High-Dimensional Econometrics:

Notions from Concentrations Inequalities & Beyond

Marcelo Ortiz-Villavicencio

EMORY UNIVERSITY

January 18, 2024

Econometrics Reading Group

Roadmap

1. Classical versus high-dimensional theory

   What can go wrong in high dimensions?

   What is the non-asymptotic viewpoint?

2. Concentration Inequalities

3. High-Dimension, Sparsity and the Lasso

   Model selection via Lasso

   Some theory of Lasso

Classical versus high-dimensional theory

Motivation

- Main question: Why do we care about *high-dimensional problems*?
- Some essential facts that motivate this discussion are the following:
  1. New datasets arising in many economic contexts have a ”high-dimensional flavor”, with d on the same order as, or possibly larger, than the sample size n
  2. The classical theory that relies on *large n, fixed d* fails to provide useful theoretical predictions.
  3. Classical methods can break down dramatically in high-dimensional settings.
- Let’s see an example to appreciate the challenges!

What can go wrong in high dimensions?

- Suppose we have a collection of random vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$
- Each $\mathbf{x}_i$ is drawn i.i.d from zero-mean distribution in $\mathbb{R}^d$.
- Our goal is to estimate $\mathbf{\Sigma} = cov(X)$
- Consider the following *sample covariance estimator*

$$\hat{\mathbf{\Sigma}} := n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$$

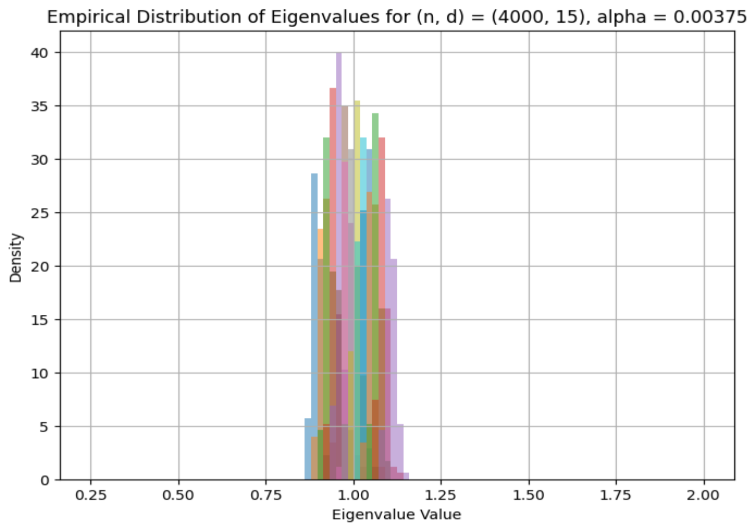
- By construction, the sample covariance matrix $\hat{\mathbf{\Sigma}}$ is an *unbiased* estimate, meaning $\mathbb{E}[\hat{\mathbf{\Sigma}}] = \mathbf{\Sigma}$

What can go wrong in high dimensions?

- A classical analysis considers the behavior of the sample covariance matrix as n increases while d stays fixed.
- We argue that the sample covariance matrix is a *consistent* estimate.
- Question: Is this type of consistency preserved if we allow the dimension d to tend to infinity?
- Suppose that we allow both n and d increase with their ratio remaining fixed, say $d/n = \alpha \in (0, 1)$
- Let $\mathbf{\Sigma} = \mathbf{I}_d$ with each $\mathbf{x}_i \sim N(0, \mathbf{I}_d)$ for $i = 1, \cdots, n$
- Using these n samples, we generated the sample covariance matrix, and then computed its vector of eigenvalues $\gamma(\widehat{\mathbf{\Sigma}}) \in \mathbb{R}^d$, say arranged in non-increasing order as

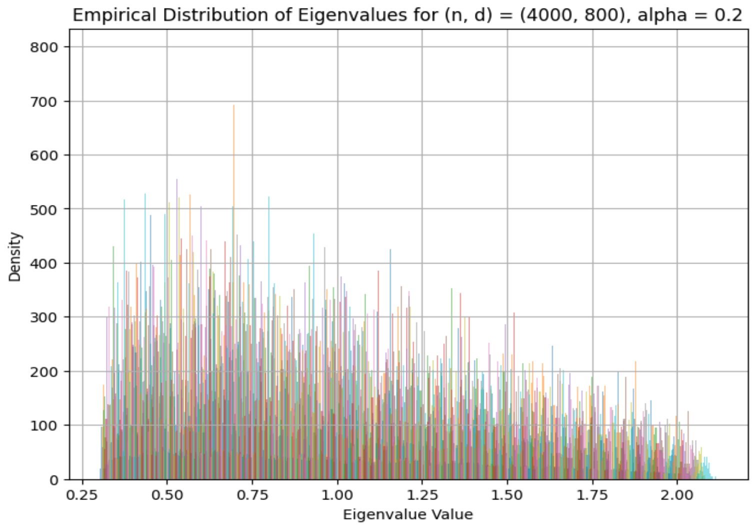
$$\gamma_{\max}(\widehat{\mathbf{\Sigma}}) = \gamma_1(\widehat{\mathbf{\Sigma}}) \geq \gamma_2(\widehat{\mathbf{\Sigma}}) \geq \cdots \geq \gamma_d(\widehat{\mathbf{\Sigma}}) = \gamma_{\min}(\widehat{\mathbf{\Sigma}}) \geq 0$$

Empirical Distribution of eigenvalues $\gamma(\widehat{\Sigma})$ with $\alpha = 0.00375$



Empirical Distribution of eigenvalues $\gamma(\widehat{\Sigma})$ with $\alpha = 0.2$

EMORY UNIVERSITY



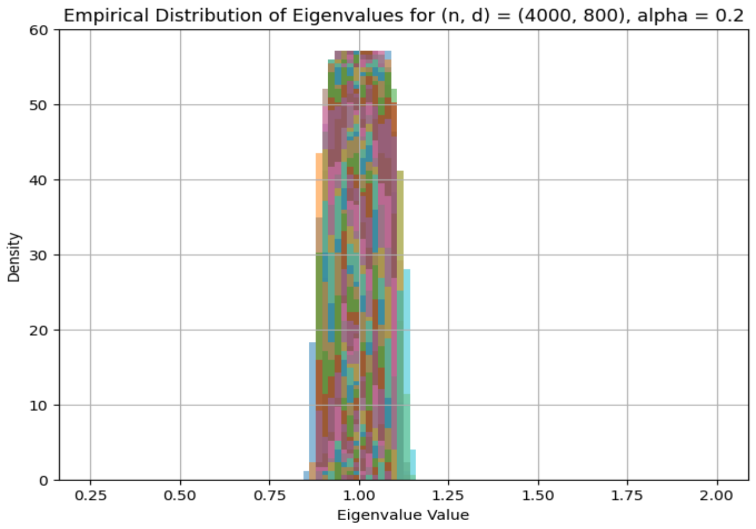
What can help us in the high-dimensional setting?

- Much of high-dimensional statistics involves constructing models of *high-dimensional phenomena* that consider some implicit form of *low-dimensional structure*!
- What types of low-dimensional structures might be appropriate for modeling covariance matrices problems?
- If we assume that the matrix is diagonal, we can improve by *imputing zeros* to *non-diagonal elements*
- $\Rightarrow$ *Sparsity*!
- Apply some form of *thresholding* by

$$T_\lambda(x) = \begin{cases} x & \text{if } |x| > \lambda \\ 0 & \text{otherwise} \end{cases}$$

- Let $\tilde{\mathbf{\Sigma}} := T_{\lambda_n}(\hat{\mathbf{\Sigma}})$ with $\lambda_n = \sqrt{\frac{2 log(d)}{n}}$

Thresholding Empirical Distribution of $\gamma(\widehat{\Sigma})$ with $\alpha = 0.2$



What is the non-asymptotic viewpoint?

- From our previous example, we can show that the maximum eigenvalue $\gamma_{\max}(\widehat{\boldsymbol{\Sigma}})$ satisfies the *upper deviation inequality*

$$\mathbb{P}[(\gamma_{\max}(\widehat{\boldsymbol{\Sigma}}) \geq (1 + \sqrt{d/n} + \delta)^2] \leq e^{-n\delta^2/2}$$

- Results of this type are what we *tail bounds* or *concentration inequalities*, and are the primary focus of *non-asymptotic theory* in high-dimensional statistics.
- The pair (n, d) is viewed as fixed, and high probability statements are made as a function of them.

Appetizer for econometrics

- Empirical Research involves crucial choices:
  - Functional forms
  - Selection of control variables
  - Choice of instruments
- In causal inference we consider the following model to estimate average treatment effect τ

$$Y_i = D_i\tau + X_i'\beta_0 + \varepsilon_i, \text{ with } \mathbb{E}\left[\varepsilon_i\right] = 0 \text{ and } \mathbb{E}\left[\varepsilon_i \mid D_i, X_i\right] = 0$$

  where X_i is a vector of p exogenous control variables, being possible $p \gg n$.
- Large dimension of X_i opens the door for selection methods such as the Lasso.
- Or even further, suppose we have access to a possibly large number of instrumental variables Z_i, all satisfying $\mathbb{E}[\varepsilon_i \mid Z_i] = 0$ (e.g., Judge IV).
- How do we deal with this?

Concentration Inequalities

Why concentration inequalities?

- Concentration inequalities are arguably some of the most important tools in modern statistical learning theory.
- Develop tools to show results that formalize the intuition for these statements:
  1. $X_1 + \cdots + X_n$ concentrates around $\mathbb{E}[X_1 + \cdots + X_n]$
  2. More general, $f(X_1, \cdots, X_n)$ concentrates around $\mathbb{E}[f(X_1, \cdots, X_n)]$
- We are interested in finite sample results and they usually take the form of two-sided bounds for the tails of deviations of a function from its mean

$$\mathbb{P}\left[\left|f\left(X_1, \ldots, X_n\right) - \mathbb{E}\left[f\left(X_1, \ldots, X_n\right)\right]\right| \geq t\right] \leq \textit{something small}$$

Going back to the basics: Classical Bounds

- The most elementary tail bound is the *Markov's inequality*.

Definition

Given a non-negative random variable X with finite mean, we have an *upper tail bound*

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \quad \text{for all } t > 0$$

Definition

For a random variable X that also has a finite variance, we define *Chebyshev's inequality* as:

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\text{var}(X)}{t^2} \quad \text{for all } t > 0.$$

where $\mu = \mathbb{E}[X]$.

- There are various extensions of Markov's inequality applicable to high orders of the form $|X - \mu|^k$ such that $\mathbb{P}[|X - \mu| \geq t] \leq \frac{\mathbb{E}[|X-\mu|^k]}{t^k}$.

Hoeffding's inequality

Let a sum of i.i.d *symmetric Bernoulli* random variables

Definition (Symmetric Bernoulli RV)

A random variable X has symmetric Bernoulli distribution (also called Rademacher distribution) if it takes values -1 and 1 with probabilities $1/2$ each, i.e.

$$\mathbb{P}\{X = -1\} = \mathbb{P}\{X = 1\} = \frac{1}{2}$$

Theorem (Hoeffding's Inequality)

Let $X_1, \ldots, X_N$ be independent symmetric Bernoulli random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for any $t \geq 0$, we have

$$\mathbb{P}\left\{\sum_{i=1}^{N} a_i X_i \geq t\right\} \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right).$$

General Hoeffding's inequality

EMORY UNIVERSITY

Theorem (Hoeffding's inequality for bounded RV)

Let $X_1, \ldots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have

$$\mathbb{P}\left\{\sum_{i=1}^{N}(X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right)$$

Remark

Unlike the classical limit theorems of Probability Theory, Hoeffding's inequality is non-asymptotic in the sense that it holds for all fixed N as opposed to $N \to \infty$. The larger N, the stronger inequality becomes.

The non-asymptotic nature of concentration inequalities like Hoeffding makes them attractive in applications in data science, where N often corresponds to sample size.

Applications?

- Widely used in statistical learning theory!
- In Supervised ML, Given n training samples, we can state bounds on the difference between the *observed* and *true error rates* for any classifier g
- In Online Learning, algorithms update their models sequentially as new data becomes available.
- Hoeffding's Inequality can be used to make statements about *how quickly* the *average loss* of the model converges to the *expected (true) average loss*.

From Markov to Chernoff

- We can generalize Markov's inequality for higher central moments of order k
- Same procedure can be applied to functions other than polynomials $|X - \mu|^k$.
- Suppose a RV X with mgf in a *neighborhood of zero*, meaning that there is some constant $b > 0$ such that the function $\varphi(\lambda) = \mathbb{E}\left[e^{\lambda(x-\mu)}\right]$ exists for all $\lambda \leq |b|$.
- We may apply Markov's inequality to the random variable $Y = e^{\lambda(X-\mu)}$
- Get the upper bound

$$\mathbb{P}[(X - \mu) \geq t] = \mathbb{P}\left[e^{\lambda(X-\mu)} \geq e^{\lambda t}\right] \leq \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}}$$

Definition (Chernoff Bound)

Optimizing our choice of λ to obtain the tightest result yields the *Chernoff bound* namely, the inequality

$$\log \mathbb{P}[(X - \mu) \geq t] \leq \inf_{\lambda \in [0,b]} \left\{\log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t\right\}.$$

Sub-Gaussian Variables

- The form of the tail bound obtained by the *Chernoff* approach depends on the *growth rate* of the mgf.
- Then we can classify RV in terms of their mgf.

Example

Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ be a Gaussian random variable with mean μ and variance σ^2. We know that X has the mgf

$$\mathbb{E}\left[e^{\lambda X}\right] = e^{\mu\lambda + \frac{\sigma^2\lambda^2}{2}}, \qquad \text{valid for all } \lambda \in \mathbb{R}.$$

Substituting this expression into the optimization problem defining the optimized Chernoff bound, we obtain

$$\inf_{\lambda \geq 0}\left\{\log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t\right\} = \inf_{\lambda \geq 0}\left\{\frac{\lambda^2\sigma^2}{2} - \lambda t\right\} = -\frac{t^2}{2\sigma^2},$$

Sub-Gaussian Variables

- We can conclude that any $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ RV satisfies the *upper deviation inequality*

$$\mathbb{P}[X \geq \mu + t] \leq e^{-\frac{t^2}{2\sigma^2}}$$

- We can introduce the following definition

Definition

A random variable X with mean $\mu = \mathbb{E}[X]$ is *sub-Gaussian* if there is a positive number σ such that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2} \quad \text{for all } \lambda \in \mathbb{R}.$$

- Combining our knowledge about Chernoff and sub-Gaussian, we claim

Proposition

Any sub-Gaussian variable satisfy the concentration inequality

$$\mathbb{P}[X \geq \mu + t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

Sub-Gaussian Variables

EMORY UNIVERSITY

- Of course there exists sub-Gaussian variables that are *non-Gaussian*.

Example

A Rademacher random variable ε takes the values $\{-1, +1\}$ equiprobably. We claim that it is sub-Gaussian with parameter $\sigma = 1$. By taking expectations and using the power-series expansion for the exponential, we obtain

$$
\begin{aligned}
\mathbb{E}\left[e^{\lambda\varepsilon}\right] = \frac{1}{2}\left\{e^{-\lambda} + e^{\lambda}\right\} &= \frac{1}{2}\left\{\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty}\frac{(\lambda)^k}{k!}\right\} \\
&= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \\
&\le 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{2^k k!} \\
&= e^{\lambda^2/2}
\end{aligned}
$$

What about functions of random variables?

EMORY UNIVERSITY

- We can apply the same principle to functions f of independent RV X_i
- $f(X_1, \ldots, X_n)$ concentrates around $\mathbb{E}[f(X_1, \ldots, X_n)]$.

Theorem (McDiarmid's inequality)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference condition: there exist constants $c_1, \ldots, c_n \in \mathbb{R}$ such that for all real numbers $x_1, \ldots, x_n$ and x_i',

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \le c_i.$$

(Intuitively, this tells us that f is not overly sensitive to arbitrary changes in a single coordinate.) Then, for any independent random variables $X_1, \ldots, X_n$,

$$\Pr[f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \ge t] \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

- Moreover, $f(X_1, \ldots, X_n)$ is $O\left(\sqrt{\sum_{i=1}^n c_i^2}\right)$-sub-Gaussian.

McDiarmid's inequality

Is this connected with previous concepts?

Remark

McDiarmid's inequality is a generalization of Hoeffding's inequality with $m_i \leq x_i \leq M_i$ *and*

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$$

Lipschitz Functions of Gaussian Variables

EMORY UNIVERSITY

Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the ℓ_2-norm if there exists a non-negative constant $L \in \mathbb{R}$ such that for all $x, y \in \mathbb{R}^n$,

$$|f(x) - f(y)| \leq L\|x - y\|_2.$$

Theorem (Sub-Gaussianity of Lipschitz functions)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to Euclidean distance, and let $X = (X_1, \ldots, X_n)$, where $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0, 1)$. Then for all $t \in \mathbb{R}$,

$$\Pr[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{t^2}{2L^2}\right).$$

It guarantees that any L-Lipschitz function of a standard normal, *regardless of the dimension*, exhibits concentration like a scalar Normal variable with variance L^2.

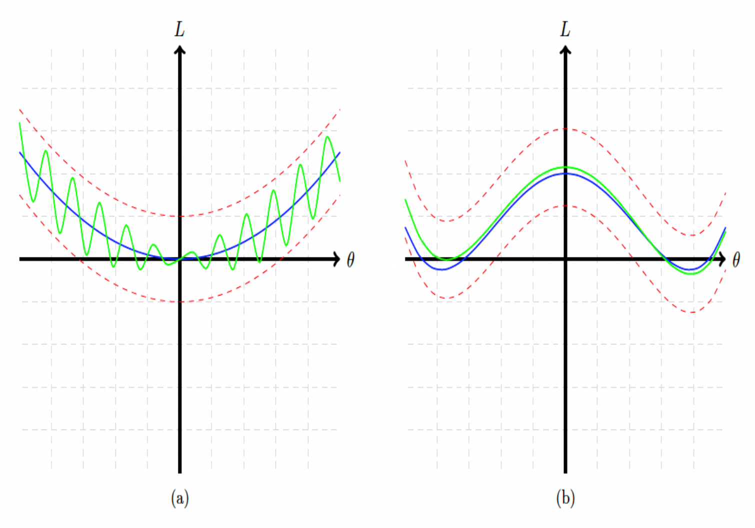
Bounds Via Uniform Convergence

- A central goal in ML theory is to bound the *excess risk* $L(\hat{\theta}) - L(\theta^*)$
- Uniform convergence is a property of a parameter set Θ, which gives us bounds of the form

$$\Pr[|\hat{L}(\theta) - L(\theta)| \geq \varepsilon] \leq \delta; , \forall\theta \in \Theta$$

- How we can do it? Concentration inequalities!
- We can use *union-bound inequality* and *Hoeffding's inequality*.

From Uniform Convergence to Error Bounds



High-Dimension, Sparsity and the Lasso

Model selection via Lasso

- Model selection and parsimony among covariates have a particular echo in statistics and econometrics
- In empirical work, applied researchers often select variables by trial and error.
- A popular machinery to perform *variable selection* is the *Lasso* (Tibshirani 1994).
- Denote $L(\beta) = n^{-1} \sum_{i=1}^{n} \left(Y_i - X_i'\beta\right)^2$ the mean-square loss function.
- The lasso estimator is given by

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} L(\beta) + \lambda_n \|\beta\|_1$$

- λ_n sets the trade-off between fit and sparsity.

Strong Assumptions

EMORY UNIVERSITY

Assumption (Sparsity in Normal Linear Model)

Let the iid sequence of random variables $(Y_i, X_i)_{i=1}^n$. *The dimension of the vector* X_i *is denoted* p *and is assumed to be larger than 1 and allowed to be* $p > n$. *We assume the following linear relation:*

$$Y_i = X_i'\beta_0 + \varepsilon_i$$

with $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right), \varepsilon_i \perp X_i, \sum_{j=1}^p \mathbf{1}\left\{\beta_j \neq 0\right\} \leq s < p$. *The covariates are bounded almost surely* $\max_{i=1,\ldots,n} \|X_i\|_\infty \leq M$

Strong Assumptions

- As we see in our introductory example, when $p > n$, $\widehat{\Sigma}$ can be degenerated in the sense that is *no positive definite*.
- We need to *restrict the eigenvalues*: all square sub-matrices contained in the empirical Gram matrix of dimension no larger than s should have a positive minimal eigenvalue.
- For a non-empty subset $S \subset \{1, \ldots, p\}$ and $\alpha > 0$, define the set:

$$\mathcal{C}[S, \alpha] := \{v \in \mathbb{R}^p : \|v_{S^c}\|_1 \leq \alpha \|v_S\|_1, v \neq 0\}$$

Assumption (Restricted Eigenvalues)

Let $\widehat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i'$*, the empirical Gram matrix, which satisfies*

$$\kappa_\alpha^2(\widehat{\Sigma}) := \min_{\substack{S \subset \{1,\ldots,p\} \\ |S| \leq s}} \min_{\delta \in \mathcal{C}[S,\alpha]} \frac{\delta' \widehat{\Sigma} \delta}{\|\delta_S\|_2^2} > 0$$

Lemma (Concentration Inequality for Gaussian RV)

Consider gaussian random variables such that for $j = 1, \ldots, p, \xi_j \sim \mathcal{N}\left(0, \sigma_j^2\right)$ *and set* $L = \max_{j=1,\ldots,p} \sigma_j$ *Then:*

$$\mathbb{E}\left[\max_{j=1,\ldots,p} |\xi_j|\right] \leq L\sqrt{2\log(2p)}$$

sketch of the proof: Use the fact that ξ is sub-Gaussian, some algebra and Jensen Inequality :)

Consistency of Lasso

EMORY UNIVERSITY

Theorem

Under previous strong assumptions and restricted eigenvalue condition with $\mathcal{C}\left[S_0, 3\right]$*, the Lasso estimator with tuning parameter* $\lambda_n = (4\sigma M/\alpha)\sqrt{2\log(2p)/n}$*, where* $\alpha \in (0,1)$*, verifies with probability greater than* $1-\alpha$ *:*

$$\left\| \widehat{\beta} - \beta_0 \right\|_1 \leq \frac{4^2 \sigma M}{\alpha \kappa_3^2(\widehat{\Sigma})} \sqrt{\frac{2s^2 \log(2p)}{n}}$$

Key takeaway: Lasso converges in ℓ_1 to the true value β_0 at rate $s\sqrt{\log(p)/n}$. The rate of convergence of OLS under full knowledge of sparsity is $s/\sqrt{n}$. Therefore, there is a ”price” to pay for ignorance which manifests itself by this $\sqrt{\log(p)}$ term.

sketch of the proof:

1. Since $\widehat{\beta}$ is a solution of the minimization problem

$$L(\widehat{\beta}) + \lambda_n \|\widehat{\beta}\|_1 \leq L(\beta_0) + \lambda_n \|\beta_0\|_1$$

2. Concentration Inequality for Gaussian RV + Markov's Inequality
3. Separate $\beta = \beta_{S_0} + \beta_{S_0^c}$
4. Use the restricted eigenvalue of the Gram matrix and Cauchy-Schwarz inequality to get

$$\left(\widehat{\beta} - \beta_0\right)' \widehat{\Sigma} \left(\widehat{\beta} - \beta_0\right) \geq \kappa_3^2(\widehat{\Sigma}) \left\|\beta_{0,S_0} - \widehat{\beta}_{S_0}\right\|_2^2 \geq \kappa_3^2(\widehat{\Sigma}) \frac{\left\|\beta_{0,S_0} - \widehat{\beta}_{S_0}\right\|_1^2}{s}$$

5. After some algebra get that with probability greater than $1 - \alpha$:

$$\left\|\beta_0 - \widehat{\beta}\right\|_1 \leq \frac{4^2 \sigma M}{\alpha \kappa_3^2(\widehat{\Sigma})} \sqrt{\frac{2s^2 \log(2p)}{n}}$$

Thanks!

marcelo.ortiz@emory.edu
marcelortiz.com
@marcelortizv

References

- Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.
- Gaillac, C., & L'Hour, J. (2021). *Machine Learning for Econometrics. ENSAE Paris – IP Paris*.
- Ma, T. (2022). *Lecture Notes for Machine Learning Theory (CS229M/STATS214)*.