High-Dimensional Statistics

Modern Inference

Aldo Solari

Outline

Classical versus high-dimensional theory

- What is meant by the term "high-dimensional", and why is it important and interesting to study high-dimensional problems?
- In order to answer these questions, we first need to understand the distinction between classical as opposed to high-dimensional theory

Reference
Wainwright (2019)
High-Dimensional Statistics: A Non-Asymptotic Viewpoint
Cambridge University Press
Chapter 1

Classical theory

- It concerns the behaviour when the *sample size n* goes to ∞
- Suppose $\{X_i\}_{i=1}^n$ i.i.d. as $X \in \mathbb{R}^m$ with mean $\mu = \mathbb{E}(X)$ and finite variance $\Sigma = \mathbb{V}\mathrm{ar}(X)$
- *Law of large numbers*: the sample mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ converges in probability to μ
- *Central limit theorem*: the rescaled deviation $\sqrt{n}(\hat{\mu}_n - \mu)$ converges in distribution to a centered Gaussian with covariance matrix Σ
- *Consistency of maximum likelihood estimation*

Question

Suppose that we are given $n = 1000$ samples from a statistical model in $m = 500$ dimensions

Will theory that requires $n \to \infty$ with the dimension m remaining fixed provide useful predictions?

High-dimensional data

- The data sets arising in many parts of modern science have a "high-dimensional flavor", with m on the same order as, or possibly larger than n

$$m \geq n$$

- Classical "large n, fixed m" theory fails to provide useful predictions
- Classical methods can break down dramatically in high-dimensional regimes

Outline

Classification problem

- Assumptions:
  - $\mathrm{P}(X|Y = A) = N(\mu_A, \Sigma)$
  - $\mathrm{P}(X|Y = B) = N(\mu_B, \Sigma)$
  - $\mathrm{P}(Y = A) = \mathrm{P}(Y = B) = 1/2$
  - For simplicity, take $\Sigma = I_m$
- Goal: determine whether an observed vector $x = (x_1, \ldots, x_m)^{\mathsf{T}} \in \mathbb{R}^m$ belongs to class A or B

Optimal decision

- Optimal decision rule: thresholding the linear statistic

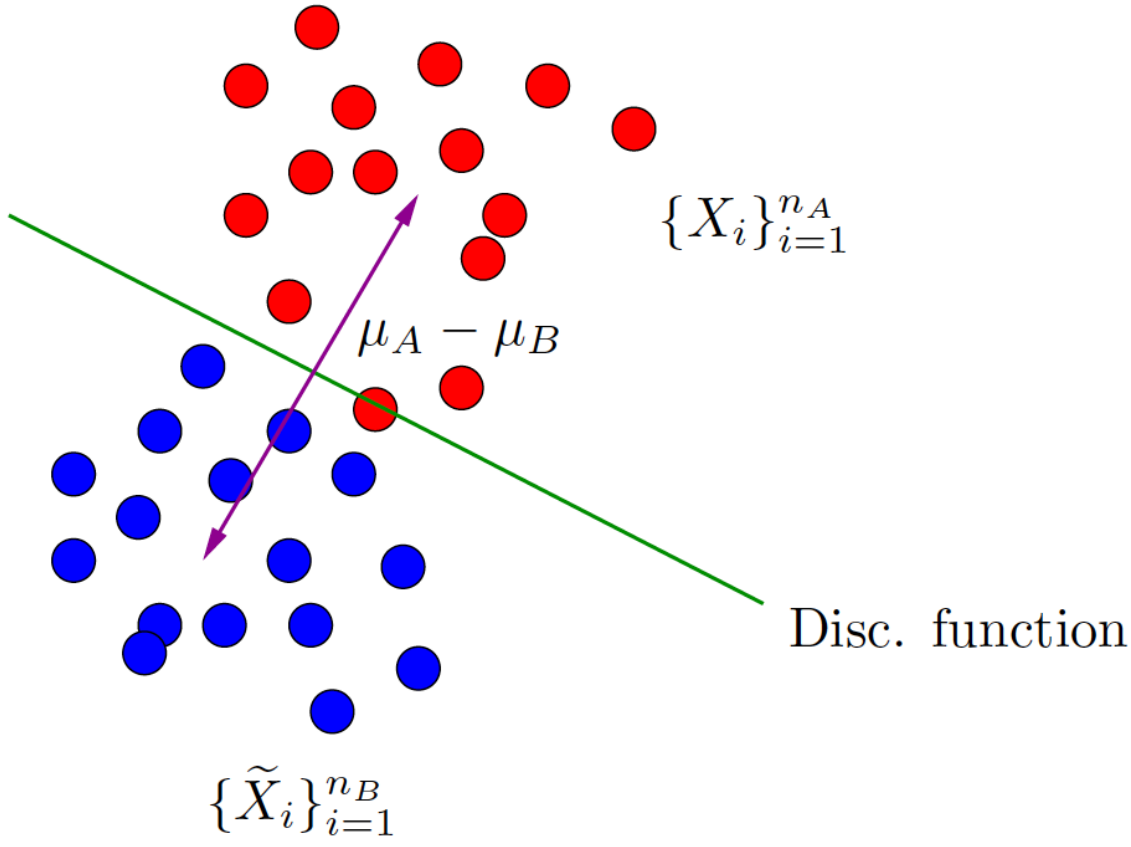
$$\Psi(x) = \langle \mu_A - \mu_B, \Big(x - \frac{\mu_A - \mu_B}{2} \Big) \rangle$$

  where $\langle x, z \rangle = x^{\mathsf{T}} z = \sum_{j=1}^{m} x_j z_j$ denotes the Euclidean inner product in $\mathbb{R}^m$

- If $\Psi(x) > 0$ then classify A, otherwise B
- Error probability of the optimal rule:

$$\mathrm{Err}(\Psi) = \frac{1}{2}\mathrm{P}(\Psi(X_A) < 0) + \frac{1}{2}\mathrm{P}(\Psi(X_B) \geq 0) = \Phi\Big(-\frac{\gamma}{2} \Big)$$

  where $X_A \equiv (X|Y = A)$, $X_B \equiv (X|Y = B)$, $\gamma = \|\mu_A - \mu_B\|_2$, $\|\mu\|_2 = \sqrt{\mu^{\mathsf{T}}\mu}$, and Φ is the cdf of a standard normal variable



$$\langle \mu_A - \mu_B, \left(x - \frac{\mu_A - \mu_B}{2} \right) \rangle = 0$$

source: Wainwright

Linear Discriminant Analysis

- Fisher's LDA: uses the plug-in principle based on n_A samples from class A and n_B samples from class B

$$\hat{\Psi}(x) = \langle \hat{\mu}_A - \hat{\mu}_B, x - \frac{\hat{\mu}_A + \hat{\mu}_B}{2} \rangle$$

- Error probability of LDA (is itself a random variable)

$$\mathrm{Err}(\hat{\Psi}) = \frac{1}{2}\mathrm{P}(\hat{\Psi}(X_A) < 0) + \frac{1}{2}\mathrm{P}(\hat{\Psi}(X_B) \geq 0)$$

- Classical theory: if $(n_A, n_B) \to \infty$ and m remains fixed, then $\hat{\mu}_A \overset{prob.}{\to} \mu_A$, $\hat{\mu}_B \overset{prob.}{\to} \mu_B$ and the asymptotic error probability is $\mathrm{Err}(\hat{\Psi}) \overset{prob.}{\to} \mathrm{Err}(\Psi) = \Phi(-\gamma/2)$

High-Dimensional Theory

- What happens if $(n_A, n_B, m) \to \infty$ with
  - $m/n_A \to \delta$ with $\delta \geq 0$
  - $m/n_B \to \delta$
  - $\|\mu_A - \mu_B\|_2 \to \gamma > 0$
- Kolmogorov (1960) showed that

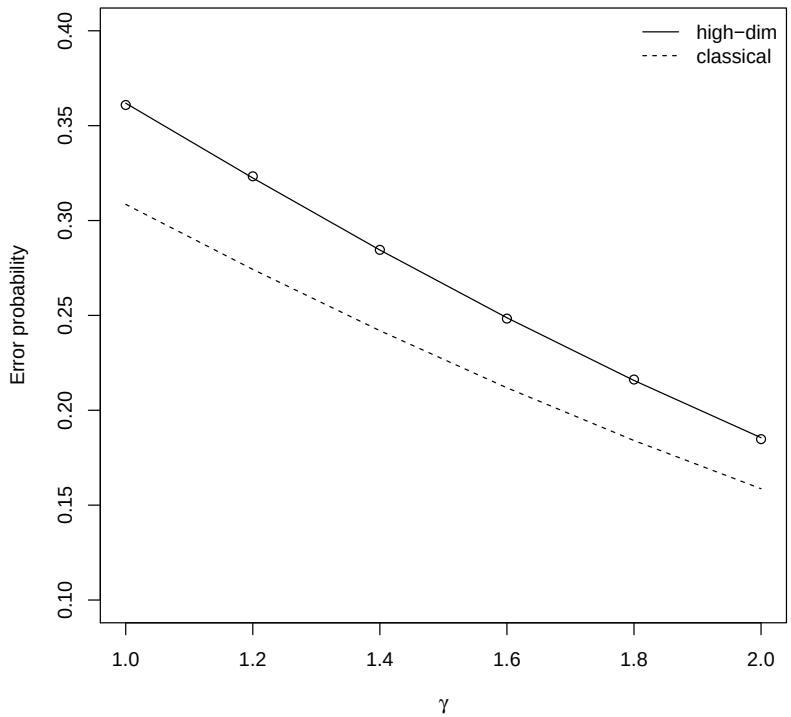
$$\mathrm{Err}(\hat{\Psi}) \overset{prob.}{\to} \Phi\Big(- \frac{\gamma^2}{2\sqrt{\gamma^2 + 2\delta}} \Big)$$

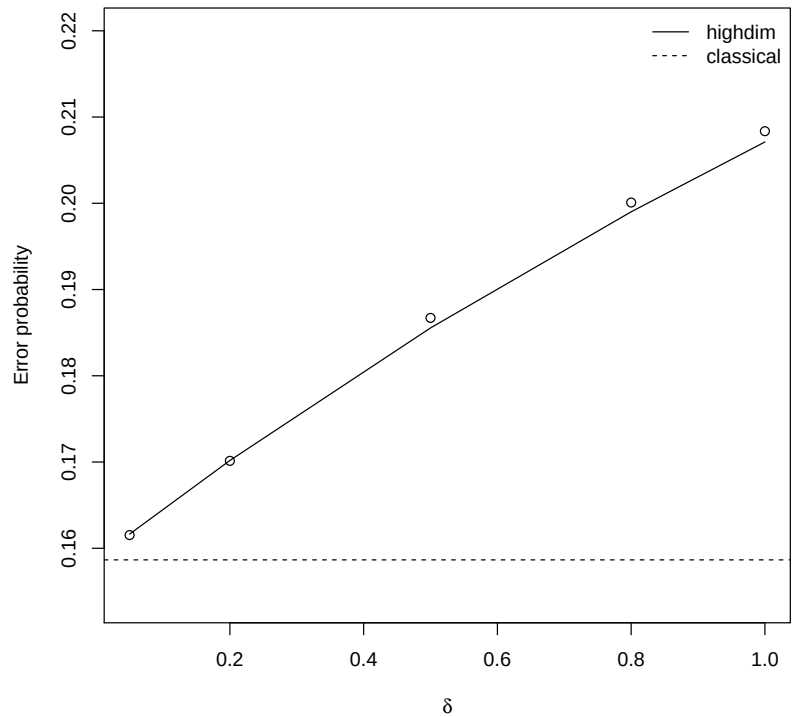
- If $m/n \to 0$, then the asymptotic error probability is $\Phi(-\gamma/2)$ as is predicted by classical theory
- If $m/n \to \delta > 0$, then the asymptotic error probability is strictly larger than $\Phi(-\gamma/2)$

It is natural to ask whether the error probability of $\hat{\Phi}$, for some finite triple

$$(m, n_A, n_B) = (400, 800, 800)$$

is better described by the classical $\Phi(-\gamma/2)$, or the high-dimensional analog $\Phi(-\gamma^2/(2\sqrt{\gamma^2 + 2\delta}))$





circles correspond to the empirical error probabilities, averaged over 10 trials

Outline

What can help us in high dimensions?

- An important fact is that high-dimensional phenomena are unavoidable
- If the ratio m/n stays bounded strictly above zero, then it is not possible to achieve the optimal classification rate
- Our only hope is that the data is endowed with some form of *low-dimensional structure*

- What is the underlying cause of the inaccuracy of the prediction for the LDA in high-dimensions?
- The squared Euclidean error

$$\|\hat{\mu} - \mu\|_2^2 = \sum_{j=1}^{m} (\hat{\mu}_j - \mu_j)^2$$

  turns out to concentrate sharply around $m/n = \delta$, i.e.

$$\mathrm{P}\left(\left|\|\hat{\mu} - \mu\|_2^2 - \frac{m}{n}\right| \geq \frac{m}{n} t\right) = \mathrm{P}\left(\left|\frac{1}{m}\sum Z_j^2 - 1\right| \geq t\right) \leq 2e^{-mt^2/8}$$

  where $\|\hat{\mu} - \mu\|_2^2 = \frac{1}{n}\sum_{j=1}^{m} Z_j^2$ and $Z_j = \sqrt{n}(\hat{\mu}_j - \mu_j) \sim N(0,1)$
- For the last inequality, see Wainwright (2019), Example 2.11

Sparsity

- The simplest form of low-dimensional structure is sparsity of vectors
- Suppose that the m-vector μ is *sparse*, with only s of its m entries being nonzero, for some sparsity parameter $s \ll m$
- In this case, we can obtain a substantially better estimator by applying some form of thresholding to the sample means

$$\tilde{\mu}_j = \hat{\mu}_j \mathbb{1}\{|\hat{\mu}_j| > \lambda\}$$

where

$$\lambda = \sqrt{\frac{2 \log m}{n}}$$

- This choice of λ is motivated by

$$\mathbb{E}\Big[\max_{j=1,\ldots,m} |\hat{\mu}_j - \mu_j|\Big] = \mathbb{E}\Big(\max_{j=1,\ldots,m} \frac{|Z_j|}{\sqrt{n}}\Big) \leq \sqrt{\frac{2 \log m}{n}} + \frac{4}{\sqrt{2 \log m}}$$

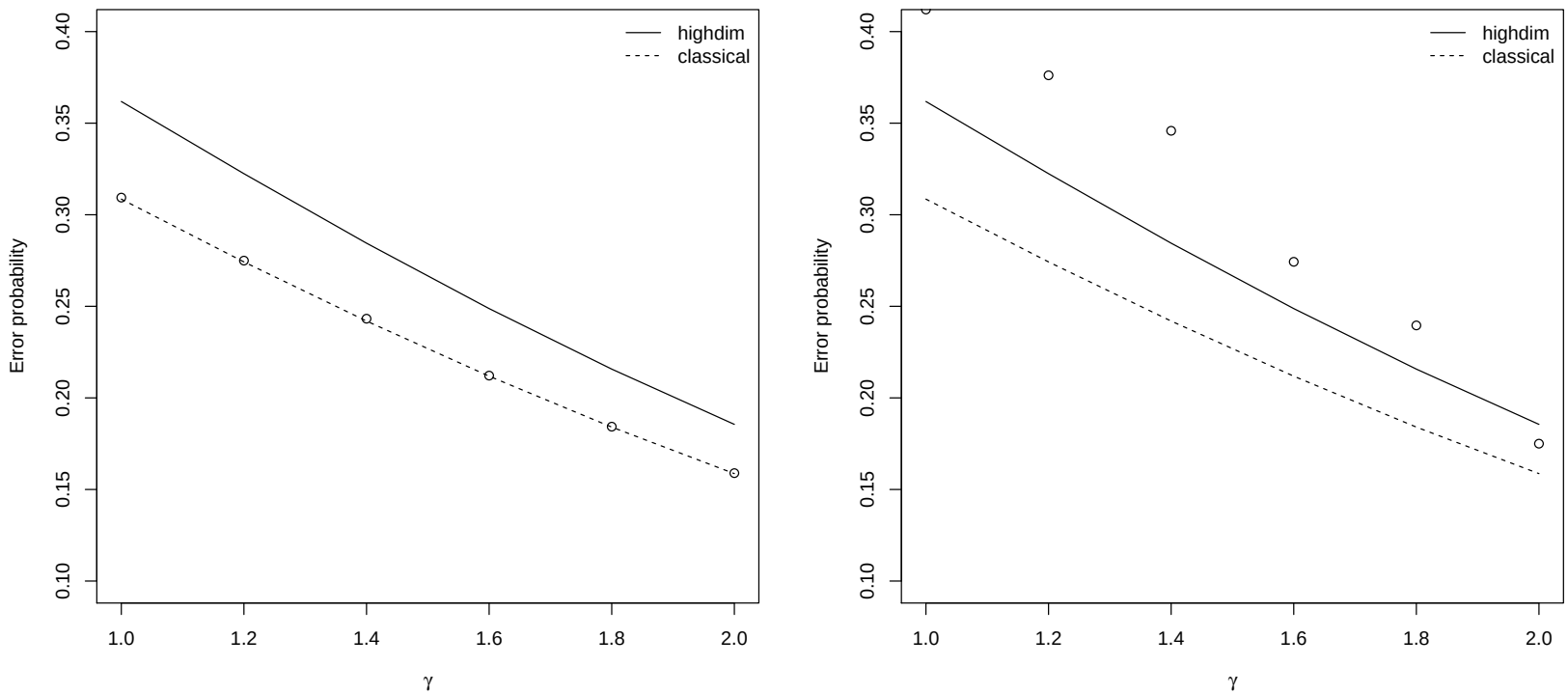
See Wainwright (2019), Examples 2.11, 2.12

Thresholded mean

Suppose that we replace $\hat{\mu}$ by the thresholded mean $\tilde{\mu}$ in LDA (both A and B), then it approaches the optimal classification rate if

$$\log \binom{m}{s} / n \to 0$$

In the previous simulation with $s = 5$ and $s = 50$



circles correspond to the empirical error probabilities, averaged over 10 trials

Homework 2

Reproduce the 4 Figures (two of page 13 and two of page 18) by computing LDA empirical probability error (averaged over 10 repetitions) and comparing it to what is described by the classical and high-dimensional theories:

- $(m, n_A, n_B) = (400, 800, 800)$ with $\delta = 1/2$ and $\gamma = 1, 1.2, 1.4, 1.6, 1.8, 2$, by using the sample averages $\hat{\mu}_A$ and $\hat{\mu}_B$
- $m = 400$, $n_A = n_B = m/\delta$ with $\gamma = 2$ and $\delta = .05, .2, .5, .8, 1$, by using the sample averages $\hat{\mu}_A$ and $\hat{\mu}_B$
- $(m, n_A, n_B) = (400, 800, 800)$ with $\delta = 1/2$, $\gamma = 1, 1.2, 1.4, 1.6, 1.8, 2$ and sparsity $s = 5$ and 50, by using the thresholded means $\tilde{\mu}_A$ and $\tilde{\mu}_B$ with $\lambda = \sqrt{\frac{2 \log m}{n}}$

The R code should be reproducible.

Outline

Linear model

- Observations $(x_i, y_i) \in \mathbb{R}^m \times \mathbb{R}$ from the linear model

$$y_i = \langle x_i, \beta^* \rangle + \epsilon_i, \qquad i = 1, \ldots, n$$

  where $\beta^* \in \mathbb{R}^m$ is an unknown coefficient vector

- In matrix form

$$y = X\beta^* + \epsilon$$

  where $\underset{n\times 1}{y} = (y_1, \ldots, y_n)^{\mathsf{T}}$ and $\underset{n\times m}{X}$ with ith row $\underset{1\times m}{x_i^{\mathsf{T}}}$

We would like our estimator $\hat{\beta}$ to possess three desirable properties:

- $\hat{\beta} \approx \beta^*$
- $X\hat{\beta} \approx X\beta^*$
- $\{j : \hat{\beta}_j = 0\} \approx \{j : \beta^*_j = 0\}$

Which is the easiest? The most difficult?

Error metrics

(In-sample) Prediction loss

$$\mathrm{MSPE} = \frac{1}{n}\|X\hat{\beta} - X\beta^*\|_2^2$$

Parameter estimation

$$\|\hat{\beta} - \beta^*\|_2^2$$

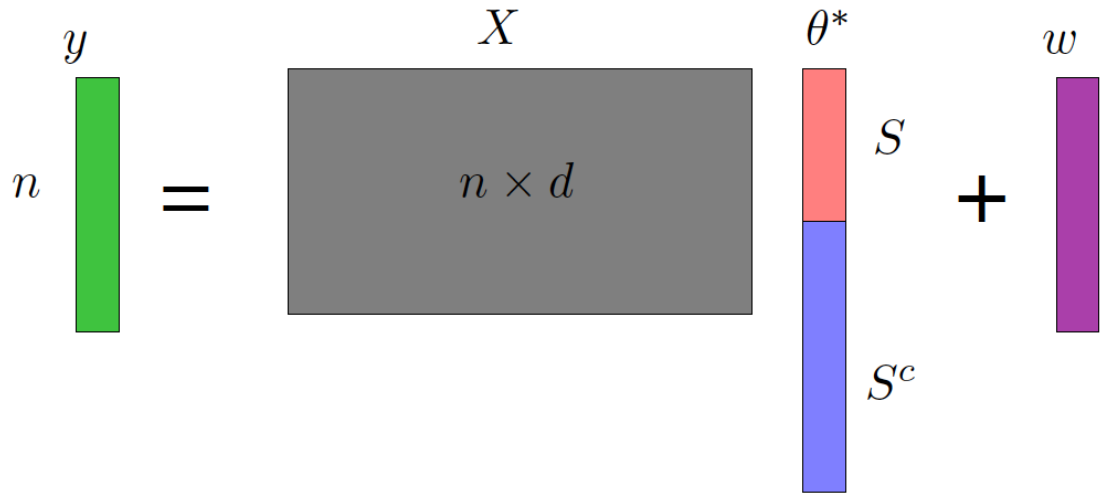
Note that $\hat{\beta} \approx \beta^* \Rightarrow X\hat{\beta} \approx X\beta^*$ but not the other way around

Variable selection

- *screening property* : $\{j : \hat{\beta}_j \neq 0\} \supseteq \{j : \beta_j^* \neq 0\}$
- *sign consistency* : $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)$

Sparsity

- If the model lacks any low-dimensional structure, then there is no hope of obtaining consistent estimators when the ratio m/n stays bounded away from zero
- *sparsity assumption*: the support set $\{j : \beta_j^* \neq 0\}$ has cardinality s substantially smaller than m



- What is the behaviour of the Lasso estimator?

$$\hat{\beta}(\lambda) = \arg\min_{\beta \in \mathbb{R}^m} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

Theoretical analysis of the lasso

Asymptotic

- Variable selection: sign consistency
- Variable selection: screening property
- Parameter estimation: ℓ_2-consistency

Non-asymptotic

- Prediction loss: MSPE bound

Assumptions

Basic assumptions

- Linear model with X fixed
- $\epsilon_1, \ldots, \epsilon_n$ i.i.d. $N(0, \sigma^2)$
- Columns of X are such that $\{n^{-1}X^{\mathsf{T}}X\}_{jj} = 1$ for all j

Additional assumptions

- β^* is sparse with s nonzero coefficients
- β^* satisfies the *beta-min assumption*
- X satisfies the *compatibility condition*
- X satisfies the *irrepresentable condition*
- ...

Sign consistency

- An estimator $\hat{\beta}$ is sign consistent if and only if

$$\mathrm{P}(\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)) \to 1 \qquad \text{as } n \to \infty$$

- The Lasso is sign consistent only if the irrepresentable condition holds. If this condition is not fulfilled, then there exists no sequence λ_n such that the Lasso estimator $\hat{\beta}$ is sign consistent
- This sheds a little doubt on whether the Lasso is a good method for identification of sparse models for both low- and high-dimensional data

Screening property

- Suppose β^* is s-sparse and X satisfies the compatibility condition for some ϕ^2. Let $\lambda \asymp \sqrt{\frac{\log m}{n}}$, then with high probability

$$\|\hat{\beta} - \beta^*\|_1 \leq \frac{4s\lambda}{\phi^2}$$

- If β^* also satisfies the beta-min condition with $\beta_{\min} > 4s\lambda/\phi^2$, then

$$\{j : \hat{\beta}_j \neq 0\} \supseteq \{j : \beta_j^* \neq 0\}$$

  with high probability
- Thus, the screening property holds when assuming the compatibility and beta-min condition

Estimation ℓ_2-consistency

- An estimator is said to be ℓ_2-consistent if

$$\|\hat{\beta} - \beta\|_2 \to 0 \qquad \text{as } n \to \infty$$

- Under specific conditions (sparse minimal eigenvalues are not decaying too fast in some sense), the requirement for ℓ_2-consistency of the lasso is

$$\frac{s_n \log m_n}{n} \to 0 \qquad \text{as } n \to \infty$$

e.g. s_n growing like $n^{1/3}$ and m_n growing linearly with s_n

Prediction bound

Theorem

For some $A > 1$ and all $\lambda > A \cdot \sqrt{\frac{8 \log(2m)\sigma^2}{n}}$ we have with probability $1 - A^{-1}$:

$$\frac{1}{n}\|X\hat{\beta} - X\beta^*\|_2^2 \leq 2\lambda\|\beta^*\|_1$$

How do we get consistency out of this? We need both of these things to go to zero as n goes to infinity:

- $A^{-1} \to 0$
- $A \cdot \sqrt{\frac{8 \log(2m)\sigma^2}{n}}\|\beta^*\|_1 \to 0$

Assuming that σ^2 and the size of the nonzero β^* stay fixed, we have consistency as long as m is dominated by e^n