Sparse and Robust Optimization and Applications

Optimization and Statistical Learning Workshop
Les Houches, 2013

Laurent El Ghaoui
with Mert Pilanci, Anh Pham

EECS Dept., UC Berkeley

January 7, 2013

Outline

Outline

Sparse Optimization

Sparse Probability Optimization
- Sparse probabilities
- Norm-ratios approach
- Some recovery results
- Applications
- Extensions

Robust Optimization for Dimensionality Reduction
- Robust low-rank LP
- Low-rank LASSO

Generic sparse optimization problem

Optimization problem with cardinality penalty:

$$\min_w \; L(X^T w) + \lambda \|w\|_0.$$

- Data: $X \in \mathbf{R}^{n \times m}$.
- Loss function L is convex.
- Cardinality function $\|w\|_0 := |\{j \;:\; w_j \neq 0\}|$ is non-convex.
- λ is a penalty parameter allowing to control sparsity.

- Arises in many applications, including (but not limited to) machine learning.
- Computationally intractable.

Classical approach

A now classical approach is to replace the cardinality function with an l_1-norm:

$$\min_w \; L(X^T w) + \lambda \|w\|_1 .$$

Pros:

- Problem becomes convex, tractable.
- Many ”recovery” results available.

Cons:

- Is neither a lower nor an upper bound in general.
- Fails completely in some cases (see next).

The l_1-norm may fail

The l_1-norm approach may fail to allow to *control* the level of sparsity of the solution.

- When the variable is restricted to be a *discrete distribution*, the l_1-norm is constant, and the level of sparsity cannot be controlled.
- If the data matrix X is *low-rank*, the cardinality of the solution may be also hard to control.

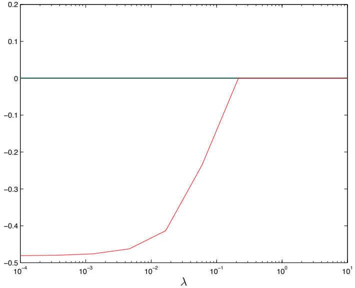
Example:

LASSO with rank-one data matrix

$$\min_w \|(qp^T)w - y\|_2 + \lambda\|w\|_1,$$

with $p \in \mathbf{R}^n$, $q, y \in \mathbf{R}^m$. Solution for $\lambda > 0$ has cardinality one or zero.

Coordinates of optimal w vs. λ



Outline

Sparse Probability Optimization

Generic sparse probability optimization problem:

$$p^* := \min_w \; L(X^T w) + \lambda \|w\|_0 \;:\; w \geq 0, \;\; \mathbf{1}^T w = 1.$$

- l_1-norm approach fails to control sparsity.
- Applications: index fund construction, examplar-based clustering.

Proposed Approach

Basic bound:

$$\|w\|_1 \leq \|w\|_0 \|w\|_\infty.$$

Yields a *lower bound* on the original problem:

$$\begin{aligned} p^* &= \min_w \; L(X^T w) + \lambda \|w\|_0 \;:\; w \geq 0, \;\; \mathbf{1}^T w = 1 \\ &\geq \hat{p} := \min_w \; L(X^T w) + \lambda \frac{\|w\|_1}{\|w\|_\infty} \;:\; w \geq 0, \;\; \mathbf{1}^T w = 1. \end{aligned}$$

Fact: The lower bound can be computed as a sequence of n uncoupled *convex* problems:

$$\hat{p} = \min_{1 \leq i \leq n} \; \min_w \; L(X^T w) + \lambda \frac{1}{w_i} \;:\; w \geq 0, \;\; \mathbf{1}^T w = 1.$$

Checking approximation quality

Let $\hat{w}$ be a solution to

$$\hat{p} = \min_{1 \le i \le n} \min_{w \,:\, w \ge 0, \mathbf{1}^T w = 1} L(X^T w) + \frac{\lambda}{w_i}.$$

We then have the bounds:

$$L(X^T \hat{w}) + \lambda \|\hat{w}\|_0 \ge p^* \ge \hat{p}. \quad (1)$$

- The quality of approximation can be easily checked when the n convex programs are solved.
- In contrast, ℓ_1 regularization does not have such a property since it is not a lower bound or upper bound. Known guarantees are not checkable in polynomial time, *e.g.* , *Restricted Isometry Property* (Candés & Tao, 2005).

Theoretical results for a special case

Problem: recover the sparsest probability measure given moment constraints:

$$p^* = \min_w \|w\|_0 \; : \; X^T w = y, \;\; w \geq 0, \;\; \mathbf{1}^T w = 1,$$

where $y \in \mathbf{R}^m$ is given. This is a special case of our generic problem, with L the indicator of the affine set $\{w \; : \; X^T w = y\}$.

Our bound is $p^* \geq \hat{p} = 1/(\max_{1 \leq i \leq n} q_i)$, where each q_i is the optimal value of a linear program:

$$q_i := \max_w \; w_i \; : \; X^T w = y, \;\; w \geq 0, \;\; \mathbf{1}^T w = 1.$$

Recovery result: geometric property

Assume that the unique solution of p^* is given by w^*; let S be the support of w^*, and let S_c be its complement. If $\mathbf{Conv}(X_{S_c})$ does not intersect an extreme point of $\mathbf{Conv}(X_S)$ then $\hat{w} = w^*$, *i.e.* the approximation is exact.

Consequence: For $X \sim$ iid Gaussian, this happens with very high probability if n is $\mathcal{O}(||w^*||_0)$.

Application: Index Tracking

Given a time-series matrix of prices of n assets over m days $X = [x_1, \ldots, x_m]$, reconstruct a given financial index time-series $y_1, \ldots, y_m$ as a convex combination of n assets using as few assets as possible:

$$p^* = \min_{w \geq 0,\, \mathbf{1}^T w = 1} \left\| X^T w - y \right\|_2^2 + \lambda ||w||_0.$$

The l_1-norm approach fails in this case.

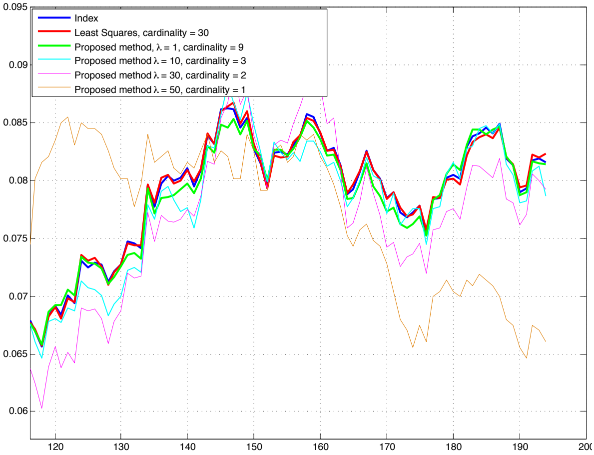
Proposed approach:

$$p^* \geq \hat{p} = \min_{1 \leq j \leq n} \min_{w \geq 0, \mathbf{1}^T w = 1} \left\| X^T w - y \right\|_2^2 + \frac{\lambda}{w_j}.$$

Solved by n second-order cone programs.

Numerical results

30 assets in 197 trading days, 2007-2008.



Application: clustering

Maximum-likelihood mixture fitting

Given data $\{z_1, \ldots, z_n\}$ of d-dimensional vectors, consider fitting a parametric iid distribution p_θ via maximizing the log-likelihood over the parameter θ

$$\max_\theta \frac{1}{n} \sum_{i=1}^n \log p_\theta(z_i).$$

Consider a mixture of Gaussians distribution with k centers:

$$p_{w,\mu_1,\ldots,\mu_k}(z) \sim \sum_{j=1}^k w_j e^{-\beta \|z-\mu_j\|_2^2}.$$

with unknown mixture weights vectors $w \in \mathbf{R}^k$, $w \geq 0$, $\mathbf{1}^T w = 1$, unknown mean vectors $\mu_j \in \mathbf{R}^d$ and fixed known covariances $\frac{1}{\beta} I_{d\times d}$.

The problem of fitting the mixture distribution becomes:

$$\max_{w,\mu_1,\ldots,\mu_k} \frac{1}{n} \sum_{i=1}^n \log \sum_{j=1}^k w_j e^{-\beta \|z_i-\mu_j\|_2^2}.$$

This is a non-convex problem and is very hard to solve.

Convex clustering

The examplar-based model

Examplar-based model (Lashkari & Golland, NIPS, 2008): assumes that each cluster mean μ_j is equal to some data point ("example") z_i.

The problem is now convex:

$$\max_{w \geq 0,\ \mathbf{1}^T w = 1} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{j=1}^{n} w_j e^{-\beta \|z_i - z_j\|_2^2}.$$

Since $k = n$ there are a maximum number of clusters in the mixture! .

Our bound

Idea: penalize the cardinality of the mixture to control the number of clusters

$$p^* := \max_{w \geq 0,\, \mathbf{1}^T w = 1} \sum_{i=1}^n \log \left[\sum_{j=1}^n w_j K_{ij} \right] - \lambda ||w||_0$$

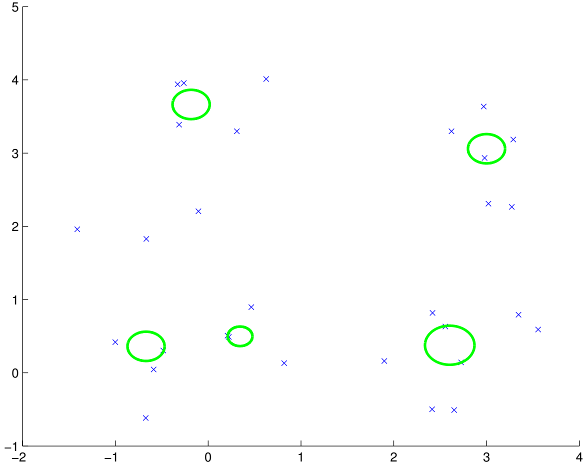
where K ($K_{ij} = e^{-\beta \|z_i - z_j\|_2^2}$) is a kernel matrix that can be pre-computed.

Our bound: $p^* \leq \hat{p}$, with

$$\hat{p} := \max_{1 \leq k \leq n} \max_{w \geq 0,\, \mathbf{1}^T w = 1} \sum_{i=1}^n \log \left[\sum_{j=1}^n w_j K_{ij} \right] - \frac{\lambda}{w_k}.$$

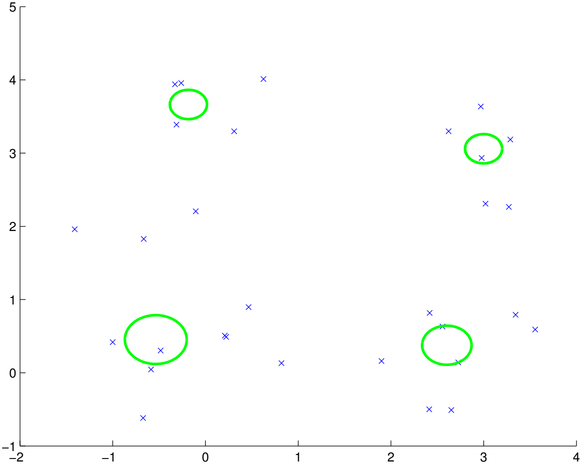
For every k, an SOCP!

Numerical results: comparison with soft k-means



Soft k-means : k = 4

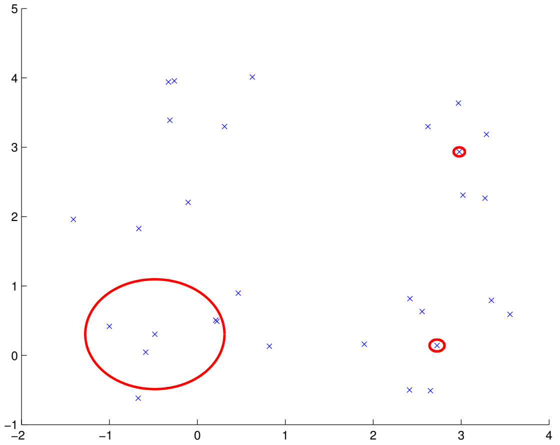
Numerical results: comparison with soft k-means



Soft k-means : k = 8

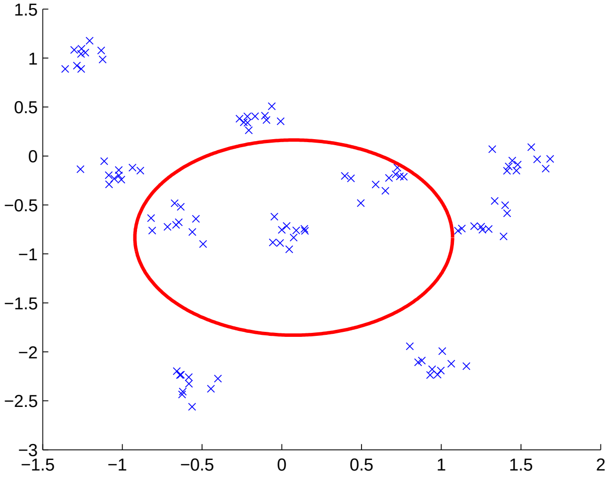
Numerical results: comparison with soft k-means

k-means can't find all the clusters!



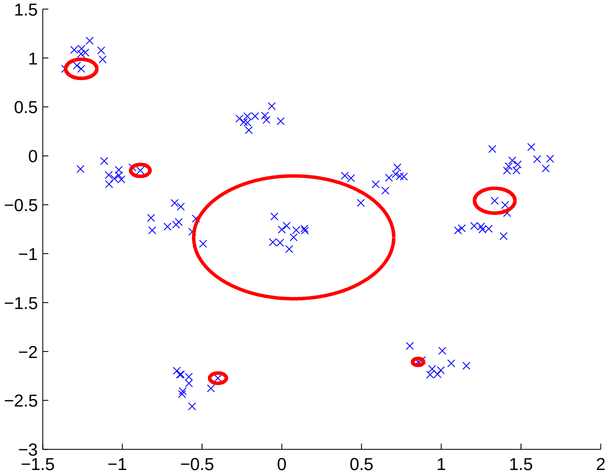
Soft k-means : k = 10

Numerical results: comparison with soft k-means



Proposed method : $\lambda = 1000$

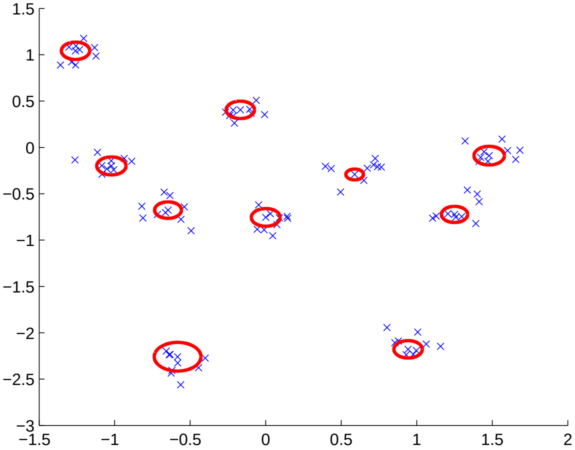
Numerical results: comparison with soft k-means



Proposed method : $\lambda = 100$

Numerical results: comparison with soft k-means

$\lambda = 45$ finds all the correct clusters!



Proposed method : $\lambda = 45$

Extensions

This strategy can be applied to almost every cardinality problem. Not only in the probability simplex!

- Basis Pursuit Denoising:

$$p^* = \min_w \|w\|_0 \;:\; \|X^T w - y\|_2 \leq \epsilon. \tag{2}$$

- Sparse Support Vector Machines:

$$p^* = \min_{w,b} \|w\|_0 \;:\; y_i(w^T x_i + b) \geq 1, \quad i = 1, \ldots, m.$$

The corresponding lower-bound approximations $\hat{p}$ can be solved using n convex programs.

- Provides a lower and upper-bound on p^* (ℓ_1 formulations are neither a lower nor an upper bound).
- In sparse recovery (2), it provably outperforms its ℓ_1 variant LASSO with Gaussian iid design.
- Preprints and code: `www.eecs.berkeley.edu/~mert`.

Outline

Sparse Optimization

Sparse Probability Optimization
- Sparse probabilities
- Norm-ratios approach
- Some recovery results
- Applications
- Extensions

Robust Optimization for Dimensionality Reduction
- Robust low-rank LP
- Low-rank LASSO

Low-rank LP

Consider a linear programming problem in n variables with m constraints:

$$\min_x \ c^T x \ : \ Ax \leq b,$$

with $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, and such that

- Many different problem instances involving the *same* matrix A have to be solved.
- The matrix A is close to low-rank.

- Clearly, we can approximate A with a low-rank matrix A_{lr} *once*, and exploit the low-rank structure to solve many instances of the LP fast.
- In doing so, we cannot guarantee that the solutions to the approximated LP are even feasible for the original problem.

Approach: robust low-rank LP

For the LP

$$\min_x \; c^T x \;:\; Ax \le b,$$

with many instances of b, c:

- Invest in finding a low-rank approximation $A_{\rm lr}$ to the data matrix A, and estimate $\epsilon := \|A - A_{\rm lr}\|$.
- Solve the *robust counterpart*

$$\min_x \; c^T x \;:\; (A_{\rm lr} + \Delta)x \le b \;\; \forall \, \Delta, \;\; \|\Delta\| \le \epsilon.$$

- Robust counterpart can be written as SOCP

$$\min_{x,t} \; c^T x \;:\; A_{\rm lr}x + t\mathbf{1} \le b, \;\; t \ge \|x\|_2.$$

- We can exploit the low-rank structure of $A_{\rm lr}$ and solve the above problem in time linear in $m + n$, for fixed rank.

Motivation: topic imaging

Task: find a short list of words that summarizes a topic in a large corpus.

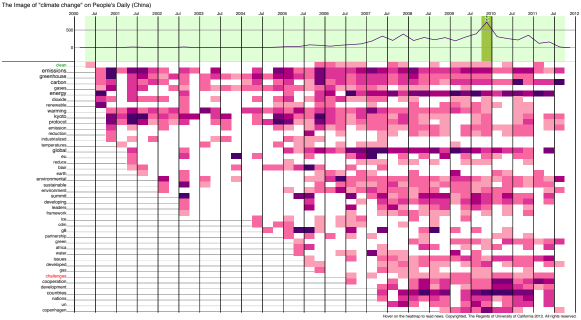


Image of topic "Climate change" over time. Each square encodes the size of regression coefficient in LASSO. *Source:* People's Daily, 2000-2011.

Interactive plot at

http://atticus.berkeley.edu/guanchengli/pd_climate_change/

Low-rank LASSO

In many learning problems, we need to solve many instances of the LASSO problem

$$\min_{w} \|X^T w - y\|_2 + \lambda \|w\|_1.$$

where

- For all the instances, the matrix X is a rank-one modification of the same matrix $\tilde{X}$.
- Matrix $\tilde{X}$ is close to low-rank (hence, X is).

In the topic imaging problem:

- $\tilde{X}$ is a term-by-document matrix that represents the whole corpus.
- y is one row of $\tilde{X}$ that encodes presence or absence of the topic in documents.
- X contains all remaining rows.

Robust low-rank LASSO

The robust low-rank LASSO

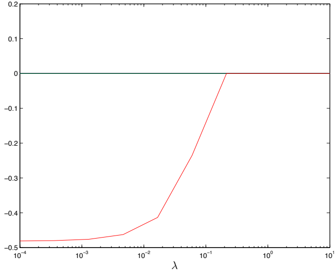
$$\min_{w} \max_{\|\Delta\| \leq \epsilon} \|(X_{\text{lr}} + \Delta)^T w - y\|_2 + \lambda \|w\|_1$$

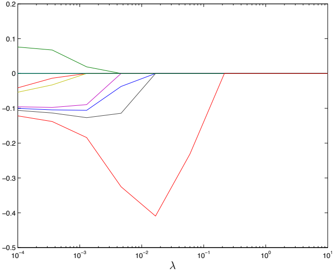
is expressed as a variant of “elastic net”:

$$\min_{w} \|X_{\text{lr}}^T w - y\|_2 + \lambda \|w\|_1 + \epsilon \|w\|_2.$$

- Solution can be found in time linear in $m + n$, for fixed rank.
- Solution has much better properties than low-rank LASSO, *e.g.* we can control the amount of sparsity.

Example





Rank-1 LASSO (left) and Robust Rank-1 LASSO (right) with random data. The plot shows the elements of the solution as a function of the l_1-norm penalty parameter.

- Without robustness ($\epsilon = 0$), the cardinality is 1 for $0 < \lambda < \lambda_{\max}$, where $\lambda_{\max}$ is a function of data. For $\lambda \geq \lambda_{\max}$, $w = 0$ at optimum. Hence the l_1-norm fails to control the solution.
- With robustness ($\epsilon = 0.01$), increasing λ allows to gracefully control the number of non-zeros in the solution.

Numerical experiments: low-rank approximation

Are real-world datasets approximately low-rank?

Dataset	TMC2007		RCV1V2		NYTIMES		PUBMED	
n		28,596		23,149		300,000		8,200,000
d		49,060		46,236		102,660		141,043
	Time (s)	σ_{k+1}/σ_1	Time (s)	σ_{k+1}/σ_1	Time (s)	σ_{k+1}/σ_1	Time (s)	σ_{k+1}/σ_1
k = 5	1	0.1539	1	0.2609	47	0.4095	187	0.4072
k = 10	1	0.1196	1	0.2100	50	0.3075	451	0.3494
k = 15	1	0.1010	1	0.1907	59	0.2709	520	0.3041
k = 20	2	0.0958	2	0.1769	73	0.2432	589	0.2793
k = 25	3	0.0909	3	0.1662	87	0.2312	687	0.2680
k = 30	4	0.0880	4	0.1615	93	0.2180	794	0.2580
k = 35	4	0.0858	4	0.1555	114	0.2098	932	0.2477
k = 40	5	0.0836	5	0.1507	130	0.2012	1150	0.2354
k = 45	6	0.0826	5	0.1475	142	0.1932	1208	0.2255
k = 50	7	0.0811	7	0.1430	158	0.1850	1862	0.2209

Runtimes[1] for computing a rank-k approximation to the whole data matrix.

[1] Experiments are conducted on a personal work station: 16GB RAM, 2.6GHz quad-core Intel.

Multi-label classification

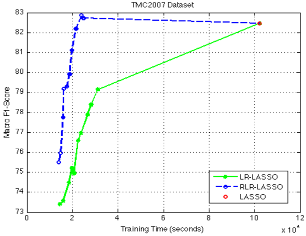
In multi-label classification, the task involves the same data matrix X, but many different response vectors y.

- Treat each label as a single classification subproblem (one-vs-all).
- Evaluation metric: Macro-F1 measure.
- Datasets:
  - RCV1-V2: 23,149 training documents; 781,265 test documents; 46,236 features; 101 labels.
  - TMC2007: 28,596 aviation safety reports; 49,060 features; 22 labels.

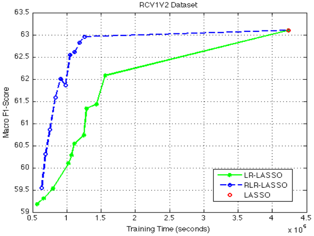
Multi-label classification

Plot performance vs. training times for various values of rank $k = 5, 10, \ldots, 50$.

TMC 2007 data set



RCV1V2 data set



In both cases, the low-rank robust counterpart allows to recover the performance obtained with full-rank LASSO (red dot), for a fraction of computing time.

Topic imaging

- Labels are columns of whole data matrix $\tilde{X}$.
- Compute low-rank approximation of $\tilde{X}$ when a column is removed.
- Evaluation: report predictive word lists for 10 queries.
- Datasets:
  - NYTimes: 300,000 documents; 102,660 features, file size is 1GB. Queries: 10 industry sectors.
  - PUBMED: 8,200,000 documents; 141,043 features, file size is 7.8GB. Queries: 10 diseases.
- In both cases we have pre-computed a rank k ($k = 20$) approximation using power iteration.

Topic imaging

automotive	agriculture	technology	tourism	aerospace	defence	financial	healthcare	petroleum	gaming
car	government	company	tourist	boeing	afghanistan	company	health	oil	game
vehicle	farm	computer	hotel	aircraft	attack	million	care	prices	gambling
auto	farmer	system	business	space	forces	stock	cost	gas	casino
sales	food	web	visitor	program	military	market	patient	fuel	player
model	water	information	economy	jet	gulf	money	corp	company	online
driver	trade	internet	travel	plane	troop	business	al_gore	barrel	computer
ford	land	american	tour	nasa	aircraft	firm	doctor	gasoline	tribe
driving	crop	job	local	flight	terrorist	fund	drug	bush	money
engine	economic	product	room	airbus	president	investment	medical	energy	playstation
consumer	country	software	plan	military	war	economy	insurance	opec	video

The New York Times data: Top 10 predictive words for different queries corresponding to industry sectors.

arthritis	asthma	cancer	depression	diabetes	gastritis	hiv	leukemia	migraines	parkinson
joint	bronchial	tumor	effect	diabetic	gastric	aid	cell	headache	treatment
synovial	asthmatic	treatment	treatment	insulin	h_pylori	infection	acute	headaches	effect
infection	children	carcinoma	disorder	level	chronic	cell	bone_marrow	pain	nerve
chronic	respiratory	cell	depressed	glucose	ulcer	hiv-1	leukemic	disorder	syndrome
pain	symptom	chemotherapy	pressure	control	acid	infected	tumor	women	disorder
treatment	allergic	survival	anxiety	plasma	stomach	antibodies	remission	chronic	neuron
fluid	infant	risk	symptom	diet	atrophic	risk	t_cell	duration	receptor
knee	inhalation	dna	drug	liver	antral	positive	antigen	symptom	alzheimer
acute	airway	malignant	neuron	renal	reflux	transmission	chemotherapy	gene	response
therapy	fev1	diagnosis	response	normal	treatment	drug	expression	therapy	brain

PubMed data: Top 10 predictive words for different queries corresponding to diseases.