

Robustness and Regularization: An Optimization Perspective

Laurent El Ghaoui (EECS/IEOR, UC Berkeley)
with help from Brian Gawalt, Onureena Banerjee

Neyman Seminar, Statistics Department, UC Berkeley

September 24, 2008

Agenda

- Robust optimization
- Example: robust SVM
- Worst-case loss function minimization
- Statistical analysis of online news

Convex optimization

“Nominal” problem:

$$\min_x f_0(x) \quad : \quad f_i(x) \leq 0, \quad i = 1, \dots, m$$

f_0, f_i 's are *convex*

- includes many problems arising in statistics
- efficient (polynomial-time) algorithms
- can use convex relaxations for non-convex problems

Robust counterpart

$$\min_x \max_{\delta \in \mathcal{D}} f_0(x, \delta) \quad : \quad \forall \delta \in \mathcal{D}, \quad f_i(x, \delta) \leq 0, \quad i = 1, \dots, m$$

- functions f_i now depend on a second variable, the “uncertainty”
- The uncertainty vector δ is constrained to lie in given set \mathcal{D}
(WLOG, \mathcal{D} is convex)
- Complexity is high in general, systematic ways to get relaxations

Penalty approach

Often, optimal value and solutions of optimization problems are sensitive to data

A common approach to deal with sensitivity is via penalization, eg

$$\min_x f_0(x) + \lambda \|Wx\|_2^2 \quad (W = \text{weighting matrix})$$

- How do we choose the penalty?
- Can we choose it in a way that reflect knowledge about problem structure?
- Does it have anything to do with uncertainty affecting data?

Goal of talk

- Explore connections between robustness and penalty design in the context of optimization problems arising in statistics
- Report on preliminary (non experimental) results
- Outline cases when robust counterpart leads to tractable problems

Related work: Bhattacharyya, Bertsimas, Caramanis et al

Example: robust SVM

Support Vector Machine (SVM) classification problem:

$$\min_{w,b} \sum_{i=1}^m (1 - y_i(z_i^T w + b))_+$$

- $Z = [z_1, \dots, z_m] \in \mathbf{R}^{n \times m}$ contains the *data points* (feature vectors)
- $y \in \{-1, 1\}^m$ contain the *labels*
- $x = (w, b)$ contains the *classifier parameters*, allowing to classify a new point z via the rule $y = \mathbf{sgn}(z^T w + b)$

Robust SVM

Assume the data matrix is only *partially known*, and address the robust optimization problem:

$$\min_{w,b} \max_{\Delta \in \mathcal{D}} \sum_{i=1}^m (1 - y_i((z_i + \delta_i)^T w + b))_+$$

where $\Delta = [\delta_1, \dots, \delta_m]$ and $\mathcal{D} \subseteq \mathbf{R}^{n \times m}$ is a set that describes additive uncertainty in the data matrix.

Measurement-wise, spherical uncertainty

Assume

$$\mathcal{D} = \{\Delta = [\delta_1, \dots, \delta_m] \in \mathbf{R}^{n \times m} : \|\delta_i\|_2 \leq \rho\},$$

where $\rho > 0$ is given.

Robust SVM reduces to

$$\min_{w,b} \sum_{i=1}^m (1 - y_i(z_i^T w + b) + \rho \|w\|_2)_+$$

Link with classical SVM

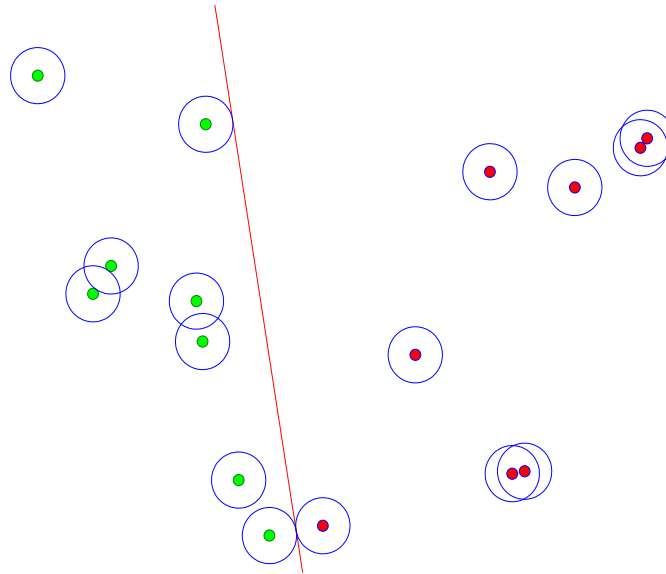
Classical SVM contains l_2 -norm regularization term:

$$\min_{w,b} \sum_{i=1}^m (1 - y_i(z_i^T w + b))_+ + \lambda \|w\|_2^2$$

where $\lambda > 0$ is a penalty parameter

With spherical uncertainty, *robust SVM is similar to classical SVM*

Separable data



Maximally robust classifier for separable data, with spherical uncertainties around each data point. In this case, the robust counterpart reduces to the familiar maximum-margin classifier problem.

Interval uncertainty

Assume

$$\mathcal{D} = \{\Delta \in \mathbf{R}^{n \times m} : \forall(i, j), |\Delta_{ij}| \leq \rho\},$$

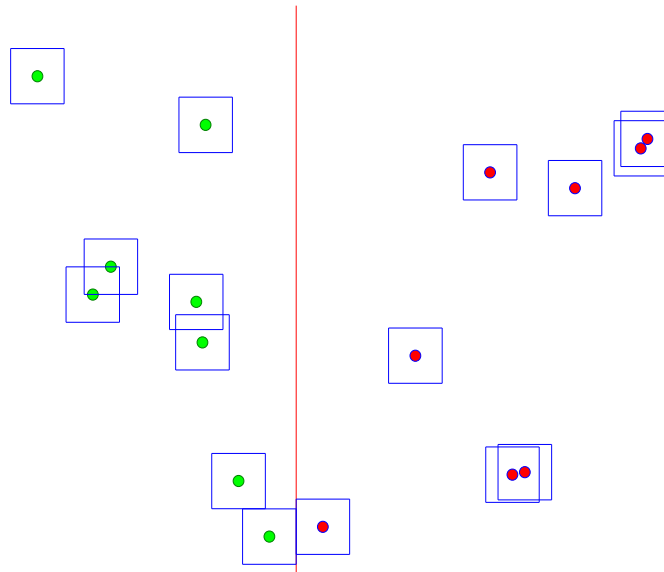
where $\rho > 0$ is given.

Robust SVM reduces to

$$\min_{w, b} \sum_{i=1}^m (1 - y_i(z_i^T w + b) + \rho \|w\|_1)_+$$

The l_1 -norm term encourages *sparsity*, may or may not regularize the solution

Separable data



Maximally robust classifier for separable data, with box uncertainties around each data point

Coupled uncertainty

- Previous uncertainty models are “measurement-wise”, with perturbations affecting each data point independent of each other
- It may be better to consider perturbations affecting the data matrix in a coupled way, for example the *norm-bound uncertainty model*

$$\Delta = \{\Delta \in \mathbf{R}^{n \times m} : \|\Delta\| \leq \rho\},$$

where $\|\cdot\|$ denotes largest singular value norm

Non-conservative formulation of the robust SVM

With norm-bound uncertainty, the robust SVM problem is equivalent to

$$\min_{w,b} \max_{k \in \{0, \dots, m\}} \rho \sqrt{k} \|w\|_2 + \sum_{i=1}^k (\mathbf{1} - D(y)(Z^T w + b\mathbf{1}))_{[i]}$$

- $D(y) = \mathbf{diag}(y)$, and $u_{[i]}$ stands for largest i -th component of u
- Can be written as a tractable second-order cone optimization problem (SOCP), and provides regularization
- Upper bounded by classical SVM with regularization parameter $\sqrt{n}\rho$

Robust classification and regression

Nominal problem:

$$\min_{\theta \in \Theta} \mathcal{L}(Z^T \theta),$$

where

- $Z := [z_1, \dots, z_m] \in \mathbf{R}^{n \times m}$ is the data matrix
- $\mathcal{L} : \mathbf{R}^m \rightarrow \mathbf{R}$ is a convex loss function
- Θ imposes “structure” (eg, sign) constraints on parameter vector θ

Loss function: assumptions

We assume that

$$\mathcal{L}(r) = \pi(\mathbf{abs}(P(r))),$$

where $\mathbf{abs}(\cdot)$ acts componentwise, $\pi : \mathbf{R}_+^m \rightarrow \mathbf{R}$ is a convex, monotone function on the non-negative orthant, and

$$P(r) = \begin{cases} r & \text{("symmetric case")} \\ r_+ & \text{("asymmetric case")} \end{cases}$$

with r_+ the vector with components $\max(r_i, 0)$, $i = 1, \dots, m$.

Loss function: examples

- l_p -norm regression
- hinge loss
- Huber, Berhu loss

Robust counterpart

$$\min_{\theta \in \Theta} \max_{\mathbf{Z} \in \mathcal{Z}} \mathcal{L}(\mathbf{Z}^T \theta).$$

where $\mathcal{Z} \subseteq \mathbf{R}^{n \times m}$ is a set of the form

$$\mathcal{Z} = \{Z + \Delta : \Delta \in \rho \mathcal{D}, \},$$

with $\rho \geq 0$ a measure of the size of the uncertainty, and $\mathcal{D} \subseteq \mathbf{R}^{l \times m}$ is given.

Preliminary analysis

For a given vector θ , we have

$$\max_{\mathbf{Z} \in \mathcal{Z}} \mathcal{L}(\mathbf{Z}^T \theta) = \max_u u^T Z^T \theta - \mathcal{L}^*(u) + \rho \phi_{\mathcal{D}}(uv^T),$$

where \mathcal{L}^* is the conjugate of \mathcal{L} , and

$$\phi_{\mathcal{D}}(X) := \max_{\Delta \in \mathcal{D}} \langle X, \Delta \rangle$$

is the support function of \mathcal{D}

Assumptions on uncertainty set \mathcal{D}

Separability condition: there exist two semi-norms ϕ, ψ such that

$$\phi_{\mathcal{D}}(uv^T) := \max_{\Delta \in \mathcal{D}} u^T \Delta v = \phi(u)\psi(v).$$

- Does not completely characterize (the support function of) the set \mathcal{D}
- Given ϕ, ψ , we can construct a set \mathcal{D}_{out} that obeys condition
- The robust counterpart only depends on ϕ, ψ

WLOG, we can replace \mathcal{D} by its convex hull

Examples

- Largest singular value model: $\mathcal{D} = \{\Delta : \|\Delta\| \leq \rho\}$, with ϕ, ψ Euclidean norms
- Any norm-bound model involving an induced norm (ϕ, ψ are then the norms dual to the norms involved)
- Measurement-wise uncertainty models, where each column of the perturbation matrix is bounded in norm, independently of the others, correspond to the case with $\psi(v) = \|v\|_1$

Other examples

- Bounded-error model: there are (at most K) errors affecting data

$$\mathcal{D} = \left\{ \Delta = [\lambda_1 \delta_1, \dots, \lambda_m \delta_m] \in \mathbf{R}^{l \times m} : \|\delta_i\| \leq 1, \quad i = 1, \dots, m, \right. \\ \left. \sum_{i=1}^m \lambda_i \leq K, \quad \lambda \in \{0, 1\}^m \right\},$$

for which $\phi(\cdot) = \|\cdot\|_*$, $\psi(v) = \text{sum of the } K \text{ largest magnitudes of the components of } v$.

Main result

For a given vector θ , we have

$$\min_{\theta} \max_{\mathbf{Z} \in \mathcal{Z}} \mathcal{L}(\mathbf{Z}^T \theta) = \min_{\theta, \kappa} \mathcal{L}_{\text{wc}}(Z^T \theta, \kappa) \quad : \quad \kappa \geq \phi(U^T \theta)$$

where

$$\mathcal{L}(r, \kappa) := \max_v v^T r - \mathcal{L}^*(v) + \kappa \psi(v)$$

is the *worst-case loss function* of the robust problem.

Worst-case loss function

The tractability of the robust counterpart is directly linked to our ability to compute optimal solutions v^* for

$$\mathcal{L}(r, \kappa) = \max_v v^T r - \mathcal{L}^*(v) + \kappa\psi(v)$$

Dual representation (assume $\psi(\cdot) = \|\cdot\|$ is a norm):

$$\mathcal{L}(r, \kappa) = \max_{\xi} \mathcal{L}(r + \kappa\xi) : \|\xi\|_* \leq 1$$

When ψ is the Euclidean norm, robust regularization of \mathcal{L} (Lewis, 2001)

Special cases

- When $\psi(\cdot) = \|\cdot\|_p$, $p = 1, \infty$, problem reduces to simple, tractable convex problem (assuming nominal problem is)
- For $p = 2$, problem can be reduced to such a simple form, for the hinge, l_q -norm and Huber loss functions

Lasso

In particular, the least-squares problem with lasso penalty

$$\min_{\theta} \|X^T \theta - y\|_2 + \rho \|\theta\|_1$$

is the robust counterpart to a least-squares problem with uncertainty on X , with additive perturbation bounded in the norm

$$\|\Delta\|_{1,2} := \max_{1 \leq i \leq l} \sqrt{\sum_{j=1}^n \Delta_{ij}^2}$$

Robust SVM with Boolean data

- Data: boolean $Z \in \{0, 1\}^{n \times m}$ (eg, co-occurrence matrix)
- Nominal problem: hinge loss minimization
- Uncertainty model: assume each data value can be flipped, total budget of flips is constrained:

$$\mathcal{D} = \left\{ \Delta = [\delta_1, \dots, \delta_m] \in \mathbf{R}^{l \times m} : \delta_i \in \{-1, 0, 1\}^l, \|\delta\|_1 \leq k \right\}$$

In this case, we have $\psi(\cdot) = \|\cdot\|_1$ and

$$\phi(u) = \|u\|_{1,k} := \min_s k\|u - s\|_\infty + \|s\|_1$$

Robust counterpart

$$\min_{w,b} \sum_{i=1}^m (1 - y_i(z_i^T w + b) + \phi(w))_+$$

- Penalty is a combination of l_1, l_∞ norms
- Problem is tractable (double number of variables over nominal)
- Still needs regularization

Refined model

We can impose $\delta_i \in \{x_i, x_i - 1\}$, leads to

$$\min_{w,b} \sum_{i=1}^m (1 - y_i(z_i^T w + b) + \phi_i(w))_+$$

with

$$\phi_i(w) =$$

Chance constraints

Theory can address problems with “chance constraints”

$$\min_{\theta} \max_{p \in \mathcal{P}} \mathbf{E}_p \mathcal{L}(Z(\delta)^T \theta)$$

where δ follows distribution p , and \mathcal{P} is a class of distributions

- Results are more limited, focused on upper bounds
- Convex relaxations are available, but more expensive
- Approach uses Bernstein approximations (Nemirovski & Ben-tal, 2006)

Robust regression with chance constraints: an example

$$\phi_p := \min_{\theta} \max_{x \sim (\hat{x}, X)} \mathbf{E}_x \|A(x)\theta - b(x)\|_p$$

- Regression variable is $\theta \in \mathbf{R}^n$
- $x \in \mathbf{R}^q$ is an uncertainty vector that enters affinely in the problem matrices: $[A(x), b(x)] = [A_0, b_0] + \sum_i x_i [A_i, b_i]$
- The distribution of uncertainty vector x is unknown, except for its mean \hat{x} and covariance X
- Objective is worst-case (over distributions) expected value of l_p -norm residual ($p = 1, 2$)

Main result

(Assume $\hat{x} = 0$, $X = I$ WLOG)

For $p = 2$, the problem reduces to least-squares:

$$\phi_2^2 = \min_{\theta} \sum_{i=0}^q \|A_i \theta - b_i\|_2^2$$

For $p = 1$, we have $(2/\pi)\psi_1 \leq \phi_1 \leq \psi_1$, with

$$\psi_1 = \min_{\theta} \sum_{i=0}^q \|A_i \theta - b_i\|_2$$

Example: robust median

As a special case, consider the *median problem*:

$$\min_{\theta} \sum_{i=1}^q |\theta - x_i|$$

Now assume that *vector x is random*, with mean \hat{x} and covariance X , and consider the robust version:

$$\phi_1 := \min_{\theta} \max_{x \sim (\hat{x}, X)} \mathbf{E}_x \sum_{i=1}^q |\theta - x_i|$$

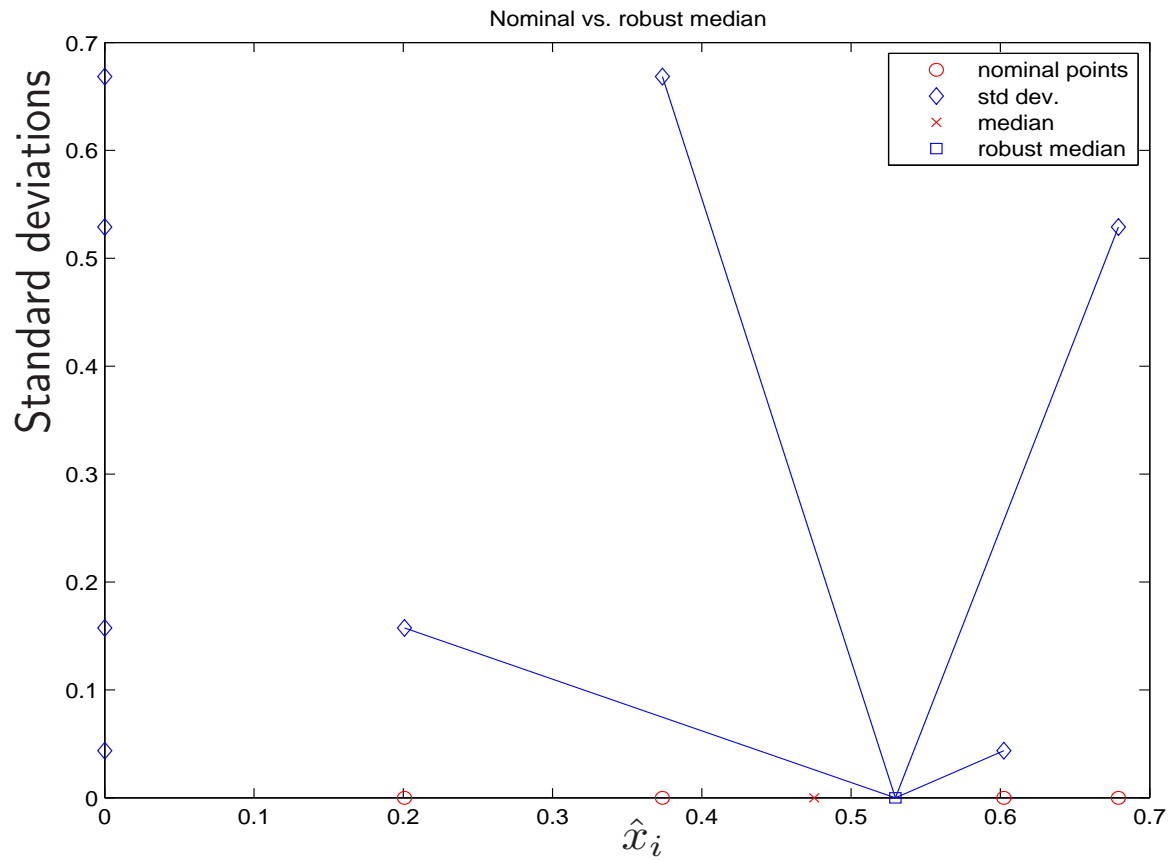
Approximate solution

We have $(2/\pi)\psi_1 \leq \phi_1 \leq \psi_1$, with

$$\psi_1 := \sum_{i=1}^n \sqrt{(\theta - \hat{x}_i)^2 + X_{ii}}$$

Amounts to find the minimum distance sum (a very simple SOCP)

Geometry of robust median problem



Statistical analysis of online news

New project started in 2007, with collaborators:

- Bin Yu (UCB, Stat), Alexandre d'Aspremont (ORFE, Princeton), Joe Hellerstein, Manesh Agrawala (CS, UCB)
- Charles Cameron (Pol Sci, Princeton), Henry Brady (Pol Sci, UC Berkeley), Suad Joseph (Anthropology, UC Davis), Sophie Clavier (Pol Sci, SFSU)
- Students Brian Gawalt, Onureena Banerjee

Funding NSF (CDI), Google

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

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

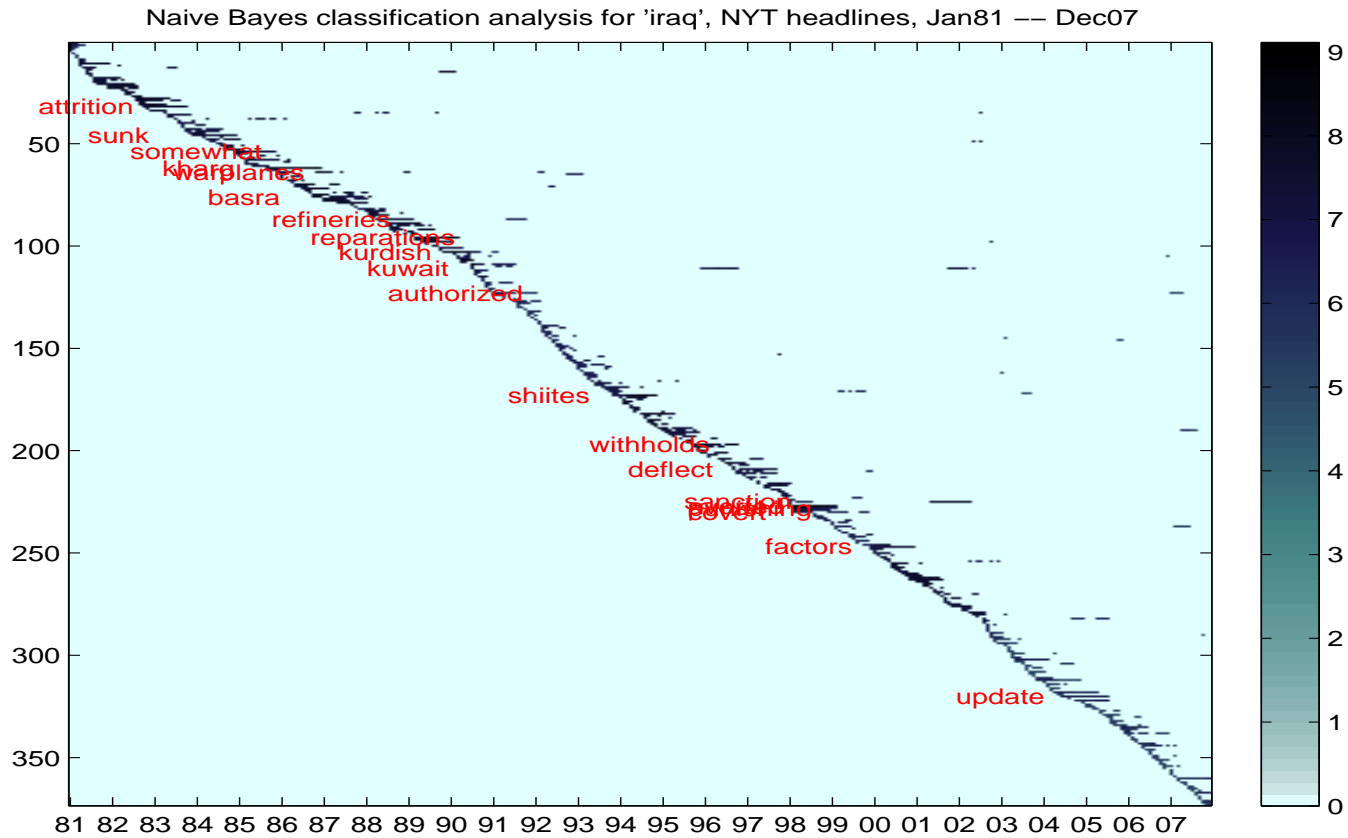
The main challenge is to connect these “soft” goals with “hard” statistical concepts and methods

Image 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

Example: 'Iraq' in NYT headlines, 1990-2007



Challenges

- Account for noise in text data (boolean/. . .)
- Group sparsity
- Robustness of solutions
- Fast updates
- Solving the problems on a cluster