8. Variations on least-squares

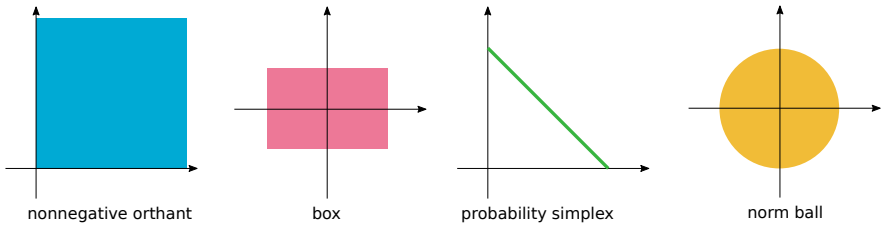
- least-squares with constraints
- ℓ_2 regularization
- ℓ_1 regularization
- generalizations of ℓ_1-regularized LS
- robust least-squares

Least-squares with constraints

$$\begin{array}{ll} \text{minimize} & \|Ax - y\| \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

$\mathcal{C}$ is a convex set (many applications fall into this case)

- used to rule out certain unacceptable approximations of y
- arise as prior knowledge of the vector x to be estimated
- same as determining the projection of y on a set more complicated than a subspace
- form a convex optimization problem with no analytical solution (typically)



nonnegativity constraints on variables

$$\mathcal{C} = \{\, x \mid x \succeq 0 \,\}$$

- parameter x known to be nonnegative, e.g., powers, rates, etc.
- finding the projection of y onto the *cone* generated by the columns of A

variable bounds

$$\mathcal{C} = \{\, x \mid l \preceq x \preceq u \,\}$$

- vector x known to lie in an interval $[l, u]$
- finding the projection of y onto the image of a box under the linear mapping induced by A

probability distribution

$$\mathcal{C} = \{\, x \;\mid\; x \succeq 0, \quad \mathbf{1}^T x = 1 \,\}$$

- arise in estimation of proportions which are nonnegative and sum to one
- approximating y by a convex combination of the columns of A

norm ball constraint

$$\mathcal{C} = \{\, x \;\mid\; \|x - x_0\| \leq d \,\}$$

where x_0 and d are problem parameters

- x_0 is a prior guess of what x should be
- d is the maximum plausible deviation from our prior guess
- the constraints $\|x - x_0\| \leq d$ can denote a **trust region**. (the linear relation $y = Ax$ is an approximation and only valid when x is near x_0)

ℓ_2-regularized least-squares

adding the 2-norm penalty to the objective function

$$\underset{x}{\text{minimize}} \quad \|Ax - y\|_2^2 + \gamma \|x\|_2^2$$

- seek for an approximate solution of $Ax \approx y$ with small norm
- also called **Tikhonov regularized least-squares** or **ridge regression**
- $\gamma > 0$ controls the trade off between the fitting error and the size of x
- has the analytical solution for any $\gamma > 0$:

$$x = (A^T A + \gamma I)^{-1} A^T y$$

  (no restrictions on shape, rank of A)

- interpreted as a MAP estimation with the log-prior of the Gaussian

ℓ_1-regularized least-squares

Idea: adding $|x|$ to a minimization problem introduces a sparse solution

consider a scalar problem:

$$\underset{x}{\text{minimize}} \quad f(x) = (1/2)(x-a)^2 + \gamma|x|$$

to derive the optimal solution, we consider the two cases:

- if $x \geq 0$ then $f(x) = (1/2)(x-(a-\gamma))^2$

$$x^\star = a - \gamma, \quad \text{provided that } a \geq \gamma$$

- if $x \leq 0$ then $f(x) = (1/2)(x-(a+\gamma))^2$

$$x^\star = a + \gamma, \quad \text{provided that } a \leq -\gamma$$

when $|a| \leq \gamma$ then $x^\star$ must be zero

the optimal solution to minimization of $f(x) = (1/2)(x-a)^2 + \gamma|x|$ is

$$x^\star = \begin{cases} (|a| - \gamma)\mathbf{sign}(a), & |a| > \gamma \\ 0, & |a| \leq \gamma \end{cases}$$

meaning: if γ is large enough, x^* will be zero

generalization to vector case: $x \in \mathbf{R}^n$

$$\underset{x}{\text{minimize}} \quad f(x) = (1/2)\|x-a\|^2 + \gamma\|x\|_1$$

the optimal solution has the same form

$$x^\star = \begin{cases} (|a| - \gamma)\mathbf{sign}(a), & |a| > \gamma \\ 0, & |a| \leq \gamma \end{cases}$$

where all operations are done in *elementwise*

ℓ_1-regularized least-squares

adding the ℓ_1-norm penalty to the least-square problem

$$\underset{x}{\text{minimize}} \;\; (1/2)\|Ax - y\|_2^2 + \gamma\|x\|_1 \tag{1}$$

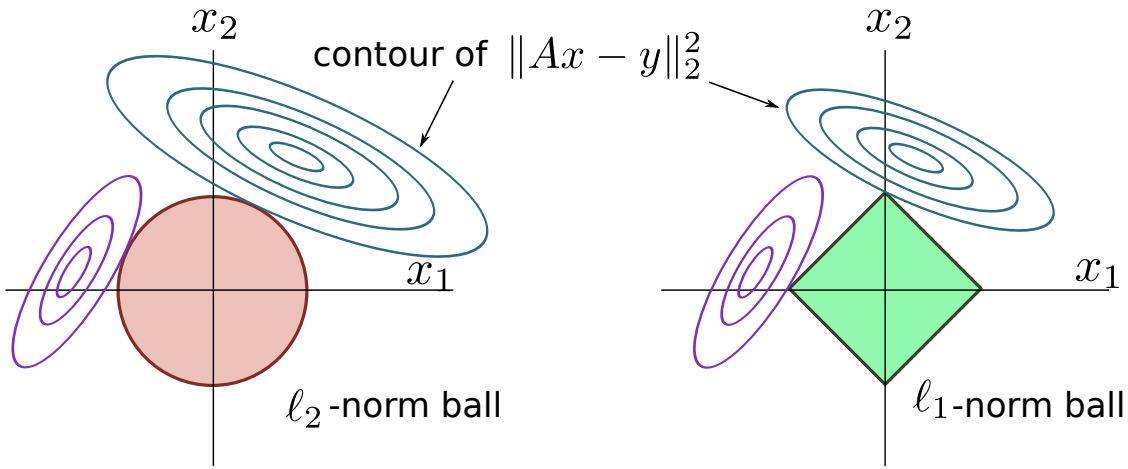
- a convex heuristic method for finding a sparse x that gives $Ax \approx y$
- also called **Lasso** or **basis pursuit**
- a nondifferentiable problem due to $\|\cdot\|_1$ term
- no analytical solution, but can be solved efficiently
- interpreted as a MAP estimation with the log-prior of the Laplacian distribution

Similar form of ℓ_1-regularized LS

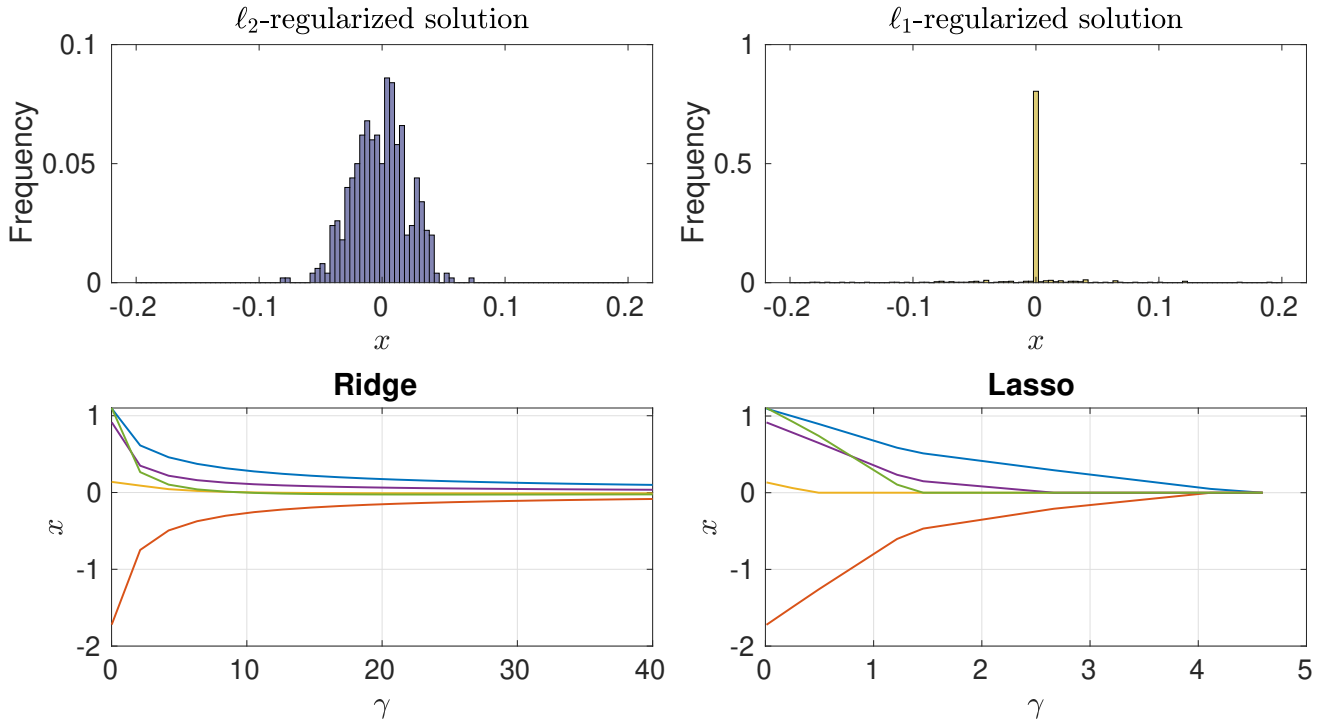
the ℓ_1-norm is an inequality constraint:

$$\underset{x}{\text{minimize}} \quad \|Ax - y\|_2^2 \quad \text{subject to} \quad \|x\|_1 \leq t \tag{1}$$

- t is specified by the user; serves as a budget of the sum of absolute values of x
- the ℓ_1-regularized LS (1) is the Lagrangian form of this problem
- for each t where $\|x\|_1 \leq t$ is active, there is a corresponding value of γ that yields the same solution from (1)



example $A \in \mathbf{R}^{m \times n}, b \in \mathbf{R}^m$ with $m = 100, n = 500, \gamma = 0.2$



- histogram of ℓ_2 solution is widely spread while ℓ_1 is more concentrated at zero
- (bottom: $n = 5$) many entries of ℓ_1 solution are exactly zero as γ varies while entries of ℓ_2 solution converge to small values

Generalizations of ℓ_1-regularized LS

many variants are proposed for acheiving particular structures in solutions

- elastic net: for highly correlated variables and lasso doesn't perform well
- group lasso: for acheiving sparsity in group
- fused lasso: for neighboring variables to be similar

Elastic net

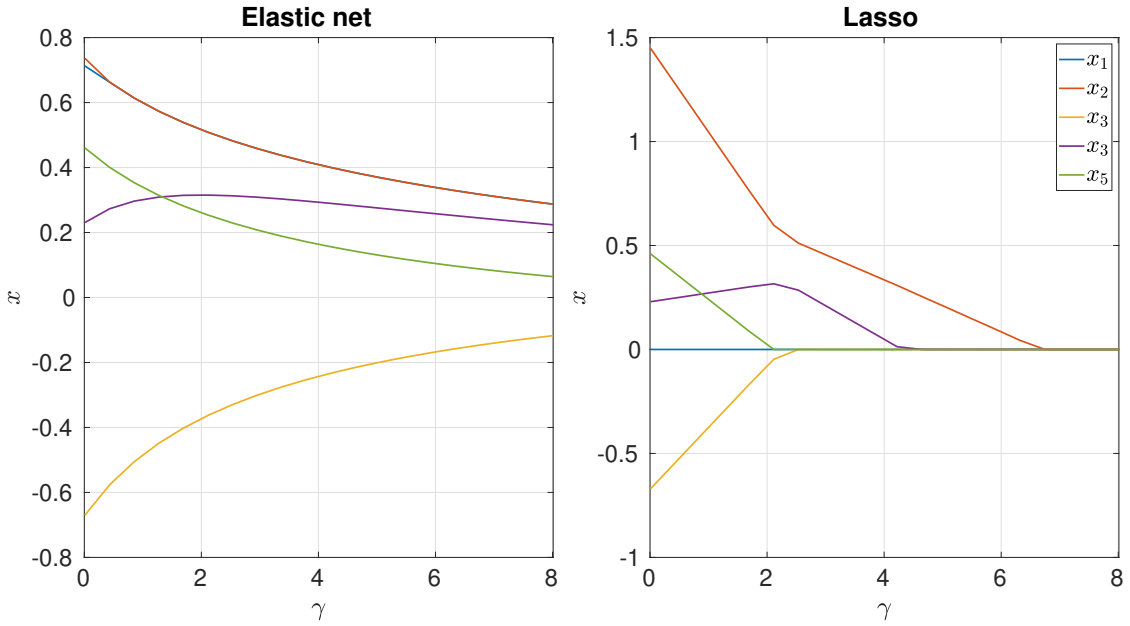
a combination between the ℓ_1 and ℓ_2 regularizations

$$\underset{x}{\text{minimize}} \;\; (1/2)\|Ax - y\|_2^2 + \gamma \left\{ (1/2)(1-\alpha)\|x\|_2^2 + \alpha\|x\|_1 \right\}$$

where $\alpha \in [0, 1]$ and γ are parameters

- the problem reduces to a lasso when $\alpha = 1$ and to a ridge regression when $\alpha = 0$
- used when we expect groups of very correlated variables (e.g. microarray, genes)
- strictly convex problem for any $\alpha < 1$ and $\lambda > 0$ (unique solution)

generate $A \in \mathbf{R}^{20\times 5}$ where a_1 and a_2 are highly correlated



- if $a_1 = a_2$, the ridge estimate of x_1 and x_2 will be equal (not obvious, please verify)
- the blue and orange lines correspond to the variables x_1 and x_2
- the lasso does not reflect the relative importance of the two variables
- using $\alpha = 0.1$, the elastic net selects the estimates of x_1 and x_2 together

Group lasso

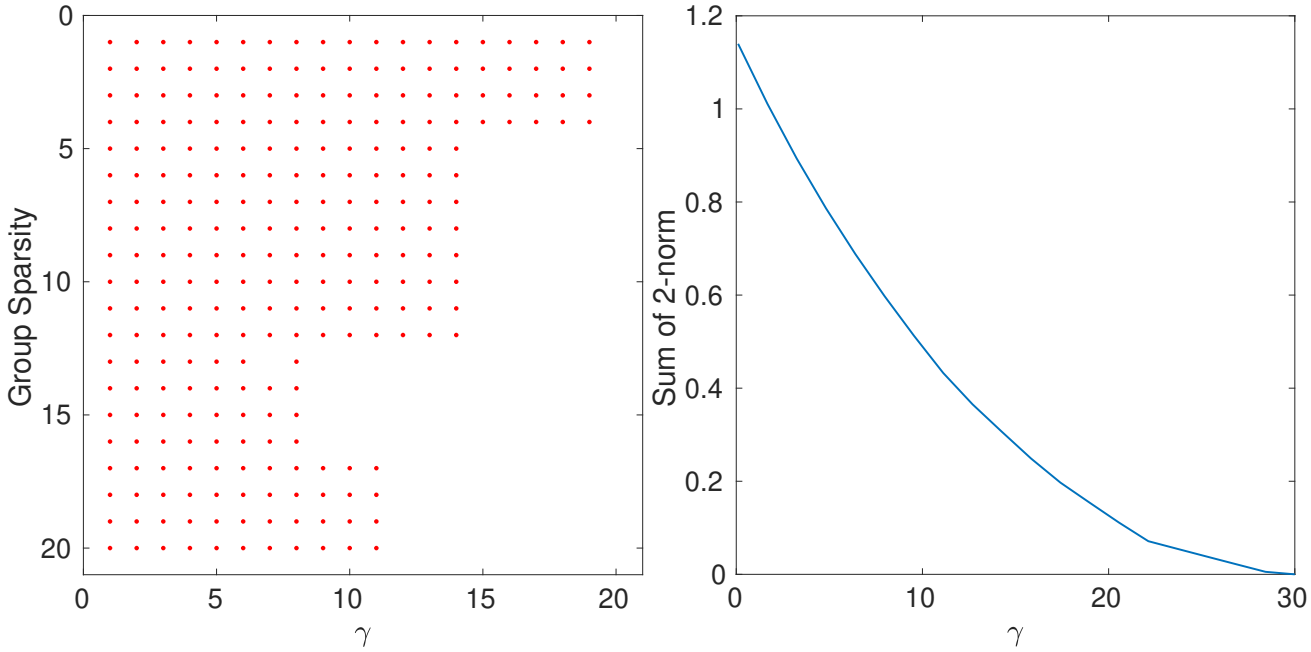
to have all entries in x within a *group* become zero simultaneously

let $x = (x_1, x_2, \ldots, x_K)$ where $x_j \in \mathbf{R}^n$

$$\text{minimize} \;\; (1/2)\|Ax - y\|_2^2 + \gamma \sum_{j=1}^{K} \|x_j\|_2$$

- the sum of ℓ_2 norm is a generalization of ℓ_1-like penalty
- as γ is large enough, either x_j is entirely zero or all its element is nonzero
- when $n = 1$, group lasso reduces to the lasso
- a non-smooth convex problem but can be solved efficiently

generate the problem with $x = (x_1, x_2, \ldots, x_5)$ where $x_i \in \mathbf{R}^4$



- as γ increases, some of partition x_i becomes entirely zero
- as the sum of 2-norm is zero, the entire vector x is zero

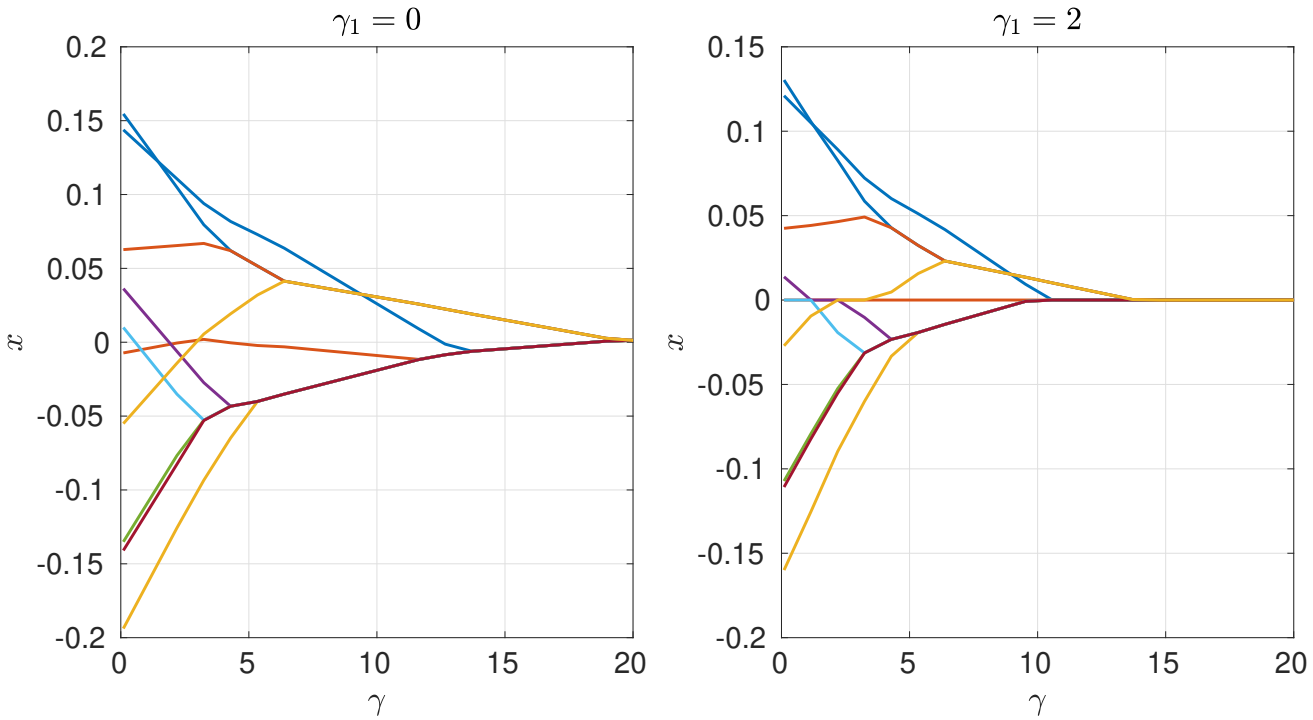
Fused lasso

to have neighboring variables similar and sparse

$$\underset{x\in\mathbf{R}^n}{\text{minimize}} \;\; (1/2)\|Ax-y\|_2^2+\gamma_1\|x\|_1+\gamma_2\sum_{j=2}^{n}|x_j-x_{j-1}|$$

- the ℓ_1 penalty serves to shrink x_i toward zero
- the second penalty is ℓ_1-type encouraging some pairs of consecutive entries to be similar
- also known as **total variation denoising** in signal processing
- γ_1 controls the sparsity of x and γ_2 controls the similarity of neighboring entries
- a nondifferentiable convex problem but can be solved efficiently

generate $A \in \mathbf{R}^{100\times 10}$ and vary γ_2 with two values of γ_1



- as γ_2, consecutive entries of x tend to be equal
- for a higher value of γ_1, some of the entries of x become zero

Robust least-squares

consider the LS problem

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|_2$$

but A may have variation or some uncertainty

we can treat the uncertainty in A in different ways

- A is deterministic but belongs to a set
- A is stochastic

Worst-case robust least-squares

describe the uncertainty by a set of possible values for A:

$$A \in \mathcal{A} \subseteq \mathbf{R}^{m \times n}$$

the problem is to minimize the worst-case error:

$$\underset{x}{\text{minimize}} \quad \sup_{A} \ \{\|Ax - y\|_2 \mid A \in \mathcal{A}\}$$

- always a convex problem
- its tractablity depends on the description of $\mathcal{A}$

example 1: given $\mathcal{A} = \{\bar{A} + E \mid \|E\|_F \leq e\}$

- meaning: each column in A corresponds to measurements of a variable recorded thru a sensor given with noise RMS
- define $w = \bar{A}x - y$, the worst-case norm-2 can be calculated by

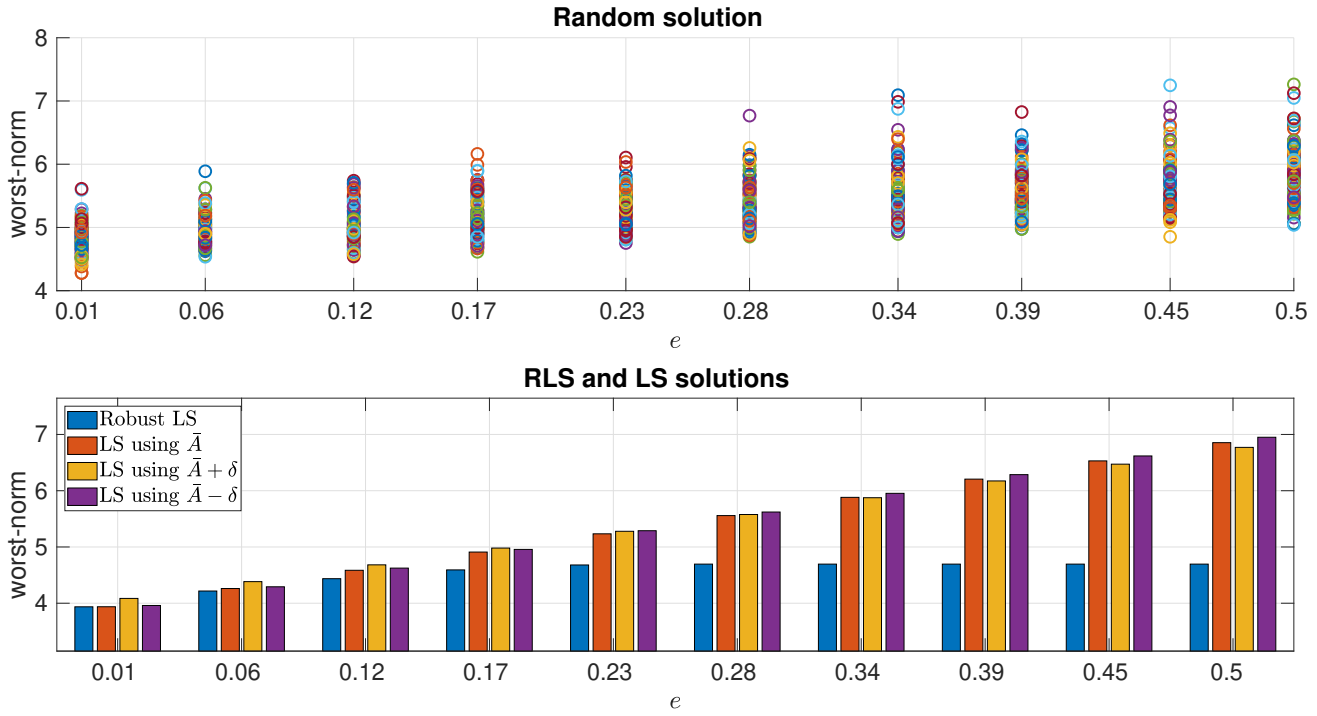
$$\begin{aligned}
\|Ax - y\|^2 = \|Ex + w\|^2 &= x^T E^T E x + 2w^T E x + \|w\|^2 \\
&\leq \lambda_{\max}(E^T E)\|x\|^2 + 2\,\mathbf{tr}((wx^T)^T E) + \|w\|^2 \qquad (1) \\
&\leq \|E\|_F^2 \|x\|^2 + 2\|wx^T\|_F \|E\|_F + \|w\|^2 \qquad (2) \\
&\leq e^2\|x\|^2 + 2e\|w\|\|x\| + \|w\|^2 = (e\|x\| + \|w\|)^2 \qquad (3)
\end{aligned}$$

- the worst-case norm is attained when $E = \alpha w x^T$ where $\alpha = e/\|w\|\|x\|$

$$\sup_{A \in \mathcal{A}} \|Ax - y\|_2 = \|\bar{A}x - y\|_2 + e\|x\|_2$$

  it is a second-order cone programming

simulation of robust LS: $\bar{A} \in \mathbf{R}^{20\times 5}$ and $e = 0.1$ (used for RLS estimation)



- compare robust LS (RLS) with LS using $\bar{A}, \bar{A}-\delta, \bar{A}+\delta$ for $\delta = 0.01$
- compute worst norm $\|\bar{A}x - y\|_2 + e\|x\|_2$ as e varies using x from various methods
- (top) worst-norms of random solution x are high and widely spread
- (bottom) worst-norms of RLS are relatively low and are not sensitive to e

example 2: let $U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}$

uncertainty in A is prescribed as upper bounds of 2-norm of each columns in U

$$\mathcal{A} = \{\bar{A} + U \mid \|u_j\|_2 \leq a_j,\ j = 1, 2, \ldots, n\}$$

it can be shown that

$$\sup_{\|u_j\|_2 \leq a_j} \|\bar{A}x - y + Ux\|_2 = a^T|x| + \|\bar{A}x - y\|_2$$

where the supremum is attained when each column of U is selected as

$$u_j = \frac{c_j \mathbf{sign}(x_j)}{\|\bar{A}x - y\|_2} \cdot (\bar{A}x - y), \quad j = 1, 2, \ldots, n$$

- the robust LS can be cast as a second-order cone programming
- the term $a^T|x|$ can be viewed as a weighted ℓ_1-regularization

Worst-case Chebyshev approximation

setting: find $\sup_U \|(Ax - y\|_\infty$ where uncertainty in A is prescribed as upper bounds of ∞-norm of each columns in U

$$\mathcal{A} = \{\bar{A} + U \mid \|u_j\|_\infty \leq a_j,\ j = 1, 2, \ldots, n\}$$

it can be shown that

$$\sup_{\|u_j\|_\infty \leq a_j} \|\bar{A}x - y + Ux\|_\infty = a^T|x| + \|\bar{A}x - y\|_\infty$$

where the supremum is attained when

- let j be the index for which $\|w\|_\infty = |w_j|$
- for each column u_k, for $k = 1, \ldots, n$, set all entries as zero, except the jth as

$$(u_k)_j = \mathbf{sign}(x_k w_j) \cdot a_k = \begin{cases} a_k, & \text{if } x_k \text{ and } w_j \text{ has the same sign} \\ -a_k, & \text{otherwise} \end{cases}$$

Stochastic robust least-squares

when A is a random variable, so we can describe A as

$$A = \bar{A} + U,$$

where $\bar{A}$ is the average value of A and U is a random matrix

use the expected value of $\|Ax - y\|$ as the objective:

$$\underset{x}{\text{minimize}} \quad \mathbf{E}\|Ax - y\|_2^2$$

expanding the objective gives

$$\begin{aligned}\mathbf{E}\|Ax - y\|_2^2 &= (\bar{A}x - y)^T(\bar{A}x - y) + \mathbf{E}x^T U^T U x \\ &= \|\bar{A}x - y\|_2^2 + x^T P x\end{aligned}$$

where $P = \mathbf{E}[U^T U]$

this problem is equivalent to

$$\underset{x}{\text{minimize}} \quad \|\bar{A}x - y\|_2^2 + \|P^{1/2}x\|_2^2$$

with solution $x = (\bar{A}^T\bar{A} + P)^{-1}\bar{A}^T y$

- a form of a regularized least-squares
- balance making $\bar{A}x - y$ small with aiming to get a small x
  (so that the variation in Ax is small)
- Tikhonov regularization is a special case of robust least-squares:

  when U has zero mean and uncorrelated variables, *i.e.*, $\mathbf{E}[U^T U] = \delta I$

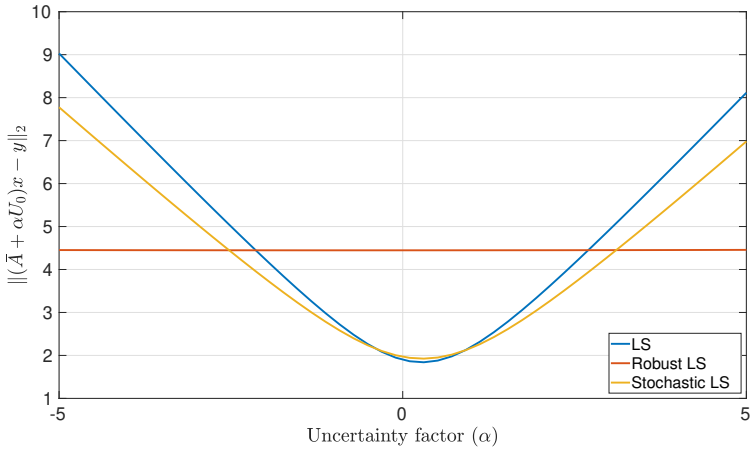
example: $u_{ij} \sim \mathcal{U}(-a_j, a_j)$ and assume columns of U are uncorrelated

$$P_{ij} = 0, \quad P_{ii} = \mathbf{E}[u_i^T u_i] = \mathbf{E}\left[\sum_{k=1}^{m} u_{ki}^2\right] = m\,\mathbf{var}[u_{ki}] = ma_j^2/3$$

Comparison between robust and stochastic LS

two comparable formulations

- robust LS: $A \in \mathcal{A} = \{\bar{A} + U \mid |u_{ij}| \leq a_j,\ j = 1, 2, \ldots, n\}$
- stochastic LS: $A = \bar{A} + U$ where $u_{ij} \sim \mathcal{U}(-a_j, a_j)$
- $\bar{A} \in \mathbf{R}^{20\times 15}$ and a_j ranges from 0.01 to 0.03



robust LS solution is most robust to uncertainty while LS solution is most sensitive to α and stochastic LS performance lies in between

Summary

- variants of least-squares problems are regarded as optimization problems with quadratic cost objective
- most of them are convex programs and can be solved by many existing algorithms
- regularized least-squares are proposed to promote a certain structure in the solutions

References

T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2009

Chapter 6 in
S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge press, 2004

G. Calafiore and L. El Ghaoui, Optimization Models, Cambridge University Press, 2014