Compressed sensing, sparsity and p-values

Sara van de Geer

April 16, 2015

Basis Pursuit

[Chen, Donoho and Saunders (1998)]

X: given $n \times p$ (sensing) matrix and

f^0: given n-vector of measurements.

We know $f^0 = X\beta^0$.

We want to recover $\beta^0 \in \mathbb{R}^p$.

There are n equations and p unknowns.

High-dimensional case: $p \gg n$.

Notation *The ℓ_1-norm is*

$$\|\beta\|_1 := \sum_{j=1}^{p} |\beta_j|, \ \beta \in \mathbb{R}^p.$$

Basis pursuit solution $\beta^* := \arg\min\{\|\beta\|_1 : \ X\beta = f^0\}$.

Let $S \subset \{1, \ldots, p\}$.

Notation

$$\beta_S := \{\beta_j \mathrm{l}\{j \in S\}\}, \; \beta_{-S} := \beta_{S^c} = \beta - \beta_S.$$

$$\beta_S = \begin{array}{cc} \begin{pmatrix} \beta_1 \\ \vdots \\ 0 \\ \beta_j \\ \vdots \\ 0 \end{pmatrix} & \begin{array}{l} \leftarrow 1 \in S \\ \vdots \\ \leftarrow j-1 \notin S \\ \leftarrow j \in S \\ \vdots \\ \leftarrow p \notin S \end{array} \end{array}, \quad \beta_{-S} = \begin{pmatrix} 0 \\ \vdots \\ \beta_{j-1} \\ 0 \\ \vdots \\ \beta_p \end{pmatrix}$$

Definition

The matrix X satisfies the null-space property at S if for all $\beta \neq 0$ in $\mathrm{null}(X)$ it holds that $\|\beta_{-S}\|_1 > \|\beta_S\|_1$.

Basis pursuit solution

$$\beta^* := \arg\min\{\|\beta\|_1 : \ X\beta = f^0\}.$$

Let $S_0 := \{j : \ \beta_j^0 \neq 0\}$ be the active set of β^0.

Loose definition *The vector β^0 is called sparse if S_0 is small.*

Theorem

Suppose X has the null-space property at S_0.
Then we have exact recovery:

$$\beta^* = \beta^0.$$

Proof.

Suppose $\beta^* \neq \beta^0$. Since $X\beta^* = X\beta^0 = f^0$ we have $\beta^* - \beta^0 \in \mathrm{null}(X)$. By the null-space property

$$\|\beta^*_{-S_0}\|_1 > \|\beta^*_{S_0} - \beta^0\|_1.$$

Since β^* minimizes $\|\cdot\|_1$ we have

$$\|\beta^*\|_1 \leq \|\beta^0\|_1.$$

We can decompose the ℓ_1-norm as

$$\|\beta^*\|_1 = \|\beta^*_{S_0}\|_1 + \|\beta^*_{-S_0}\|_1.$$

Hence

$$\|\beta^*_{S_0}\|_1 + \|\beta^*_{-S_0}\|_1 \leq \|\beta^0\|_1.$$

But then by the triangle inequality

$$\|\beta^*_{-S_0}\|_1 \leq \|\beta^*_{S_0} - \beta^0\|_1.$$

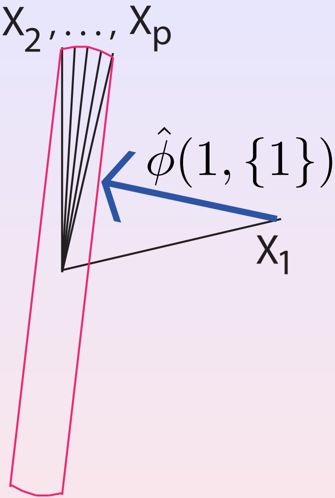
Thus we arrived at a **contradiction**. $\square$

Definition [vdG (2007)]
The compatibility constant for the set S and the stretching constant $L > 0$ is

$$\hat{\phi}^2(L, S) = \min\left\{ \frac{|S|}{n} \|X\beta_S - X\beta_{-S}\|_2^2 : \ \|\beta_{-S}\|_1 \le L, \ \|\beta_S\|_1 = 1 \right\}.$$

We have:

$$X \text{ satisfies the null-space property at } \ S \Leftrightarrow \hat{\phi}(1, S) > 0.$$



The compatibility constant $\hat{\phi}(1, S)$ for the case $S = \{1\}$.

Regularized formulation

$$\beta_\lambda := \arg\min \Big\{ \|X\beta - f^0\|_2^2/n + 2\lambda\|\beta\|_1 \Big\}.$$

Lemma

We have

$$\|X(\beta_\lambda - \beta^0)\|_2^2/n \le \frac{\lambda^2 |S_0|}{\hat{\phi}^2(1, S_0)}.$$

Adding noise

Let

$$Y = f^0 + \epsilon$$

with ϵ unobservable noise.

Let β^0 be a solution of $f^0 = X\beta^0$.

Definition *The Lasso is*

$$\hat{\beta} := \hat{\beta}_\lambda := \arg\min_\beta \left\{ \|Y - X\beta\|_2^2/n + 2\lambda\|\beta\|_1 \right\}.$$

Theorem (prediction error of the Lasso) *Let*

$$\lambda_\epsilon \geq \|X^T \epsilon\|_\infty / n.$$

Take $\lambda > \lambda_\epsilon$*. Then for*

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \ \bar{\lambda} := \lambda + \lambda_\epsilon, \ L := \frac{\bar{\lambda}}{\underline{\lambda}}$$

we have

$$\|X(\hat{\beta} - \beta^0)\|_2^2 / n \leq \frac{\bar{\lambda}^2 |S_0|}{\hat{\phi}^2(L, S_0)}.$$

Note 1 $\|\cdot\|_\infty$ is the dual norm of $\|\cdot\|_1$.

Note 2 Suppose $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 I)$ and $\mathrm{diag}(X^T X)/n = I$.

Then

$$\mathbb{P}\left(\|X^T \epsilon\|_\infty / n \geq \sigma_0 \sqrt{\frac{2\log(2p/\alpha)}{n}} \right) \leq \alpha.$$

Note 3 Under compatibility conditions Lasso thus has prediction error

$$\|X(\hat{\beta} - \beta^0)\|_2^2 / n \sim \sigma_0^2 \log p \times \frac{|S_0|}{n}$$

$$= \sigma_0^2 \log p \times \frac{\text{number of active parameters}}{\text{number of observations}}.$$

= oracle inequality

= adaptation

Note 1 $\|\cdot\|_\infty$ is the dual norm of $\|\cdot\|_1$.

Note 2 Suppose $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 I)$ and $\mathrm{diag}(X^T X)/n = I$.

Then
$$\mathbb{P}\left(\|X^T\epsilon\|_\infty/n \geq \sigma_0\sqrt{\frac{2\log(2p/\alpha)}{n}}\right) \leq \alpha.$$

Note 3 Under compatibility conditions Lasso thus has prediction error

$$\|X(\hat\beta - \beta^0)\|_2^2/n \sim \sigma_0^2 \log p \times \frac{|S_0|}{n}$$

$$= \sigma_0^2 \log p \times \frac{\text{number of active parameters}}{\text{number of observations}}.$$

= oracle inequality

= adaptation

Note 1 $\|\cdot\|_\infty$ is the dual norm of $\|\cdot\|_1$.

Note 2 Suppose $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 I)$ and $\mathrm{diag}(X^T X)/n = I$.

Then

$$\mathbb{P}\left(\|X^T \epsilon\|_\infty / n \geq \sigma_0 \sqrt{\frac{2\log(2p/\alpha)}{n}} \right) \leq \alpha.$$

Note 3 Under compatibility conditions Lasso thus has prediction error

$$\|X(\hat\beta - \beta^0)\|_2^2/n \sim \sigma_0^2 \log p \times \frac{|S_0|}{n}$$

$$= \sigma_0^2 \log p \times \frac{\text{number of active parameters}}{\text{number of observations}}.$$

= oracle inequality

= adaptation

What if β^0 is only approximately sparse?

Theorem (trade-off approximation error and sparsity) *Let*

$$\lambda_\epsilon \ge \|X^T \epsilon\|_\infty / n.$$

Take $\lambda > \lambda_\epsilon$*. Then for*

$$\underline{\lambda} := \lambda - \lambda_\epsilon,\ \bar{\lambda} := \lambda + \lambda_\epsilon,\ L := \frac{\bar{\lambda}}{\underline{\lambda}}$$

we have for all β *and* S

$$\|X(\hat{\beta} - \beta^0)\|_2^2 / n \le \underbrace{\|X(\beta - \beta^0)\|_2^2 / n + 4\lambda \|\beta_{-S}\|_1}_{\textit{approximation error}} + \underbrace{\frac{\bar{\lambda}^2 |S|}{\hat{\phi}^2(L, S)}}_{\textit{“effective sparsity”}} .$$

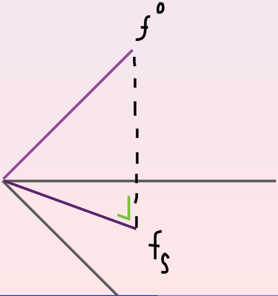
Corollary

Let $S \subset \{1, \ldots, p\}$ be arbitrary.
Let f_S be the projection of f^0 on the space spanned by $\{X_j\}_{j \in S}$.
Then

$$\| X(\hat{\beta} - \beta^0) \|_2^2 / n \le \| f_S - f^0 \|_2^2 / n + \frac{\bar{\lambda}^2 |S|}{\hat{\phi}^2(L, S)}.$$

So

$$\| X(\hat{\beta} - \beta^0) \|_2^2 / n \le \min_S \left\{ \| f_S - f^0 \|_2^2 / n + \frac{\bar{\lambda}^2 |S|}{\hat{\phi}^2(L, S)} \right\}.$$



What about the ℓ_1-estimation error?

Theorem(including the ℓ_1-error) *Let*

$$\lambda_\epsilon \geq \|X^T \epsilon\|_\infty / n.$$

Take $\lambda > \lambda_\epsilon$. Then for

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \ \bar{\lambda} := \lambda + \lambda_\epsilon + \delta\underline{\lambda}, \ L := \frac{\bar{\lambda}}{(1-\delta)\underline{\lambda}}$$

we have for all β and S

$$2\delta\underline{\lambda}\|\hat{\beta}-\beta\|_1 + \|X(\hat{\beta}-\beta^0)\|_2^2/n \leq \|X(\beta-\beta^0)\|_2^2/n + \frac{\bar{\lambda}^2|S|}{\hat{\phi}^2(L,S)} + 4\lambda\|\beta_{-S}\|_1.$$

Corollary (weak sparsity)
Let

$$\rho_r^r := \sum_{j=1}^{p} |\beta_j^0|^r, \ 0 < r < 1,$$

$$S_* := \{ j : \ |\beta_j^0| > 3\lambda_\epsilon \}.$$

We have (with $\delta = 1/5$, $\lambda = 2\lambda_\epsilon$)

$$\| \hat{\beta} - \beta^0 \|_1 \le 2^8 \lambda_\epsilon^{1-r} \frac{\rho_r^r}{\hat{\phi}^2(4, S_*)}.$$

Asymptopia
Suppose $1/\hat{\phi}^2(4, S_*) = \mathcal{O}(1)$.
Let $\lambda_\epsilon \asymp \sqrt{\log p / n}$.
When $\rho_r^r = o((n/\log p)^{\frac{1-r}{2}})$ we have $\| \hat{\beta} - \beta^0 \|_1 = o_{\mathbb{P}}(1)$.

Question

What is so special about the ℓ_1-norm?
Why does it lead to exact recovery and oracle inequalities?

Answer

Its decomposability:

$$\|\beta\|_1 = \|\beta_S\|_1 + \|\beta_{-S}\|_1 .$$

Question

What is so special about the ℓ_1-norm?
Why does it lead to exact recovery and oracle inequalities?

Answer

Its decomposability:

$$\|\beta\|_1 = \|\beta_S\|_1 + \|\beta_{-S}\|_1 .$$

Question

What is so special about the ℓ_1-norm?
Why does it lead to exact recovery and oracle inequalities?

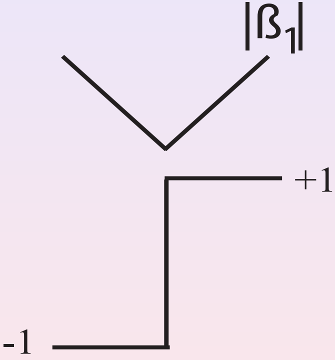
Answer

Its decomposability:

$$\|\beta\|_1 = \|\beta_S\|_1 + \|\beta_{-S}\|_1.$$

Definition *The sub-differential of* $\beta \mapsto \|\beta\|_1$ *is*

$$\partial\|\beta\|_1 = \{z : \ \|z\|_\infty = 1, \ z^T\beta = \|\beta\|_1\}.$$



subdifferential calculus

We invoke decomposability actually as the triangle property

$$\max_{z \in \partial \|\beta^0\|_1} z^T \beta \geq \|\beta_{-S_0}\|_1 - \|\beta_{S_0}\|_1 .$$

Other norms

Let Ω be a norm on $\mathbb{R}^p$.

Definition *The* *dual norm* *of* Ω *is*

$$\Omega_*(z) := \max_{\Omega(\beta)\le 1} z^T\beta, \ z \in \mathbb{R}^p.$$

Definition *The* *sub-differential* *of* $\beta \mapsto \Omega(\beta)$ *is*

$$\partial\Omega(\beta) := \{z : \ \Omega_*(z) = 1, \ z^T\beta = \Omega(\beta)\}.$$

Other norms

Let Ω be a norm on $\mathbb{R}^p$.

Definition *The dual norm of* Ω *is*

$$\Omega_*(z) := \max_{\Omega(\beta)\le 1} z^T\beta,\ z \in \mathbb{R}^p.$$

Definition *The sub-differential of* $\beta \mapsto \Omega(\beta)$ *is*

$$\partial\Omega(\beta) := \{z :\ \Omega_*(z) = 1,\ z^T\beta = \Omega(\beta)\}.$$

Definition *We say that Ω is* ***weakly decomposable*** *at β^0 if there exists semi-norms Ω^+ and Ω^- (depending on β^0) with $\Omega^-(\beta^0) = 0$ such that for all β*

$$\Omega(\beta) \geq \Omega^+(\beta) + \Omega^-(\beta).$$

Definition *We say that Ω satisfies the* ***triangle property*** *at β^0 if there exists semi-norms Ω^+ and Ω^- (depending on β^0) such that for all β*

$$\max_{z_0 \in \partial\Omega(\beta^0)} z^T(\beta - \beta^0) \geq \Omega^-(\beta) - \Omega^+(\beta - \beta^0) \quad .$$

Definition *We say that Ω is* **weakly decomposable** *at β^0 if there exists semi-norms Ω^+ and Ω^- (depending on β^0) with $\Omega^-(\beta^0) = 0$ such that for all β*

$$\Omega(\beta) \geq \Omega^+(\beta) + \Omega^-(\beta).$$

Definition *We say that Ω satisfies the* **triangle property** *at β^0 if there exists semi-norms Ω^+ and Ω^- (depending on β^0) such that for all β*

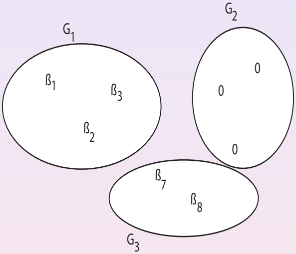
$$\max_{z_0 \in \partial\Omega(\beta^0)} z^T(\beta - \beta^0) \geq \Omega^-(\beta) - \Omega^+(\beta - \beta^0) \quad .$$

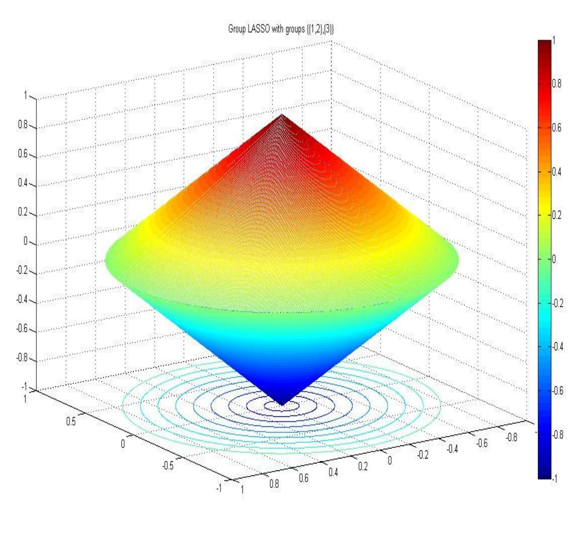
Example 1: group penalty

$$\Omega(\beta) := \sum_{k=1}^{m} \|\beta_{G_k}\|_2.$$

$$\Omega_*(z) = \max_k \|z_{G_k}\|_2.$$

Let $S_0 \subset \cup_{k \in T_0} G_k$.
Then
$\Omega^+(\beta) = \sum_{k \in T_0} \|\beta_{G_k}\|_2$,
$\Omega^-(\beta) = \sum_{k \notin T_0} \|\beta_{G_k}\|_2$





Unit ball of the group penalty

Norms generated from cones

Let $\mathcal{A} \subset \mathbb{R}_+^p$ be a convex cone and

$$\Omega(\beta) := \min_{a \in \mathcal{A},\ \|a\|_1 = 1} \sqrt{\sum_{j=1}^p \frac{\beta_j^2}{a_j}}$$

Then

$$\Omega_*(z) = \max_{a \in \mathcal{A},\ \|a\|_1 = 1} \sqrt{\sum_{j=1}^p a_j z_j^2}.$$

Suppose $a_{S_0} \in \mathcal{A}$ for all $a \in \mathcal{A}$.
Then Ω is weakly decomposable at β^0, with

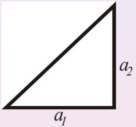
$$\Omega^+(\beta) = \min_{a_{S_0} \in \mathcal{A}_{S_0},\ \|a_{S_0}\|_1 = 1} \sqrt{\sum_{j \in S_0} \frac{\beta_j^2}{a_j}},$$

and

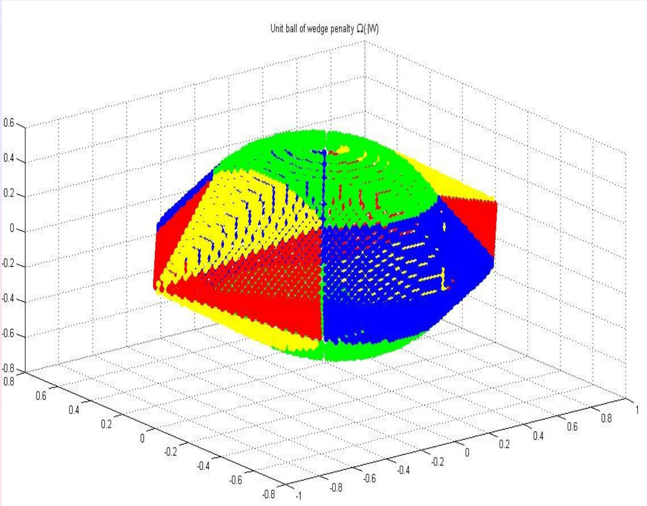
$$\Omega^-(\beta) = \min_{a_{-S_0} \in \mathcal{A}_{-S_0}, \|a_{-S_0}\|_1 = 1} \sqrt{\sum_{j \notin S_0} \frac{\beta_j^2}{a_j}}.$$

Example 2: wedge penalty

$$\mathcal{A} := \{a_1 \geq a_2 \geq \cdots\}$$



Then Ω is decomposable at $\beta^0 = (\beta_1^0, \cdots, \beta_{s_0}^0, 0, \cdots, 0)^T$.



Unit ball of the wedge penalty

Example 3: nuclear norm penalty

Let $\beta^0 = \text{vec}(B^0)$ and

$$\Omega(\beta) = \|B\|_{\text{nuclear}}.$$

Then

$$\Omega_*(z) = \Lambda_{\max}(Z),$$

where $\Lambda^2_{\max}(Z)$ is the largest eigenvalue of $Z^T Z$.
Write the SVD of B^0 as

$$B^0 = P_0 \Lambda^0 Q_0^T, \quad P_0^T P_0 = I, \ Q_0^T Q_0 = I, \ \Lambda^0 = \begin{pmatrix} \Lambda_1^0 & & \\ & \ddots & \\ & & \Lambda_{s_0}^0 \end{pmatrix}.$$

Then

$$\partial\Omega(\beta^0) = \{Z = P_0 Q_0^T + (I - P_0 P_0^T) W (I - Q_0 Q_0^T) : \ \Lambda_{\max}(W) \le 1\}.$$

We have the triangle property with

$$\Omega^+(B) = \|P_0 P_0^T B Q_0 Q_0^T\|_{\text{nuclear}}, \Omega^-(B) = \|(I - P_0 P_0^T) B (I - Q_0 Q_0^T)\|_{\text{nuclear}}.$$

Definition

Suppose Ω is weakly decomposable at β^0
- or alternatively has the triangle property at β^0 -
The effective sparsity with stretching constant $L > 0$ is

$$\hat{\Gamma}(L, \beta^0) := \left(\min \left\{ \|X\beta\|_2^2/n : \ \Omega^-(\beta) \leq L, \ \Omega^+(\beta) = 1 \right\} \right)^{-1}$$

.

Ω-basis pursuit

$$\beta^*_\Omega := \arg\min\{\Omega(\beta) :\ X\beta = f^0\}.$$

Lemma

Suppose Ω is weakly decomposable at β^0.

If $\Gamma(1, \beta^0) < \infty$ we have $\beta^*_\Omega = \beta^0$.

Ω-regularized formulation

$$\beta_{\Omega,\lambda} := \arg\min \left\{ \|X\beta - f^0\|_2^2/n + 2\lambda\Omega(\beta) \right\}.$$

Lemma

Suppose Ω is weakly decomposable at β^0
- or alternatively has the triangle property at β^0 -
Then

$$\|X(\beta_{\Omega,\lambda} - \beta^0)\|_2^2/n \leq \hat{\Gamma}(1,\beta^0)^2\lambda^2.$$

Adding noise leads the requirement $\lambda > \underline{\Omega}_*(X^T\epsilon)/n$ where $\underline{\Omega}_*$ is the dual norm of $\underline{\Omega} := \Omega^+ + \Omega^-$

For approximately decomposable β^0 we have sharp oracle inequalities

Increasing the stretching constant further leads to bounds for the $\underline{\Omega}$-estimation error.

everything as for the Lasso

[Adding noise] leads the requirement $\lambda > \underline{\Omega}_*(X^T\epsilon)/n$ where $\underline{\Omega}_*$ is the dual norm of $\underline{\Omega} := \Omega^+ + \Omega^-$

For approximately decomposable β^0 we have sharp oracle inequalities

Increasing the stretching constant further leads to bounds for the $\underline{\Omega}$-estimation error.

everything as for the Lasso

Adding noise leads the requirement $\lambda > \underline{\Omega}_*(X^T\epsilon)/n$ where $\underline{\Omega}_*$ is the dual norm of $\underline{\Omega} := \Omega^+ + \Omega^-$

For approximately decomposable β^0 we have sharp oracle inequalities

Increasing the stretching constant further leads to bounds for the $\underline{\Omega}$-estimation error.

everything as for the Lasso

Adding noise leads the requirement $\lambda > \underline{\Omega}_*(X^T\epsilon)/n$ where $\underline{\Omega}_*$ is the dual norm of $\underline{\Omega} := \Omega^+ + \Omega^-$

For approximately decomposable β^0 we have sharp oracle inequalities

Increasing the stretching constant further leads to bounds for the $\underline{\Omega}$-estimation error.

everything as for the Lasso

General loss and norms

Let $R_n(\beta)$, $\beta \in \mathbb{R}^p$ be some (observable) *empirical risk*.
Let $R(\beta)$, $\beta \in \mathbb{R}^p$ be (unobservable) *theoretical risk*.

We assume R_n and R to be differentiable w.r.t. β.
Denote their derivatives as $\dot{R}_n$ and $\dot{R}$.

Ω-penalized empirical risk minimizer

$$\hat{\beta} := \arg\min \left\{ R_n(\beta) + \lambda \Omega(\beta) \right\}.$$

Two point margin condition

There is a strictly convex function G with $G(0) = 0$ and a semi-norm τ on $\mathbb{R}^p$ such that for all β and β' we have

$$R(\beta) - R(\beta') \geq \dot{R}(\beta')^T(\beta - \beta') + G(\tau(\beta - \beta')).$$

Definition *The convex conjugate of G is*

$$H(v) = \sup_{u \geq 0} \left\{ uv - G(u) \right\}, \; v \geq 0.$$

Example

$$G(u) = u^2/2 \; \Rightarrow \; H(v) = v^2/2.$$

Two point margin condition

There is a strictly convex function G with $G(0) = 0$ and a semi-norm τ on $\mathbb{R}^p$ such that for all β and β' we have

$$R(\beta) - R(\beta') \geq \dot{R}(\beta')^T(\beta - \beta') + G(\tau(\beta - \beta')).$$

Definition *The convex conjugate of G is*

$$H(v) = \sup_{u \geq 0} \left\{ uv - G(u) \right\}, \ v \geq 0.$$

Example

$$G(u) = u^2/2 \ \Rightarrow \ H(v) = v^2/2.$$

Definition Let τ be a semi-norm, Ω be a norm and $L > 0$ a stretching constant. Assume Ω is weakly decomposable - or has the triangle property - at β. The effective sparsity at β is

$$\Gamma_\Omega(L, \beta, \tau) := \left(\min\{\tau(\beta') : \ \Omega_\beta^-(\beta') \leq L, \ \Omega_\beta^+(\beta') = 1\} \right)^{-1} .$$

Let be given some "target"

$$\beta = \beta^+ + \beta^-$$

with

1. Ω weakly decomposable - or having the triangle property - at β^+
2. with $\Omega^+_{\beta^+}(\beta^-) = 0$.

Let

$$\Omega^+ := \Omega^+_{\beta^+},\ \Omega^- := \Omega^-_{\beta^+}, \qquad \underline{\Omega} := \Omega^+ + \Omega^-.$$

Write the dual norm of $\underline{\Omega}$ as $\underline{\Omega}_*$.

Theorem (sharp oracle inequality) *Let*

$$\lambda_\epsilon \geq \underline{\Omega}_*(\dot{R}_n(\hat{\beta}) - \dot{R}(\hat{\beta})).$$

Take $\lambda > \lambda_\epsilon$ *and define*

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \bar{\lambda} := \lambda + \lambda_\epsilon + \delta\underline{\lambda},\ L := \frac{\bar{\lambda}}{(1-\delta)\underline{\lambda}}.$$

Then

$$\delta\underline{\lambda}\ \underline{\Omega}(\hat{\beta} - \beta) + R(\hat{\beta}) \leq R(\beta) + H(\bar{\lambda}\Gamma_\Omega(L, \beta, \tau)) + 2\lambda\Omega(\beta^-).$$

Example: matrix completion

Let

$$Y_i = \mathrm{trace}(X_i^T B^0) + \epsilon_i, \ i = 1, \ldots, n,$$

where $X_1, \ldots, X_n$ are i.i.d. $p \times q$ matrices with

$$\mathbb{P}(X_i = e_j e_k^T) = \frac{1}{pq} \ (i = 1, \ldots, n).$$

Let $\| \cdot \|_2$ be the Fobenius norm, and

$$R_n(B) := -pq \sum_{i=1}^n Y_i \mathrm{trace}(X_i^T B)/n + \frac{1}{2} \|B\|_2^2.$$

Let

$$R(B) := \mathbb{E} R_n(B) = -\mathrm{trace}(B^T B^0) + \frac{1}{2} \|B\|_2^2 = \frac{1}{2} \|B - B^0\|_2^2 - \frac{1}{2} \|B^0\|_2^2.$$

Then

$$\dot{R}(B) = (B - B^0).$$

Checking the two point margin condition

We have

$$R(B) - R(B') = \frac{1}{2}\|B - B^0\|_2^2 - \frac{1}{2}\|B' - B^0\|_2^2$$

$$= \frac{1}{2}\|B - B'\|_2^2 + \frac{1}{2}\|B' - B^0\|_2^2 + \mathrm{trace}((B - B')^T(B' - B^0)) - \frac{1}{2}\|B' - B^0\|_2^2$$

$$= \mathrm{trace}(\dot{R}(B')^T(B - B')) + \frac{1}{2}\|B - B'\|_2^2.$$

So we may take

$$\tau(B) := \|B\|_2, \; G(u) = u^2/2$$

Hence

$$H(v) = v^2/2.$$

We moreover find

$$\Gamma^2(L, B, \|\cdot\|_2) \leq \mathrm{rank}(B).$$

Let

$$W_{j,k} := \left(\sqrt{\frac{pq}{n}} \sum_{i=1}^{n} X_{i,j,k}\epsilon_i \right), \; 1 \leq j \leq p, \; 1 \leq k \leq q.$$

Theorem [Koltchinskii et al. (2011)] *Let*

$$\lambda_\epsilon \geq \Lambda_{\max}(W).$$

Take $\lambda > \lambda_\epsilon$ *and define*

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \bar{\lambda} := \lambda + \lambda_\epsilon + \delta\underline{\lambda},\ L := \frac{\bar{\lambda}}{(1-\delta)\underline{\lambda}}.$$

Then

$$\delta\underline{\lambda}\|\hat{B}-B\|_{\text{nuclear}}+\frac{1}{2}\|\hat{B}-B^0\|_2^2 \leq \frac{1}{2}\|B-B^0\|_2^2+\bar{\lambda}^2\text{rank}(B^+)+2\lambda\|B^-\|_{\text{nuclear}}.$$

Note Inequality for random matrix $\rightsquigarrow$ $\lambda_\epsilon \sim \sqrt{pq\log(pq)/n}$.

p-values

As before we consider some empirical risk R_n.
We use the one step estimator

$$\hat{b} = \hat{\beta} - \hat{\Theta}^T \dot{R}_n(\hat{\beta})$$

where $\hat{\Theta}$ is some approximation of the inverse Fisher information matrix.

Let $\hat{W}$ be a diagonal matrix of weights.

We have

$$\hat{W}(\hat{b} - \beta^0) = \hat{W}(\hat{\beta} - \beta^0) - \hat{W}\hat{\Theta}^T \dot{R}_n(\hat{\beta})$$

$$= -\underbrace{\hat{W}\hat{\Theta}^T \dot{R}_n(\beta^0)}_{\text{main term}} + \underbrace{\hat{W}\left(I - \hat{\Theta}^T \ddot{R}_n(\tilde{\beta})\right)(\hat{\beta} - \beta^0)}_{\text{remainder}}$$

Hence to show: for some surrogate inverse $\hat{\Theta}$ and matrix of weights $\hat{W}$:

$$\hat{W}(I - \hat{\Theta}^T \ddot{R}_n(\tilde{\beta})) \text{ is “small”.}$$

In addition, we want studentization:

$$\mathrm{diag}\left(\hat{W}\hat{\Theta}^T \mathrm{Cov}(\dot{R}_n(\beta^0))\hat{\Theta}\hat{W}\right) \approx I.$$

P-values using the Lasso

$$Y = X\beta^0 + \epsilon.$$

$$R_n(\beta) := \frac{1}{2n}\|Y - X\beta\|_2^2.$$

$$\dot{R}_n(\beta) = -X^T(Y - X\beta)/n, \ \dot{R}_n(\beta^0) = -X^T\epsilon/n$$

$$\ddot{R}_n(\beta) = X^TX/n =: \hat{\Sigma}.$$

So we need a surrogate inverse for $\hat{\Sigma}$.

Inverting a matrix Σ_0

Suppose $\Theta_0 := \Sigma_0^{-1}$ exists.
Then

$$\Theta_0 = \begin{pmatrix} \theta_1^0 & \theta_2^0 & \cdots & \theta_p^0 \end{pmatrix}$$

where

$$\theta_j^0 = \frac{1}{\tau_j^2} \begin{pmatrix} -\gamma_{1,j} \\ \vdots \\ 1 \\ \vdots \\ -\gamma_{p,j} \end{pmatrix} \begin{array}{l} \\ \\ \leftarrow j^{\text{th}} \text{ row} \\ \\ \\ \end{array}$$

with
$\{\gamma_{k,j}\}_{k \neq j}$: coefficients of the projection of the j^{th} variable on all others,
τ_j: the length of the residual.

Square-root Lasso

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \| Y - X\beta \|_2 / \sqrt{n} + \lambda_0 \|\beta\|_1 \right\}.$$

The surrogate inverse

Let $\hat{\gamma}_j$ be the square-root Lasso with tuning parameter $\lambda_\sharp$ for the regression of X_j on X_{-j}.
Define the residuals

$$\hat{\tau}_j := \|X_j - X_{-j}\hat{\gamma}_j\|_2/\sqrt{n} = \|X\hat{C}_j\|_2/\sqrt{n}.$$

Let $\tilde{\tau}_j^2 := \hat{\tau}_j(\hat{\tau}_j + \lambda_\sharp \|\hat{\gamma}\|_1)$.
Define $\hat{\theta}_j := \hat{C}_j/\tilde{\tau}_j^2$.
Surrogate inverse of the Gram matrix $\hat{\Sigma} := X^T X/n$:

$$\hat{\Theta} := (\hat{\theta}_1, \ldots, \hat{\theta}_p)$$

Let

$$\hat{W} := \frac{\sqrt{n}}{\hat{\sigma}} \begin{pmatrix} \hat{\tau}_1 + \lambda_\sharp \|\hat{\gamma}_1\|_1 & & \\ & \ddots & \\ & & \hat{\tau}_1 + \lambda_\sharp \|\hat{\gamma}_1\|_1 \end{pmatrix}$$

Then

$$\begin{aligned}
&\|\hat{W}(I - \hat{\Theta}^T\hat{\Sigma})(\hat{\beta} - \beta^0)\|_\infty \\
&\le \|\hat{W}(I - \hat{\Theta}^T\hat{\Sigma})\|_\infty \|\hat{\beta} - \beta^0\|_1 \\
&\le \sqrt{n}\lambda_\sharp \|\hat{\beta} - \beta^0\|_1 \hat{\sigma}.
\end{aligned}$$

Moreover

$$\mathrm{diag}(\hat{W}\hat{\Theta}^T \mathrm{Cov}(X^T\epsilon/n)\hat{\Theta}\hat{W}) = \frac{\sigma_0^2}{\hat{\sigma}^2} I.$$

Let the *de-sparsified Lasso* be the one step estimator

$$\hat{b} := \hat{\beta} + \hat{\Theta}^T \overbrace{X^T(Y - X\hat{\beta})/n}^{-\dot{R}_n(\hat{\beta})}$$

Asymptotic linearity *We have*

$$\hat{W}(\hat{b} - \beta^0) = \underbrace{\hat{W}\hat{\Theta}^T X^T \epsilon / n}_{\textit{studentized linear term}} + \mathrm{rem},$$

where $\|\mathrm{rem}\|_\infty \le \sqrt{n}\lambda_\sharp \|\hat{\beta} - \beta^0\|_1 / \hat{\sigma}$.

Conclusion

∘ One can derive sharp oracle inequalities for empirical risk minimizers penalized by an appropriate norm.
∘ The choice of the norm depends on the sparsity structure one has in mind.
∘ Examples include exponential families, support vector machines, trace regression, graphical models , ...
∘ For certain cases these oracle estimators can serve as initial estimators in a one step procedure.
∘ The one-step procedure removes the asymptotic bias but yields non-sparse estimators....
∘ which serve as pivot for asymptotic p-values.

THANK YOU!