

Model:

$$\begin{matrix} & \\ & \\ & \\ & \\ \text{y} & \end{matrix} = \begin{matrix} & & & \\ & z_1 & z_2 & \dots \\ & 43 & 2 & \\ & \vdots & \vdots & \\ & \text{A} & & \end{matrix} + \begin{matrix} & \\ & \beta_1 \\ & \beta_2 \\ & \vdots \\ & \text{B} \\ & \vdots \\ & \text{E} \\ & \vdots \\ & \text{C} \end{matrix}$$

$\beta_j = 0 \Rightarrow$ column A_j does not explain y .

measurement error

1 row = 1 subject

Problem: Recover β from measurements y .

minimize F $\|AB - y\|_2^2$

i.e. minimize the ℓ_2 norm of the error

Convex optim problem $\xrightarrow{\text{solution}}$ $F^{\text{ols}} = \underbrace{(A^T A)^{-1}}_{\text{exists?}} A^T y$

$$\text{rank}(A^T A) = \text{rank}(A) ?$$

$$\text{rank}(A^T A) = \text{rank}(N(A))$$

If $N(A) = N(\bar{A}^T A)$, $\text{rank}(A^T A) = \text{rank}(A)$

$$x \in N(A) \Rightarrow A^T A x = 0$$

$$\Rightarrow x \in N(A^T A)$$

$$\begin{aligned}
 x \in N(A^T A) &\Rightarrow A^T A x = 0 \\
 &\Rightarrow x^T A^T A x = 0 \\
 &\Rightarrow \|A x\|_2 = 0 \\
 &\Rightarrow x \in N(A)
 \end{aligned}$$

$$\beta^{\text{OLS}} = (A^T A)^{-1} A^T Y \quad \text{Exists?}$$

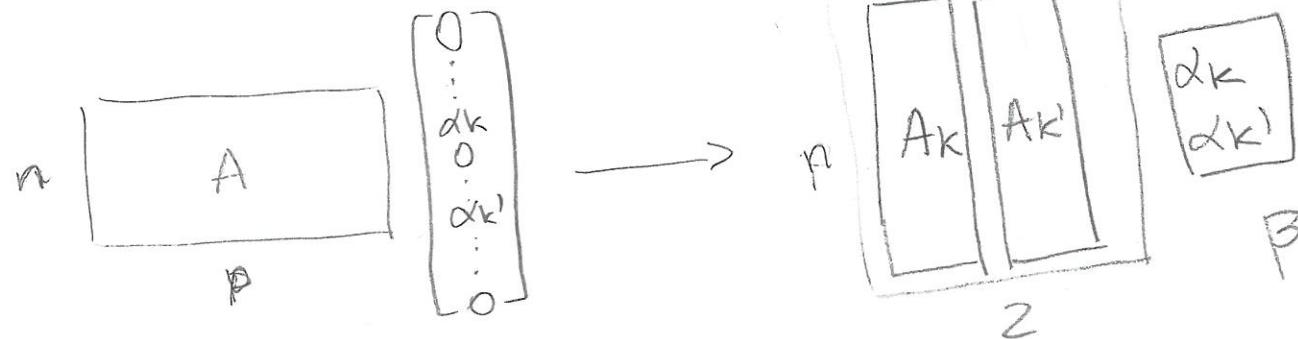
4

$$P > n \geq \text{rank}(A) = \text{rank}(A^T A)$$

$$\Rightarrow N(A^T A) \neq 304$$

$\Rightarrow A^T A$ is singular

Problem when $P > n$



\rightarrow Solve OLS problem.

We consider two options:

- ① minimize (w.r.t \mathbb{R}^n_+) $(\|A\beta - y\|_2^2, \|\beta\|_0)$

where $\|\beta\|_0 = |\{i : \beta_i \neq 0\}|$

→ LASSO selector

- ② Modify compressed sensing LP

→ Dantzig selector

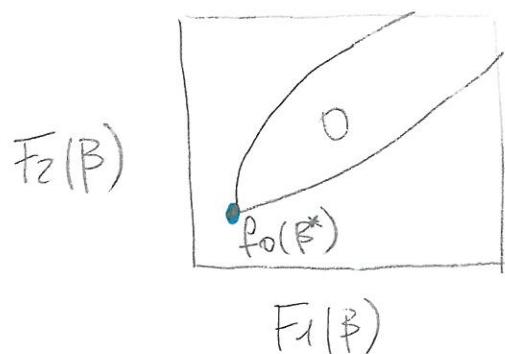
- minimize (w.r.t \mathbb{R}^n) $(\|AB - Y\|_2^2, \|\beta\|_1)$

$$f_0(\beta) = (F_1(\beta), F_2(\beta)) = (\|AB - Y\|_2^2, \|\beta\|_1)$$

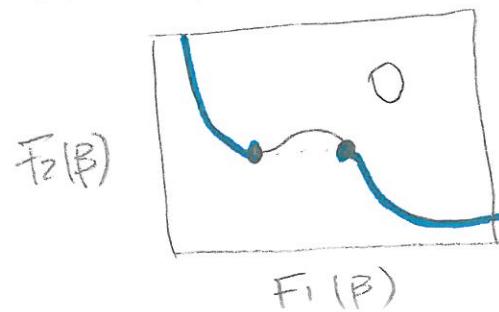
multi-criterium optimization: [Boyd]

$$O = \{ f(\beta) \mid \beta \text{ feasible} \}$$

- β^* is optimal if $f_0(\beta^*)$ is better than any other point

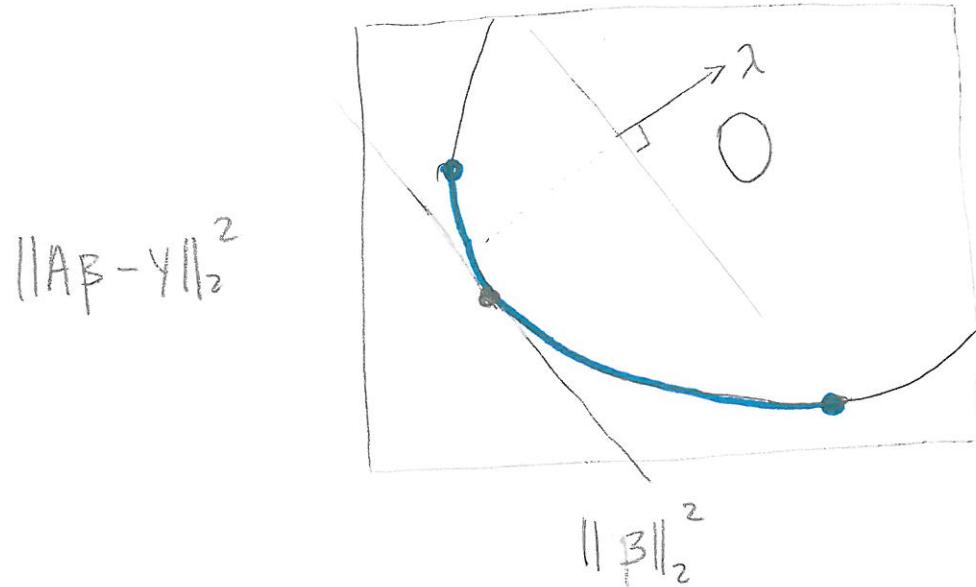


- β^{PO} is pareto optimal if no point is better.



Regularized Least-Squares

minimize (w.r.t. \mathbb{R}^2) $(\|AB - Y\|_2^2, \|\beta\|_2^2)$



Usually we set $\gamma = (1, \frac{\lambda_2}{\lambda_1})$ and solve

$$\text{minimize } \|AB - Y\|_2^2 + \gamma \|\beta\|_2^2 \quad (1)$$

Find a β^0 by solving

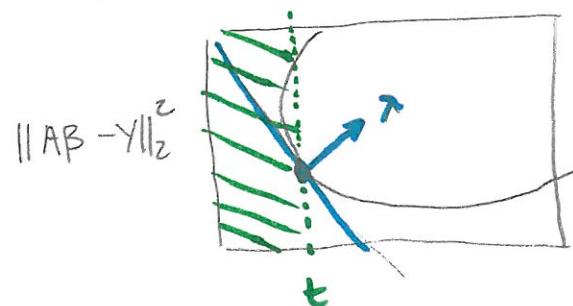
$$\text{minimize } \gamma_1 \|AB - Y\|_2^2 + \gamma_2 \|\beta\|_2^2$$

for a fixed $\lambda = (\lambda_1, \lambda_2)$

Remark: pb (1) is equivalent

to
 minimize $\|AB - Y\|_2^2$

subject to $\|\beta\|_2^2 \leq \epsilon$



Ideally we would like to solve

$$\text{minimize } \|A\beta - y\|_2^2 + \gamma \|\beta\|_0 \rightarrow (\text{NP hard})$$



8

Use the best convex approximation $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$

LASSO [R.Tibshirani, 1996]

$$\text{minimize } \|A\beta - y\|_2^2 + \gamma \|\beta\|_1 \quad \text{or equivalently}$$

$$\begin{aligned} & \text{minimize } \|A\beta - y\|_2^2 \\ & \text{subject to } \|\beta\|_1 \leq t. \end{aligned}$$

Does the ℓ_1 -norm induces sparsity? [Mairal PhD thesis, 2010]

not always

What is known:

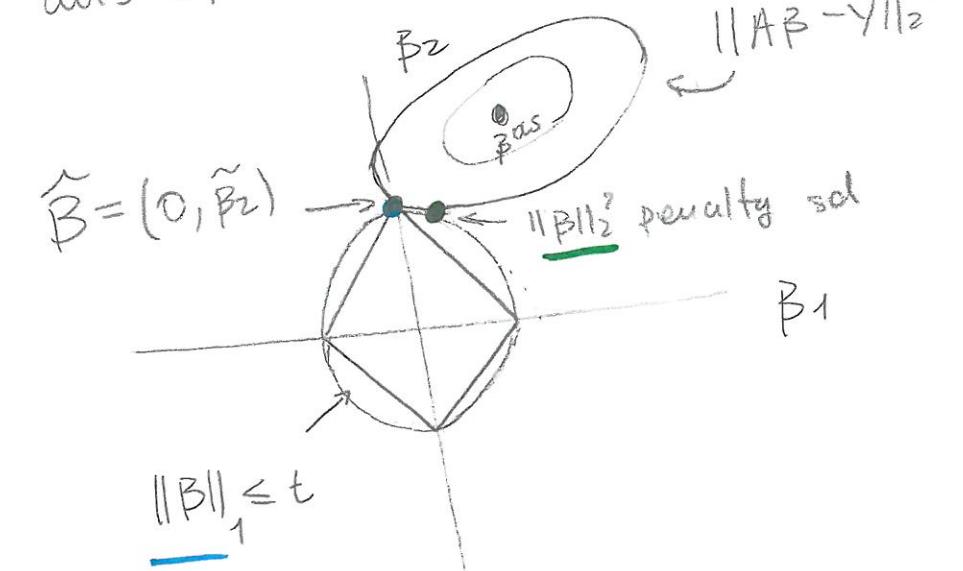
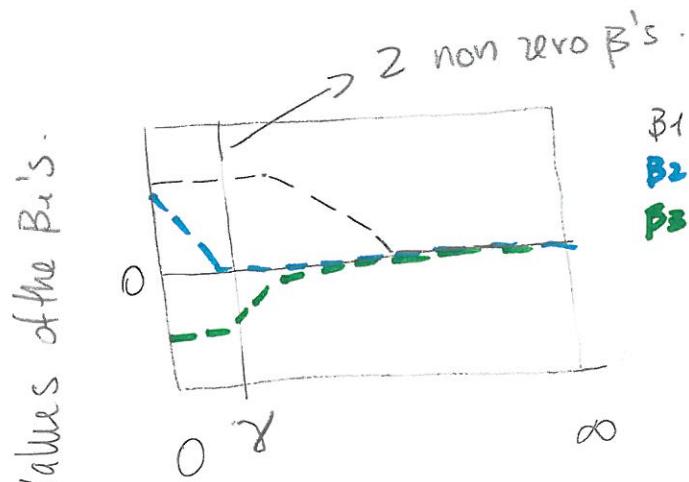
When $F_1(\beta) (= \|A\beta - y\|_2^2)$ is differentiable at 0, $\hat{\beta} = 0$ for large values of γ .

Between 0 and $+\infty$:

Empirically: l_1 induces sparsity (often)

Analytically: l_1 induced sparsity not clear.

Why does l_1 -norm can induce sparsity?



solution is more likely to be sparse - specially as γ grows.

Compressed Sensing

$$n \begin{array}{|c|} \hline A \\ \hline \end{array} \begin{array}{|c|} \hline \beta \\ \hline \end{array} = \begin{array}{|c|} \hline y \\ \hline \end{array} \quad (\text{no noise.})$$

P

minimize $\|\beta\|_1$

subject to $A\beta = y$.

• Noise \rightarrow relax the constraint:

$$\|\beta^*\|_1 = \sup_{i \in \mathcal{I}^c} |(A\beta - y)_i| < \lambda^\sigma \quad [\text{V. Rivoirard, 2011}]$$

Residuals within noise level.

The **Dantzig selector** is the solution of

minimize $\|\beta\|_1$

subject to $\|A^*(A\beta - y)\|_\infty \leq \lambda^\sigma$

why A^* ?

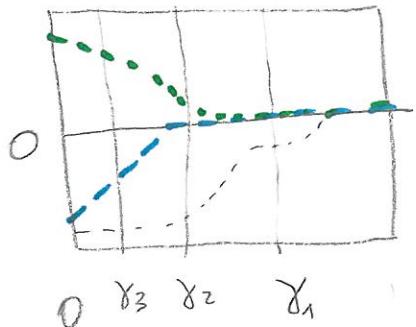
- relaxation of $A^*(A\beta^{\text{OLS}} - y) = 0$.
 $(\beta^{\text{OLS}} = (A^*A)^{-1}A^*y)$

- connection to LASSO [Karush-Kuhn-Tucker]
 1st order KKT cond:
 (optimality cond) $\|A^*(\hat{A}\beta - y)\|_\infty \leq \gamma$

By KKT conditions λ and γ are related.

11

LASSO tuning γ is important



γ_1 : only $B_1 \neq 0$

γ_2 : non zero B_1 and B_2

γ_3 : all B_i 's are non zero

Dantzig: How to choose λ ?

λ such that B (the real value) is feasible with high probability.

$$Y = AB + \varepsilon, \quad \varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

If $\|A\|_2 = 1$, $A_j^T \varepsilon = A_{1j}\varepsilon_1 + \dots + A_{nj}\varepsilon_n \sim N(0, \sigma^2)$ and

$$P \left(\max_{1 \leq j \leq p} \|A_j^T \varepsilon\| \leq \sigma \sqrt{2 \log p} \right) \geq 1 - \frac{1}{2\sqrt{\pi \log p}} \quad \xrightarrow[p \uparrow \infty]{} 1$$

$$\text{set } \lambda = \sigma \sqrt{2 \log p}$$

- S -restricted isometry constant of A

smallest δ_S such that for any S -sparse x

$$(1 - \delta_S) \|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_S) \|x\|_2^2$$

- S, S' restricted orthogonality constant of A

x, x' are S, S' -sparse. $\theta_{S,S'}$ smallest constant such that

$$|\langle Ax, Ax' \rangle| \leq \theta_{S,S'} \|x\|_2 \|x'\|_2$$

[Candès and Tao, 2005]

$$\theta_{S,S'} \leq \delta_{S+S'} \leq \theta_{S,S'} + \max\{\delta_S, \delta_{S'}\}.$$

Theorem [Candès and Tao, 2007]

β is S -sparse and $\delta_{2S} + \theta_{S,2S} < 1$.

with probability larger than $1 - \frac{1}{2\sqrt{\pi \log P}}$ the Dantzig selector $\hat{\beta}$ obeys

$$\|\hat{\beta} - \beta\|_2^2 \leq \underline{\underline{C_1}} \frac{(2 \log P) S \sigma^2}{\underline{\underline{S}}} \quad ; \quad C_1 = \frac{1}{1 - \delta_{2S} - \theta_{S,2S}}$$

How good is this estimation?

If we knew the support T^* of β in advance, then

$$\underline{\underline{\frac{1}{1 + \delta_S} S \sigma^2}} \leq E \|\hat{\beta}_{T^*}^{ds} - \beta\|_2^2 \leq \frac{1}{1 - \delta_S} S \sigma^2$$

Remarks:

- compare bound in probability with the Expected value of a risk.
- same up to a logarithmic factor.

Why $S_{2S} + \theta_{S,2S} < 1$?

Back to CS: $AB = Y$

If every $2S$ columns of A are linearly independent.

Then an S -sparse vector β can be reconstructed uniquely from AB .

Proof

If you can't: there are two vectors β and β' such that $AB = AB'$

$$AB = AB' \Rightarrow A(\beta - \beta') = 0 \quad (1)$$

Because $\beta - \beta'$ is $2S$ -sparse, A has $2S$ linearly dependent columns. \blacksquare

Theorem

β is S -sparse and $\delta_{2S} + \theta_{S,2S} < 1 - \epsilon$ ($\epsilon > 0$),

with probability larger than $1 - \frac{1}{2\sqrt{2\pi}\log P}$ the Dantzig selector $\hat{\beta}$ obeys

$$\|\hat{\beta} - \beta\|_{\ell_2}^2 \leq C_2^2 \lambda_P^2 \left(\sigma^2 + \sum_{i=1}^P \min(\beta_i^2, \sigma^2) \right)$$

$$\sum_{i=1}^P \min(\beta_i^2, \sigma^2) = E\|\beta^* - \beta\|_{\ell_2}^2 \quad \text{where} \quad \beta_i^* = \begin{cases} y_i & \text{if } |\beta_i| > \sigma, \\ 0 & \text{otherwise} \end{cases}$$

Asymptotic Analysis of l_0 - l_1 equivalence

[Donoho, 2011, 2013].

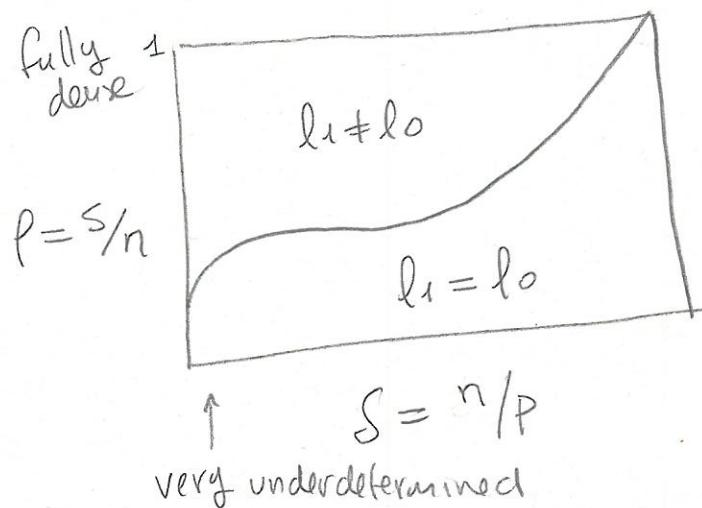
- S, n, P

- Proportional Growth $S \sim P^n, n \sim \delta P$ as $P \rightarrow \infty$.

- Phase diagram $(\delta, p) \in [0, 1]^2$
↑ how underdetermined the pb is
how dense the vector is.

Phase transition

$p_{CG}(\delta, \pm)$ function derived from combinatorial geometry



$$p < p_{CG}(\delta, \pm) ; P \setminus B_0 = \beta_1 \rightarrow 1 \text{ as } P \rightarrow \infty$$

$$p > p_{CG}(\delta, \pm) ; P \setminus B_0 = \beta_1 \rightarrow 0 \text{ as } P \rightarrow \infty$$

- For random A , if there is a sufficient (strictly) sparse solution
 ℓ_1 gives lo solution
- Precise tradeoff between required sparsity and allowed degree of undersampling

Strict sparsity is too strong.

ℓ_p balls as weak sparsity constraints.

- (weak) sparsity: ℓ_q norms, $q \leq 1$, are sparsity measures-

$$\|\beta - \beta_k\|_2 \leq c \|\beta\|_q K^{1/2 - 1/q}$$

Motivation (?) $f \in BV[0,1]$: wavelet coefficients are weak- $\ell_{1/2}$.

:

Compressed Sensing over ℓ_q balls.

- $\beta_0 \in \ell_q$ ball
- A Gaussian iid entries
- $y = AB$ measurements.

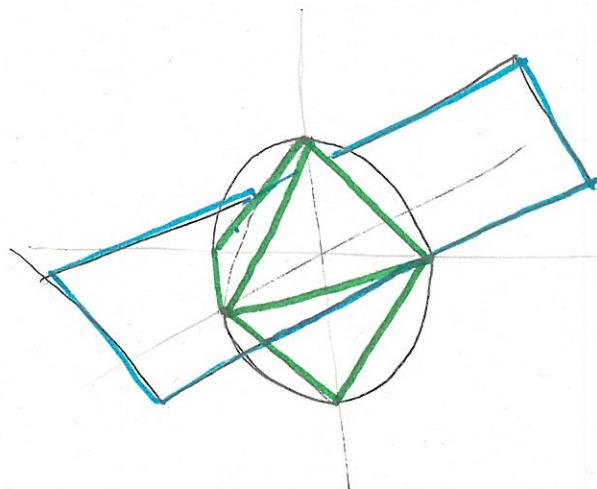
Geometry: • β_0 is a point in a ℓ_q ball: $B_{q,P}$

• Knowledge of y gives us a codimension- n section: $H_{y,P-n}$

• Remaining Uncertainty about β_0 is $\text{diam}(H_{y,P-n} \cap B_{q,P})$

Documented using Gelfand widths

$$d^n(B_{2,P} | l_2) = \sup_y \text{diam}(H_{y,P-n} \cap B_{q,P})$$



$$\begin{aligned} d^n(B_{2,P} | l_2) &= 1 \gg \sqrt{\frac{\log(P/n)}{n}} \\ &\gg d^n(B_{1,P} | l_2) \end{aligned}$$

- Kashin, 1977:

$$d^n(B_{1,P} \| l_2) \leq c (1 + \log(P/n))^{3/2} n^{-1/2}$$

:

- Geometric functional analysis [late 80's]

$$d^n(B_{q,P} \| l_2) \leq c (1 + \log(P/n))^{1/q - 1/2} n^{-1/q}$$

l_1 -minimization

$$\|B_1 - B_0\|_2 \leq c \underbrace{(1 + \log(P/n))^{1/q - 1/2}}_{\text{widths}} n^{-1/q}$$

you could control this bound by changing the number of measurements n .

Donoho Comment on Compressed Sensing:

Theory:

- A lot of progress
 - Emphasis on harmonic analysis and functional analysis. (RIP)
 - You get qualitative bounds and order estimates
no constants.
- Need more precise.

Algorithms:

- Known algorithms applied to the problem.
 - But why?
- Needed: algorithms derived from CS

How to do this?

Minimax decision theory [30's, 40's]

a game:

- Two players: Nature vs Researcher



Forced to have
a sparse object
(constraints)



Forced to pick
small number of
observations.

Basic Results

- Y_1, \dots, Y_n iid $N(\mu, \sigma^2)$

- $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$

After Game theory result by Von Neumann:
↳ Statisticians knew that \bar{Y}_n was an ok
estimate for the mean. We didn't
have a good way to articulate.

$$\min_T \max_{\mu} E(T(Y_1, \dots, Y_n) - \mu)^2$$

↑ nature can choose any mean

If a statistician could choose any measurable procedure of the data
he should play \bar{Y}_n

First it was applied to known estimators

Then researchers began to come up with estimators
from a minimax decision problem.

Bounded Normal Mean Problem

[Early 1980's]

$$\begin{array}{l} \cdot Y \sim N(\mu, 1) \\ \cdot |\mu| \leq \xi \end{array} \quad \left. \right\} \rightarrow M^*(\xi; \infty) = \min_{T} \max_{|\mu| \leq \xi} E(T(Y) - \mu)^2$$

solution is not \bar{Y}_n . It is
 • nonlinear and nonobvious.

Thresholding

Restrict the problem to decision

$$n_\theta(y) = (|y| - \theta)_+ \operatorname{sign}(y) \rightarrow \text{minimax threshold risk}$$

$$\min_{\theta} \max_{F, \epsilon} E_F(T(Y) - X)$$

$$Y \sim N(X, 1) \text{ and } X \sim F \text{ with } \mathcal{F}_{\theta, \epsilon} := \{F : P_F\{X \neq \theta\} \leq \epsilon\}, \quad 0 \leq \epsilon \leq 1$$

What do you get from Minimax Decision theory?

$$\bullet f_{C_0}(\delta) = M^{-1}(s)/s \quad ; \quad \text{where } M(\varepsilon) := M(\varepsilon; \theta)$$

ε -sparse normal mean.

For the LASSO

$$\min_B \sum_i (y_i - (AB)_i)^2 + \lambda \sum_i |x_{ii}|$$

$$\bullet \text{performance } MSE(\hat{\beta}, \beta) = \frac{1}{p} \mathbb{E} \|\hat{\beta} - \beta\|_2^2$$

$n/p \rightarrow \infty$ as $n, p \rightarrow \infty$ you get

$$\text{A.MSE}(\delta; F, \lambda) = \lim_{p \rightarrow \infty} MSE(\hat{\beta}_{1,\lambda}, \beta) \quad \text{a formula}$$

\downarrow \uparrow
penalization
undersampling

Using the formula:

$$\min_{\lambda} \max_{F \in \mathcal{F}_{\varepsilon, q}} \text{AMSE}(\delta, F, \lambda) = \frac{\varepsilon \delta}{M^{-1}(\delta)}$$

\circ p -th moment sparsity:

$$E_{\mathcal{F}_N} |X|^q \leq \varepsilon^q$$

This formula can tell the evolution of the λ AMSE in function of δ .

Noisy observations

control of the effect of the noise.

Other properties:

AMSE as $\delta \rightarrow 0$, ...

- The Dantzig Selector : Statistical estimation when $P \gg n$.
Candès and Tao ; 2007.

Phase Transition

- Observed Universality of Phase Transition in
High dimensional Geometry, ...

[DoTa Universality]
By Donoho and Tanner
2009.

- The Gel'Fand widths of ℓ_p -balls for $0 < p \leq 1$. By Foucart et al.
2010

AMP

PHD thesis of Arian Maleki.

LASSO

- Regression Shrinkage and Selection via the LASSO. By R.Tibshirani ; 1996
- The LASSO risk for Gaussian Matrices. By Bayati and Montanari 2012

(Computer Science) Numerical methods

- Toward a Unified Theory of Sparse Dimensionality reduction in Euclidean space. By J.Bourgain and J. Nelson ; 2014
- PhD J. mairal ; 2011 → about structured sparsity.