Extracting Systematic Social Science Meaning from Text (& Cause-Specific Mortality Rates from Symptom Data)

Gary King
Harvard University

March 20, 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`

Content Analysis: Past and Future

- Dates to the 1600s: The Church tracked nonreligious texts by classifying newspaper stories
- Prominent early social scientists used it: Berelson, de Grazia, etc.
- Spread to vast array of fields
- Automated methods now joining hand coding
- Use increased six-fold 1980–2000
- Huge potential for new applications: explosive increase in web pages, blogs, emails, digitized books and articles, audio recordings (automatically converted to text), and government reports, legislative hearings and records, electronic medical records, etc.
- Infeasible to expand hand coding efforts much further
- Automated methods are essential

Inputs and Target Quantities of Interest

- Available inputs:
  - Large set of text documents
  - A set of mutually exclusive and exhaustive categories
  - A small subset of documents hand-coded into the categories
- Quantities of interest
  - Computer Science: individual document classification
  - Social Science: proportion of documents in each category
  - *Can* get the 2nd by aggregating the 1st (turns out not to be necessary!)
  - E.g., classify constituents' letters to a member of congress by policy area, or estimate proportion of letters in each policy area
  - E.g., classify emails as spam or not, or estimate proportion of email that is spam
- Maximizing one goal won't get you the other: high classification accuracy can coexist with huge biases in category proportions

Our Approach

- Gives unbiased estimates of population proportions
- Works better than aggregating the best classification method
- No problem if classification accuracy is low
- (And individual classification is not necessary)
- No parametric modeling assumptions
- The hand coded subset need not be a random sample
- Scales to large numbers of documents
- Separately: propose correction for imperfect inter-coder reliability (i.e., should work better than hand coding everything if that were feasible)

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
- Explosive growth: ≈ 0 in 2000 to 39–100 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

- Subject: the grand conversation about the American presidency
- Question: opinions 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"
  - Language ranges from "my crunchy gf thinks dubya hid the wmd's!" to the Queen's English
  - Little common internal structure (no inverted pyramid)

Representing Text as Numbers

- Filter: choose English language blogs that mention Bush ("Bush", "George W.", "Dubya", "King George", etc.), Hillary Clinton ("Senator Clinton", "Hillary", "Hitlery", "Mrs. Clinton"), etc.
- Preprocess: convert to lower case, remove punctuation, perform stemming (reduce "consist", "consisted", "consistency", "consistent", "consistently", "consisting", and "consists", to their stem: "consist")
- Code variables as presence or absence of unique unigrams, bigrams, trigrams, etc.
- Example:
  - Our 10,771 blog posts about Bush and Clinton: 201,676 unigrams, 2,392,027 bigrams, 5,761,979 trigrams.
  - 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
   - Classification of population documents not necessary
   - Biased if hand-coded documents are not random sample of population
   - nonrandomness common due to population drift, studying data subdivisions, etc.
2. Aggregation of model-based individual classifications
   - Biased if not 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)
   - Even optimal classification with high % correctly classified can produce biased estimates of proportions

Using Misclassification Rates to Correct Proportions

- Divide labeled set into training and test sets
- Use training set to classify test set (ignoring D) and determine misclassification rates (using D)
- Use entire labeled set to classify all unlabeled documents and aggregate to category proportions
- Use misclassification rates to correct:
  - Suppose we find that 12% of test set documents in category 2 should really have been in category 1
  - Correct proportions for the unlabeled set: subtract 12% from category 2 and add 12% to category 1
- Assumes only that misclassification rates were estimated well
- Estimates of category proportions: vastly improved

Formalization from Epidemiology (Levy and Kass, 1970)

- An accounting identity:

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

- Solve:

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

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

Generalize to J categories (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 the intermediate $\hat{D}$ 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:

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

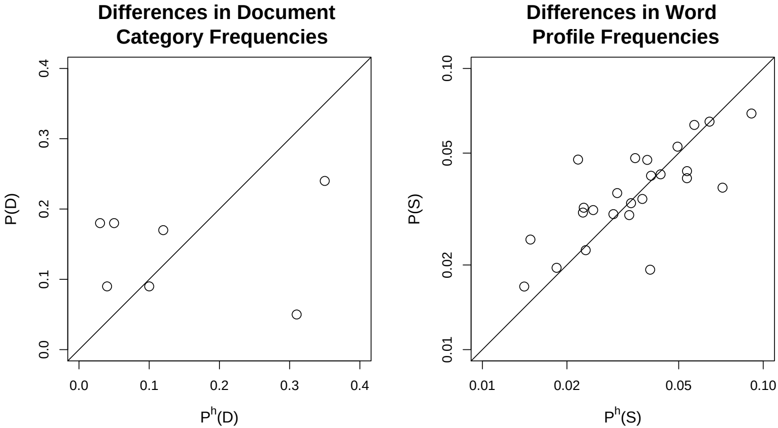
$$\Longrightarrow \; Y = X\beta \quad \Longrightarrow \quad \beta = (X'X)^{-1}X'y$$

Document category proportions (quantity of interest) Word stem profile

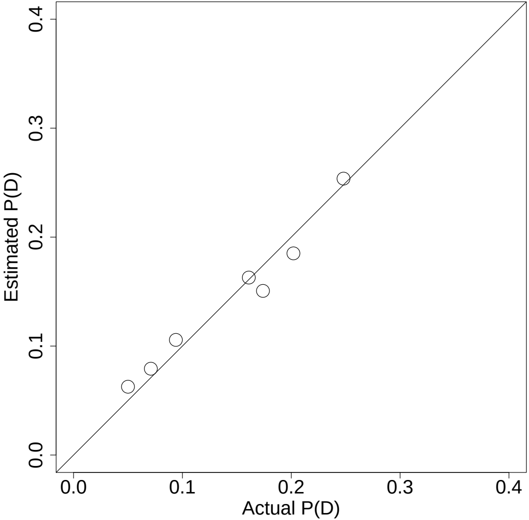
proportions (estimate in unlabeled set by tabulation) Word stem profiles, by category (estimate in *labeled* set by tabulation) Alternative symbols (to emphasize the linear equation) Solve for quantity of interest (with no error term)

- 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

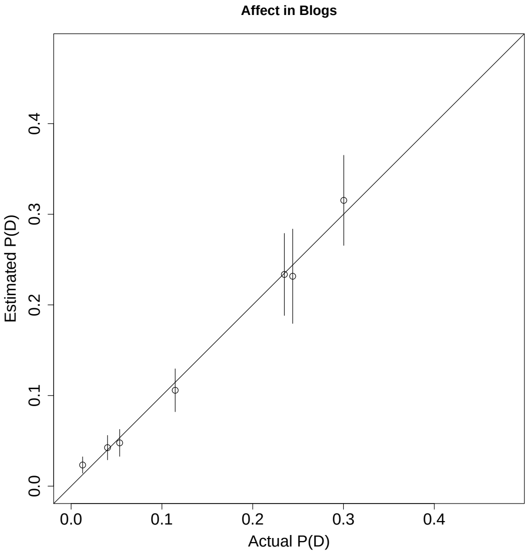
A Simulation with a Nonrandom Hand-coded Sample



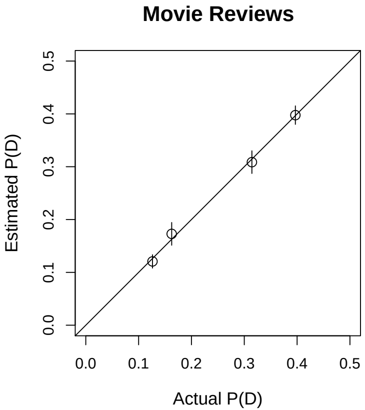
A Simulation: Accurate Estimates

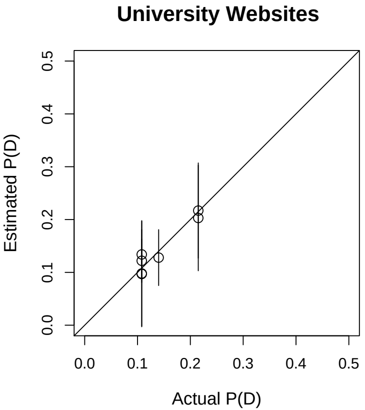


Out of Sample Validation: Blogs

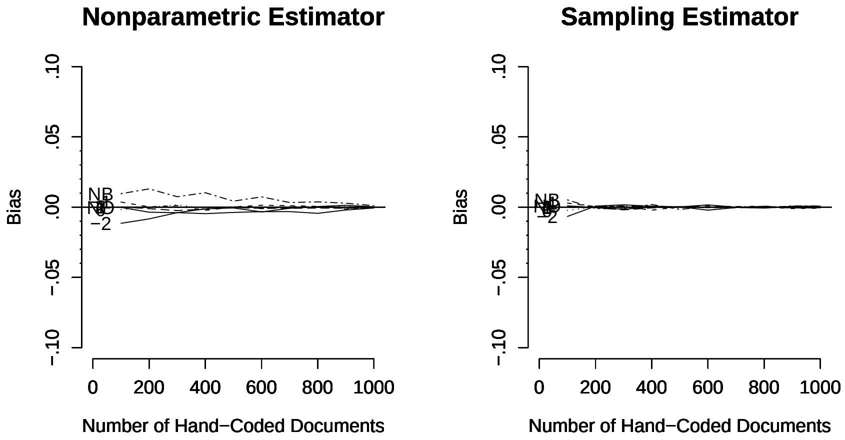


Out of Sample Validation: Other Examples

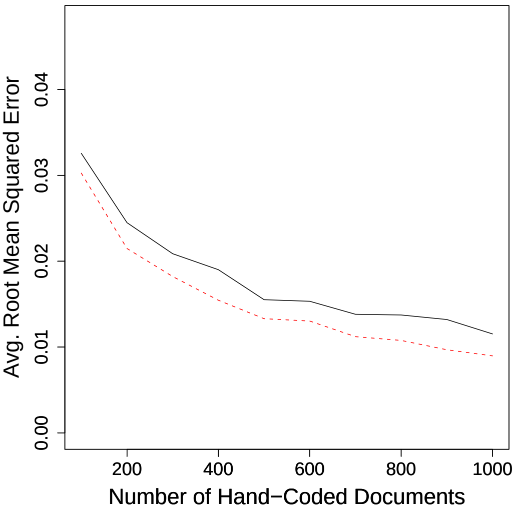




Bias by Number of Hand Coded Documents



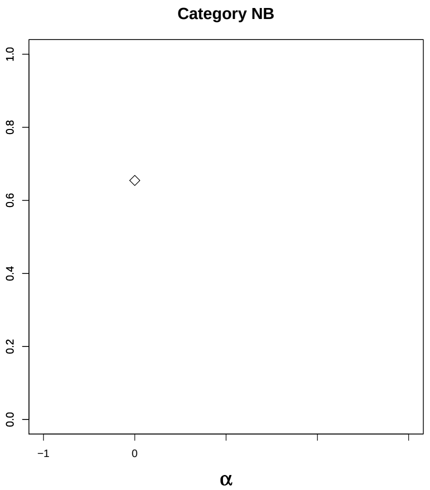
Average RMSE by Number of Hand Coded Documents



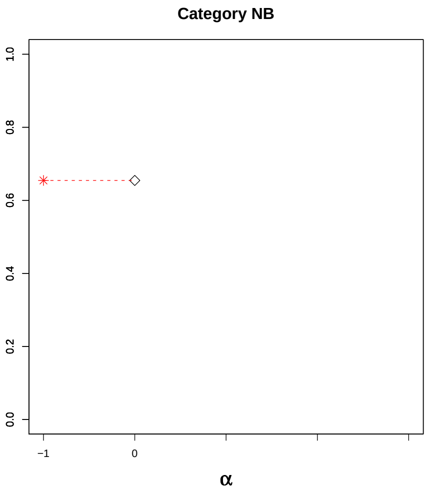
Misclassification Matrix for Blog Posts

	-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

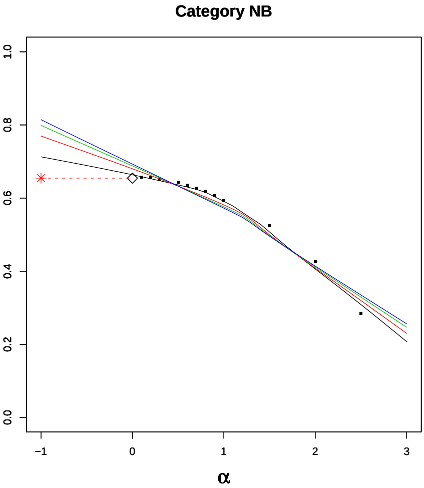
SIMEX Analysis of Not a Blog Category



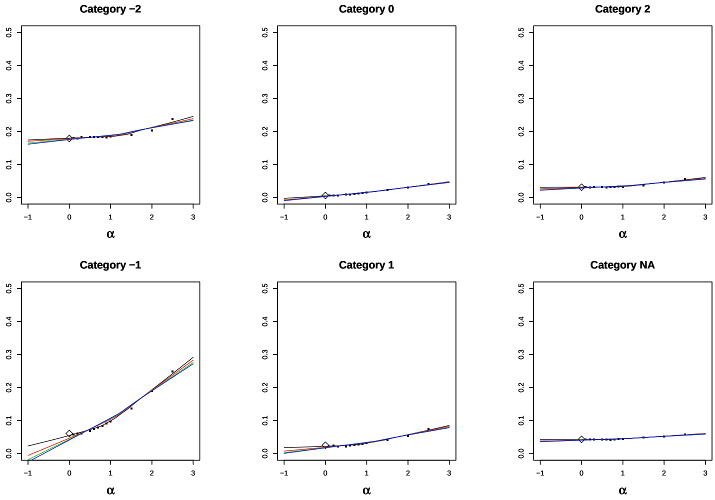
SIMEX Analysis of Not a Blog Category



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

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
  - 75 have no death registration at all
- The Approach
  - 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)

An Alternative Approach

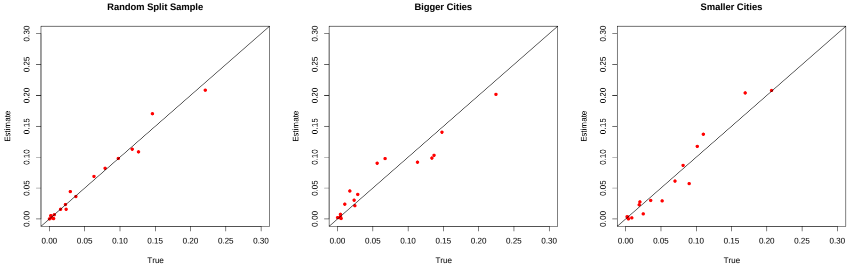
- ~~Document Category~~, Cause of Death,

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

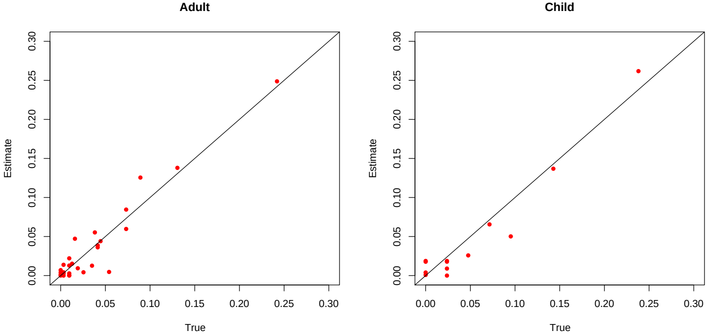
- ~~Word Stem Profile~~, Symptoms:

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

Validation in China



Validation in Tanzania



For more information

http://GKing.Harvard.edu