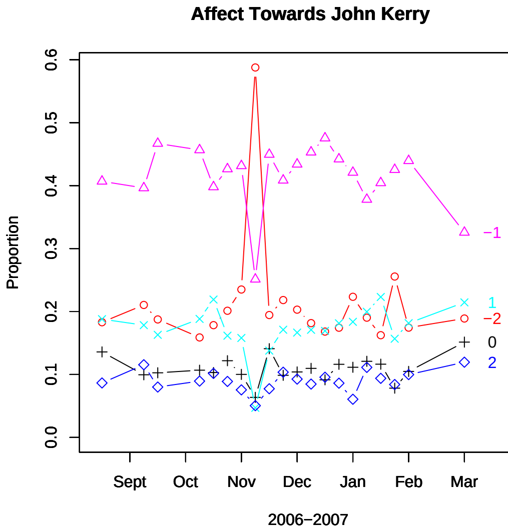
- Affect about President Bush and 2008 candidates
- Specific categories:

Label	Category
-2	extremely negative
-1	negative
0	neutral
1	positive
2	extremely positive
NA	no opinion expressed
NB	not a blog

- Hard case:
  - Part ordinal, part nominal categorization
  - "Sentiment categorization is more difficult than topic classification"
  - Informal language: "my crunchy gf thinks dubya hid the wmd's, :)!"
  - Little common internal structure (no inverted pyramid)

The Conversation about John Kerry's Botched Joke

You know, education — if you make the most of it . . . you can do well. If you don't, you get stuck in Iraq.



Representing Text as Numbers

- Filter: choose English language blogs that mention Bush
- Preprocess: convert to lower case, remove punctuation, keep only word stems ("consist", "consisted", "consistency" $\rightsquigarrow$ "consist")
- Code variables: presence/absence of unique unigrams, bigrams, trigrams
- Our Example:
  - Our 10,771 blog posts about Bush and Clinton: 201,676 unigrams, 2,392,027 bigrams, 5,761,979 trigrams.
  - keep only unigrams in $> 1\%$ or $< 99\%$ of documents: 3,672 variables
  - Groups infinite possible posts into "only" $2^{3,672}$ distinct types

Notation

- Document Category

$$D_i = \begin{cases} \text{-2} & \text{extremely negative} \\ \text{-1} & \text{negative} \\ 0 & \text{neutral} \\ 1 & \text{positive} \\ 2 & \text{extremely positive} \\ \text{NA} & \text{no opinion expressed} \\ \text{NB} & \text{not a blog} \end{cases}$$

- Word Stem Profile:

$$\mathbf{S}_i = \begin{cases} S_{i1} = 1 & \text{if ``awful'' is used, 0 if not} \\ S_{i2} = 1 & \text{if ``good'' is used, 0 if not} \\ \vdots & \vdots \\ S_{iK} = 1 & \text{if ``except'' is used, 0 if not} \end{cases}$$

Quantities of Interest

- Computer Science: individual document classifications

$$D_1, D_2 \ldots, D_L$$

- Social Science: proportions in each category

$$P(D) = \begin{pmatrix} P(D=-2) \\ P(D=-1) \\ P(D=0) \\ P(D=1) \\ P(D=2) \\ P(D=\text{NA}) \\ P(D=\text{NB}) \end{pmatrix}$$

Issues with Existing Statistical Approaches

1. Direct Sampling
   - Biased without a random sample
   - nonrandomness common due to population drift, data subdivisions, etc.
   - (Classification of population documents not necessary)
2. Aggregation of model-based individual classifications
   - Biased without a random sample
   - Models $P(D|\mathbf{S})$, but the world works as $P(\mathbf{S}|D)$
   - Bias unless
     - $P(D|\mathbf{S})$ encompasses the "true" model.
     - $\mathbf{S}$ spans the space of all predictors of D (i.e., all information in the document)
   - Bias even with optimal classification and high % correctly classified

Using Misclassification Rates to Correct Proportions

- Use some method to classify unlabeled documents
- Aggregate classifications to category proportions
- Use labeled set to estimate misclassification rates (by cross-validation)
- Use misclassification rates to correct proportions
- Result: vastly improved estimates of category proportions
- (No new assumptions beyond that of the classifier)
- (still requires random samples, individual classification, etc)

Formalization from Epidemiology

(Levy and Kass, 1970)

- Accounting identity for 2 categories:

$$P(\hat{D}=1) = (\text{sens})P(D=1) + (1-\text{spec})P(D=2)$$

- Solve:

$$P(D=1) = \frac{P(\hat{D}=1) - (1-\text{spec})}{\text{sens} - (1-\text{spec})}$$

- Use this equation to correct $P(\hat{D})$

Generalizations: J Categories, No Individual Classification
(King and Lu, 2007)

- Accounting identity for J categories

$$P(\hat{D}=j)=\sum_{j'=1}^{J}P(\hat{D}=j|D=j')P(D=j')$$

- Drop $\hat{D}$ calculation, since $\hat{D}=f(\mathbf{S})$:

$$P(\mathbf{S}=s)=\sum_{j'=1}^{J}P(\mathbf{S}=s|D=j')P(D=j')$$

- Simplify to an equivalent matrix expression:

$$P(\mathbf{S})=P(\mathbf{S}|D)P(D)$$

Estimation

The matrix expression again:

$$\underset{2^K \times 1}{P(\mathbf{S})} = \underset{2^K \times J}{P(\mathbf{S}|D)}\underset{J \times 1}{P(D)}$$

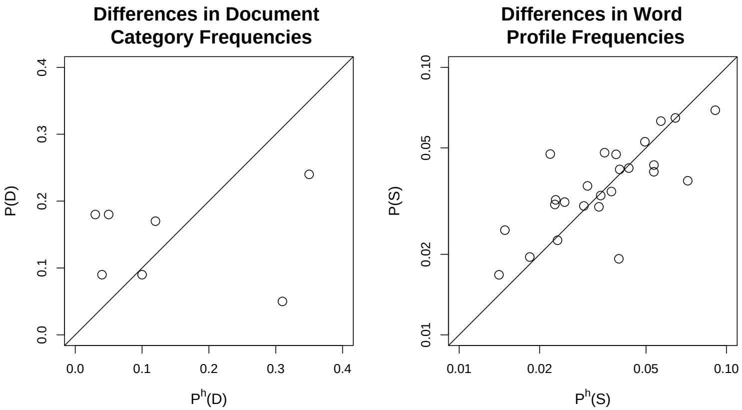
$$\implies Y = X\beta \quad \implies \quad \beta = (X'X)^{-1}X'y$$

Document category proportions (quantity of interest) Word stem profile

proportions (estimate in unlabeled set by tabulation) Word stem profiles, by category (estimate in *labeled* set by tabulation) Alternative symbols (to emphasize the linear equation) Solve for quantity of interest (with no error term)

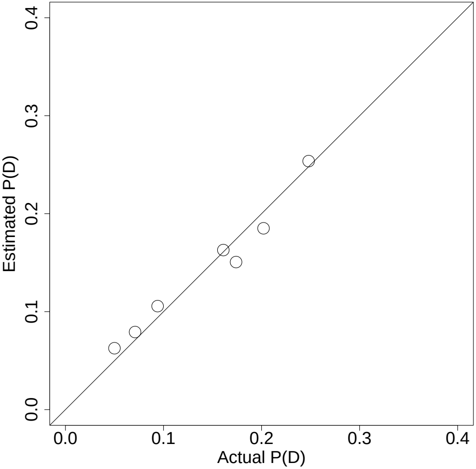
- Technical estimation issues:
  - 2^K is enormous, far larger than any existing computer
  - $P(\mathbf{S})$ and $P(\mathbf{S}|D)$ will be too sparse
  - Elements of P(D) must be between 0 and 1 and sum to 1
- Solutions

A Nonrandom Hand-coded Sample

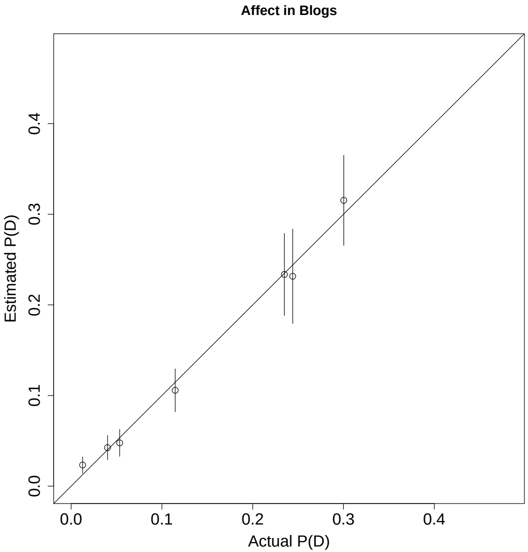


All existing methods would fail with these data.

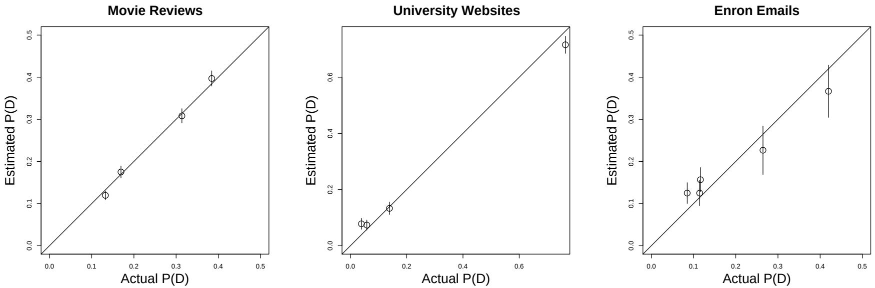
Accurate Estimates



Out of Sample Validation: Blogs



Out of Sample Validation: Other Examples



Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality rate to set research goals, budgetary priorities, and ameliorative policies
  - High quality death registration: only 23/192 countries
- Existing Approaches
  - Ask relatives or caregivers 50-100 symptom questions
  - Ask physicians to determine cause of death (low intercoder reliability)
  - Apply expert algorithms (high reliability, low validity)
  - Find deaths with medically certified causes from a local hospital, trace caregivers to their homes, ask the same symptom questions, and statistically classify deaths in population (model-dependent, low accuracy)

An Alternative Approach

- ~~Document Category~~, Cause of Death,

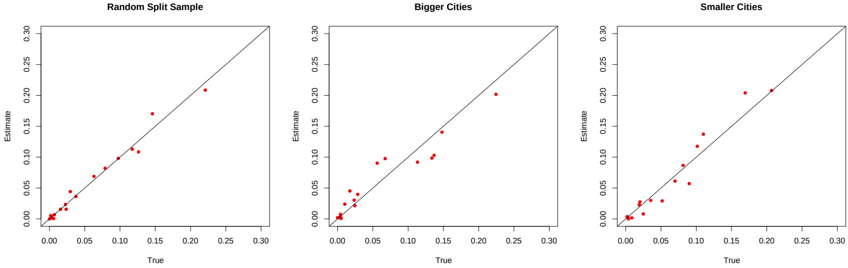
$$D_i = \begin{cases} 1 & \text{if bladder cancer} \\ 2 & \text{if cardiovascular disease} \\ 3 & \text{if transportation accident} \\ \vdots & \vdots \\ J & \text{if infectious respiratory} \end{cases}$$

- ~~Word Stem Profile~~, Symptoms:

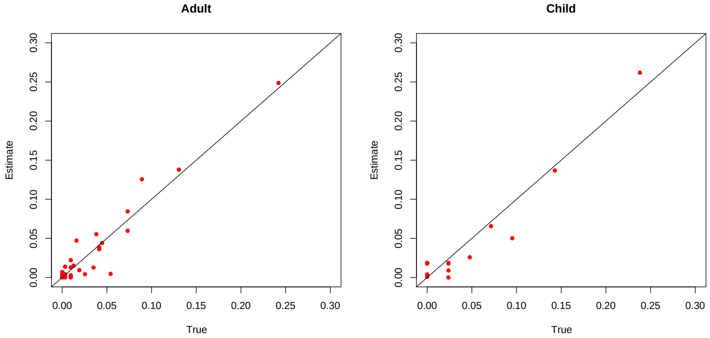
$$\mathbf{S}_i = \begin{cases} S_{i1} = 1 & \text{if ``breathing difficulties'', 0 if not} \\ S_{i2} = 1 & \text{if ``stomach ache'', 0 if not} \\ \vdots & \vdots \\ S_{iK} = 1 & \text{if ``diarrhea'', 0 if not} \end{cases}$$

- Apply the same methods

Validation in China



Validation in Tanzania



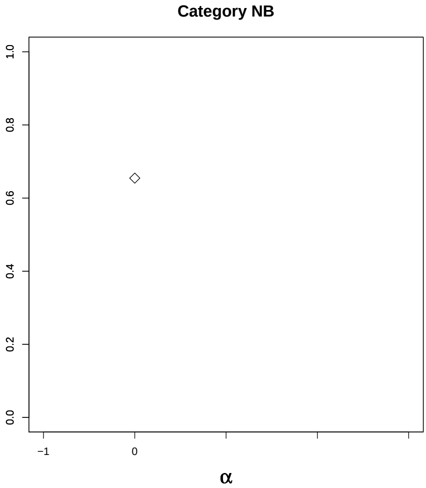
For more information

http://GKing.Harvard.edu

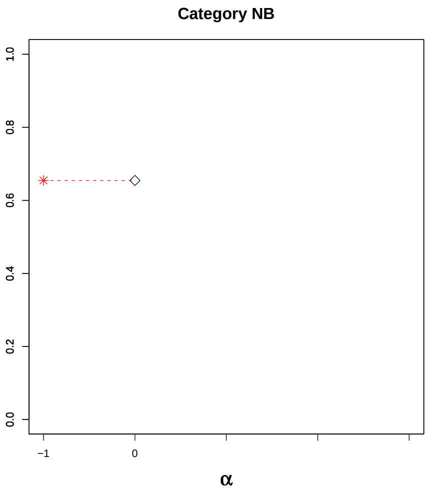
Misclassification Matrix for Blog Posts

	-2	-1	0	1	2	NA	NB	$P(D_1)$
-2	**.70**	.10	.01	.01	.00	.02	.16	.28
-1	.33	**.25**	.04	.02	.01	.01	.35	.08
0	.13	.17	**.13**	.11	.05	.02	.40	.02
1	.07	.06	.08	**.20**	.25	.01	.34	.03
2	.03	.03	.03	.22	**.43**	.01	.25	.03
NA	.04	.01	.00	.00	.00	**.81**	.14	.12
NB	.10	.07	.02	.02	.02	.04	**.75**	.45

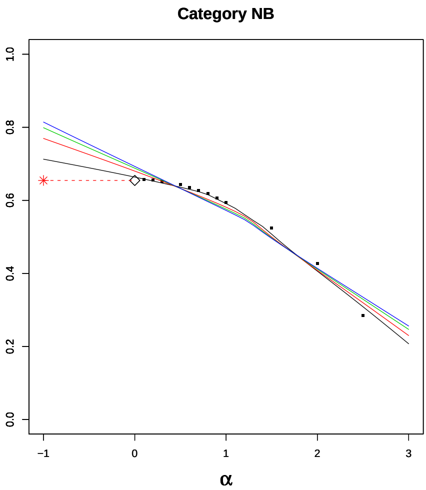
SIMEX Analysis of "Not a Blog" Category



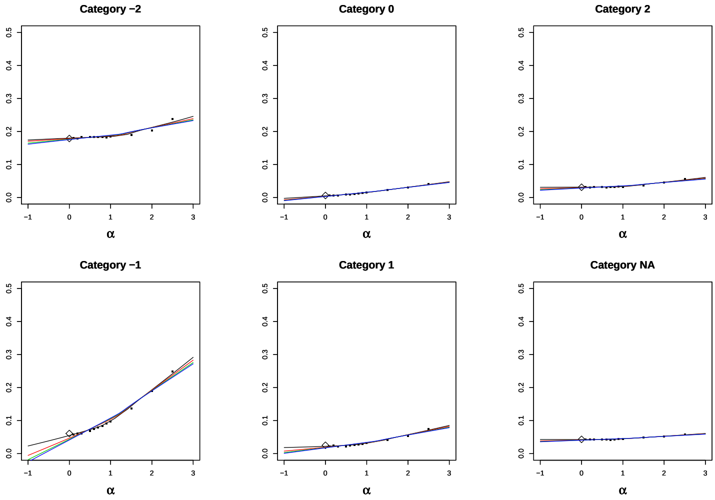
SIMEX Analysis of "Not a Blog" Category



SIMEX Analysis of “Not a Blog” Category



SIMEX Analysis of Other Categories



What can go wrong?

- We assume $P^h(\mathbf{S}|D) = P(\mathbf{S}|D)$
- Must choose word stem subset size (a smoothing parameter)
- Need enough labeled documents in each category (can hand code more if CI's are too large, perhaps via case-control methods)
- Need sufficient information in: documents, categorization scheme, numerical summaries of the documents, and hand-codings
- Use additional hand coding to verify assumptions