

Statistical Analysis of Online News

Laurent El Ghaoui
EECS, UC Berkeley

Information Systems Colloquium, Stanford University

November 29, 2007

Collaborators

Joint work with

Alexandre d'Aspremont (ORFI, Princeton)
Bin Yu (Stat/EECS, Berkeley),
Babak Ayazifar (EECS, Berkeley)
Suad Joseph (Anthropology, UC Davis)
Sophie Clavier (International Relations, SFSU)

and UCB students:

Onureena Banerjee, Brian Gawalt, Vahab A. Pournaghshband

Online data

Online data:

- online news (text, video)
- voting records (Senate, UN, . . .)
- demographic data
- economic data

Now *widely* available . . . Or, easy to scrape!

What about statistics?

Progresses in statistical learning:

- efficient algorithms for large-scale optimization
- better understanding of sparsity (interpretability) issues
(LASSO and variants, compressed sensing, etc)

Current hot application topics in Stat, Applied Math: *biology, finance*

StatNews project

Our data:

- online text news
- voting and other political records (PAC contributions, etc)
- International bodies voting records, such as UN General Assembly votes

StatNews project

Goals:

- Provide open source tools for fetching, assembling data, and perform statistical analyses
- Show compressed (sparse) views of data
- Ultimately foster a forum where such views are discussed

Project is in its infancy

Example: Senate voting analysis

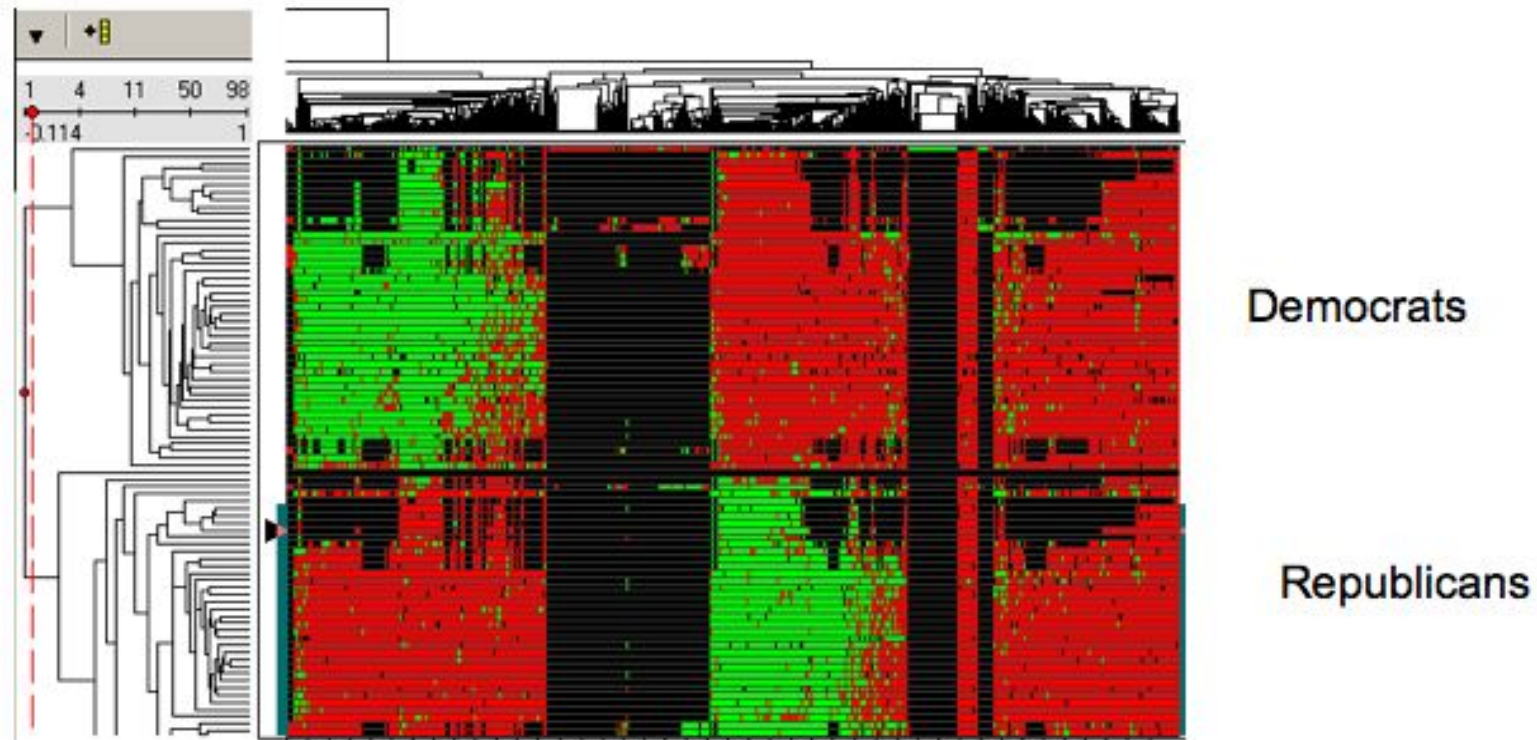
(Courtesy Georges Natsoulis, Stanford Genome Technology Center)

Data set: US Senate voting records from the 109th Senate (2004 - 2006)

- 100 variables, 542 samples, each sample is a bill that was put to a vote
- Records 1 for yes, -1 for no/abstention on each bill

The next slide shows the result of hierarchical clustering, using off-the-shelf commercial software

No=green Yes=Red



3 completely different voting patterns on 3 sets of bills

Hierarchical clustering: analysis

The data appears to have *structure*:

- As expected, Senators are divided by party line
- Perhaps more surprisingly, bills appear to fall into three distinct categories, of comparable size

Now let's learn more about the categories . . .

Rep=Y Dem=N

37314Equal Protection of Voting Rights Act of 2001-
 36923John Ashcroft for Attorney General
 34781Pay-As-You-Go Amendment
 35180Foreign Student Public Education Amendment
 35047Veterans Affairs, HUD FY '96 Appropriations bill
 35019Continuing Appropriation bill
 34760Balanced Budget Proposed Constitutional Amendment
 35535Yucca Mountain Alternate Sites amendment
 35970Education Savings Accounts bill
 38666Detainees at Guantanamo Bay Amendment
 38702USA PATRIOT and Terrorism Prevention Reauthoriza...
 38182Federal Marriage Amendment
 37706Budget Resolution FY2004
 38659Budget Reconciliation bill
 38792Debt Limit Increase Resolution
 38512Janice R. Brown, US Circuit Court
 38848Tax Reconciliation bill
 38386Alberto R. Gonzales, for Attorney General
 35235Overseas Abortion Amendment
 35143Medical Professionals Amendment
 37915Prohibit Partial-Birth Abortion bill
 35045Flag Desecration bill
 38750Tax Reconciliation bill
 35047Interior Department FY96 Appropriations bill
 38855English As National Language Amendment
 38930Gulf of Mexico Energy Security Act of 2006
 35145Product Liability bill
 38496Motion to invoke cloture on Priscilla R. Owen
 37644Consolidated Appropriations Resolution, 2003
 38561CAFTA Implementation Bill
 38378Condoleezza Rice, Secretary of State
 35586Emergency Supplemental Appropriations bill
 38778USA PATRIOT and Terrorism Prevention Reauthoriza...
 35608Budget Reconciliation Bill
 38863Michael Hayden Confirmation
 35151Line-Item Veto bill
 34961Welfare Reform Bill
 35020National Highway System Designation Act of 1995

Dem=Y Rep=N

35187Public Assistance to Legal Immigrants Amendment
 38608EPA's Clean Air Mercury Rule
 38428Tax Subsidy for Domestic Companies Amendment
 38673Additional Funding For Veterans Amendment
 38428Native American Funding Amendment
 38666Investigating Contracts in Iraq Amendment
 38750Tax Rate Extension Amendment
 38559Corporate Financing of Terrorism Amendment
 38923Teen Pregnancy Education Amendment
 38519Reduction in Dependence on Foreign Oil
 38792Education Funding Amendment
 38889Minimum Wage Adjustment Amendment
 38659Pay As You Go Amendment
 38428Prescription Drugs Amendment
 38652AIDS Drug Assistance Program Amendment
 38651Low-Income Home Energy Assistance Program Amen...
 38659ANWR Amendment
 38519Renewable Portfolio Standard (RPS) Amendment
 38118Unemployment Benefits Amendment
 38428Homeland Security Grant Program Amendment
 35600Intelligence Appropriations Declassification amendment
 36006Federal Election Commission term length
 36720Death/Estate Tax Amendment
 35255Minimum Wage Increase Amendment
 35318Employment Nondiscrimination Act of 1996
 35236Nomination of Alice M. Rivlin
 37428Military Abortion Amendment
 36697Hate Crimes Amendment
 38862Immigration Reform Bill
 38462Future Military Funding for Iraq Amendment
 35046Deployment of US Armed Forces in Bosnia-Herzegovina
 36300Juvenile Crime bill
 37425Terrorism Insurance Bill
 35279Minimum Wage Increase bill
 36556Violent Protestors Amendment
 36419Appropriations bill FY2000, Treasury, Postal Service
 36483District of Columbia FY2000 Appropriations bill

Both vote Y

36278Withdrawal of U.S. Troops from the Balkans resolution
 37371Securing America's Future Energy (SAFE) Act of 2001-
 38489Transportation Equity Act: A Legacy for Users
 34865Telecommunications bill
 36657Africa Free Trade bill
 36298Religious Memorials at Schools Amendment
 35698Defense Department FY98 Appropriations bill
 35325Appropriations bill FY97, Energy and Water Developm...
 35039Housing for Older Persons Act of 1995
 35187Immigration Reform bill
 35279Health Insurance Portability bill
 35740Defense Department FY98-99 Authorization bill
 36425FY 2000 Defense Authorization-Conference Report
 35937Iran Missile Sanctions bill
 35742Appropriations bill FY98, Labor, HHS, Education
 36041Appropriations bill FY99, Treasury, Postal Service
 36069Defense Department FY99 Authorization bill
 35467Secretary of Transportation Nomination
 35501Secretary of Energy Nomination
 35642Budget Reconciliation bill
 36342Lawrence H. Summers for Secretary of the Treasury
 36230Education Flexibility Partnership Act of 1999
 37243No Child Left Behind Act
 37469Department of Defense Appropriations, FY2003 bill
 37238National Defense Authorization Act for Fiscal Year 2002
 37357Equal Protection of Voting Rights Act of 2001-
 37155Air Transportation Safety and System Stabilization Act
 36915Norman Y. Mineta for Secretary of Transportation
 37210Agriculture FY2002 Appropriations bill
 38085Pension Funding Equity Act of 2004
 38106Internet Access Tax bill
 37931Consolidated Appropriations Act, 2004
 38266National Intelligence Reform Act of 2004
 37931Department of Labor, HHS, and Education Appropriati...
 37889Terrorism Information Awareness bill
 37916Reduction of SPAM bill
 37937Fiscal 2004 Defense Authorization - Conference Report

Challenging the results

As a statistician, we can easily *challenge these results*:

- The number of samples may not be sufficient, but we don't see it on the plot!
- There might be better (more robust) methods for clustering
- What could be the underlying model, and what are the simplifying assumptions? (stationarity, complexity, etc)
- The word frequency count method can be improved

Many approaches can be thrown at the problem—whatever the method, it will always only provide a particular, *biased* view of data

Online news

Current data sets:

- New York Times, since August 2007
- Reuters corpus, 1996-7
- Reuters “Significant Development” corpus, 2000-2007

Image of Presidential candidates

Adverbs in Obama vs. McCain:

- Gather 200 NYT articles mentioning the candidates' names in the past 6 months
- perform sparse logistics regression, with features the 2300 words ending in 'ly', and label +1 if "Obama" appears more than "McCain", -1 otherwise
- then look at the non-zero coefficients of the classifier, > 0 ones correspond to Obama, < 0 ones to McCain

OBAMA

<u>Word</u>	<u>Coefficient</u>
nearly	0.00281
commonly	0.00100
utterly	0.00086
lovely	0.00073
highly	0.00061
family	0.00058
previously	0.00047
recently	0.00042
especially	0.00011

McCAIN

<u>Word</u>	<u>Coefficient</u>
really	-0.00149
aggressively	-0.00140
actually	-0.00120
early	-0.00110
beautifully	-0.00106
rarely	-0.00102
emily	-0.00096
arrestingly	-0.00091
relatively	-0.00077
imply	-0.00066
closely	-0.00050
certainly	-0.00050
only	-0.00035
hopelessly	-0.00006

Statistical learning: the pandora box is open

Following Bin Yu (2007): statistical learning is now deeply linked to

- distributed (web) databases, networks
- large-scale optimization
- compressed sensing and sparsity
- visualization methods

We need to design statistical learning algorithms with these interactions in mind

Challenges

- Sparse multivariate statistics (sparse PCA, sparse covariance selection, etc)
- Discrete random Markov field modelling (e.g. for voting data)
- Large-scale computations: distributed, online (recursive updates)
- Heterogeneous data and kernel optimization methods (handling text and images)
- Visualization of statistical results
(e.g., how to visualize a graph and the level of confidence we can associate to it)

Sparsity

Consider the problem of representing features on a graph: to be interpretable, the graph must not be too dense

Here, “interpretable” often involves *sparsity*

- Find a few keywords that best explain the Senators votes
- Find a sparse representation of the joint distribution of votes
- Find the few keywords that are important in predicting the appearance of a reference word

Sparse Covariance Selection

- Draw n independent samples $y_i \sim \mathcal{N}_p(0, \Sigma)$, where Σ is unknown.
- *Prior belief*: many conditional independencies among the variables in this distribution.
- Zeros in inverse covariance correspond to conditional independence properties among variables.
- *Covariance estimation*:: From y_1, \dots, y_n , recover the covariance matrix Σ .
- *Covariance selection*: choosing which elements of our estimate $\hat{\Sigma}^{-1}$ to set to zero.

Penalized Maximum-Likelihood Approach

Penalized ML problem:

$$\max_{X \succ 0} \log \det X - \mathbf{Tr}(SX) - \rho \|X\|_1$$

- $\rho > 0$ is regularization parameter, and $\|X\|_1 := \sum_{i,j} |X_{ij}|$.
- Convex, non-smooth problem, can be solved in $O(n^{4.5})$ with first-order methods.
- Same idea used in l_1 -norm penalized regression (LASSO), for example.

Sparse Principal Component Analysis

Principal Component Analysis (PCA) is a classic tool in multivariate data analysis.

- *Input*: a $n \times m$ data matrix $A = [a_1, \dots, a_m]$, containing m observations (samples) $a_i \in \mathbf{R}^n$.
- *Output*: a sequence of *factors* ranked by *variance*, where each factor is a *linear* combination of the problem variables

Typical use: *reduce the number of dimensions* of a model while *maximizing the information* (variance) contained in the simplified model.

Variational formulation of PCA

We can rewrite the PCA problem as a sequence of problems of the form

$$\max_x x^T \Sigma x : \|x\|_2 = 1,$$

where $\Sigma = AA^T$ is (akin to) the covariance matrix of the data. This finds a direction of *maximal variance*.

The problem is *easy*, its solution is $\lambda^{\max}(\Sigma)$, with $x^* =$ any associated eigenvector.

Sparse PCA

We seek to increase the sparsity of "principal" directions, while maintaining a good level of explained variance.

Sparse PCA problem:

$$\phi := \max_x x^T \Sigma x - \rho \mathbf{Card}(x) : \|x\|_2 = 1.$$

where $\rho > 0$ is given, and $\mathbf{Card}(x)$ denotes the cardinality (number of non-zero elements) of x .

This is non-convex and *NP-hard*.

Lower bound

The Cauchy-Schwartz inequality:

$$\forall x : \|x\|_1 \leq \sqrt{\mathbf{Card}(x)} \cdot \|x\|_2$$

yields the lower bound:

$$\phi \geq \phi_1 := \max_x x^T \Sigma x - \rho \|x\|_1^2 : \|x\|_2 = 1.$$

Above problem is *still not convex*

Relaxation of l_1 -norm bound

Using the lifting $X = xx^T$ we obtain the SDP approximation

$$\phi_1 \leq \psi_1 := \max_X \langle \Sigma, X \rangle - \rho \|X\|_1 : X \succeq 0, \mathbf{Tr} X = 1,$$

where $\|X\|_1$ is the sum of the absolute value of the components of matrix X .

Above approximation can be interpreted as a *robust PCA* problem:

$$\psi_1 = \max_{X : X \succeq 0, \mathbf{Tr} X = 1} \min_{\|U\|_\infty \leq \rho} \langle (\Sigma + U), X \rangle = \min_{\|U\|_\infty \leq \rho} \lambda_{\max}(\Sigma + U).$$

Kernel optimization for supervised problems

Many problems in text corpora analysis involve regression or classification with *heterogeneous* data

- Sentiment detection (“is this piece of news good or bad?”)
- Classification approaches to clustering
- In some cases, we need to predict a value based on text (and possibly other information, such as prices)

We can represent text, images, and in general, heterogeneous data with numbers (e.g. bag-of-words), but there are many such representations—which is the best?

Linear regression

Linear regression model for prediction:

$$y(t) = \theta^T x(t) + e(t)$$

where $X = [x(1), \dots, x(T)]$ is the feature matrix, y is the vector of observations, and e contains the noise.

Regularized least-squares solution:

$$\min_w \|X^T \theta - y\|_2^2 + \rho^2 \|w\|_2^2,$$

where ρ is given.

Solution

The dual to the LS problem writes

$$\max_{\alpha} \alpha^T y - \alpha^T K_{\rho} \alpha,$$

where $K_{\rho} := X^T X + \rho^{-2} I$.

- The optimal dual variable is $\alpha = K_{\rho}^{-1} y$
- The optimal value of the LS problem is $y^T K_{\rho}^{-1} y$
- The prediction at a test point x is $w^T x = \rho^{-2} x^T X \alpha^*$

The kernel matrix

The solution (optimal value, and prediction) depends only on the “kernel matrix” \mathbf{K} containing the scalar products between training points, and those between training points and the test point.

$$\mathbf{K} := \begin{pmatrix} K & X^T x \\ x^T X & x^T x \end{pmatrix}, \text{ with } K := X^T X.$$

This matrix is positive semidefinite, and the optimal value of the LS problem, $y^T K^{-1} y$, is convex in that matrix.

Kernel optimization

Let \mathcal{K} be a subset of the set of positive, semidefinite matrices of order $T + 1$ ($T =$ number of samples).

Kernel optimization problem:

$$\min_{\mathbf{K} \in \mathcal{K}} y^T \mathbf{K}^{-1} y$$

The above problem is convex.

Rank-one kernel optimization

Choose

$$\mathcal{K} = \left\{ K(\mu, \lambda) = \rho^2 \sum_{i=1}^n \mu_i e_i e_i^T + \sum_{i=1}^m \lambda_i k_i k_i^T, \sum_{i=1}^n \mu_i = \sum_{i=1}^m \lambda_i = 1, \mu \geq 0, \lambda \geq 0 \right\},$$

where e_i 's are the unit vectors in \mathbf{R}^n , and k_i 's are given vectors.

The kernel optimization problem writes

$$\phi^2 = \min_{\lambda, \mu} y^T K(\mu, \lambda)^{-1} y : \lambda \geq 0, \mu \geq 0, \sum_{i=1}^n \mu_i + \sum_{i=1}^m \lambda_i = 1,$$

LP solution

The problem reduces to the LP

$$\min_u \left\| y - \sum_{i=1}^m u_i k_i \right\|_1 + \rho \|u\|_1.$$

The corresponding optimal kernel weights are given by

$$\mu_i = \frac{|v_i|}{\rho\phi}, \quad i = 1, \dots, n, \quad \lambda_i = \frac{|u_i|}{\phi}, \quad i = 1, \dots, m,$$

where $v = y - \sum_{i=1}^m u_i k_i$.

Kernel optimization in practice

In the context of text corpora analysis, the approach can be applied as follows:

- We select a collection of Kernels, each of which provides a representation of data (e.g. a bag-of-words kernel, another based on some other feature, such as prices)
- We compute the eigenvectors of all the kernel matrices, which gives us a collection of rank-one kernels $k_i k_i^T$, $i = 1, \dots, m$.
- We include the dyads $e_i e_i^T$ for regularization purposes.

Ising models of binary distributions

Second-order Ising model: distribution on a binary random variable x parametrized by

$$p(x; Q, q) = \exp(x^T Q x + q^T x - Z(Q, q)), \quad x \in \{0, 1\}^n,$$

where (Q, q) are the model parameters, and $Z(Q, q)$ is the normalization term.

WLOG $q = 0$, and define the *log-partition function*

$$Z(Q) := \log \left(\sum_{x \in \{0, 1\}^n} \exp[x^T Q x] \right).$$

Maximum-likelihood problem

Given and empirical covariance matrix S , solve

$$\min_{Q \in \mathcal{Q}} Z(Q) - \mathbf{Tr} QS$$

where , and \mathcal{Q} is a subset of the set \mathcal{S}^n of $n \times n$ symmetric matrices

When $\mathcal{Q} = \mathcal{S}^n$, the above corresponds to the *maximum entropy* problem

$$\max_p H(p) : p \geq 0, p^T \mathbf{1} = 1, S = \sum_{x \in \{0,1\}^n} p(x) x x^T,$$

where H is the discrete entropy function, $H(p) = - \sum_{x \in \{0,1\}^n} p(x) \log p(x)$.

Bounds on the log-partition function

- Due to its *exponential number of terms*, computing or optimizing the log-partition function is NP-hard
- We are interested in finding tractable, convex upper bounds on $Z(Q)$
- such bounds yield suboptimal points for the ML problem

Cardinality bound

Let Δ_k be the set of vectors in $\{0, 1\}^n$ with cardinality k . Since $(\Delta_k)_{k=0}^n$ forms a partition of $\{0, 1\}^n$, we have

$$Z(Q) = \log \left(\sum_{k=0}^n \sum_{x \in \Delta_k} \exp[x^T Q x] \right).$$

Thus,

$$Z(Q) \leq \log \left(\sum_{k=0}^n |\Delta_k| \exp[\psi_k(Q)] \right)$$

where $\psi_k(Q)$ is any upper bound on the maximum of $x^T Q x$ over Δ_k

Cardinality bound

Use

$$\psi_k(Q) = \max_{X \succeq d(X)d(X)^T} \mathbf{Tr} QX \quad : \quad d(X)^T d(X) = k, \quad \mathbf{1}^T X \mathbf{1} = k^2,$$

with $d(X)$ the diagonal matrix formed by zeroing out all off-diagonal elements in X .

- This results in a new bound, the *cardinality bound*, which can be computed in $O(n^4)$
- The corresponding maximum-likelihood problem is also *tractable* ($O(n^4)$)

Approximation error

Standard Ising models: $Q = \mu I + \lambda \mathbf{1}\mathbf{1}^T$, with λ, μ scalars

(such models describe node-to-node interactions via a mean-field approximation)

- The cardinality bound is *exact* on standard Ising models
- The approximation error is controlled by the l_1 -distance to the class of standard Ising models:

$$0 \leq Z_{\text{card}}(Q) - Z(Q) \leq 2D_{\text{st}}(Q), \quad D_{\text{st}}(Q) := \min_{\lambda, \mu} \|Q - \mu I - \lambda \mathbf{1}\mathbf{1}^T\|_1.$$

Comparison with the log-determinant bound

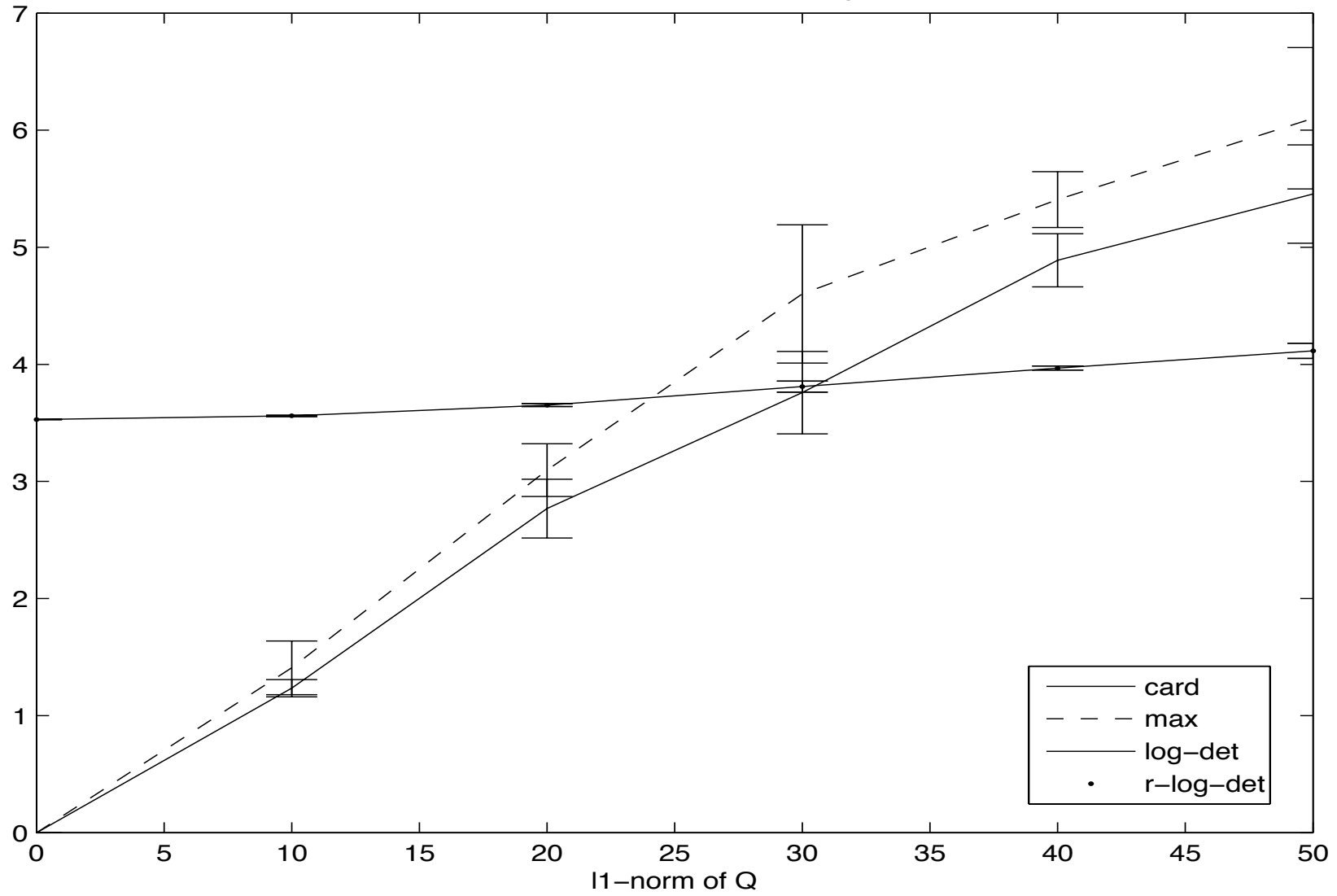
Wainwright and Jordan's log-determinant bound (2004):

$$Z_{\text{rld}}(Q) := (n/2) \log(2\pi e) + \max_{d(X)=x} \mathbf{Tr} QX + \frac{1}{2} \log \det(X - xx^T + \frac{1}{12}I)$$

Fact: The cardinality bound is better than the log-determinant one, provided $\|Q\|_1 \leq 0.08n$

(In practice, the condition is *very* conservative)

approximation error for upper bounds on the exact log-partition function (n=20)



Consequences for the maximum-likelihood problem

Including the convex bound $D_{\text{st}}(Q) \leq \epsilon$ in the maximum-likelihood problem makes a lot of sense:

- It ensures that the approximation error is less than 2ϵ
- It will tend to produce an optimal Q^* that has *few* off-diagonal elements differing from their median

Hence, the model is “interpretable” — we can display a graph showing only those non-median terms, the user needs to know that there is an overall “mean-field” effect

Concluding remarks

- Online news analysis, and more generally, the *analysis of social data* found on the web, constitute a new frontier for statistics and optimization, as were biology and finance in the last decade
- This raises new *fundamental challenges for machine learning*, especially in the areas of sparsity, online learning and binary data models
- Calls for a renewed interaction between engineering and social sciences