Sparse Statistical Analysis of Online News

Laurent El Ghaoui
(EECS/IEOR, UC Berkeley)

with help from
Onureena Banerjee & Brian Gawalt
(EECS, UCB)

BCNM Intro Talk

August 27, 2008
Multivariate statistics in context

Context: explosion in available data:

- heterogeneous (numerical, text, image, video)
- noisy (missing data, outliers, noise, etc)
- streaming (online learning)
- distributed (across networks)

Results in information overload
New challenges

Avalanche of data raises *new challenges*:

- very large-scale problems
- database issues (what to store and where, what to pre-compute)
- distributed computing (how to solve and where)
- online learning (how to update fast)
- robustness & regularization (handling noise, uncertainty)
- *interpretability* requirements, via e.g. sparsity constraints
Statistics for the “happy few”

Do you remember those times when statistics involved . . .

• proprietary data, given in batch, moderate size, centralized

• statistical expert-assisted problem solving (model setup, feature selection, nonlinear optimization, confidence analysis, etc)

• solvers using nonlinear (often, unconstrained) optimization, on single machines

• experts from field of application to interpret the results

• no sparsity constraints—all we cared about was statistical performance
A more modern point of view

Large-scale, efficient statistical analysis "at everyone’s fingertips":

• "automated" feature selection

• automated confidence analysis & regularization

• a more analytical view of search (from the list to the short list)

• emphasis on *interpretable* results, visualization

• emphasis on computational complexity constraints, and (sometimes) on real-time updates
A few challenging applications

• computational biology
• financial markets
• health and medicine
• public social data analysis: online news, voting records, etc
New project started in 2007, with collaborators:

- **In statistics, optimization**: Bin Yu (UCB, Stat), Alexandre d’Aspremont (ORFE, Princeton)

- **In social sciences**: Charles Cameron (Pol Sci, Princeton), Henry Brady (Pol Sci, UC Berkeley), Suad Joseph (Anthropology, UC Davis), Sophie Clavier (Pol Sci, SFSU)
Data

Current data sets:

- New York Times articles, 1981-2007 (2.5 Million articles)
- Reuters corpus, 1996-7
- Reuters “Significant Development” corpus, 2000-2007
- Voting data from VoteWorld
- More to come?
Goals

• Understand the *image* (statistical associations) of a word or term as painted in the news

• Form a *graph of words* as they relate to each other

• Observe the *evolution* of the image or graph across time

• Understand news sources *relative* to each other, the *propagation* of concepts across news sources, and its dynamics

The main challenge is to connect these “soft” goals with “hard” statistical concepts and methods
Image of a term in the news: possible approaches

Given a dictionary of words or terms and a corpus of news documents

- **Counting**: raw word frequencies, tf-idf scores, co-occurrence in same unit (sentence, paragraph, document, headline)

- **Sparse regression analysis**: non-zero regression coefficients correspond to relevant words

- **Sparse covariance analysis, sparse PCA**: allow to build a sparse representation of words/terms (unsupervised learning)
Image dynamics visualization

Sparse regressor matrix plot:

• Each row in the plot represents a word which, at some point in time, was statistically associated with the query word

• Each column to a month

• Columns show the classification weights assigned to the associated words by a classifier (computed over the past year, in rolling horizon fashion)

• Classification method: sparse logistic regression

Sparse logistic regression analysis for 'gay', NYT headlines, Jan81 — Jan07

- Plots shows evolving image over time (from "Gay Men Chorus" to "Right" to "Pride" to "Marriage")

- Identifies "sticky" words (those which, once around, stay around a long time, eg, "Marriage") vs. "transient" ones (eg, "ROTC", "Virus")

- Allows to highlight when shifts occur, and the overall dynamic nature of the query (fewer sticky words recently)

- The plot could be interactive and fun to manipulate!