Model:

\[
\begin{align*}
0.31 & \\
1 & \\
\vdots & \\
\end{align*}
\]

\[
= \begin{bmatrix}
0.4 & 0.2 \\
0.6 & 0.4 \\
\vdots & \vdots \\
\end{bmatrix}
\begin{bmatrix}
B_1 \\
B_2 \\
\vdots \\
\end{bmatrix}
\begin{bmatrix}
E_1 \\
E_2 \\
\vdots \\
\end{bmatrix}
\]

\[
R_j = 0 \implies \text{column } A_j \text{ does not explain } Y.
\]

1 row = 1 subject.

Problem

Recover \( B \) from measurements \( Y \).

\[
\minimize_{B} \| A B - Y \|_2^2
\]

i.e. minimize the norm of the error.

Convex optimization problem \( \xrightarrow{\text{solution}} \) \( P_{\text{LS}} = (A^T A)^{-1} A Y \).

\( \exists \)?
\[ A = \begin{bmatrix} A_1 & A_2 & \cdots \end{bmatrix} \]

\[ n - \text{subjects (observations)} \]

\[ p - \text{variables} \]

\[ \text{Rank}(A^TA) = \text{rank}(A)? \]

\[ \text{rank}(A) + \text{dim}(\mathcal{N}(A)) = p = \text{rank}(A^TA) + \text{dim}(\mathcal{N}(A^TA)) \]

If \( \mathcal{N}(A) = \mathcal{N}(A^TA) \), \( \text{rank}(A^TA) = \text{rank}(A) \)

\[ x \in \mathcal{N}(A) \Rightarrow A^TAx = 0 \]

\[ \Rightarrow x \in \mathcal{N}(A^TA) \]

\[ \Rightarrow A^TAx = 0 \]

\[ \Rightarrow x^TA^TAx = 0 \]

\[ \Rightarrow \|Ax\|_2 = 0 \]

\[ \Rightarrow x \in \mathcal{N}(A) \]
\[ B_{ols} = (A^T A)^{-1} A Y \]  Exists?

\[ p > n \quad \Rightarrow \quad \text{rank}(A) = \text{rank}(A^T A) \]

\[ \Rightarrow N(A^T A) \neq 0 \]

\[ \Rightarrow A^T A \text{ is singular} \]

**Problem when \( p > n \)**

\[ \begin{pmatrix} 0 & \vdots & a_k & \cdots & a_k' \end{pmatrix} \]

\[ \begin{pmatrix} n \mid A \\ \mid p \end{pmatrix} \rightarrow \begin{pmatrix} \begin{pmatrix} A_k & A_k' \end{pmatrix} & \begin{pmatrix} a_k \ a_k' \end{pmatrix} \end{pmatrix} \]

\[ \Rightarrow \text{Solve OLS problem.} \]
We consider two options:

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

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

  $\rightarrow$ **Lasso selector**

- Modify compressed sensing LP

  $\rightarrow$ **Dantzig selector**
minimize \((w.r.t \; \|x\|_2^2, \|y\|_2^2)\)

\[ f_0(\beta) = (f_1(\beta), f_2(\beta)) = (\|A\beta - y\|_2^2, \|b\|_1) \]

multi-criteria optimization. [Boyd]

\[ 0 = \nabla f(\beta) \mid \beta \text{ feasible} \]

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

\[ F_2(\beta) \]

\[ F_1(\beta) \]

- \(\beta^0\) is pareto optimal if no point is better.

\[ F_2(\beta) \]

\[ F_1(\beta) \]
Regularized Least-Squares

\[
\text{minimize (w.r.t. } \mathbb{R}_+^2) \quad (\| A \beta - y \|^2_2, \| \beta \|^2_2)
\]

Usually we set \( \lambda = (\lambda_1, \lambda_2) \) and solve

\[
\text{minimize } \| A \beta - y \|^2_2 + \gamma \| \beta \|^2_2
\]

Find a \( \beta^0 \) by solving

\[
\text{minimize } \lambda_1 \| A \beta - y \|^2_2 + \lambda_2 \| \beta \|^2_2
\]

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

\[
\text{Remark: P6 (1) is equivalent to }
\]

\[
\text{minimize } \| A \beta - y \|^2_2
\]

subject to \( \| \beta \|^2_2 \leq \epsilon \)
Ideally we would like to solve

\[
\text{minimize } \| A \beta - y \|_2^2 + \lambda \| \beta \|_1 \rightarrow (N \geq \text{hard})
\]

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

\[\text{LASSO} \quad \text{[R. Tibshirani, 1996]}\]

\[
\text{minimize } \| A \beta - y \|_2^2 + \lambda \| \beta \|_1 \quad \text{or equivalently minimize } \| A \beta - y \|_2^2 \\
\text{subject to } \| \beta \|_1 \leq t.
\]

Does the \( l_1 \)-norm induce sparsity? \[\text{Mairal PhD thesis, 2010}\]

not always
what is known:

When $F_1(B) = \|AB - Y\|_2^2$ is differentiable at 0, $\hat{B} = 0$ for large values of $\lambda$.

Between 0 and $+\infty$:

- Empirically: $l_1$ induces sparsity (often)
- Analytically: $l_1$ induced sparsity not clear.

Why does $l_1$-norm can induce sparcity?

Solution is more likely to be sparse - specially as $p$ grows.
Compressed Sensing

\[ n \begin{bmatrix} A \\ P \end{bmatrix} \begin{bmatrix} \beta \\ p \end{bmatrix} = \begin{bmatrix} Y \end{bmatrix} \quad (\text{no noise}) \]

- minimize \( ||B||_1 \)
- subject to \( AB = Y \)

- Noise \( \rightarrow \) relax the constraint.
  - \( \| AB - Y \|_\infty = \sup_{i \neq j} | (AB - Y)_{ij} | < \lambda \sigma \quad \text{Residuals within noise level.} \)

The Dantzig selector is the solution of

- minimize \( ||B||_1 \)
- subject to \( \| A^*(AB - Y) \|_\infty \leq \lambda \sigma \)

- Why \( A^* \)?
  - relaxation of \( A^* (A \beta_{OLS} - Y) = 0 \). 
    \( \beta_{OLS} = (A^*A)^{-1} A^* Y \)
  - connection to LASSO [Karush-Kuhn-Tucker]

1st order KKT cond:
- optimality cond: \( \| A^*(A \hat{\beta} - Y) \|_2 \leq \lambda \)

[ V. Rivoirard, 2011 ]
By KKT conditions, $x$ and $y$ are related.

**LASSO tuning $\lambda$ is important**

\[ y_1: \text{only } B_1 \neq 0 \]
\[ y_2: \text{non zero } B_1 \text{ and } B_2 \]
\[ y_3: \text{all } B_i's \text{ are non zero} \]

**Dantzig:** How to choose $\lambda$?

$\lambda$ such that $B$ (the real value) is feasible with high probability.

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

If $\|A\|_2 = 1$, $A^T \varepsilon = A_{1d} \varepsilon_1 + \ldots + A_{kd} \varepsilon_k \sim N(0, \sigma^2)$ and

\[
P \left( \max_{1 \leq d \leq p} \|A_d^T \varepsilon\| \leq \sigma \sqrt{2 \log p} \right) \geq 1 - \frac{1}{2 \sqrt{11 \log p}}
\]

set $\lambda = \sigma \sqrt{2 \log p}$
Theoretical results.

- **S-restricted isometry constant** of A
  smallest \( s \) 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

  \[
  | \langle Ax, Ax' \rangle | \leq \Theta_{S, S'} \| x \|_2 \| x' \|_2
  \]

  [Candes and Tao, 2005]

  \[
  \Theta_{S, S'} \leq \delta_{S+S'} \leq \Theta_{S, S'} + \max \{ \delta_s, \delta_{S'} \}
  \]
Theorem [Candès and Tao, 2007]

\[ B \text{ is } S\text{-sparse and } 2s + \Theta s, 2s < 1. \]

With probability larger than \( 1 - \frac{1}{2 \sqrt{\pi \log p}} \) the Dantzig selector \( \hat{B} \) obeys

\[ \| \hat{B} - B \|_2^2 \leq C_1 (2 \log p) S_0^2 \]

\[ \text{where } C_1 = \frac{1}{1 - s} \]

How good is this estimation?

If we knew the support \( T^* \) of \( B \) in advance, then

\[ \frac{1}{1 + S_0} S_0^2 \leq \mathbb{E} \| \beta_{T^*} - \beta \|_2^2 \leq \frac{1}{1 - S_0} S_0^2 \]

Remarks:
- Compare bound in probability with the Expected value at risk.
- Some up to a logarithmic factor.
Why $S_2S + g_{S,2S} < 1$?

Back to CS: $AB = Y$

If every $2S$ columns of $A$ are linearly independent, then an $S$-sparse vector $B$ can be reconstructed uniquely from $AB$.

**Proof**

If there are two vectors $B$ and $B'$ such that $AB = AB'$,

$$A\hat{B} = A\hat{B}' \Rightarrow A(\hat{B} - \hat{B}') = 0 \quad (1)$$

Because $\hat{B} - \hat{B}'$ is $2S$-sparse, $A$ has $2S$ linearly dependent columns.
Theorem

$\beta$ is $s$-sparse and $s_{z^c} + \theta s_{z^c} < 1 - t$ ($t > 0$).

With probability larger than $1 - \frac{1}{2 \sqrt{2 \pi \log p}}$ the Dantzig selector $\beta$ obeys

$$\|\hat{\beta} - \beta\|_2^2 \leq C_2^2 \lambda P \left( \sigma^2 + \sum_{i=1}^{p} \min \{ \beta_i^2, \sigma^2 \} \right)$$

$$P \min_{i=1} \{ \beta_i^2, \sigma^2 \} = \mathbb{E} \| \hat{\beta}^* - \beta \|_2^2$$

where

$$\beta_i^* = \begin{cases} y_i & \text{if } |\beta_i| > \sigma, \\ 0 & \text{otherwise} \end{cases}$$
Asymptotic Analysis of $l_0 - l_1$ equivalence

- $S, n, p$
- Proportional Growth $S \sim pn$, $n \sim Sp$ as $p \to \infty$.
- Phase diagram $(S, p) \in [0, 1]^2$
  
  $\uparrow$ how underdetermined the problem is
  
  $\downarrow$ how dense the vector is.

Phase transition:

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

\[ p < p_{CG}(S, \pm) \quad ; \quad P \prod_{0} = B_{1 \frac{1}{4}} \to 1 \quad \text{as} \quad p \to \infty \]

\[ p > p_{CG}(S, \pm) \quad ; \quad P \prod_{0} = B_{1 \frac{1}{4}} \to 0 \quad \text{as} \quad p \to \infty \]

$\frac{S}{n}$

\[ S = \frac{n}{p} \]

\[ \uparrow \quad \text{very underdetermined} \]
• For random $A$, if there is a sufficient (strictly) sparse solution $\ell_1$ gives $\ell_0$ solution

• Precise tradeoff between required sparsity and allowed degree of undersampling

Strict Sparsity is too strong.

$\ell_p$ balls as weak sparsity constraints.

• (Weak) sparsity: $\ell_q$ norms, $q \leq 1$, are sparsity measures.

$$\|\beta - B_k\|_2 \leq c \|\beta\|_q k^{\frac{1}{2} - \frac{1}{q}}$$

Motivation (?) $f \in BV [0,1]$; wavelet coefficients are weak-$\ell_{1/2}$. 
Compressed Sensing over $l_q$ balls.

- $B_0 \in l_q$ ball
- A Gaussian i.i.d entries
- $Y = AB$ measurements.

**Geometry:**
- $B_0$ 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 $B_0$ is $\text{diam} (H_Y, p-n \cap B_{q,p})$

Documented using Gelfand widths

$$d^n(B_2, p l_2) = \sup_Y \text{diam} (H_Y, p-n \cap B_{q,p})$$

$$d^n(B_2, p l_2) \approx 1 \gg \sqrt{\frac{\log (P/n)}{n}}$$
$$\gg d^n(B_1, p l_2)$$
• Kashin, 1977:

\[ d^n(B_1, p \| l_2) \leq C (1 + \log (p/n))^{3} n^{-1/2} \]

• Geometric functional analysis [late 80's]

\[ d^n(B_q, p \| l_2) \leq C (1 + \log (p/n))^{1/4 - 1/2} n^{-1/4} \]

\[ \| B_1 - B_{\text{ball}} \|_2 \leq C (1 + \log (p/n))^{1/4 - 1/2} n^{-1/4} \]

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.
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
- $Y_1, \ldots, Y_n \text{ iid } N(\mu, \sigma^2)$
- $\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$

After game theory result by von Neumann, statisticians knew that $\overline{Y}_n$ was an ok estimate for the mean. We didn't have a good way to articulate.
\[ \min_{T} \max_{\mu} \mathbb{E} \left( T(Y_1, \ldots, Y_n) - \mu \right)^2 \]

\[ \uparrow \quad \text{nature can choose any mean} \]

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]

- \( Y \sim N(\mu, 1) \)
- \( |\mu| \leq \xi \)

\( \xi \rightarrow \mathcal{M}^*(\xi; \infty) = \min_{T} \max_{|\mu| \leq \xi} E( (T(Y) - \mu)^2 \)

Solution is not \( \bar{Y} \). It is
- non-linear and non-obvious.

---

**Thresholding**

Restrict the problem to decision

\( p_0(y) = (|y| - \theta)_+ \cdot \text{sign}(y) \rightarrow \min \max \text{ threshold risk} \)

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

\( Y \sim N(x, 1) \) and \( X \sim F \) with \( F_{\theta, \epsilon} := \{ F : P_F (x \neq 0 \mid x \geq 3 \epsilon) \leq 0 \} \), \( 0 \leq \epsilon \leq 1 \)
What do you get from Minimax Decision theory?

- \( p_c(\delta) = M^{-1}(\delta) / \delta \) \quad \text{where} \quad M(\varepsilon) := M(\varepsilon; \theta) \) 
  \( \varepsilon \)-sparse normal mean.

For the LASSO

- \( \min_{\beta} \operatorname{Av}_{i} (y_i - (A\beta)_i)^2 + \lambda \operatorname{Av}_{i} |x_i| \) 

- Performance \( \operatorname{MSE}(\hat{\beta}, \beta) = \frac{1}{p} \operatorname{E} \| \hat{\beta} - \beta \|_2^2 \)

- \( n/p \to 8 \) as \( n, p \to \infty \) you get
  - out of \( \hat{\beta} \) penalization
  \[ \operatorname{AMSE}(\delta; F, \gamma) = \lim_{p \to \infty} \operatorname{MSE}(\hat{\beta}_i, \lambda; \beta) \]
  \( \uparrow \) undersampling
  \( \downarrow \)
Using the formula:

$$\min \max AMSE(\delta, F, \lambda) = \frac{\delta S}{M^{-1}(S)}$$

- Optimal moment sparsity:
  $$E_{F_n, l \times 1} \leq \delta \lambda$$

- This formula can tell the evolution of the AMSE in function of $S$.

**Noisy Observations**

- Control of the effect of the noise.

**Other Properties:**

$$AMSE as \delta \to 0$$
The Dantzig Selector: Statistical estimation when $P \gg n$.
Candes and Tao; 2007.

Phase Transition

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

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

AMP

PhD thesis of Arjan Maleki.
Lasso

- Regression Shrinkage and Selection via the Lasso. By R. Tibshirani; 1996

- The LASSO risk for Gaussian matrices. By Bayati and Montanari; 2012

(Co)Computational Science) Numerical methods


- PhD J. Mairal; 2011 – about structured sparsity.