Session 10: Linear regression

Optimization and Computational Linear Algebra for Data Science

Textbook: Boyd & Vandenberghe: Introduction to applied linear algebra - Chaps 12 & 13

Marylou Gabrié (based on material by Léo Miolane)

Contents

1. Ordinary least squares
2. Penalized linear regression
3. Matrix norms

Introduction

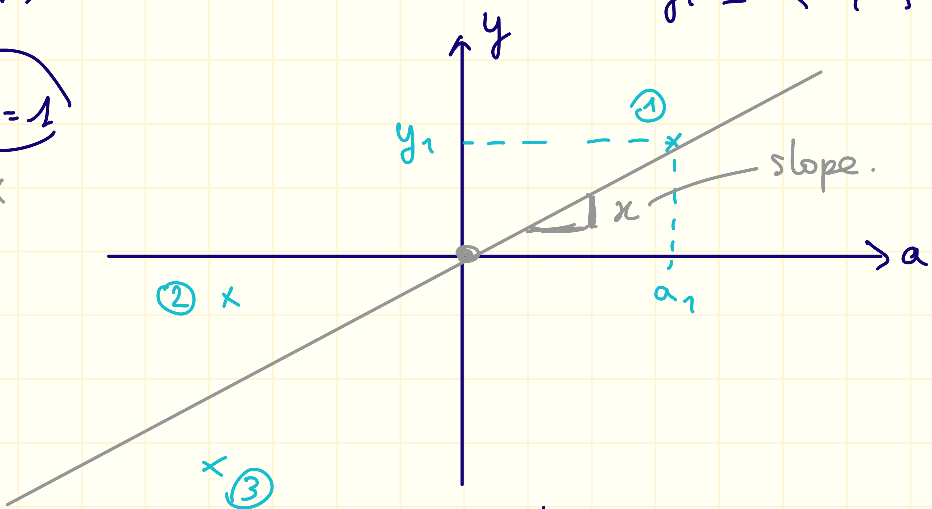
examples $\left\{ \begin{array}{l} d \text{ physiological measures } a_i \\ 1 \text{ output } y_i - \text{diabete stage} \end{array} \right.$ d features

- We have n « feature vectors » $a_1, \dots, a_n \in \mathbb{R}^d$.
- Each point a_i comes with a « target variable » $y_i \in \mathbb{R}$.
- **Goal.** Find a linear relation between the a_is and the y_is:

↳ Find $x \in \mathbb{R}^d$ such that $y_i \simeq \langle x, a_i \rangle$ for all $i = 1 \dots n$

$d = 1$

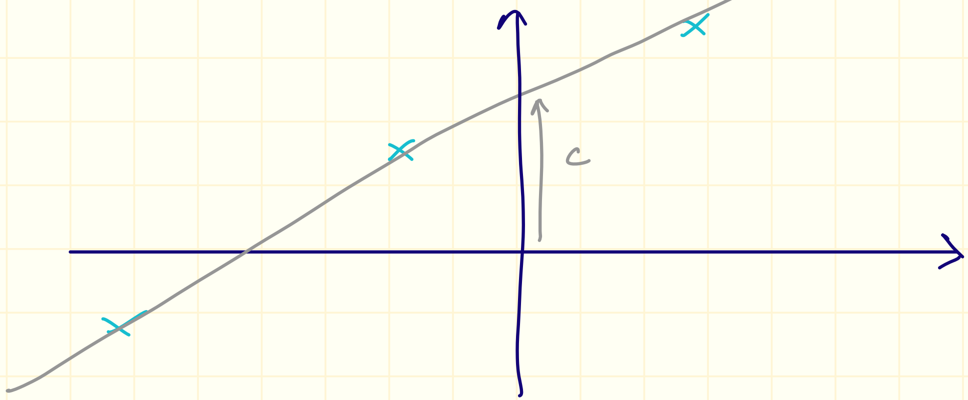
$y_i = a_i x$



- **Prediction:** New $a \in \mathbb{R}^d$ $\hat{y} = \langle x, a \rangle$

Can we have an intercept?

Find $x \in \mathbb{R}^d$ such that $y_i = \langle x, a_i \rangle + \mathbf{c}$. $= \langle \tilde{x}, \tilde{a}_i \rangle$



$\tilde{a}_i \in \mathbb{R}^{d+1}$ — $\tilde{a}_i = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{id} \\ 1 \end{pmatrix}$ $\qquad \tilde{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \\ c \end{pmatrix}$

=> <u>Write</u>: $A = \begin{pmatrix} - & a_1^T & - \\ & \vdots & \\ - & a_n^T & - \end{pmatrix} \in \mathbb{R}^{n \times (d+1)}$ $\quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$ $\quad$ Find x in $\mathbb{R}^{d+1}$

"$Ax = y$"

Solving $Ax = y$ is a bad idea

From now on "no intercept"

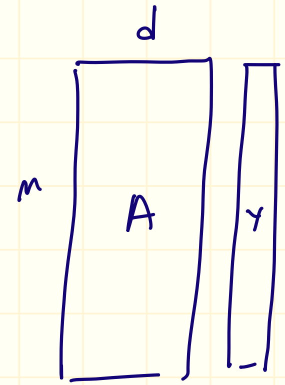
The system $Ax = y$ may have:

- No solution.

<u>example</u>: A is a "tall" matrix $(n > d)$

$\Rightarrow \dim \mathrm{Im}(A) \leqslant d < n$

$\Rightarrow y \in \mathbb{R}^n$ is unlikely to belong to $\mathrm{Im}(A) \subseteq \mathbb{R}^n$.

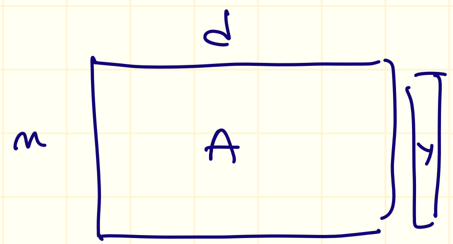


- Infinitely many solutions.

<u>example</u>: A is a "fat" matrix $(d > n)$

$\dim(\mathrm{Ker}(A)) \geqslant d - n > 0$

$\Rightarrow$ typically infinitely many solutions ....



1. Ordinary least squares

Least squares problem

Euclidian, ℓ2-norm

(LS) Minimize $f(x) = \|Ax - y\|^2$ with respect to $x \in \mathbb{R}^d$.

f is convex (HW9) therefore:

x minimizes $f \iff \nabla f(x) = 0$

$\iff 2(A^T A x - A^T y) = 0$

$\iff A^T A x = A^T y$

Conclusion: the minimizers of f are the solutions $x \in \mathbb{R}^d$ of the linear system

$$\boxed{A^T A x = A^T y.}$$

CASE 1 : $A^T A$ is invertible: $\boxed{x = (A^T A)^{-1} A^T y.}$

The Moore-Penrose pseudo-inverse

What if $A^\top A$ is not invertible?

"dagger"

Definition

Let $A = U\Sigma V^\intercal$ be the SVD of A. The matrix $A^\dagger \overset{\text{def}}{=} V\Sigma' U^\intercal$ is called the **(Moore-Penrose) pseudo-inverse** of A, where $\Sigma' \in \mathbb{R}^{d\times n}$ is

$$\Sigma'_{i,i} = \begin{cases} 1/\Sigma_{i,i} & \text{if } \Sigma_{i,i} \neq 0 \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \quad \Sigma'_{i,j} = 0 \ \text{ for } i \neq j$$

$$A = U \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & 0 & \\ & & & & & \ddots \ 0 \end{pmatrix} V^T$$

$$A^+ = V \begin{pmatrix} 1/\sigma_1 & & & & \\ & \ddots & & & \\ & & 1/\sigma_r & & \\ & & & 0 & \\ & & & & 0 \end{pmatrix} U^T$$

Exercise: Check that if A is invertible then $A^{-1} = A^\dagger$.

Solving $A^\top A x = A^\top y$

A

minimize $\|y - Ax\|^2$

Claim: The vector $x^{\mathrm{LS}} \stackrel{\text{def}}{=} A^\dagger y$ is a $\boxed{\text{solution of } A^\top A x = A^\top y}$

$$A^\top A x^{LS} = V\Sigma^\top \cancel{U^\top U} \Sigma \cancel{V^\top V} \Sigma' U^\top y$$

$$= V \underbrace{\Sigma^\top \Sigma \Sigma'}_{\Sigma^\top} U^\top y = V \Sigma^\top U^\top y = A^\top y.$$

Theorem

The set of the minimizers of $f(x) = \|Ax - y\|^2$ is

$$\left\{ x^{\mathrm{LS}} + v \,\middle|\, v \in \underbrace{\mathrm{Ker}(A)}_{\mathrm{Ker}(A^\top A)} \right\}.$$

2. Penalized least squares

Ridge regression

Ridge regression consists in adding a « ℓ_2 penalty » :

(Ridge) Minimize $f(x) = \|Ax - y\|^2 + \lambda\|x\|^2$ w.r.t. $x \in \mathbb{R}^d$.

for some fixed $\lambda > 0$.

- CONVEXITY: f is strongly convex $\Rightarrow$ strictly convex
  f admits a unique minimizer.
  $$\boxed{x^{\text{Ridge}} = (A^T A + \lambda I_d)^{-1} A^T y.}$$ HW!

- WHY $\ell 2$ penalty?

  Tradeoff:
  - promotes solution x with small norm
    Ex $\|x^{\text{Ridge}}\| \leq \|x^{LS}\|$.
    $\Rightarrow$ small norm solution can be robust.
  - issue $\|Ax^{LS} - y\| \leq \|Ax^{\text{Ridge}} - y\|$

Lasso

$$\|x\|_1 = \sum_{i=1}^{d} |x_i|$$

The Lasso adds a « ℓ_1 penalty » :

(Lasso) Minimize $f(x) = \|Ax - y\|^2 + \lambda \|x\|_1$ w.r.t. $x \in \mathbb{R}^d$.

for some fixed $\lambda > 0$.

- CONVEXITY: f is not strictly convex, no unique minimizer in general.

  In practice, LASSO minimizer is still unique.

- WHY ℓ_1 PENALTY:
  - promotes sparse vectors: x^{LASSO} will have "a lot" of coordinates equal to 0

$$x^{LASSO} = \begin{pmatrix} x_1 \\ 0 \\ x_2 \\ 0 \\ 0 \\ \vdots \end{pmatrix}$$

→ Features selection.

- Tradeoff.

Intuition behind feature selection

Lemma

Let x^{Lasso} be a minimizer of the Lasso cost function and let $r = \|x^{\text{Lasso}}\|_1$. Then x^{Lasso} is a solution to the constrained optimization problem:

$$\text{minimize} \quad \|Ax - y\|^2 \quad \text{subject to} \quad \|x\|_1 \leq r.$$

Proof: By contradiction: assume that there exist x such that

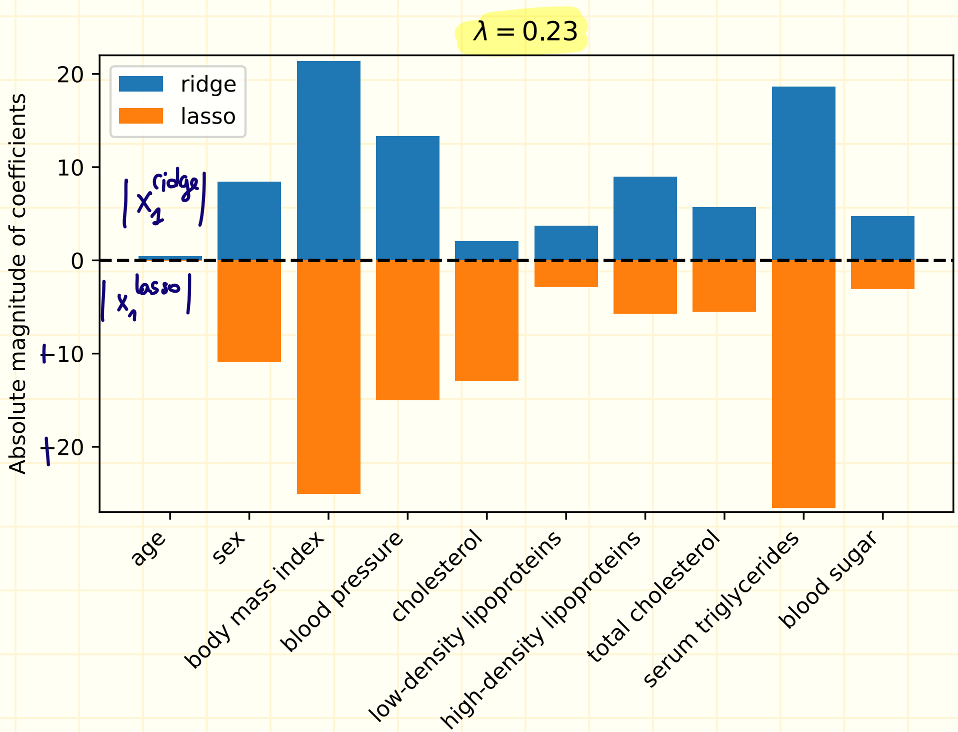
$$\begin{cases} \|Ax-y\|^2 < \|Ax^{lasso} - y\|^2 \\ \|x\|_1 \leq \|x^{LASSO}\|_1 = r. \end{cases}$$

$$\Rightarrow \|Ax-y\|^2 + \lambda\|x\|_1 < \underbrace{\|Ax^{LASSO} - y\|^2 + \lambda\|x^{LASSO}\|_1}_{f(x^{LASSO}).}$$

$\longrightarrow$ Contradiction since x^{LASSO} minimizer of $f(\cdot)$

Effect of Regularization

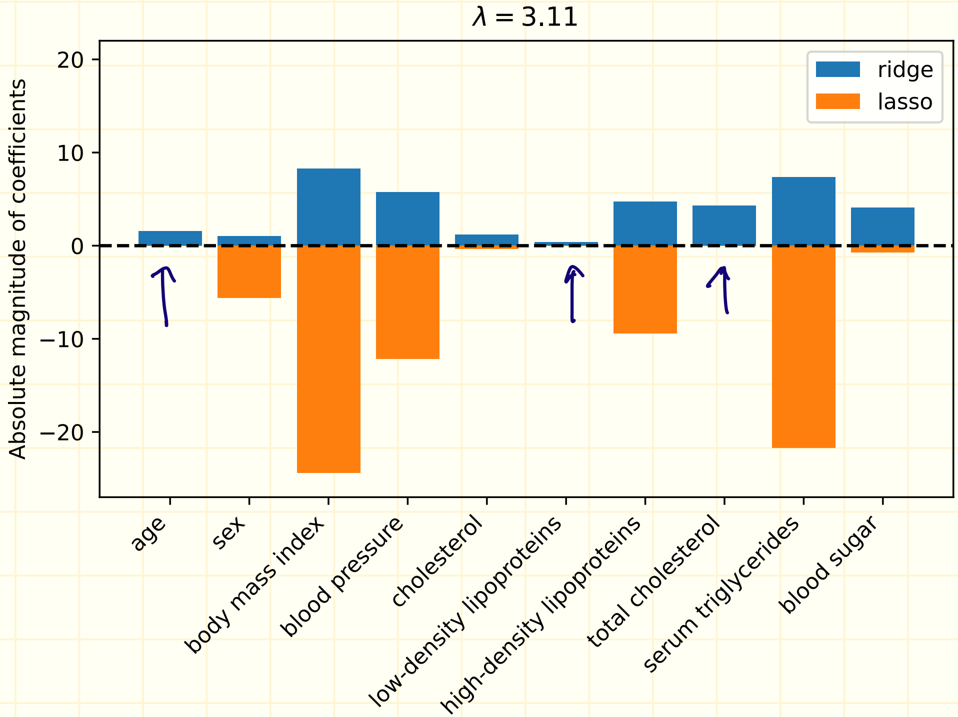
Consider a data set with $n = 442$ patients, with $d = 10$ dimensions feature vectors and the prediction of diabetese disease progression



https://scikit-learn.org/stable/datasets/toy_dataset.html

Effect of Regularization

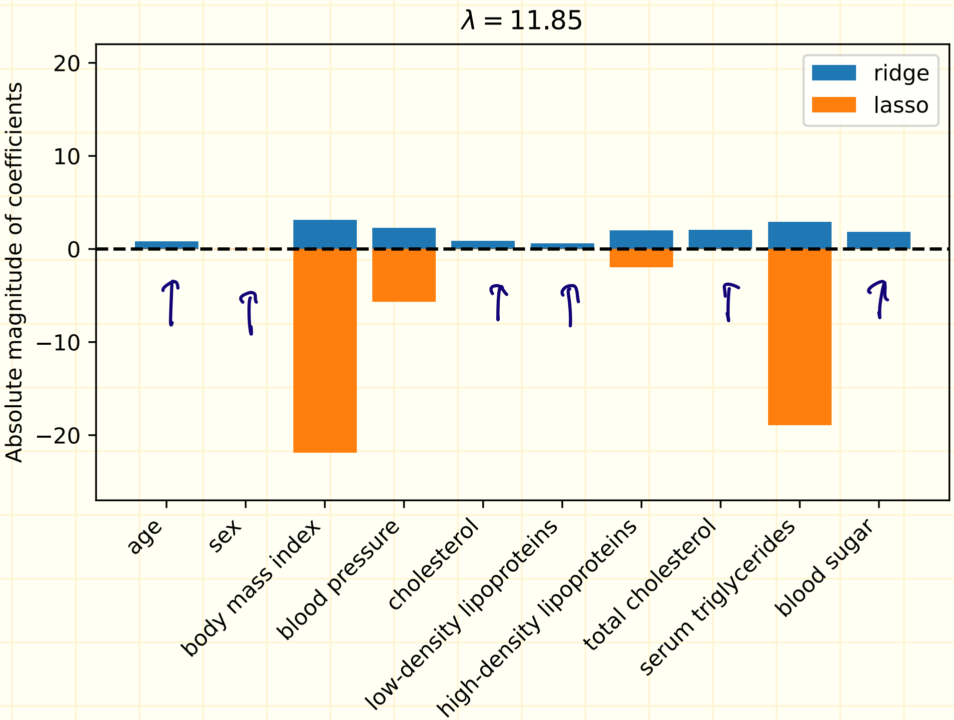
Consider a data set with $n = 442$ patients, with $d = 10$ dimensions feature vectors and the prediction of diabetese disease progression



https://scikit-learn.org/stable/datasets/toy_dataset.html

Effect of Regularization

Consider a data set with $n = 442$ patients, with $d = 10$ dimensions feature vectors and the prediction of diabetese disease progression



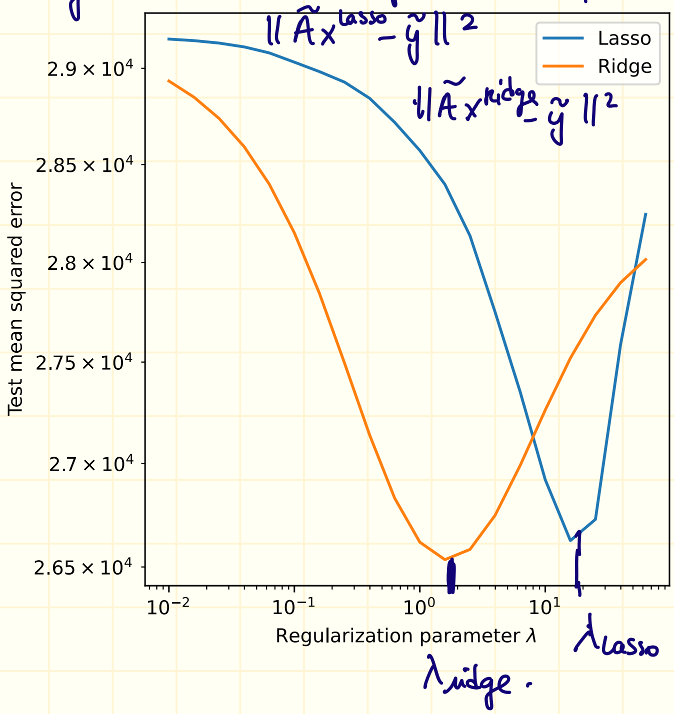
https://scikit-learn.org/stable/datasets/toy_dataset.html

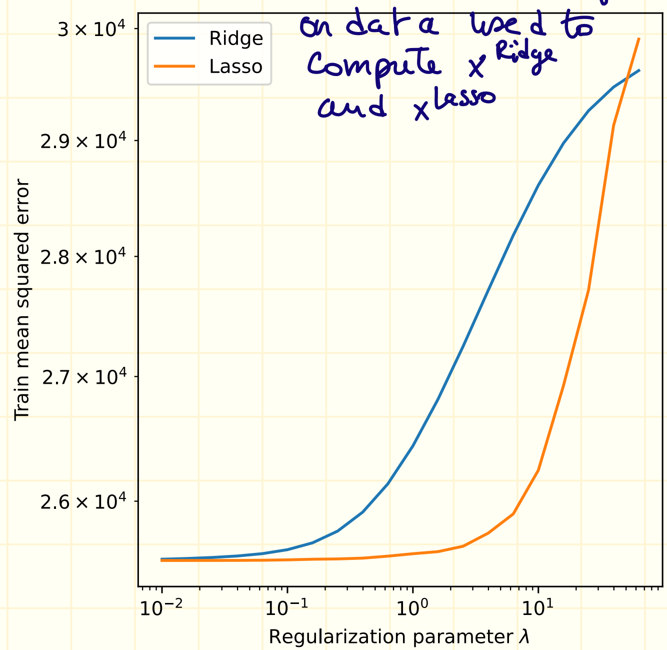
Effect of Regularization

Consider a data set with $n = 442$ patients, with $d = 10$ dimensions feature vectors and the prediction of diabetese disease progression

$\tilde{A}$ $\tilde{y}$ not used to compute x^{Ridge}, x^{lasso}

$\|Ax^{lasso} - y\|^2$ $\|Ax^{Ridge} - y\|^2$





https://scikit-learn.org/stable/datasets/toy_dataset.html

3. Matrix norms

Frobenius norm

Definition

The Frobenius norm of a matrix $A \in \mathbb{R}^{n\times m}$ is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}^2} = \sqrt{\mathrm{Tr}(A^TA)}$$

Proposition

$$\|A\|_F = \sqrt{\sum_{i=1}^{\min(n,m)} \sigma_i(A)^2}$$

$$\|A\|_F^2 = \mathrm{Tr}(AA^T) = \mathrm{Tr}(U\Sigma \cancel{V^T V} \Sigma^T U^T) = \mathrm{Tr}(\Sigma\Sigma^T) = \sigma_1^2 + \dots + \sigma_r^2$$

The spectral norm

Definition

The spectral norm of a matrix $A \in \mathbb{R}^{n \times m}$ is defined as

$$\|A\|_{\mathrm{Sp}} = \max_{\|x\|=1} \|Ax\|.$$

Proposition

largest singular values

$$\|A\|_{\mathrm{Sp}} = \sigma_1(A).$$

Proof: $\|A\|_{sp}^2 = \max_{\|x\|=1} \|Ax\|^2$

$= \max_{\|x\|=1} x^T A^T A x$

$= \lambda_1(A^T A) = \sigma_1^2(A)$.

max eigenvalue of $A^T A$

The nuclear norm

Definition

The nuclear norm of a matrix $A \in \mathbb{R}^{n \times m}$ is defined as

$$\|A\|_\star = \sum_{i=1}^{\min(n,m)} \sigma_i(A).$$

"ℓ_1 norm of matrices"

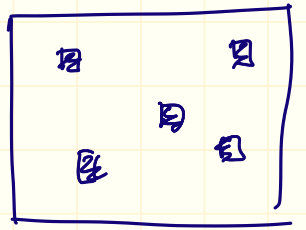
Application to matrix completion

We have a data matrix $M \in \mathbb{R}^{n \times m}$ that we only observe partially. That is we only have access to

observed entries
indices.

$$M_{i,j} \quad \text{for } (i,j) \in \Omega,$$

M



where $\Omega \subset \{1, \ldots, n\} \times \{1, \ldots m\}$ is a subset of the complete set of the entries.

↳ minimize $\boxed{\text{rank}(X) \text{ w.r.t } X \in \mathbb{R}^{n \times m} \text{ s.t. } M_{i,j} = X_{i,j} \text{ for all } (i,j) \in \Omega}$

(such that)

→ NP-HARD

→ Instead Solve $\boxed{\text{minimize } \|X\|_* \text{ w.r.t } X \in \mathbb{R}^{n \times m} \text{ s.t } M_{i,j} = X_{i,j} \text{ for all } (i,j) \in \Omega.}$

Application to matrix completion

Questions?

Questions?