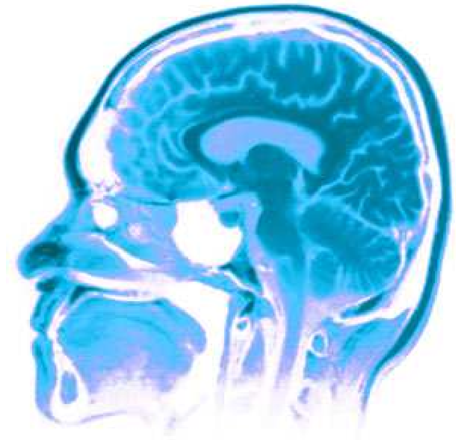
CPSC340



Sparse regularization and feature selection

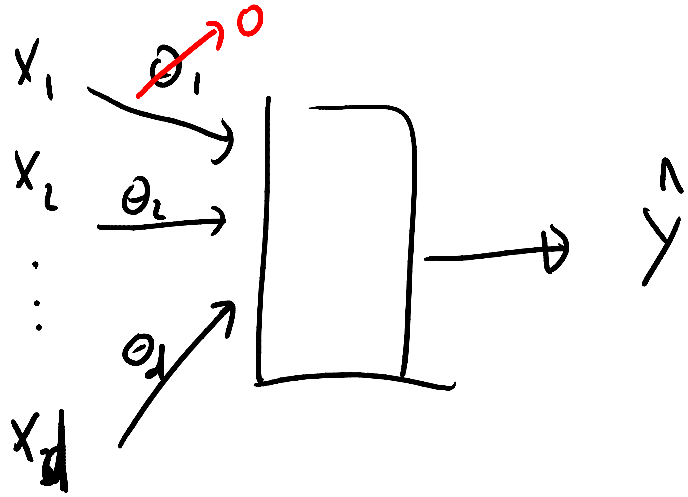
Nando de Freitas
October, 2012
University of British Columbia

Outline of the lecture

This lecture introduces one of the most popular modern techniques for regression and variable selection: Sparse regularization. The goal is for you to:

- ❑ Learn regularization with the L1 norm.
- ❑ Understand how regularizers can be used to automatically select input features.
- ❑ Understand the concept of sub-gradients as part of the derivation of an algorithm.
- ❑ Understand the pseudo-code of a coordinate descent algorithm to estimate the parameters of linear sparse models.

Selecting features for prediction



$$\hat{y} = x_1 \underbrace{\cancel{\theta_1}}_{\to 0} + x_2\theta_2 + \dots + x_d\theta_d$$

x_1 is expensive

x_1 does not contribute to good predictions $\hat{y}$

Then we want $\theta_1 \to 0$

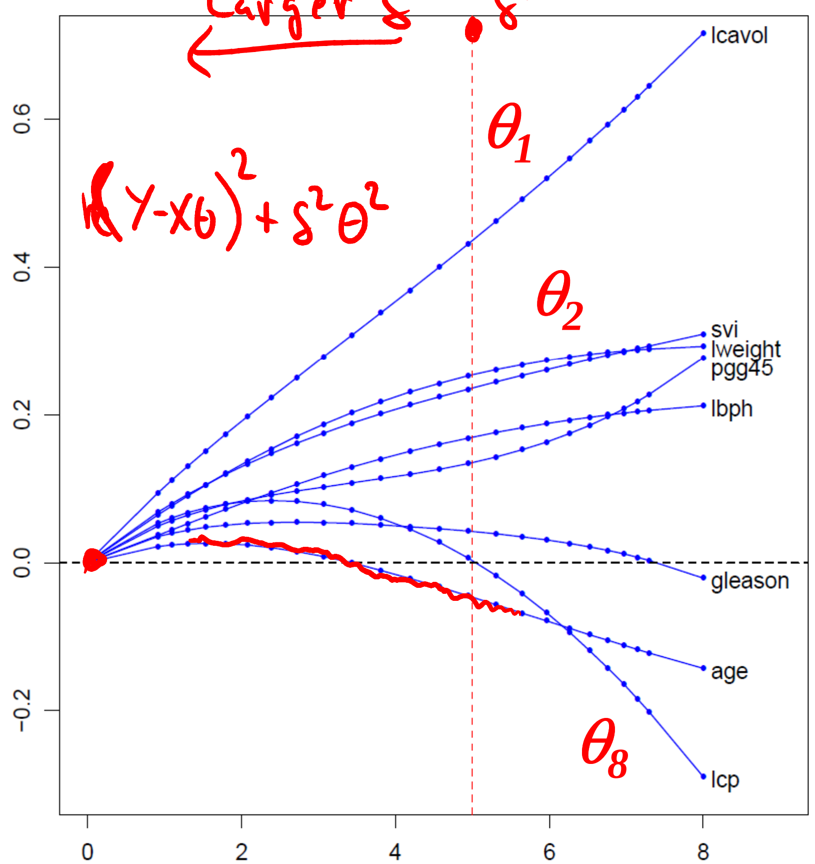
Selecting features for prediction

As δ increases, $t(\delta)$ decreases and each θ_i goes to zero, but too slowly for ridge. Lasso will ensure that irrelevant features $\mathbf{x}_i$ have weight $\theta_i = 0$.

Larger δ^2 δ^*

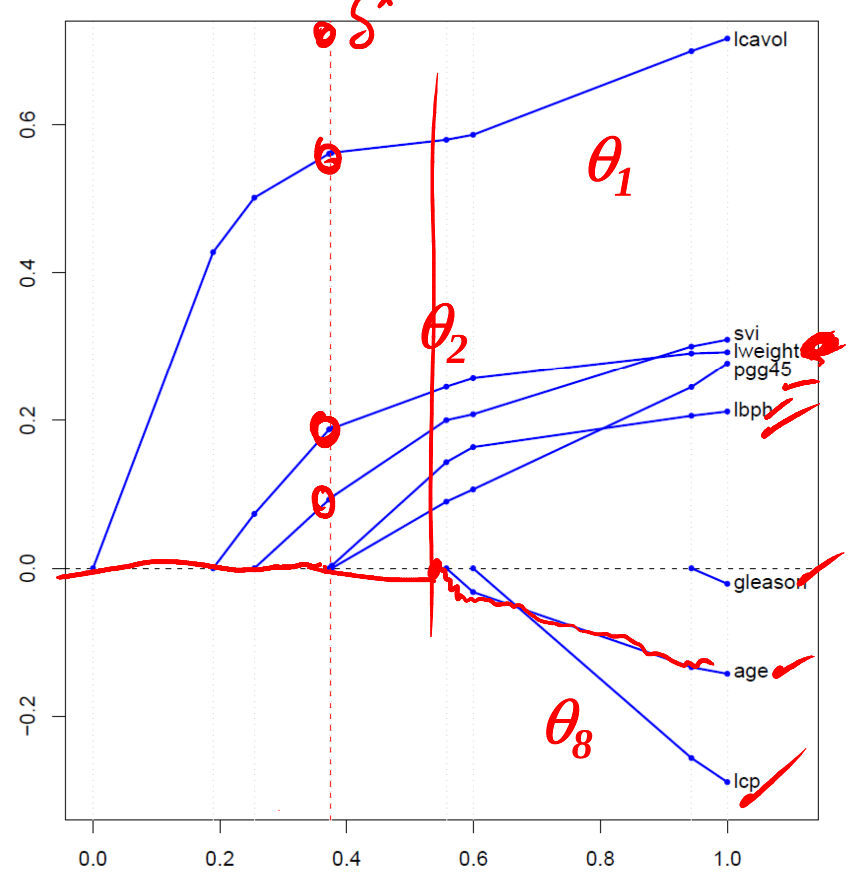
$(Y - X\theta)^2 + \delta^2\theta^2$

θ



$t(\delta)$ e.g. $\delta = 1/\delta^2$

δ^*



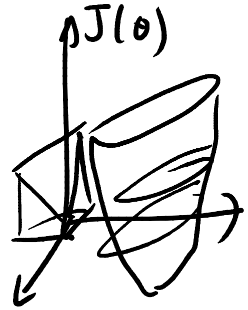
$t(\delta)$

[Hastie, Tibshirani & Friedman book]

The Lasso: least absolute selection and shrinkage operator

$$J(\theta) = \underbrace{(Y - X\theta)^T (Y - X\theta)} + \delta^2 \underbrace{\sum_{j=1}^{d} |\theta_j|}$$

L_1 Norm



in 2D $\qquad \theta = (\theta_1, \theta_2)$

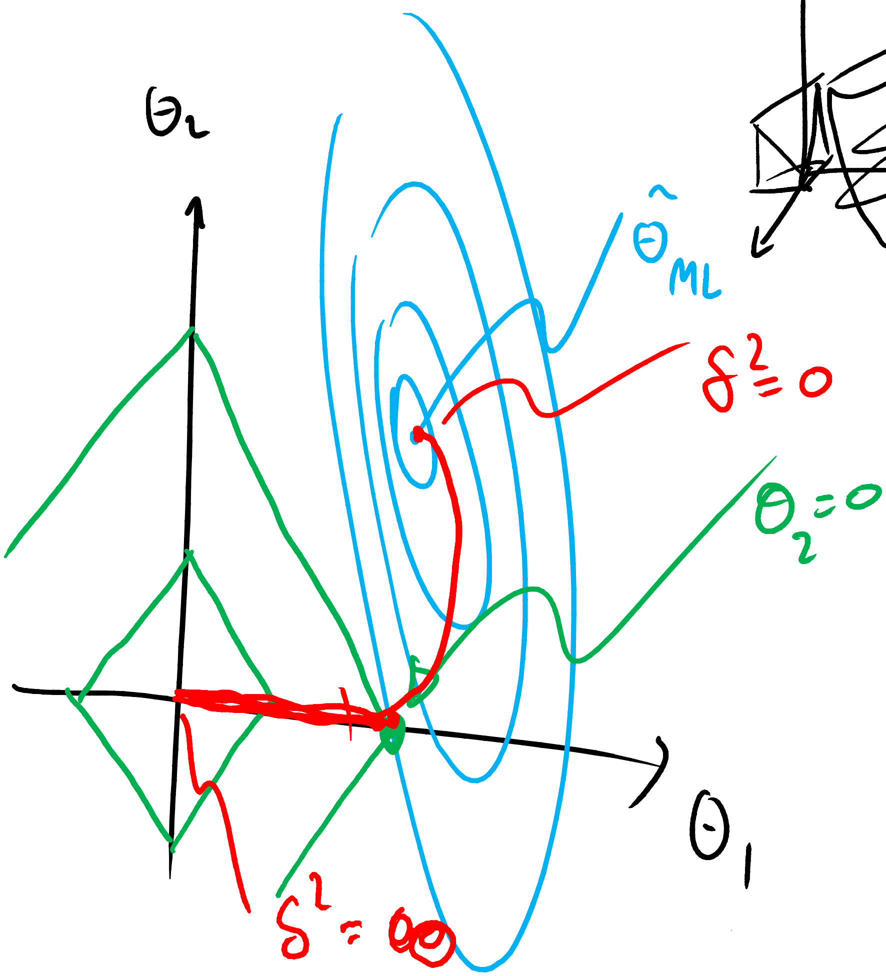
$|\theta_1| + |\theta_2| = \text{const}$

$\theta_1 + \theta_2 = \text{const}$

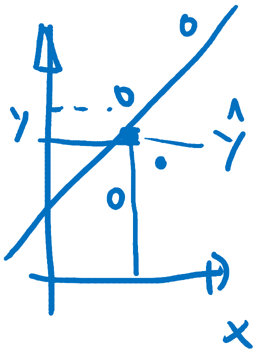
$\theta_1 - \theta_2 = \text{const}$

$-\theta_1 - \theta_2 = \text{const}$

$-\theta_1 + \theta_2 = \text{const}$



Differentiating the objective function



$$J(\boldsymbol{\theta}) = (\mathbf{y} - \overset{n\times d}{\mathbf{X}}\overset{d\times 1}{\boldsymbol{\theta}})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \delta^2 \sum_{j'=1}^{d} |\theta_{j'}| \qquad X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

$$= \sum_{i=1}^{n} \big(y_i - \underbrace{x_i^T\theta}_{\hat{y}_i}\big)^2 + \boxed{\delta^2 \sum_{j'=1}^{d} |\theta_{j'}|}$$

$\underline{\theta} = (\theta_1, \theta_2, \theta_3)$

$\underline{\theta}_2 = (\theta_1, \theta_3)$

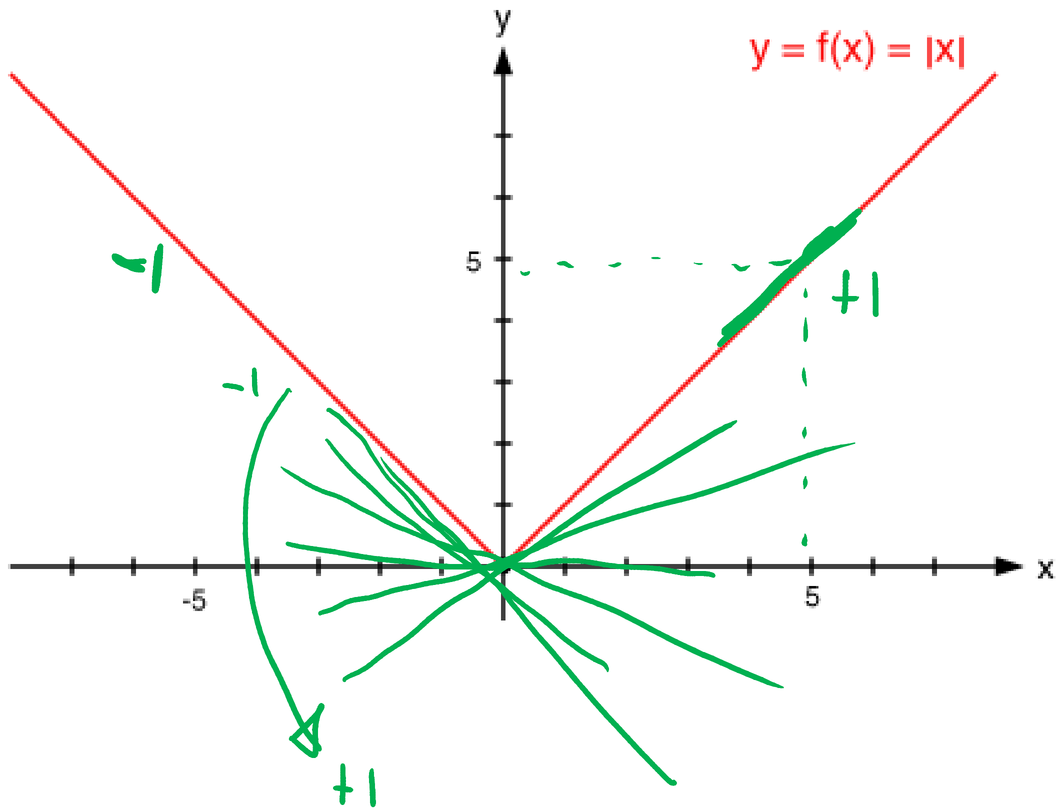
$$= \left[\sum_{i=1}^{n} \big(y_i - x_{ij}\theta_j - x_{i-j}^T\theta_{-j}\big)^2\right] + \boxed{\delta^2 \sum_{j'=1}^{d} |\theta_{j'}|}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} 2\big(y_i - x_{ij}\theta_j - x_{i-j}^T\theta_{-j}\big)(-x_{ij}) + \delta^2 \frac{\partial}{\partial \theta_j}\big(|\theta_1| + \dots + |\theta_d|\big)$$

$$= \overbrace{\boxed{2\sum_{i=1}^{n} x_{ij}^2}}^{a_j > 0}\theta_j - \overbrace{\boxed{2\sum_{i=1}^{n} \big(y_i - x_{i-j}^T\theta_{-j}\big)x_{ij}}}^{c_j} + \delta^2 \frac{\partial}{\partial \theta_j}|\theta_j|$$

Differentiating the objective function

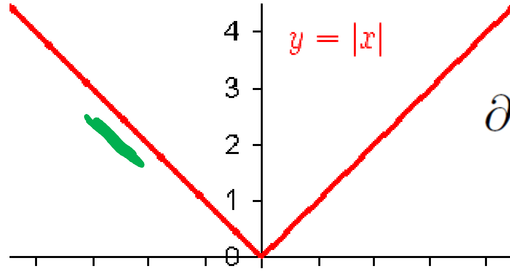
Subdifferentials



$$f(x) = |x|$$

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ \{+1\} & \text{if } x > 0 \end{cases}$$

[Wikipedia]



$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ \{+1\} & \text{if } x > 0 \end{cases}$$

$$\boxed{\partial_{\theta_j} J(\boldsymbol{\theta}) = a_j\theta_j - c_j + \delta^2 \partial_{\theta_j} |\theta_j|}$$

$$= \begin{cases} \{a_j\theta_j - c_j - 1\delta^2\} & \text{if } \theta_j < 0 \leftarrow \\ [-c_j - 1\delta^2, -c_j + 1\delta^2] & \text{if } \theta_j = 0 \\ \{a_j\theta_j - c_j + 1\delta^2\} & \text{if } \theta_j > 0 \leftarrow \end{cases}$$

$\frac{\partial|\theta_j|}{\partial\theta_j} = -1$

$\frac{\partial|\theta_j|}{\partial\theta_j} = +1$

*Hence, the estimate of the **j**-th parameter, **given the other parameters**, is*

$$\widehat{\theta}_j = \begin{cases} \boxed{(c_j + \delta^2)/a_j} & \text{if } c_j < -\delta^2 \\ 0 & \text{if } c_j \in [-\delta^2, \delta^2] \\ (c_j - \delta^2)/a_j & \text{if } c_j > \delta^2 \end{cases}$$

when $\theta_j < 0$

$a_j\theta_j - c_j - \delta^2 = 0$

$a_j\theta_j = c_j + \delta^2$

$\hat{\theta}_j = \frac{c_j + \delta^2}{a_j}$

Coordinate descent algorithm for sparse prediction

1. Initialize $\underline{\theta}$, e.g. $\underline{\theta} = (\underline{\underline{X}}^T \underline{\underline{X}} + \delta^2 I)^{-1} \underline{\underline{X}}^T \underline{y}$ (ridge)

2. REPEAT UNTIL CONVERGED

% $\underline{\underline{X}}$ is n by d

% $\underline{Y}$ is n by 1

3. FOR $j = 1, 2, \ldots, d$ DO

4. $a_j = 2 \sum_{i=1}^{n} x_{ij}^2$ ✓

5. $c_j = 2 \sum_{i=1}^{n} x_{ij} (y_i - \underline{x}_i^T \underline{\theta} + x_{ij} \theta_j)$ ✓

6. If $c_j < -\delta^2$

7. $\theta_j = (c_j + \delta^2) / a_j$

8. Elseif $c_j > \delta^2$

9. $\theta_j = (c_j - \delta^2) / a_j$

10. Else

11. $\theta_j = 0$

$X^{*T} \hat{\theta} = \hat{y}$

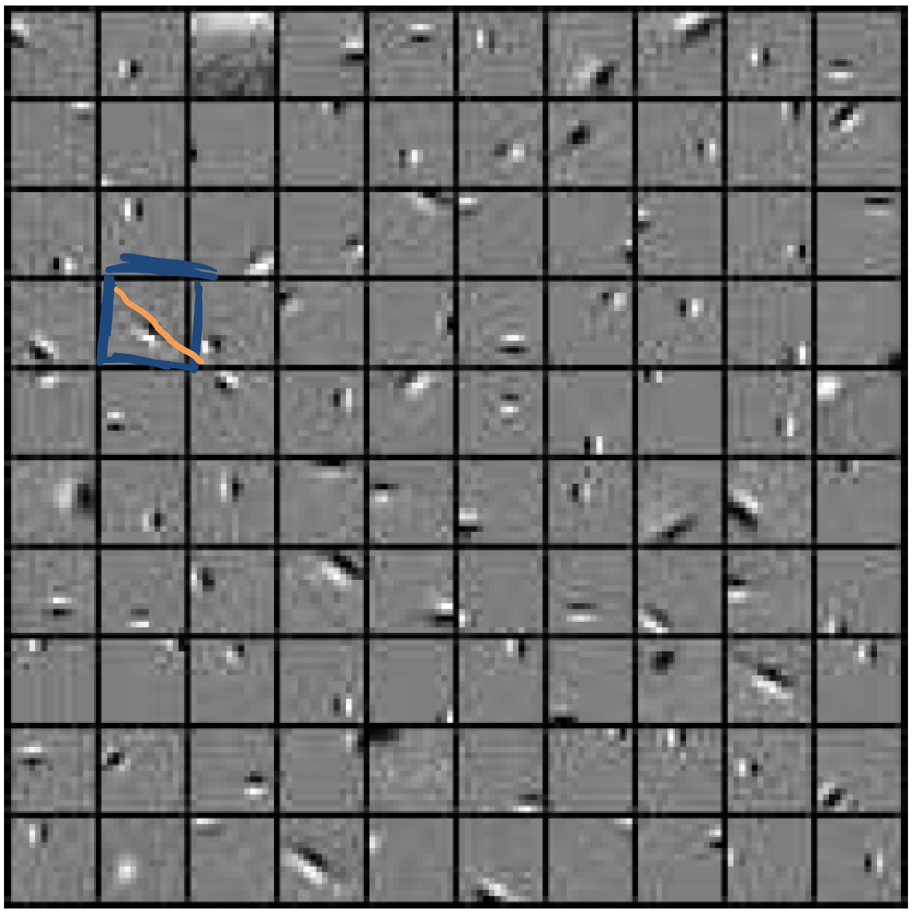
The effect of L1 regularization on PCA

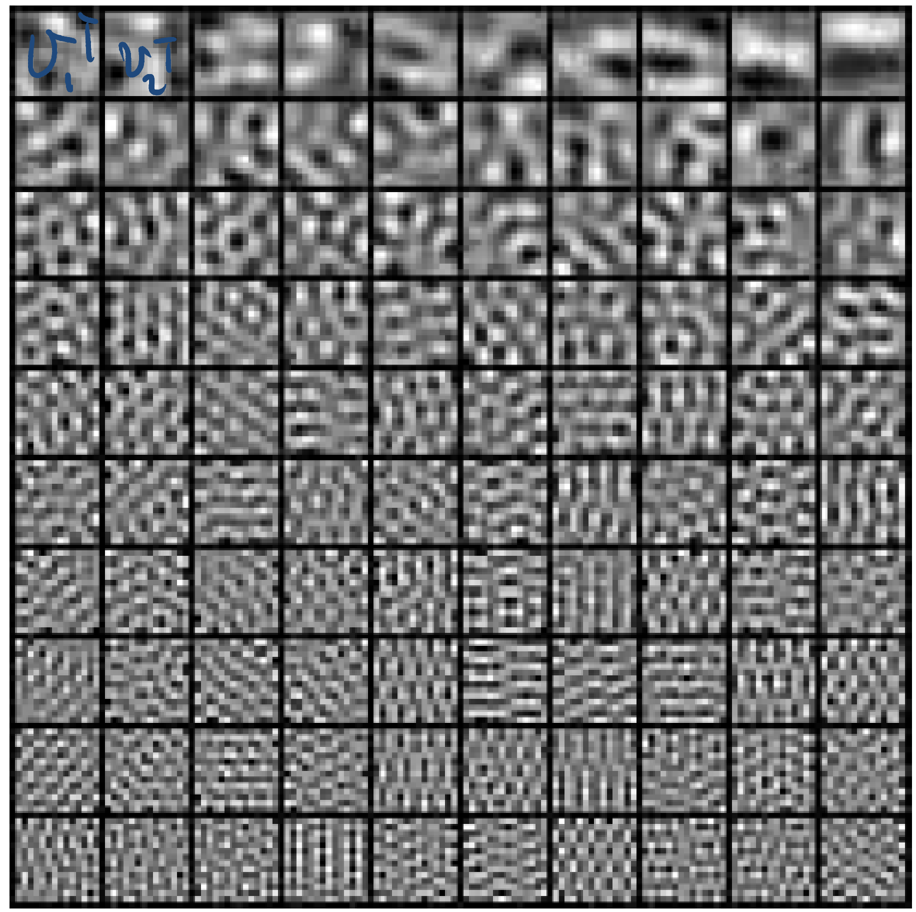
$$\mathbf{B}^*, \mathbf{C}^* = \boxed{\arg\min_{\mathbf{B},\mathbf{C}} ||\mathbf{X} - \mathbf{BC}||_2^2} + \lambda ||\mathbf{C}||_1$$

$$s.t. \quad ||\mathbf{b}_j||_2^2 = 1, \ \forall j.$$

$B = U\Sigma$

$C = V^T$





Next lecture

In the next lecture, we go back to probability so as to get enough background to understand classification.