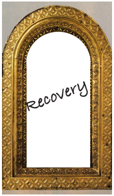
High-dimensional statistics: a triptych

Sara van de Geer

Panel I: Oracle inequalities

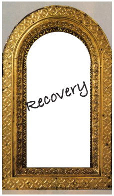
Panel II: Asymptotic normality and CRLB's



Panel III : Lower bounds for restricted eigenvalues

Panel I: Oracle inequalities

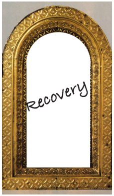
Panel II: Asymptotic normality and CRLB's



Panel III : Lower bounds for restricted eigenvalues

Panel I: Oracle inequalities

Panel II: Asymptotic normality and CRLB's



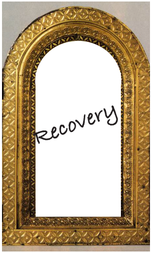
Panel III : Lower bounds for restricted eigenvalues

High-dimensional statistics: a triptych

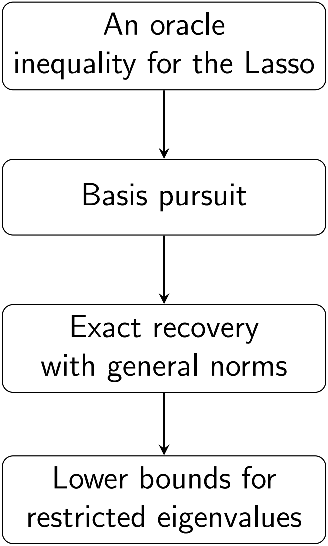
Sara van de Geer

July 15, 2016

Panel III:
Lower bounds for restricted eigenvalues

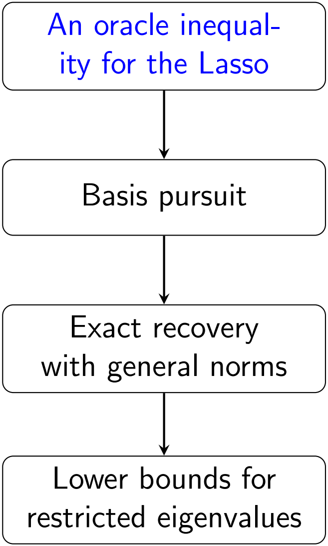


Joint work with:
Andreas Elsener, Alan Muro, Jana Janková, Benjamin Stucky



Concepts:

Compatibility constant
Effective sparsity
Isotropy
Small ball property



Concepts:

Compatibility constant
Effective sparsity
Isotropy
Small ball property

- $Y \in \mathbb{R}^n$ response
- $X \in \mathbb{R}^{n \times p}$ co-variables

Linear model:

$$Y = X\beta^0 + \epsilon, \;\; \epsilon \sim \mathcal{N}_n(0, I)$$

Lasso

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2/n + 2\lambda\|\beta\|_1 \right\}$$

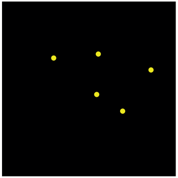
Notation

Active set:

$S_0 := \{j : \ \beta_j^0 \neq 0\}$

Sparsity:

$s_0 := |S_0|$



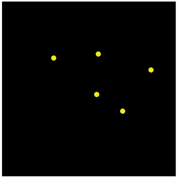
Coefficients of a vector $\beta \in \mathbb{R}^p$ on a set $S \subset \{1, \ldots, p\}$:

$$\beta_S := \begin{pmatrix} \beta_1 \\ \vdots \\ 0 \\ 0 \\ \beta_{j+1} \\ \vdots \\ 0 \end{pmatrix} \qquad \begin{array}{l} \leftarrow \in S \\ \vdots \\ \leftarrow \notin S \\ \leftarrow \notin S \\ \leftarrow \in S \\ \vdots \\ \leftarrow \notin S \end{array}$$

Notation

Active set:
$S_0 := \{j : \ \beta_j^0 \neq 0\}$

Sparsity:
$s_0 := |S_0|$



Coefficients of a vector $\beta \in \mathbb{R}^p$ on a set $S \subset \{1, \ldots, p\}$:

$$\beta_S := \begin{pmatrix} \beta_1 \\ \vdots \\ 0 \\ 0 \\ \beta_{j+1} \\ \vdots \\ 0 \end{pmatrix} \qquad \begin{array}{c} \leftarrow \in S \rightarrow \\ \vdots \\ \leftarrow \notin S \rightarrow \\ \leftarrow \notin S \rightarrow \\ \leftarrow \in S \rightarrow \\ \vdots \\ \leftarrow \notin S \rightarrow \end{array} \qquad \begin{pmatrix} 0 \\ \vdots \\ \beta_{j-1} \\ \beta_j \\ 0 \\ \vdots \\ \beta_p \end{pmatrix} =: \beta_{-S}$$

Note:

$\{X\beta_S : \ \beta \in \mathbb{R}^p\}$ is the linear space spanned by $\{X_j\}_{j \in S}$

$\{X\beta_{-S} : \ \beta \in \mathbb{R}^p\}$ is the linear space spanned by $\{X_j\}_{j \notin S}$

Definition compatibility constant:

$$\hat{\phi}^2(L, S) := \min\left\{ s\|X\beta_S - X\beta_{-S}\|_2^2/n : \ \underbrace{\|\beta_S\|_1 = 1, \ \|\beta_{-S}\|_1 \leq L}_{\text{"cone condition"}} \right\}$$

with

- $S \subset \{1, \ldots, p\}$, $s := |S|$
- $L \geq 1$ *a "stretching factor"*

Lemma

[vdG 2007, Bickel et al. 2009, Koltchinskii et al. 2011, ...]
Let $\lambda_0 := \sqrt{2\log(2p/\alpha)}$ *and* $\lambda > \lambda_0$.
Then with probability at least $1-\alpha$

$$\|X(\hat\beta - \beta^0)\|_2^2/n + \lambda_0\|\hat\beta - \beta^0\|_1/c \le \frac{\lambda^2 s_0}{\hat\phi^2(L, S_0)}.$$

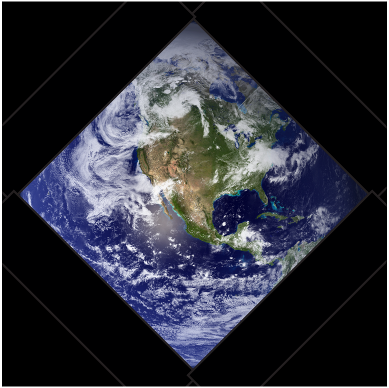
where

$$L := C \ \ \frac{\lambda+\lambda_0}{\lambda-\lambda_0}.$$

□

We will now show that the compatibility constant is related to canonical correlation ...

... in the ℓ_1-world

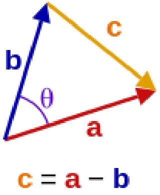


A note about distances and inner products

Let a and b be two vectors in $\mathbb{R}^n$.
Suppose $\|a\|_2 = \|b\|_2 = 1$.
Write

- $\rho := a^T b \; (= cos(\theta))$
- $c := a - b$

Then $\|c\|_2^2 = \|a\|_2^2 + \|b\|_2^2 - 2a^T b = 2(1-\rho) =: \phi^2$



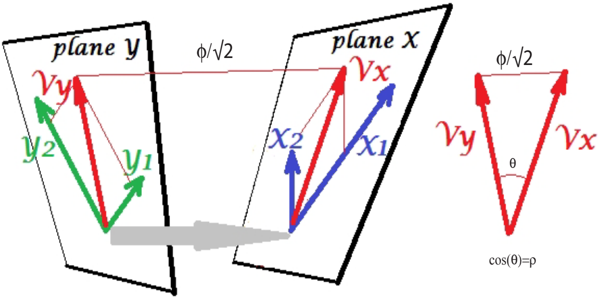
Canonical correlation in the ℓ_2-world

The canonical correlation between X_S and X_{-S} is

$$\hat{\rho}_{\text{can}}(S) := \max\left\{ (X\beta_{-S})^T (X\beta_S) : \|X\beta_S\|_2 = 1, \ \|X\beta_{-S}\|_2 = 1 \right\}.$$

We let $\hat{\phi}^2_{\text{can}}(S) = 2(1 - \hat{\rho}_{\text{can}}(S))$.

Canonical correlation: (in $\mathbb{R}^4$)



$$\mathcal{Y} = \{X\beta_{-S} : \ \beta \in \mathbb{R}^p\} \quad \mathcal{X} = \{X\beta_S : \ \beta \in \mathbb{R}^p\}$$

Note:

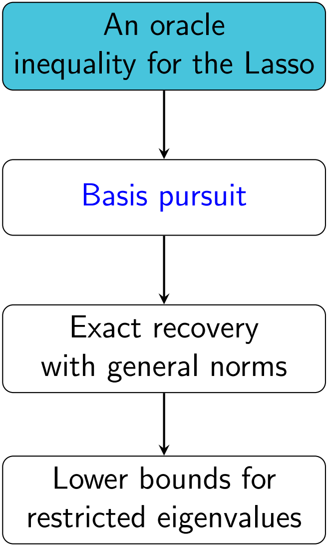
Assume (WLG) $X_S^T X_S = I$ and $X_{-S}^T X_{-S} = I$.

Then

$$\hat{\rho}_{\mathrm{can}}(S) = \max\left\{ (X\beta_{-S})^T (X\beta_S) : \ \|\beta_S\|_2 = 1, \ \|\beta_{-S}\|_2 = 1 \right\}.$$

Moreover then

$$\begin{aligned} \hat{\phi}^2_{\mathrm{can}}(S) &= 2(1 - \hat{\rho}_{\mathrm{can}}(S)) \\ &= \min\left\{ \|X\beta_S - X\beta_{-S}\|_2^2 : \ \|\beta_S\|_2 = 1, \ \|\beta_{-S}\|_2 = 1 \right\} \end{aligned}$$



Concepts:

Compatibility constant
Effective sparsity
Isotropy
Small ball property

Canonical correlation and exact recovery

X given $n \times p$ matrix with $p \leq n$.

Observed: $f^0 := X\beta^0$.

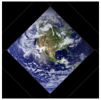
Let

$$\beta^* := \min\{\|\beta\|_2 : \ X\beta = f^0\}.$$

Then

$$\boxed{\hat{\phi}_{\rm can}(S_0) > 0 \ \Rightarrow \ \beta^* = \beta^0}$$

Canonical correlation in the ℓ_1-world



Note:

$$\boxed{\max\{\|\beta_S\|_1^2 : \ \|\beta_S\|_2^2 = 1\} = s} \text{ where } s := |S|.$$

$$\underset{\ell_1}{\uparrow} \qquad \underset{\ell_2}{\uparrow}$$

Notation

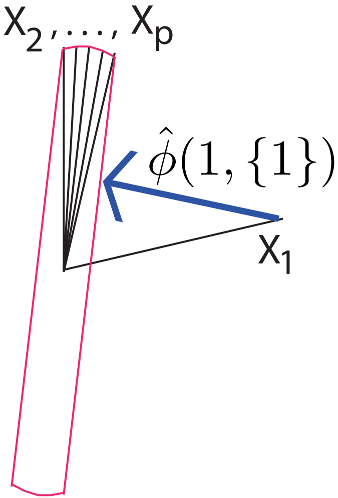
For a vector $v \in R^n$: $\|v\|_n{}^2 := v^T v / n$

abuse of notation

Definition *The compatibility constant is*

$$\hat{\phi}^2(S) := \min\Big\{ s\|X\beta_S - X\beta_{-S}\|_n^2 : \ \underbrace{\|\beta_S\|_1 = 1, \ \|\beta_{-S}\|_1 \le 1}_{\ell_1 \text{ instead of } \ell_2} \Big\}.$$

Compatibility constant: (in $\mathbb{R}^2$)



$\hat{\phi}(S) = \hat{\phi}(1, S)$ for the case $S = \{1\}$

Basis Pursuit

X given $n \times p$ matrix with $p \gg n$.

Observed: $f^0 := X\beta^0$.
Basis Pursuit [Chen, Donoho and Saunders (1998)]:

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

Exact recovery

$$\hat{\phi}(S_0) > 0 \Rightarrow \beta^* = \beta^0$$

Effective sparsity

Note:

$$\hat{\phi}^2(L,S) = \min\Big\{ s\|X\beta\|_n^2 : \overbrace{\|\beta_S\|_1 = 1,\ \|\beta_{-S}\|_1 \le L}^{\text{“cone condition”}} \Big\}$$

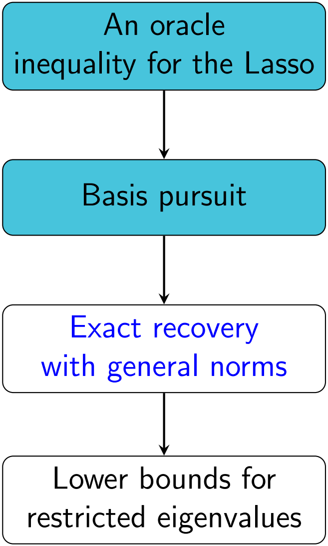
$$= \min\Big\{ s\frac{\|X\beta\|_n^2}{\|\beta_S\|_1^2} : \ \|\beta_{-S}\|_1 \le L\|\beta_S\|_1 \Big\}$$

$$\Rightarrow$$

$$\frac{s}{\hat{\phi}^2(L,S)} = \max\Big\{ \frac{\|\beta_S\|_1^2}{\|X\beta\|_n^2} : \ \|\beta_{-S}\|_1 \le L\|\beta_S\|_1 \Big\}$$

$$:= \hat{\Gamma}^2(L,S) := \text{effective sparsity}$$

We call $\hat{\Gamma}(L,S)$ an $\boxed{\ell_1 \backslash \|\cdot\|_n\text{-comparison}}$



Concepts:

Compatibility constant
Effective sparsity
Isotropy
Small ball property

General norms

Let Ω be a norm on $\mathbb{R}^p$

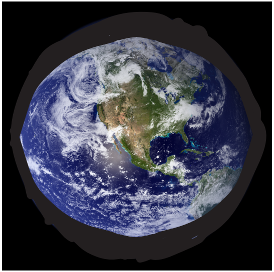
Examples:

ℓ_1-norm

sorted ℓ_1-norm

variational norms

nuclear norm



The $\Omega-$world

Definition *The triangle property holds at β^0 if $\exists$ semi-norms Ω^+ and Ω^- such that*

$$\boxed{\Omega(\beta^0) - \Omega(\beta) \le \Omega^+(\beta - \beta^0) - \Omega^-(\beta)}$$

.

Example: ℓ_1-norm

$$\boxed{\|\beta^0\|_1 - \|\beta\|_1 \le \|\beta_{S_0} - \beta^0\|_1 - \|\beta_{-S_0}\|_1}$$

Definition *The* *Ω-effective sparsity* *is*

$$\hat{\Gamma}^2(L,\beta^0) := \max\left\{ \frac{(\Omega^+(\beta))^2}{\|X\beta\|_n^2} : \underbrace{\Omega^-(\beta) \le L\Omega^+(\beta)}_{\text{“cone condition”}} \right\}.$$

We call $\hat{\Gamma}(L,\beta^0)$ an $\boxed{\Omega \backslash \|\cdot\|_n\text{-comparison}}$

Example: ℓ_1-norm

$$\hat{\Gamma}^2(L,\beta^0) = \frac{s_0}{\hat{\phi}^2(L,S_0)}$$

Notation

For stretching factor $L = 1$:

$$\hat{\Gamma}(\beta^0) := \hat{\Gamma}(1,\beta^0)$$

Exact recovery using general norms

Let

$$\beta^* := \min\left\{\Omega(\beta) : \; X\beta = f^0\right\}.$$

Then

$$\boxed{\hat{\Gamma}(\beta^0) < \infty \;\Rightarrow\; \beta^* = \beta^0}$$

Ω-exact recovery results
and
oracle results for Ω-penalized least squares

involve the effective sparsity
$\hat{\Gamma}^2(L, \beta^0)$

Problem setting

Question

$$\text{when is } \hat{\Gamma}(L, \beta^0) < \infty?$$

or

how large is it?

Notation

$$\underline{\Omega} := \Omega^+ + \Omega^-$$

Example: ℓ_1-norm

$$\underline{\Omega} = \Omega = \| \cdot \|_1$$

Ω^+/ℓ_2-comparison

$$\gamma_0 := \max \left\{ \frac{\Omega^+(\beta)}{\|\beta\|_2} : \ \beta \in \mathbb{R}^p \right\}$$

Example: ℓ_1-norm

$$\gamma_0^2 = \frac{1}{s_0}$$

Note

$$\begin{array}{l} \Omega^+(\beta) = 1 \\ \Omega^-(\beta) \le L \end{array} \quad \Rightarrow \quad \begin{array}{l} \|\beta\|_2 \ge 1/\gamma_0 \\ \underline{\Omega}(\beta) \le L + 1 \end{array}$$

cone condition

$\rightsquigarrow$ study of <u>restricted eigenvalue</u>

$$\boxed{\min\left\{ \|X\beta\|_n^2 : \ \|\beta\|_2 = 1, \ \underline{\Omega}(\beta) \le M \right\}^1}$$

Notation $\hat{\Sigma} := X^T X / n$ Gram matrix

Note: As

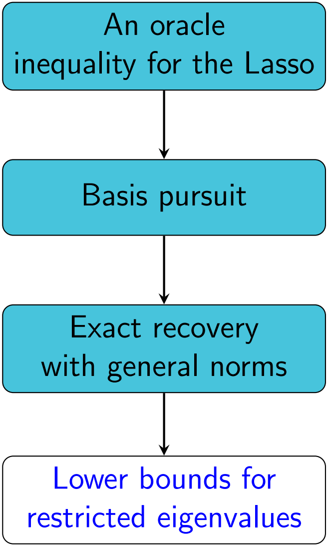
$$\|X\beta\|_n^2 = \beta^T \hat{\Sigma} \beta$$

the minimal eigenvalue of $\hat{\Sigma}$ is

$$\boxed{\min\left\{ \|X\beta\|_n^2 : \ \|\beta\|_2 = 1 \right\}}$$

which is **zero** when $p > n$.

[1] with $M := L\gamma_0$



Concepts:

Compatibility constant
Effective sparsity
Isotropy
Small ball property

Random matrix

$$X := \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \quad n \times p \ \text{ data matrix}$$

Model:
Rows $X_i = (X_{i,1}, \ldots, X_{i,p})$ i.i.d. copies of a random vector $X_0 \in \mathbb{R}^p$.

Empirical inner product matrix $\hat{\Sigma} := X^T X / n$.
Theoretical inner product matrix $\Sigma_0 := \mathbb{E}\hat{\Sigma} = \mathbb{E}X_0^T X_0$.

We will provide lower bounds for

$$\min \left\{ \beta^T \hat{\Sigma} \beta : \ \beta^T \Sigma_0 \beta = 1, \ \Omega(\beta) \leq M \right\}.$$

A simple bound for ℓ_1-case

Let Ω be the ℓ_1-norm.
We have

$$\max\left\{ \left| \beta^T(\hat{\Sigma} - \Sigma_0)\beta \right| : \ \|\beta\|_1 \le M \right\} \le M^2 \|\hat{\Sigma} - \Sigma_0\|_\infty .$$

Hence

$$\min\left\{ \beta^T \hat{\Sigma} \beta : \ \|\beta\|_1 \le M, \ \beta^T \Sigma_0 \beta = 1 \right\} \ge 1 - M^2 \|\hat{\Sigma} - \Sigma_0\|_\infty .$$

Asymptotics

$$\|\hat{\Sigma} - \Sigma_0\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n}) \ \rightsquigarrow \ \text{requiring } M = o((n/\log p)^{\frac{1}{4}})$$

We will improve this to

$$M = o(\sqrt{n/\log p})$$

Definition

Let $m \geq 2$.

The random vector $X_0 \in \mathbb{R}^p$ is m-th order isotropic with constant C if for all $\beta \in \mathbb{R}^p$ with $\beta^T \Sigma_0 \beta = 1$ it holds that

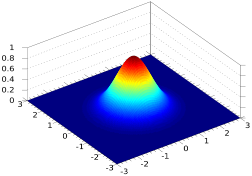
$$\mathbb{P}(|X_0 \beta| > t) \leq (C/t)^m \ \forall \ t > 0.$$

Example

Suppose $X_0 \sim \mathcal{N}_p(0, \Sigma_0)$

Then $\forall$ m

X_0 is m-th order isotropic with universal constant C



Let $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence independent of X.
Let Ω_* be the dual norm of Ω.
Theorem 1 [vdG, 2016]
Suppose that for some $m > 2$ the random vector X_0 is weakly m-th order isotropic.
Then i $\Rightarrow$ ii:

$$i: \ \mathbb{E}\Omega_*(X^T\epsilon)/n \times M = o(1)$$

$$\Downarrow$$

$$ii: \ \min\left\{ \beta^T \hat{\Sigma} \beta : \ \beta^T \Sigma_0 \beta = 1, \ \Omega(\beta) \le M \right\} \ge 1 - o_{\mathbb{P}}(1).$$

Relation with oracle inequalities

Consider[2] $\Omega = \|\cdot\|_1$

Lemma [vdG 2007, Bickel et al. 2009, ...]

Let[3] $\lambda_0 := \sqrt{2\log(2p/\alpha)}$ *and* $\lambda > \lambda_0$.
Then with probability at least $1-\alpha$

$$\|X(\hat\beta - \beta^0)\|_2^2/n + \underbrace{\lambda_0}_{\sim\Omega_*(X^T\epsilon)/n} \underbrace{\|\hat\beta - \beta^0\|_1/c}_{\sim\hat M} \le \underbrace{\frac{\lambda^2 s_0}{\phi^2(L, S_0)}}_{o_{\mathbb{P}}(1)}.$$

where

$$L := C \ \frac{\lambda + \lambda_0}{\lambda - \lambda_0}.$$

□

[2]Similar relations for general norms
[3]$\mathbb{P}(\|X^T\epsilon\|_\infty/n \ge \lambda_0) \le \alpha$

Convergence of the compatibility constant

Theoretical compatibility constant:

$$\phi_0^2(L, S_0) := \min\left\{ s_0 \beta^T \Sigma_0 \beta : \ \|\beta_{S_0}\|_1 = 1, \ \|\beta_{-S_0}\|_1 \leq L \right\}.$$

Empirical compatibility constant:

$$\hat{\phi}^2(L, S_0) := \min\left\{ s_0 \beta^T \hat{\Sigma} \beta : \ \|\beta_{S_0}\|_1 = 1, \ \|\beta_{-S_0}\|_1 \leq L \right\}.$$

By the theorem,
under isotropy and assuming $\|X^T \epsilon\|_\infty / n = \mathcal{O}_{\mathbb{P}}(\sqrt{\log p / n})$:

$$\frac{s_0}{\phi_0^2(L, S_0)} = o\left(\frac{n}{\log p} \right) \Rightarrow \hat{\phi}^2(L, S_0) = \phi_0^2(L, S_0)(1 - o_{\mathbb{P}}(1)).$$

Convergence of the compatibility constant

Theoretical compatibility constant:

$$\phi_0^2(L, S_0) := \min\left\{ s_0 \beta^T \Sigma_0 \beta : \ \|\beta_{S_0}\|_1 = 1, \ \|\beta_{-S_0}\|_1 \leq L \right\}.$$

Empirical compatibility constant:

$$\hat{\phi}^2(L, S_0) := \min\left\{ s_0 \beta^T \hat{\Sigma} \beta : \ \|\beta_{S_0}\|_1 = 1, \ \|\beta_{-S_0}\|_1 \leq L \right\}.$$

By the theorem,
under isotropy and assuming $\|X^T \epsilon\|_\infty / n = \mathcal{O}_{\mathbb{P}}(\sqrt{\log p / n})$:

$$\frac{s_0}{\phi_0^2(L, S_0)} = o\left(\frac{n}{\log p} \right) \Rightarrow \hat{\phi}^2(L, S_0) \gg 0 \text{ whp}$$

Let $\Lambda_{\min}(\Sigma_0)$ be the smallest eigenvalue of Σ_0.

Corollary

Since $\phi^2(L, S_0) \geq \Lambda_{\min}(\Sigma_0)$ *we have under the conditions of the previous*

$$\Lambda_{\min}(\Sigma_0) \gg 0 \ \Rightarrow \ \hat{\phi}^2(L, S_0) \gg 0 \text{ whp}$$

Bounds for the compatibility constant using the transfer principle

Transfer principle (Oliveira[2013])
Let A be a symmetric $p \times p$ matrix with $A_{j,j} \geq 0$ for all $j \in \{1, \ldots, p\}$.
Let $s \in \{2, \ldots, p\}$ and suppose that for all $S \subset \{1, \ldots, p\}$ with cardinality $|S| = s$ one has

$$\beta_S^T A \beta_S \geq 0 \qquad \forall \ \beta \in \mathbb{R}^p.$$

Then

$$\beta^T A \beta \geq - \max_j A_{j,j} \|\beta\|_1^2 / (s-1) \qquad \forall \ \beta \in \mathbb{R}^p.$$

WLG: $\mathrm{diag}(\Sigma_0) = I$:

$$\Sigma_0 = \begin{pmatrix} 1 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{1,2} & 1 & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,p} & \sigma_{2,p} & \cdots & 1 \end{pmatrix}$$

Normalized design

Define

$$\hat{R} := \tilde{X}^T \tilde{X} / n, \ \ \tilde{X}_j := X_j / \hat{\sigma}_j, \ \ j = 1, \ldots, p.$$

Define the (empirical) compatibility constant for normalized design

$$\hat{\tilde{\phi}}^2(L, S_0) := \min \left\{ s_0 \beta^T \hat{R} \beta : \ \ \| \beta_{S_0} \|_1 = 1, \ \ \| \beta_{-S_0} \|_1 \leq L \right\}.$$

The (theoretical) adaptive restricted eigenvalue is defined as

$$\kappa_0^2(L, S_0) := \min \left\{ \beta^T \Sigma_0 \beta : \ \ \| \beta_{S_0} \|_2 = 1, \ \ \| \beta_{-S_0} \|_1 \leq L \sqrt{s_0} \right\}.$$

Note

$$\phi^2(L, S_0) \geq \kappa_0^2(L, S_0) \geq \Lambda_{\min}(\Sigma_0)$$

Theorem [vdG and Muro, 2014]
Suppose that for some $m > 2$ the random vector X_0 is weakly m-th order isotropic. Then for some $c > 1$

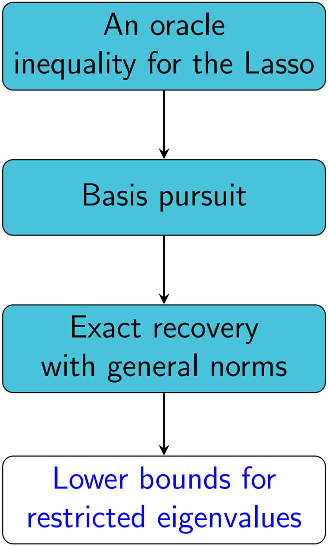
$$\frac{s_0}{\kappa_0^2(cL, S_0)} = o\left(\frac{n}{\log p}\right) \Rightarrow \hat{\tilde{\phi}}^2(L, S_0) \geq \frac{\kappa_0^2(cL, S_0)(1 - o_{\mathbb{P}}(1))}{c}.$$

Theorem [vdG and Muro, 2014]
Suppose that for some $m > 2$ the random vector X_0 is weakly m-th order isotropic. Then for some $c > 1$

$$\frac{s_0}{\kappa_0^2(cL, S_0)} = o\left(\frac{n}{\log p}\right) \Rightarrow \hat{\tilde{\phi}}^2(L, S_0) \gg 0 \text{ whp.}$$

Conclusion of this

With normalized design one can show that the compatibility constant $\tilde{\phi}^2(L, S)$ stays away from zero assuming only higher order isotropy and no further (moment) conditions.



Concepts:

Compatibility constant
Effective sparsity
Isotropy
Small ball property

Definition

The random vector $X_0 \in \mathbb{R}^p$ satisfies the small ball property with constants $C_1 > 0$ and $C_2 > 0$ if for all $\beta \in \mathbb{R}^p$ with $\beta^T \Sigma_0 \beta = 1$ it holds that

$$\mathbb{P}(|X_0 \beta| \geq 1/C_1) \geq 1/C_2.$$

It can be shown that for appropriate constants one has (for $m > 2$)

X_0 Gaussian $\Rightarrow$ m-th order isotropy $\Rightarrow$ small ball property.

E.g. see [Mendelson & Koltchinskii (2013)].

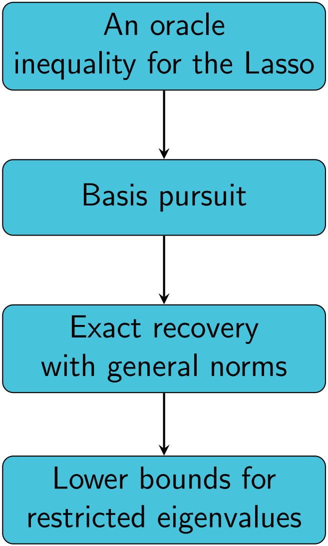
Bounds using the small ball property

Theorem [Lecué and Mendelson, 2015]
Suppose
∘ X_0 *satisfies the small ball property*
∘ $\max_j \hat{\sigma}_j = 1 + \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n})$.
Then for a constant $c > 1$ $i \Rightarrow ii$:

$$i: \; M = o\left(\frac{n}{\log p}\right)$$

$$\Downarrow$$

$$ii: \; \min\left\{\beta^T \hat{\Sigma} \beta : \; \|\beta\|_2 = 1, \; \|\beta\|_1 \leq M\right\} \geq \frac{1 - o_{\mathbb{P}}(1)}{c}$$



Concepts:

Compatibility constant
Effective sparsity
Isotropy
Small ball property

Conclusion

the compatibility constant
stays away from zero
when $s_0 = o(\sqrt{n/\log n})+$
distributional assumptions
or when $s_0 = o(n/\log n)+$
stronger
distributional assumptions

for norms more general than ℓ_1 the results go through under isotropy conditions but not (yet) under the small ball property


THANK YOU!

Higher order isotropy

Suppose the graph of X_0 has a directed acyclic graph (DAG) structure that is, satisfying (after an appropriate permutation of the indexes) the structural equations model

$$X_{0,1} = \epsilon_{0,1}, \tag{1}$$

$$X_{0,j} = \sum_{k=1}^{j-1} X_{0,k}\beta_{k,j} + \epsilon_{0,j},\ j = 2, \ldots, p$$

where $\{\epsilon_{0,j}\}_{j=1}^{p}$ is a martingale difference array for the filtration $\{\mathcal{F}_j\}_{j=0}^{p-1}$.

We assume:
- $X_{0,j}$ is $\mathcal{F}_j$-measurable, $j = 1, \ldots, p$
- $\omega_j^2 := \mathrm{var}(\epsilon_{0,j}) = \mathbb{E}\mathrm{var}(X_j|\mathcal{F}_{j-1})$ exists for all j.

Lemma *Assume the structural equations model.. Assume in addition that for some constant* C *and for all* $\lambda \in \mathbb{R}$

$$\mathbb{E}(\exp[\lambda\epsilon_{0,j}/\omega_j]|\mathcal{F}_{j-1}) \leq \exp[\lambda^2 C^2/2], j = 1, \ldots, p.$$

Then X_0 *is sub-Gaussian with constant* C.

More generally let $\{\mathcal{F}_j\}_{j=0}^{p}$ be a filtration and for $j = 1, \ldots, p$, let $X_{0,j}$ be $\mathcal{F}_j$-measurable and V_j be $\mathcal{F}_{j-1}$-measurable and satisfying for some $m > 2$,

$$\max_{1 \le j \le p} \|V_j\|_m := \mu_m < \infty.$$

Lemma *Suppose that for some constant K and all j*

$$\mathbb{E}(X_{0,j}|\mathcal{F}_{j-1}) = 0,\ \mathbb{E}(|X_{0,j}|^k|\mathcal{F}_{j-1}) \le \frac{k!}{2} K^{k-2} V_j^2,\ \ k = 2, 3, \ldots.$$

We have for all $2 < m_0 < m$ and all $\|u\|_2 = 1$

$$\|X_0 u\|_{m_0} \le \sqrt{\frac{2m}{m - m_0}} \left(\frac{3m\Gamma(m_0/2 + 1)}{m - m_0} \right)^{m_0/2+1} \mu_m + \Big(3\Gamma(m_0+1) \Big)^{1/}$$