High-Dimensional Statistics & Sparsity

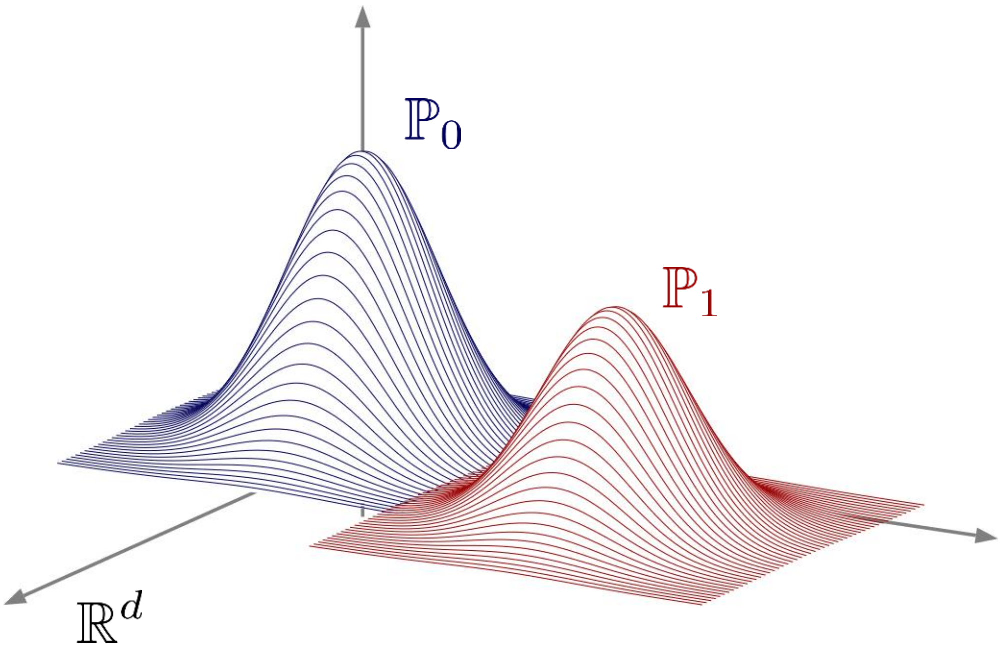
UDRC Summer School

João F. C. Mota

Heriot-Watt University, Edinburgh, UK

Motivation

Hypothesis testing in high-dimensions



Problem

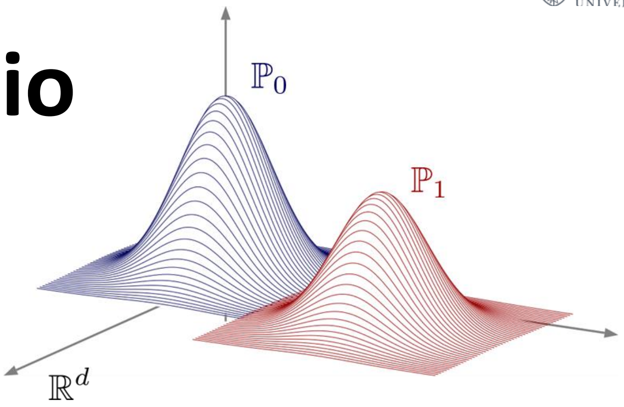
Observe a random vector $X \in \mathbb{R}^d$

$X \sim \mathbb{P}_0$ or $X \sim \mathbb{P}_1$?

False positive: $\alpha = \mathbb{P}\big(\text{decide } X \sim \mathbb{P}_1 \,\big|\, X \sim \mathbb{P}_0\big)$

False negative: $\beta = \mathbb{P}\big(\text{decide } X \sim \mathbb{P}_0 \,\big|\, X \sim \mathbb{P}_1\big)$

Decision Rule: Likelihood Ratio



$x_1, \ldots, x_n$: i.i.d. realizations of X

For a given $T \geq 0$,

If $\dfrac{\mathbb{P}_1(x_1, \ldots, x_n)}{\mathbb{P}_0(x_1, \ldots, x_n)} > T$, then decide $X \sim \mathbb{P}_1$

If $\dfrac{\mathbb{P}_1(x_1, \ldots, x_n)}{\mathbb{P}_0(x_1, \ldots, x_n)} \leq T$, then decide $X \sim \mathbb{P}_0$

MAP rule / minimizes risk when

$$T = \frac{\mathbb{P}(X \sim \mathbb{P}_0)}{\mathbb{P}(X \sim \mathbb{P}_1)}$$

$$\frac{\mathbb{P}_1(x_1, \ldots, x_n)}{\mathbb{P}_0(x_1, \ldots, x_n)}$$

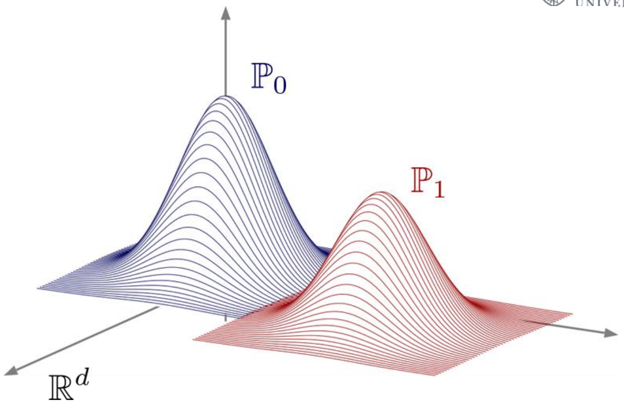
False positive: $\alpha_L = \mathbb{P}\big(L(x_1, \ldots, x_n) > T \,\big|\, X \sim \mathbb{P}_0 \big)$

False negative: $\beta_L = \mathbb{P}\big(L(x_1, \ldots, x_n) \leq T \,\big|\, X \sim \mathbb{P}_1 \big)$

Neyman-Pearson Lemma

$$\alpha_L = \mathbb{P}\big(L(x_1, \dots, x_n) > T \,\big|\, X \sim \mathbb{P}_0 \big)$$

$$\beta_L = \mathbb{P}\big(L(x_1, \dots, x_n) \le T \,\big|\, X \sim \mathbb{P}_1 \big)$$



Neyman-Pearson Lemma

The likelihood ratio test is *optimal:*

If there is another (possibly random) decision rule $D(x_1, \dots, x_n)$ such that

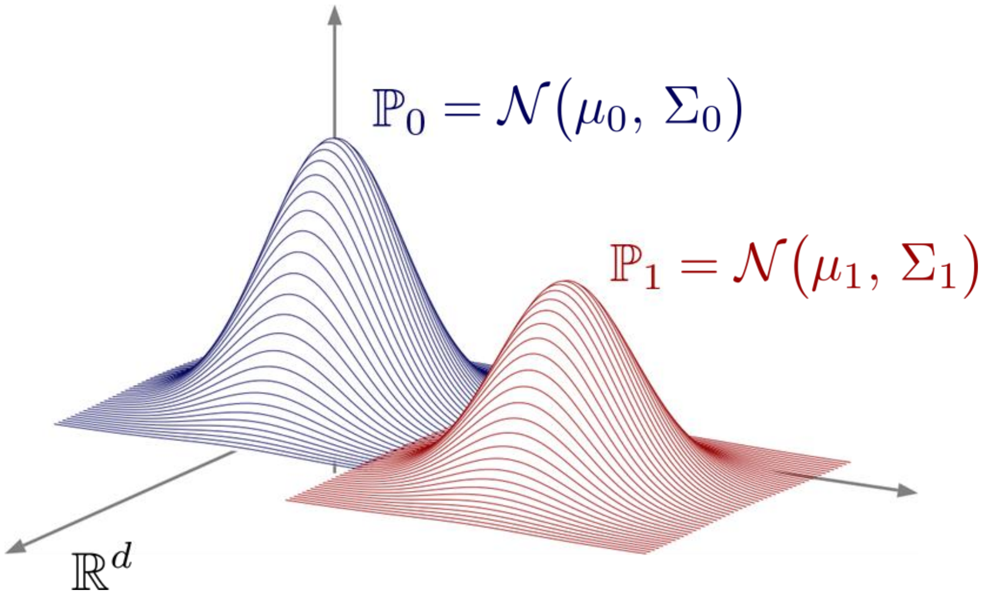
$$\mathbb{P}\big(D(x_1, \dots, x_n) \text{ decides } X \sim \mathbb{P}_1 \,\big|\, X \sim \mathbb{P}_0 \big) \le \alpha_L \,,$$

then

$$\mathbb{P}\big(D(x_1, \dots, x_n) \text{ decides } X \sim \mathbb{P}_0 \,\big|\, X \sim \mathbb{P}_1 \big) \ge \beta_L \,.$$

And vice-versa.

Linear Discriminant Analysis



$$X \sim \mathcal{N}(\mu, \Sigma) \quad \Longrightarrow \quad f_X(x) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big)$$

Assume (to simplify): $\Sigma_0 = \Sigma_1 = \Sigma$ and $n = 1$ (one observation)

$$L(x) > T \quad \Longleftrightarrow \quad \Psi(x) := \left(x - \frac{\mu_0 + \mu_1}{2}\right)^\top \Sigma^{-1}(\mu_1 - \mu_0) \ > \ \log T$$

$$\Psi(x) := \left(x - \frac{\mu_0 + \mu_1}{2}\right)^\top \Sigma^{-1}(\mu_1 - \mu_0) \; > \; \log T$$

Probability of error (assuming $\mathbb{P}_0$ and $\mathbb{P}_1$ are equally likely)

$$\begin{aligned}
\mathrm{Err}(\Psi) &= \mathbb{P}\big(\Psi(X) > 0 \;\&\; \mathbb{P}_0 \text{ true}\big) \;+\; \mathbb{P}\big(\Psi(X) \le 0 \;\&\; \mathbb{P}_1 \text{ true}\big) \\
&= \mathbb{P}\big(\Psi(X) > 0 \mid \mathbb{P}_0\big) \cdot \mathbb{P}\big(\mathbb{P}_0\big) \;+\; \mathbb{P}\big(\Psi(X) \le 0 \mid \mathbb{P}_1\big) \cdot \mathbb{P}\big(\mathbb{P}_1\big) \\
&= \frac{1}{2}\,\mathbb{P}\big(\Psi(X) > 0 \mid \mathbb{P}_0\big) \;+\; \frac{1}{2}\,\mathbb{P}\big(\Psi(X) \le 0 \mid \mathbb{P}_1\big)
\end{aligned}$$

(using Gaussianity and manipulating...)

$$= \boxed{\Phi\left(-\frac{\gamma}{2}\right) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\gamma/2} e^{-t^2/2}\, dt}$$

classical error expression

$$\gamma = \sqrt{(\mu_0 - \mu_1)^\top \Sigma^{-1} (\mu_0 - \mu_1)}$$

need to be estimated

n_0 and n_1 samples

<u>High-dimensional regime</u>: n_0 and n_1 same order as d

Fisher Linear Discriminant

$$\Psi(x) = \left(x - \frac{\mu_0 + \mu_1}{2}\right)^\top \Sigma^{-1}(\mu_1 - \mu_0)$$

$$\gamma = \sqrt{(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1)} \qquad \mathrm{Err}(\Psi) = \Phi\Big(-\frac{\gamma}{2}\Big)$$

Unbiased estimators:

$$\widehat{\mu}_0 := \frac{1}{n_0}\sum_{i=1}^{n_0} x_i \qquad \widehat{\mu}_1 := \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$$

$$\widehat{\Sigma} := \frac{1}{n_0 - 1}\sum_{i=1}^{n_0}(x_i - \widehat{\mu}_0)(x_i - \widehat{\mu}_0)^\top + \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_i - \widehat{\mu}_1)(y_i - \widehat{\mu}_1)^\top$$

Plug estimators into log-likelihood ratio:

$$\widehat{\Psi}(x) := \left(x - \frac{\widehat{\mu}_0 + \widehat{\mu}_1}{2}\right)^\top \widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_0)$$

Fisher linear discriminant function

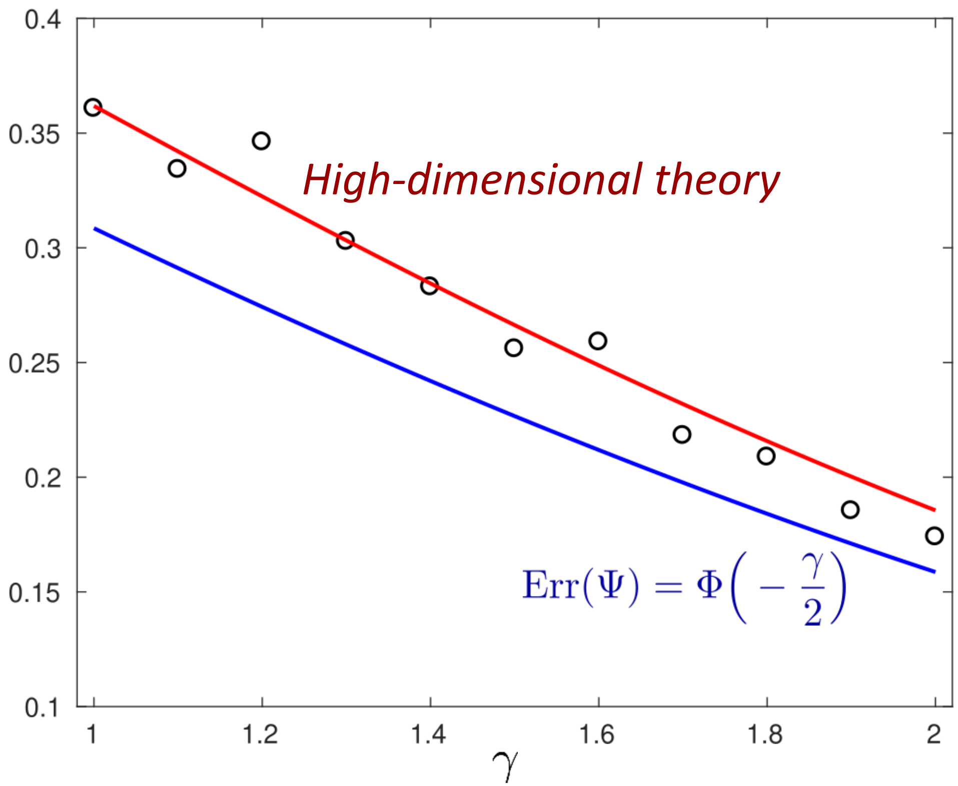
Fisher Linear Discriminant

Assume $\mathbb{P}(\mathbb{P}_0) = \mathbb{P}(\mathbb{P}_1)$ $\quad \Sigma = I_d$ $\quad d = 400$ $\quad n_0 = n_1 = 800$

Vary $\gamma = \|\mu_0 - \mu_1\|_2$ between 1 and 2

Test with $\widehat{\Psi}(x)$ over 5000 random trials

Probability of error



[Kolmogorov]

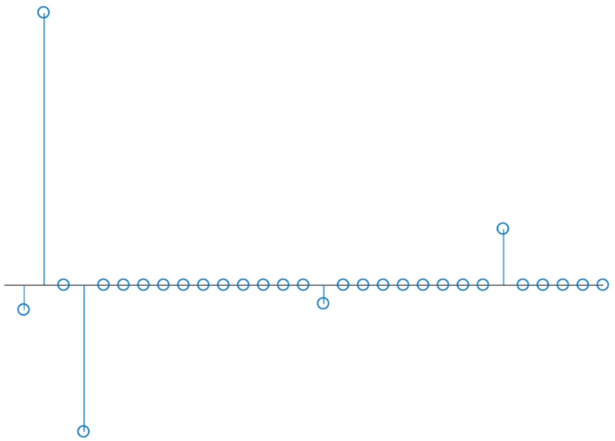
$$(d, n_0, n_1) \to \infty \qquad \frac{d}{n_0}, \frac{d}{n_1} \to \alpha$$

$$\mathrm{Err}(\widehat{\Psi}) \xrightarrow{\text{prob.}} \Phi\left(-\frac{\gamma^2}{2\sqrt{\gamma^2 + 2\alpha}}\right)$$

What can help in high-dimensions?

Structure e.g., sparsity

Suppose μ_0 and μ_1 are sparse: only have $s \ll d$ nonzero entries



Procedure: *hard-threshold* entries of estimates

$$\widehat{\mu}_0' = \frac{1}{n_0}\sum_{i=1}^{n_0} x_i \longrightarrow (\widehat{\mu}_0)_i = H_\lambda\Big(\big(\widehat{\mu}_0'\big)_i\Big)$$

(same for μ_1)

hard-thresholding operator

$$H_\lambda(x) := \begin{cases} x & , \text{ if } |x| > \lambda \\ 0 & , \text{ if } |x| \le \lambda \end{cases}$$

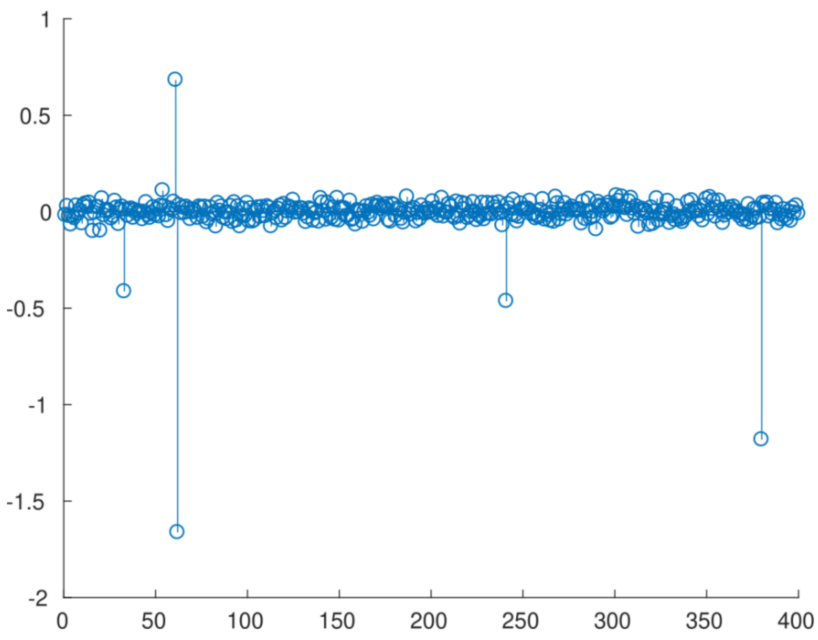
Example

hard-thresholding operator

$$H_\lambda(x) := \begin{cases} x & , \text{ if } |x| > \lambda \\ 0 & , \text{ if } |x| \leq \lambda \end{cases}$$

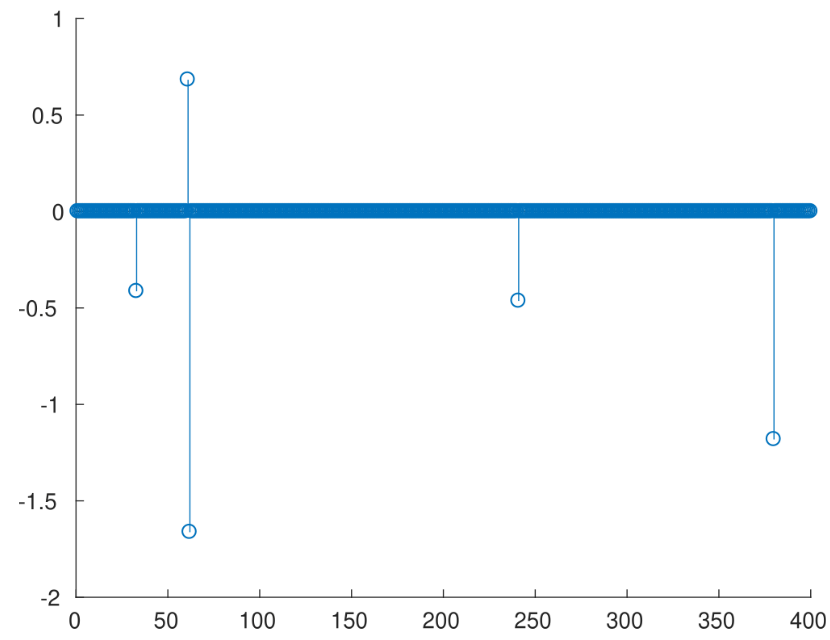
$$d = 400 \qquad n = 800 \qquad s = 5 \qquad \lambda = \sqrt{\frac{2 \log\, d}{n}} = 0.1224$$

$\boldsymbol{\widehat{\mu}'_0}$



$$\xrightarrow{H_\lambda(\cdot)}$$

$\boldsymbol{\widehat{\mu}_0}$



Same Experiments

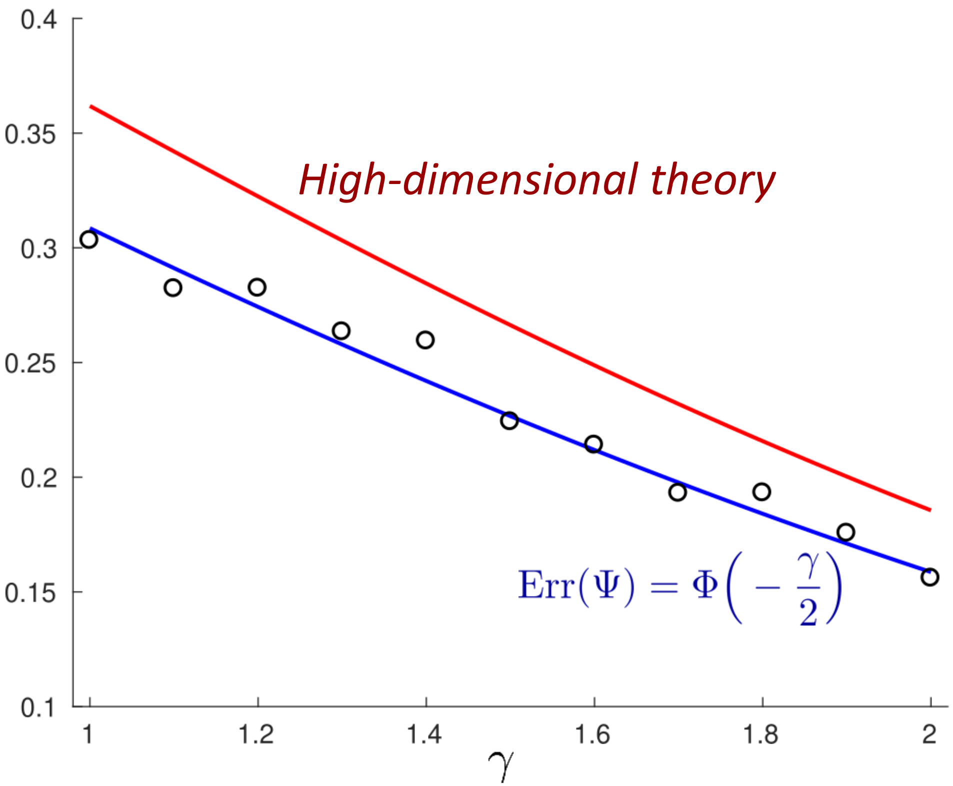
Assume $\mathbb{P}(\mathbb{P}_0) = \mathbb{P}(\mathbb{P}_1)$ $\quad \Sigma = I_d \quad d = 400 \quad n_0 = n_1 = 800$

Vary $\gamma = \|\mu_0 - \mu_1\|_2$ between 1 and 2

Test with $\widehat{\Psi}(x)$ over 5000 random trials

$s = 5 \qquad \lambda = \sqrt{\frac{2\log d}{n}}$

Probability of error



Sparsity makes problem low-dimensional

Outline

- *Motivation: Hypothesis Testing in High-Dimensions*
- Introduction to LASSO and other sparsity problems
- Gaussian graphical model selection
- Matrix completion

A Crime Problem

City	Police funding/ resident ($/ya)	% with 4 years high-school (+25)	% not in high-school (16-19)	% in college (18-24)	% with 4+ years college (25+)	Crimes/million
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
6	25	68	8	32	15	603
7	34	68	12	24	14	484
8	33	62	13	28	11	546
9	36	69	7	25	12	424
⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	66	67	26	18	16	940

Goal: Predict *# crimes / million* based on the other indicators

Linear Regression

Linear/affine model

$d = 5$ predictors

$n = 50$ samples

n

City	Police funding per resident	% with 4 years high-school (+25)	% in high-school (16-19)	% in college (18-24)	% with 4+ years college (25+)	Crimes/million
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
6	25	68	8	32	15	603
7	34	68	12	24	14	484
8	33	62	13	28	11	546
9	36	69	7	25	12	424
⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	66	67	26	18	16	940

d

$$\boxed{y_i \simeq x_0 + \sum_{j=1}^{d} a_{ij} x_j} \quad i = 1, \ldots, n \qquad \Longleftrightarrow \qquad \begin{aligned} y &\simeq x_0 1_n + Ax \\ &= \underbrace{\begin{bmatrix} 1_n & A \end{bmatrix}}_{\overline{A}} \underbrace{\begin{bmatrix} x_0 \\ x \end{bmatrix}}_{\overline{x}} \end{aligned}$$

offset

coefficient to be determined

response variable (crime rate)

table entry

Find coefficients: *least-squares*

$$\boxed{\underset{\overline{x}}{\text{minimize}} \quad \frac{1}{2} \left\| y - \overline{A}\,\overline{x} \right\|_2^2} \qquad \Longleftrightarrow \qquad \overline{x}_{\text{LS}}^{\star} = \left(\overline{A}^{\top} \overline{A} \right)^{-1} \overline{A}\, y$$

Linear Regression

$$\underset{\overline{x}}{\text{minimize}} \quad \frac{1}{2}\left\|y - \overline{A}\overline{x}\right\|_2^2$$

$$\overline{x}^\star_{\text{LS}} = \left(\overline{A}^\top \overline{A}\right)^{-1} \overline{A}\, y$$

$$= \begin{array}{ll} \begin{bmatrix} 489.6486 \\ 10.9807 \\ -6.0885 \\ 5.4803 \\ 0.3770 \\ 5.5005 \end{bmatrix} & \begin{array}{l} \textit{offset} \\ \textit{police funding / resident (\$/year)} \\ \textit{\% of 25+ year-olds with 4+ years of high-school} \\ \textit{\% of 16-19 year-olds not in high-school} \\ \textit{\% of 18-24 year-olds in college} \\ \textit{\% of 25+ year-olds with 4+ years of college} \end{array} \end{array}$$

Problems with least-squares

little interpretability — all coefficients contribute to prediction

small bias, large variance — zeroing coefficients can improve mean-squared error

LASSO

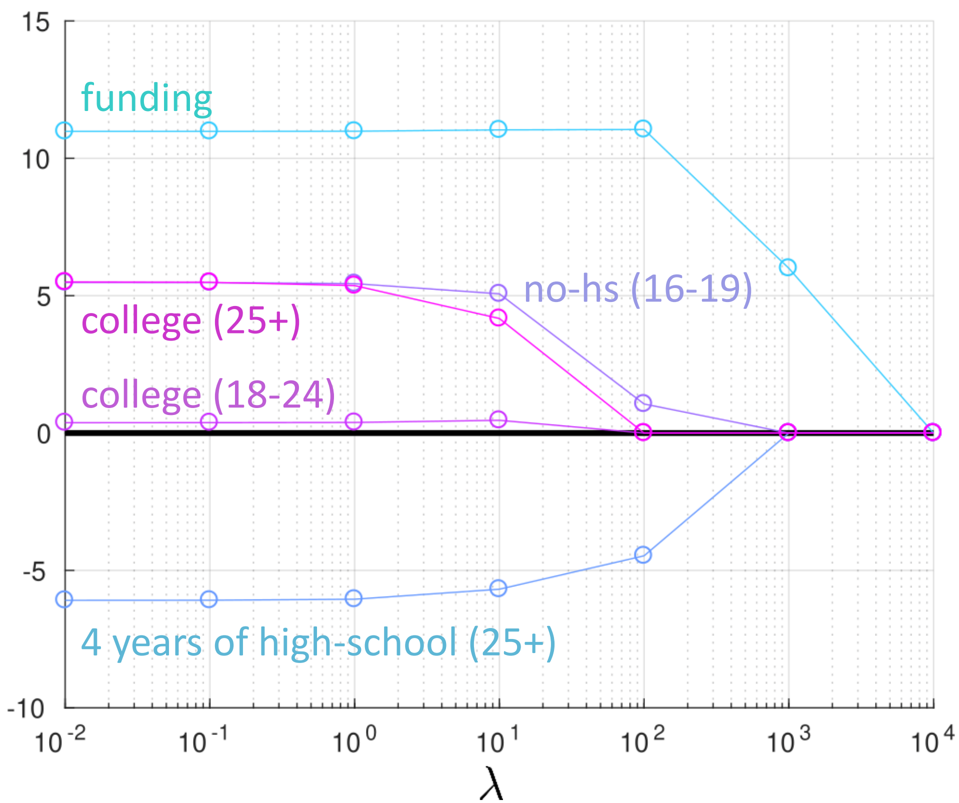
least absolute selection and shrikange operator

$$\underset{x}{\text{minimize}} \quad \frac{1}{2n}\|y - Ax\|_2^2 + \lambda\|x\|_1$$

L1-norm $\|x\|_1 = |x_1| + |x_2| + \cdots + |x_d|$

regularization parameter ≥ 0

Coefficient value



In reality, we solved ...

$$\min_{x_0,\, x} \quad \frac{1}{2n}\|y - x_0 1_n - Ax\|_2^2 + \lambda\|x\|_1$$

necessary because $\displaystyle\frac{1}{n}\sum_{i=1}^{n} y_i \neq 0$

L1-Norm Induces Sparsity

$$\underset{x}{\text{minimize}} \quad \frac{1}{2n}\|y - Ax\|_2^2 + \lambda\|x\|_1$$

LASSO (aka Basis Pursuit Denoising)

$\Updownarrow$ for some τ depending on λ

$$\begin{aligned} &\underset{x}{\text{minimize}} && \|y - Ax\|_2^2 \\ &\text{subject to} && \|x\|_1 \leq \tau \end{aligned}$$

Constrained LASSO

$\Updownarrow$ for some σ depending on τ

$$\begin{aligned} &\underset{x}{\text{minimize}} && \|x\|_1 \\ &\text{subject to} && \|y - Ax\|_2 \leq \sigma \end{aligned}$$

Relaxed Basis Pursuit

Basis Pursuit when $\sigma = 0$

L1-Norm Induces Sparsity

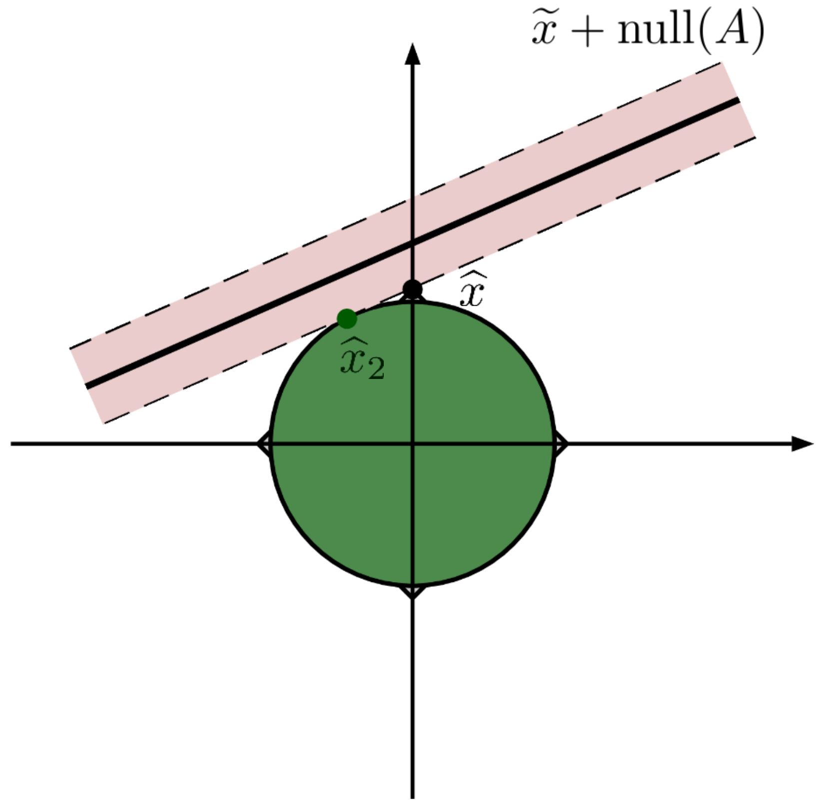
$$\widehat{x} \in \arg\min_x \; \|x\|_1 \quad \text{s.t.} \quad \|y - Ax\|_2 \leq \sigma$$

Assume $\sigma = 0$:

$y = Ax$ has solutions $\widetilde{x} + \mathrm{null}(A)$

$$y = A\widetilde{x} \qquad \{d \, : \, Ad = 0\}$$

Assume $\sigma > 0$: *margin around* $\widetilde{x} + \mathrm{null}(A)$

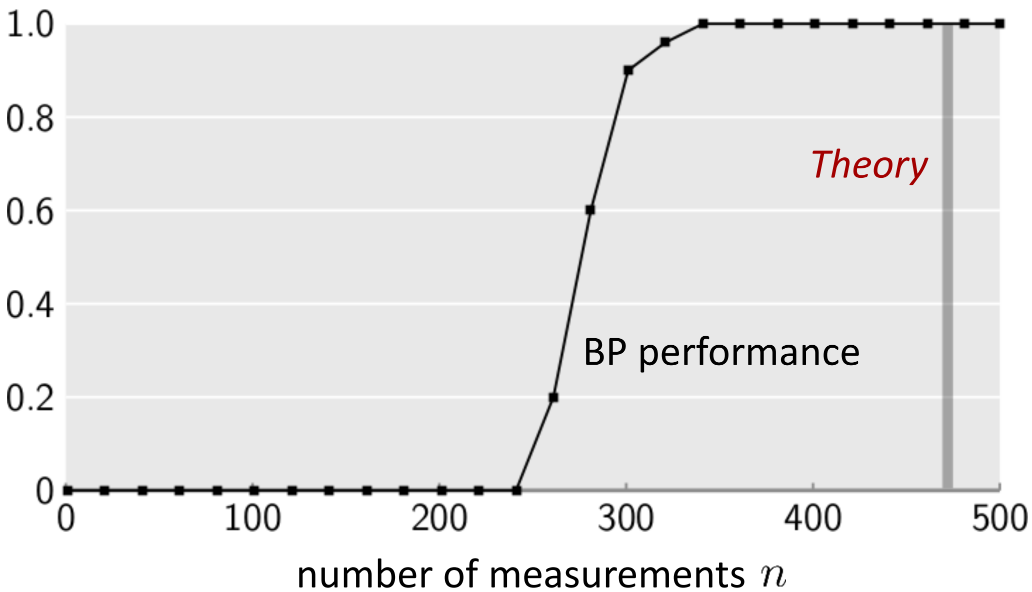


What about the L2-norm?

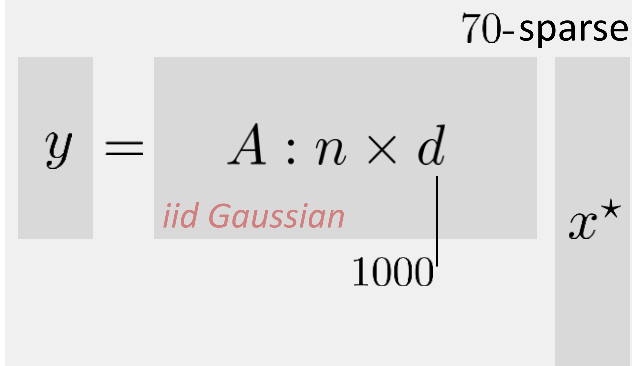
$$\widehat{x}_2 \in \arg\min_x \; \|x\|_2^2 \quad \text{s.t.} \quad \|y - Ax\|_2 \leq \sigma$$

Example: Compressed Sensing

Sucess rate (50 trials)



Compressed Sensing (CS)



Basis Pursuit

$$\widehat{x} = \arg\min_{x} \ \|x\|_1 \quad \text{s.t.} \quad y = Ax$$

Theorem [Chandrasekaran et al. 12']

$x^\star \in \mathbb{R}^d$ *unknown*, but s-sparse

iid entries $\mathcal{N}(0, 1/n)$

$y = Ax^\star$ *measurements*

$$\boxed{n \geq 2s \log\left(\frac{d}{s}\right) + \frac{7}{5}s + 1} \quad \Longrightarrow \quad x^\star = \operatorname*{argmin}_{x} \ \|x\|_1 \quad \text{s.t.} \quad y = Ax \qquad \text{w.h.p.}$$

Application: Image Reconstruction

$256 \times 496 \quad \Longrightarrow \quad z^\star \in \mathbb{R}^{126976}$

not sparse

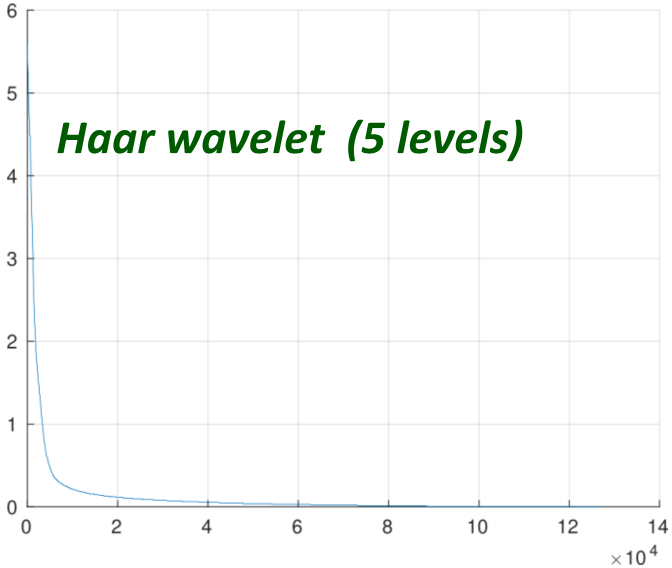
Natural images have sparse representations

$$z^\star = \Psi \, x^\star$$

sparse or near-sparse

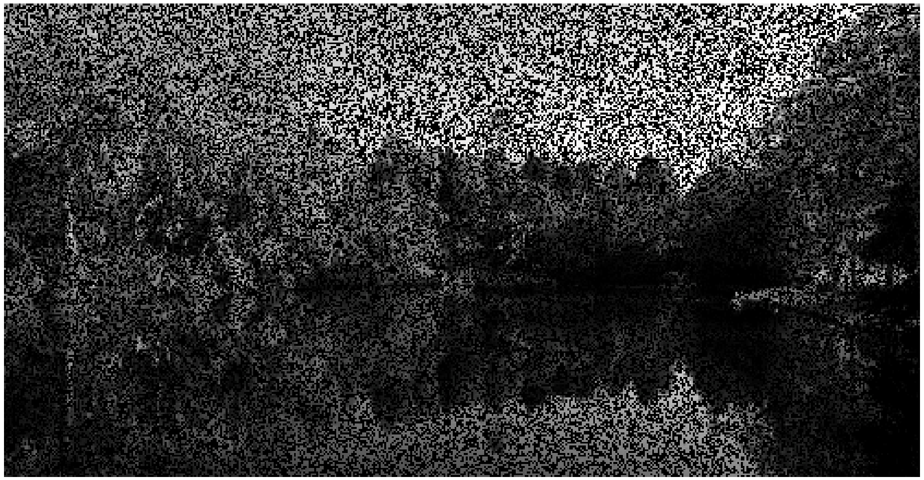
dictionary (wavelet, DCT, gradient space)

Ordered coefficients: $|x_i^\star|$



Application: Image Reconstruction

Suppose we observe *only 50%* of pixels



y

Solve

$$\widehat{x} = \arg\min_{x} \|x\|_1 \quad \text{s.t.} \quad y = \Phi\,\Psi\,x$$

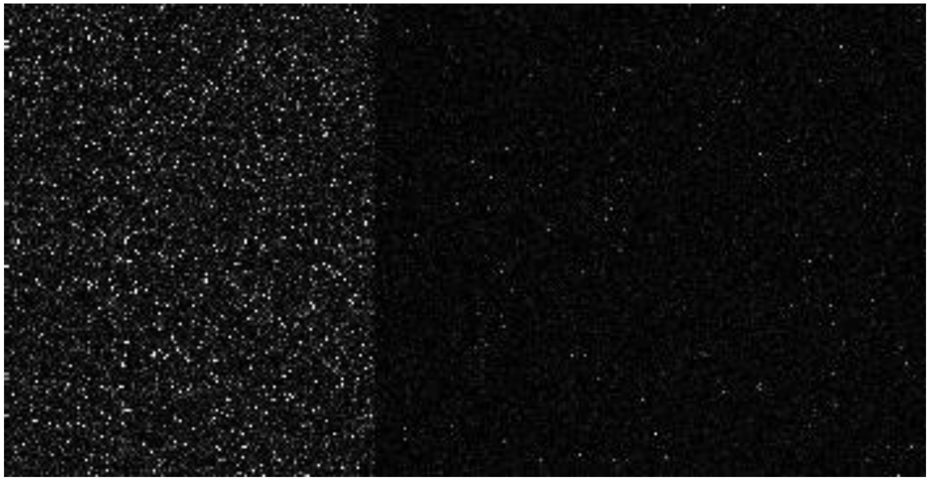
Φ: observed indices

Ψ: wavelet

$\widehat{x}$

PSNR: 21.31 dB

Application: Image Reconstruction



y

Solve

$$\widehat{x} = \arg\min_{x} \; \|x\|_1 \quad \text{s.t.} \quad y = \Phi \, \Psi \, x$$

wavelet

partial DFT

each entry of y has info from entire image

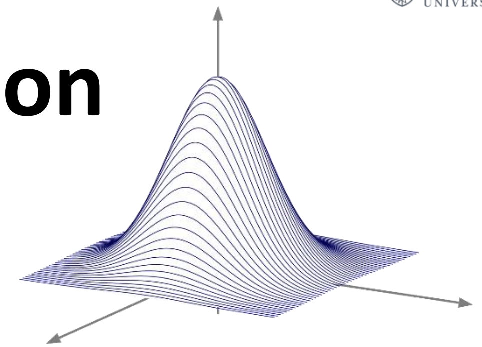
$\widehat{x}$

PSNR: 24.93 dB

Outline

- *Motivation: Hypothesis Testing in High-Dimensions*
- *Introduction to LASSO and other sparsity problems*
- Gaussian graphical model selection
- Matrix completion

Gaussian Graphical Model Selection



$$X = (X_1, X_2, \dots, X_d) \;\sim\; \mathcal{N}(0_d\,,\,\Sigma^\star)$$

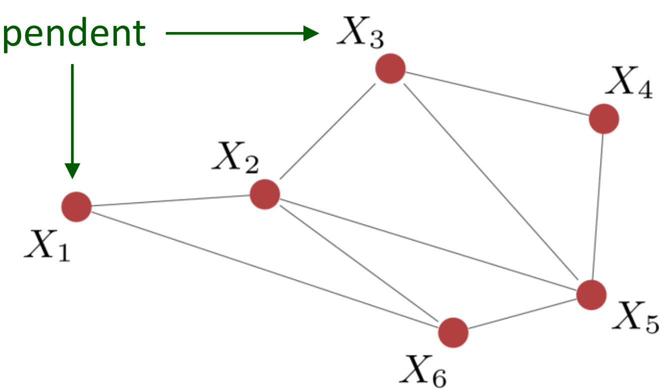
Problem: given n idd observations of X, denoted $X^{(1)}, \dots, X^{(n)}$, estimate $\Sigma^\star$

Assumption: most pairs of coordinates (X_i, X_j) are conditionally independent

$\Downarrow$

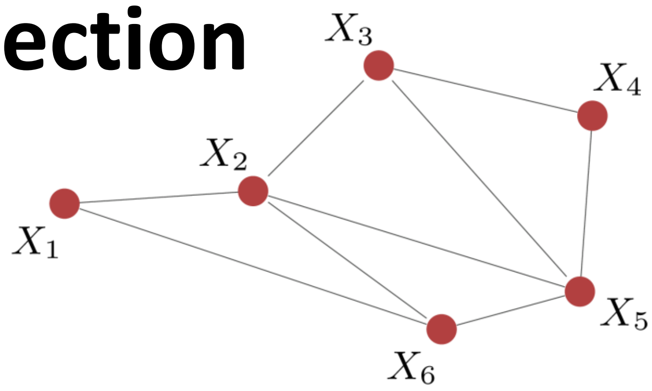
precision matrix $\Theta^\star := \left(\Sigma^\star\right)^{-1}$ is *sparse*

independent



$$\Longrightarrow \qquad \Theta^\star = \begin{bmatrix} \star & \star & 0 & 0 & 0 & \star \\ \star & \star & \star & 0 & \star & \star \\ 0 & \star & \star & \star & \star & 0 \\ 0 & 0 & \star & \star & \star & 0 \\ 0 & \star & \star & \star & \star & \star \\ \star & \star & 0 & 0 & \star & \star \end{bmatrix}$$

Gaussian Graphical Model Selection



$$X = (X_1, X_2, \ldots, X_d) \ \sim \ \mathcal{N}(0_d\,,\,\Sigma^\star)$$

pdf : $$f_X(x\,;\,\Theta^\star) = \frac{\sqrt{\det\ \Theta^\star}}{(2\pi)^{\frac{d}{2}}} \exp\Big(-\frac{1}{2} x^\top \Theta^\star x\Big)$$

Maximum likelihood estimator of $\Theta^\star$

$$\widehat{\Theta}_{\mathrm{ML}} = \arg\max_{\Theta} \ \log \prod_{i=1}^{n} f\Big(x^{(1)}, \ldots, x^{(n)}\,;\,\Theta\Big)$$

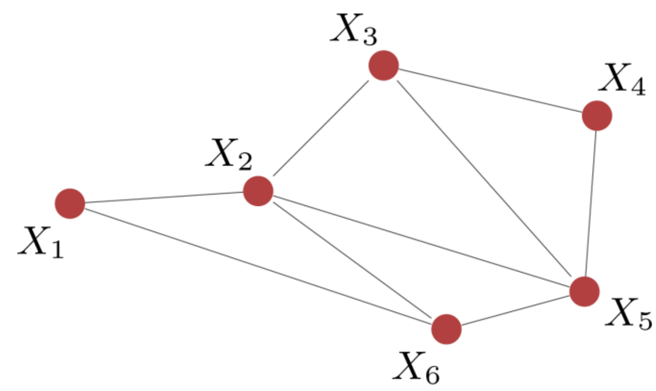
$$= \arg\min_{\Theta} - \log\ \det\ \Theta + \mathrm{tr}\Big(\Theta\,\widehat{\Sigma}_n\Big)$$

sample covariance matrix $\widehat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^{n} x^{(i)} {x^{(i)}}^\top$

$$= \widehat{\Sigma}_n^{-1}$$

(assuming it is invertible $\implies$ $n > d$ *)*

Graphical LASSO



Maximum likelihood estimator

$$\widehat{\Theta}^{\star}_{\mathrm{ML}} = \arg\min_{\Theta} - \log\det\Theta + \mathrm{tr}\Big(\Theta\,\widehat{\Sigma}_n\Big)$$

Assumption: most pairs of coordinates (X_i, X_j) are conditionally independent

Graphical LASSO

$$\widehat{\Theta}_{\mathrm{GL}} = \arg\min_{\Theta} - \log\det\Theta + \mathrm{tr}\Big(\Theta\,\widehat{\Sigma}_n\Big) + \underbrace{\lambda\big\|\Theta\big\|_{1,\text{off-d}}}_{\sum_{i\neq j}|\Theta_{ij}|}$$

applies L1-norm only to off-diagonal entries

$$\Theta^{\star} = \begin{bmatrix} \star & \star & 0 & 0 & 0 & \star \\ \star & \star & \star & 0 & \star & \star \\ 0 & \star & \star & \star & \star & 0 \\ 0 & 0 & \star & \star & \star & 0 \\ 0 & \star & \star & \star & \star & \star \\ \star & \star & 0 & 0 & \star & \star \end{bmatrix}$$

sensible estimators even for non-Gaussian RVs

Example: Dow Jones

Price of stock of *6 companies* at beginning of each week of 2011 (Jan-Jun)

Week	Alcoa	American Express	Boeing	Bank of America	Caterpillar	Cisco Systems
1	14.67	43.30	66.15	10.59	100.25	14.94
2	15.29	43.73	69.26	10.89	101.30	15.14
3	15.82	43.86	69.42	11.18	102.59	16.04
4	15.87	43.86	70.29	11.47	102.72	16.41
5	15.92	43.96	70.86	11.87	103.42	16.59
6	15.95	44.13	71.17	11.89	103.56	16.82
7	15.96	44.20	71.43	12.28	104.86	16.88
8	16.18	44.75	71.52	12.32	105.58	16.93
9	16.19	44.94	71.60	12.36	105.87	17.01
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
24	17.42	50.74	79.31	14.77	96.93	21.22
25	18.06	51.39	80.35	15.08	99.62	22.11

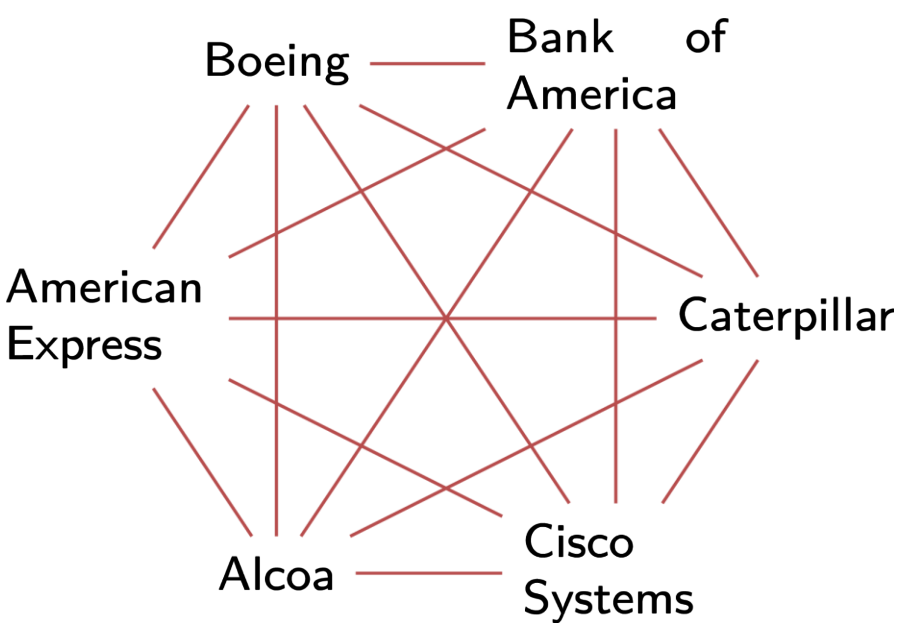
	$\widetilde{X}_1$	$\widetilde{X}_2$	$\widetilde{X}_3$	$\widetilde{X}_4$	$\widetilde{X}_5$	$\widetilde{X}_6$
remove mean	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
	X_1	X_2	X_3	X_4	X_5	X_6

$d = 6$

$n = 25$

Example: Dow Jones

$$\widehat{\Theta}_{\mathrm{ML}} = \widehat{\Sigma}_n^{-1} = \begin{bmatrix} 60.82 & 4.85 & -6.21 & -21.34 & -0.07 & -4.81 \\ 4.85 & 7.34 & -1.22 & -5.50 & 0.37 & -4.78 \\ -6.21 & -1.22 & 3.03 & 2.08 & -0.08 & -2.69 \\ -21.34 & -5.50 & 2.08 & 14.31 & -0.42 & 1.69 \\ -0.07 & 0.37 & -0.08 & -0.42 & 0.06 & 0.07 \\ -4.81 & -4.78 & -2.69 & 1.69 & 0.07 & 11.38 \end{bmatrix}$$



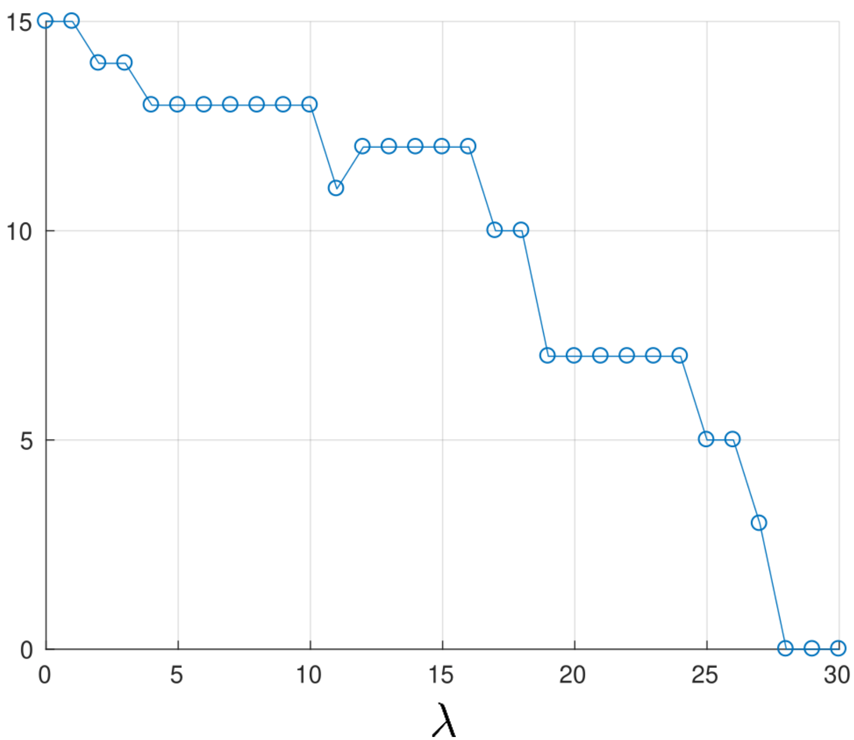
Example: Dow Jones

$$\widehat{\Theta}_{\mathrm{GL}} = \arg\min_{\Theta} - \log\det\Theta + \mathrm{tr}\left(\Theta\,\widehat{\Sigma}_n\right) + \lambda \big\|\Theta\big\|_{1,\text{off-d}}$$

graphical LASSO

$\left|(\widehat{\Theta}_{\mathrm{GL}})_{ij}\right| \leq 10^{-3} \quad \Longrightarrow$ we assume no correlation, i.e., *no edge* (i,j)

Number of edges



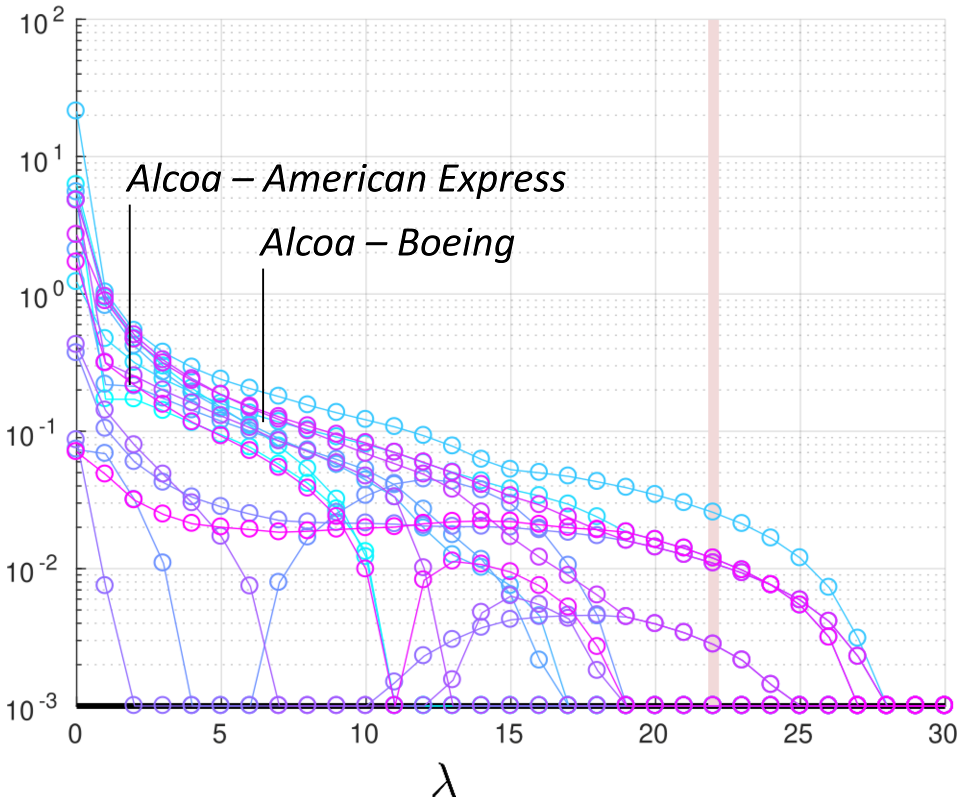
Example: Dow Jones

$$\widehat{\Theta}_{\mathrm{GL}} = \arg\min_{\Theta} -\log\det\Theta + \mathrm{tr}\left(\Theta\,\widehat{\Sigma}_n\right) + \lambda\|\Theta\|_{1,\text{off-d}}$$

graphical LASSO

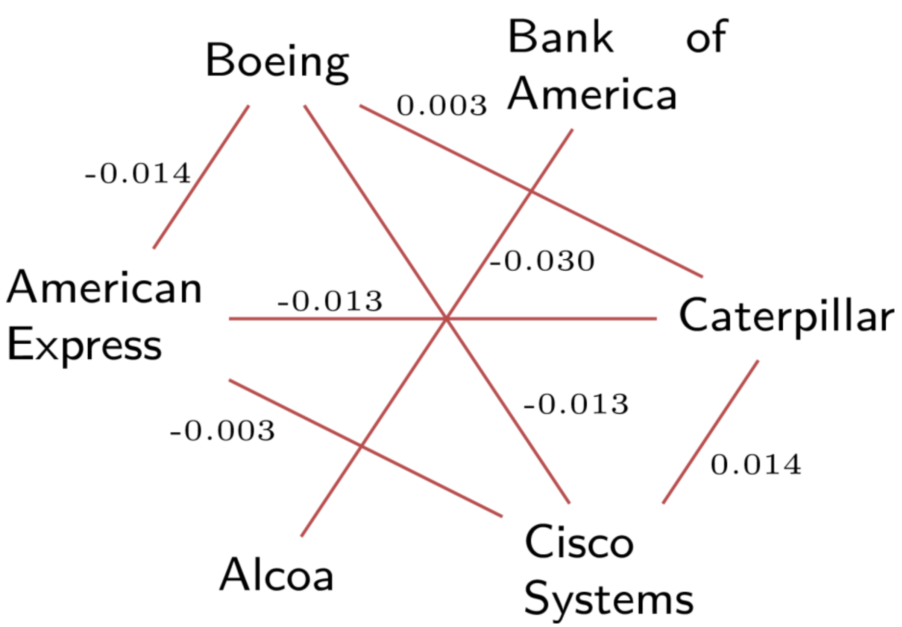
$\left|(\widehat{\Theta}_{\mathrm{GL}})_{ij}\right| \leq 10^{-3} \quad \Longrightarrow \quad$ we assume no correlation, i.e., *no edge* (i,j)

$\left|(\widehat{\boldsymbol{\Theta}}_{\mathrm{GL}})_{\boldsymbol{ij}}\right|$



Example: Dow Jones

$$\widehat{\Theta}_{\mathrm{GL}} = \begin{bmatrix} 1.814 & 0 & 0 & -0.030 & 0 & 0 \\ 0 & 0.186 & -0.014 & 0 & 0.013 & -0.003 \\ 0 & -0.014 & 0.096 & 0 & 0.003 & -0.013 \\ -0.030 & 0 & 0 & 0.601 & 0 & 0 \\ 0 & 0.013 & 0.003 & 0 & 0.026 & 0.014 \\ 0 & -0.003 & -0.013 & 0 & 0.014 & 0.282 \end{bmatrix}$$



Outline

- *Motivation: Hypothesis Testing in High-Dimensions*
- *Introduction to LASSO and other sparsity problems*
- *Gaussian graphical model selection*
- Matrix completion

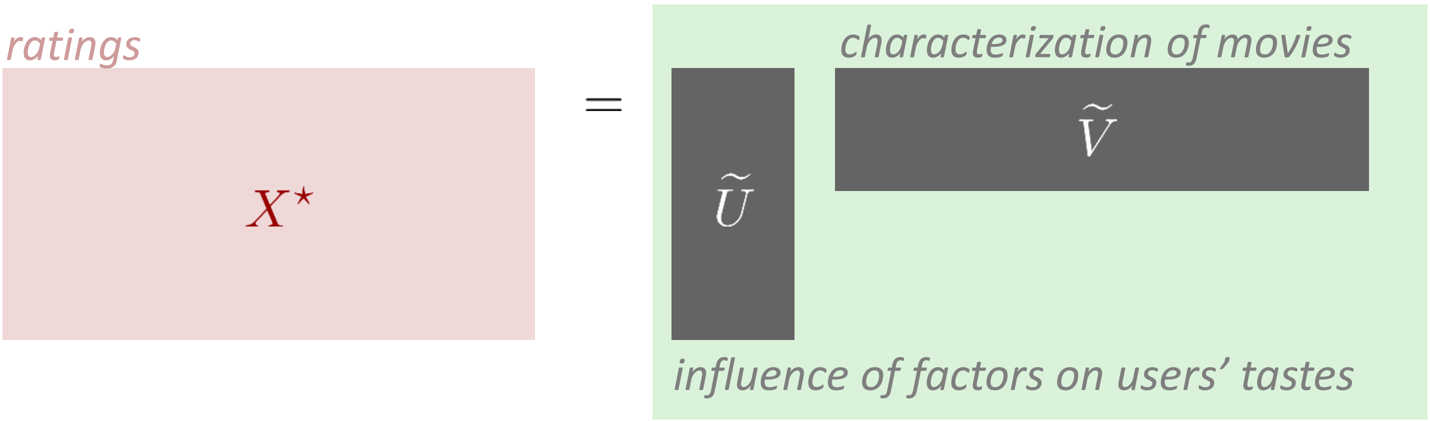
Matrix Completion

Suppose someone gave you \$1M for *completing* a table like this...

$$\begin{array}{cc} & \text{movies} \\ \text{users} & \begin{bmatrix} \star & 0.1 & \star & \star & 0.9 & \star & 0.9 & \star \\ 0.5 & 0.4 & \star & \star & \star & 0.6 & \star & 0.6 \\ \star & \star & \star & 0.7 & \star & \star & \star & 0.2 \\ 0.4 & 0.2 & 0.9 & \star & \star & \star & \star & \star \\ \star & \star & 0.7 & 0.7 & \star & \star & 0.3 & \star \end{bmatrix} \end{array}$$

rating user i gives to movie j

Key insight: only a few factors may explain users' tastes (genre, actors, ads, ...)



Fewer degrees of freedom

Singular value decomposition: any real $m \times n$ matrix can be decomposed as

$$A = U\Sigma V^\top = \underbrace{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix}}_{m\times k} \underbrace{\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k \end{bmatrix}}_{k\times k} \underbrace{\begin{bmatrix} - & v_1^\top & - \\ & \vdots & \\ - & v_k^\top & - \end{bmatrix}}_{k\times n}$$

orthogonal

rank

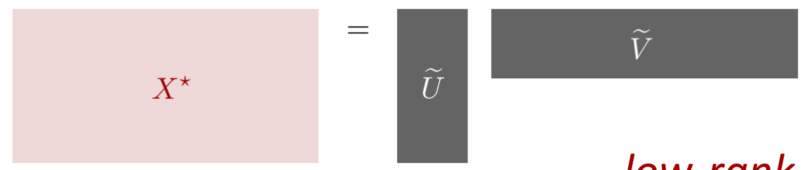
orthogonal

$$= \underbrace{\begin{bmatrix} | & & | & | & & | \\ u_1 & \cdots & u_k & u'_{k+1} & \cdots & u'_m \\ | & & | & | & & | \end{bmatrix}}_{m\times m} \underbrace{\left[\begin{array}{cccc|c} \sigma_1 & 0 & \cdots & 0 & \\ 0 & \sigma_2 & \cdots & 0 & \\ \vdots & \vdots & \ddots & \vdots & 0_{k\times(n-k)} \\ 0 & 0 & \cdots & \sigma_k & \\ \hline & 0_{(m-k)\times k} & & & 0_{(m-k)\times(n-k)} \end{array}\right]}_{m\times n} \underbrace{\begin{bmatrix} - & v_1^\top & - \\ & \vdots & \\ - & v_k^\top & - \\ - & {v'_1}^\top & - \\ & \vdots & \\ - & {v'_k}^\top & - \end{bmatrix}}_{n\times n}$$

sparse vector if matrix is low-rank

$$= \underbrace{U\Sigma^{\frac{1}{2}}}_{m\times k} \underbrace{\Sigma^{\frac{1}{2}} V^\top}_{k\times n}$$

$$X^\star = \underbrace{U\Sigma^{\frac{1}{2}}}_{m\times k}\underbrace{\Sigma^{\frac{1}{2}}V^\top}_{k\times n}$$



low-rank

Our problem

$$\begin{array}{ll} \underset{X\in\mathbb{R}^{m\times n}}{\text{minimize}} & \text{rank}(X) \\ \text{subject to} & X_{ij} = a_{ij}\,, \quad (i,j)\in\mathcal{O} \end{array}$$

nonconvex

observed entries

relax

nuclear norm

$$\begin{array}{ll} \underset{X\in\mathbb{R}^{m\times n}}{\text{minimize}} & \big\|(\sigma_1(X),\sigma_2(X),\ldots,\sigma_r(X))\big\|_1 \\ \text{subject to} & X_{ij} = a_{ij}\,, \quad (i,j)\in\mathcal{O} \end{array} = \|X\|_\star$$

$$\begin{array}{ll} \underset{X\in\mathbb{R}^{m\times n}}{\text{minimize}} & \|X\|_\star \\ \text{subject to} & \text{tr}(XM_l) = a_l\,, \quad l = 1,\ldots,p \end{array}$$

$$\begin{array}{cc} & \text{movies} \\ \text{users} & \begin{bmatrix} \star & 0.1 & \star & \star & 0.9 & \star & 0.9 & \star \\ 0.5 & 0.4 & \star & \star & \star & 0.6 & \star & 0.6 \\ \star & \star & \star & 0.7 & \star & \star & \star & 0.2 \\ 0.4 & 0.2 & 0.9 & \star & \star & \star & \star & \star \\ \star & \star & 0.7 & 0.7 & \star & \star & 0.3 & \star \end{bmatrix} \end{array}$$

Example Result

Theorem [Chandrasekaran et al. 12']

$X^\star \in \mathbb{R}^{m \times n}$ *unknown*, but rank k

iid entries $\mathcal{N}(0,1)$

$a_l = \mathrm{tr}(X M_l)\,, \quad l = 1, \ldots, p$ *measurements*

$$\boxed{p \geq 3k\,(m+n-k)+1} \quad \Longrightarrow \quad \begin{array}{rll} X^\star = & \underset{X}{\mathrm{argmin}} & \|X\|_\star \\ & \text{s.t.} & \mathrm{tr}(XM_l) = a_l\,, \quad l = 1, \ldots, p \end{array}$$

w.h.p.

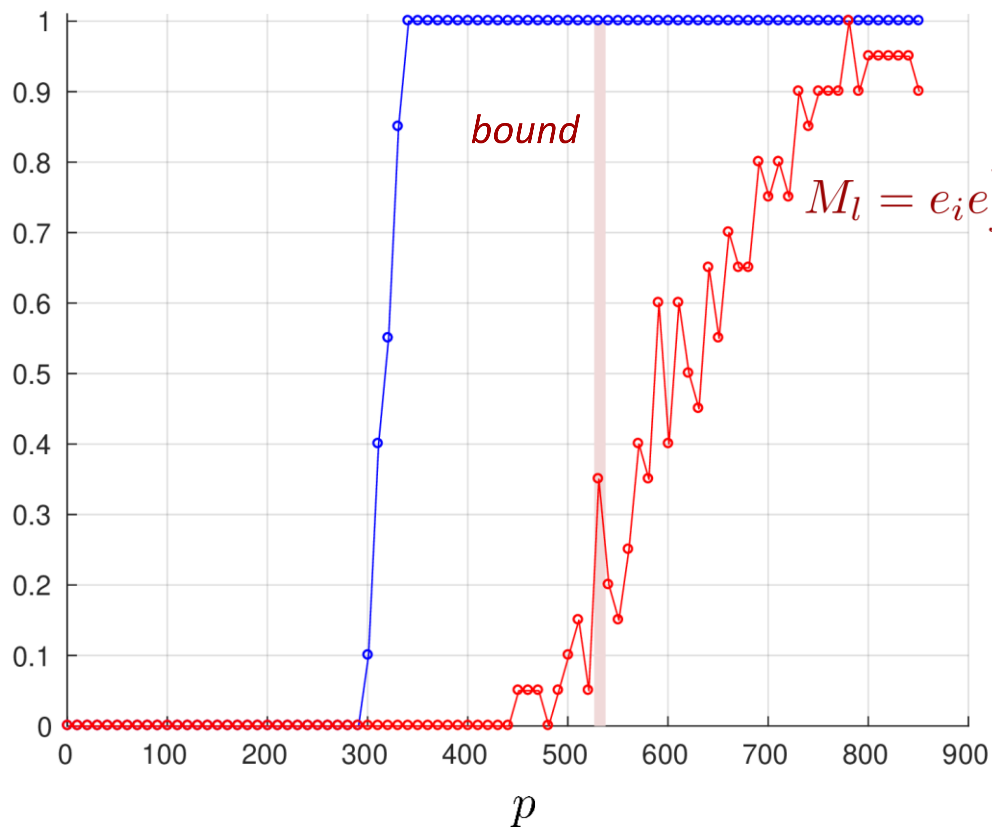
Experiments

$X^\star : 30 \times 30$ $\mathrm{rank}(X^\star) = 3$

$$\begin{array}{ll} \underset{X \in \mathbb{R}^{m\times n}}{\text{minimize}} & \|X\|_\star \\ \text{subject to} & \mathrm{tr}(XM_l) = a_l\,, \quad l = 1,\ldots,p \end{array}$$

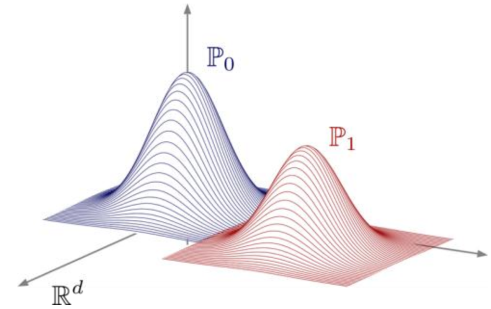
iid entries $\mathcal{N}(0,1)$

Sucess rate (20 trials)

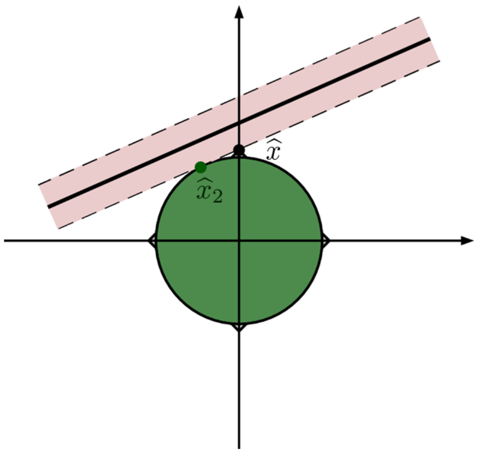


$M_l = e_i e_j^\top$ — *all-zeros, and 1 at random entry*

Conclusions

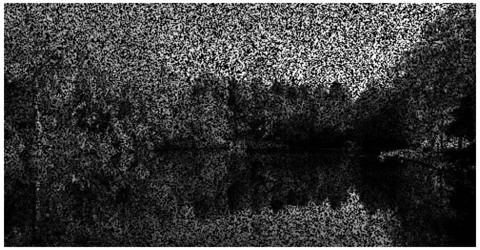


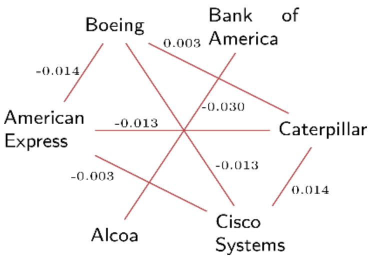
- Structure is key in *high-dimensional* problems
- *Sparsity* encodes several types of structure

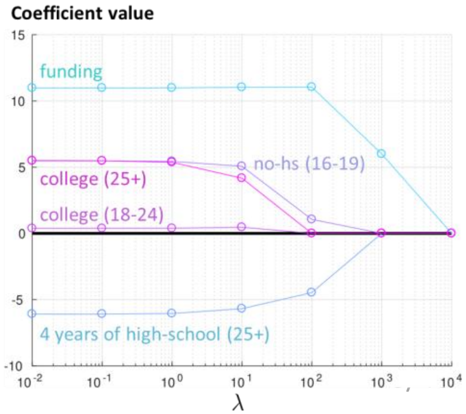


- Several applications (and theory)
- LASSO, basis pursuit, ... improve *interpretability* and (often) *performance*
- Didn't cover: optimization theory and *algorithms*

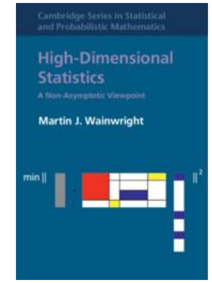
City	Police funding/ resident ($/yr)	% with 4 years high-school (+25)	% not in high-school (16-19)	% in college (18-24)	% with 4+ years college (25+)	Crimes/million
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
6	25	68	8	32	15	603
7	34	68	12	24	14	484
8	33	62	13	28	11	546
9	36	69	7	25	12	424
⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	66	67	26	18	16	940

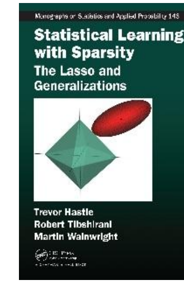






References





M. J. Wainwright
High-Dimensional Statistics: A Non-Asymptotic Viewpoint
Cambridge University Press, 2019

T. Hastie, R. Tibshirani, M. Wainwright
Statistical Learning with Sparsity: The Lasso and Generalizations
CRC Press, 2016 https://web.stanford.edu/~hastie/StatLearnSparsity/

V. Chandrasekaran, B. Recht, P. A. Parrilo, A. S. Willsky
The Convex Geometry of Linear Inverse Problems
Foundations of Computational Mathematics, Vol. 12, pp. 805-849, 2012

J. F. C. Mota, N. Deligiannis, M. R. D. Rodrigues
Compressed sensing with side information: Geometrical interpretation and performance bounds
IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2014

P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu
High-dimensional covariance estimation by minimizing L1-penalized log-determinant divergence
Electronic Journal of Statistics, Vol. 5, pp. 935-980, 2011

References

V. Chandrasekaran
Convex Optimization Methods for Graphs and Statistical Modeling
PhD thesis, MIT, 2011

M. S. Brown, M. Pelosi, H. Dirska
Dynamic-radius species-conserving genetic algorithm for the financial forecasting of Dow Jones index stocks
Machine Learning and Data Mining in Pattern Recognition, Vol. 7988, pp. 27-41, 2013

Code & presentation

https://github.com/joaofcmota/udrc-summerschool

http://jmota.eps.hw.ac.uk/documents/Mota21-HighDimensionalStatsAndSparsity-UDRC.pdf