Taming the Beast:
Sparse Machine Learning for Large Text Corpora

Laurent El Ghaoui

Berkeley Center for New Media & EECS Dept., UC Berkeley

with help from
Guan-Cheng Li, Vu Pham, Viet-An Duong, Xinyu Dai

New Directions in Management Science and Engineering Lecture
MS& E Department
Stanford University, May 15, 2012

Outline

Outline

Information Overload

Avalanche of "information" in text format, *e.g.*

- News articles, press releases, RSS feeds, TV captioning data.
- 10-K filings, marketing brochures, financial analyst reports, and other company-related documents.
- Consumer reviews, blogs, emails, and other social media content.
- Scientific papers, patents, law documents, bills, literature.

Information Overload

Avalanche of "information" in text format, *e.g.*

- News articles, press releases, RSS feeds, TV captioning data.
- 10-K filings, marketing brochures, financial analyst reports, and other company-related documents.
- Consumer reviews, blogs, emails, and other social media content.
- Scientific papers, patents, law documents, bills, literature.

The top *20* most important news sources have generated $\sim$ *40,000* news articles yesterday.

What might be useful?

- *Summarize* large text databases.
- Detect and visualize *trends* in term usage.
- *Compare* how topics of interest are treated across different sources.
- Allow for quick *translation* of summaries if original data is in foreign-language.
- *Cluster* text documents.
- Provide interpretable *visualizations* .

What might be useful?

- *Summarize* large text databases.
- Detect and visualize *trends* in term usage.
- *Compare* how topics of interest are treated across different sources.
- Allow for quick *translation* of summaries if original data is in foreign-language.
- *Cluster* text documents.
- Provide interpretable *visualizations* .

Approach: *sparse machine learning* tools to help in these tasks.

Example

Discovery of emerging issues in flight security

After each commercial flight in the US, pilots generate “ASRS reports” to document flight-related issues.

Key problem: detect emerging issues that are not being classified into existing categories, *e.g.*:

- “Wake vortex” problem of the Boeing 757.
- Increased number of runway incursions at LAX.

Example

Discovery of emerging issues in flight security

After each commercial flight in the US, pilots generate “ASRS reports” to document flight-related issues.

Key problem: detect emerging issues that are not being classified into existing categories, *e.g.*:

- “Wake vortex” problem of the Boeing 757.
- Increased number of runway incursions at LAX.

Don’t search for a needle — picture the haystack!

StatNews project

Statistical Analysis of News

Project started in 2007, with collaborators:

- *In statistics, optimization:* Bin Yu (Stat, UCB), Alexandre d'Aspremont (Ecole Polytechnique), Francis Bach (INRIA).
- *In social sciences:* Lee Fleming (IEOR), Sophie Clavier (International Relations, SFSU).

Sponsors: NSF, Google, CITRIS and INRIA.

StatNews web site
Data

- *Archives:*
  - New York Times, 1987-2007 (2.5 Million articles).
  - NYT headlines from 1851 to present.
  - headlines from 5 other sources since 1996.
- *English-speaking current news* (from April 2011-present):

  BBC, Ha'aretz, Moscow Times, Reuters, USA Today, Associated Press, The Australian, China Daily, CNN, Financial Times, The Guardian, India Times, Jerusalem Post, New York Times, Russian Times, Washington Post.

- *Chinese-speaking current news* (People's Daily).

StatNews project

Goals

- *Occurence analysis:* Picture the relative weight (frequency) given to different topics over time.
- Visualize the *image* (statistical associations) of a word or term as painted in the news, and visualize the *evolution* of the image, over time.
- Visualize news sources *relative* to each other, the *propagation* of concepts across news sources, and its dynamics.
- Provide a *web-based service* to analyze our text data, and allowing users to upload their own (medium-size) databases.

Outline

Information Overload

Topic imaging
- Predictive approach
- Visualizaations
- Beyond co-occurence
- Examples

Research Agenda
- Sparse PCA
- SAFE for LASSO
- Contextual applications

Topic imaging

Task: *topic imaging* (subject-specific summarization) in a given corpus.

- Sparse statistical prediction as surrogate.
- Human experiments to validate and find robust pre-processing schemes.

Topic imaging

Task: *topic imaging* (subject-specific summarization) in a given corpus.

- Sparse statistical prediction as surrogate.
- Human experiments to validate and find robust pre-processing schemes.

Result: a short list of terms that summarizes the topic as treated in the corpus.

What is topic imaging?

Topic image: A small set of terms that are semantically related to a given topic ("the query").

As a predictive problem: predict appearance of query term in a document given the term use in that document.

What is topic imaging?

Topic image: A small set of terms that are semantically related to a given topic ("the query").

As a predictive problem: predict appearance of query term in a document given the term use in that document.

- Predictive model must be *interpretable:* number of predictors (other terms) must be few (sparse modeling).
- Model must be obtained *fast* .

Visualizations

From the StaNews server:

- Compare different topics in a single source: http://statnews.org/pcaa8
- Compare same topic across different sources: http://atticus.berkeley.edu/guanchengli/showcase/chi/pd_hum_rig/ and http://atticus.berkeley.edu/guanchengli/showcase/chi/wapo_hum_rig/
- Compare sources: http://statnews2.eecs.berkeley.edu/snapdragon/showcase/spca_country_3month/

Visualizations

From the StaNews server:

- Compare different topics in a single source: http://statnews.org/pcaa8
- Compare same topic across different sources: http://atticus.berkeley.edu/guanchengli/showcase/chi/pd_hum_rig/ and http://atticus.berkeley.edu/guanchengli/showcase/chi/wapo_hum_rig/
- Compare sources: http://statnews2.eecs.berkeley.edu/snapdragon/showcase/spca_country_3month/

How did we get those word lists?

Co-occurence analysis

To capture the "image" of a term, we can use *co-occurence analysis:*

- We count the words that occur within the same unit of text (say, paragraph) as the term queried.
- We retain the top (say, 10) words co-occurring most frequently.
- The image is the corresponding list.

Implemented on our server: `http://statnews.org/`

Co-occurence analysis

To capture the “image” of a term, we can use *co-occurence analysis:*

- We count the words that occur within the same unit of text (say, paragraph) as the term queried.
- We retain the top (say, 10) words co-occurring most frequently.
- The image is the corresponding list.

Implemented on our server: `http://statnews.org/`

- *Pros:* fast, often revealing.

Co-occurence analysis

To capture the "image" of a term, we can use *co-occurence analysis:*

- We count the words that occur within the same unit of text (say, paragraph) as the term queried.
- We retain the top (say, 10) words co-occurring most frequently.
- The image is the corresponding list.

Implemented on our server: `http://statnews.org/`

- *Pros:* fast, often revealing.
- Cons: does not allow to compare two corpora.

Example

Two NYT op-ed columnists

Data: columns from *The New York Times* opinion Editors, Nicholas Kristof and Roger Cohen, between October 23, 2008 and March 31, 2009.

Questions:

- What are these authors talking about?
- What makes them different?

The ten most common words

Nicholas Kristof	Roger Cohen
mr	obama
people	iran
obama	said
said	american
president	president
world	iranian
new	israel
american	states
years	new
united	united

Both talk about the American elections . . .

The ten most common words

Nicholas Kristof	Roger Cohen
mr	obama
people	iran
obama	said
said	american
president	president
world	iranian
new	israel
american	states
years	new
united	united

So there's a lot of common words . . .

The ten most common words

Nicholas Kristof	Roger Cohen
mr	obama
people	iran
obama	said
said	american
president	president
world	iranian
new	israel
american	states
years	new
united	united

And some words are not very descriptive.

Sparse classification approach

To obtain the image of a term in a given corpus:

- *Separate* the corpus in two classes, one with all the documents (paragraphs) that contain the term, and the other without.

Sparse classification approach

To obtain the image of a term in a given corpus:

- *Separate* the corpus in two classes, one with all the documents (paragraphs) that contain the term, and the other without.
- Apply a *sparse classification algorithm* that uses words as features to predict the appearance of the given term in any given paragraph.

Sparse classification approach

To obtain the image of a term in a given corpus:

- *Separate* the corpus in two classes, one with all the documents (paragraphs) that contain the term, and the other without.
- Apply a *sparse classification algorithm* that uses words as features to predict the appearance of the given term in any given paragraph.
- The algorithm *assigns a weight to each term* that ever appears in the entire corpus.

Sparse classification approach

To obtain the image of a term in a given corpus:

- *Separate* the corpus in two classes, one with all the documents (paragraphs) that contain the term, and the other without.
- Apply a *sparse classification algorithm* that uses words as features to predict the appearance of the given term in any given paragraph.
- The algorithm *assigns a weight to each term* that ever appears in the entire corpus.
- *Most of the weights are zero*, which singles out a few important terms with high predictive power.

Example

Classification of the two NYT op-ed columnists

Nicholas Kristof	Roger Cohen
videos	olmert
darfur	persian
antibiotics	chemical
facebook	mohammad
sudanese	ali
janjaweed	dialogue
youtube	cease
sudan	iranian
sweatshops	tehran
invite	holocaust

The classification approach complements co-occurence analysis: it finds what is *unique* to each columnist.

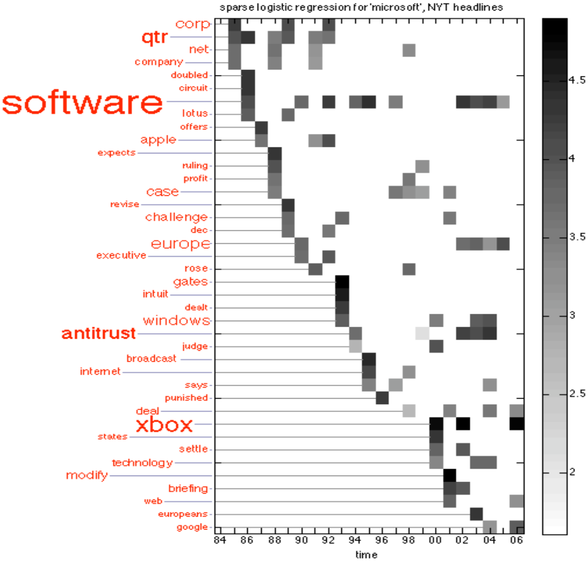
Evolution of image across time

- Proceed in sliding window fashion, with window size of say a year, and increments of one month.
- For each time window, use sparse classification to find a short list of words relevant to the query. (Thus we have a list of words for each year.)
- Visualize the matrix of classifier weights, ranking words by order of appearance, with font proportional to overall weights across time.

Provides a *summary* and a *timeline* .

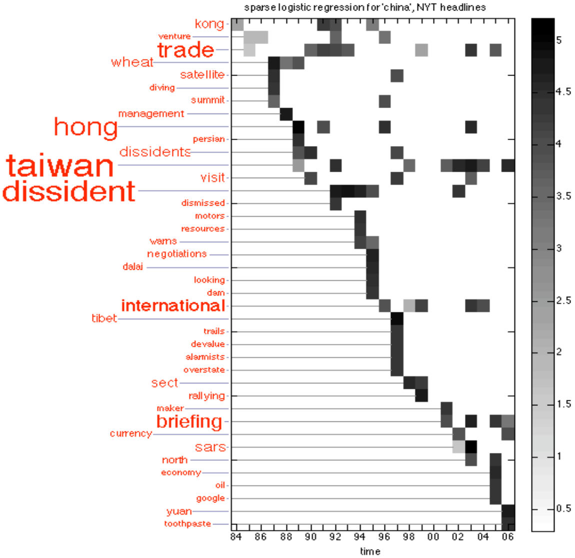
“Microsoft”

Data: The New York Times headlines, 1985-2007



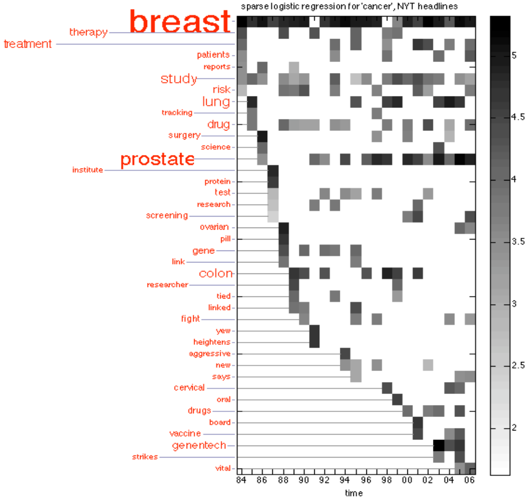
"China"

Data: The New York Times headlines, 1985-2007



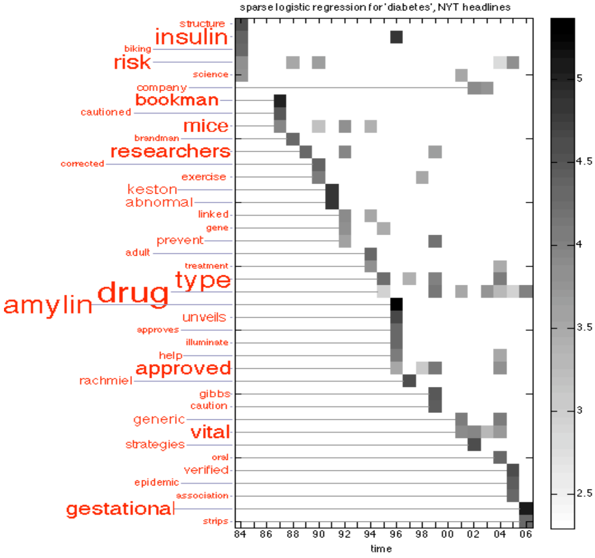
“Cancer”

Data: The New York Times headlines, 1985-2007



"Diabetes"

Data: The New York Times headlines, 1985-2007



Topic imaging in foreign languages

- Run topic imaging task on foreign press data in original language.
- Translate the *few* terms in the resulting list.

Avoids huge translation task!

Topic imaging in foreign languages

- Run topic imaging task on foreign press data in original language.
- Translate the *few* terms in the resulting list.

Avoids huge translation task!

Query: can you guess?

Source: People's Daily, Feb-Apr 2011.

利比亚 欧佩克 opec
利比亚 武力 force
利比亚 局势 situation
利比亚 行动 action
利比亚 平民 civilians
利比亚 撤出 withdrawal
利比亚 空袭 airstrike
利比亚 北非 french-speaking
利比亚 瓦莱塔 valletta
利比亚 撤离 evacuate
利比亚 军机 planes
利比亚 人道主义 humanitarianism
利比亚 卡扎菲 qadhafi

Outline

Research agenda

- *High-dimensional sparse machine learning:*
  - Safe feature elimination.
  - Data thresholding.
  - Kernel optimization for text classification.
  - Sparse PCA (allows interpretability of principal directions).
- *Visualization* and interactions with machine learning methods.
- *Contextual applications* (see next).

Sparse PCA

$$\max_{x\; x^T x=1} \; x^T C x - \lambda \mathbf{Card}(x).$$

- C covariance matrix.
- **Card** denotes cardinality (number of non-zero elements).
- $|lambda > 0$ penalty parameter.
- Allows to obtain *interpretable* results (in contrast to classical PCA).

Safe feature elimination: if a_i is the i-th feature vector

$$\max_{u\; u^T u=1} \sum_{i=1}^{m} ((a_i^T u)^2 - \lambda)_+$$

Allows to declare $x_i = 0$ whenever $\|a_i\|_2 \leq \lambda$.

Sparse PCA

1st PC (6 words)	2nd PC (5 words)	3rd PC (5 words)	4th PC (4 words)	5th PC (4 words)
million	point	official	president	school
percent	play	government	campaign	program
business	team	united_states	bush	children
company	season	u_s	administration	student
market	game	attack		
companies				

- *Data* : New York Times articles, 2009-2011, available at the UCI Machine Learning Repository. Corpus has 300K articles and has a dictionary of 100 K unique words.
- *Method* : Sparse PCA. This is an unsupervised method: Information about article section is not provided to the algorithm.
- SAFE allowed to reduce # features down to about 1000.

SAFE for LASSO

A (variant of) LASSO:

$$\min_x \|Ax - y\|_2 + \lambda \|x\|_1$$

with $A = [a_1, \ldots, a_n]$ the data matrix (each column is a feature).

Dual :

$$\max_u u^T y \ : \ \|A^T u\|_\infty \leq \lambda, \ \ \|u\|_2 \leq 1.$$

From optimality conditions, if $\|a_i\|_2 < \lambda$ then $x_i = 0$.

Perception Risk in Finance

(with Gah-Yi Vanh, LSE, and Sophia Chami, MS London).

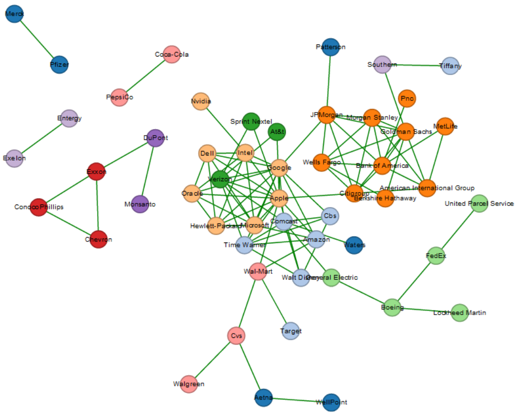
- Text data (news, financial reports) now actively used in finance.
- Most approaches focus on price movement estimation (*e.g.*sentiment analysis).
- Project focuses on using news data to better estimate *risk* (*e.g.*, covariance matrix).
- Initial results demonstrate news data contains useful information about risk.

Basic idea: estimate covariance matrix as a mix of price- and news-based ones:

$$C = tC^{\text{price}} + (1 - tC^{\text{news}}).$$

with $t \in [0, 1]$ estimated via cross-validation.

Sparse graphical model



Gaussian graphical model via l_1-penalized maximum-likelihood. Data: $\approx 300K$ Bloomberg full articles spanning 2010-2011.

News-based covariance recovers structure of the data (GICS sectors).

Active Collaborations

- "Emerging issues" in pilot-generated flight reports (with A. Srivastava, Machine Learning Group, NASA).
- Dynamics of innovation (with Lee Fleming, IEOR, UCB): study of diffusion of scientific innovation across scientific literature (PubMed), patents and news.
- Tracking of National Vulnerability Database (with Dawn Song, UCB).
- Image of countries and international institutions in foreign and US media (with S. Clavier, International Relations, SFSU). Focus: US-China relations.
- Monitoring of maintenance logs (with Piero Bonissone, GE Global Research).
- Perception risk in finance (with Terrance Odean, Haas, UCB).
- Discrete choice models with text data: analysis of an App Store database (with Denis Nekipelov, Econ, Minjung Park, Haas).
- Cervical cancer screening in social media (with Courtney Lyles & Urmimala Sarkar, UCSF's Center for Vulnerable Populations).

In the wings . . .

- Analysis of the tobacco litigation database (with Robert Proctor, History, Stanford).
- Analysis of historical Foreign news archives (with Mairi McLaughlin, French, UCB).
- Vote prediction based on text and campaign contributions (with Henry Brady, Pol Sci & Public Policy, UCB).