# Mining Sparse Representations: Theory, Algorithms, and Applications

Jun Liu, Shuiwang Ji, and Jieping Ye

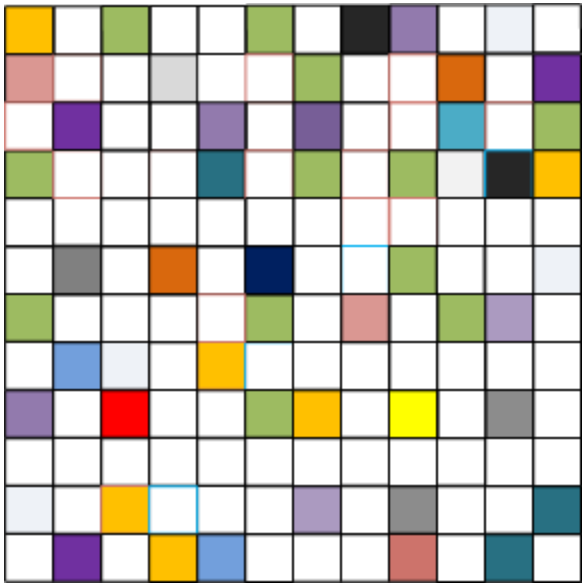
Computer Science and Engineering

The Biodesign Institute

Arizona State University

# What is Sparsity?

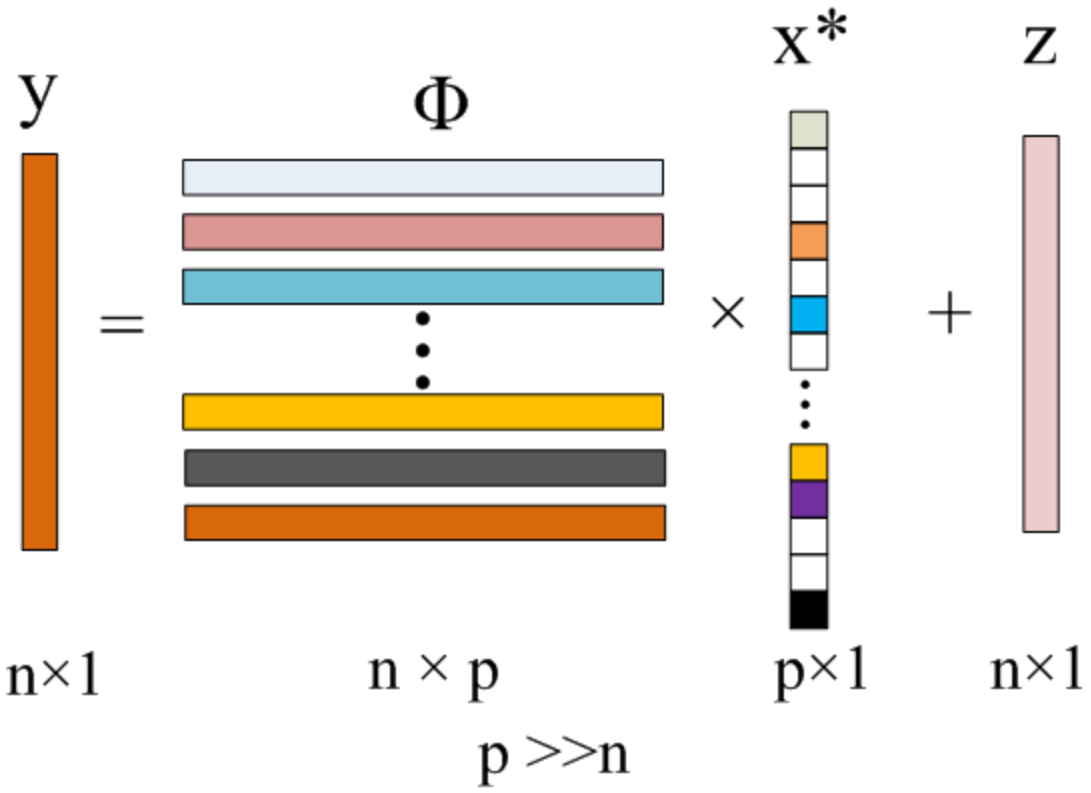
- Many data mining tasks can be represented using a vector or a matrix.
- "Sparsity" implies many zeros in a vector or a matrix.



# Sparsity in Data Mining

- Regression
- Classification
- Multi-Task Learning
- Collaborative Filtering
- Network Construction

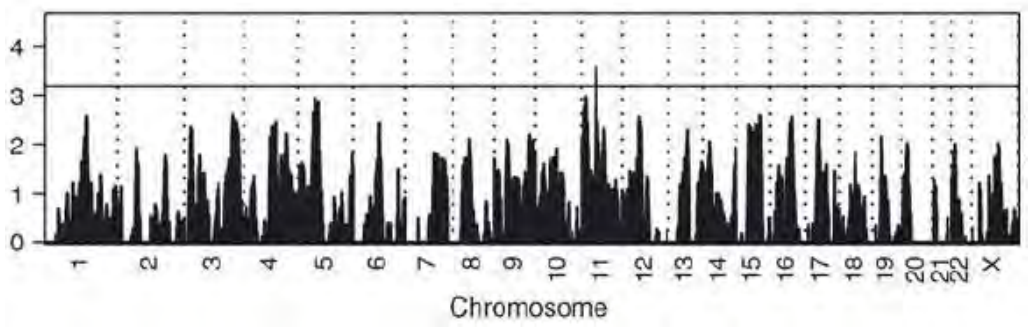
# Regression




- Select a small number of features
  - Lasso

# Application: Genetics

- Find the locations on the genome that influence a trait: e.g. cholesterol level




- Model: y = Ax + z (x is very sparse)

$y_i$ : measured cholesterol level of patient $i$

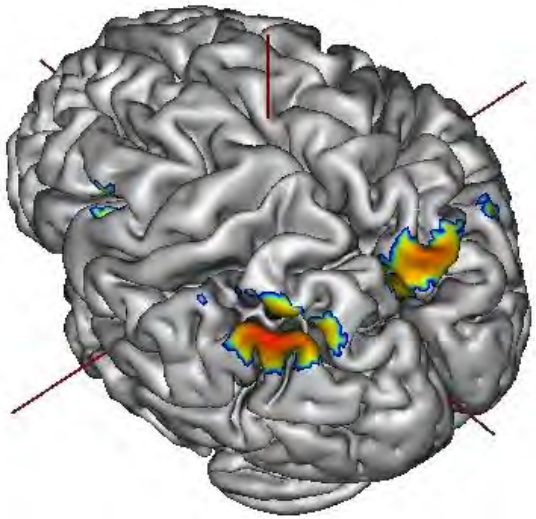
$A$: genotype matrix; $i$th row $\rightarrow$ genotype of patient $i$
e.g. $A_{ij}$ number of alleles of type $a$ at location $j$

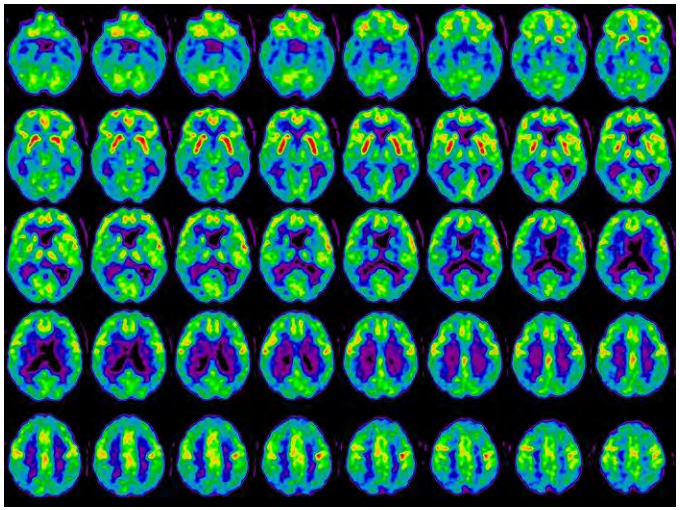
$z_i$ environmental factors (not accounted by genetics) for patient $i$

Typical dimensions:

Number of columns: about 500,000, number of rows: a few thousands

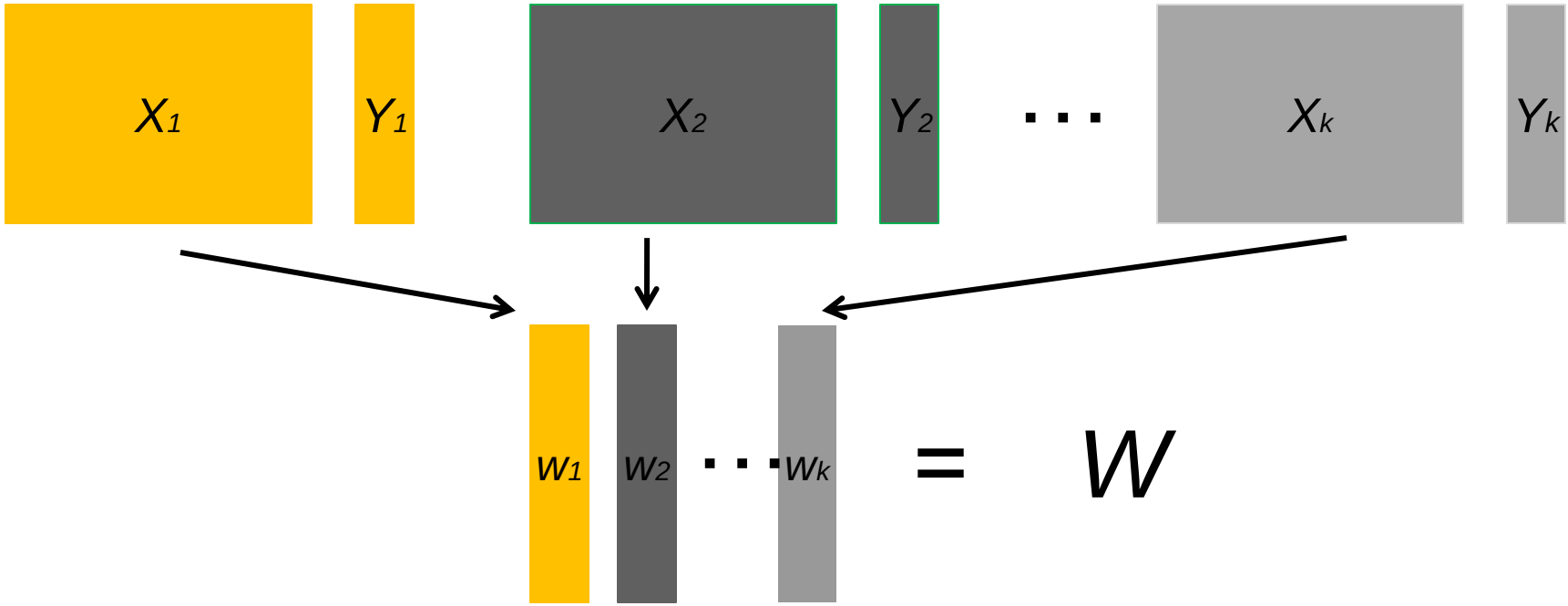
# Application: Neuroscience





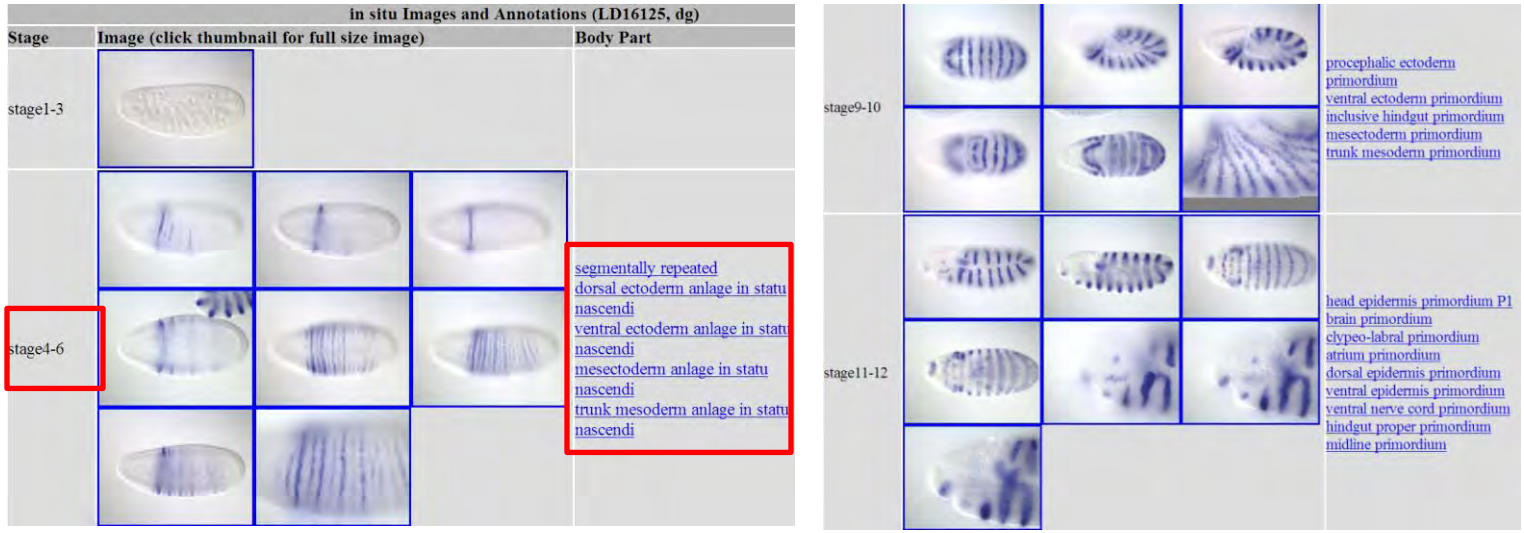
- Neuroimages for studying Alzheimer's Disease

# Multi-Task/Class Learning




- Key Challenge: How to capture the shared information among multiple tasks?
  - Shared features (group Lasso)
  - Shared low-dimensional subspace (trace norm)

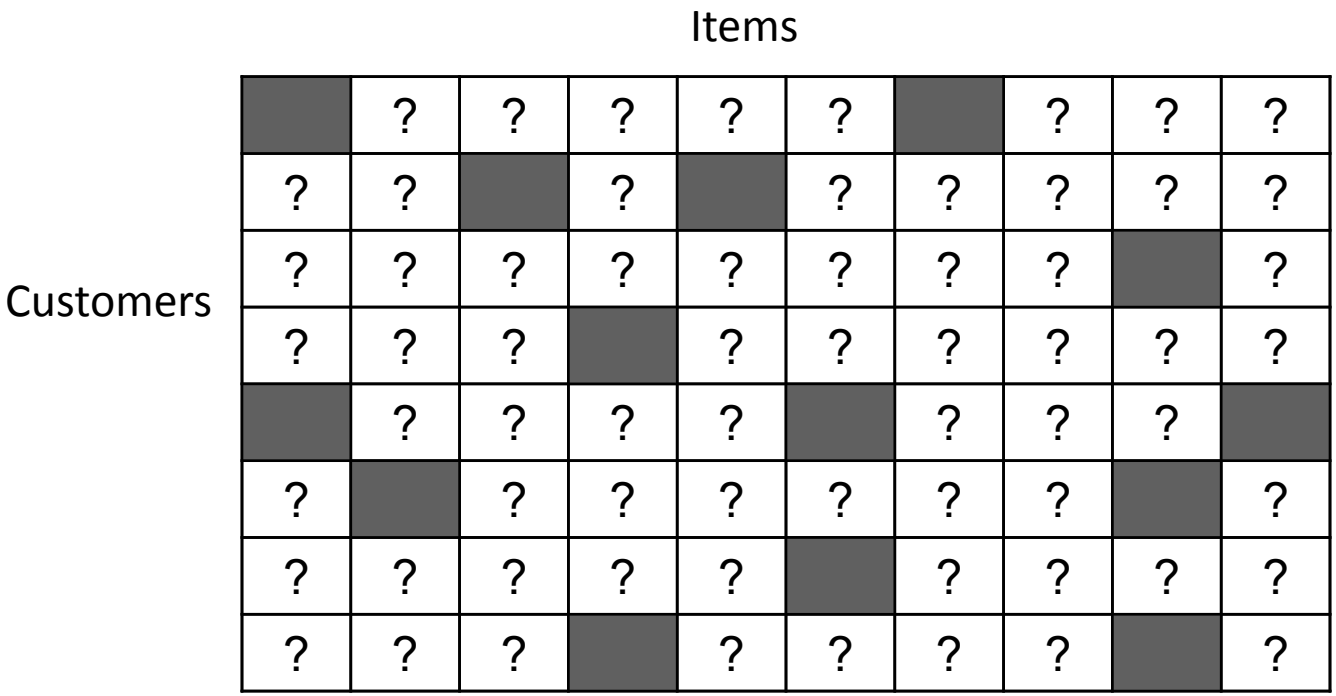
# Application: Biological Image Annotation




http://www.fruitfly.org/cgi-bin/ex/insitu.pl

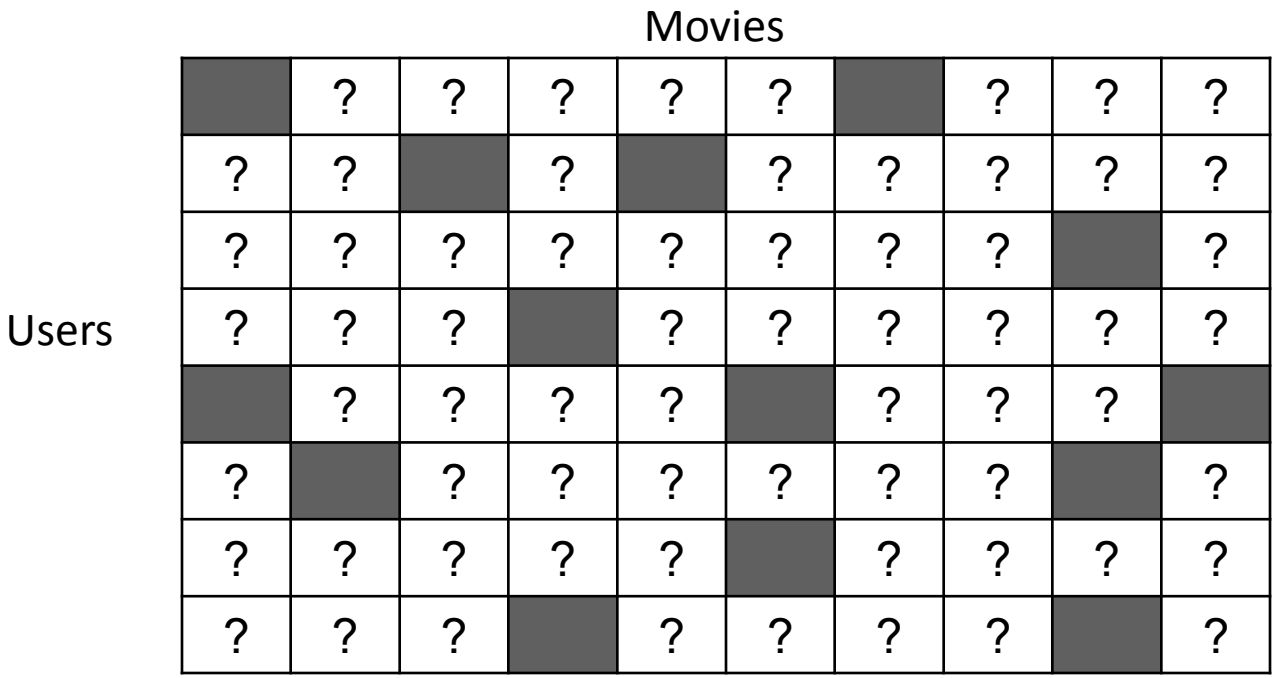
- Document expression patterns over 6,000 genes with over 70,000 images
- Annotated with body part keywords

# Collaborative Filtering



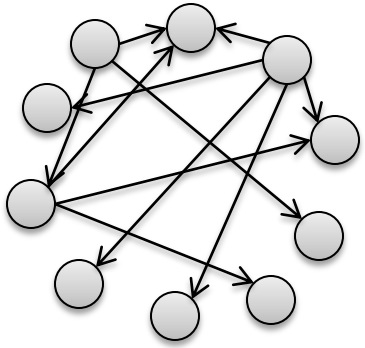

- Customers are asked to rank items
- Not all customers ranked all items
- Predict the missing rankings

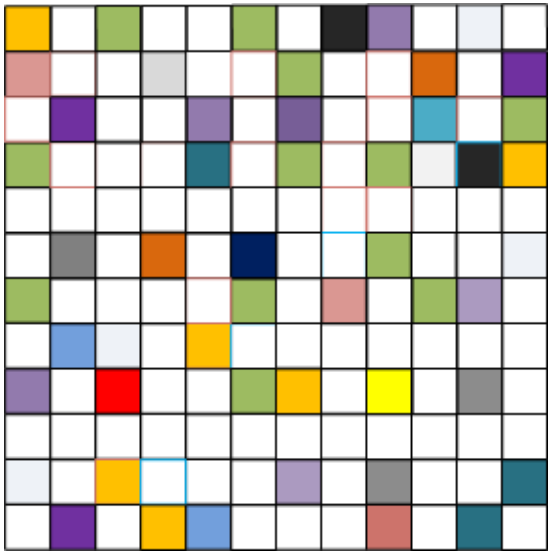
# Application: The Netflix Challenge




- About a million users and 25,000 movies
- Known ratings are sparsely distributed
- Predict unknown ratings

# Network Construction

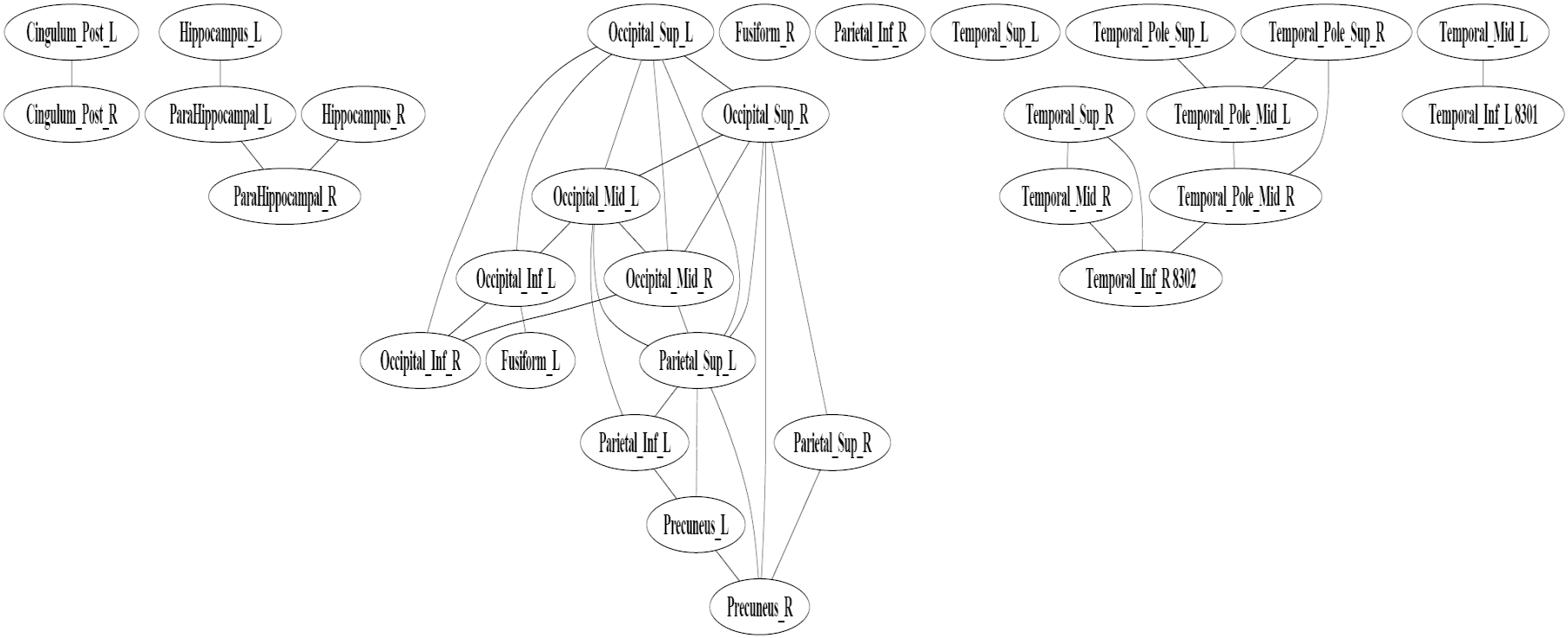




Equivalent matrix representation

Sparsity: Each node is linked to a small number of neighbors in the network.

# Application: Brain Network




Brain Connectivity of different regions for Alzheimer's Disease (Sun et al., KDD 2009)

# A Unified Model for Sparse Learning

- Let x be the model parameter to be estimated. A commonly employed model for estimating x is

$$\min \ \text{loss}(x) + \lambda \ \text{penalty}(x) \qquad (1)$$

- (1) is equivalent to the following model:

$$\begin{aligned} &\min \ \text{loss}(x) \\ &\text{s.t.} \quad \text{penalty}(x) \leq z \end{aligned} \qquad (2)$$

# Loss Functions

- Least squares
- Logistic loss
- Hinge loss
- …

# Penalty Functions: Nonconvex

- Zero norm

  penalty(x)=the number of nonzero elements

- Advantages

  The sparsest solution

- Disadvantages:

  Not a valid norm, nonconvex, NP-hard

- Extension to the matrix case: rank

# Penalty Functions: Convex

- $L_1$ norm

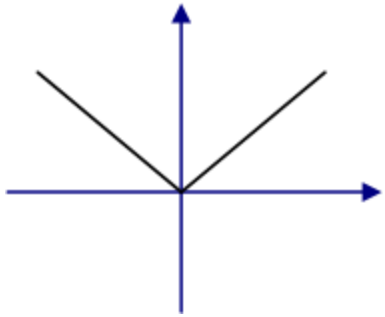
$$\text{penalty}(x)=\|x\|_1=\sum_i |x_i|$$

- Advantages:
  - Valid norm
  - Convex
  - Computationally tractable
  - Theoretical properties
  - Many applications
  - Various Extensions

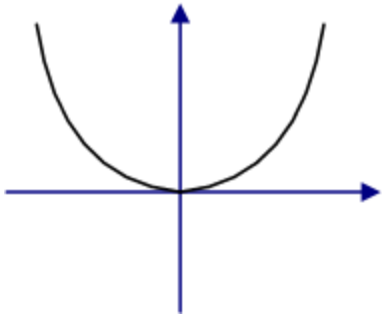
In this tutorial, we discuss *sparse representation based on $L_1$ and its extensions.*

# Why does $L_1$ Induce Sparsity?

Analysis in 1D (comparison with $L_2$)



$0.5\times(x-v)^2 + \lambda|x|$



$0.5\times(x-v)^2 + \lambda x^2$

| | |
|---|---|
| If $v \geq \lambda$, $x=v-\lambda$<br>If $v \leq -\lambda$, $x=v+\lambda$<br>Else, $x=0$ | $x=v/(1+2\lambda)$ |
| Nondifferentiable at 0 | Differentiable at 0 |

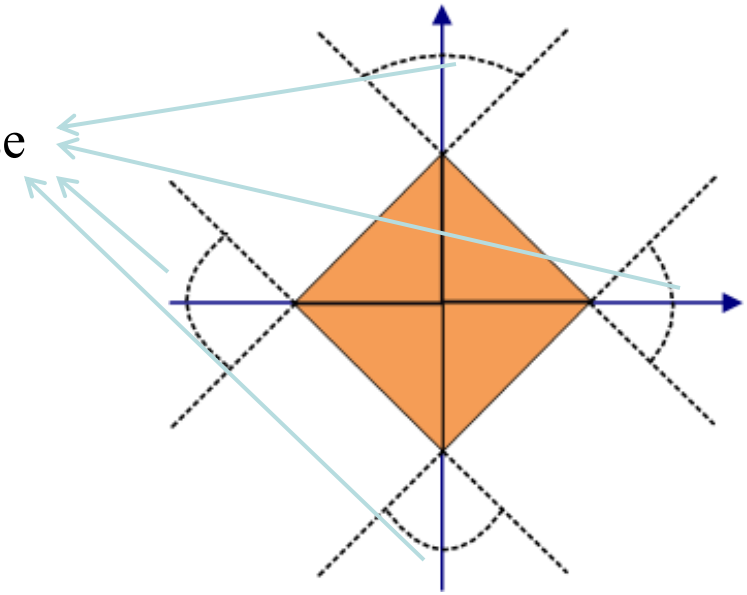
# Why does $L_1$ Induce Sparsity?

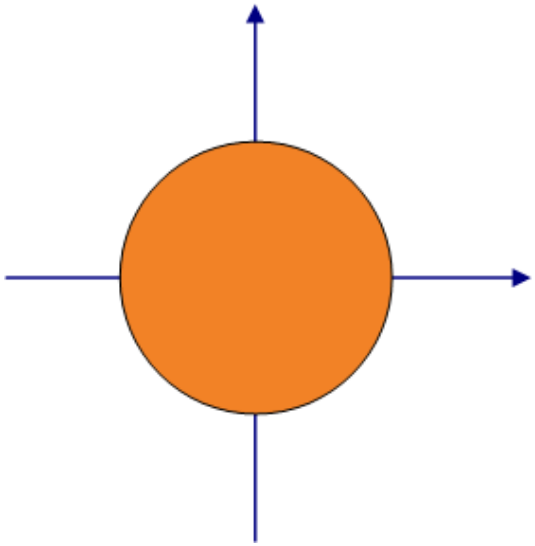
Understanding from the projection

$\min \text{loss}(x)$ $\quad$ $\min 0.5\|x-v\|^2$
s.t. $\|x\|_1 \leq 1$ $\quad$ s.t. $\|x\|_1 \leq 1$

$\min \text{loss}(x)$ $\quad$ $\min 0.5\|x-v\|^2$
s.t. $\|x\|_2 \leq 1$ $\quad$ s.t. $\|x\|_2 \leq 1$

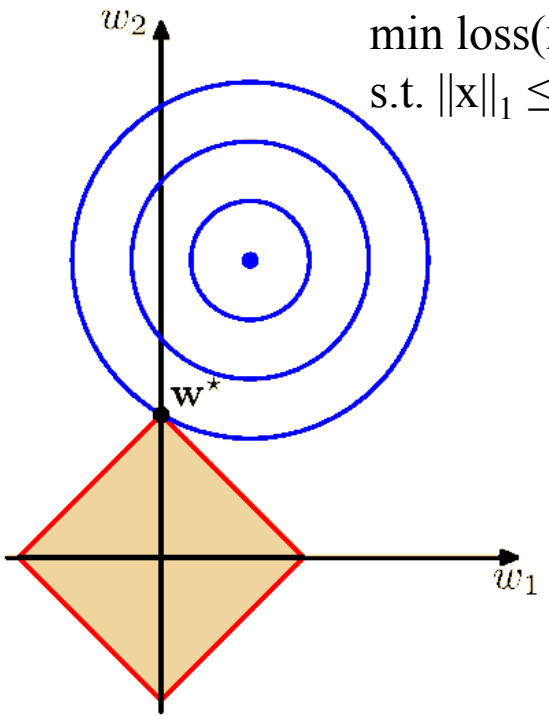
Sparse

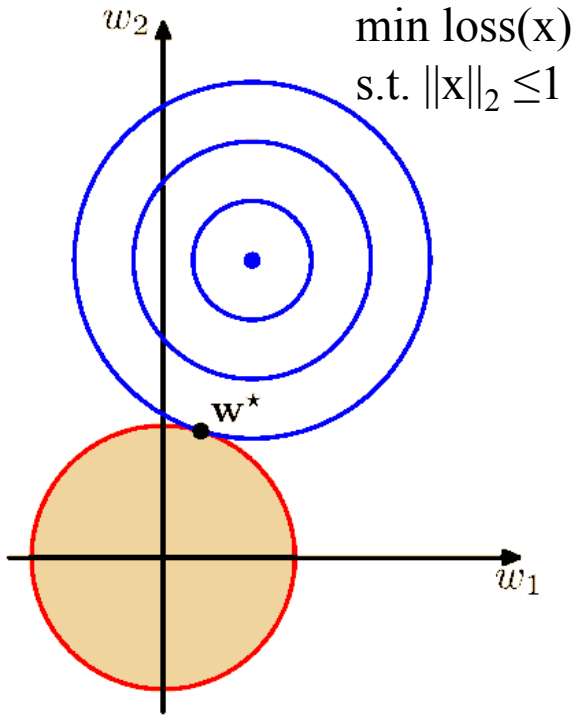




# Why does $L_1$ Induce Sparsity?

Understanding from constrained optimization




(Bishop, 2006)

# Sparsity via $L_1$

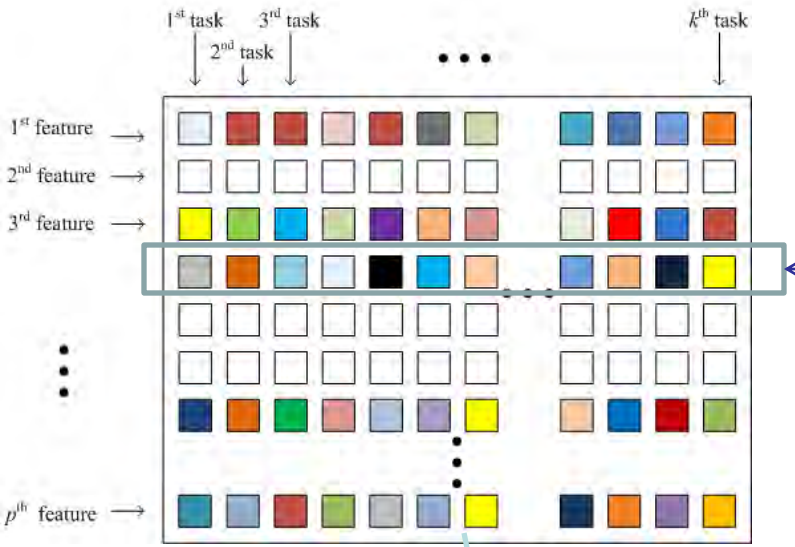
Lasso,
Dantzig selector

$\|x\|_1$

Summation of the absolute values

Application: Regression/Classification

# Sparsity via $L_1/L_q$

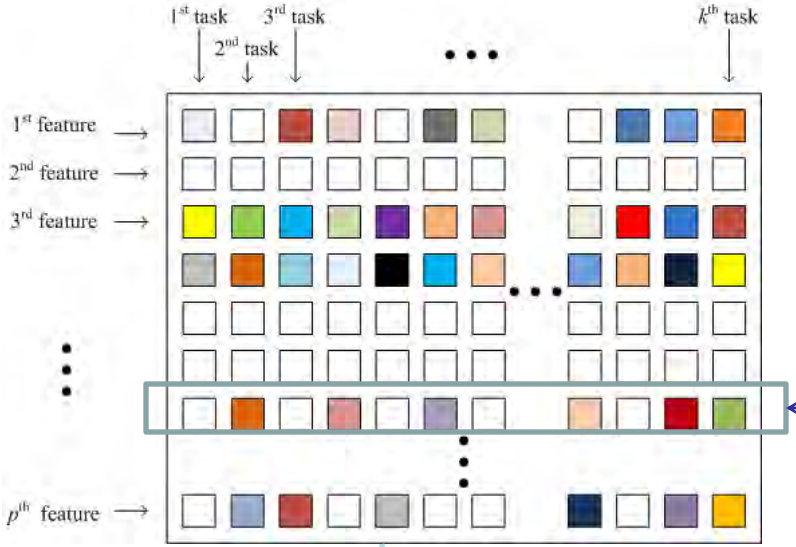



$$\sum_{i=1}^{k} \|X^i\|_q$$

Group Lasso,
Group Dantzig

Application: Multi-Task Learning

# Sparsity via $L_1$+ $L_1/L_q$




$$\lambda_1 \sum_{i=1}^{k} \|X^i\| + \lambda_q \sum_{i=1}^{k} \|X^i\|_q$$

Sparse Group Lasso

Application: Multi-Task Learning

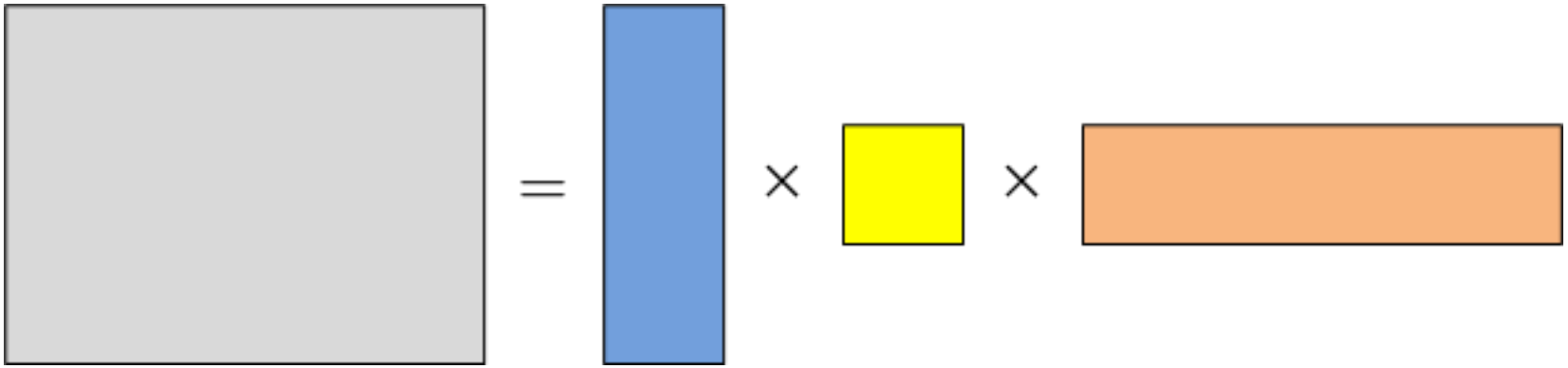
# Sparsity via Fused Lasso

$$\lambda_1 \sum_{i=1}^{p} |x_i| + \lambda_2 \sum_{i=1}^{p-1} |x_i - x_{i+1}|$$

Fused Lasso,
Time Varying Network

Application: Regression with Ordered Features

# Sparsity via Trace Norm



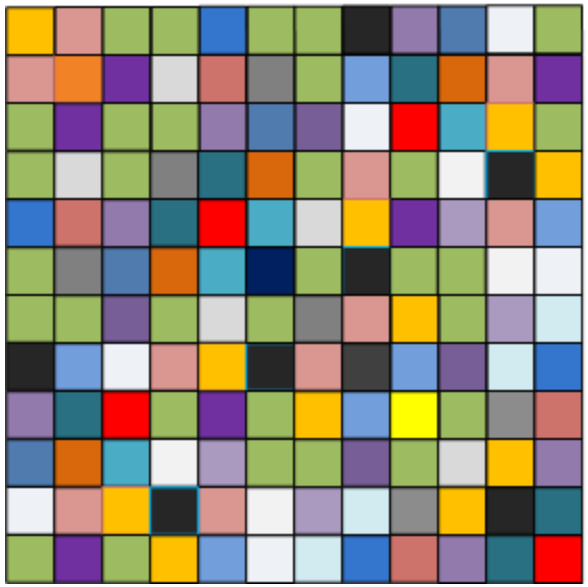
$$\|X\|_* = \sum_{i=1}^{k} \sigma_i \qquad X = U\Sigma V^{\mathrm{T}}$$

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)$$

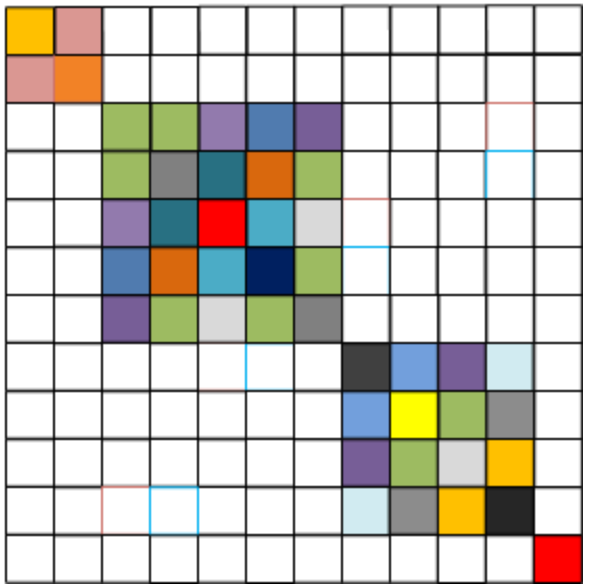
Application: Multi-Task Learning, Matrix Completion

# Sparse Inverse Covariance

Inverse Covariance Matrix



Sparse Inverse Covariance Estimation



Connectivity Study

$$\|X\|_1, X \succeq 0$$

# Goal of This Tutorial

- Introduce various sparsity-induced norms
  - Map sparse models to different applications
- Efficient implementations
  - Scale sparse learning algorithms to large-size problems

# Outline

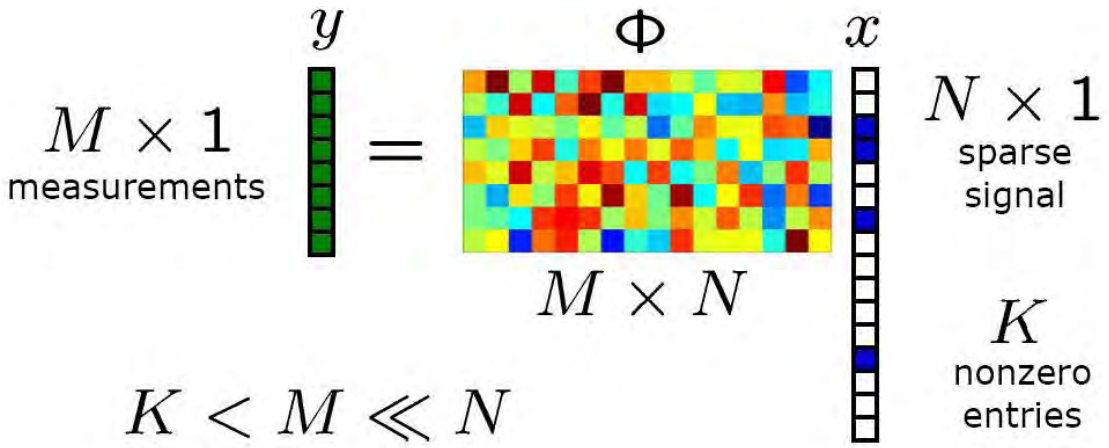
- Sparsity via $L_1$
- Sparsity via $L_1/L_q$
- Sparsity via Fused Lasso
- Sparse Inverse Covariance Estimation
- Sparsity via Trace Norm
- Implementations and the SLEP package
- Trends in Sparse Learning

# Compressive Sensing

(Donoho, 2004; Candes and Tao, 2008; Candes and Wakin, 2008)

## Principle:

"sparse signal statistics can be recovered from a small number of ***nonadaptive linear measurements***"




**NP hard**

$P_0$

$$\min \quad \|x\|_0 \qquad \text{s.t.} \quad \Phi x = y$$

$P_1$

$$\min \quad \|x\|_1 \qquad \text{s.t.} \quad \Phi x = y$$

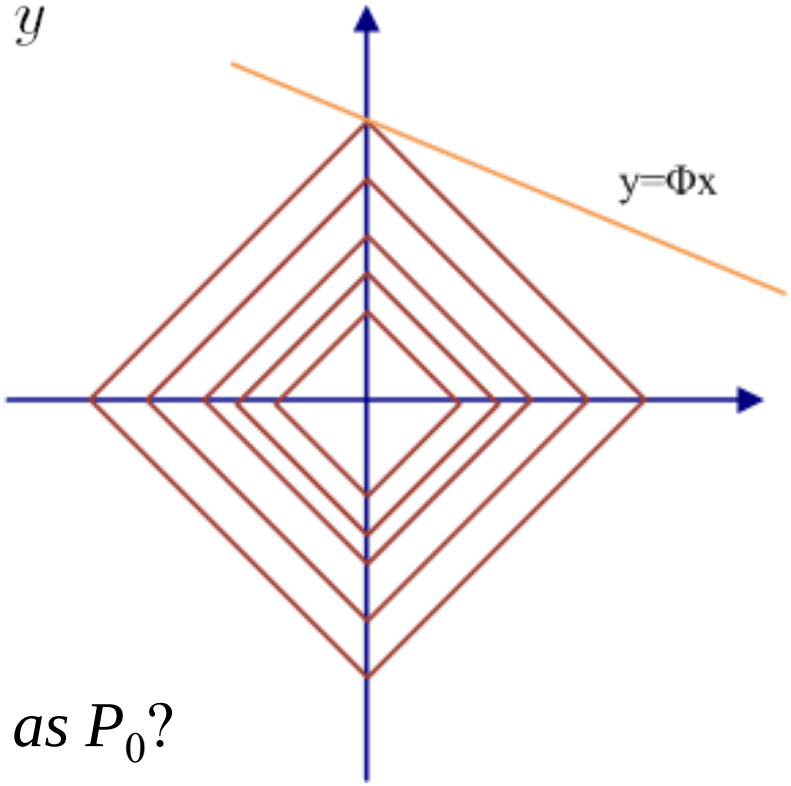
# Basis Pursuit

## (Chen, Donoho, and Saunders, 1999)

$$\begin{aligned} \min \quad & \|x\|_1 \\ \text{s.t.} \quad & \Phi x = y \end{aligned}$$

Linear Programming

min <1, *t*>
s.t. $y=\Phi x$, $-t \le x \le t$




Question:
When *can* $P_1$ *obtain the same result as* $P_0$?

# Sparse Recovery and RIP

$\Phi$ satisfies the *K*-restricted isometry property with constant $\delta_K$ if $\delta_K$ is the smallest constant satisfying

$$(1 - \delta_K)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_K)\|x\|_2^2$$

for every *K*-sparse vector *x*.

**Theorem (E. J. Candès (2008))**

*If $\delta_{2k} < \sqrt{2} - 1$, then for all k-sparse vectors x such that $\Phi x = b$, the solution of $(P_1)$ is equal to the solution of $(P_0)$.*

# Extensions to the Noisy Case

$$\begin{aligned} \min \quad & \|x\|_1 \\ \text{s.t.} \quad & \Phi x = y \end{aligned}$$

noise

$$y = \Phi x + z$$

$$\begin{aligned} \min \quad & \|x\|_1 \\ \text{s.t.} \quad & \|\Phi x - y\|_2 \leq \epsilon \end{aligned}$$

Basis pursuit De-Noising
(Chen, Donoho, and Saunders, 1999)

$$\begin{aligned} \min \quad & \frac{1}{2}\|\Phi x - y\|_2^2 \\ \text{s.t.} \quad & \|x\|_1 \leq \rho \end{aligned}$$

Lasso (Tibshirani, 1996)

$$\frac{1}{2}\|\Phi x - y\|_2^2 + \lambda \|x\|_1$$

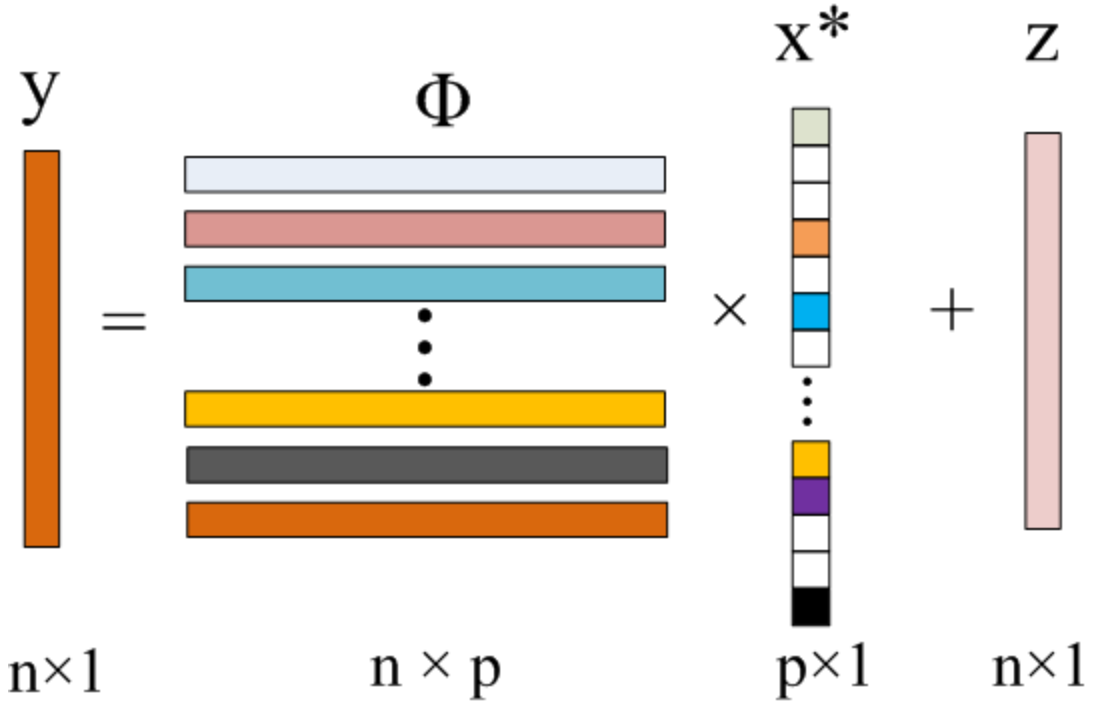
Regularized counterpart of Lasso

$$\begin{aligned} \min \quad & \|x\|_1 \\ \text{s.t.} \quad & \|\Phi^{\mathrm{T}}(\Phi x - y)\|_\infty \leq \epsilon \end{aligned}$$

Dantzig selector (Candes and Tao, 2007)

# Lasso

## (Tibshirani, 1996)




$p >> n$

$$\frac{1}{2}\|\Phi x - y\|_2^2 + \lambda\|x\|_1$$

# Theory of Lasso

(Zhao and Yu, 2006; Wainwright 2009;
Meinshausen and Yu, 2009; Bickel, Ritov, and Tsybakov, 2009)

- Support Recovery

  $\sup(x)=\sup(x^*)$?

- Sign Recovery

  $\text{sign}(x)=\text{sign}(x^*)$?

- $L_1$ Error

  $\|x-x^*\|_1$

- $L_2$ Error

  $\|x-x^*\|_2$

# $L_1$ Error

(Bickel, Ritov, and Tsybakov, 2009)

**Theorem (Bickel, Ritov, and Tsybakov, 2009)**

*Let $z_i$ be independent $\mathcal{N}(0, \sigma^2)$ random variables. Let all the diagonal elements of the matrix $\Phi^{\mathrm{T}}\Phi/n$ be equal to 1, and $M(x^*) = s$ where $1 \leq s \leq M$. Let Assumption $RE(s, 3)$ be satisfied. Let*

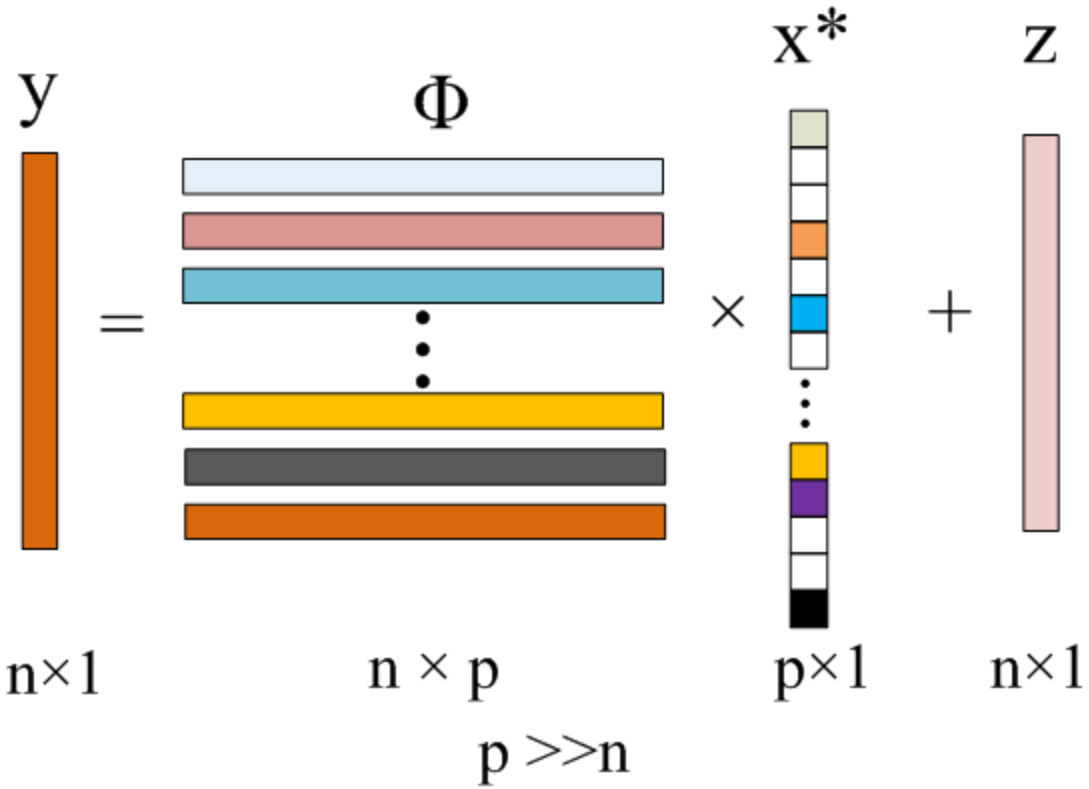
$$\hat{x} = \arg\min\{\frac{1}{n}\|y - \Phi x\|_2^2 + 2\lambda\|x\|_1\},$$

$$\lambda = A\delta\sqrt{\frac{\log M}{n}}$$

*and $A > 2\sqrt{2}$. Then for all $n \geq 1$, with probability at least $1 - M^{1-A^2/8}$, we have*

$$\|\hat{x} - x^*\|_1 \leq \frac{16A}{\kappa^2}\delta s\sqrt{\frac{\log M}{n}}.$$

# Dantzig Selector




$$\begin{aligned} \min \quad & \|x\|_1 \\ \text{s.t.} \quad & \|\Phi^{\mathrm{T}}(\Phi x - y)\|_\infty \leq \epsilon \end{aligned}$$

# Theory of Dantzig Selector

(Candes and Tao, 2007; Bickel, Ritov, and Tsybakov, 2009)

**Theorem (Bickel, Ritov, and Tsybakov, 2009)**

*Let $z_i$ be independent $\mathcal{N}(0, \sigma^2)$ random variables. Let all the diagonal elements of the matrix $\Phi^T\Phi/n$ be equal to 1, and $M(x^*) = s$ where $1 \leq s \leq M$. Let Assumption $RE(s, 1)$ be satisfied. Let*

$$\hat{x} = \arg \min_{\|\frac{1}{n}\Phi^T(y-\Phi x)\|_\infty \leq \lambda} \|x\|_1,$$

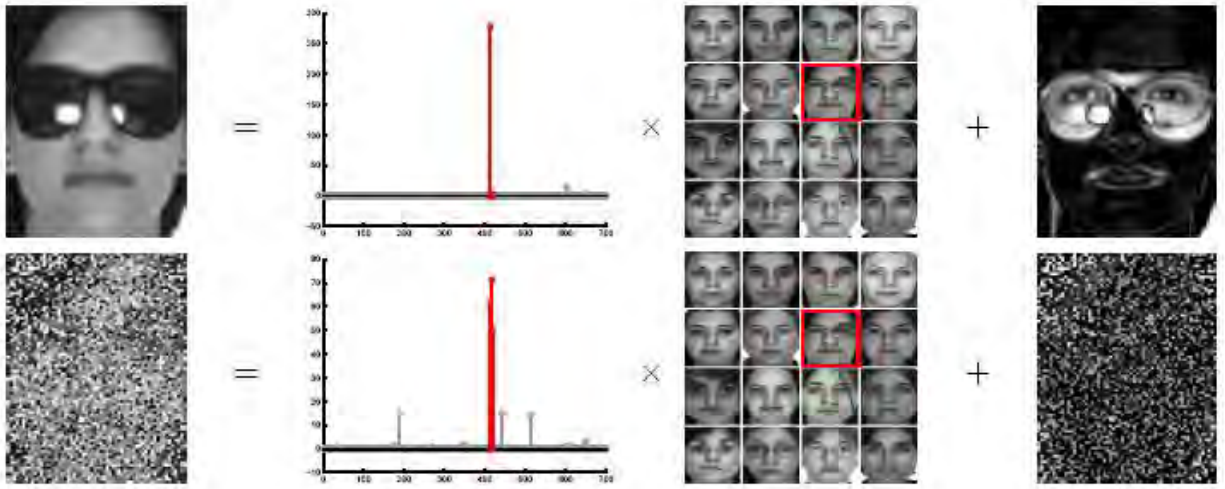
$$\lambda = A\delta\sqrt{\frac{\log M}{n}}$$

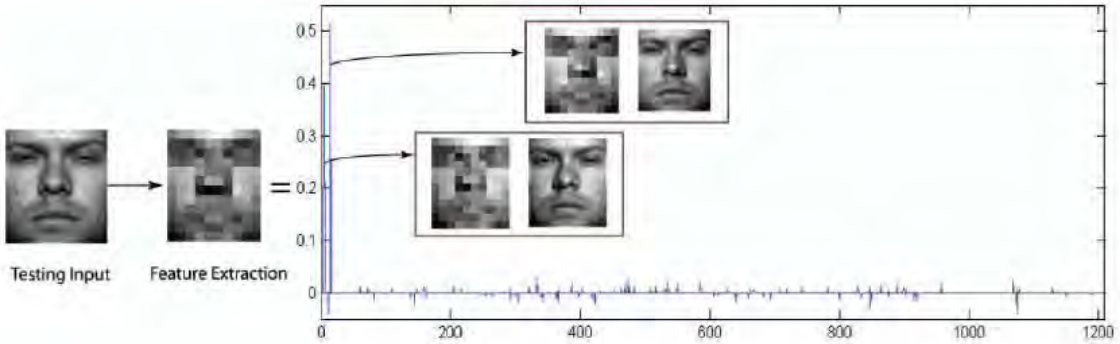
*and $A > \sqrt{2}$. Then for all $n \geq 1$, with probability at least $1 - M^{1-A^2/2}$, we have*

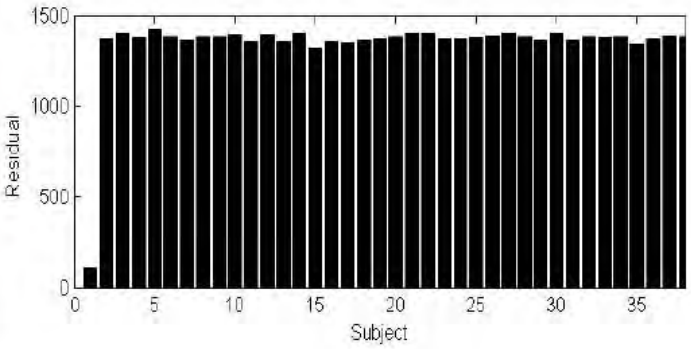
$$\|\hat{x} - x^*\|_1 \leq \frac{8A}{\kappa^2}\delta s\sqrt{\frac{\log M}{n}}.$$

# Face Recognition

(Wright et al. 2009)





# 1-norm SVM

(Zhu et al. 2003)

$$\min_{\beta_0,\beta} \quad \sum_{i=1}^{n}\left[1 - y_i\left(\beta_0 + \sum_{j=1}^{q}\beta_j h_j(x_i)\right)\right]_+$$

Hinge Loss

$$\text{s.t.} \quad \|\beta\|_1 = |\beta_1| + \cdots + |\beta_q| \le s,$$

1-norm constraint

Table 1: Simulation results of 1-norm and 2-norm SVM

| | | Test Error (SE) | | | | |
|---|---|---|---|---|---|---|
| | **Simulation** | **1-norm** | **2-norm** | **No Penalty** | $\lvert\mathcal{D}\rvert$ | **# Joints** |
| 1 | No noise input | 0.073 (0.010) | 0.08 (0.02) | 0.08 (0.01) | 5 | 94 (13) |
| 2 | 2 noise inputs | 0.074 (0.014) | 0.10 (0.02) | 0.12 (0.03) | 14 | 149 (20) |
| 3 | 4 noise inputs | 0.074 (0.009) | 0.13 (0.03) | 0.20 (0.05) | 27 | 225 (30) |
| 4 | 6 noise inputs | 0.082 (0.009) | 0.15 (0.03) | 0.22 (0.06) | 44 | 374 (52) |
| 5 | 8 noise inputs | 0.084 (0.011) | 0.18 (0.03) | 0.22 (0.06) | 65 | 499 (67) |

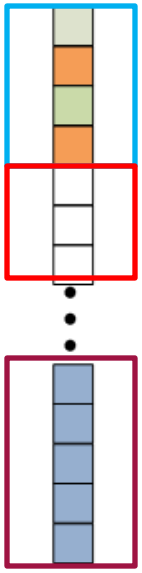
# Outline

- Sparsity via $L_1$
- **Sparsity via $L_1/L_q$**
- Sparsity via Fused Lasso
- Sparse Inverse Covariance Estimation
- Sparsity via Trace Norm
- Implementations and the SLEP package
- Trends in Sparse Learning

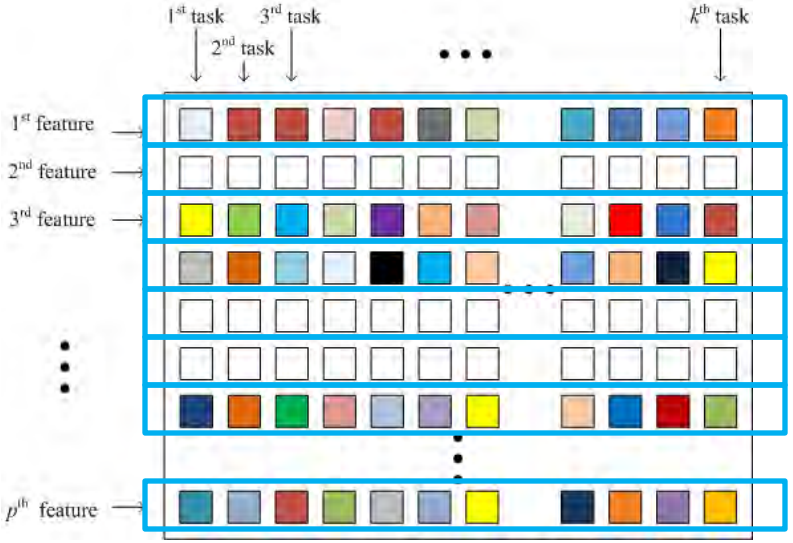
# From $L_1$ to $L_1/L_q$ ($q$>1)?

$L_1$

$L_1/L_q$



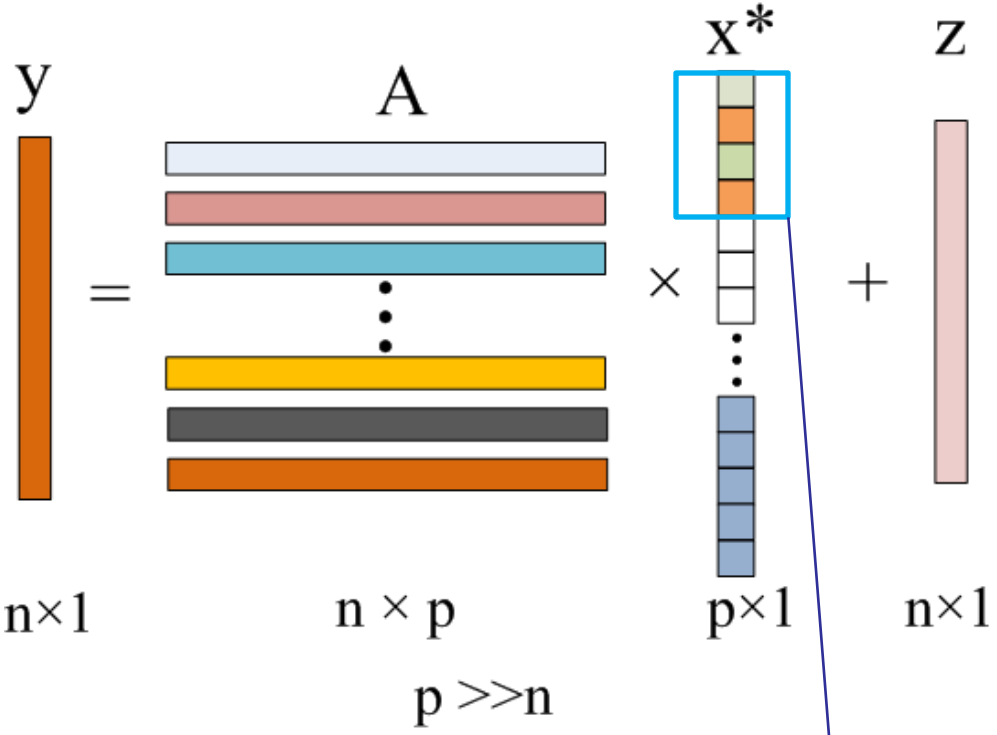
$L_1/L_q$




Most existing work focus on $q$=2, ∞

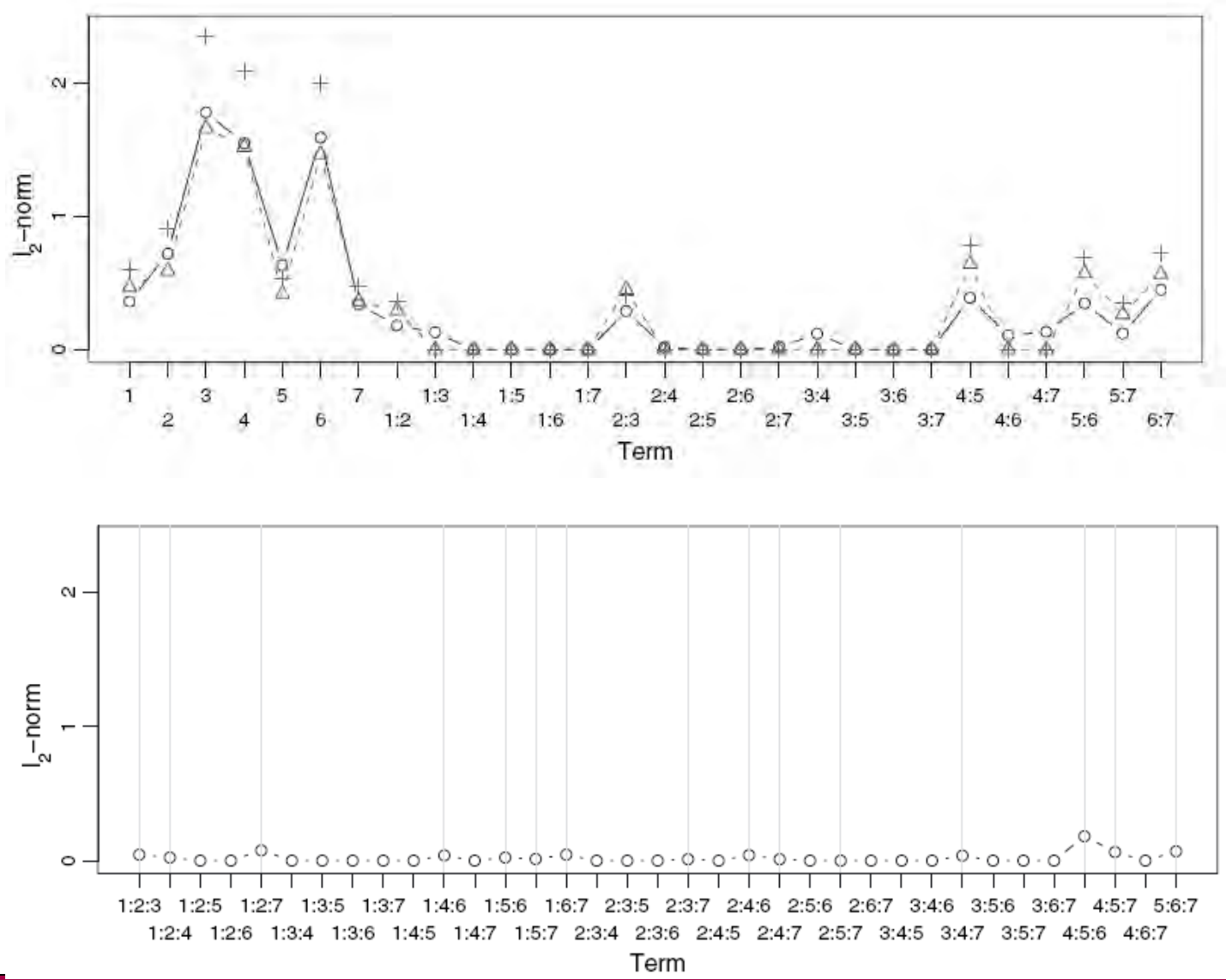
# Group Lasso

## (Yuan and Lin, 2006; Meier et al., 2008)

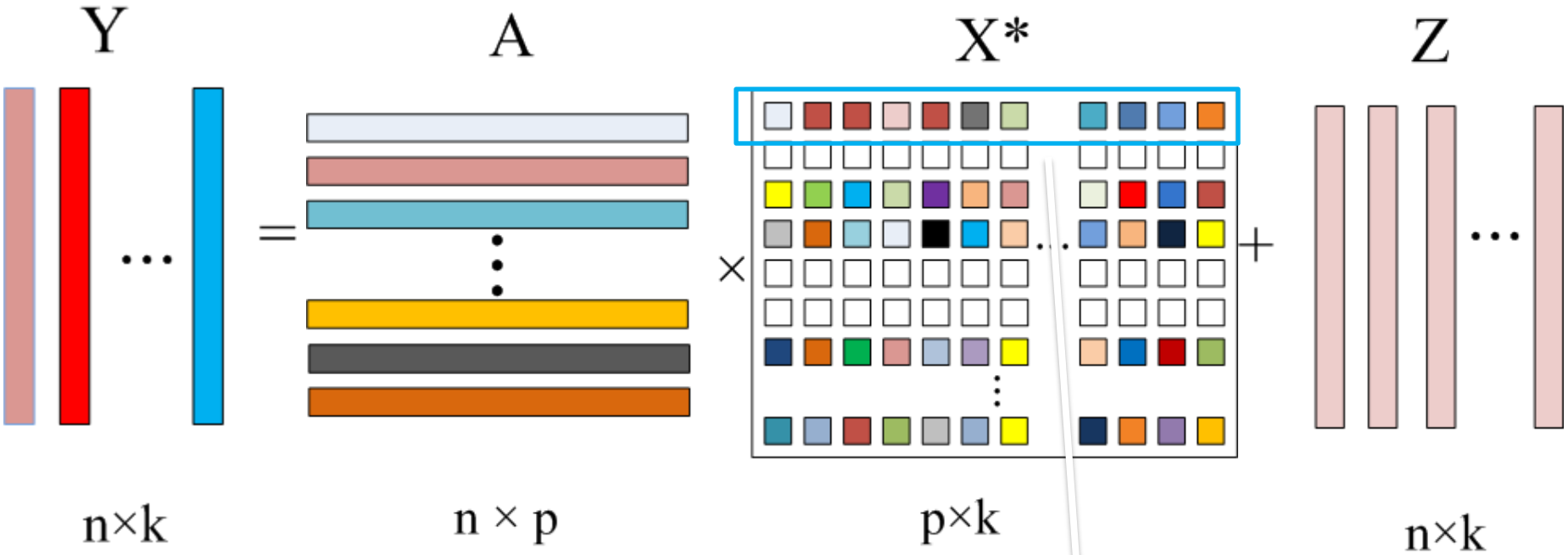



$$\min \frac{1}{2}\|Ax - y\|_2^2 + \lambda \sum_{i=1}^{g} d_i \|x_{G_i}\|_2$$

# Splice Site Detection



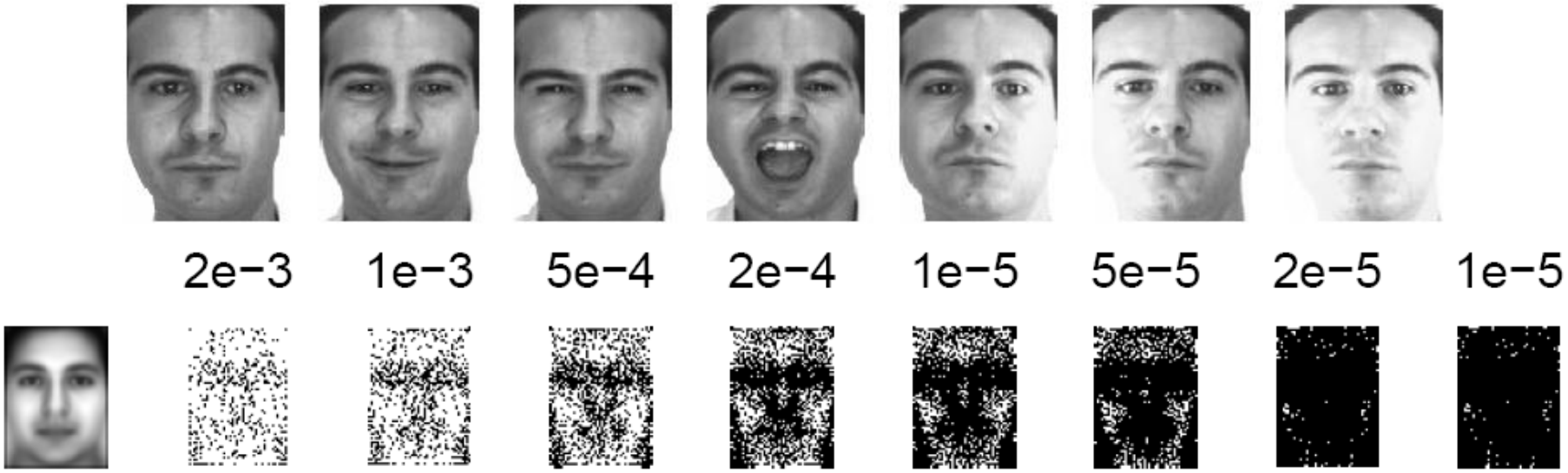
# Multi-task Learning via $L_1/L_q$ Regularization

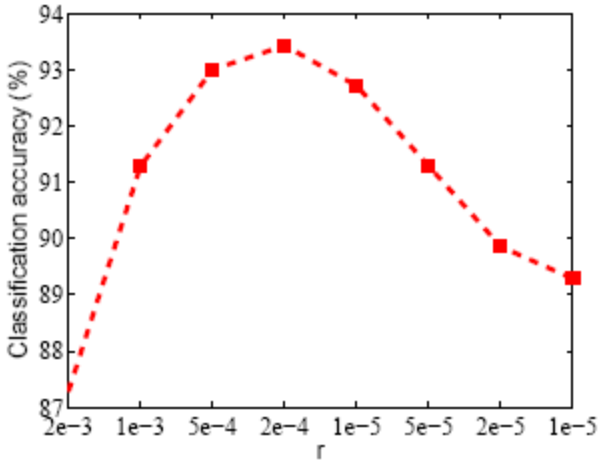


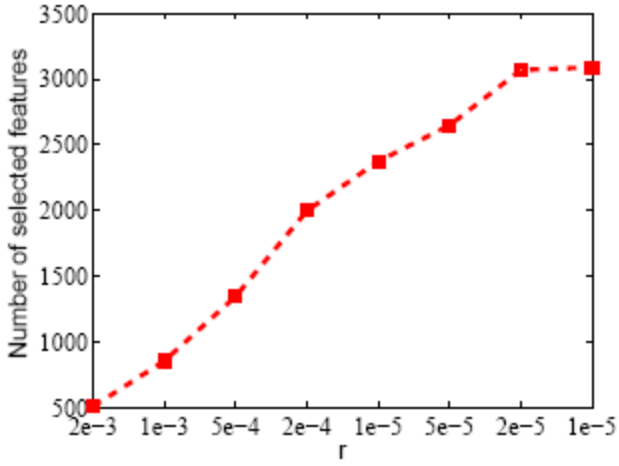

$$\min \frac{1}{2}\|AX - Y\|_2^2 + \lambda \sum_{i=1}^{k} \|X^i\|_q$$

# Face Recognition

## (Liu, Yuan, Chen, and Ye, 2009)

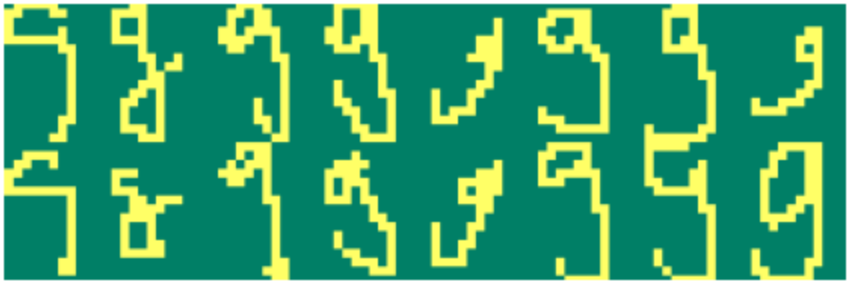


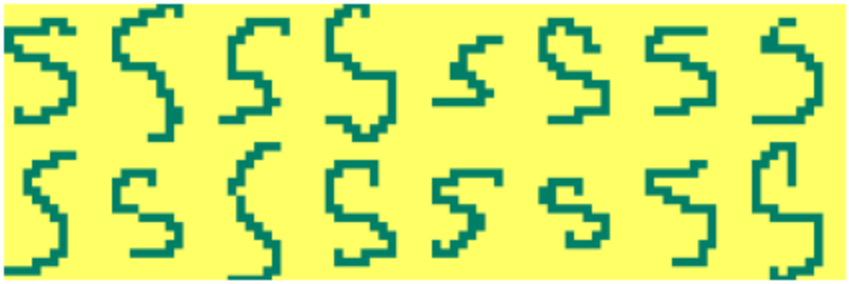



# Writer-specific Character Recognition

## (Obozinski, Taskar, and Jordan, 2006)

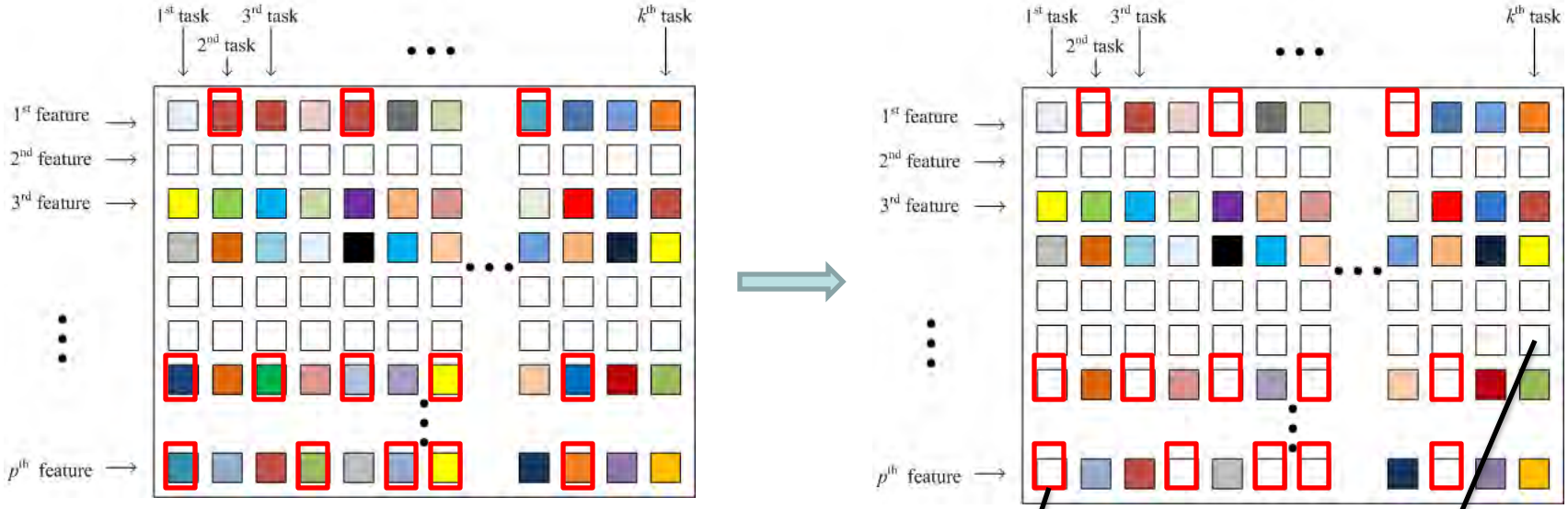




| | strokes : error(%) | | | | pixels: error (%) | | | |
|---|---|---|---|---|---|---|---|---|
| Task | $\ell_1/\ell_2$ | $\ell_1/\ell_1$ | id.$\ell_1$ | pool | $\ell_1/\ell_2$ | $\ell_1/\ell_1$ | id.$\ell_1$ | pool |
| $c/e$ | **2.5** | 3.0 | 3.3 | 3.0 | **4.0** | 8.5 | 9.0 | 4.5 |
| $g/y$ | 8.4 | 11.3 | **8.1** | 17.8 | **11.4** | 16.1 | 17.2 | 18.6 |
| $g/s$ | 3.3 | 3.8 | **3.0** | 10.7 | **4.4** | 10.0 | 10.3 | 6.9 |
| $m/n$ | 4.4 | 4.4 | **3.6** | 4.7 | **2.5** | 6.3 | 6.9 | 4.1 |
| $a/g$ | **1.4** | 2.8 | 2.2 | 2.8 | **1.3** | 3.6 | 4.1 | 3.6 |
| $i/j$ | **8.9** | 9.5 | 9.5 | 11.5 | 12.0 | 14.0 | 14.0 | **11.3** |
| $a/o$ | **2.0** | 2.9 | 2.3 | 3.8 | **2.8** | 4.8 | 5.2 | 4.2 |
| $f/t$ | **4.0** | 5.0 | 6.0 | 8.1 | **5.0** | 6.7 | 6.1 | 8.2 |
| $h/n$ | **0.9** | 1.6 | 1.9 | 3.4 | **3.2** | 14.3 | 18.6 | 5.0 |

# Sparse Group Lasso

## (Peng et al., 2010; Friedman, Hastie, and Tibshirani, 2010; Liu and Ye, 2010)

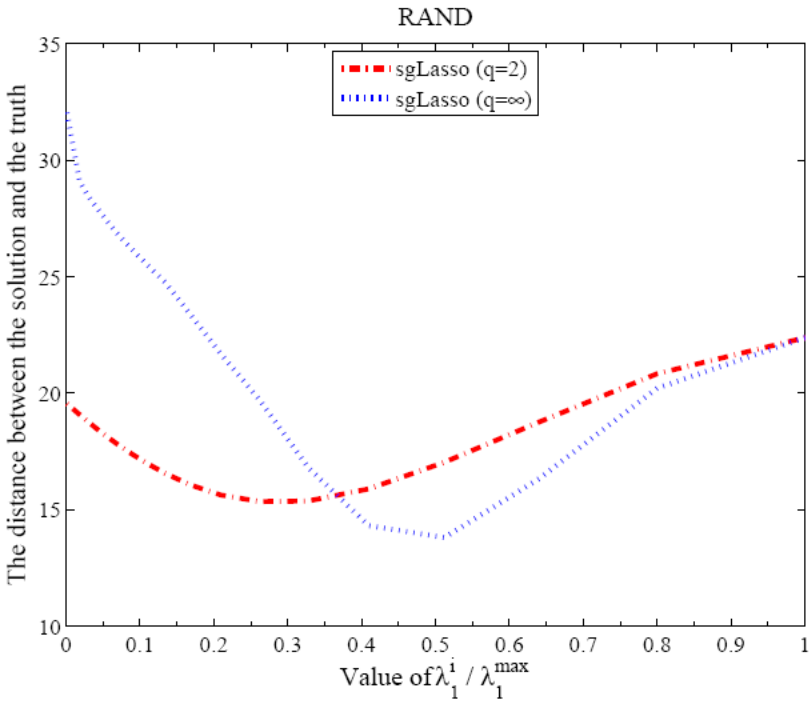


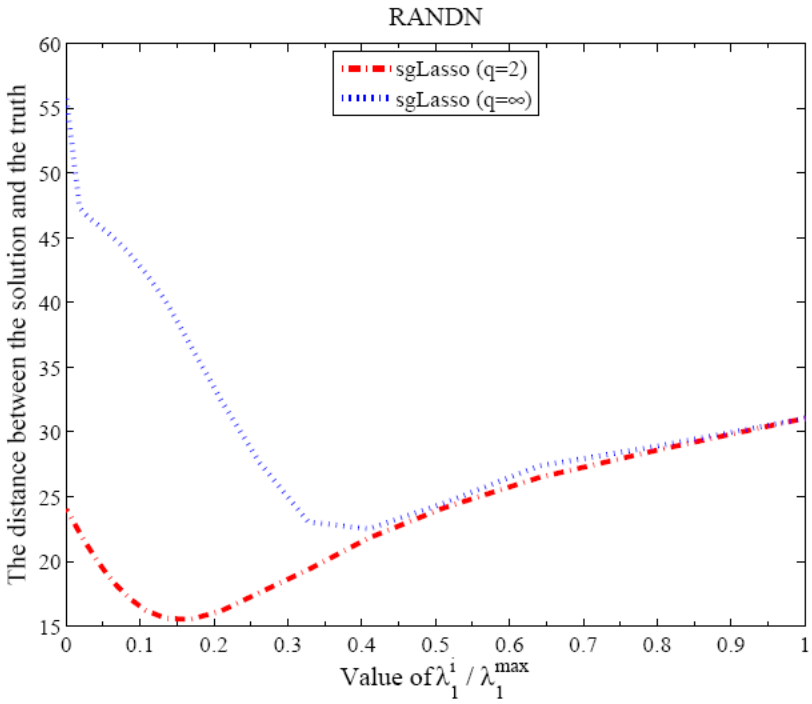

**Simultaneous within and between group sparsity**

$$\min_X \text{loss}(X) + \lambda_1 \sum_{i=1}^{p} \|X^i\|_1 + \lambda_q \sum_{i=1}^{p} \|X^i\|_q$$

# Simulation Study

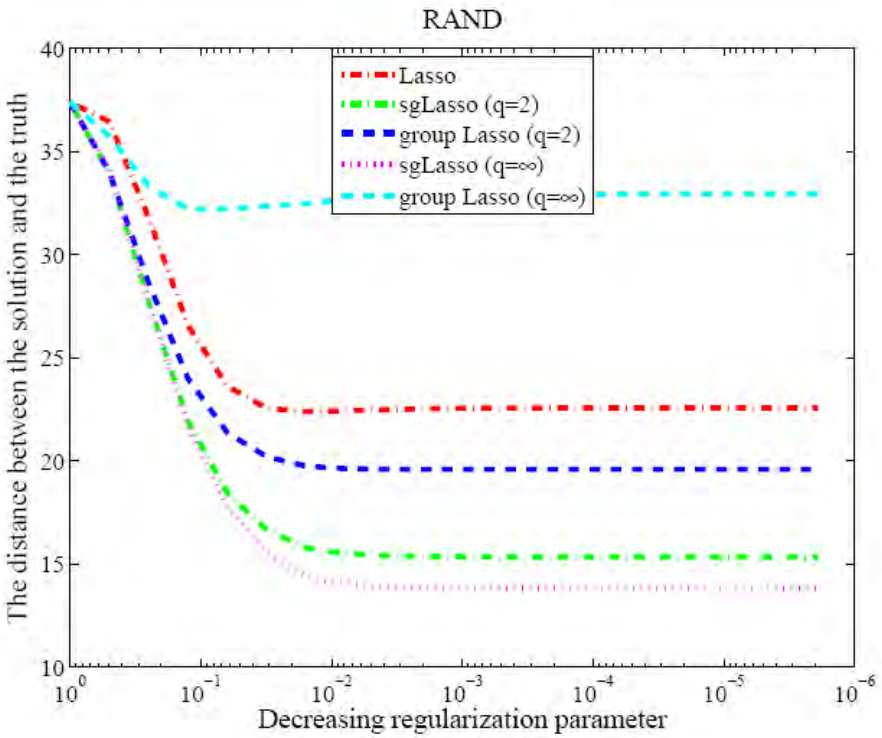
## (Liu and Ye, 2010)

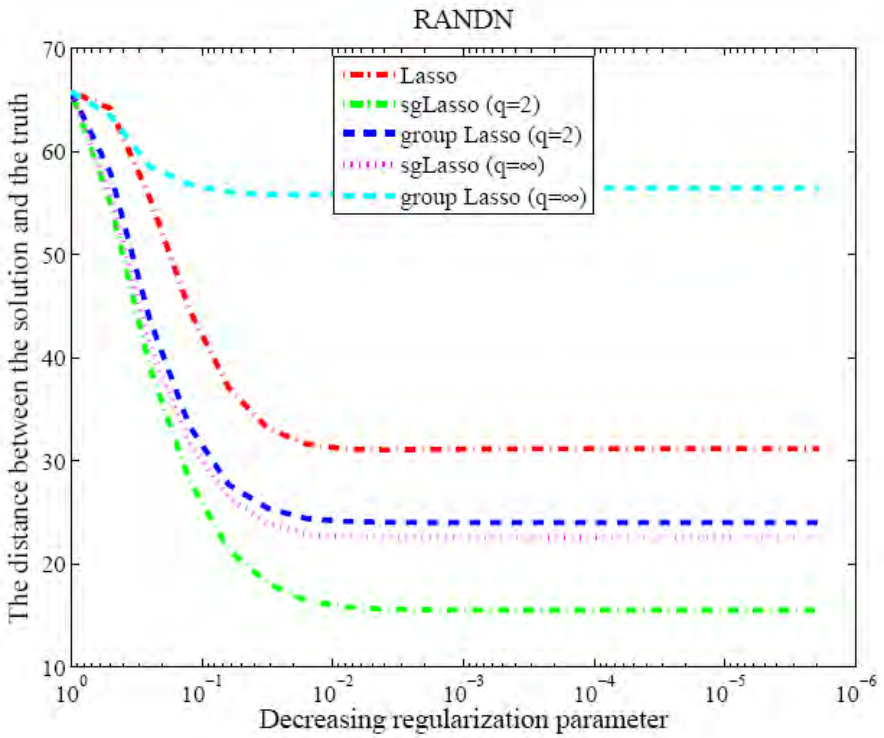


# Simulation Study

## (Liu and Ye, 2010)



# Integrative Genomics Study of Breast Cancer

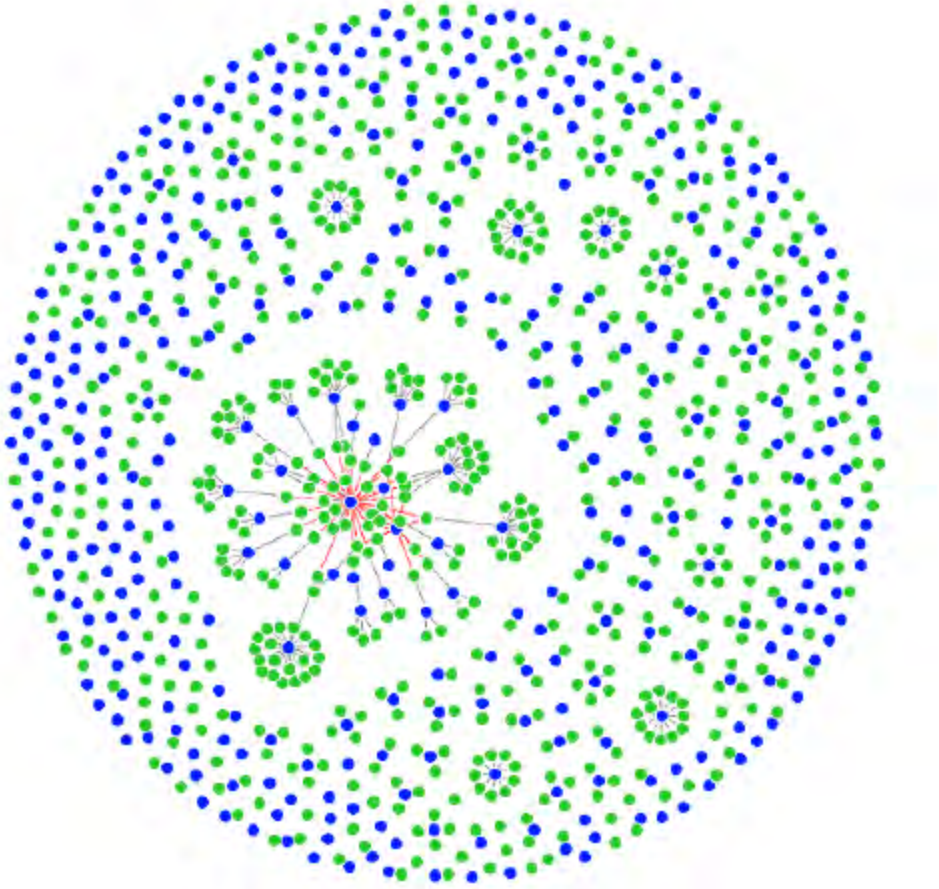
## (Peng et al., 2010)

- Response variables: 654 gene expressions.
- Predictor variables:
  (1) Expressions of genes connected to the current response in ***Exp.Net***. (2) copy number of the 384 CNAIs.
- Penalties are imposed on the coefficients for unlinked CNAIs.
- Identified 43 trans edges correspond to 3 CNAIs from **17q12-q21.2**
  - This region was highly amplified in 19% (33) of the samples.
  - Amplification of this region influences the expression levels of 31 unlinked genes

  ➔ **17q12-q21.2 may harbor transcriptional factors** whose activities closely relate to breast cancer.
- Indeed, this region has 4 known transcription factors (NEUROD2, IKZF3, THRA and NR1D1) and 2 transcription co-activators (MED1,MED24)

# Integrative Genomics Study of Breast Cancer

## (Peng et al., 2010)

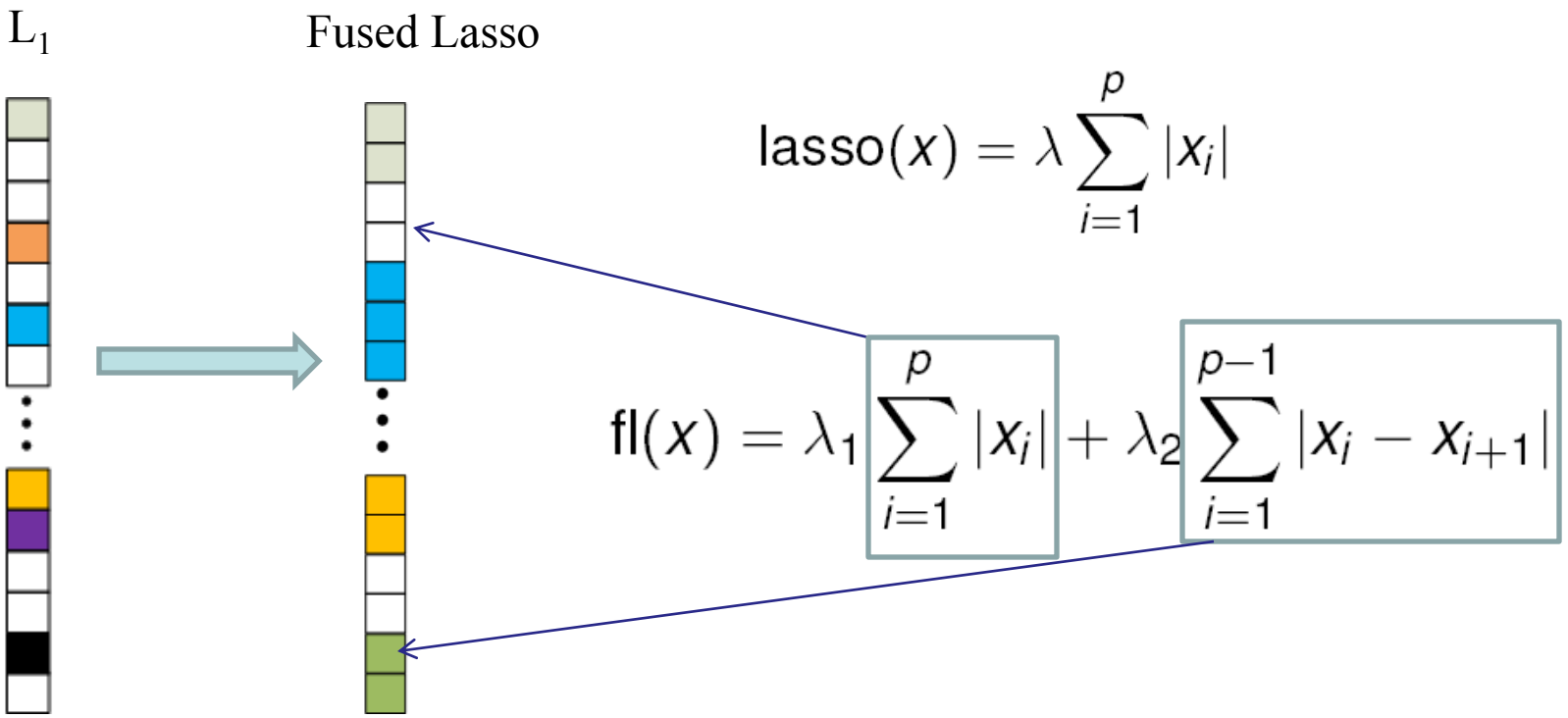


# Outline

- Sparsity via $L_1$
- Sparsity via $L_1/L_q$
- **Sparsity via Fused Lasso**
- Sparse Inverse Covariance Estimation
- Sparsity via Trace Norm
- Implementations and the SLEP package
- Trends in Sparse Learning

# Fused Lasso

## (Tibshirani et al., 2005; Tibshirani and Wang, 2008; Friedman et al., 2007)

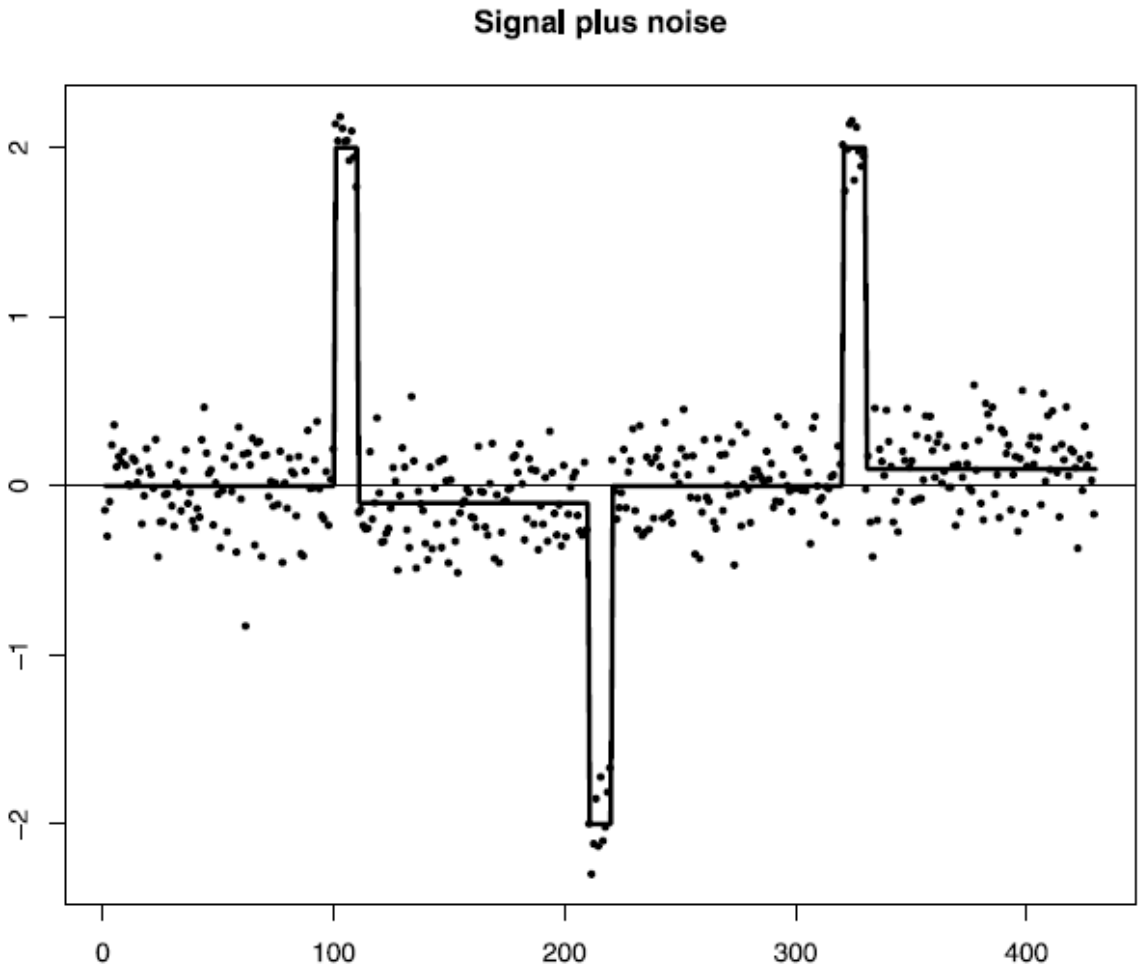


# Illustration of Fused Lasso

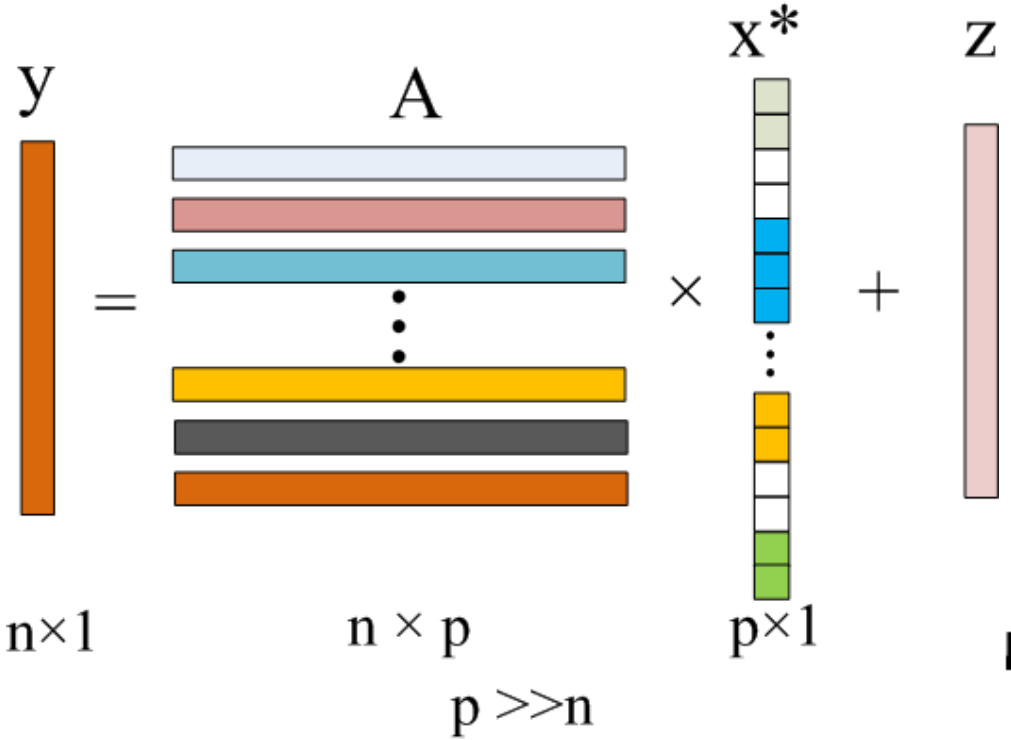
## (Rinaldo, 2009)

$$y = x + z$$

Noise



# Fused Lasso




$$\min_x \frac{1}{2}\|Ax - y\|_2^2 + \mathrm{fl}(x)$$

$$\mathrm{fl}(x) = \lambda_1 \sum_{i=1}^{p} |x_i| + \lambda_2 \sum_{i=1}^{p-1} |x_i - x_{i+1}|$$

# Asymptotic Property

## (Tibshirani et al., 2005)

$$y_i = \sum_j x_{ij}\beta_j + \varepsilon_i \qquad \sum_{i=1}^{N}(y_i - \mathbf{x}_i^{\mathrm{T}}\beta)^2 + \lambda_N^{(1)}\sum_{j=1}^{p}|\beta_j| + \lambda_N^{(2)}\sum_{j=2}^{p}|\beta_j - \beta_{j-1}|$$

If $\lambda_N^{(l)}/\sqrt{N} \to \lambda_0^{(l)} \geqslant 0$ $(l = 1, 2)$ and

$$C = \lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)$$

is non-singular then

$$\sqrt{N}(\hat{\beta}_N - \beta) \xrightarrow[\mathrm{d}]{} \arg\min(V),$$

where

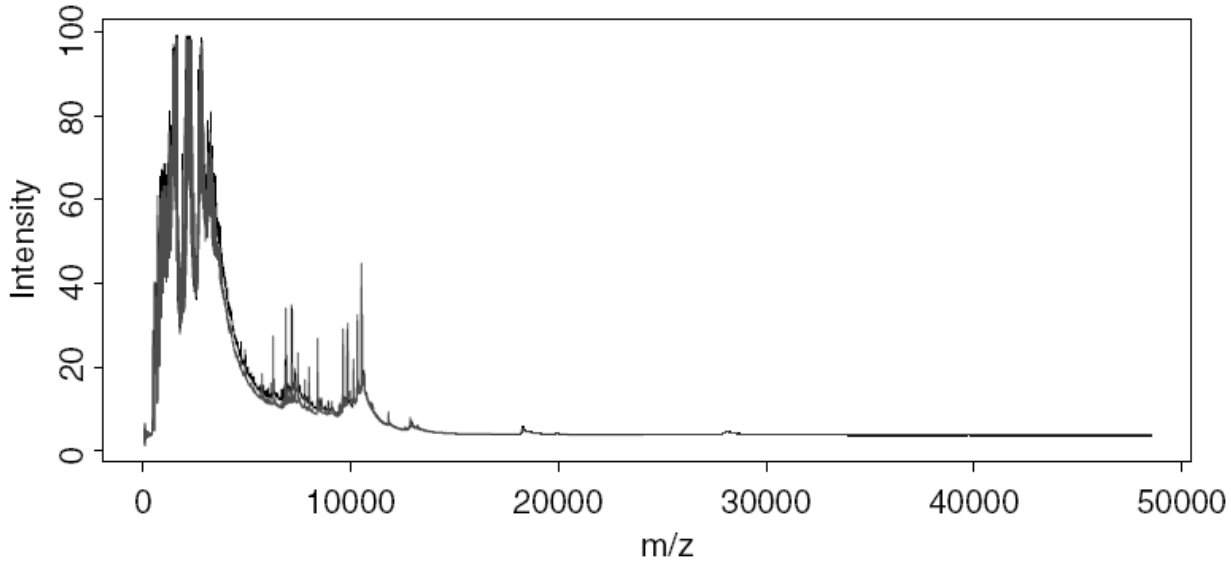
$$V(\mathbf{u}) = -2\mathbf{u}^{\mathrm{T}}\mathbf{W} + \mathbf{u}^{\mathrm{T}}C\mathbf{u} + \lambda_0^{(1)}\sum_{j=1}^{p}\{u_j\,\mathrm{sgn}(\beta_j)\,I(\beta_j \neq 0) + |u_j|\,I(\beta_j = 0)\}$$

$$+\lambda_0^{(2)}\sum_{j=2}^{p}\{(u_j - u_{j-1})\,\mathrm{sgn}(\beta_j - \beta_{j-1})\,I(\beta_j \neq \beta_{j-1}) + |u_j - u_{j-1}|\,I(\beta_j = \beta_{j-1})\}$$

and $\mathbf{W}$ has an $\mathcal{N}(\mathbf{0}, \sigma^2 C)$ distribution.

# Prostate Cancer

## (Tibshirani et al., 2005)

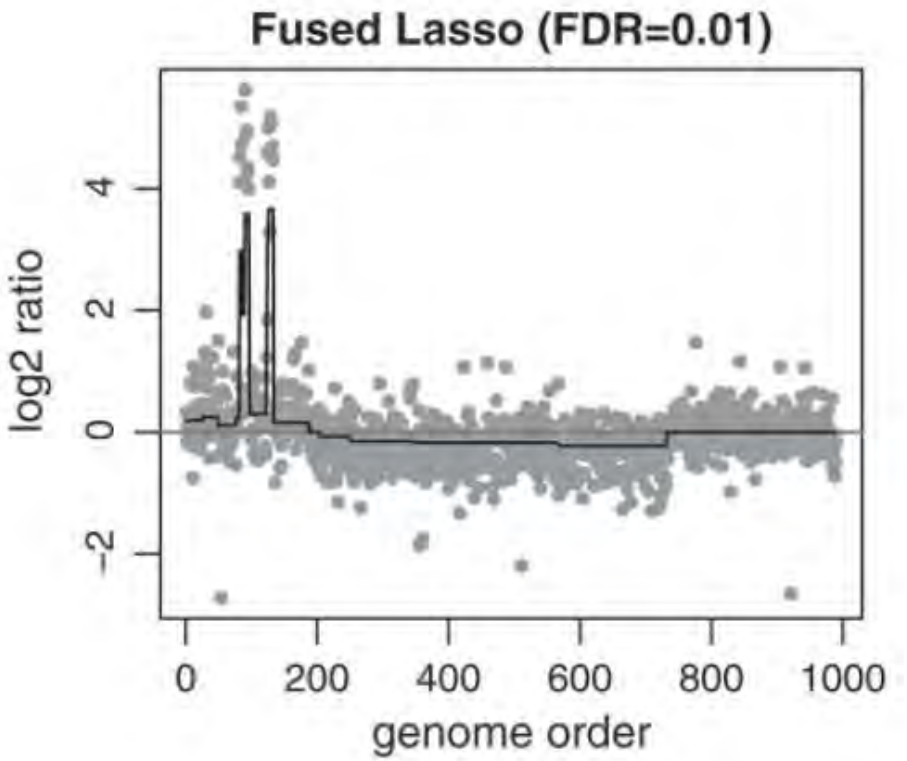


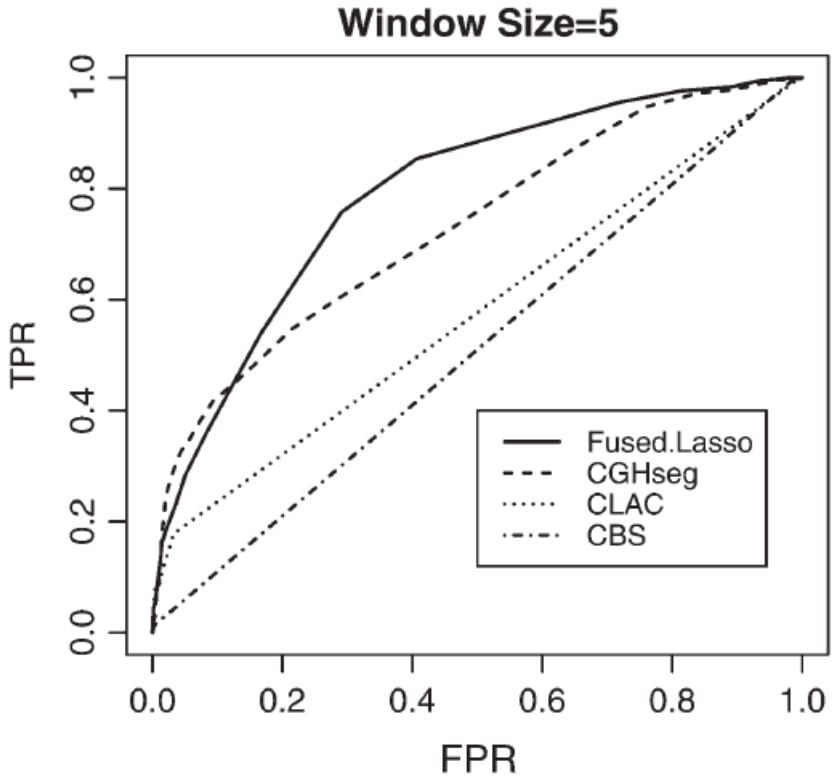

| *Method* | *Validation errors/108* | *Degrees of freedom* | *Number of sites* | $s_1$ | $s_2$ |
|---|---|---|---|---|---|
| Nearest shrunken centroids | 30 | | 227 | | |
| Lasso | 16 | 60 | 40 | 83 | 164 |
| Fusion | 18 | 102 | 2171 | 16 | 32 |
| Fused lasso | 16 | 103 | 218 | 113 | 103 |

# Leukaemia Classification Using Microarrays
## (Tibshirani et al., 2005)

| *Method* | *10-fold cross-validation error* | *Test error* | *Number of genes* |
|---|---|---|---|
| (1) Golub *et al.* (1999) (50 genes) | 3/38 | 4/34 | 50 |
| (2) Nearest shrunken centroid (21 genes) | 1/38 | 2/34 | 21 |
| (3) Lasso, 37 degrees of freedom ($s_1 = 0.65, s_2 = 1.32$) | 1/38 | 1/34 | 37 |
| (4) Fused lasso, 38 degrees of freedom ($s_1 = 1.08, s_2 = 0.71$) | 1/38 | 2/34 | 135 |
| (5) Fused lasso, 20 degrees of freedom ($s_1 = 1.35, s_2 = 1.01$) | 1/38 | 4/34 | 737 |
| (6) Fusion, 1 degree of freedom | 1/38 | 12/34 | 975 |

# Arracy CGH


Fused Lasso (FDR=0.01)
log2 ratio
-2
0
2
4
0
200
400
600
800
1000
genome order

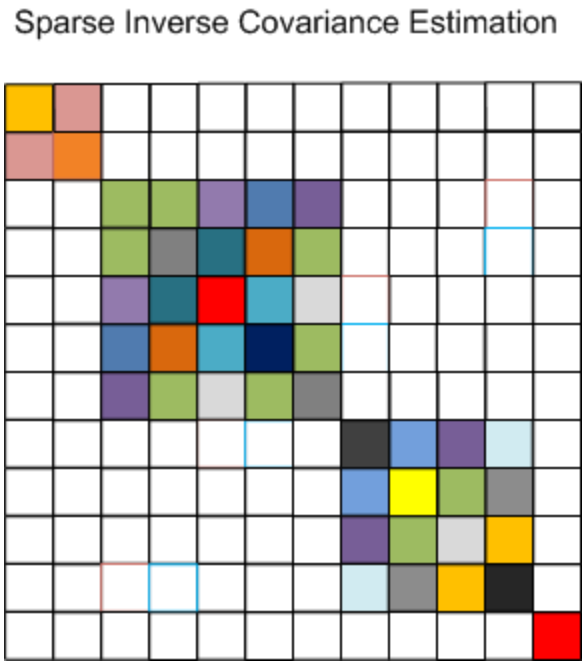

Window Size=5
TPR
0.0
0.2
0.4
0.6
0.8
1.0
FPR
Fused.Lasso
CGHseg
CLAC
CBS

# Outline

- Sparsity via $L_1$
- Sparsity via $L_1/L_q$
- Sparsity via Fused Lasso
- **Sparse Inverse Covariance Estimation**
- Sparsity via Trace Norm
- Implementations and the SLEP package
- Trends in Sparse Learning

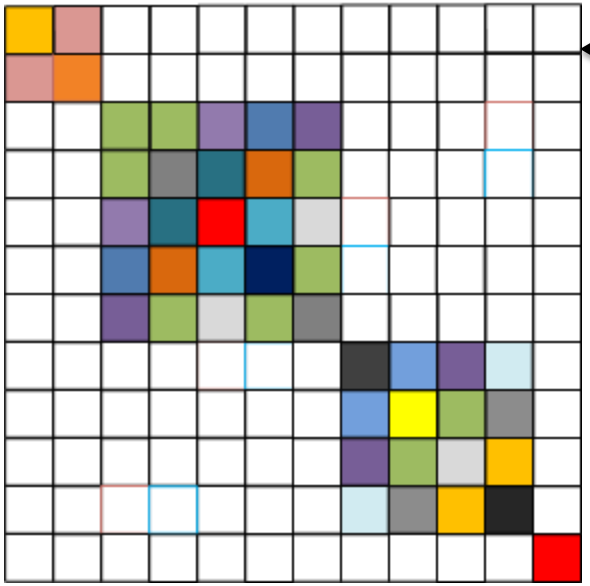
# Sparse Inverse Covariance Estimation

*The pattern of zero entries in the inverse covariance matrix of a multivariate normal distribution corresponds to conditional independence restrictions between variables (Meinshausen & Buhlmann, 2006).*



# The SICE Model

Sparse Inverse Covariance Estimation



Sparsity

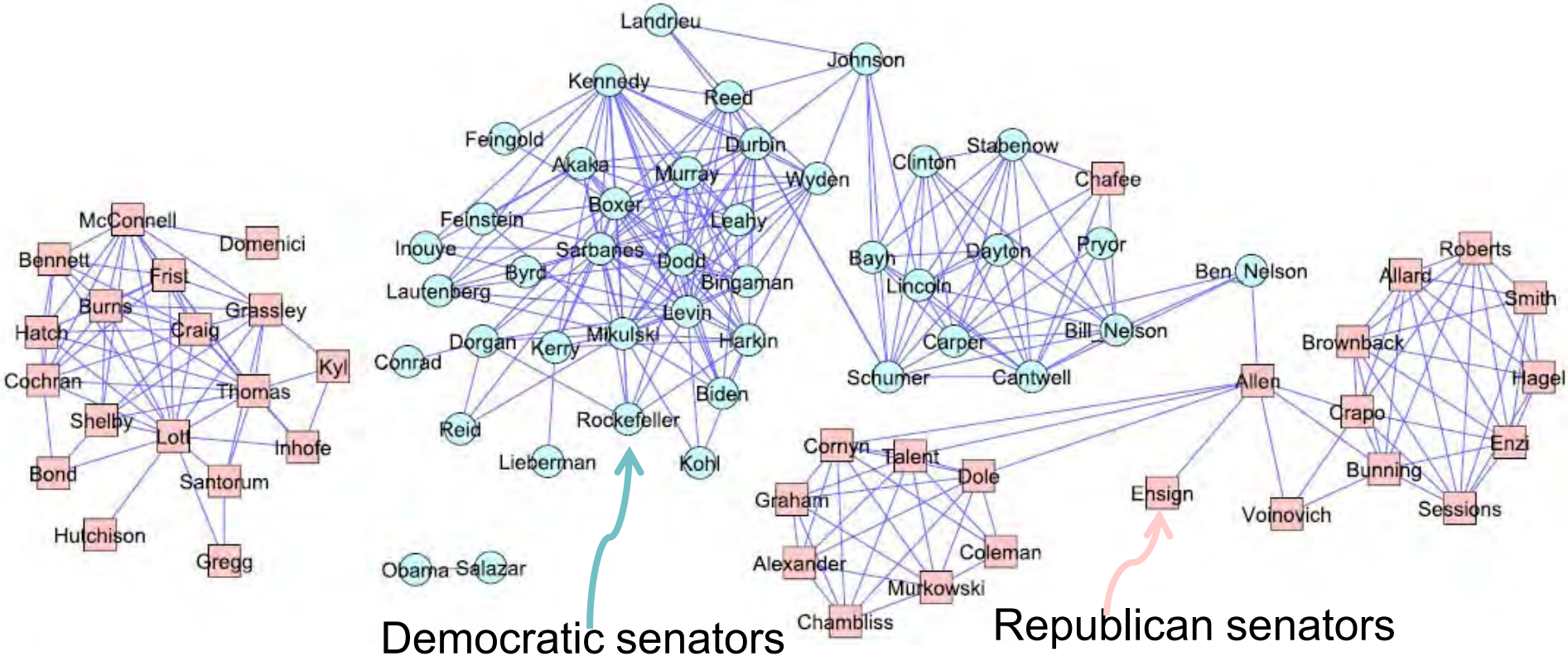
$$\arg\max_{X \succ 0} \log\det X - \mathrm{trace}(SX) - \lambda\|X\|_1$$

Maximum Likelihood Estimation

When S is invertible, directly maximizing the likelihood gives

$X=S^{-1}$

# Example: Senate Voting Records Data (2004-06)




Chafee (R, RI) has only Democrats as his neighbors, an observation that supports media statements made by and about Chafee during those years.

# The Monotone Property

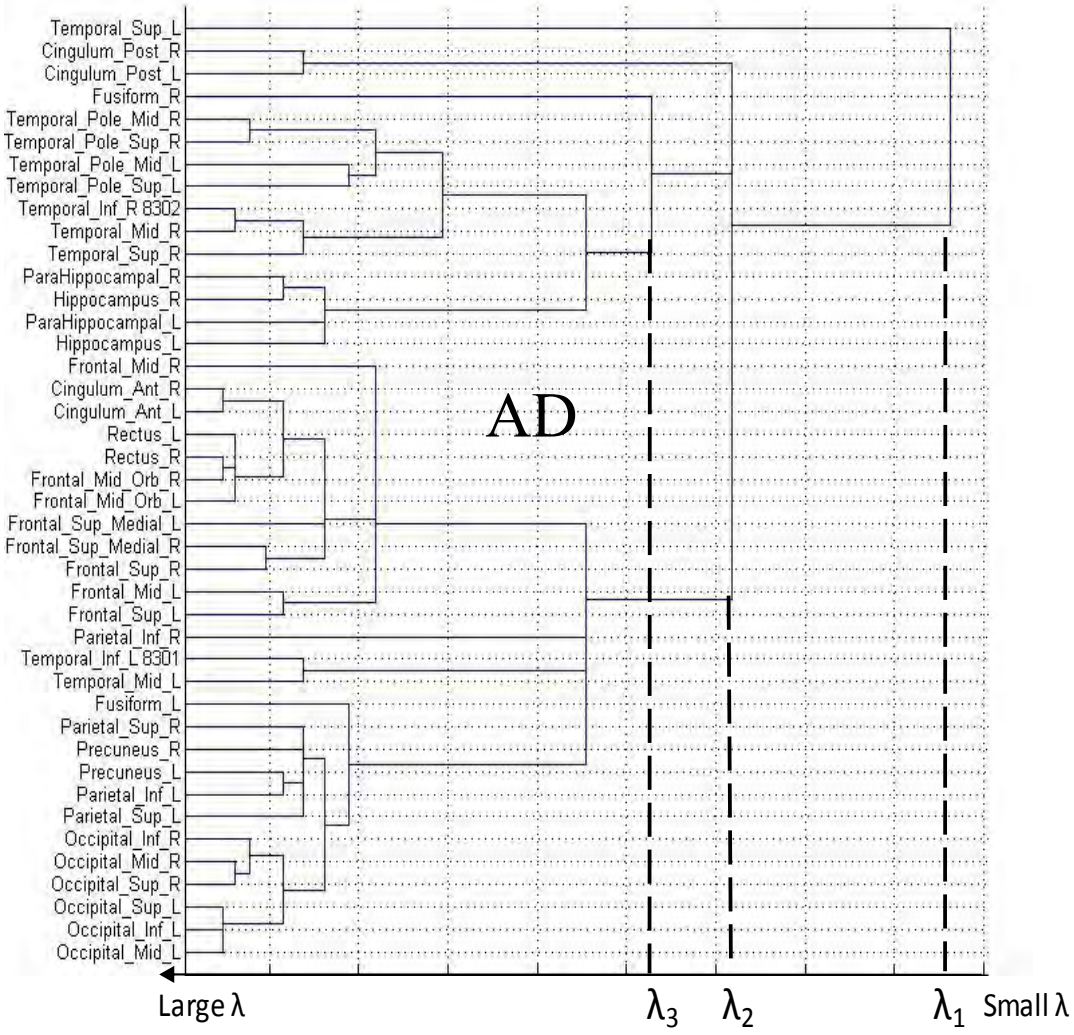
## (Huang et al., NIPS 2009)

**Monotone Property**

Let $C_k(\lambda_1)$ and $C_k(\lambda_2)$ be the sets of all the connectivity components of $X_k$ with $\lambda = \lambda_1$ and $\lambda = \lambda_2$ respectively.
If $\lambda_1 < \lambda_2$, then $C_k(\lambda_1) \supseteq C_k(\lambda_2)$.

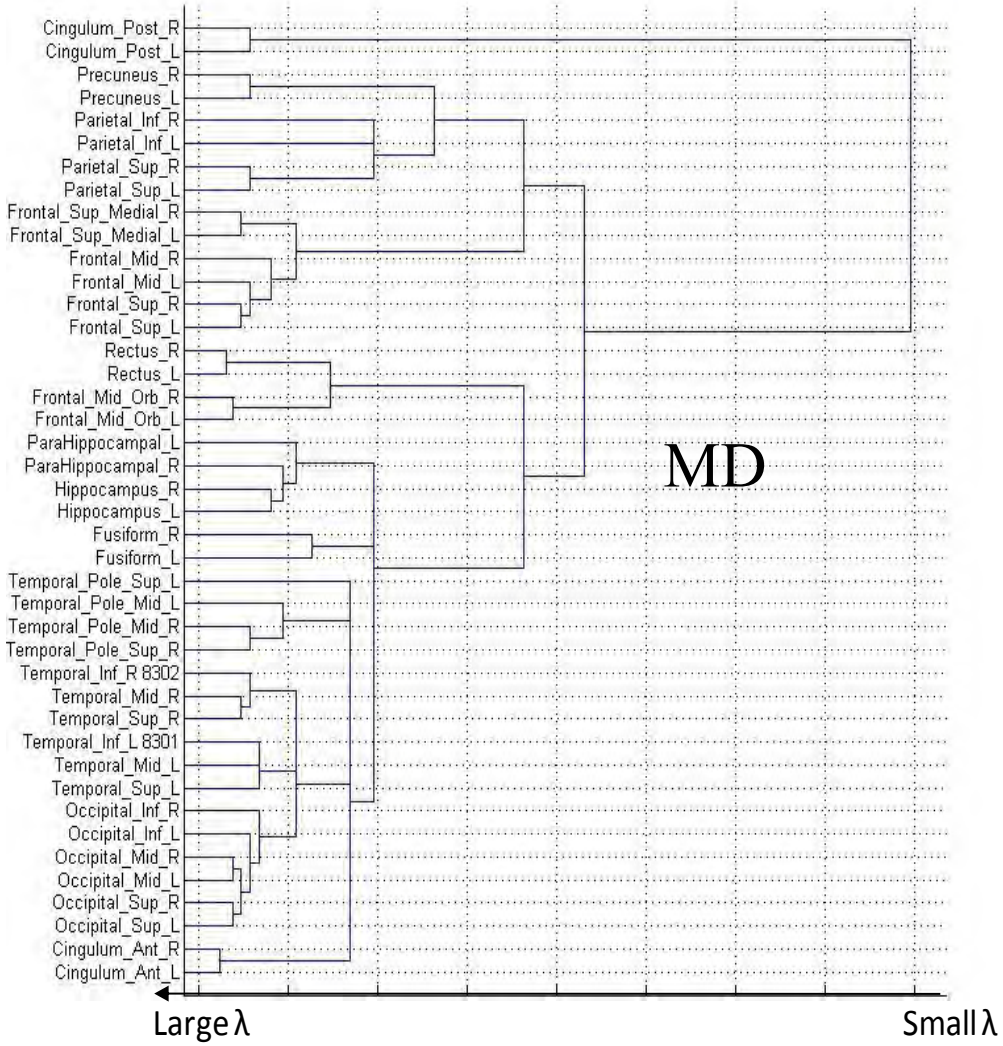
*Intuitively, if two nodes are connected (either directly or indirectly) at one level of sparseness, they will be connected at all lower levels of sparseness.*

# Brain Network for Alzheimer's Disease



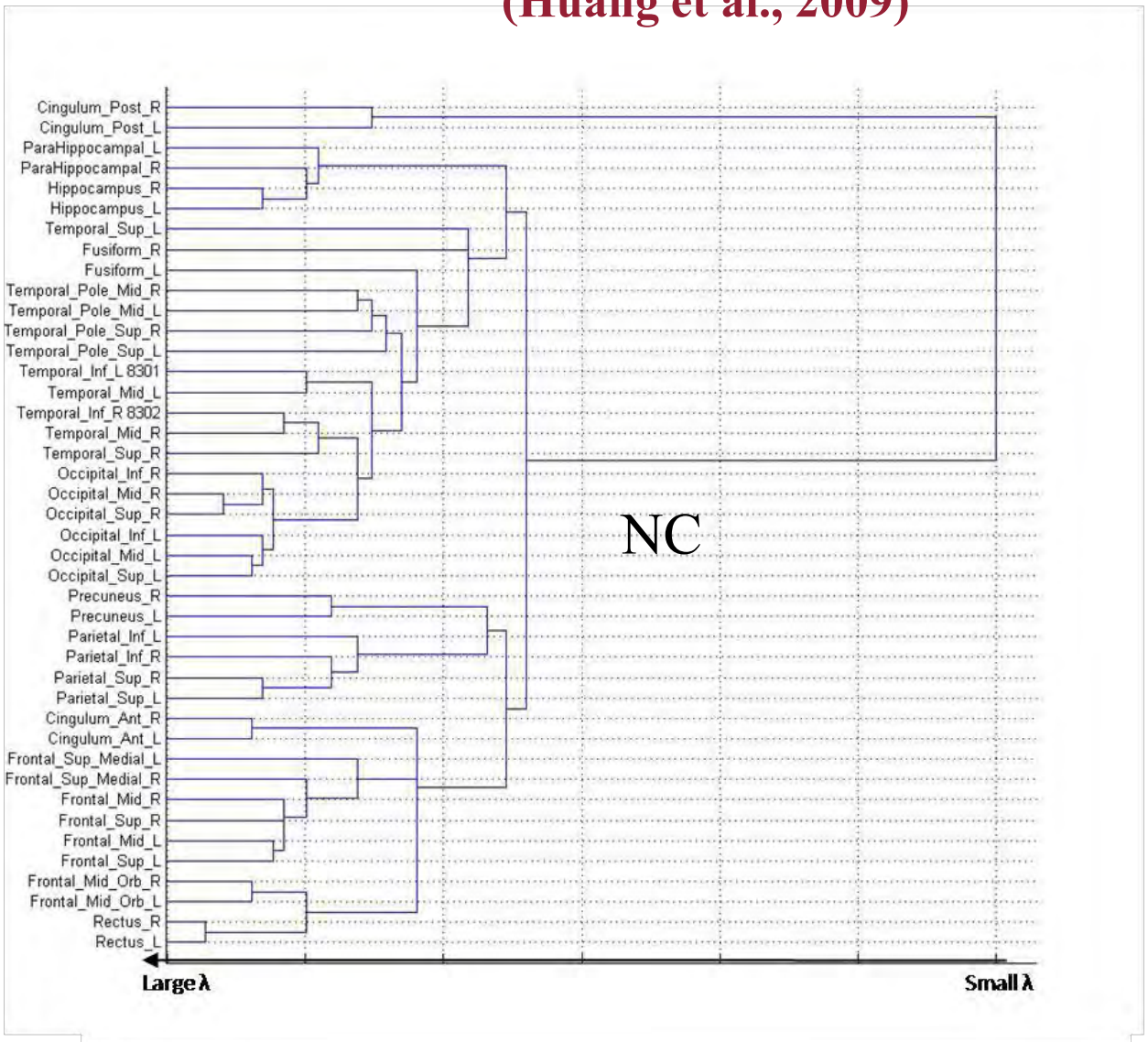
# Brain Network for Alzheimer's Disease

## (Huang et al., 2009)



# Brain Network for Alzheimer's Disease

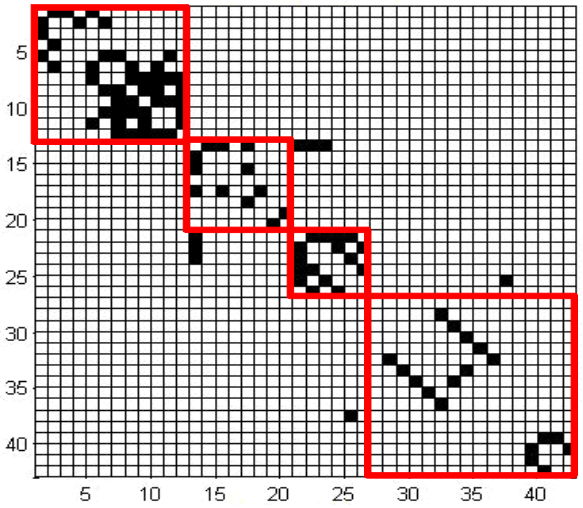
## (Huang et al., 2009)



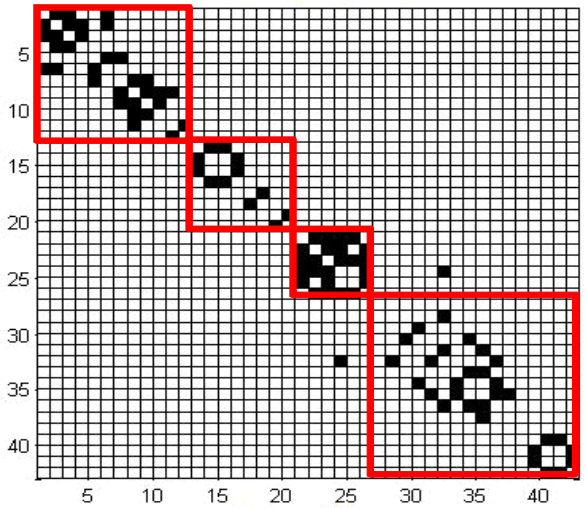
# Brain Network for Alzheimer's Disease

## (Huang et al., 2009)

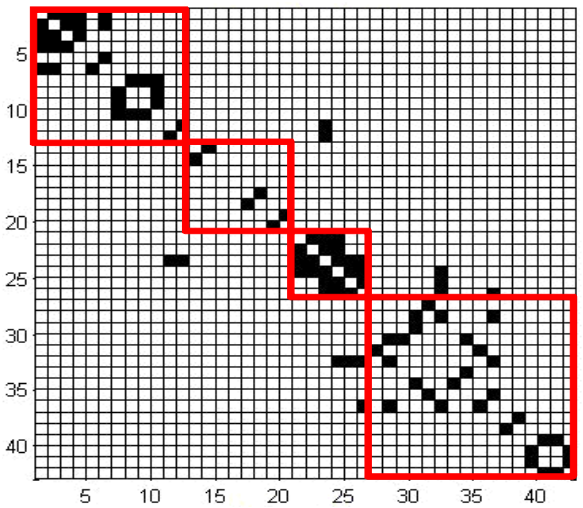
AD




MD




NC

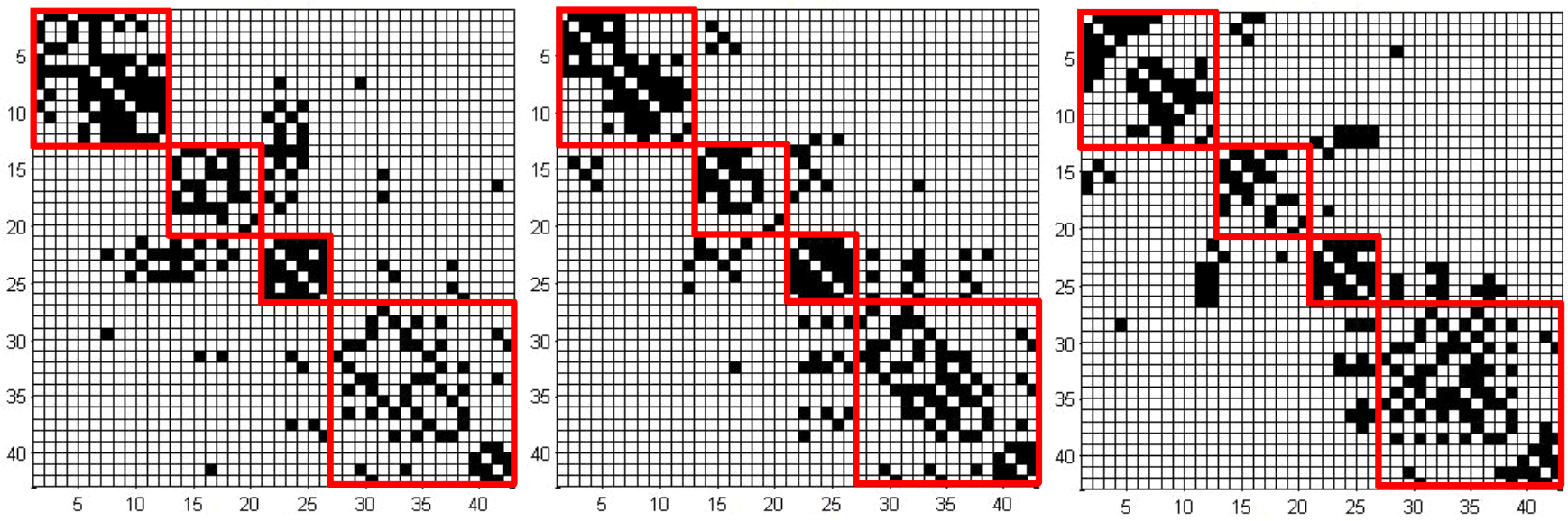



Strong Connectivity

# Brain Network for Alzheimer's Disease

## (Huang et al., 2009)

AD MD NC

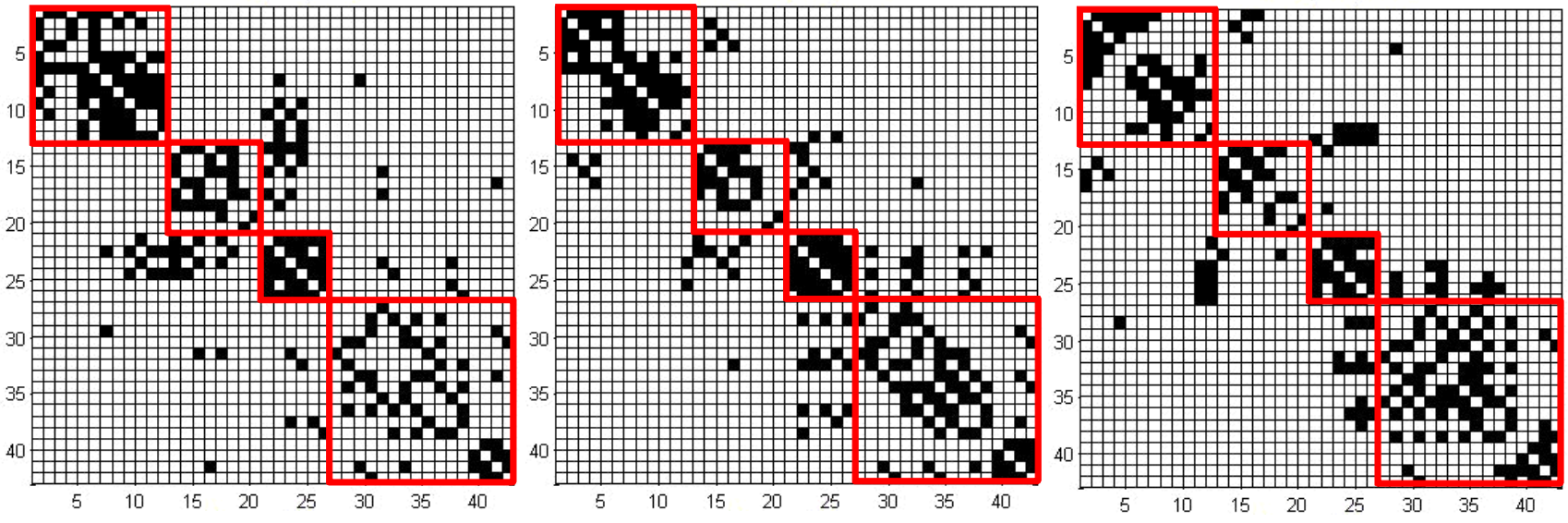



Mild Connectivity

# Brain Network for Alzheimer's Disease

## (Huang et al., 2009)

AD MD NC




Weak Connectivity

# Brain Network for Alzheimer's Disease

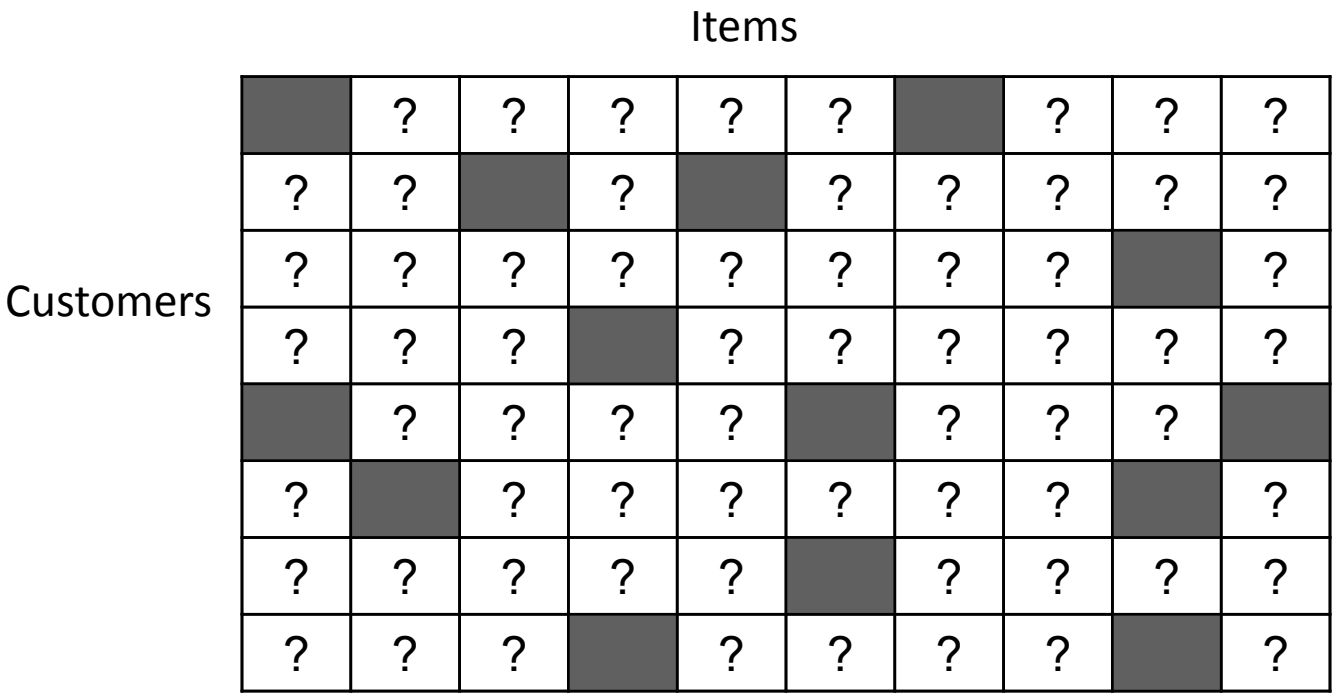
**(Huang et al., 2009)**

- Temporal: decreased connectivity in AD, decrease not significant in MCI.
- Frontal: increased connectivity in AD (compensation), increase not significant in MCI.
- Parietal, occipital: no significant difference.
- Parietal-occipital: increased weak/mild con. in AD.
- Frontal-occipital: decreased weak/mild con. in MCI.
- Left-right: decreased strong con. in AD, not MCI.

# Outline

- Sparsity via $L_1$
- Sparsity via $L_1/L_q$
- Sparsity via Fused Lasso
- Sparse Inverse Covariance Estimation
- **Sparsity via Trace Norm**
- Implementations and the SLEP package
- Trends in Sparse Learning

# Collaborative Filtering




- Customers are asked to rank items
- Not all customers ranked all items
- Predict the missing rankings

# The Netflix Problem

Movies

Users

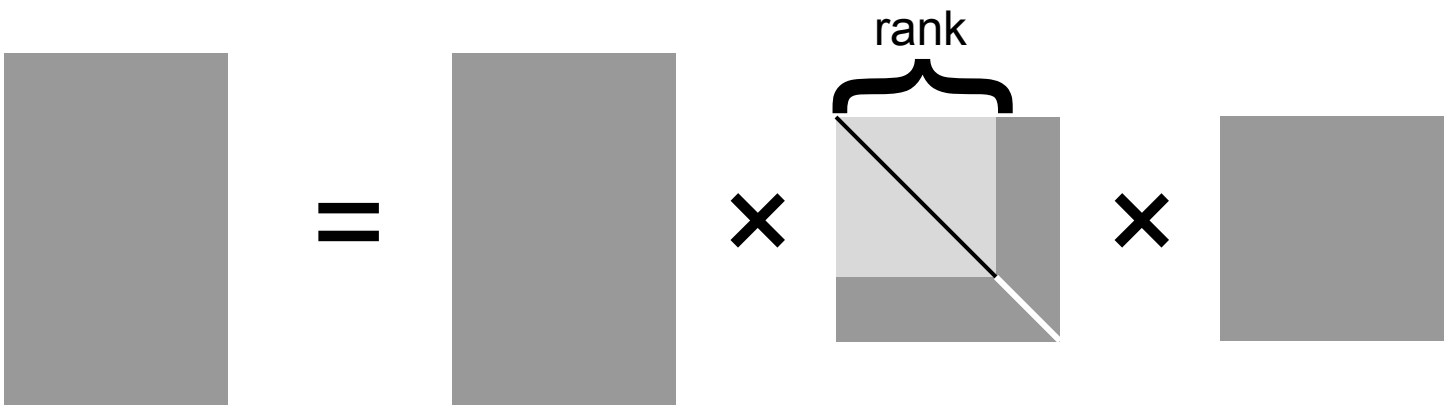
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ? | ? | ? | ? | ? | | ? | ? | ? |
| ? | ? | | ? | | ? | ? | ? | ? | ? |
| ? | ? | ? | ? | ? | ? | ? | ? | | ? |
| ? | ? | ? | | ? | ? | ? | ? | ? | ? |
| | ? | ? | ? | ? | | ? | ? | ? | |
| ? | | ? | ? | ? | ? | ? | ? | | ? |
| ? | ? | ? | ? | ? | | ? | ? | ? | ? |
| ? | ? | ? | | ? | ? | ? | ? | | ? |

- About a million users and 25,000 movies
- Known ratings are sparsely distributed
- Predict unknown ratings

**Preferences of users are determined by a small number of factors → low rank**

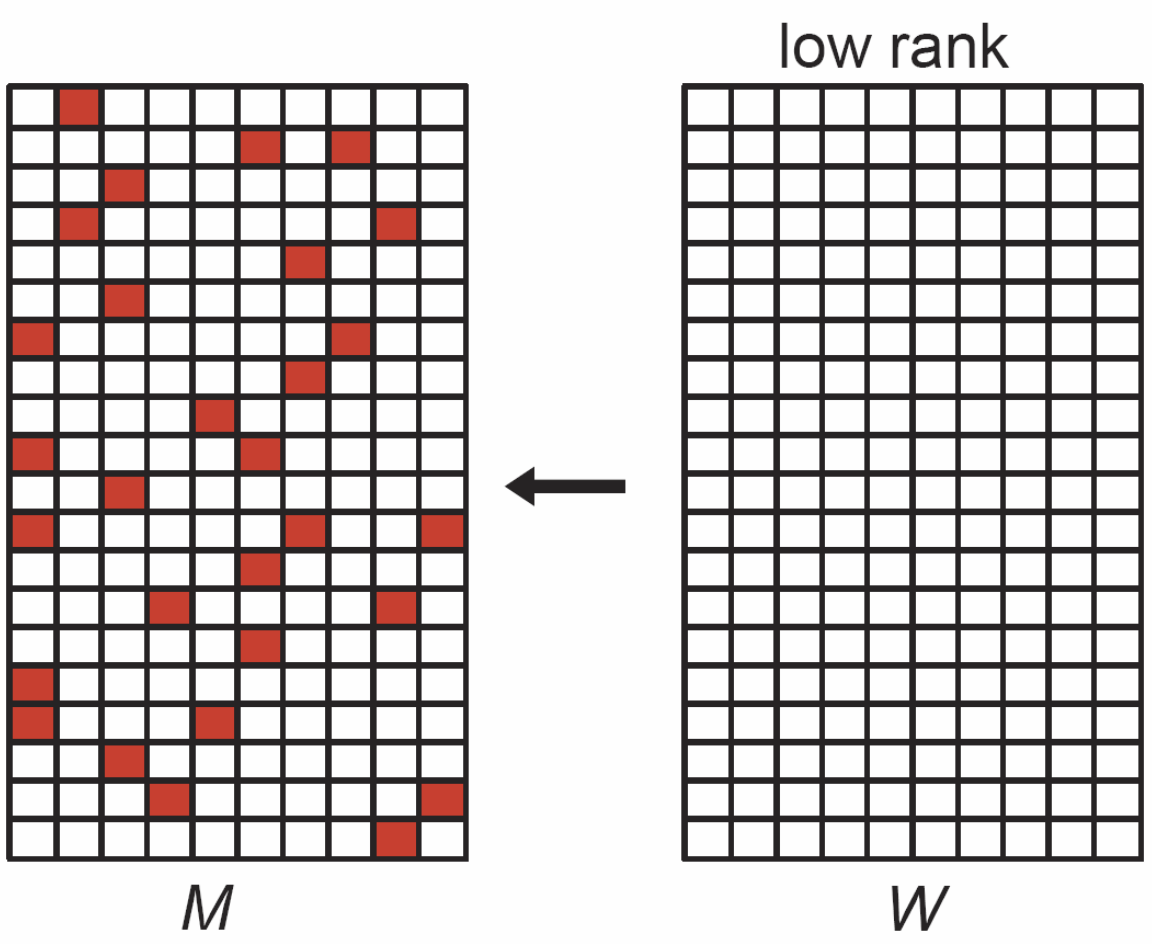
# Matrix Rank

- The number of independent rows or columns
- The singular value decomposition (SVD):

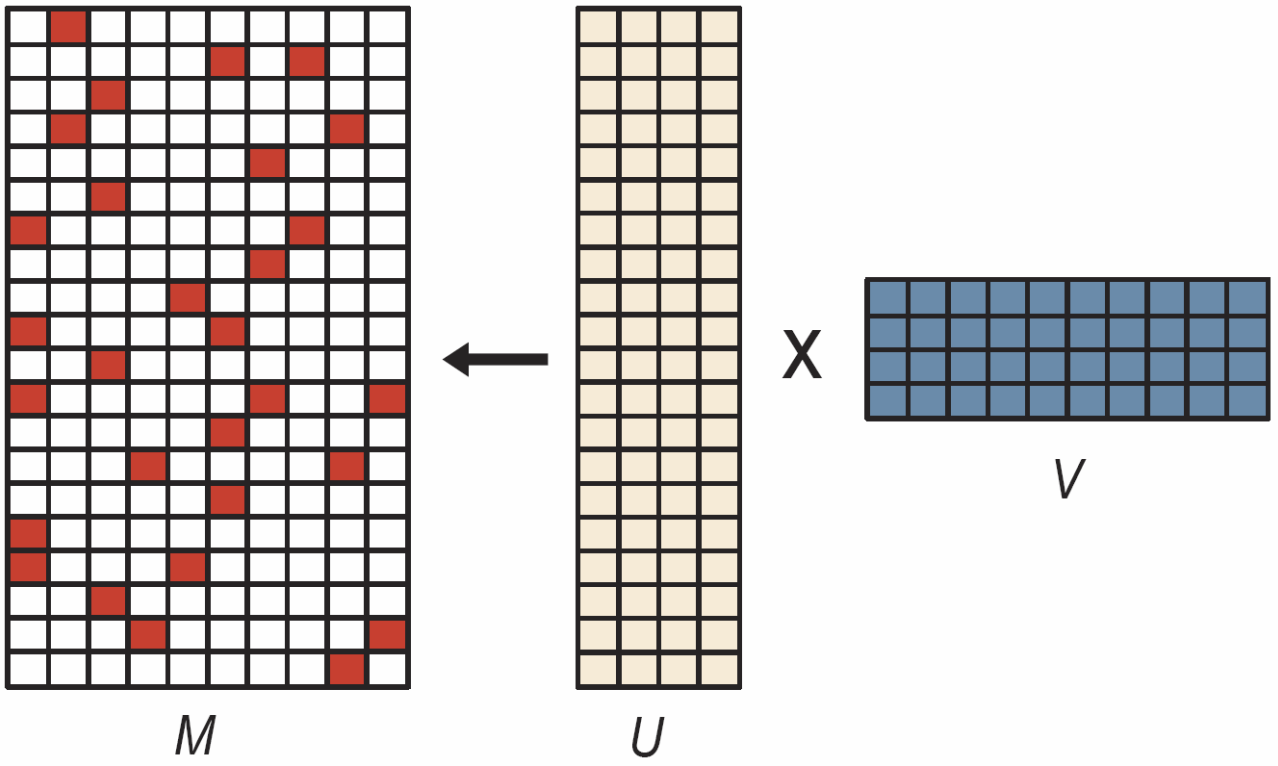


# The matrix Completion Problem

$$\min_{W} \sum_{i,j \in \text{observed}} \ell(M_{ij}, W_{ij}) + \lambda * \text{rank}(W)$$



# The Alternating Approach




- Optimize over *U* and *V* iteratively
- Solution is locally optimal

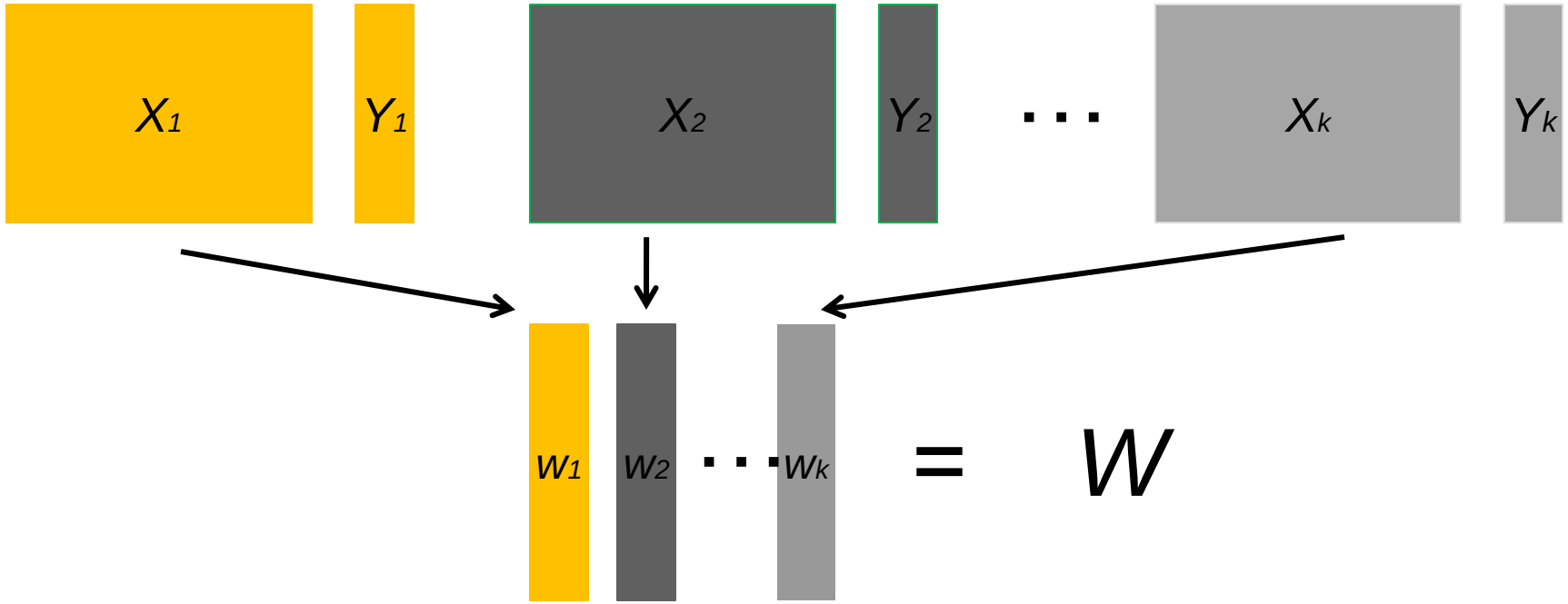
# Fundamental Questions

- Can we recover a matrix M of size $n_1$ by $n_2$ from m sampled entries, m << $n_1$ $n_2$?
- In general, it is impossible.

- Surprises (Candes & Recht'08):
  - Can recover matrices of interest from incomplete sampled entries
  - Can be done by convex programming

$$\min \|\boldsymbol{X}\|_* \quad \text{s. t.} \quad X_{ij} = M_{ij},\ (i,j) \in \Omega$$

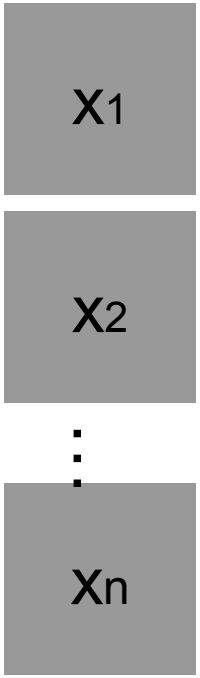
# Multi-Task/Class Learning

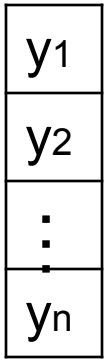


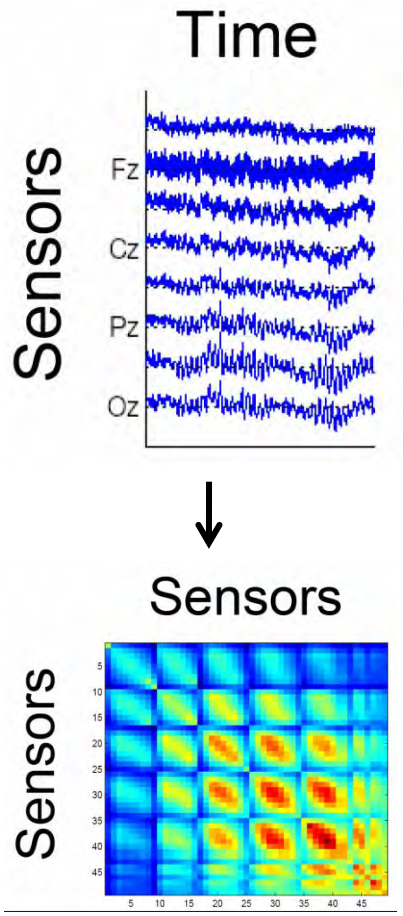

- The multiple tasks/classes are usually related
- The matrix *W* is of low rank

$$\sum_{i=1}^{T}\sum_{j=1}^{n_i} l(y_i^j, w_i^T x_i^j) + \lambda * rank(W)$$

# Matrix Classification




$$\min_{W} \sum_{i=1}^{n} l(y_i, Tr(W^T X_i)) + rank(W)$$

[Tomioka & Aihara (2007), ICML]

# Other Low-Rank Problems

- Image compression
- System identification in control theory
- Structure-from-motion problem in computer vision
- Other settings:
  - low-degree statistical model for a random process
  - a low-order realization of a linear system
  - A low-order controller for a plant
  - a low-dimensional embedding of data in Euclidean space

# Two Formulations for Rank Minimization

$$\min \quad f(W) + \lambda * \text{rank}(W)$$

$$\begin{aligned} \min \quad & \text{rank}(W) \\ \text{subject to} \quad & f(W) = b \end{aligned}$$

Rank minimization is NP-hard

# Trace Norm (Nuclear Norm)

Trace norm of a matrix is the sum of its singular values:

$$W = U \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_k \end{pmatrix} V^T$$

$$||W||_* = \sum_{i=1}^{k} \sigma_i$$

- trace norm $\Leftrightarrow$ 1-norm of the vector of singular values
- Trace norm is the convex envelope of the rank function over the unit ball of spectral norm $\Rightarrow$ a convex relaxation
- In some sense, this is the tightest convex relaxation of the NP-hard rank minimization problem

# Two Formulations for Nuclear Norm

$$\min \quad f(W) + \lambda^* \|W\|_*$$

$$\begin{aligned} \min \quad & \|W\|_* \\ \text{subject to} \quad & f(W)=b \end{aligned}$$

Nuclear norm minimization is convex

- Consistent estimation can be obtained
- Can be solved by
  - Semi-definite programming
  - Gradient-based methods

# Sparsity with Vectors and Matrices

| **Parsimony concept** | Cardinality | Rank |
|---|---|---|
| **Hilbert space norm** | Euclidean | Frobenius |
| **Sparsity inducing norm** | $\ell_1$ | Trace norm |
| **Convex optimization** | Linear programming | Semi-definite programming |

[Recht *et al*. (2010) SIAM Review]

# Rank Minimization and CS

Rank minimization

$$\min \quad \text{rank}(W)$$
$$\text{s.t.} \quad f(W) = b$$

Convex relaxation

$$\min \quad \|W\|_*$$
$$\text{s.t.} \quad f(W) = b$$

$W$ is diagonal, linear constraint

Rank minimization

$$\min \quad \|w\|_0$$
$$\text{s.t.} \quad Aw = b$$

Convex relaxation

$$\min \quad \|w\|_1$$
$$\text{s.t.} \quad Aw = b$$

# Theory of Matrix Completion

$$\min \|\boldsymbol{X}\|_* \quad \text{s. t.} \quad X_{ij} = M_{ij},\ (i,j) \in \Omega$$

- $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ *of rank* $r$ *(obeying* $\mu_0 r \lesssim n^{1/5}$*)*

$$\mu_0 := \max\left(\text{coh}(\text{col. space}), \text{coh}(\text{row. space})\right)$$

- *random set of entries of size* $m$

*The minimizer is unique and equal to* $\boldsymbol{M}$ *w.p. at least* $1 - n^{-3}$ *if*

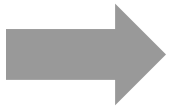
$$m \gtrsim \mu_0\, n^{6/5}\, r \log n, \qquad n = \max(n_1, n_2)$$

Candès and Recht (2008)

# Semi-definite programming (SDP)

$$\min_{W} \|W\|_*$$

$$\text{s.t.} \quad f(W) = b$$



$$\min_{X, A_1, A_2} \frac{1}{2}(Tr(A_1) + Tr(A_2))$$

$$\text{s.t.} \quad \begin{bmatrix} A_1 & X \\ X^T & A_2 \end{bmatrix} \geq 0$$

$$f(W) = b$$

- SDP is convex, but computationally expensive
- Many recent efficient solvers:
  - Singular value thresholding (Cai et al, 2008 )
  - Fixed point method (Ma et al, 2009)
  - Accelerated gradient descent (Toh & Yun, 2009, Ji & Ye, 2009)

# Outline

- Sparsity via $L_1$
- Sparsity via $L_1/L_q$
- Sparsity via Fused Lasso
- Sparsity via Trace Norm
- Sparse Inverse Covariance Estimation
- **Implementations and the SLEP package**
- Trends in Sparse Learning

# Optimization Algorithm

$$\min \; f(x) = \text{loss}(x) + \lambda \times \text{penalty}(x)$$

$$\begin{aligned} \min \quad & \text{loss}(x) \\ \text{s.t.} \quad & \text{penalty}(x) \le z \end{aligned}$$

loss(x) is convex and smooth,
penalty(x) is convex but nonsmooth

- Smooth Reformulation – general solver
- Coordinate descent
- Subgradient descent
- Gradient descent
- Accelerated gradient descent
- Online algorithms
- Stochastic algorithms
- …

# Smooth Reformulations: $L_1$

$$\min_x \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_1$$

$$\begin{aligned} \min_{x,t} \quad & \frac{1}{2}\|Ax - y\|_2^2 + \lambda \sum_{i=1}^{p} t_i \\ \text{s.t.} \quad & -t_i \leq x_i \leq t_i, \forall i \end{aligned}$$

Linearly constrained quadratic programming

# Smooth Reformulation: $L_1/L_2$

$$\min_x \frac{1}{2}\|Ax - y\|_2^2 + \lambda \sum_{i=1}^{g} \|x_{G_i}\|_2$$

$$\begin{aligned} \min_{x,t} \quad & \frac{1}{2}\|Ax - y\|_2^2 + \lambda \sum_{i=1}^{g} t_i \\ \text{s.t.} \quad & \|x_{G_i}\|_2 \le t_i, \forall i \end{aligned}$$

Second order cone programming

# Smooth Reformulation: Fused Lasso

$$\min_x \frac{1}{2}\|Ax - y\|_2^2 + \lambda_1 \|x\|_1 + \lambda_2 \sum_{i=1}^{p-1} |x_i - x_{i-1}|$$

$$\min_{x,t,s} \quad \frac{1}{2}\|Ax - y\|_2^2 + \lambda_1 \sum_{i=1}^{p} t_i + \lambda_2 \sum_{i=1}^{p-1} s_i$$

$$\text{s.t.} \quad -t_i \le x_i \le t_i, -s_i \le x_i - x_{i-1} \le s_i, \forall i$$

Linearly constrained quadratic programming

# Summary of Reformulations

$$\min_{x,t} \ \frac{1}{2}\|Ax-y\|_2^2+\lambda\sum_{i=1}^{p} t_i \quad \text{s.t.} \ -t_i \le x_i \le t_i, \forall i$$

$$\min_{x,t} \ \frac{1}{2}\|Ax-y\|_2^2+\lambda\sum_{i=1}^{g} t_i \quad \text{s.t.} \ \|x_{G_i}\|_2 \le t_i, \forall i$$

$$\min_{x,t,s} \ \frac{1}{2}\|Ax-y\|_2^2+\lambda_1\sum_{i=1}^{p} t_i+\lambda_2\sum_{i=1}^{p-1} s_i \quad \text{s.t.} \ -t_i \le x_i \le t_i, -s_i \le x_i - x_{i-1} \le s_i, \forall i$$

## Advantages:

- Easy to incorporate existing solvers
- Fast and high precision for small size problems

## Disadvantages:

- Does not scale well for large size problems, due to 1) many additional variables and constraints are introduced; 2) the computation of Hessian is demanding
- Does not utilize well the "structure" of the nonsmooth penalty
- Not applicable to all the penalties discussed in this tutorial, say, $L_1/L_3$.

# Coordinate Descent

## (Tseng, 2002)

$$\min_{x=(x_1,\ldots,x_n)} f(x)$$

Given $x \in \Re^n$, choose $i \in \{1, \ldots, n\}$. Update

$$x^{\text{new}} = \arg\min_{u | u_j = x_j \ \forall j \neq i} f(u).$$

Repeat until "convergence".

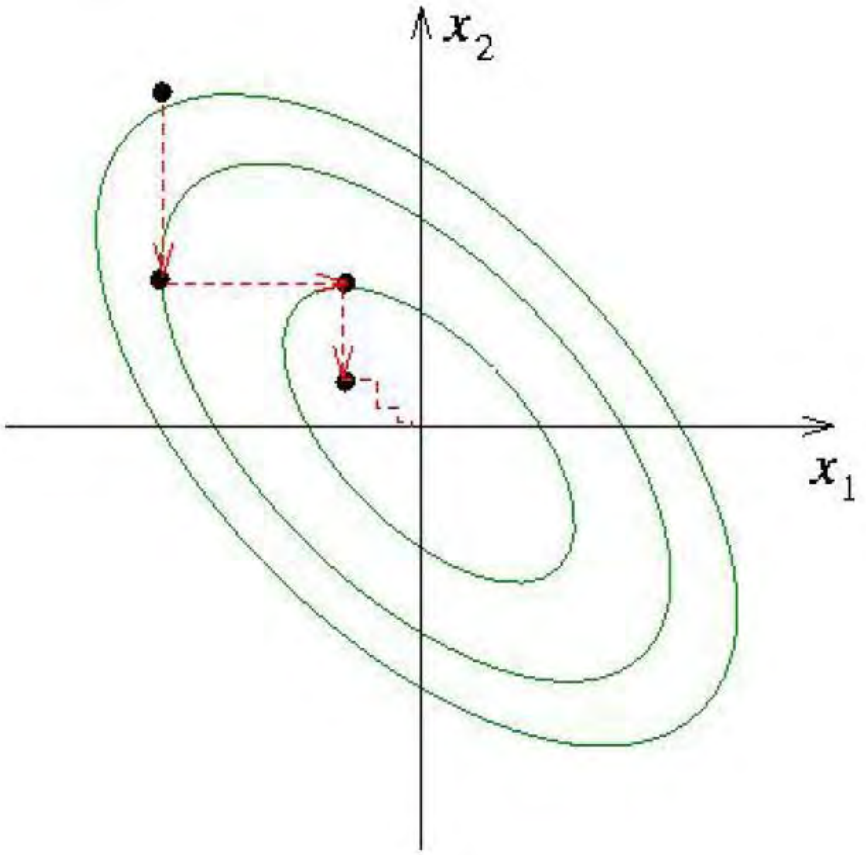
Gauss-Seidel: Choose $i$ cyclically, 1, 2,..., $n$, 1, 2,...

Gauss-Southwell: Choose $i$ with $|\frac{\partial f}{\partial x_i}(x)|$ maximum.

# Coordinate Descent: Example

## (Tseng, 2002)

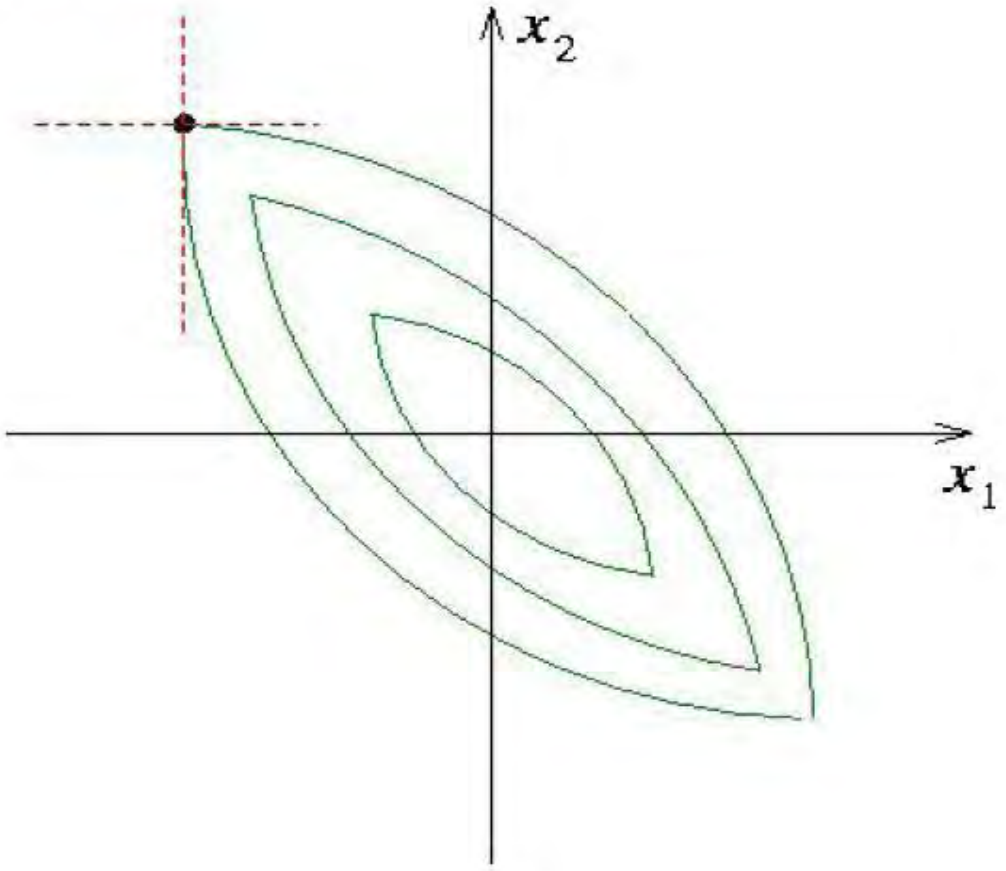
$$\min_{x=(x_1,x_2)} (x_1 + x_2)^2 + \frac{1}{4}(x_1 - x_2)^2$$



# Coordinate Descent: Convergent?

Example:

$$\min_{x=(x_1,x_2)} (x_1 + x_2)^2 + \frac{1}{4}(x_1 - x_2)^2 + |x_1 + x_2|$$



# Coordinate Descent: Convergent?

## (Tseng, 2002)

Given $x \in \Re^n$, choose $i \in \{1, ..., n\}$. Update

$$x^{\text{new}} = \arg\min_{u | u_j = x_j \ \forall j \neq i} f(u).$$

Repeat until "convergence".

- If $f(x)$ is smooth, then the algorithm is guaranteed to converge.
- If $f(x)$ is nonsmooth, the algorithm can get stuck.
- If the nonsmooth part is separable, convergence is guaranteed.

$$\min \ f(x) = \text{loss}(x) + \lambda \times \text{penalty}(x)$$

$$\text{penalty}(x) = \|x\|_1$$

# Coordinate Descent

Given $x \in \Re^n$, choose $i \in \{1, ..., n\}$. Update

$$x^{\text{new}} = \underset{u | u_j = x_j \ \forall j \neq i}{\arg\min} f(u).$$

Repeat until "convergence".

- Can $x^{new}$ be computed efficiently?

$$\min \ f(x) = \text{loss}(x) + \lambda \times \text{penalty}(x)$$

$$\text{penalty}(x) = \|x\|_1$$

$$\text{loss}(x) = 0.5 \times \|Ax - y\|^2_2$$

# CD in Sparse Representation

- Lasso (Fu, 1998; Friedman et al., 2007)
- L1/$L_q$ regularized least squares & logistic regression (Yuan and Lin, 2006,Liu et al., 2009; Argyriou et al., 2008; Meier et al., 2008)
- Sparse group Lasso for $q$=2 (Peng et al., 2010; Friedman et al., 2010)
- Fused Lasso Signal Approximator (Friedman et al., 2007; Hofling, 2010)
- Sparse inverse covariance estimation (Banerjee et al., 2008; Friedman et al., 2007)

# Summary of CD

Advantages:

- Easy for implementation, especially for the least squares loss
- Can be fast, especially the solution is very sparse

Disadvantages:

- No convergence rate
- Can be hard to derive $x^{new}$ for general loss
- Can get stuck when the penalty is non-separable

# Subgradient Descent

## (Nemirovski, 1994; Nesterov, 2004)

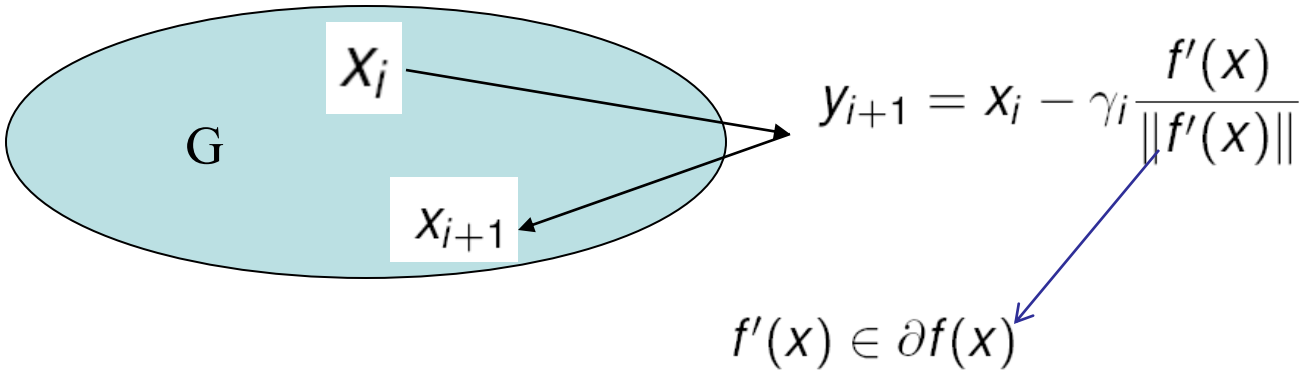
$$\min_{x \in G} f(x)$$

Repeat

$$y_{i+1} = x_i - \gamma_i \frac{f'(x)}{\|f'(x)\|}$$

$$x_{i+1} = \pi_G(y_{i+1}) \equiv \arg\min_{x \in G} \frac{1}{2}\|x - y_{i+1}\|_2^2$$

Until "convergence"

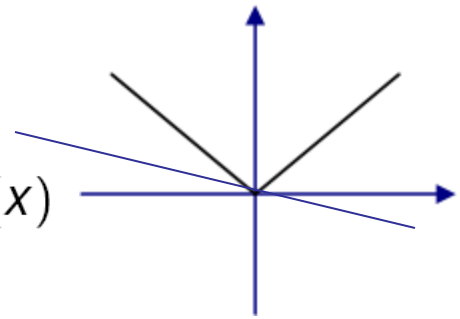


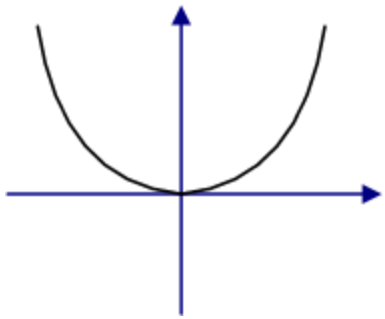

Subgradient: one element in the subdifferential set

# Subgradient Descent: Subgradient

## (Nemirovski, 1994; Nesterov, 2004)

$h'(x) \in \partial h(x)$





$h(x)=0.5\times(x-v)^2 + \lambda|x|$

$g(x)=0.5\times(x-v)^2 + \lambda x^2$

$$\partial h(x) = x - v + \lambda \mathrm{SGN}(x)$$

$$\mathrm{SGN}(x) = \begin{cases} \{1\} & x > 0 \\ \{1\} & x > 0 \\ [-1, 1] & x = 0 \end{cases}$$

g(x) is differentiable for all x

h(x) is non-differentiable at x=0

# Subgradient Descent: Convergent?

## (Nemirovski, 1994; Nesterov, 2004)

Repeat

$$y_{i+1} = x_i - \gamma_i \frac{f'(x)}{\|f'(x)\|}$$

$$x_{i+1} = \pi_G(y_{i+1}) \equiv \arg\min_{x \in G} \frac{1}{2}\|x - y_{i+1}\|_2^2$$

Until "convergence"

If $f$(x) is Lipschitz continuous with constant $L(f)$, the set $G$ is closed convex, and the step size is set as follows

$$\gamma_i = DN^{-1/2}, \ \ i = 1, ..., N,$$

then, we have

$$\varepsilon_N \leq O(1)\frac{L(f)D}{\sqrt{N}}$$

# SD in Sparse Representation

- $L_1$ constrained optimization (Duchi et al., 2008)
- $L_1/L_\infty$ constrained optimization (Quattoni et al., 2009)

$$\pi_G(v) \equiv \arg\min_{x \in G} \frac{1}{2}\|x - v\|_2^2$$

Advantages:

- Easy implementation
- Guaranteed global convergence

Disadvantages

- It converges slowly
- It does not take the structure of the non-smooth term into consideration

# Gradient Descent

$$\min_x f(x)$$

f(x) is smooth

Repeat

$$x_{i+1} = x_i - \gamma_i f'(x_i)$$

Until "convergence"

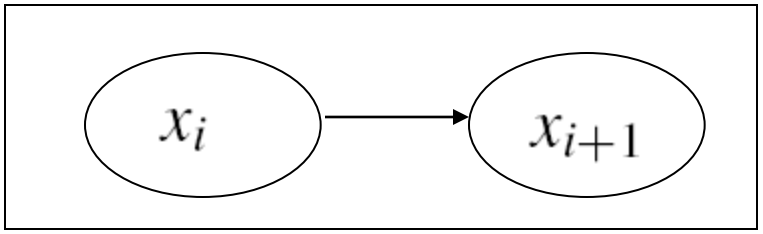
$$\min_{x \in G} f(x)$$

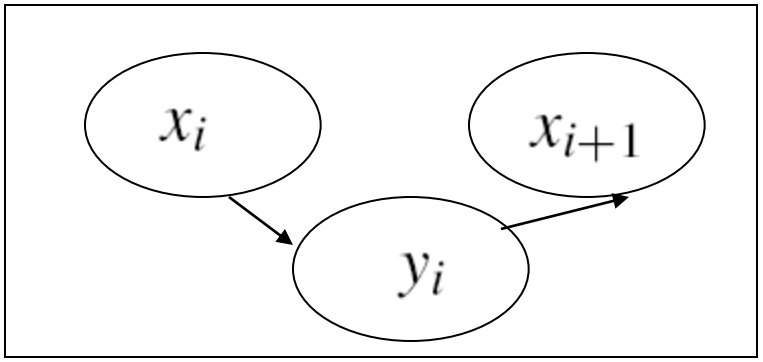
Repeat

$$y_i = x_i - \gamma_i f'(x_i)$$

$$x_{i+1} = \pi_G(y_i)$$

Until "convergence"




If $\gamma_i \leq \frac{1}{L}$, we can establish the convergence rate of $O(1/N)$.

# Gradient Descent:

## The essence of the gradient step

$$\min_x f(x)$$

Repeat

$$x_{i+1} = x_i - \gamma_i f'(x_i)$$

Until "convergence"

$$\min_{x \in G} f(x)$$

Repeat

$$y_i = x_i - \gamma_i f'(x_i)$$
$$x_{i+1} = \pi_G(y_i)$$

Until "convergence"

$$x_{i+1} = \arg\min_x \mathcal{M}(x_i, \gamma_i)$$

$$x_{i+1} = \arg\min_{x \in G} \mathcal{M}(x_i, \gamma_i)$$

$$\mathcal{M}(x_i, \gamma_i) = [f(x_i) + \langle f'(x_i), x - x_i \rangle] + \frac{1}{2\gamma_i}\|x - x_i\|_2^2$$

1st order Taylor expansion

Model

Regularization

# Gradient Descent:

## Extension to the composite model (Nesterov, 2007; Beck and Teboulle, 2009)

$$\min\ f(x) = \text{loss}(x) + \lambda \times \text{penalty}(x)$$

**Model**

$$\mathcal{M}(x_i, \gamma_i) = [\text{loss}(x_i) + \langle \text{loss}'(x_i), x - x_i \rangle] + \frac{1}{2\gamma_i}\|x - x_i\|_2^2 + \lambda \times \text{penalty}(x)$$

1st order Taylor expansion

Regularization

Nonsmooth part

Repeat

$$x_{i+1} = \arg\min \mathcal{M}(x_i, \gamma_i)$$

Until "convergence"

Convergence rate $O(1/N)$

Much better than $O(1/\sqrt{N})$ Subgradient descent

# Gradient Descent:

## Extension to the composite model (Nesterov, 2007; Beck and Teboulle, 2009)

### Model

$$\mathcal{M}(x_i, \gamma_i) = [\text{loss}(x_i) + \langle \text{loss}'(x_i), x - x_i \rangle] + \frac{1}{2\gamma_i}\|x - x_i\|_2^2 + \lambda \times \text{penalty}(x)$$

Repeat

$$x_{i+1} = \arg\min \mathcal{M}(x_i, \gamma_i)$$

Until "convergence"

$$\pi_{\text{penalty}}(v) = \frac{1}{2}\|x - v\|_2^2 + \rho \times \text{penalty}(x)$$

proximal operator (Moreau, 1965)

$$v = x_i - \gamma_i \text{loss}'(x_i)$$

$$\rho = \gamma_i \lambda$$

# Accelerated Gradient Descent:

## unconstrained version (Nesterov, 2007; Beck and Teboulle, 2009)

The lower complexity bound shows that, the first-order methods can not achieve a better convergence rate than $O(1/N^2)$.

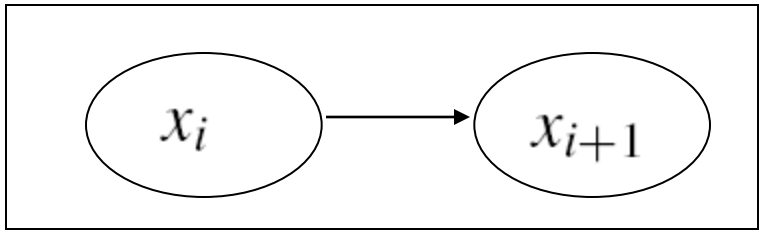
Can we develop a method that can achieves the optimal convergence rate?

GD $\min_x f(x)$

Repeat

$$x_{i+1} = x_i - \gamma_i f'(x_i)$$

Until "convergence"



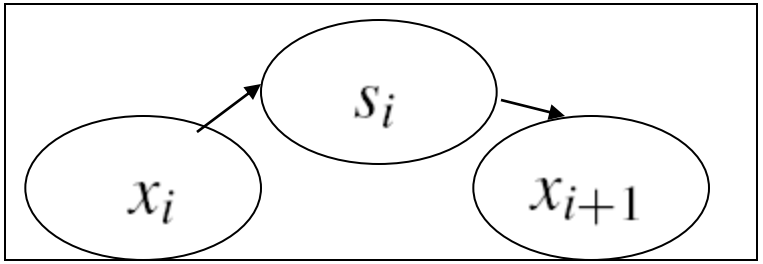

$O(1/N)$

AGD

Repeat

$$s_i = x_i + \alpha_i(x_i - x_{i-1})$$
$$x_{i+1} = s_i - \gamma_i f'(s_i)$$

Until "convergence"




$O(1/N^2)$

# Accelerated Gradient Descent:

## constrained version (Nesterov, 2007; Beck and Teboulle, 2009)

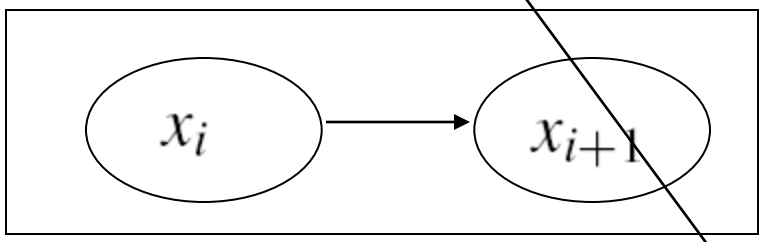
GD

$$\min_{x \in G} f(x)$$

Repeat

$$x_{i+1} = \pi_G(x_i - \gamma_i f'(x_i))$$

Until "convergence"




$O(1/N)$

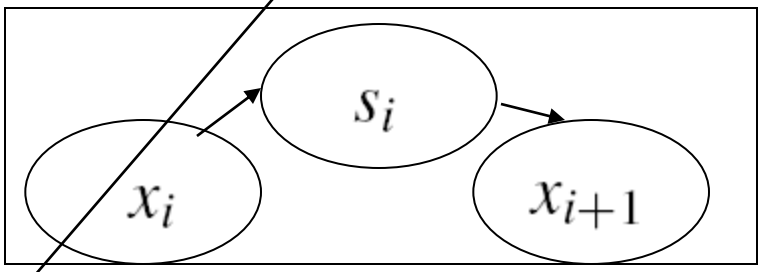
AGD

Repeat

$$s_i = x_i + \alpha_i(x_i - x_{i-1})$$

$$x_{i+1} = \pi_G(s_i - \gamma_i f'(s_i))$$

Until "convergence"




$O(1/N^2)$

Can the projection be computed efficiently?

# Accelerated Gradient Descent:

## composite model (Nesterov, 2007; Beck and Teboulle, 2009)

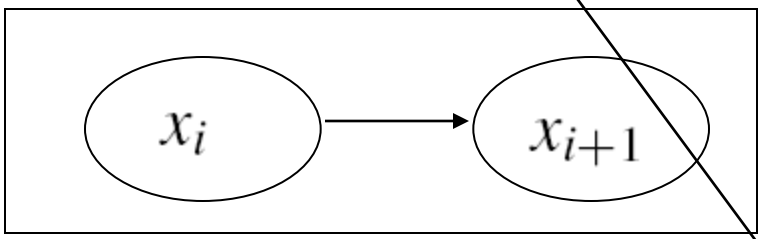
$$\mathcal{M}(x_i, \gamma_i) = [\text{loss}(x_i) + \langle \text{loss}'(x_i), x - x_i \rangle] + \frac{1}{2\gamma_i}\|x - x_i\|_2^2 + \lambda \times \text{penalty}(x)$$

GD min $f(x) = \text{loss}(x) + \lambda \times \text{penalty}(x)$

Repeat

$$x_{i+1} = \arg\min \mathcal{M}(x_i, \gamma_i)$$

Until "convergence"




$O(1/N)$

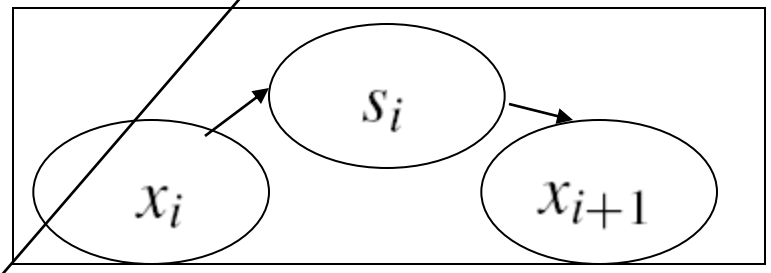
AGD

Repeat

$$s_i = x_i + \alpha_i(x_i - x_{i-1})$$

$$x_{i+1} = \arg\min \mathcal{M}(s_i, \gamma_i)$$

Until "convergence"




$O(1/N^2)$

Can the proximal operator be computed efficiently?

# Accelerated Gradient Descent in Sparse Representations

- Lasso (Nesterov, 2007; Beck and Teboulle, 2009)
- $L_1/L_q$ (Liu, Ji, and Ye, 2009; Liu and Ye, 2010)
- Sparse group Lasso (Liu and Ye, 2010)
- Trace Norm (Ji and Ye, 2009; Pong et al., 2009; Toh and Yun, 2009; Lu et al., 2009)
- Fused Lasso (Liu, Yuan, and Ye, 2010)

Advantages:

- Easy for implementation
- Optimal convergence rate
- Scalable to large sample size problem

# Accelerated Gradient Descent in Sparse Representations

Advantages:

- Easy for implementation
- Optimal convergence rate
- Scalable to large sample size problem

Key computational cost

- Gradient and functional value
- The projection (for constrained smooth optimization)
- The proximal operator (for composite function)

# Euclidean projection onto the $L_1$ ball

## (Duchi et al., 2008; Liu and Ye, 2009)

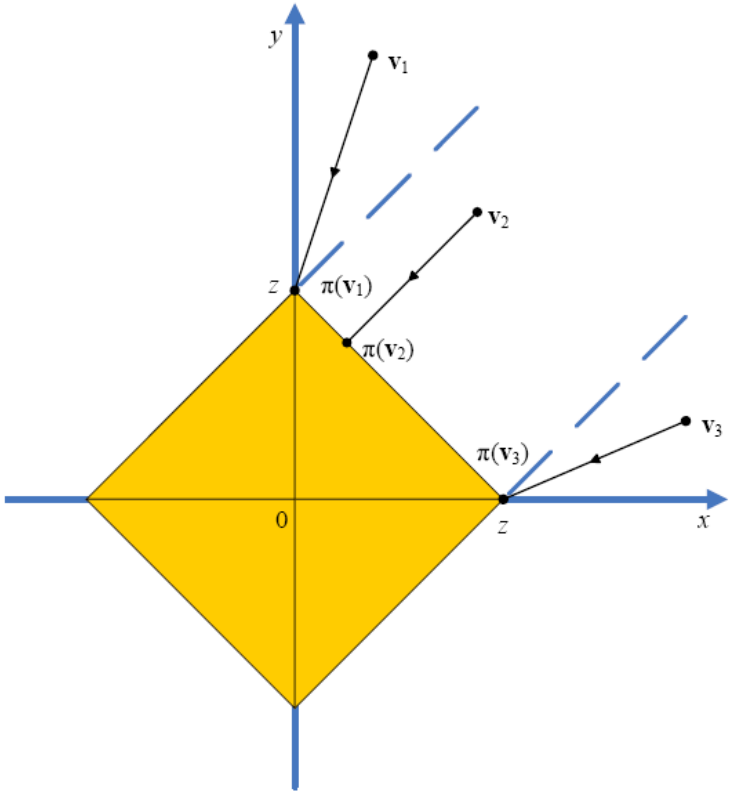
$$\min_{\|x\|_1 \le z} \text{loss}(x)$$

$$\pi(v) = \min_{\|x\|_1 \le z} \frac{1}{2}\|x - v\|_2^2$$

zero finding

$$x_i^\star = \text{sgn}(v_i)\max(|v_i| - \lambda^\star, 0)$$

$$f(\lambda) = \sum_{i=1}^{n} \max(|v_i| - \lambda, 0) - z$$



# Efficient Computation of the Proximal Operators

## (Liu and Ye, 2010; Liu and Ye, 2010; Liu, Yuan and Ye, 2010)

$$\min \ f(x) = \text{loss}(x) + \lambda \times \text{penalty}(x)$$

$$\pi_{\text{penalty}}(v) = \frac{1}{2}\|x - v\|_2^2 + \rho \times \text{penalty}(x)$$

- $L_1/L_q$
- Sparse group Lasso
- Fused Lasso

# Proximal Operator Associated with $L_1/L_q$

**Optimization problem:**

$$\min_{\mathbf{W}\in\mathbb{R}^p} f(\mathbf{W}) \equiv l(\mathbf{W}) + \lambda \sum_{i=1}^{k} \|\mathbf{w}_i\|_q$$

**Associated proximal operator:**

$$\pi_{1q}(\mathbf{V},\lambda) = \arg\min_{\mathbf{X}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{X}-\mathbf{V}\|_2^2 + \lambda \sum_{i=1}^{k} \|\mathbf{x}_i\|_q$$

It can be decoupled into the following $q$-norm regularized Euclidean projection problem:

$$\pi_q(\mathbf{v}) = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left( g(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{v}\|_2^2 + \lambda\|\mathbf{x}\|_q \right)$$

# Proximal Operator Associated with $L_1/L_q$

$$\pi_q(\mathbf{v}) = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left( g(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{v}\|_2^2 + \lambda\|\mathbf{x}\|_q \right)$$

**Method:**

Convert it to two simple zero finding algorithms

**Characteristics:**

1. Suitable to any $q \geq 1$
2. The proximal plays a key building block in quite a few methods such as the accelerated gradient descent, coordinate gradient descent(Tseng, 2008), forward-looking subgradient (Duchi and Singer, 2009), and so on.

# Achieving a Zero Solution

For the $q$-norm regularized Euclidean projection

$$\pi_q(\mathbf{v}) = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left( g(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{v}\|_2^2 + \lambda\|\mathbf{x}\|_q \right)$$

We have

$$\pi_q(\mathbf{v}) = \mathbf{0} \textit{ if and only if } \lambda \geq \|\mathbf{v}\|_{\bar{q}}. \qquad \frac{1}{q}+\frac{1}{\bar{q}}=1$$

We restrict our following discussion to $\mathbf{v} > \mathbf{0}$

$$g'(\mathbf{x}^*) = 0 \qquad \Rightarrow \qquad \mathbf{x}^* + \lambda\|\mathbf{x}^*\|_q^{1-q}\mathbf{x}^{*(q-1)} = \mathbf{v}$$

*where* $\mathbf{y} \equiv \mathbf{x}^{(q-1)}$ *is defined component-wisely as:* $y_i = \mathrm{sgn}(x_i)|x_i|^{q-1}$

# Fixed Point Iteration?

One way to compute $\pi_q(\mathbf{v})$ is to apply the following fixed point iteration

$$\mathbf{x}^* = \mathbf{v} - \lambda \|\mathbf{x}^*\|_q^{1-q} \mathbf{x}^{*(q-1)}$$

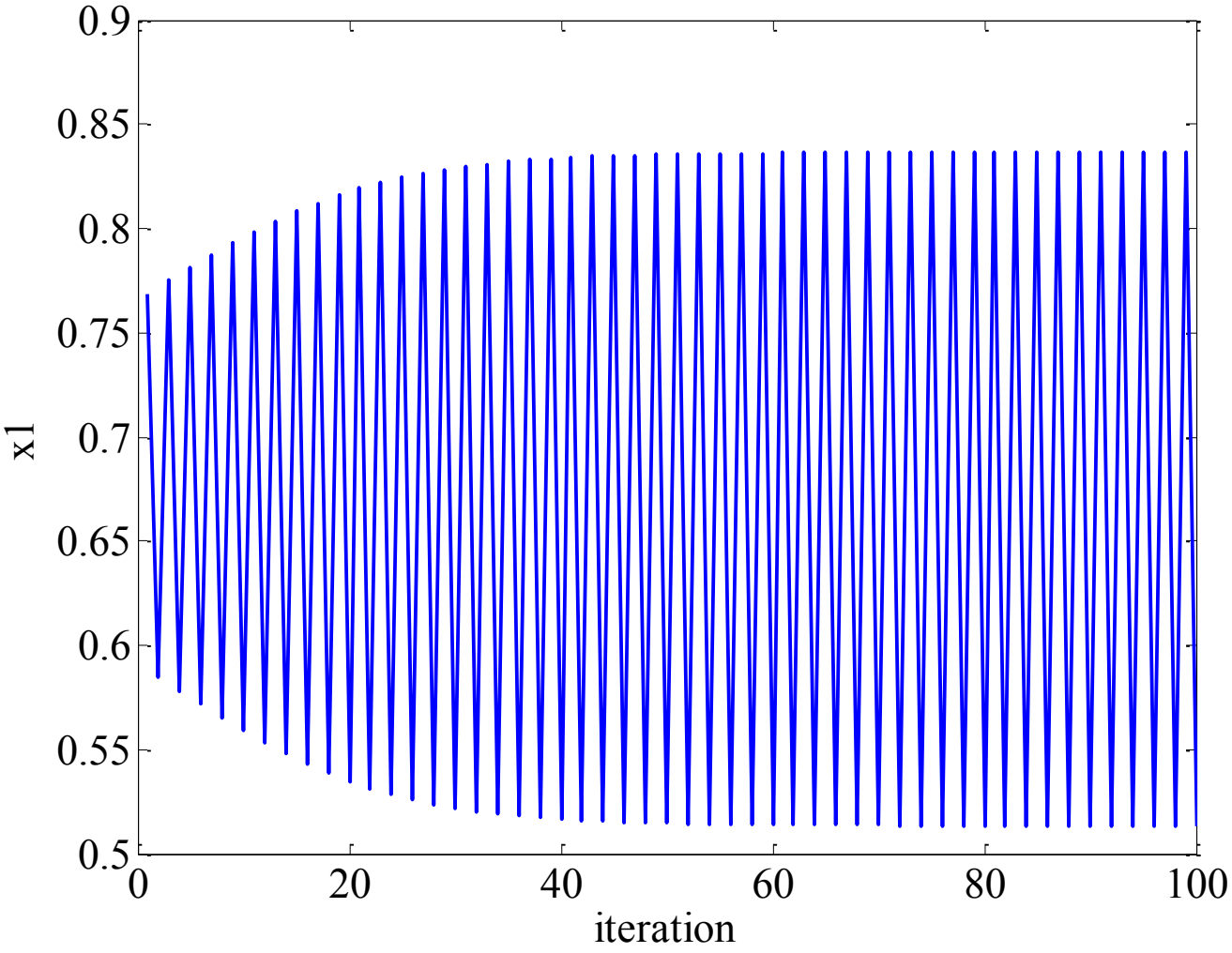
Let us consider the two-dimensional case.

$$\mathbf{v} = [1; 2],\ q = 3 \text{ and } \lambda = 1$$

We start from $[0; 0]$

Fixed point iteration is not guaranteed to converge.

# Fixed Point Iteration?



# Solving Two Zero Finding Problems

The methodology is also based on the following relationship

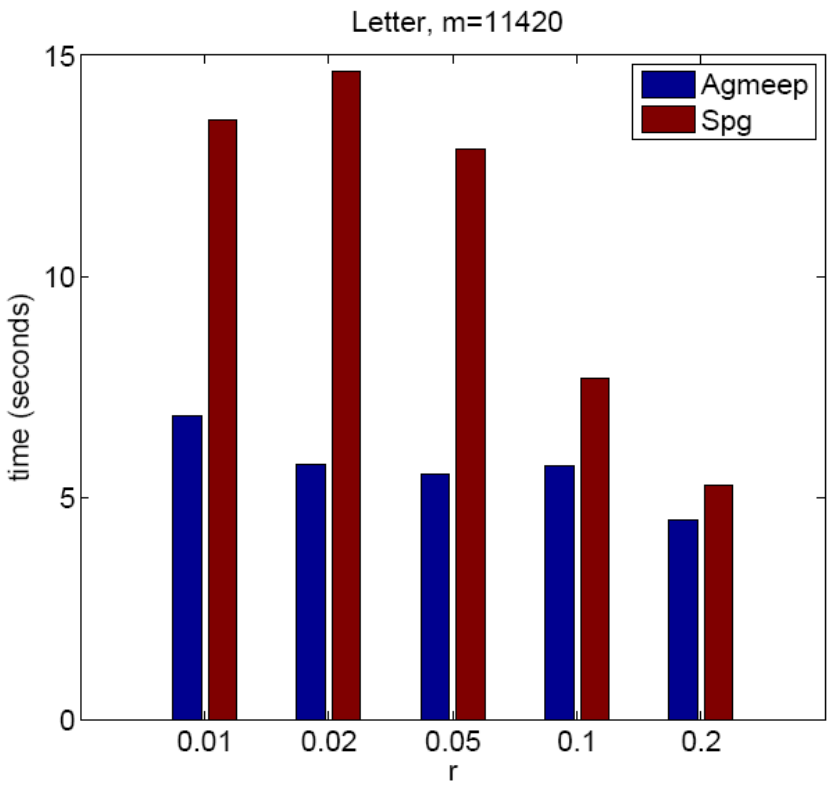
$$\mathbf{x}^* + \lambda \|\mathbf{x}^*\|_q^{1-q} \mathbf{x}^{*(q-1)} = \mathbf{v}$$

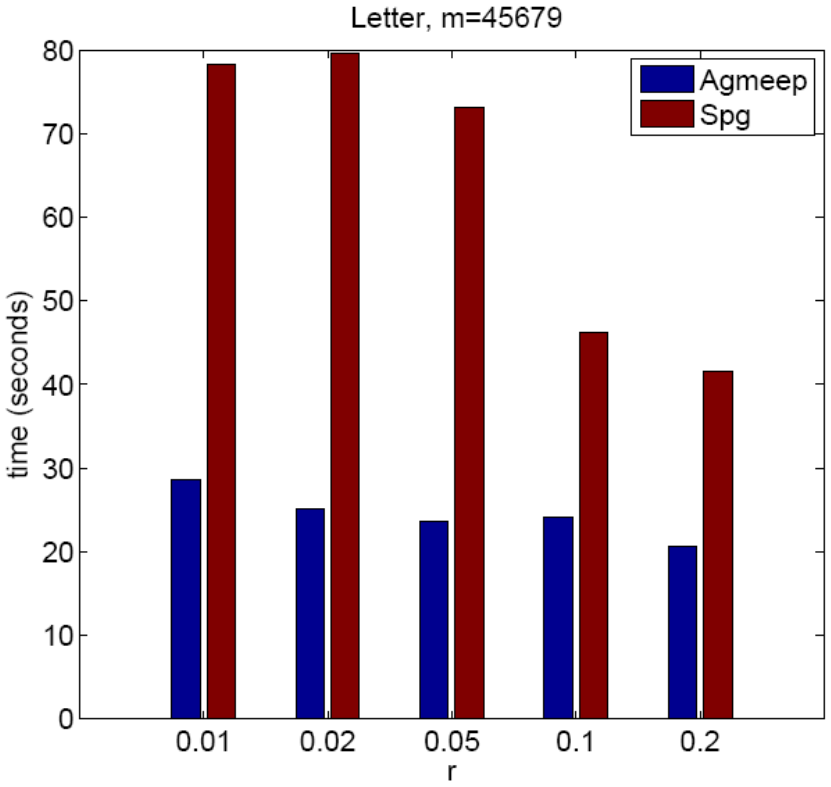
Construct three auxiliary functions.

Solve the above equation by two simple one-dimensional zero finding problems (associated with the constructed auxiliary functions).

# Efficiency

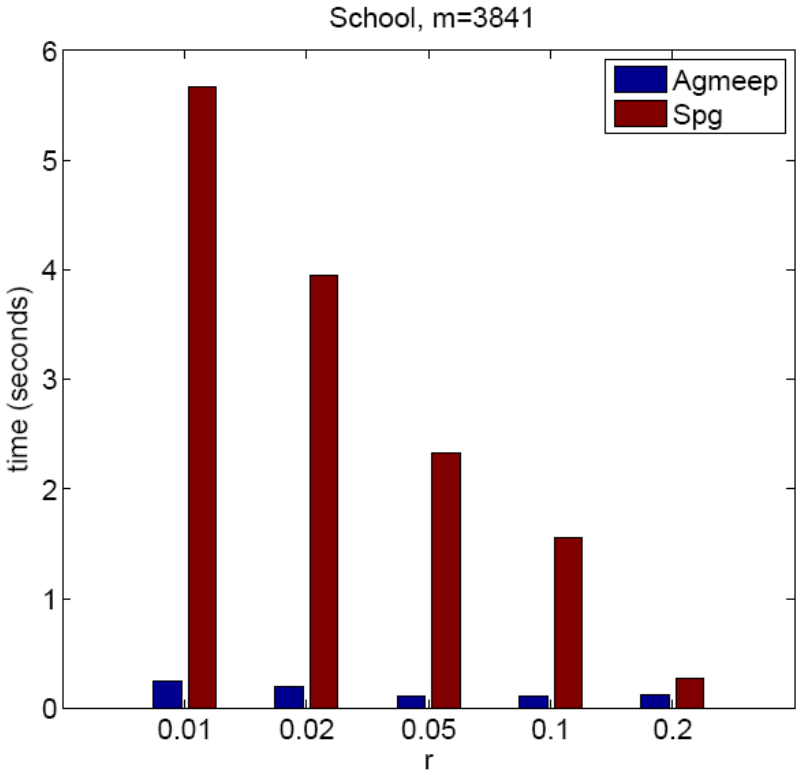
## (compared with spectral projected gradient)

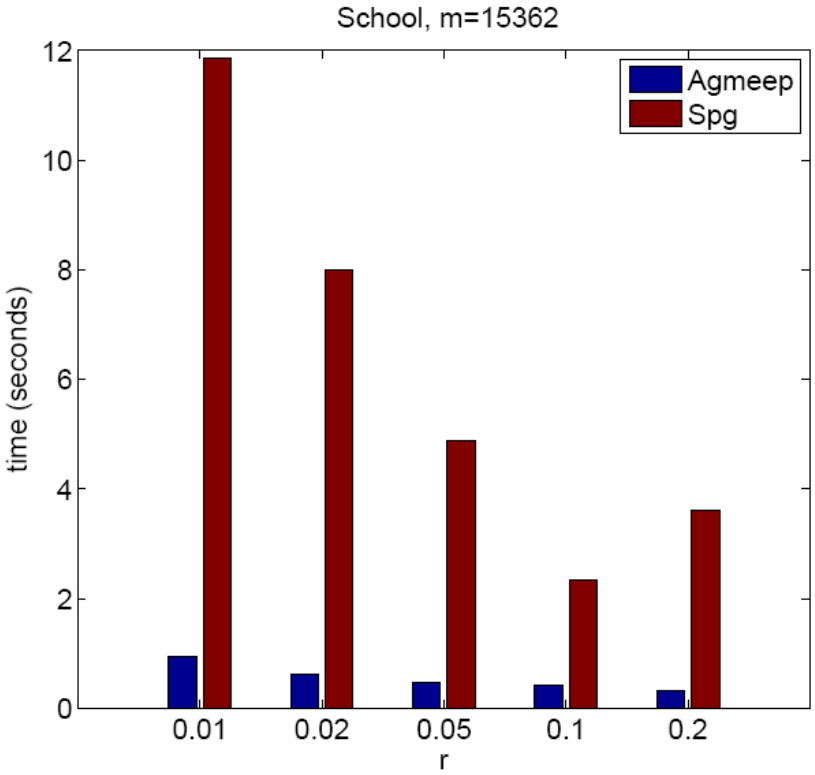


# Efficiency

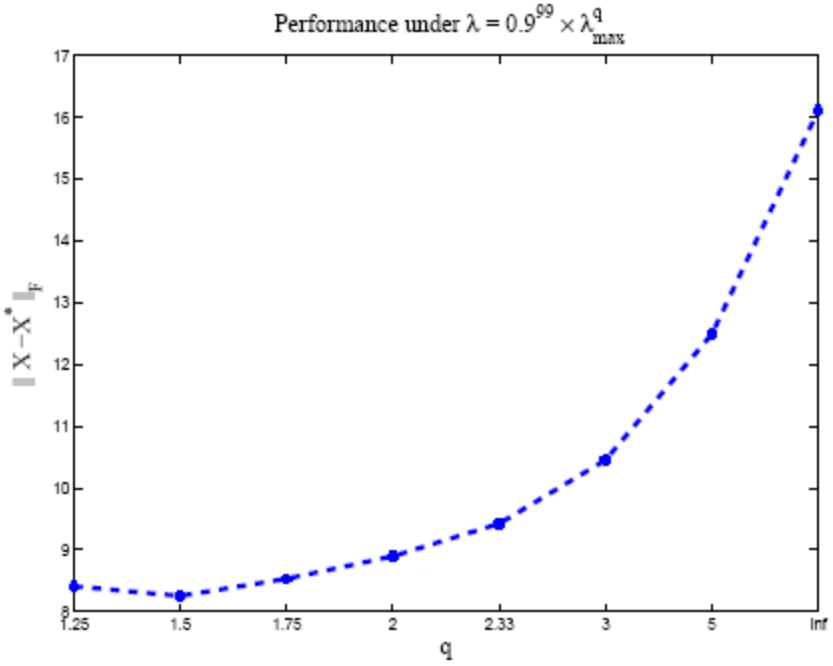
## (compared with spectral projected gradient)



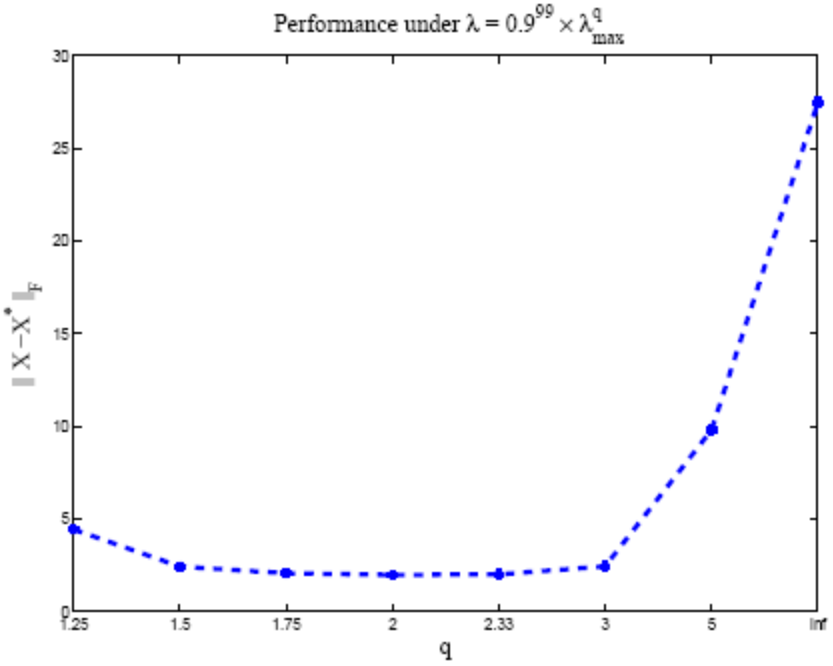



# Effect of $q$ in $L_1/L_q$

Multivariate linear regression



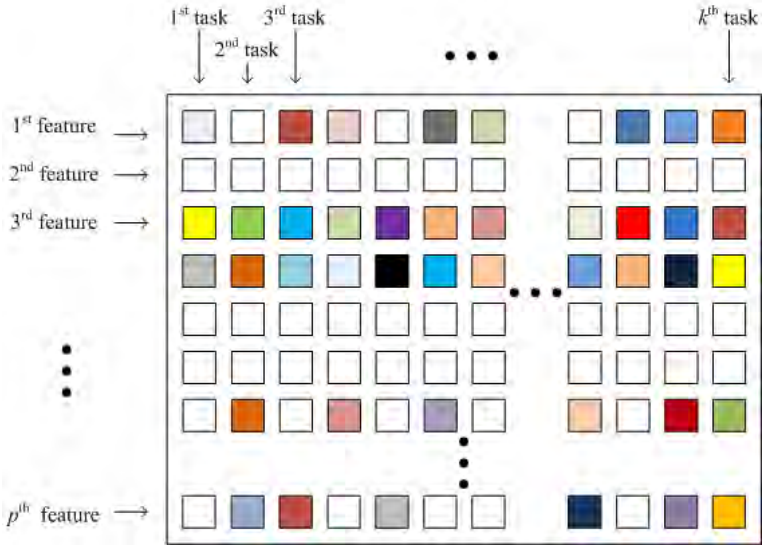

RAND




RANDN

# Proximal Operator Associated with $L_1$+$L_1/L_q$

## (Liu and Ye, 2010)




**Optimization problem:**

$$\min_X \text{loss}(X) + \lambda_1 \|X\|_1 + \lambda_q \|X\|_{\ell_1/\ell_q}$$

**Associated proximal operator:**

$$\pi^{\lambda_1}_{\lambda_q}(V) = \arg\min_X \frac{1}{2}\|X - V\|_F^2 + \lambda_1 \|X\|_1 + \lambda_q \|X\|_{\ell_1/\ell_q}$$

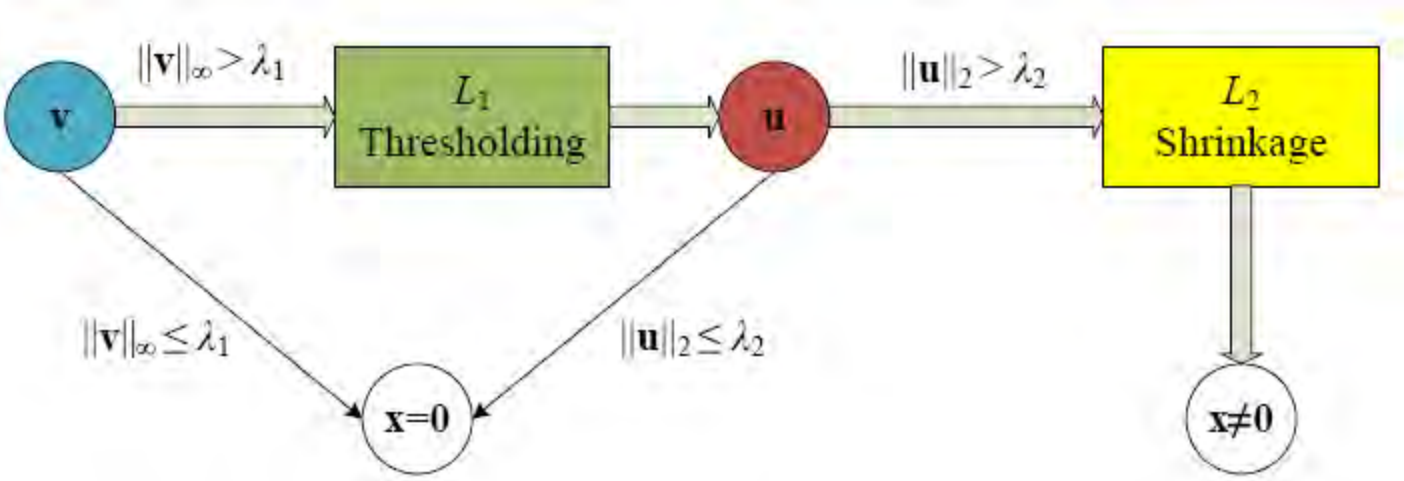
$$\pi^{\lambda_1}_{\lambda_q}(v) = \arg\min_x \frac{1}{2}\|x - v\|_F^2 + \lambda_1 \|x\|_1 + \lambda_q \|x\|_{\ell_q}$$

# Proximal Operator Associated with $L_1$+$L_1/L_2$

## (Liu and Ye, 2010)

$$\pi_{\lambda_q}^{\lambda_1}(v) = \arg\min_x \frac{1}{2}\|x - v\|_F^2 + \lambda_1\|x\|_1 + \lambda_q\|x\|_{\ell_q}$$

$$\pi_{\lambda_2}^{\lambda_1}(v) = \begin{cases} \mathbf{0} & if\ \|\mathbf{u}\|_2 \le \lambda_2 \\ \dfrac{\|\mathbf{u}\|_2 - \lambda_2}{\|\mathbf{u}\|_2}\mathbf{u} & if\ \|\mathbf{u}\|_2 > \lambda_2. \end{cases}$$



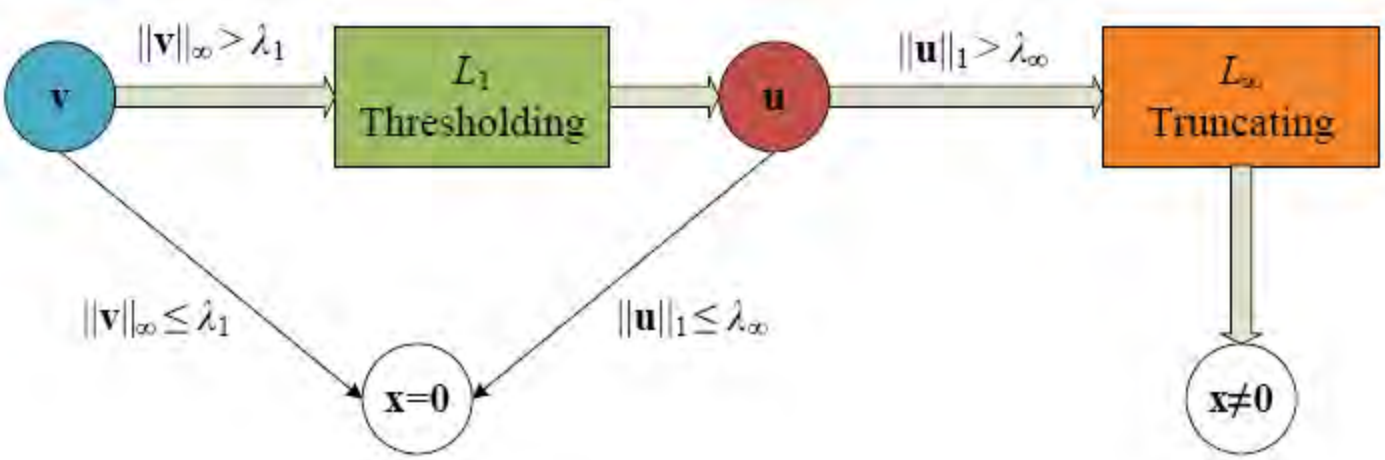
# Proximal Operator Associated with $L_1$+$L_1$/$L_{inf}$

## (Liu and Ye, 2010)

$$\pi_{\lambda_q}^{\lambda_1}(v) = \arg\min_x \frac{1}{2}\|x - v\|_F^2 + \lambda_1\|x\|_1 + \lambda_q\|x\|_{\ell_q}$$

$$\pi_{\lambda_\infty}^{\lambda_1}(v) = \begin{cases} \mathbf{0} & \|\mathbf{u}\|_1 \le \lambda_\infty \\ \mathrm{sgn}(\mathbf{u}) \odot \min(|\mathbf{u}|, t^*) & \|\mathbf{u}\|_1 > \lambda_\infty. \end{cases}$$

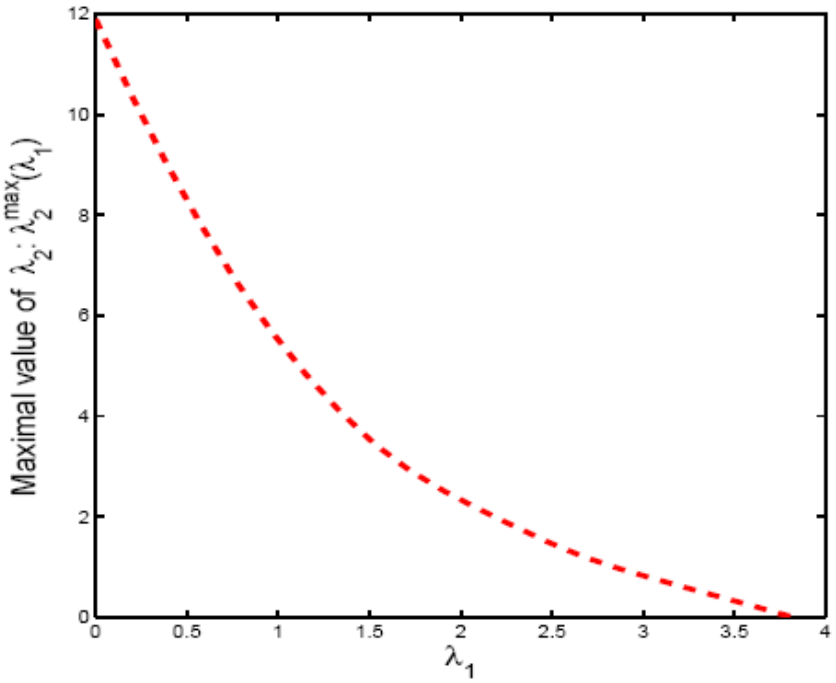
t* is the root of $h(t) = \sum_{i=1}^{n} \max(|u_i| - t, 0) - \lambda_\infty$

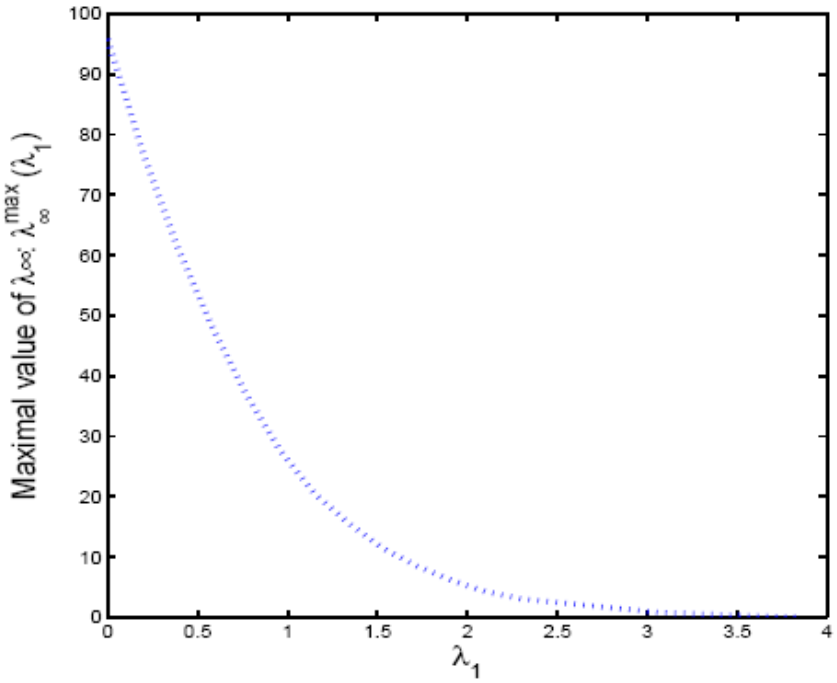


# Effective Interval for Sparse Group Lasso

## (Liu and Ye, 2010)

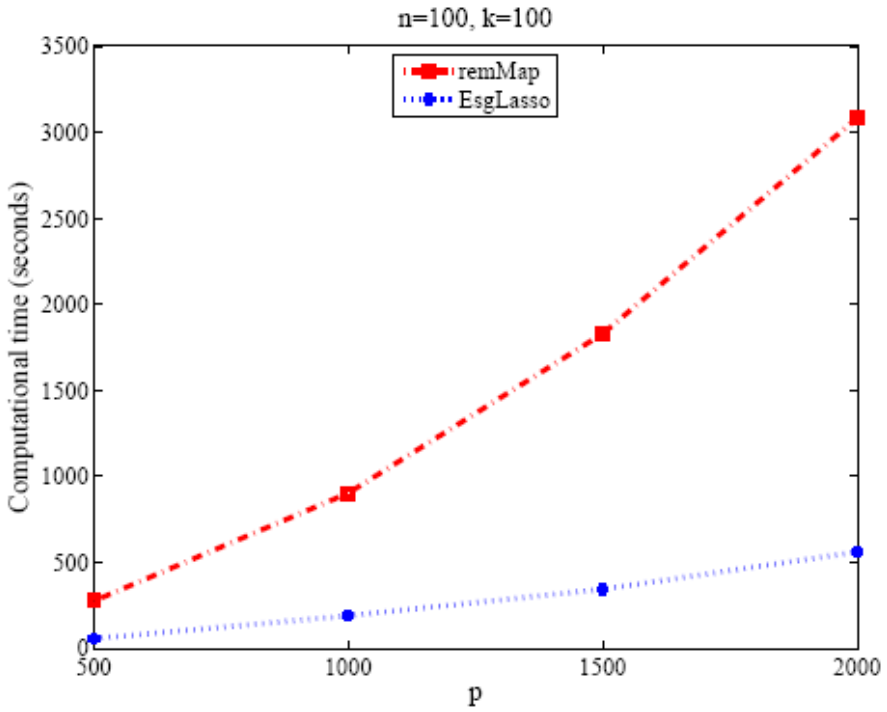
$$\min_X \text{loss}(X) + \lambda_1 \|X\|_1 + \lambda_q \|X\|_{\ell_1/\ell_q}$$

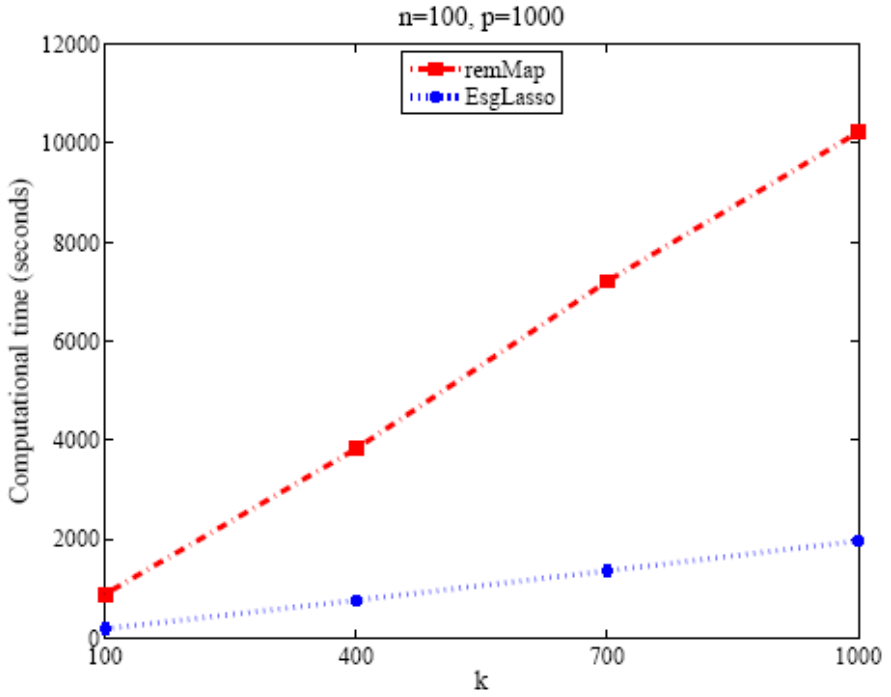


# Efficiency

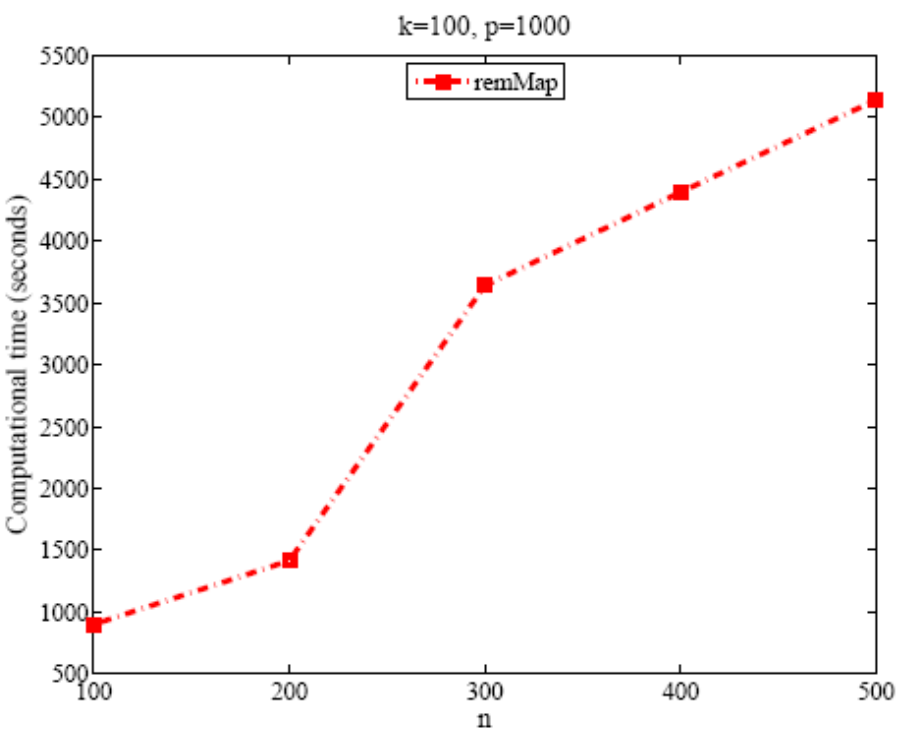
## (comparison with remMap via coordiante descent)

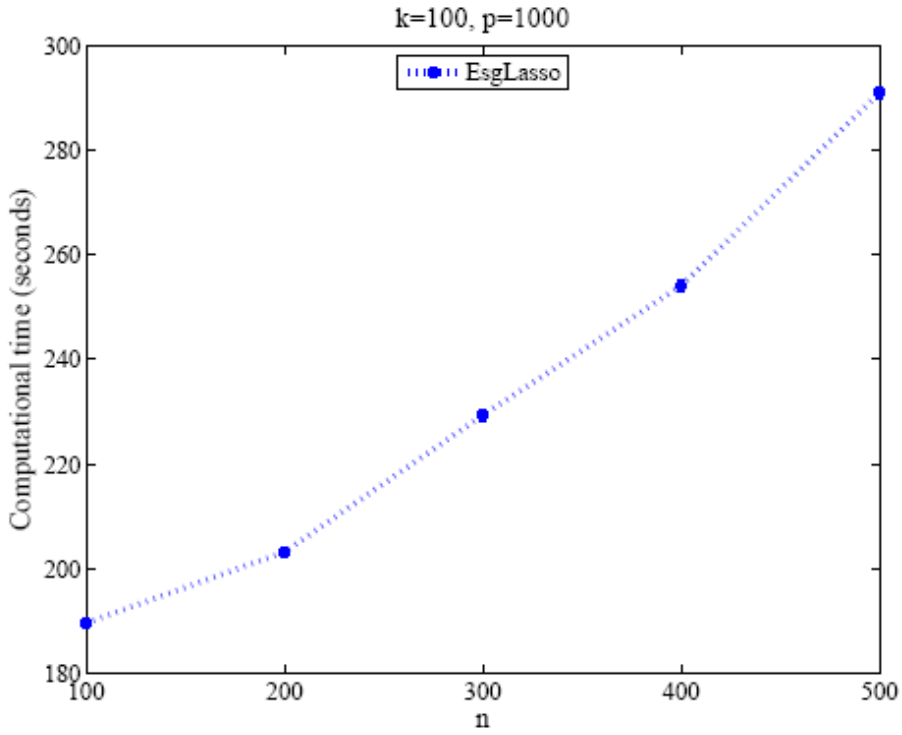


# Efficiency

## (comparison with remMap via coordiante descent)



# Proximal Operator Associated with Fused Lasso

**Optimization problem:**

$$\min_x \text{loss}(x) + \lambda_1 \sum_{i=1}^{n} |x_i| + \lambda_2 \sum_{i=1}^{n-1} |x_i - x_{i+1}|$$

**Associated proximal operator:**

$$\pi_{\lambda_2}^{\lambda_1}(\mathbf{v}) = \arg\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|^2 + \lambda_1 \sum_{i=1}^{n} |x_i| + \lambda_2 \sum_{i=1}^{n-1} |x_i - x_{i+1}|$$

Fused Lasso Signal Approximator

# Fused Lasso Signal Approximator

## (Liu, Yuan, and Ye, 2010)

$$\pi_{\lambda_2}^{\lambda_1}(\mathbf{v}) = \arg\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{x}-\mathbf{v}\|^2 + \lambda_1 \sum_{i=1}^{n} |x_i| + \lambda_2 \sum_{i=1}^{n-1} |x_i - x_{i+1}|$$

THEOREM 1. *For any* $\lambda_1, \lambda_2 \geq 0$, *we have*

$$\pi_{\lambda_2}^{\lambda_1}(\mathbf{v}) = \mathrm{sgn}(\pi_{\lambda_2}^{0}(\mathbf{v})) \odot \max(|\pi_{\lambda_2}^{0}(\mathbf{v})| - \lambda_1, 0).$$

Let $$R_{ij} = \begin{cases} -1 & j = i, i = 1, 2, \ldots, n-1 \\ 1 & j = i+1, i = 1, 2, \ldots, n-1 \\ 0 & \text{otherwise,} \end{cases}$$

We have $$\pi_{\lambda_2}(\mathbf{v}) = \arg\min f_{\lambda_2}(\mathbf{v}) \equiv \frac{1}{2}\|\mathbf{x}-\mathbf{v}\|^2 + \lambda_2 \|R\mathbf{x}\|_1$$

# Fused Lasso Signal Approximator

## (Liu, Yuan, and Ye, 2010)

$$\pi_{\lambda_2}(\mathbf{v}) = \arg\min f_{\lambda_2}(\mathbf{v}) \equiv \frac{1}{2}\|\mathbf{x}-\mathbf{v}\|^2 + \lambda_2\|R\mathbf{x}\|_1$$

Method:

- Subgradient Finding Algorithm (SFA)---looking for an appropriate and unique subgradient of $\|R\mathbf{x}\|_1$ at the minimizer (motivated by the proof of Theorem 1).

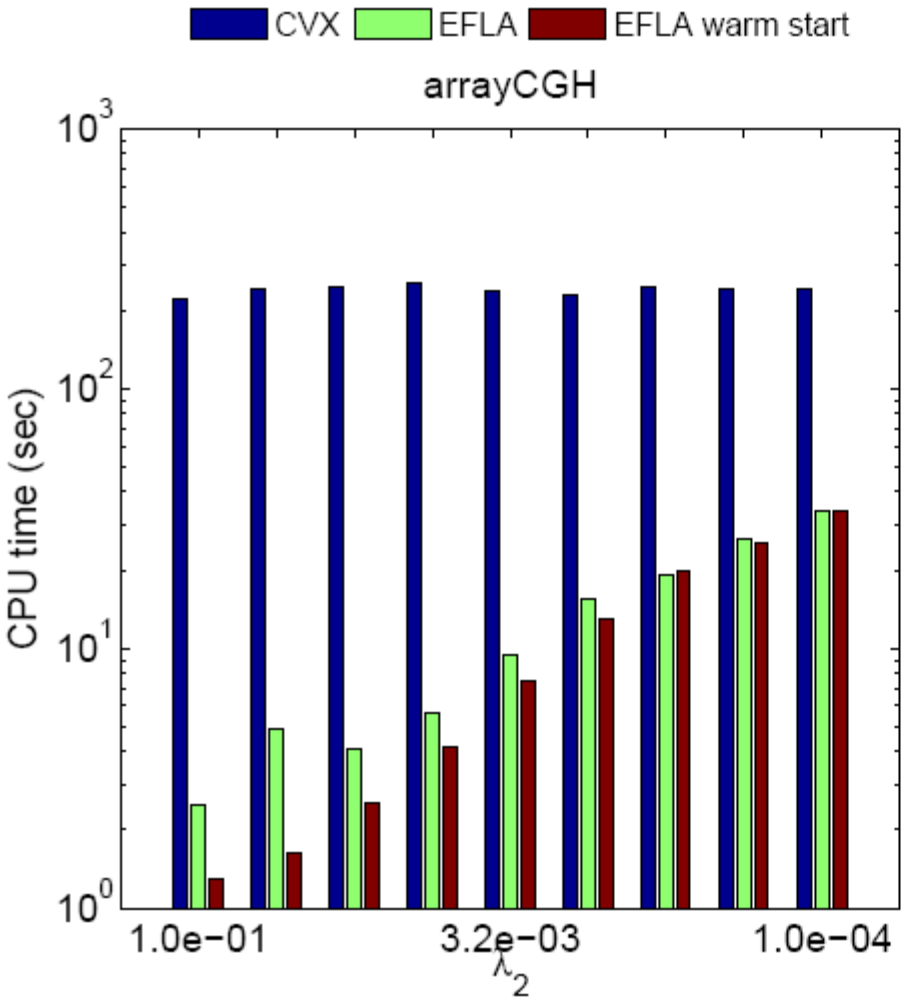
$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\|\mathbf{z}\|_\infty \le \lambda_2} \phi(\mathbf{x},\mathbf{z}) \equiv \frac{1}{2}\|\mathbf{x}-\mathbf{v}\|^2 + \langle R\mathbf{x},\mathbf{z}\rangle.$$

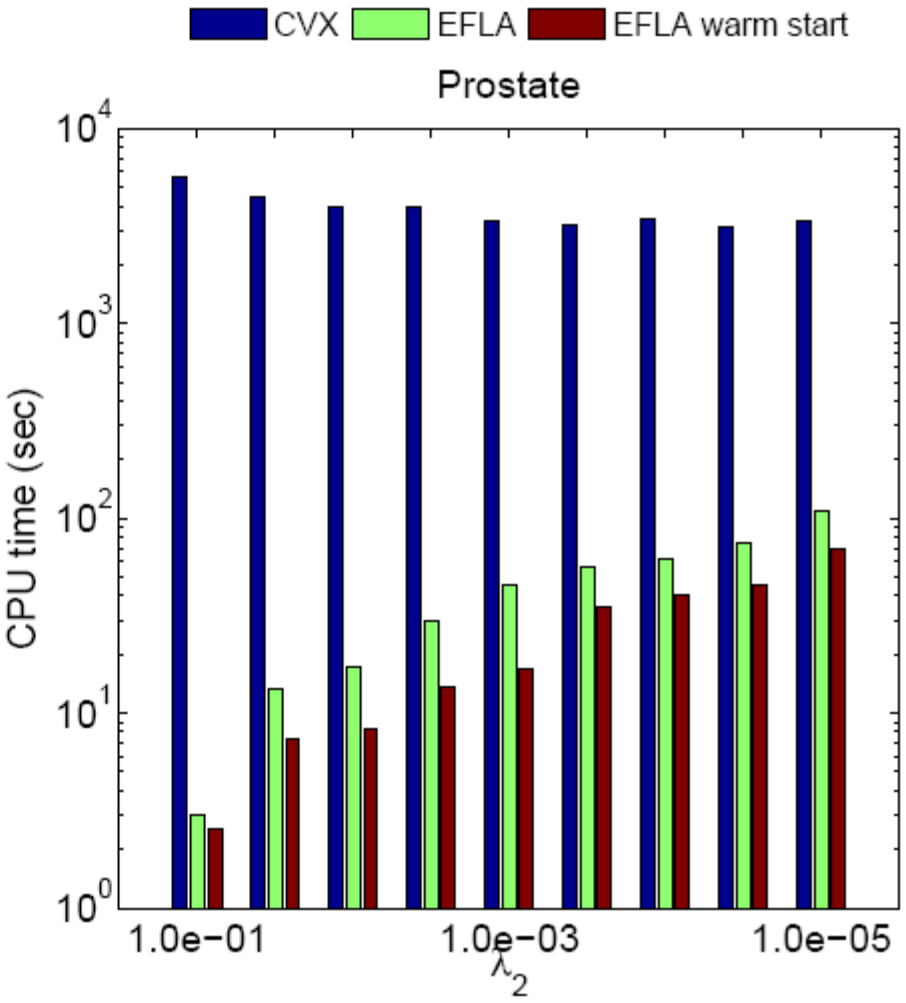
$$\mathbf{x} = \mathbf{v} - R^{\mathrm{T}}\mathbf{z}.$$

$$\min_{\|\mathbf{z}\|_\infty \le \lambda_2} \psi(\mathbf{z}) \equiv -\phi(\mathbf{v}-R^{\mathrm{T}}\mathbf{z},\mathbf{z}) = \frac{1}{2}\|R^{\mathrm{T}}\mathbf{z}\|^2 - \langle R^{\mathrm{T}}\mathbf{z},\mathbf{v}\rangle.$$

# Efficiency

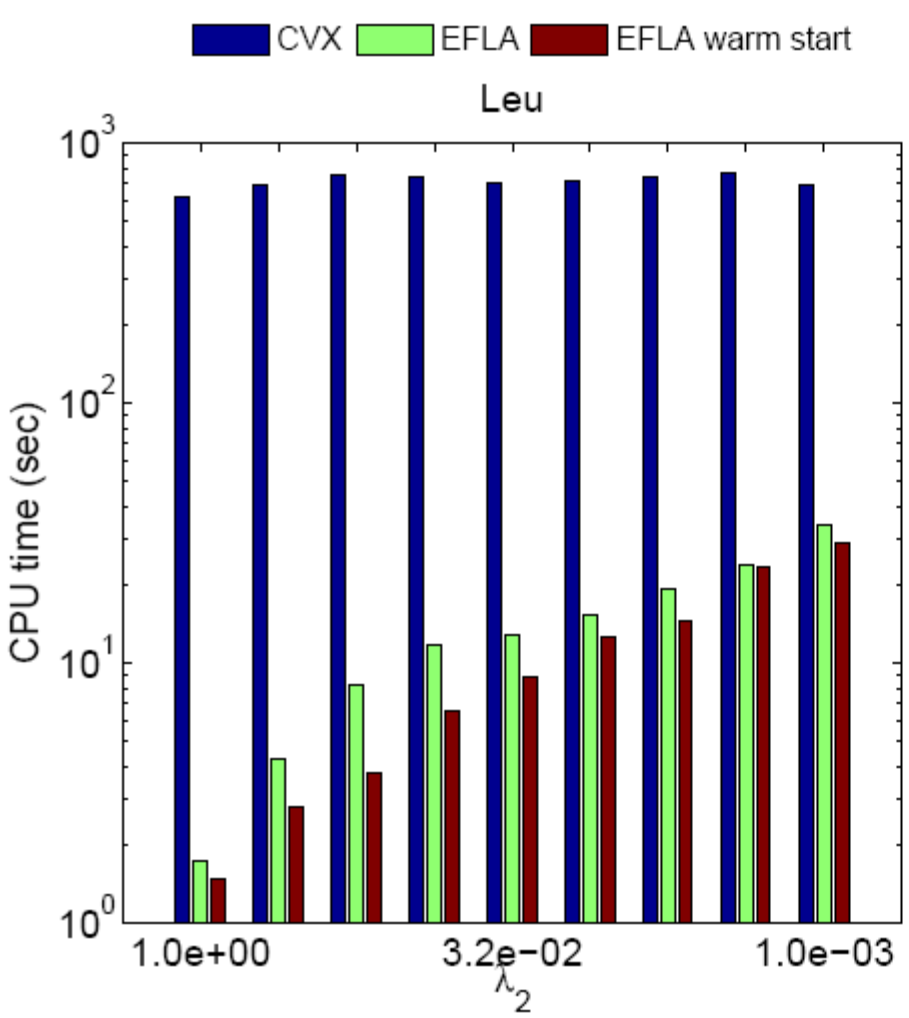
## (Comparison with the CVX solver)

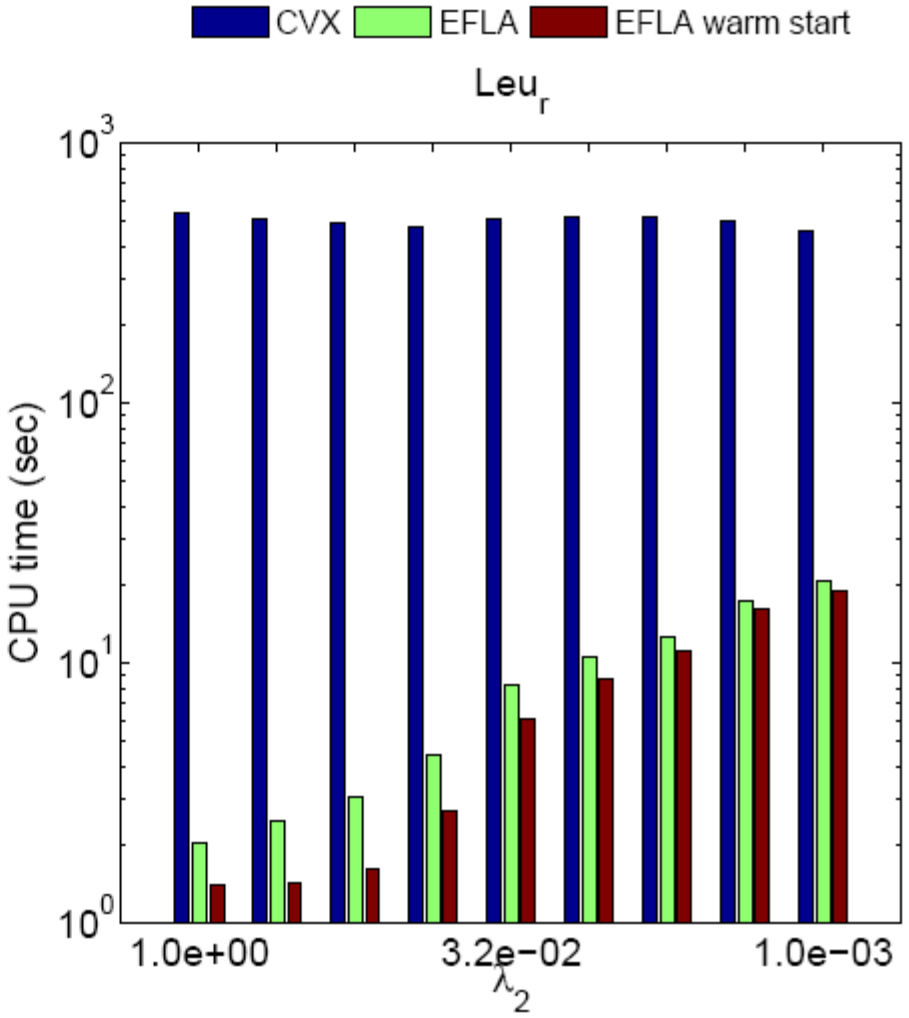


# Efficiency

## (Comparison with the CVX solver)



# Summary of Implementations

- The accelerated gradient descent, which is an optimal first-order method, is favorable for large-scale optimization.
- To apply the accelerated gradient descent, the key is to develop efficient algorithms for solving either the projection or the associated proximal operator.

**SLEP Package**

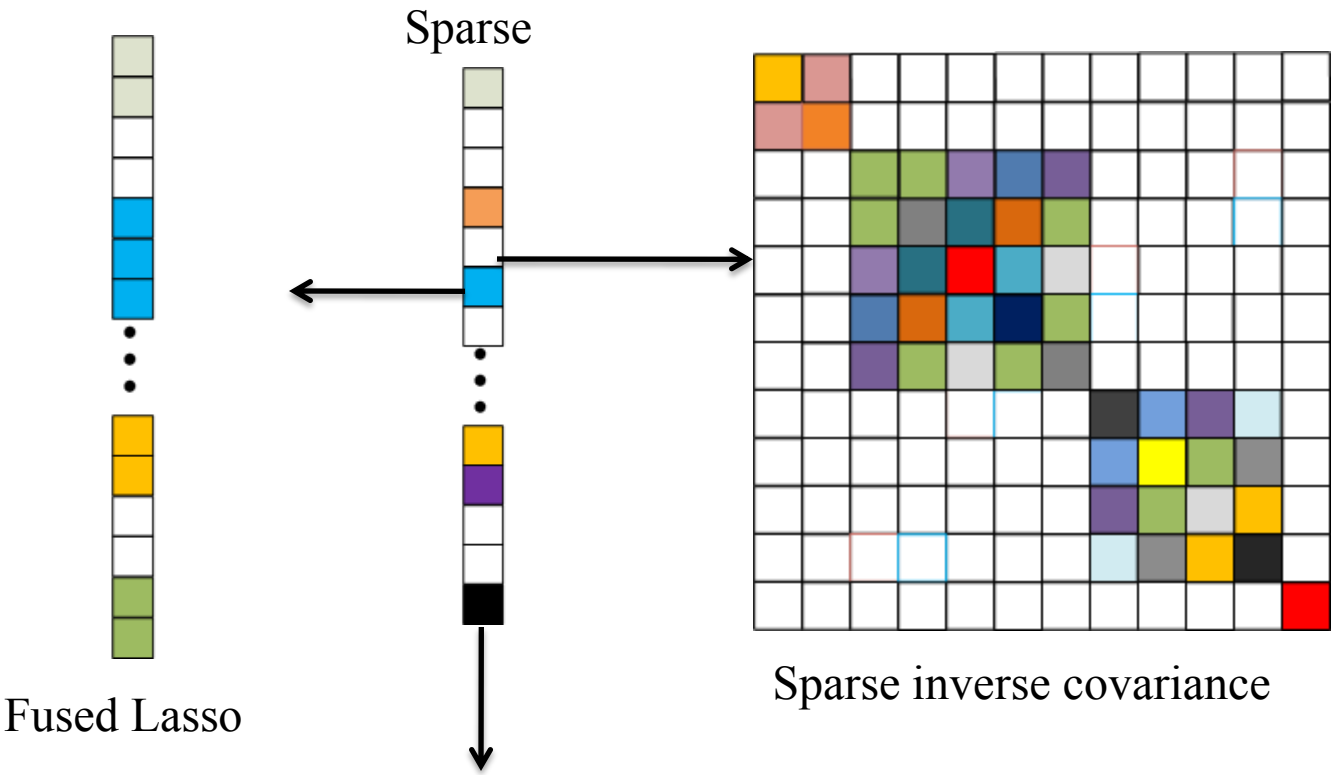
http://www.public.asu.edu/~jye02/Software/SLEP

# Outline

- Sparsity via $L_1$
- Sparsity via $L_1/L_q$
- Sparsity via Fused Lasso
- Sparsity via Trace Norm
- Sparse Inverse Covariance Estimation
- Implementations and the SLEP package
- **Trends in Sparse Learning**

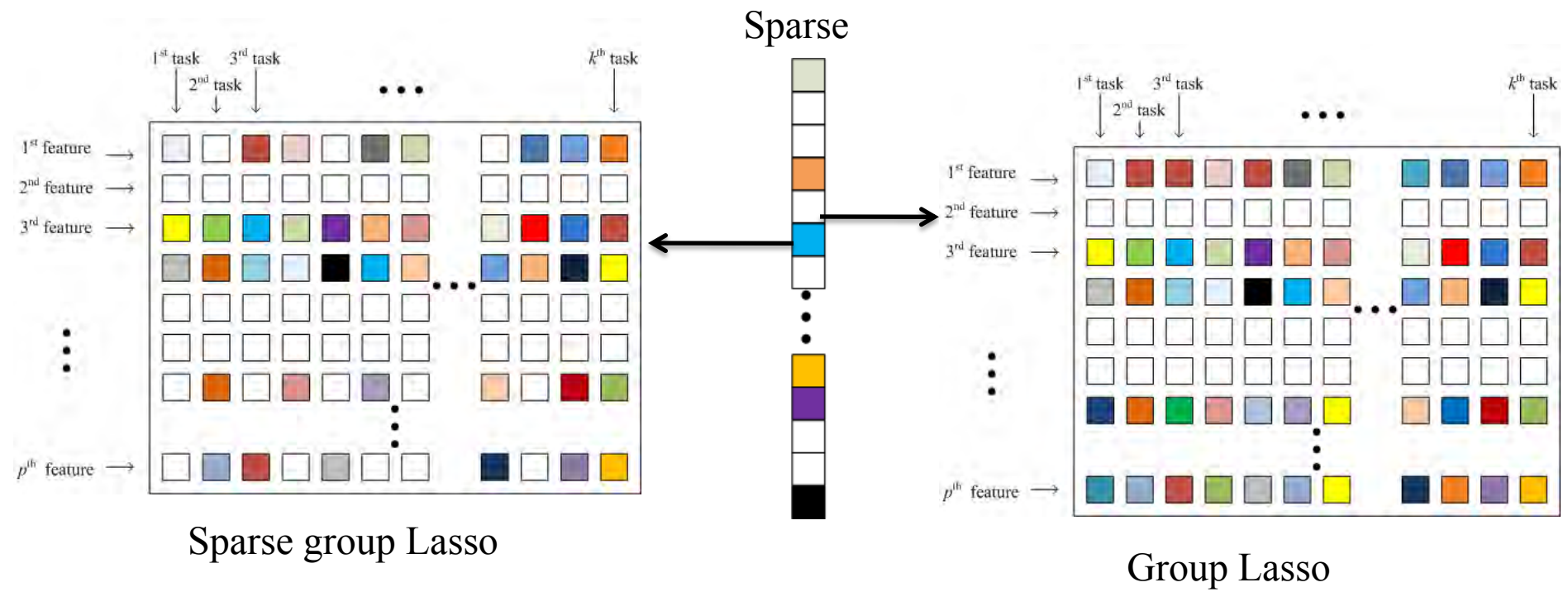
# New Sparsity Induced Penalties?

$$\min \; f(x) = \text{loss}(x) + \lambda \times \text{penalty}(x)$$



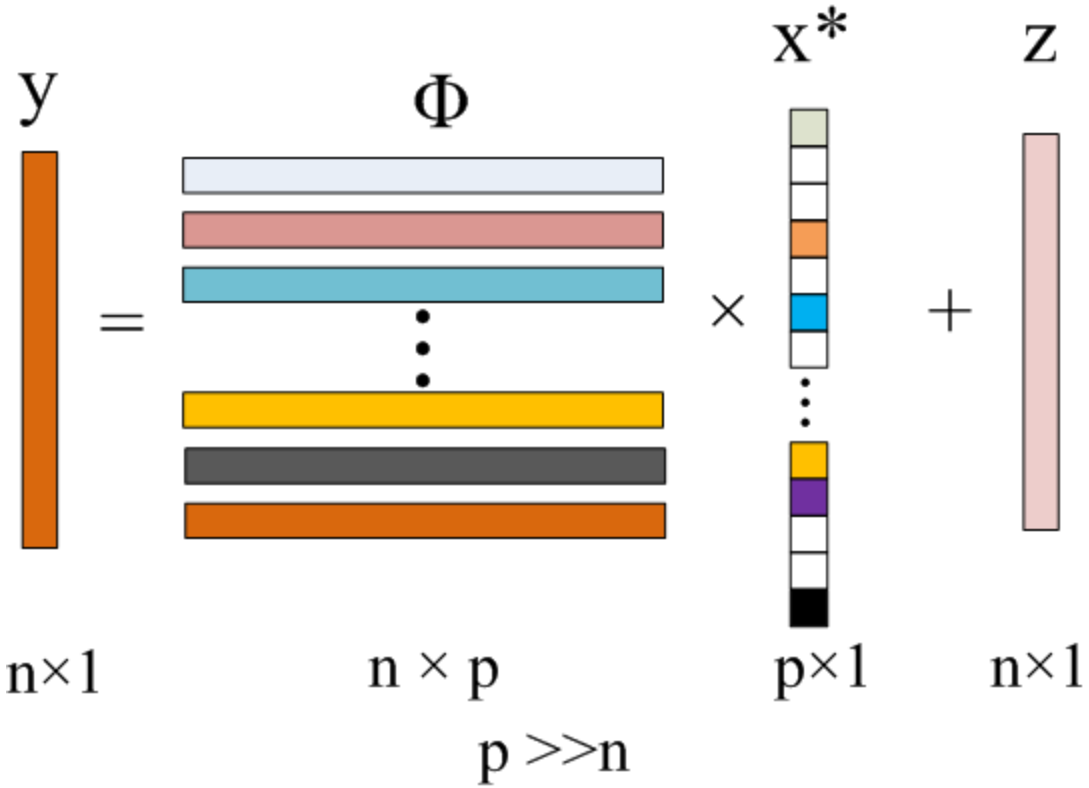
# Overlapping Groups?

$$\min \ f(x) = \text{loss}(x) + \lambda \times \text{penalty}(x)$$




- How to learn groups?
- How about the overlapping groups?

# Efficient Algorithms for Huge-Scale Problems




Algorithms for $p>10^8$, $n>10^5$?

It costs over 1 **Terabyte** to store the data.

# References

## (Compressive Sensing and Lasso)

Bajwa, W., Haupt, J., Sayeed, A., & Nowak, R. (2006). Compressive wireless sensing. *International Conference on Information Processing in Sensor Networks*.

Berg, E., Schmidt, M., Friedlander, M. P., & Murphy, K. (2008). *Group sparsity via linear-time projection*Tech. Rep. TR-2008-09). Department of Computer Science, University of British Columbia, Vancouver.

Bickel, P., Ritov, Y., , & Tsybakov, A. (2007). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*.

Candes, E., & Romberg, J. (2006). Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations of Computational Mathematics*, *6*, 227–254.

Candès, E., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, *52*, 489–509.

# References

## (Compressive Sensing and Lasso)

Candès, E., & Tao, T. (2004). Rejoinder: the dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, *35*, 2392–2404.

Candès, E., & Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, *51*, 4203–4215.

Candès, E., & Tao, T. (2006). Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, *52*, 5406–5425.

Candès, E., & Tao, T. (2007). The dantzig selector: statistical estimation when $p$ is much larger than $n$. *Annals of Statistics*, *35*, 2392–2404.

Candès, E., & Wakin, M. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine*, *25*, 21–30.

# References

## (Compressive Sensing and Lasso)

Chen, S., Donoho, D. L., & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, *43*, 129–159.

Daubechies, I., Defrise, M., & De Mol, C. (2005). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, *4*, 1168–1200.

Daubechies, I., Fornasier, M., & Loris, I. (2008). Accelerated projected gradient method for linear inverse problems with sparsity constraints. *Journal of Fourier Analysis and Applications*, *14*, 764–792.

Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, *52*, 1289–1306.

# References

## (Compressive Sensing and Lasso)

Duchi, J., Shalev-Shwartz, S., Singer, Y., & Tushar, C. (2008). Efficient projection onto the $\ell_1$-ball for learning in high dimensions. *International Conference on Machine Learning*.

Duchi, J., & Singer, Y. (2009). Boosting with structural sparsity. *International Conference on Machine Learning*.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*, 407–499.

Figueiredo, M., Nowak, R. D., & Wright, S. J. (2007). Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, *1*, 586–597.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, *1*, 302–332.

# References

## (Compressive Sensing and Lasso)

Hale, E., Yin, W., & Zhang, Y. (2007). *A fixed-point continuation method for $l_1$-regularized minimization with applications to compressed sensing* (Technical Report). CAAM TR07-07.

Kim, S.-J., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2007). An interior point method for large-scale l1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, *1*, 606–617.

Koh, K., Kim, S., & Boyd, S. (2007). An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, *8*, 1519–1555.

Liu, J., Chen, J., & Ye, J. (2009). Large-scale sparse logistic regression. *ACM SIGKDD International Conference On Knowledge Discovery and Data Mining*.

# References

## (Compressive Sensing and Lasso)

Liu, J., & Ye, J. (2009). Efficient euclidean projections in linear time. *International Conference on Machine Learning*.

Roth, V., & Fischer, B. (2008). The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. *International conference on Machine learning* (pp. 848–855).

Schmidt, M., Fung, G., & Rosales, R. (2007). Fast optimization methods for l1 regularization: A comparative study and two new approaches. *European Conference on Machine Learning* (pp. 286–297).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B, 58*, 267–288.

# References

## (Compressive Sensing and Lasso)

Yin, W., Osher, S., Goldfarb, D., & Darbon, J. (2008). Bregman iterative algorithms for $\ell_1$-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, *1*, 143–168.

Zhao, P., Rocha, G., & Yu., B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, *37*, 3468–3497.

Zhao, P., & Yu, B. (2004). *Boosted lasso* (Technical Report). Statistics Department, UC Berkeley.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zhu, J., Rosset, S., Hastie, T., & Tibshirani, R. (2003). 1-norm support vector machines. *Neural Information Processing Systems* (pp. 49–56).

# References

## (Group Lasso and Sparse Group Lasso)

Argyriou, A., Evgeniou, T., & Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, *73*, 243–272.

Bach, F. (2008). Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, *9*, 1179–1225.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, *2*, 183–202.

Duchi, J., & Singer, Y. (2009). Boosting with structural sparsity. *International Conference on Machine Learning*.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). *A note on the group lasso and a sparse group lasso* (Technical Report). Department of Statistics, Stanford University.

Höfling, H. (2009). A path algorithm for the fused lasso signal approximator. *arXiv*.

# References

## (Group Lasso and Sparse Group Lasso)

Liu, H., Palatucci, M., & Zhang, J. (2009a). Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. *International Conference on Machine Learning*.

Liu, H., & Zhang, J. (2009a). Estimation consistency of the group lasso and its applications. *International Conference on Artificial Intelligence and Statistics*.

Liu, H., & Zhang, J. (2009b). *On the $\ell_1$-$\ell_q$ regularized regression* (Technical Report). Department of Statistics, Carnegie Mellon University.

Liu, J., Ji, S., & Ye, J. (2009b). Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. *The 25th Conference on Uncertainty in Artificial Intelligence*.

# References

## (Group Lasso and Sparse Group Lasso)

Liu, J., & Ye, J. (2010b). Efficient $\ell_1/\ell_q$-norm regularization. *preprint*.

Liu, J., & Ye, J. (2010c). Efficient sparse group lasso with between- and within-group sparsity. *preprint*.

Meier, L., Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, *70*, 53–71.

Negahban, S., Ravikumar, P., Wainwright, M., & Yu, B. (2009). A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. In *Advances in neural information processing systems*, 1348–1356.

# References

## (Group Lasso and Sparse Group Lasso)

Negahban, S., & Wainwright, M. (2008). Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization. In *Advances in neural information processing systems*, 1161–1168.

Nemirovski, A. (1994). *Efficient methods in convex programming*. Lecture Notes.

Nesterov, Y. (2004). *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. *CORE Discussion Paper*.

Obozinski, G., Taskar, B., & Jordan, M. I. (2007). *Joint covariate selection for grouped classification* (Technical Report). Statistics Department, UC Berkeley.

Obozinski, G., Wainwright, M., & Jordan, M. (2008). High-dimensional support union recovery in multivariate regression. In *Advances in neural information processing systems*, 1217–1224.

# References

## (Group Lasso and Sparse Group Lasso)

Peng, J., Zhu, J., A., B., Han, W., Noh, D.-Y., Pollack, J. R., & Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics*, to appear.

Quattoni, A., Carreras, X., Collins, M., & Darrell, T. (2009). An efficient projection for $\ell_{1,\infty}$,infinity regularization. *International Conference on Machine Learning*.

Tseng, P. (2001). Convergence of block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, *109*, 474–494.

Tseng, P., & Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, *117*, 387–423.

# References

## (Group Lasso and Sparse Group Lasso)

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal Of The Royal Statistical Society Series B*, *68*, 49–67.

Zhao, P., Rocha, G., & Yu., B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, *37*, 3468–3497.

Zhao, P., & Yu, B. (2006). *Boosted lasso* (Technical Report). Statistics Department, UC Berkeley.

# References

## (Fused Lasso)

Ahmed, A., & Xing, E. P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, *106*, 11878–11883.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, *1*, 302–332.

Liu, J., Yuan, L., & Ye, J. (2010). An efficient algorithm for a class of fused lasso problems. *preprint*.

Rinaldo, A. (2009). Properties and refinements of the fused lasso. *Annals of Statistics*, *37*, 2922–2952.

# References

## (Fused Lasso)

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal Of The Royal Statistical Society Series B*, *67*, 91–108.

Tibshirani, R., & Wang, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, *9*, 18–29.

# References

## (Trace Norm)

Cai, J., Candès, E. J., & Shen, Z. (2008). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, to appear.

Candès, E. J., & Tao, T. (2009). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, to appear.

Goldfarb, D., & Ma, S. (2009). Convergence of fixed point continuation algorithms for matrix rank minimization. *arXiv:0906.3499v2*.

Ji, S., & Ye, J. (2009). An accelerated gradient method for trace norm minimization. *International Conference on Machine Learning*.

# References

## (Trace Norm)

Jin, R., Wang, S., & Zhou, Y. (2009). Regularized distance metric learning:theory and algorithm. In *Neural information processing systems*.

Liu, Z., & Vandenberghe, L. (2009). Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, *31*, 1235–1256.

Lu, Z., Monteiro, R. D. C., & Yuan, M. (2009). Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *arXiv:0904.0691v1*.

Ma, S., Goldfarb, D., & Chen, L. (2009). Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming Series A*, to appear.

# References

## (Trace Norm)

Meka, R., Jain, P., & Dhillon, I. S. (2009). Guaranteed rank minimization via singular value projection. *arXiv:0909.5457v3*.

Pong, T., Tseng, P., Ji, S., & Ye, J. (2009). Trace norm regularization: Reformulations, algorithms, and multi-task learning. *submitted to SIAM Journal on Optimization*.

Toh, K.-C., & Yun, S. (2009). An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. Preprint, Department of Mathematics, National University of Singapore, March 2009.

Ying, Y., Huang, K., & Campbell, C. (2009). Sparse metric learning via smooth optimization. In *Neural information processing systems*.

# References

## (Sparse Inverse Covariance)

Banerjee, O., Ghaoui, L., & D'Aspremont, A. (2007). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research.*

Friedman, J., Hastie, T., & Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *report.*

Honorio, J., Ortiz, L., Samaras, D., Paragios, N., & Goldstein, R. Sparse and locally constant gaussian graphical models. In *Advances in neural information processing systems 22.*

Huang, S., Li, J., Sun, L., Liu, J., Wu, T., Chen, K., Fleisher, A., Reiman, E., & Ye, J. Learning brain connectivity of alzheimer's disease from neuroimaging data. In *Advances in neural information processing systems 22.*

# References

## (Sparse Inverse Covariance)

Sun, L., Patel, R., Liu, J., Chen, K., Wu, T., Li, J., Reiman, E., & Ye, J. (2009). Mining brain region connectivity for alzheimers disease study via sparse inverse covariance estimation. In *Acm sigkdd international conference on knowledge discovery and data mining*.