CS 6740/INFO 6300: A preface[1]

Polonius	What do you read, my lord?
Hamlet	Words, words, words.
Polonius	What is the matter, my lord?
Hamlet	Between who?
Polonius	I mean, the matter that you read, my lord.
Hamlet	Slanders, sir: for the satirical rogue says here that old men have grey beards....
Polonius	[*Aside*] Though this be madness, yet there is method in't.

–*Hamlet*, Act II, Scene ii.

[1]Students are not responsible for this material.

What is the matter?

Text categorization (broadly construed): identification of “similar” documents.

Similarity criteria include:

- **topic** (e.g., news aggregation sites)
- **source** (authorship or genre identification)
- **relevance** to a query (ad hoc information retrieval)
- **sentiment polarity**, or author’s overall opinion(data mining)
- **quality** (writing and language/learning aids/evaluators, user interfaces, plagiarism detection)

Method to the madness

For computers, understanding natural language is hard! What can we achieve within a "knowledge-lean" (but "data-rich") framework?

Act I: Iterative Residual Re-scaling: a generalization of Latent Semantic Indexing (LSI) that creates improved representations for topic-based categorization [Ando SIGIR '00, Ando & Lee SIGIR '01]

Act II: Sentiment analysis via minimum cuts: optimal incorporation of pair-wise relationships in a more semantically-oriented task using politically-oriented data [Pang & Lee ACL 2004, Thomas, Pang & Lee EMNLP 2006]

Act III How online opinions are received: an Amazon case study: discovery of new social/psychological biases that affect human quality judgments [Danescu-Niculescu-Mizil, Kossinets, Kleinberg, &Lee WWW 2009]

Words, words, words: the vector-space model

Documents:

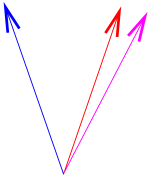
make
hidden
Markov
model
probabilities
normalize

car
emissions
hood
make
model
trunk

car
engine
hood
tires
truck
trunk

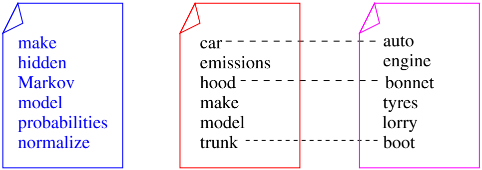
Term−document matrix D:

$$\begin{array}{cl} \left[\begin{matrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{matrix}\right] & \begin{matrix} \text{car} \\ \text{emissions} \\ \text{engine} \\ \text{hidden} \\ \text{hood} \\ \text{make} \\ \text{Markov} \\ \text{model} \\ \text{normalize} \\ \text{probabilities} \\ \text{tires} \\ \text{truck} \\ \text{trunk} \end{matrix} \end{array}$$



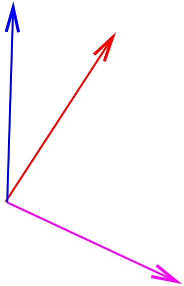
Problem: Synonymy

Documents:



Term−document matrix D:

$$\begin{array}{l} \left[\begin{array}{ccc} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{l} \text{auto} \\ \text{bonnet} \\ \text{boot} \\ \text{car} \\ \text{emissions} \\ \text{engine} \\ \text{hidden} \\ \text{hood} \\ \text{lorry} \\ \text{make} \\ \text{Markov} \\ \text{model} \\ \text{normalize} \\ \text{probabilities} \\ \text{tires} \\ \text{trunk} \\ \text{tyres} \end{array}$$



One class of approaches: Subspace projection

Project the document vectors into a lower-dimensional subspace.

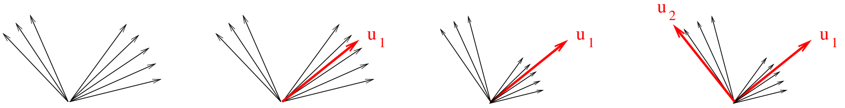
$\triangleright$ Synonyms no longer correspond to orthogonal vectors, so topic and directionality may be more tightly linked.

Most popular choice: Latent Semantic Indexing (LSI) [Deerwester et al., 1990]

- Pick some number k that is smaller than the rank of the term-document matrix D.
- Compute the first k *left singular vectors* $u_1, u_2, \ldots, u_k$ of D.
- Create $\mathbf{D}' :=$ the projection of $\mathbf{D}$ onto $\mathrm{span}(\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k})$.

Motivation: D' is the two-norm-optimal rank-k approximation to D [Eckart and Young, 1936].

A geometric view



Start with document vectors

Choose direction u maximizing projections

Compute ***residuals*** (subtract projections)

Repeat to get next u (orthogonal to previous u_i's)

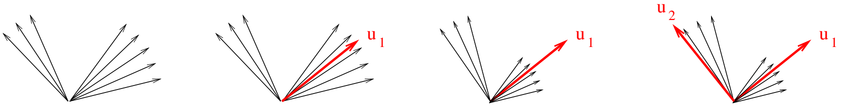
That is, in each of k rounds, find

$u = \arg\max_{x:|x|=1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(x, r_j))$ ("weighted average")

But, is the induced optimum rank-k approximation to the original term-document matrix *also* the optimal representation of the documents?

Results are mixed; e.g., Dumais et al. (1998).

A geometric view



Start with document vectors

Choose direction u maximizing projections

Compute ***residuals*** (subtract projections)

Repeat to get next u (orthogonal to previous u_i's)

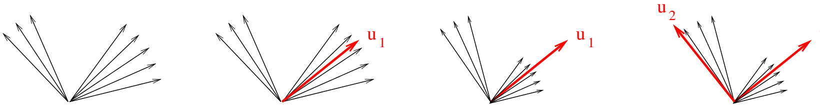
That is, in each of k rounds, find

$u = \arg\max_{x:|x|=1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(x, r_j))$ ("weighted average")

But, is the induced optimum rank-k approximation to the original term-document matrix *also* the optimal representation of the documents?

Results are mixed; e.g., Dumais et al. (1998).

A geometric view



Start with document vectors

Choose direction u maximizing projections

Compute ***residuals*** (subtract projections)

Repeat to get next u (orthogonal to previous u_i's)

That is, in each of k rounds, find

$$u = \arg\max_{x:|x|=1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(x, r_j)) \qquad (\text{“weighted average”})$$

But, is the induced optimum rank-k approximation to the original term-document matrix *also* the optimal representation of the documents?

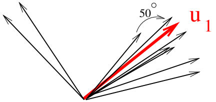
Results are mixed; e.g., Dumais et al. (1998).

Arrows of outrageous fortune

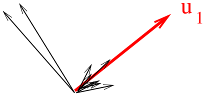
Recall: in each of k rounds, LSI finds

$u = \arg\max_{x:|x|=1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(x, r_j))$

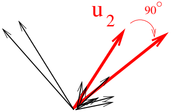
Problem: Non-uniform distributions of topics among documents



Choose direction u
maximizing projections



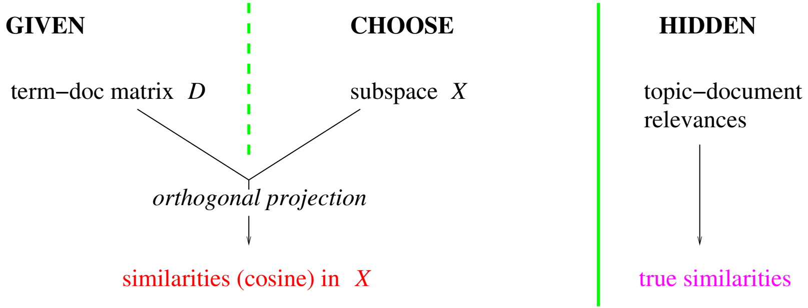
Compute residuals



Repeat to get next u
(orthogonal to previous u_i's)

dominant topics bias the choice

Gloss of main analytic result



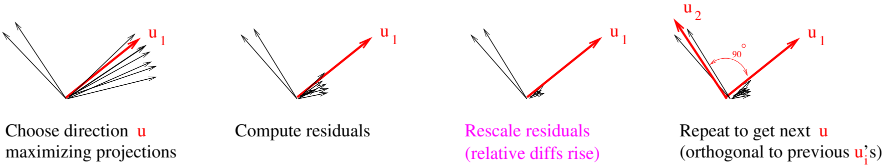
Under mild conditions, the distance between X^{LSI} and $X^{optimal}$ is bounded by a function of the topic-document distribution's non-uniformity and other reasonable quantities, such as D's "distortion".

Cf. analyses based on generative models [Story, 1996; Ding, 1999; Papadimitriou et al., 1997, Azar et al., 2001] and empirical observations comparing X^{LSI} with an optimal subspace [Isbell and Viola, 1998].

By indirections find directions out

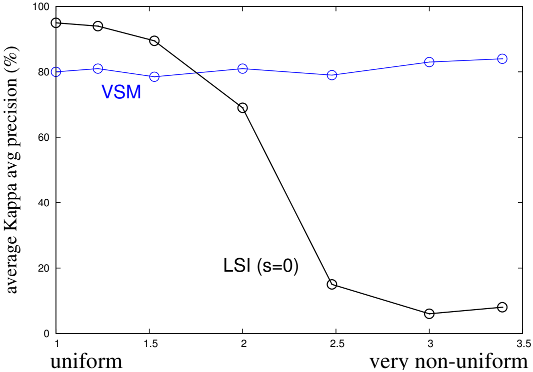
Recall: $u = \arg\max_{x:|x|=1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(x, r_j))$.

We can compensate for non-uniformity by re-scaling the residuals: $r_j \to |r_j|^s \cdot r_j$, where s is a *scaling factor* [Ando, 2000].



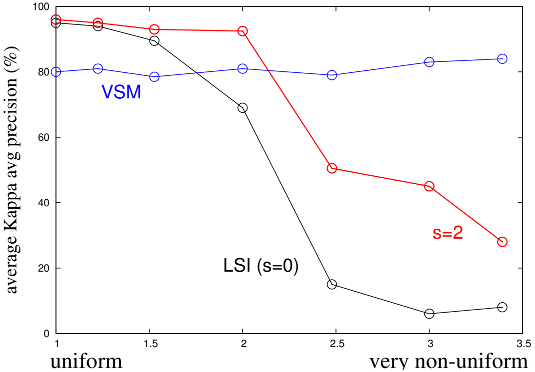
The **Iterative Residual Re-scaling** algorithm (IRR) estimates the (unknown) non-uniformity to *automatically* set the scaling factor s.

One set of experiments



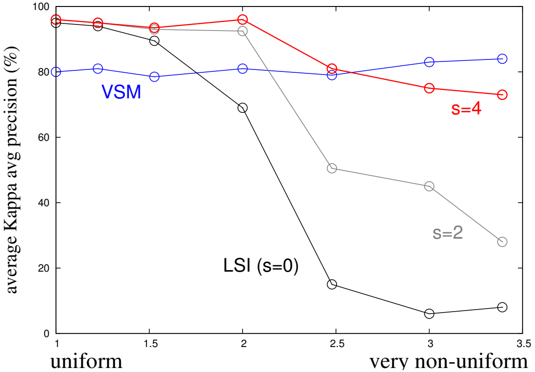
Each point: average over 10 different single-topic TREC-document datasets of the given non-uniformity.
(Analysis does not assume single-topic documents)

One set of experiments



Each point: average over 10 different single-topic TREC-document datasets of the given non-uniformity.
(Analysis does not assume single-topic documents)

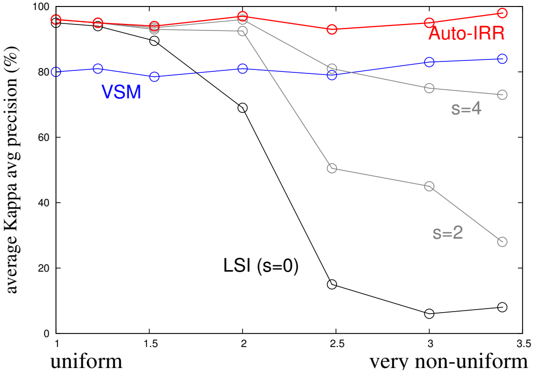
One set of experiments



Each point: average over 10 different single-topic TREC-document datasets of the given non-uniformity.

(Analysis does not assume single-topic documents)

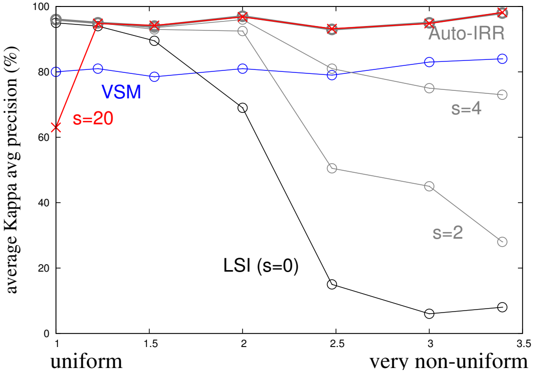
One set of experiments



Each point: average over 10 different single-topic TREC-document datasets of the given non-uniformity.

(Analysis does not assume single-topic documents)

One set of experiments



Each point: average over 10 different single-topic TREC-document datasets of the given non-uniformity.

(Analysis does not assume single-topic documents)

Act II: Nothing either good or bad, but thinking makes it so

We've just explored improving text categorization based on *topic*.

An interesting alternative: sentiment polarity — an author's overall opinion towards his/her subject matter ("thumbs up" or "thumbs down").[2]

Applications include:

- organizing opinion-oriented text for IR or question-answering systems
- providing summaries of reviews, customer feedback, and surveys

Much recent interest: for example, one 2002 paper has over 800 citations. See Pang and Lee (2008) monograph for an extensive survey.

[2] This represents one restricted sub-problem within the field of sentiment analysis.

More matter, with less art

State-of-the-art methods using bag-of-words-based feature vectors have proven less effective for sentiment classification than for topic-based classification [Pang, Lee & Vaithyanathan, 2002].

- 1. This laptop is a great deal.
  2. A great deal of media attention surrounded the release of the new laptop.
  3. If you think this laptop is a great deal, I've got a nice bridge you might be interested in.
- This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.
- Read the book. [Bob Bland]

More matter, with less art

State-of-the-art methods using bag-of-words-based feature vectors have proven less effective for sentiment classification than for topic-based classification [Pang, Lee & Vaithyanathan, 2002].

- 1. This laptop is a great deal.
  2. A great deal of media attention surrounded the release of the new laptop.
  3. If you think this laptop is a great deal, I've got a nice bridge you might be interested in.
- This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.
- Read the book. [Bob Bland]

More matter, with less art

State-of-the-art methods using bag-of-words-based feature vectors have proven less effective for sentiment classification than for topic-based classification [Pang, Lee & Vaithyanathan, 2002].

- 1. This laptop is a great deal.
  2. A great deal of media attention surrounded the release of the new laptop.
  3. If you think this laptop is a great deal, I've got a nice bridge you might be interested in.
- This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.
- Read the book. [Bob Bland]

Broader implications: politics

The on-line availability of politically-oriented documents, both official (e.g., parliamentary debates) and non-official (e.g., blogs), means:

The "[alteration of] the citizen-government relationship" [Shulman and Schlosberg 2002]
"The transformation of American politics" [*The New York Times*, 2006]
"The End of News?" [*The New York Review of Books*, 2005]

More opportunities for sentiment analysis!

Recall: people are searching for political news and perspectives.

Broader implications: politics

The on-line availability of politically-oriented documents, both official (e.g., parliamentary debates) and non-official (e.g., blogs), means:

The "[alteration of] the citizen-government relationship" [Shulman and Schlosberg 2002]
"The transformation of American politics" [*The New York Times*, 2006]
"The End of News?" [*The New York Review of Books*, 2005]

More opportunities for sentiment analysis!
Recall: people are searching for political news and perspectives.

Broader implications: politics

The on-line availability of politically-oriented documents, both official (e.g., parliamentary debates) and non-official (e.g., blogs), means:

The "[alteration of] the citizen-government relationship" [Shulman and Schlosberg 2002]
"The transformation of American politics" [*The New York Times*, 2006]
"The End of News?" [*The New York Review of Books*, 2005]

More opportunities for sentiment analysis!

Recall: people are searching for political news and perspectives.

One ought to recognize that the present political chaos is connected with the decay of language, and that one can probably bring about some improvement by starting at the verbal end.

Broader implications: politics

The on-line availability of politically-oriented documents, both official (e.g., parliamentary debates) and non-official (e.g., blogs), means:

The "[alteration of] the citizen-government relationship" [Shulman and Schlosberg 2002]
"The transformation of American politics" [*The New York Times*, 2006]
"The End of News?" [*The New York Review of Books*, 2005]

More opportunities for sentiment analysis!

Recall: people are searching for political news and perspectives.

> *One ought to recognize that the present political chaos is connected with the decay of language, and that one can probably bring about some improvement by starting at the verbal end.*

— George Orwell, "Politics and the English language", 1946

NLP for opinionated politically-oriented language

Sentiment analysis applied to this domain can enable:

- the summarization of un-solicited commentary and evaluative statements, such as editorials, speeches, and blog posts
  - (these may contain complex language, but not as complex as in the legislative proposals themselves ...)

- Governmental eRulemaking initiatives (e.g., www.regulations.gov) directly solicit citizen comments on potential new rules
  - 400,000 received for a single rule on labeling organic food

NLP for opinionated politically-oriented language

Sentiment analysis applied to this domain can enable:

- the summarization of un-solicited commentary and evaluative statements, such as editorials, speeches, and blog posts
  - (these may contain complex language, but not as complex as in the legislative proposals themselves ...)

- Governmental eRulemaking initiatives (e.g., www.regulations.gov) directly solicit citizen comments on potential new rules
  - 400,000 received for a single rule on labeling organic food

Our task

Given: transcripts of Congressional floor debates
Goal: classify each *speech segment* (uninterrupted sequence of utterances by a single speaker) as supporting or opposing the proposed legislation

Important characteristics:

1. Ground-truth labels can be determined automatically (speaker votes)
2. Very wide range of topics: flag burning, the U.N., "Recognizing the 30th anniversary of the victory of United States winemakers at the 1976 Paris Wine Tasting"
3. Presentation of evidence rather than opinion

   "Our flag is sacred!": is it pro-ban or contra-ban-revocation?
4. Discussion context: some speech segments are responses to others

Using discussion structure

Two sources of information (details suppressed):

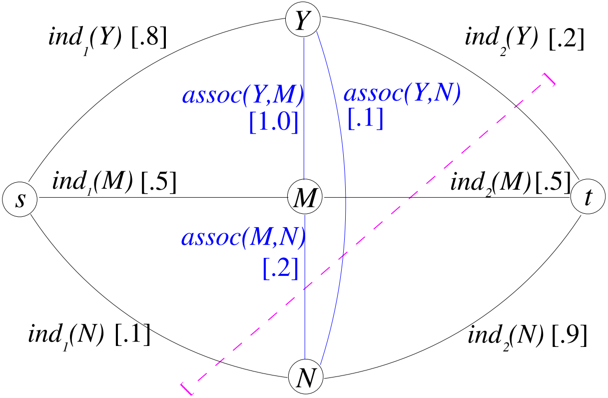
- An individual-document classifier that scores each speech segment x in isolation
- An *agreement classifier* for *pairs* of speech segments, trained to score by-name references (e.g., "I believe Mr. Smith's argument is persuasive") as to how much they indicate agreement

Optimization problem: find a classification c that minimizes:

$$\sum_{x} ind(x, \overline{c}(x)) + \sum_{x,x' \colon c(x) \neq c(x')} agree(x, x')$$

(the items' desire to switch classes due to individual or associational preferences)

Graph formulation and minimum cuts



Each labeling corresponds to a partition, or cut, whose cost is the sum of weights of edges with endpoints in different partitions (for symmetric assoc.).

Solving

Using network-flow techniques, computing the minimum cut...

- takes polynomial time, worst case, and little time in practice [Ahuja, Magnanti & Orlin, 1993]
- special case: finding the maximum a posteriori labeling in a Markov random field [Besag 1986; Greig, Porteous & Seheult, 1989]

Incorporating relationships leads to large improvements over SVMs run on individual documents alone.

Previous applications of the min-cut paradigm: vision; computational biology; Web mining; learning with unlabeled data
Examples of other methods incorporating relationship information:

Graph partitioning, e.g., normalized cut, correlation clustering, spectral graph transduction; Probabilistic relational models and related "collective classification" formalisms

Act III: Broader implications: sociology/social psychology

What opinions are influential?

⟶ proxy question: which Amazon reviews are rated helpful?

[Danescu-Niculescu-Mizil, Kossinets, Kleinberg, and Lee '09]

Prior work has focused on features of the *text* of the reviews, and has not been in the context of sociological inquiry. [Kim et al. '06, Zhang and Varadarajan '06, Ghose and Ipeirotis '07, Jindal and B. Liu '07, J. Liu et al '07].

Our focus: how about *non-textual* features (social aspects, biases)?

Our corpus: millions of Amazon book reviews.

Act III: Broader implications: sociology/social psychology

What opinions are influential?

$\longrightarrow$ proxy question: which Amazon reviews are rated helpful?

[Danescu-Niculescu-Mizil, Kossinets, Kleinberg, and Lee '09]

Prior work has focused on features of the *text* of the reviews, and has not been in the context of sociological inquiry. [Kim et al. '06, Zhang and Varadarajan '06, Ghose and Ipeirotis '07, Jindal and B. Liu '07, J. Liu et al '07].

Our focus: how about *non-textual* features (social aspects, biases)?

Our corpus: millions of Amazon book reviews.

Some social factors boosting helpfulness scores

- using "real name"

Our focus: What about the review's star rating in relationship to others?

Theories from social psychology:

- conform (to the average rating) [Bond and Smith '96]
- "brilliant but cruel" [Amabile '83]

Some social factors boosting helpfulness scores

- using "real name"
- being from New Jersey (for science books)

Our focus: What about the review's star rating in relationship to others?

Theories from social psychology:

- conform (to the average rating) [Bond and Smith '96]
- "brilliant but cruel" [Amabile '83]

Some social factors boosting helpfulness scores

- using "real name"
- being from New Jersey (for science books)
- not being from Guam

Our focus: What about the review's star rating in relationship to others?

Theories from social psychology:

- conform (to the average rating) [Bond and Smith '96]
- "brilliant but cruel" [Amabile '83]

Some social factors boosting helpfulness scores

- using "real name"
- being from New Jersey (for science books)
- not being from Guam

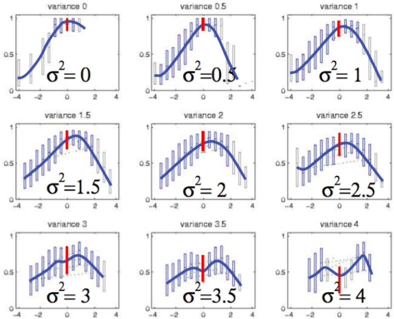
Our focus: What about the review's star rating in relationship to others?

Theories from social psychology:

- conform (to the average rating) [Bond and Smith '96]
- "brilliant but cruel" [Amabile '83]

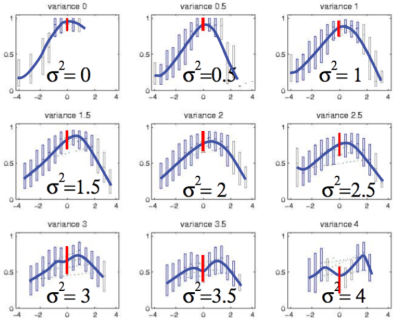
New observation: effect of variance

As *variance* among reviews increases, be *slightly above* the mean

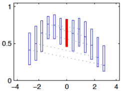


New observation: effect of variance

As *variance* among reviews increases, be *slightly above* the mean



... except in Japan, where it's best to be *slightly below*.

Example: $\sigma^2 = 3$: 

Are the social effects just textual correlates?

We would like to control for the actual quality of a review's text. (Maybe people from NJ inherently write better reviews about science books?)

How should we determine the "real" helpfulness, in order to control for it?

- manual annotation? Tedious, subjective.
- automatic classification? Need extremely high accuracy guarantees.

Are the social effects just textual correlates?

We would like to control for the actual quality of a review's text. (Maybe people from NJ inherently write better reviews about science books?)

How should we determine the "real" helpfulness, in order to control for it?

- manual annotation? Tedious, subjective.
- automatic classification? Need extremely high accuracy guarantees.

It turns out that 1% of Amazon reviews are *plagiarized*! (see also David and Pinch ['06]).

Our social-effects findings regarding position relative to the mean hold on plagiarized pairs, which *by definition* have the same textual quality.

The undiscovered country

We discussed:

- Better choice of feature vectors for document representation via IRR
  - Bounds analogous to those for LSI on IRR?
  - Alternative ways to compensate for non-uniformity?

- Sentiment classification incorporating pairwise agreement constraints using a minimum-cut paradigm
  - Other constraints, either knowledge-lean or knowledge-based?
  - Transductive learning for selecting association-constraint parameters?

- Non-textual factors affecting judgment of review quality
  - Other such factors?
  - Construction of review-aggregation systems that compensate for such biases?