Mining Sparse Representations: Formulations, Algorithms, and Applications

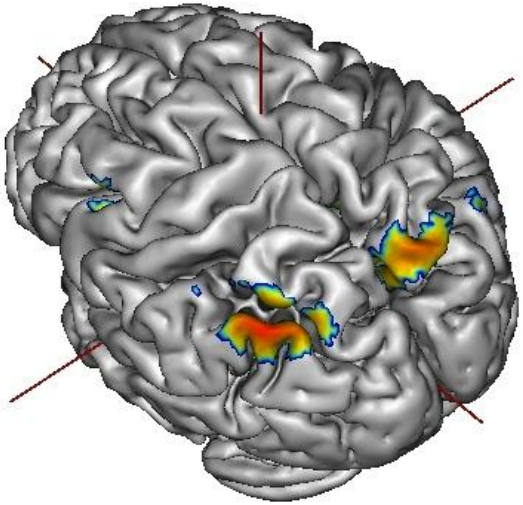
Jun Liu, Shuiwang Ji, and Jieping Ye

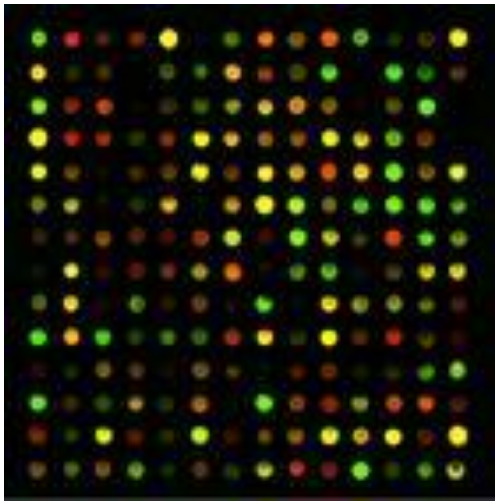
Computer Science and Engineering

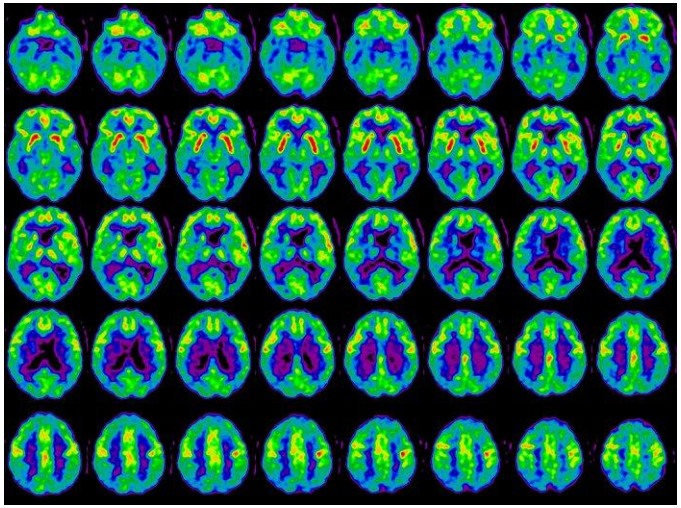
The Biodesign Institute

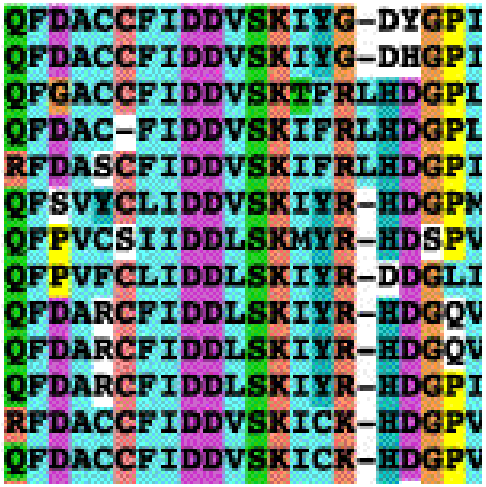
Arizona State University

Mining High-Dimensional Data



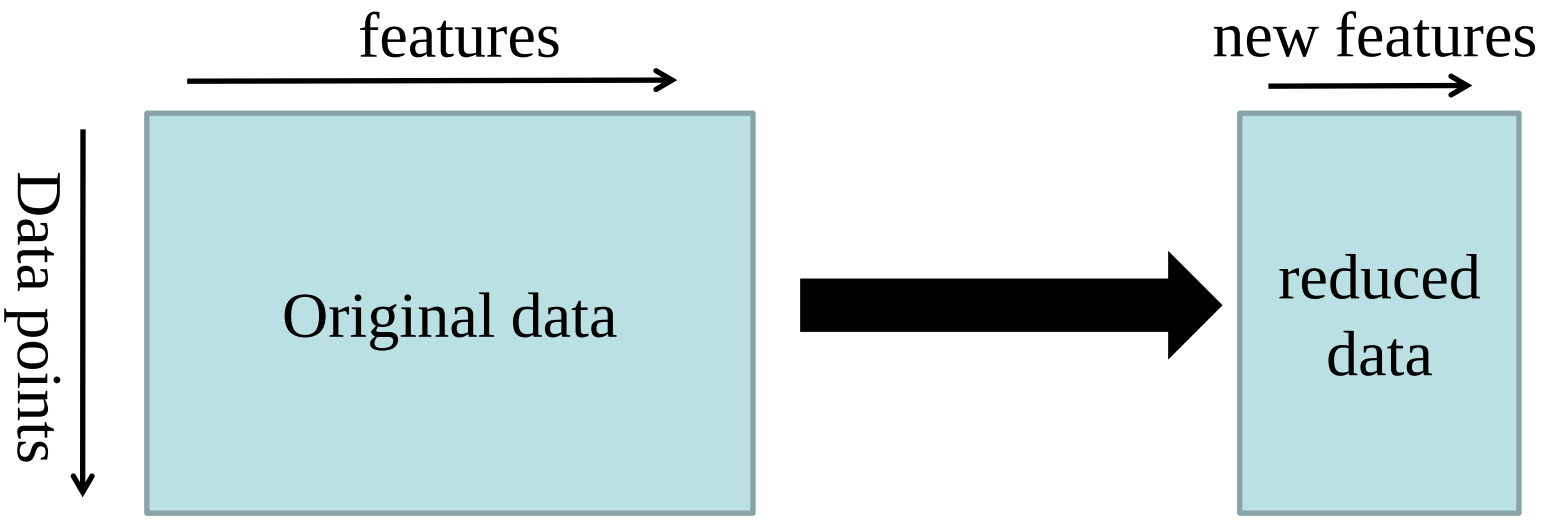






Dimensionality Reduction

- Dimensionality reduction algorithms
  - Feature extraction
  - Feature selection



SIAM Data Mining 2007 Tutorial (Yu, Ye, and Liu):
"Dimensionality Reduction for Data Mining - Techniques, Applications, and Trends"

Dimensionality Reduction

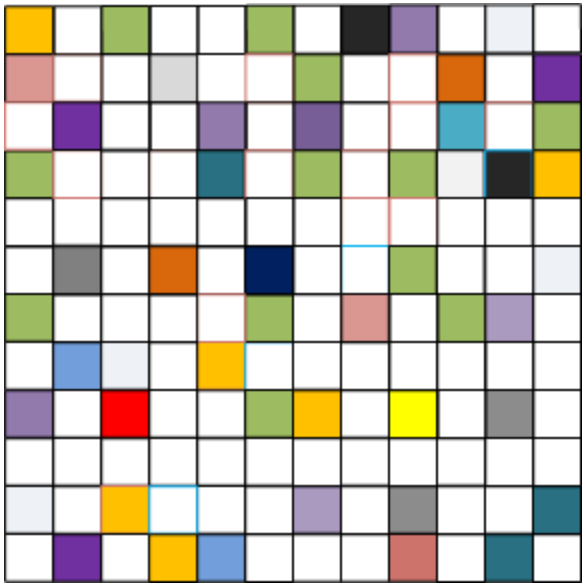
- Dimensionality reduction algorithms
  - Feature extraction
  - Feature selection

- We focus on sparse learning in this tutorial
  - Embed dimensionality reduction into data mining tasks
  - Flexible models for complex feature structures
  - Strong theoretical guarantee
  - Empirical success in many applications
  - Recent progress on efficient implementations

SIAM Data Mining 2007 Tutorial (Yu, Ye, and Liu):
"Dimensionality Reduction for Data Mining - Techniques, Applications, and Trends"

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.



Motivation: Signal Acquisition (1)

- Wish to acquire a digital object $x \in \mathbb{R}^p$ from n measurements:

$$y_i = \langle x, \varphi_i \rangle, i = 1, 2, \ldots, n$$

- Waveforms φ_i
  - Dirac delta functions (spikes)
    - y is a vector of sampled values of x in the time or space domain
  - Indicator functions of pixels
    - y is the image data typically collected by sensors in a digital camera
  - Sinusoids
    - y is a vector of Fourier coefficients (e.g., MRI)

Motivation: Signal Acquisition (1)

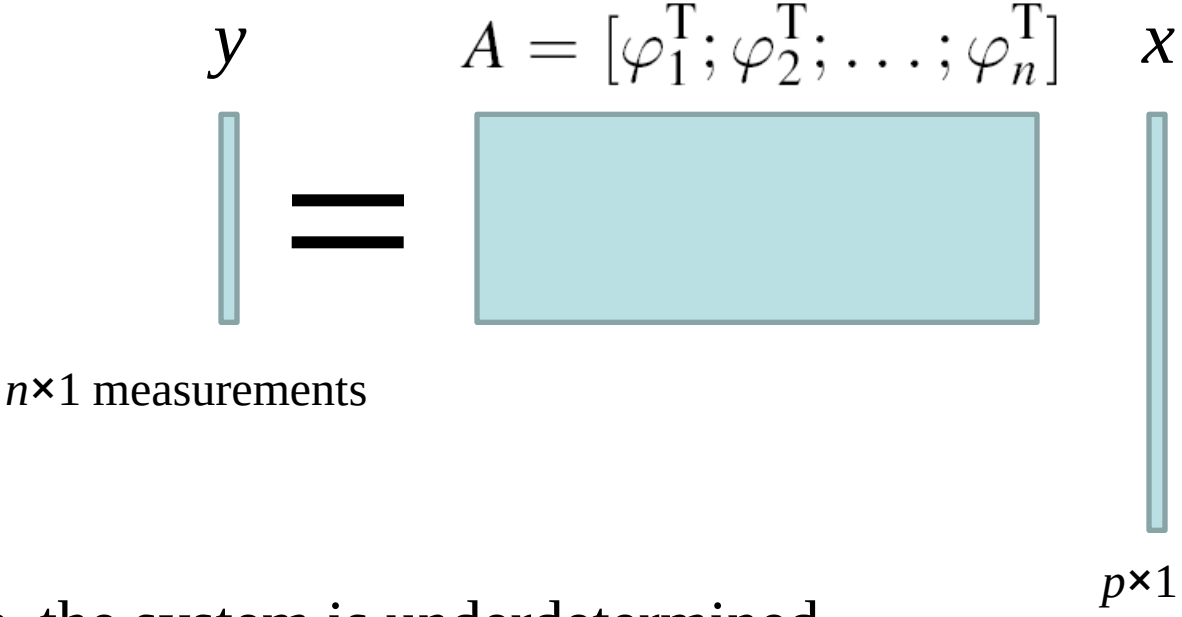
- Wish to acquire a digital object $x \in \mathbb{R}^p$ from n measurements:

$$y_i = \langle x, \varphi_i \rangle, i = 1, 2, \ldots, n$$

- Is accurate reconstruction possible from $n << p$ measurements only?
  - Few sensors
  - Measurements are very expensive
  - Sensing process is slow

Motivation: Signal Acquisition (2)

- Conventional wisdom: reconstruction is impossible
  - Number of measurements must match the number of unknowns

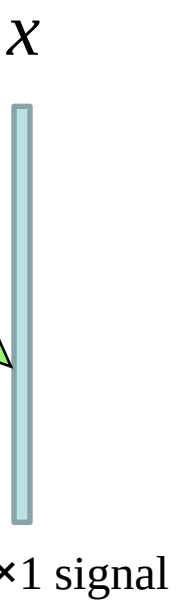


If $n << p$, the system is underdetermined.

Motivation: Signal Acquisition (2)

- Conventional wisdom: reconstruction is impossible
  - Number of measurements must match the number of unknowns

- If x is known to be sparse, i.e., most entries are zero, then with a large probability we can recover x exactly by solving a linear programming.

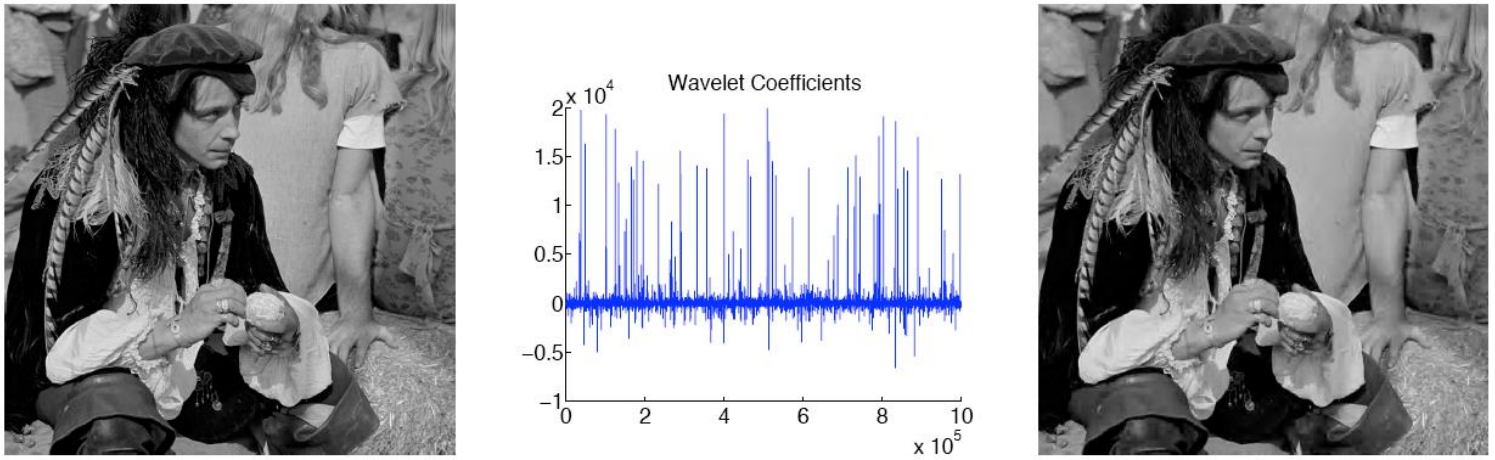


If $n<<p$, the system is underdetermined.

Motivation: Signal Acquisition (3)

- Many natural signals are sparse or compressible in the sense that they have concise representations when expressed in the proper basis

Megapixel image represented as 2.5% largest wavelet coefficients



(*Candes and Wakin, 2008*)

Sparsity

- Dominant modeling tool
  - Genomics
  - Genetics
  - Signal and audio processing
  - Image processing
  - Neuroscience (theory of sparse coding)
  - Machine learning
  - Data mining
  - …

Sparse Learning Models

- 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:

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

Sparse Learning Models

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

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

- Least squares
- Logistic loss
- Hinge loss
- …

- Zero norm is the natural choice
  - The number of nonzero elements of x
  - Not a valid norm, nonconvex, NP-hard

The L_1 Norm Penalty

In this tutorial, we focus on sparse learning based on L_1 and its extensions:

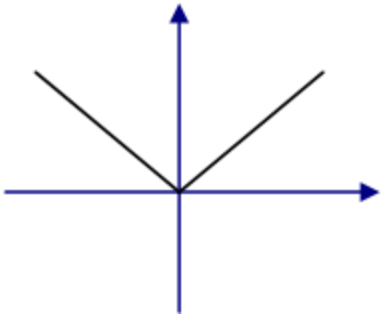
- penalty(x)=$\|x\|_1=\sum_i|x_i|$
  - Valid norm
  - Convex
  - Computationally tractable
  - Sparsity induced norm
  - Theoretical properties
  - Various Extensions

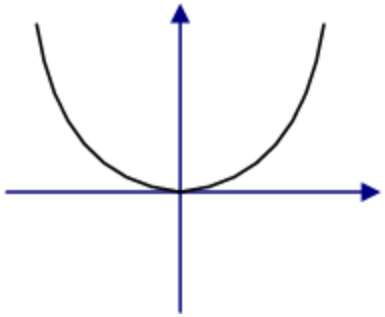
$$\min \ \text{loss}(x) + \lambda \|x\|_0$$

$$\min \ \text{loss}(x) + \lambda \|x\|_1$$

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 $v \geq \lambda$, $x=v-\lambda$
If $v \leq -\lambda$, $x=v+\lambda$
Else, $x=0$

$x=v/(1+2\lambda)$

Nondifferentiable at 0

Differentiable at 0

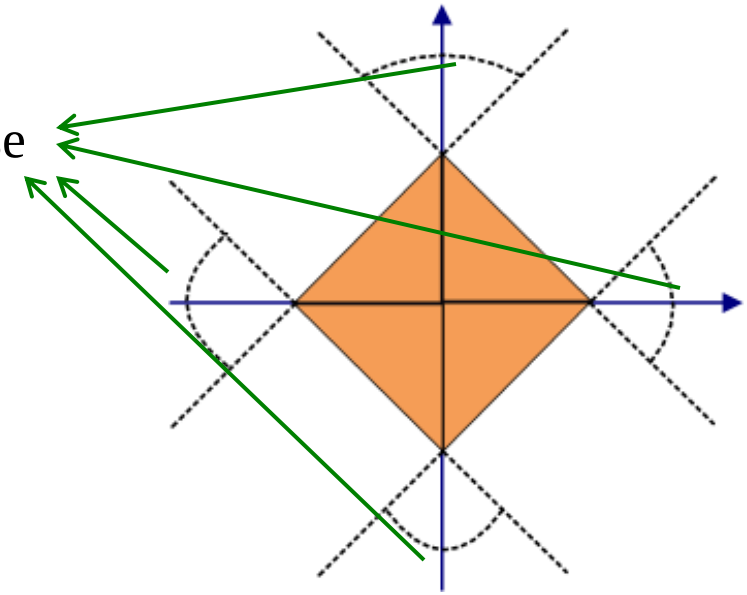
Why does L_1 Induce Sparsity?

Understanding from the projection

min loss(x)
s.t. $\|x\|_1 \leq 1$

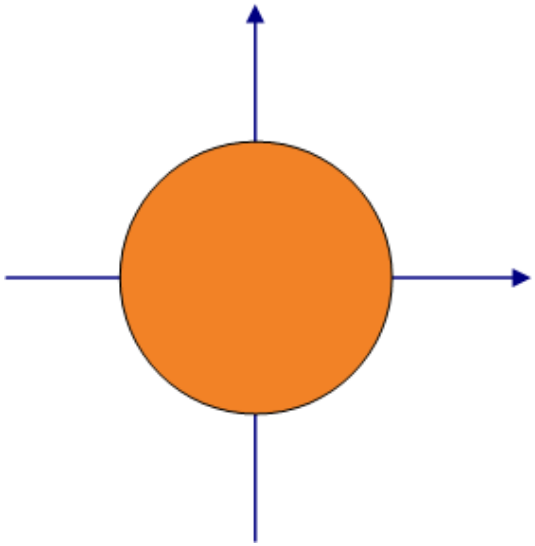
min $0.5\|x-v\|^2$
s.t. $\|x\|_1 \leq 1$

Sparse



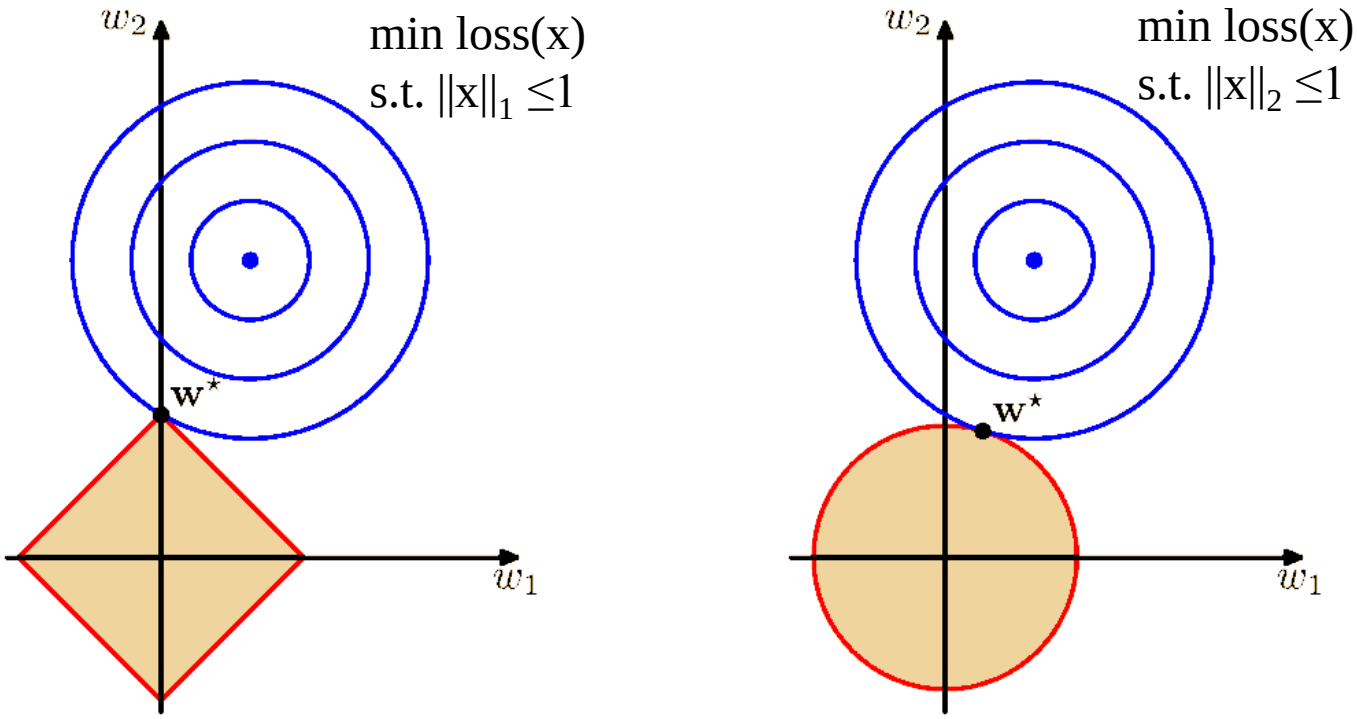
min loss(x)
s.t. $\|x\|_2 \leq 1$

min $0.5\|x-v\|^2$
s.t. $\|x\|_2 \leq 1$



Why does L_1 Induce Sparsity?

Understanding from constrained optimization



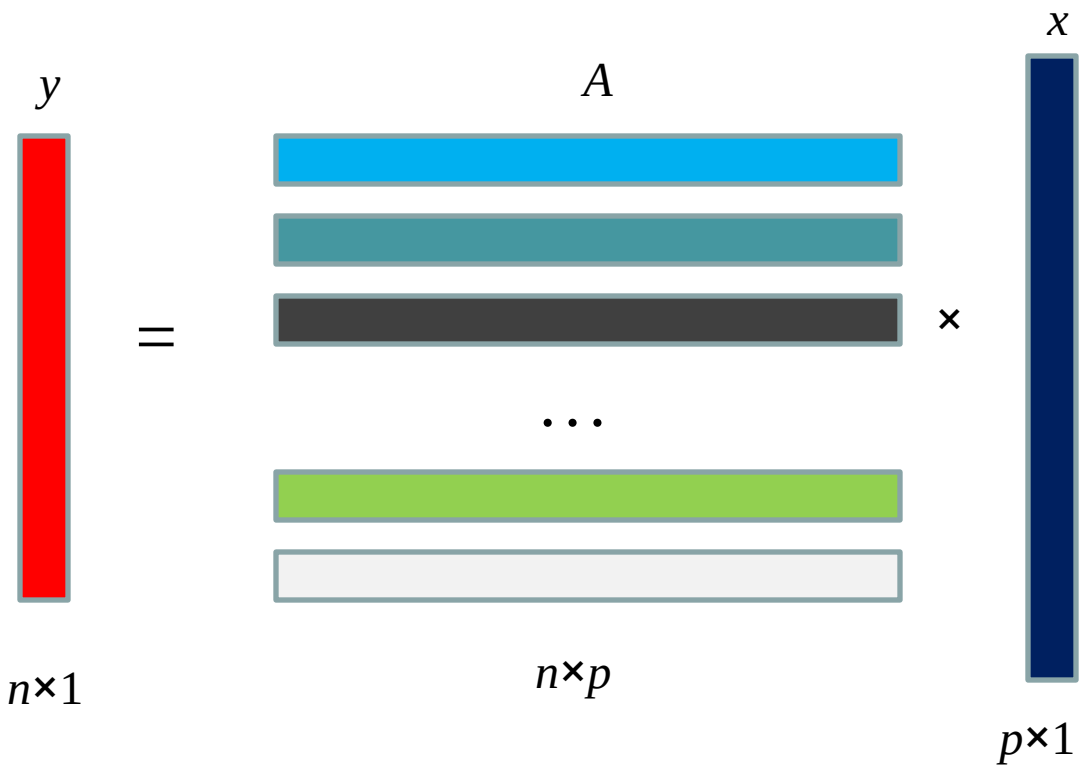
(Bishop, 2006, Hastie et al., 2009)

Outline

- Sparse Learning Models
  - 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)



- *x* is sparse
- *p*>>*n*
- *A is a measurement matrix satisfying certain conditions*

NP hard

$$P_0 \quad \min \ \|x\|_0 \quad \text{s.t.} \ \ Ax = y$$

$$P_1 \quad \min \ \|x\|_1 \quad \text{s.t.} \ \ Ax = y$$

Sparse Recovery

$$P_0 \quad \min \ \|x\|_0 \quad \text{s.t.} \ Ax = y$$

$$P_1 \quad \min \ \|x\|_1 \quad \text{s.t.} \ Ax = y$$

The solution to P_1 is the unique optimal solution to P_0 if $\delta_{2K} < \sqrt{2} - 1$. (Recent improvement : $\delta_K < 0.307$)

The measurement matrix A satisfies the K-restricted isometry property with constant δ_K if δ_K is the smallest number satisfying

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

for every K-sparse vector x.

Extensions to the Noisy Case

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

$$y = Ax + z$$

noise

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

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

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

Lasso (Tibshirani, 1996)

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

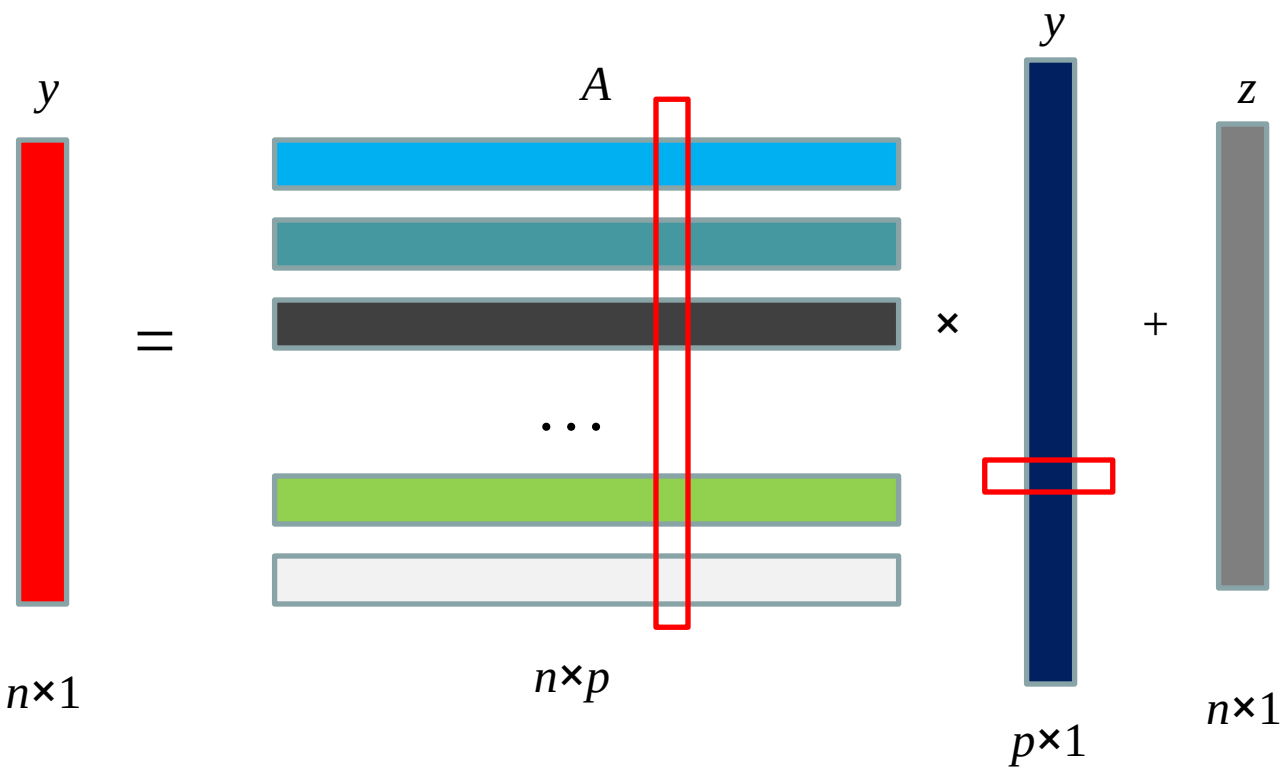
Regularized counterpart of Lasso

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

Dantzig selector (Candes and Tao, 2007)

Lasso

(Tibshirani, 1996)



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

Simultaneous feature selection and regression

Lasso Theory

(Bickel, Ritov, and Tsybakov, 2009)

Restricted eigenvalue conditions

Assumption $\mathrm{RE}(s, c_0)$:

$$\kappa(S, c_0) \triangleq \min_{J_0 \subseteq \{1,2,\ldots,p\}: |J_0| \leq s} \quad \min_{\triangle \neq 0: \|\triangle_{J_0^C}\|_1 \leq c_0 \|\triangle_{J_0}\|_1} \frac{\|A\triangle\|_2}{\sqrt{n}\|\triangle_{J_0}\|_2} > 0$$

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 $A^{\mathrm{T}}A/n$ be equal to 1, and $M(x^*) = s$ where $1 \leq s \leq p$. Let Assumption $RE(s, 3)$ be satisfied. Let

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

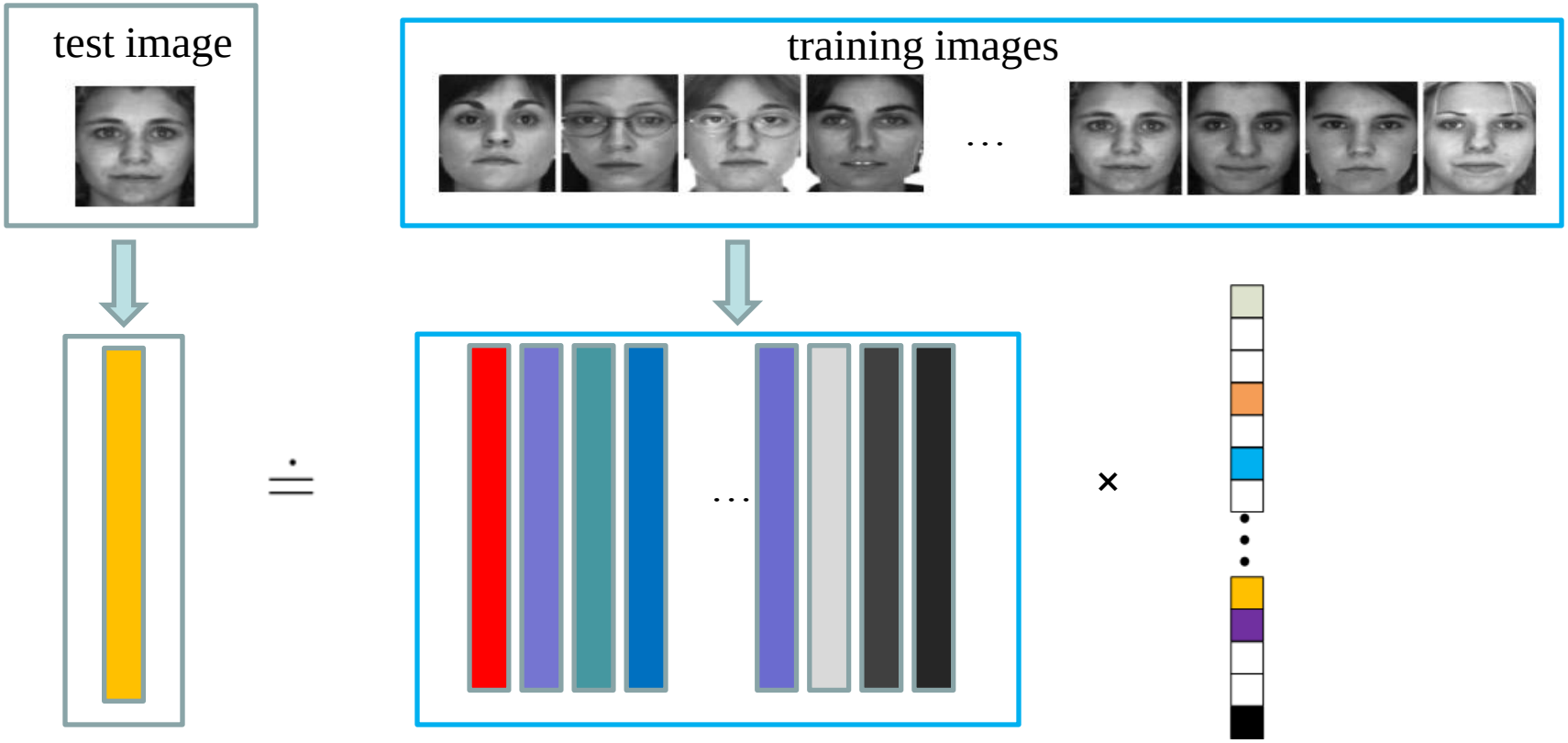
$$\lambda = A\delta\sqrt{\frac{\log p}{n}}$$

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

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

Application: Face Recognition

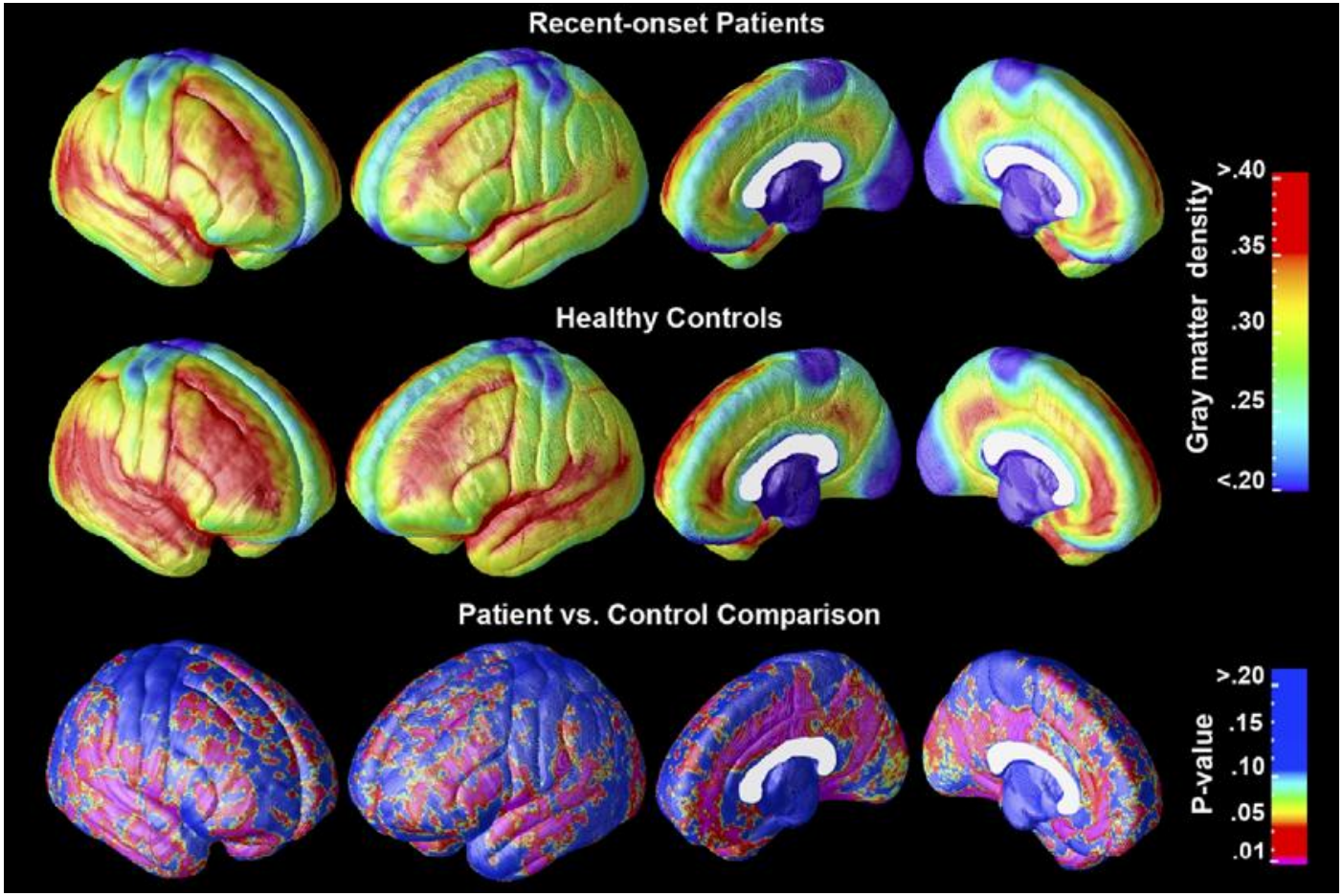
(Wright et al. 2009)



Use the computed sparse coefficients for classification

Application: Biomedical Informatics

(Sun et al. 2009)



Elucidate a Magnetic Resonance Imaging-Based Neuroanatomic Biomarker for Psychosis

Application: Bioinformatics

- T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel and K. Lange. Genome-wide association analysis by Lasso penalized logistic regression. ***Bioinformatics,*** 2009.

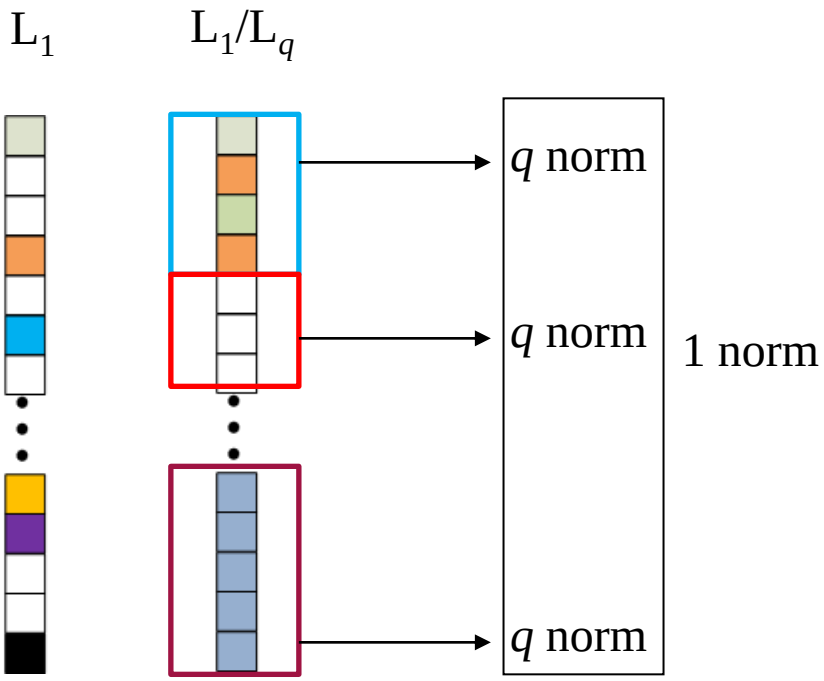
- W. Shi, K.E. Lee, and G. Wahba. Detecting disease causing genes by LASSO-Pattern search algorithm. ***BMC Proceedings***, 2007.

- S.K. Shevade and S.S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. ***Bioinformatics,*** 2003.

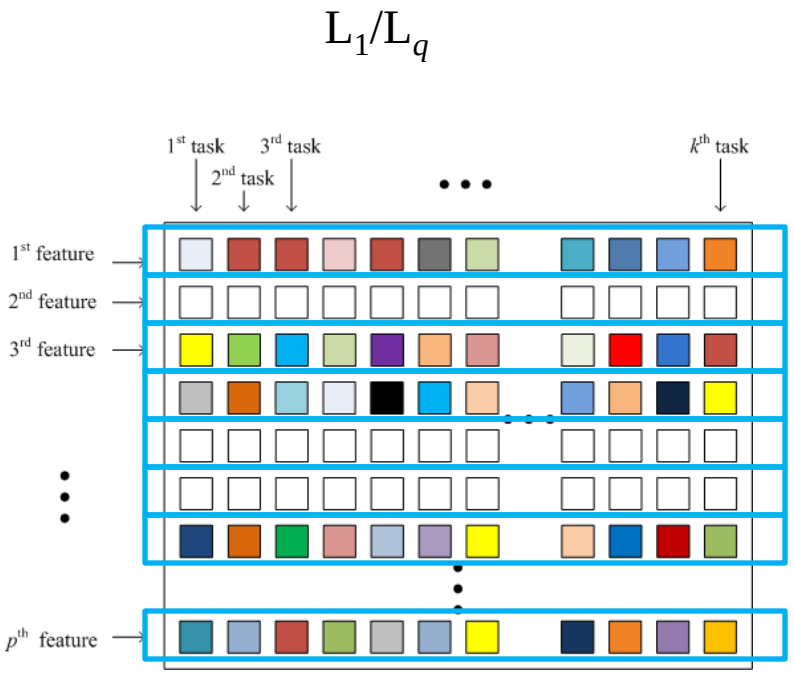
Outline

- Sparse Learning Models
  - 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)?



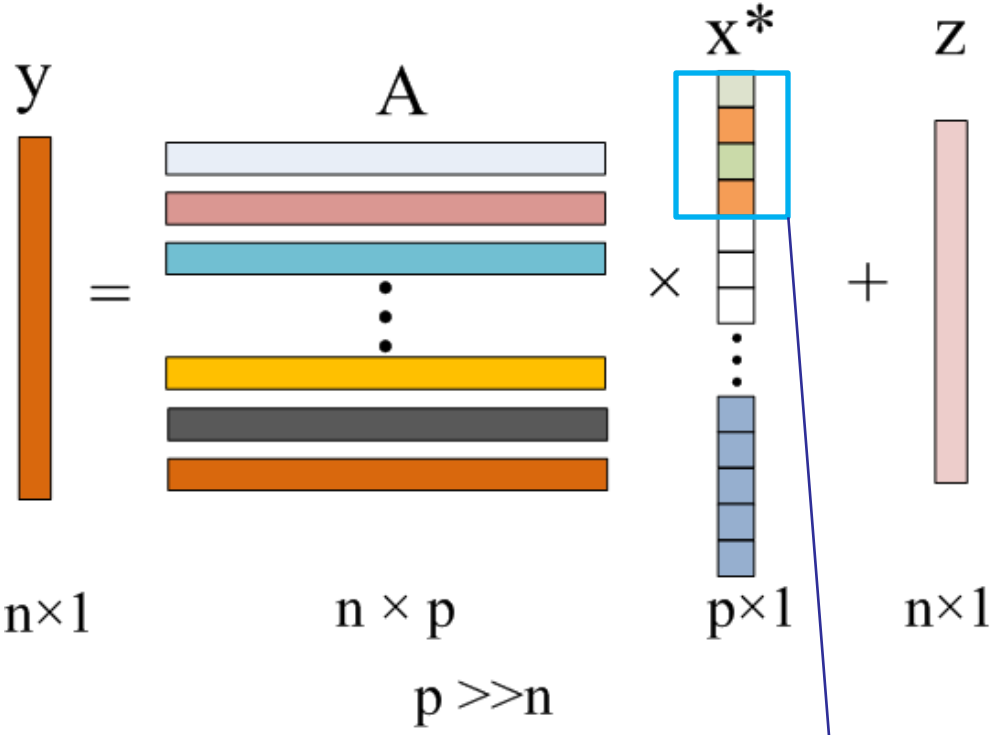
$$\|X\|_{q,1} = \sum_i \|X_{G_i}\|_q$$



Most existing work focus on q=2, ∞

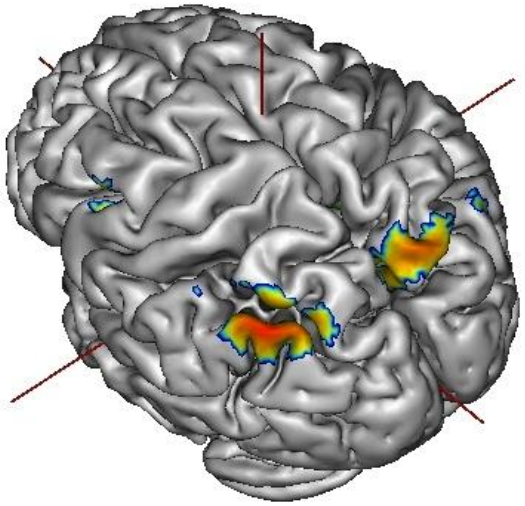
Group Lasso

(Yuan and Lin, 2006)

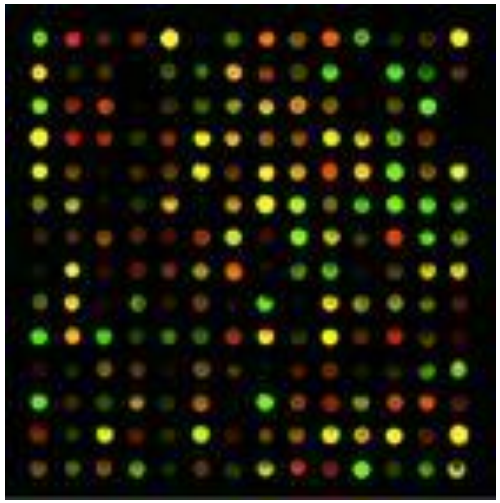


$$\min \frac{1}{2}\|Ax - y\|_2^2 + \lambda \sum_{i=1}^{g} d_i \|x_{G_i}\|_2$$

Group Feature Selection



brain region



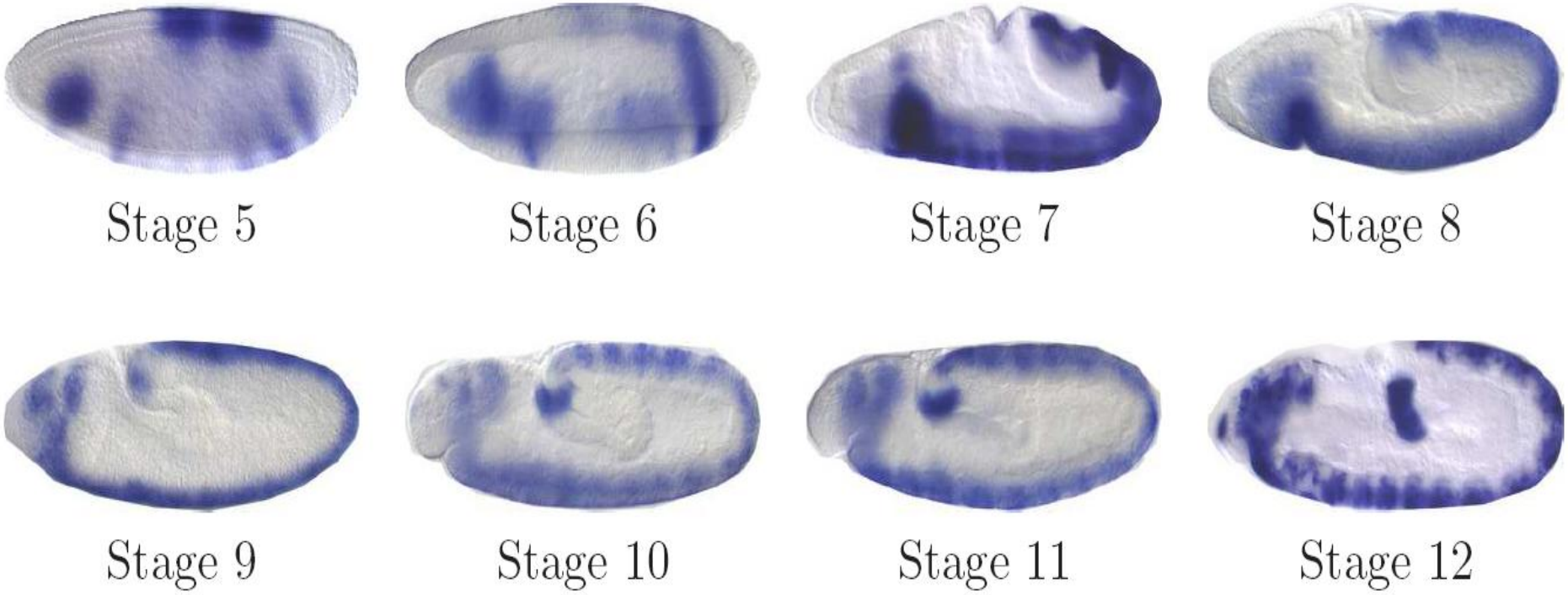
functional group

group

A	**1**	**0**	**0**	**0**
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

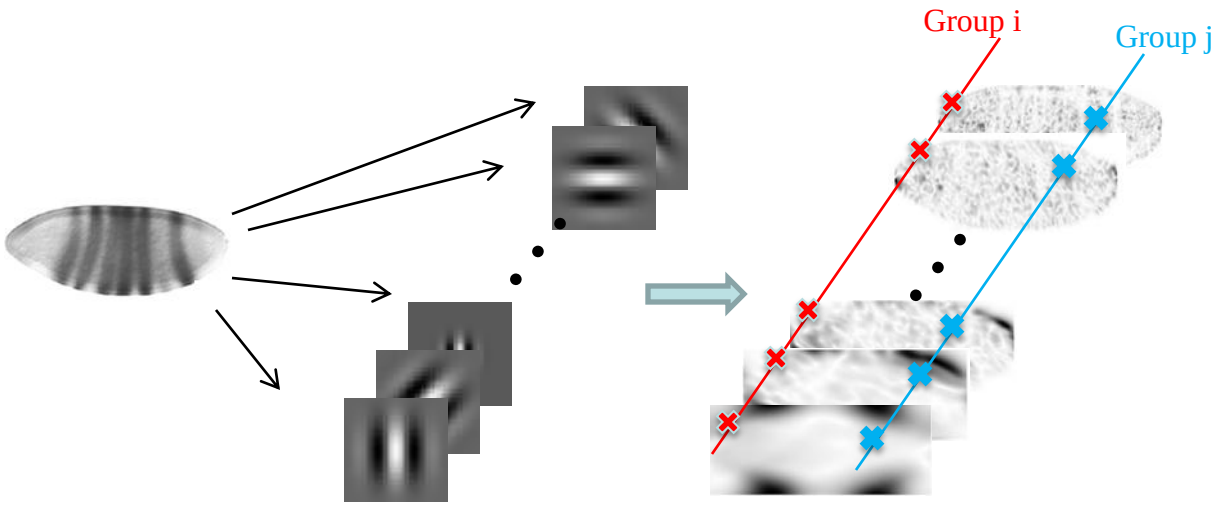
categorical variable

Developmental Stage Annotation (1)



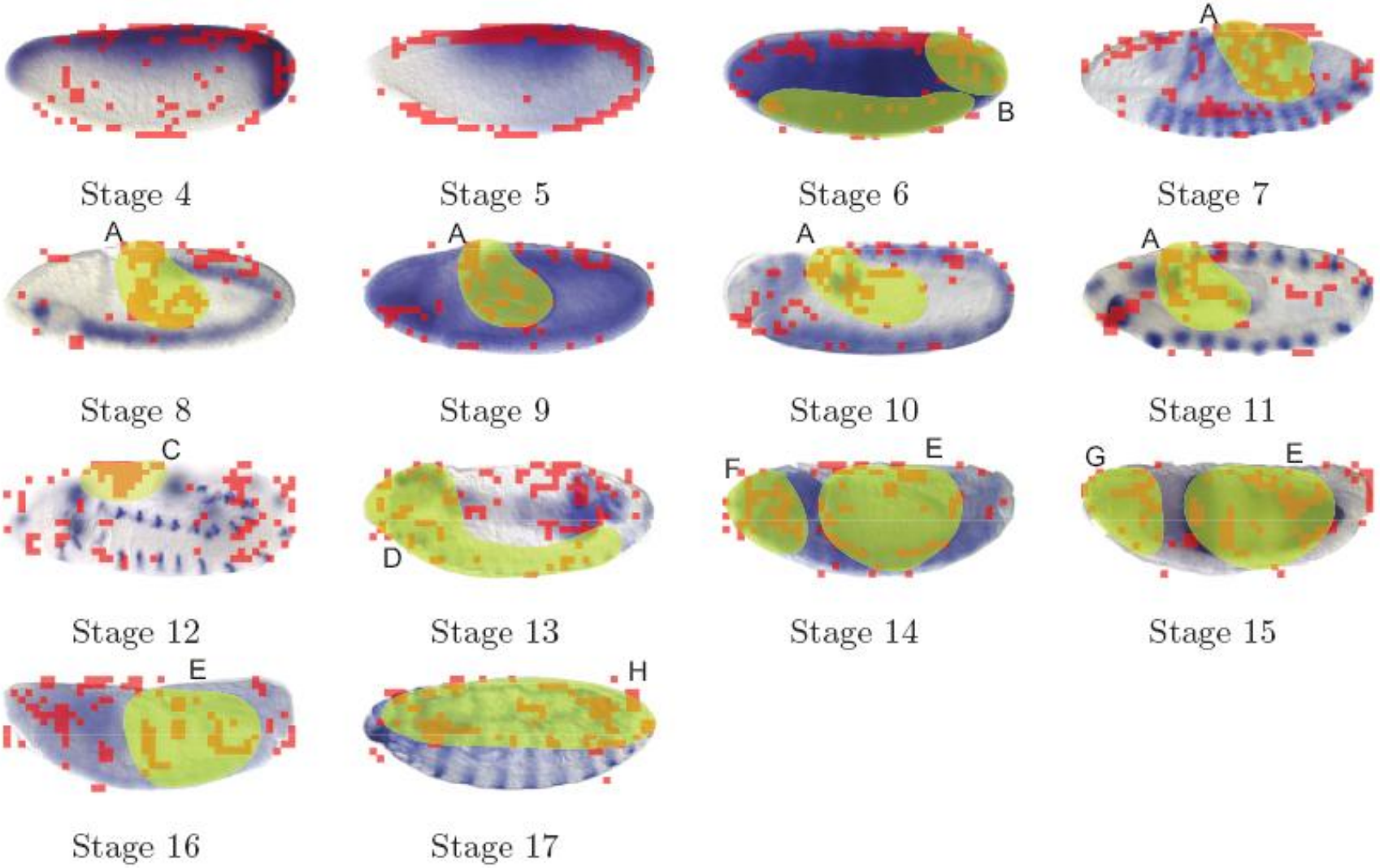
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
25	40	15	50	40	10	10	30	40	60	120	120	60	60	100	360	720

Developmental Stage Annotation (2)

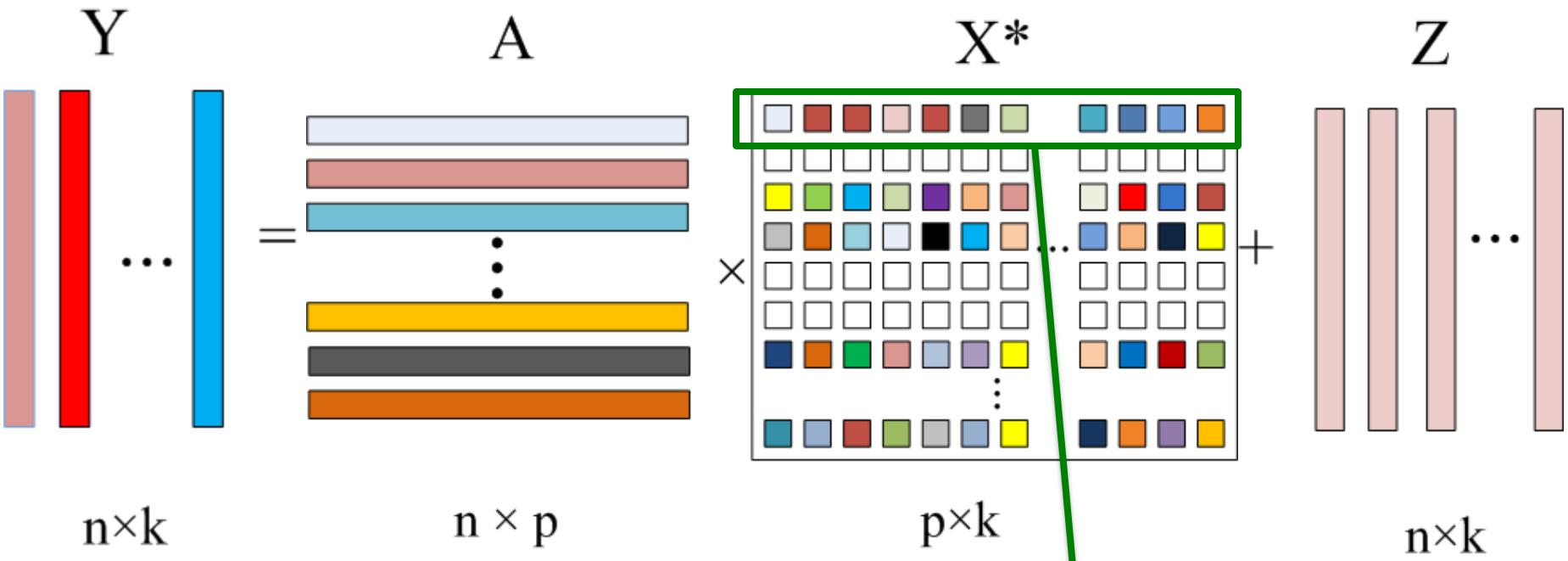


A group of 24 features is associated with a single region of the image. → Group selection

Developmental Stage Annotation (3)



Multi-Task/Class Learning via L_1/L_q



$$\min_X \frac{1}{2}\|AX - Y\|_F^2 + \lambda \sum_{i=1}^{p} \|\mathbf{x}_i\|_q$$

Writer-specific Character Recognition

(Obozinski, Taskar, and Jordan, 2006)

Letter data set:

1) The letters are from more than 180 different writers

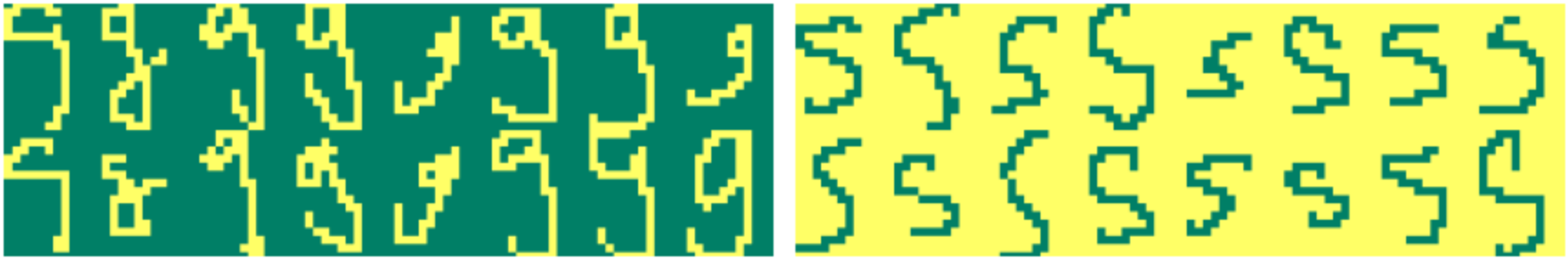
2) It has 8 tasks for discriminating letter c/e, g/y, g/s, m/n, a/g, i,/j, a/o. f/t, and h/n

The letter '*a' written by 40 different people*

Writer-specific Character Recognition

(Obozinski, Taskar, and Jordan, 2006)

Samples of the letters *s and g for one writer*

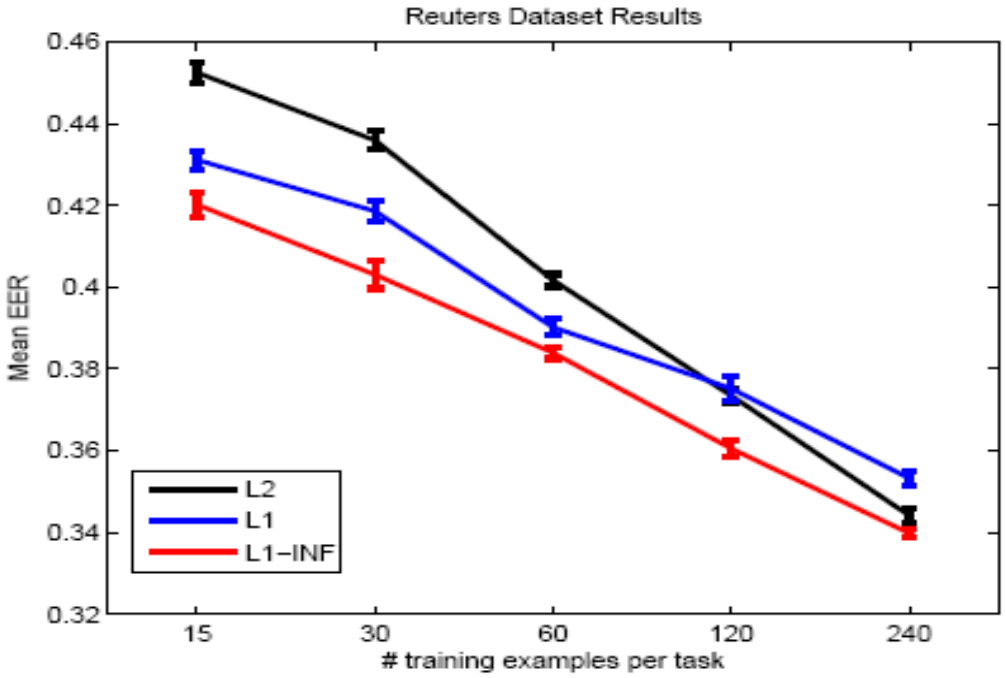


	pixels: error (%)			
Task	ℓ_1/ℓ_2	ℓ_1/ℓ_1	id.ℓ_1	pool
c/e	**4.0**	8.5	9.0	4.5
g/y	**11.4**	16.1	17.2	18.6
g/s	**4.4**	10.0	10.3	6.9
m/n	**2.5**	6.3	6.9	4.1
a/g	**1.3**	3.6	4.1	3.6
i/j	12.0	14.0	14.0	**11.3**
a/o	**2.8**	4.8	5.2	4.2
f/t	**5.0**	6.7	6.1	8.2
h/n	**3.2**	14.3	18.6	5.0

Visual Category Recognition

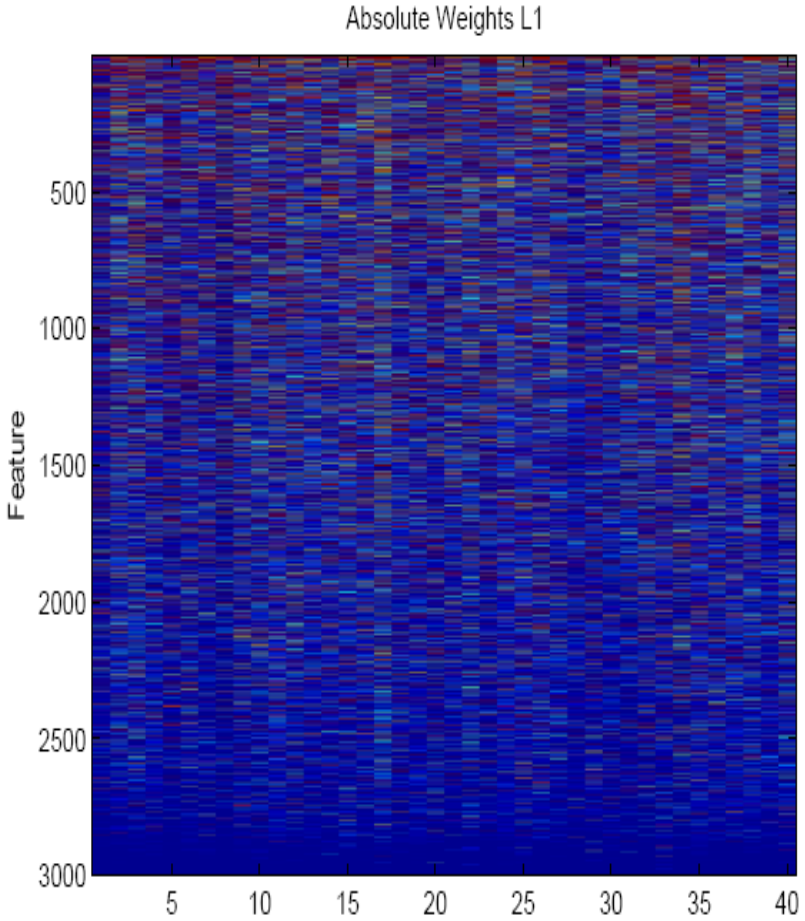
(Quattoni et al., 2009)

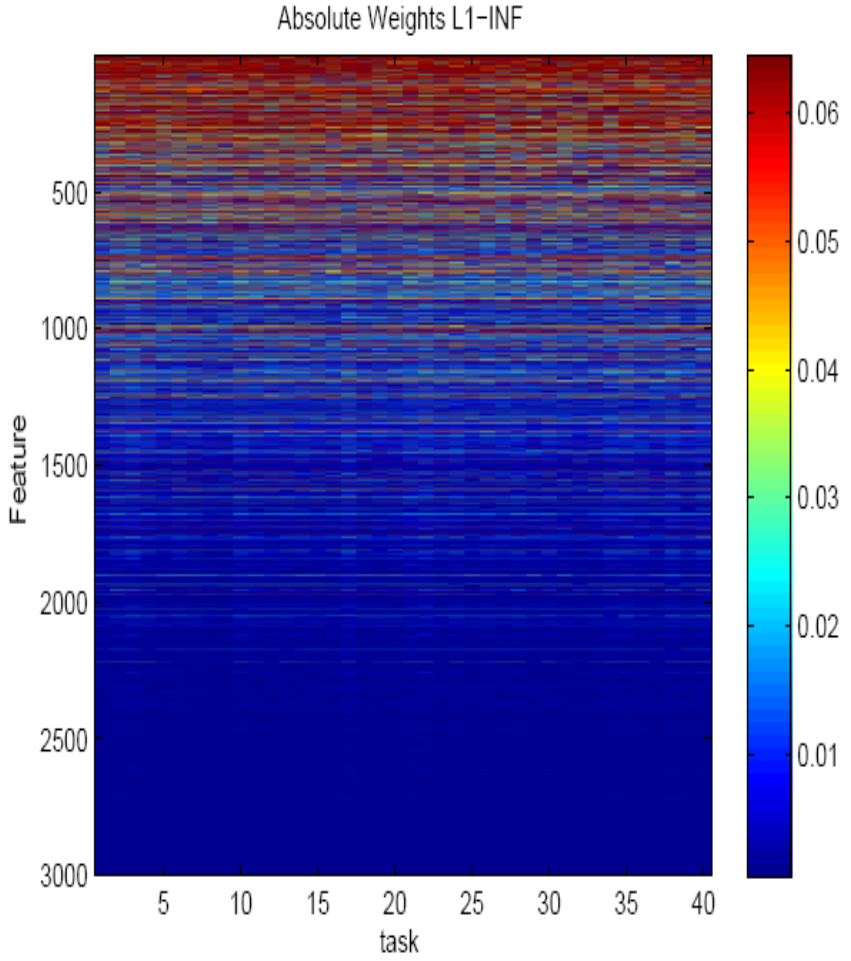
- Images on the Reuters website have associated story or topic labels, which correspond to different stories in the news.
  - An image can belong to one or more stories.
  - Binary prediction of whether an image belonged to one of the 40 most frequent stories.



Visual Category Recognition

(Quattoni et al., 2009)





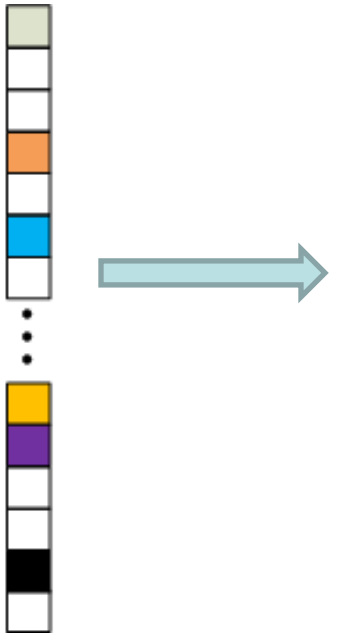
Outline

- Sparse Learning Models
  - 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

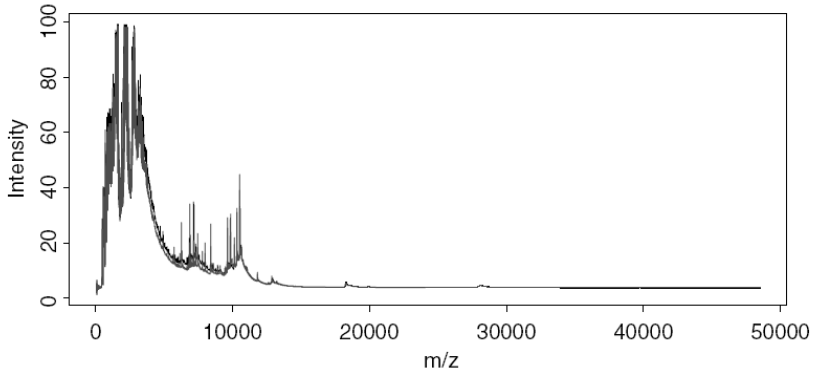
(Tibshirani et al., 2005; Tibshirani and Wang, 2008; Friedman et al., 2007)

L_1 Fused Lasso

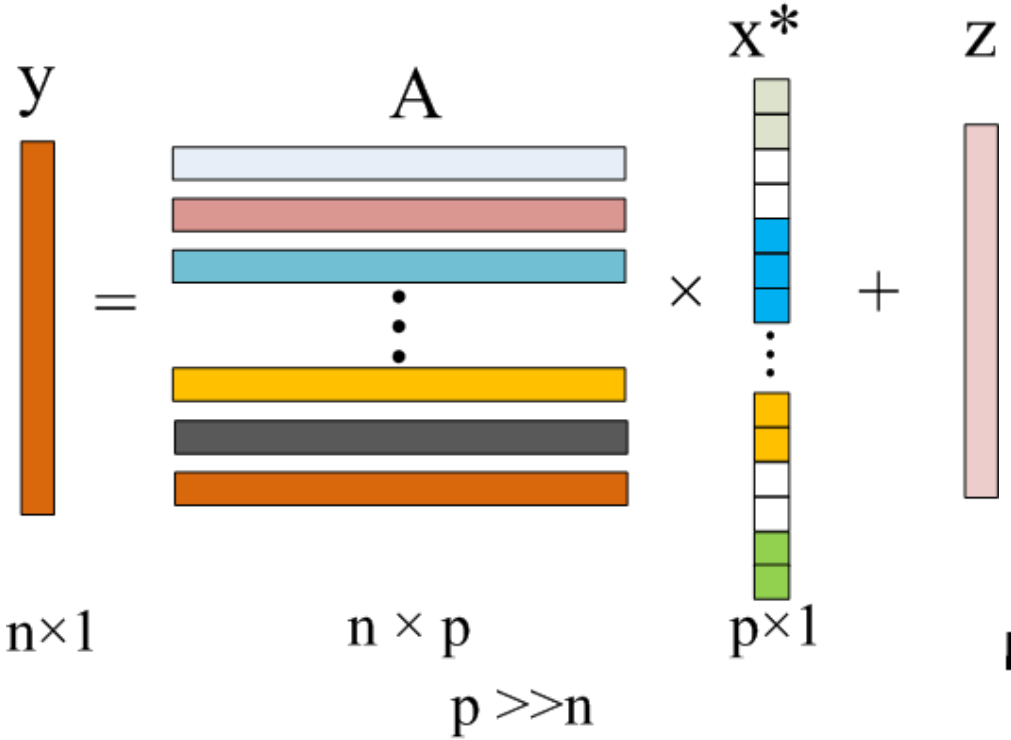


$$\text{lasso}(x) = \lambda \sum_{i=1}^{p} |x_i|$$

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



Fused Lasso



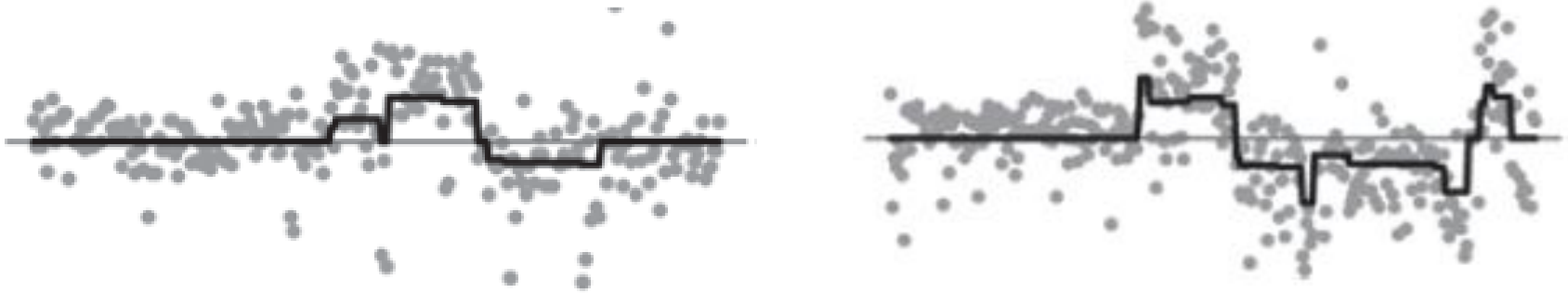
$$\min_x \frac{1}{2}\|Ax - y\|_2^2 + \mathrm{fl}(x)$$

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

Application: Arracy CGH Data Analysis

(Tibshirani and Wang, 2008)

- Comparative genomic hybridization (CGH)
  - Measuring DNA copy numbers of selected genes on the genome
  - In cells with cancer, mutations can cause a gene to be either deleted or amplified
- Array CGH profile of two chromosomes of breast cancer cell line MDA157.



Application to Unordered Features

- Features in some applications are not ordered, e.g., genes in a microarray experiment have no pre-specified order
  - Estimate an order for the features using **hierarchical clustering**

- The leukaemia data [Golub et al. 1999]
  - 7129 genes and 38 samples: 27 in class 1 (acute lymphocytic leukaemia) and 11 in class 2 (acute mylogenous leukaemia)
  - A test sample of size 34

Application to Unordered Features

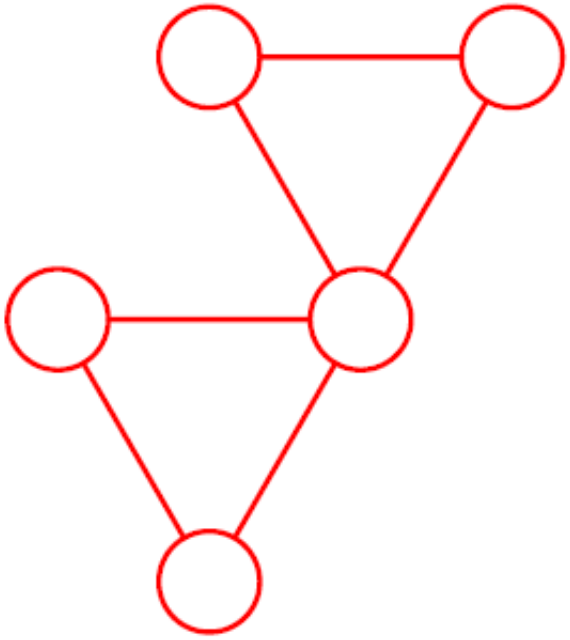
- Features in some applications are not ordered, e.g., genes in a microarray experiment have no pre-specified order
  - Estimate an order for the features using **hierarchical clustering**

	Fused Lasso	Lasso
Array CGH	88%	82%
Prostate Cancer	98%	98%
Leukemias	96%	94%
Leukemias Reordered	97%	94%

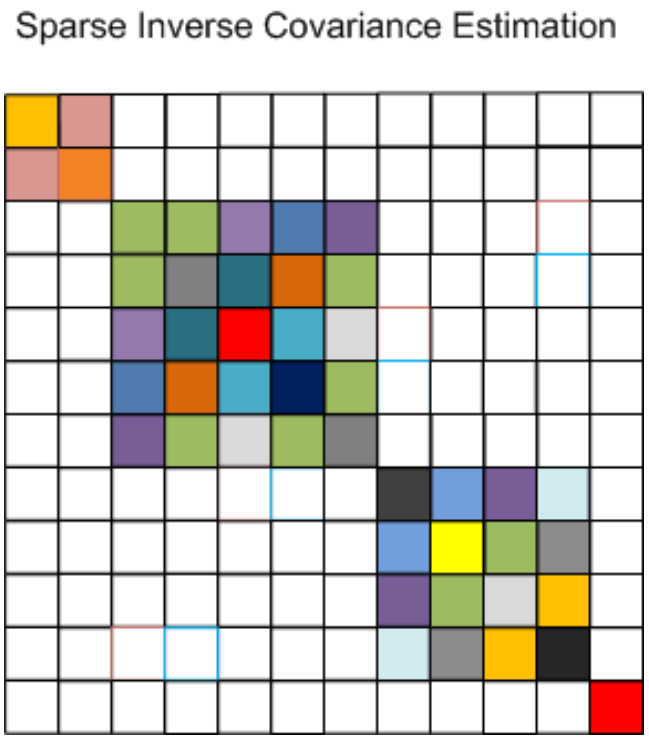
Outline

- Sparse Learning Models
  - 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



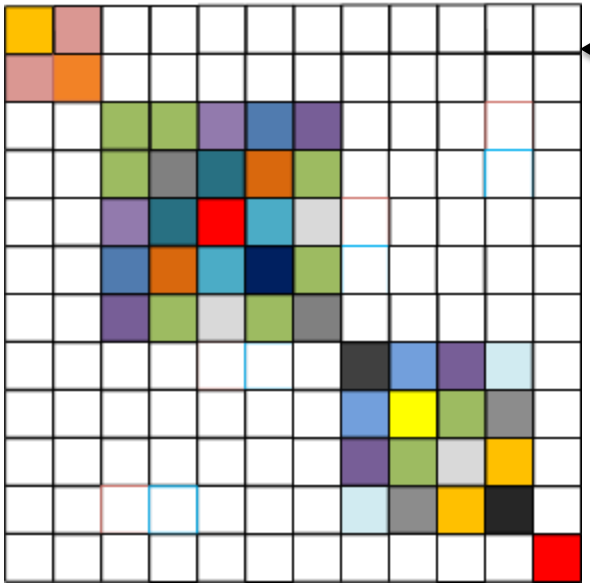
Undirected graphical model
(Markov Random Field)



The pattern of zero entries in the inverse covariance matrix of a multivariate normal distribution corresponds to conditional independence restrictions between variables.

The SICE Model

Sparse Inverse Covariance Estimation



Sparsity

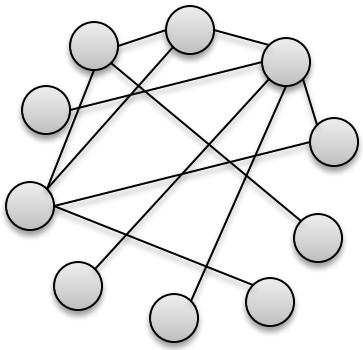
$$\arg\max_{X \succ 0} \log\det X - \mathrm{trace}(SX) - \lambda\|X\|_1$$

Log-likelihood

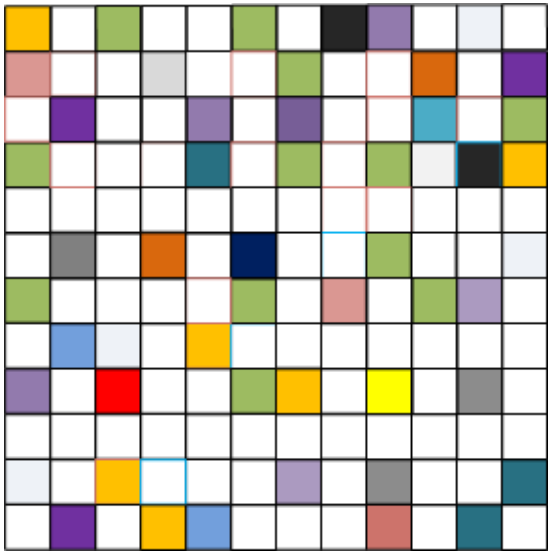
When *S* is invertible, directly maximizing the likelihood gives

$X=S^{-1}$

Network Construction



- Biological network
- Social network
- Brain network



Equivalent matrix representation

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

The Monotone Property (1)

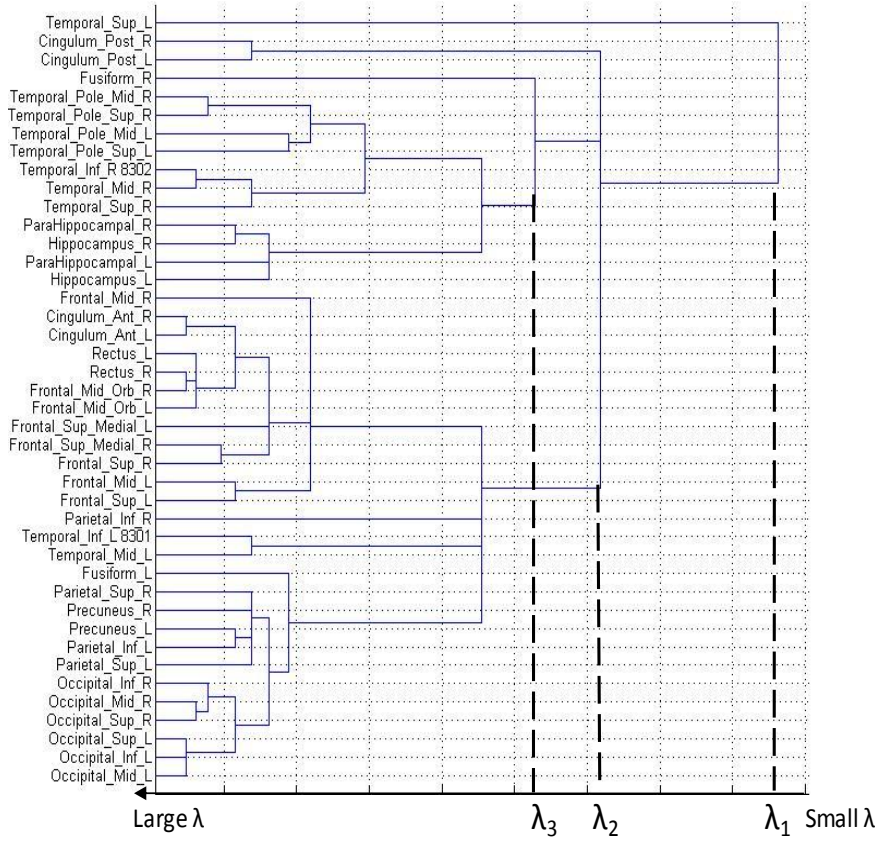
$$\arg\max_{X \succ 0} \log\det X - \mathrm{trace}(SX) - \lambda \|X\|_1$$

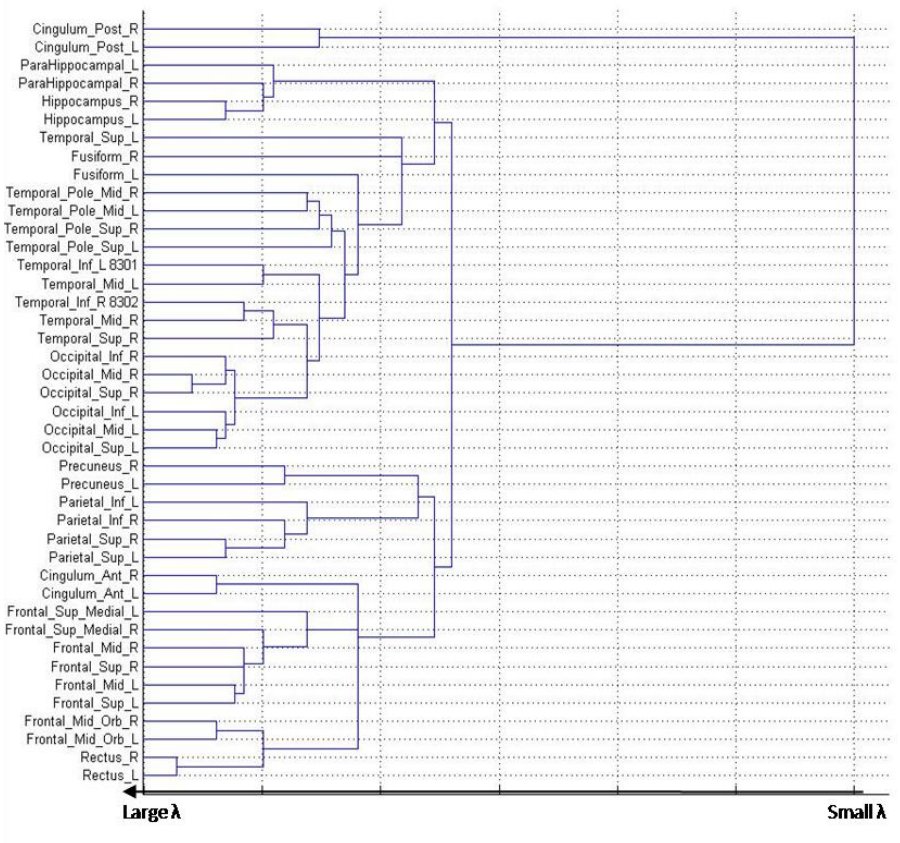
Monotone Property

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.

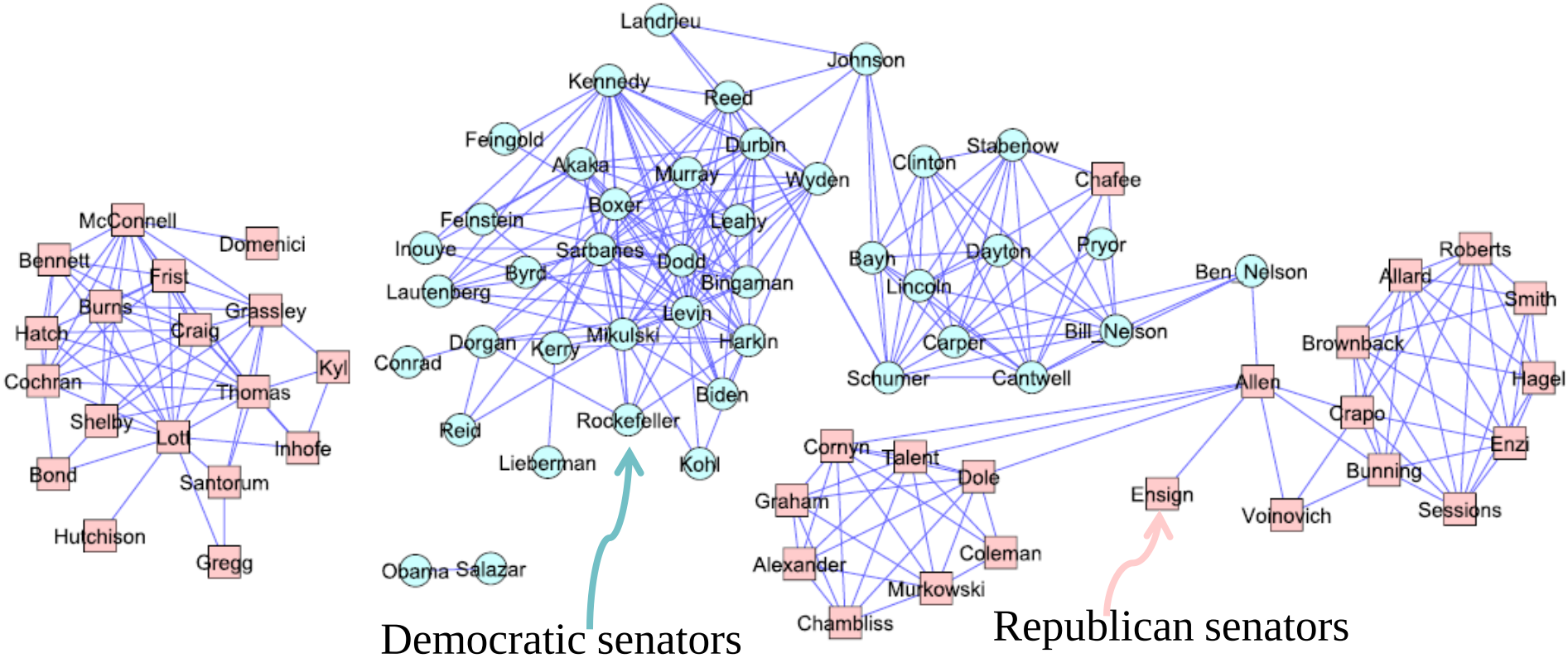
The Monotone Property (2)





Example: Senate Voting Records Data (2004-06)

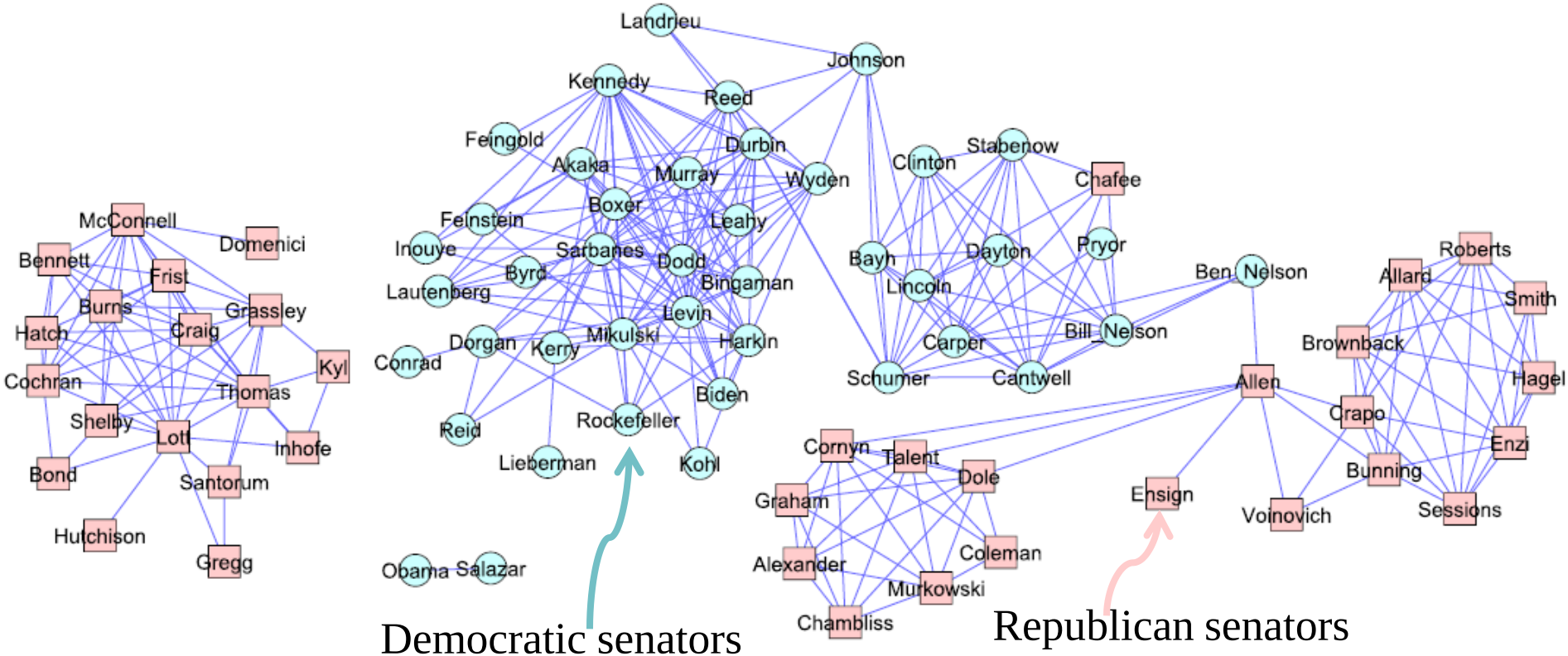
(Banerjee et al., 2008)



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

Example: Senate Voting Records Data (2004-06)

(Banerjee et al., 2008)



Senator Allen (R, VA) unites two otherwise separate groups of Republicans and also provides a connection to the large cluster of Democrats through Ben Nelson (D, NE), which also supports media statements made about him prior to his 2006 re-election campaign.

Brain Connectivity using Neuroimages (1)

- AD is closely related to the alternations of the brain network, i.e., the connectivity among different brain regions
  - AD patients have decreased hippocampus connectivity with prefrontal cortex (Grady et al. 2001) and cingulate cortex (Heun et al. 2006).
- Brain regions are moderately or less inter-connected for AD patients, and cognitive decline in AD patients is associated with disrupted functional connectivity in the brain
  - Celone et al. 2006, Rombouts et al. 2005, Lustig et al. 2006.

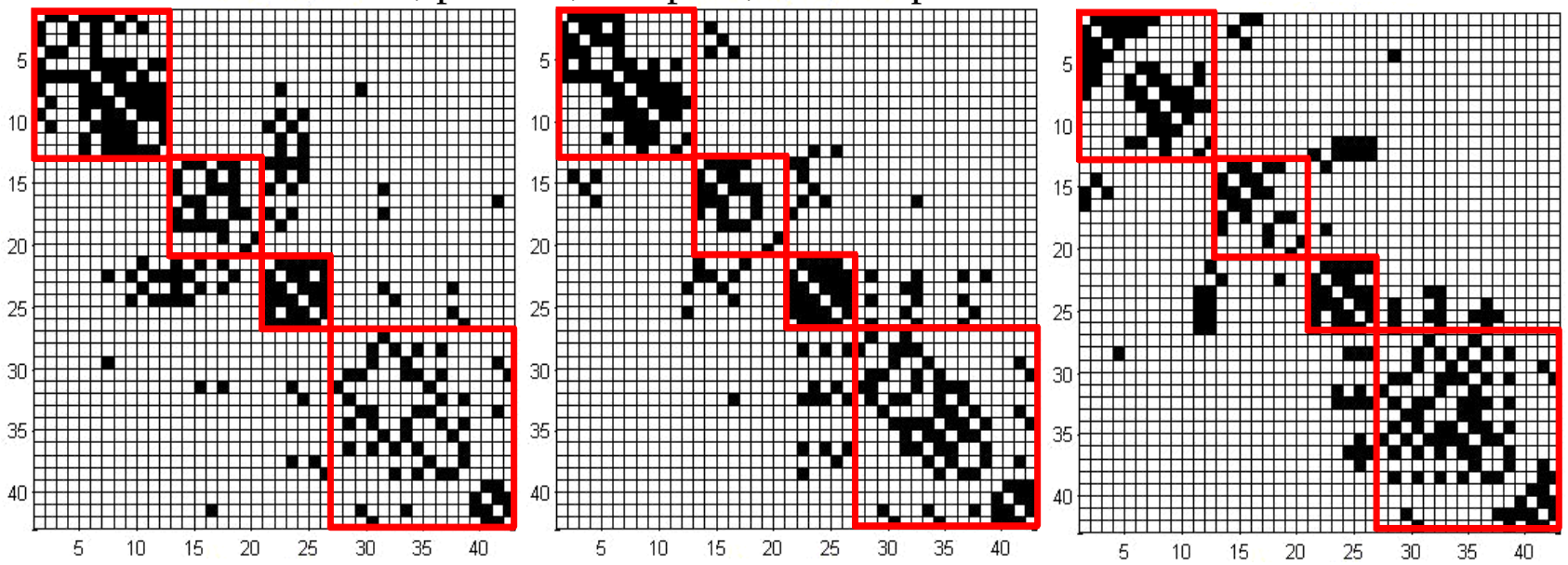
- PET images (49 AD, 116 MCI, 67 NC)
  - AD: Alzheimer's Disease, MCI: Mild Cognitive Impairment, NC: Normal Control
  - http://www.loni.ucla.edu/Research/Databases/

Brain Connectivity using Neuroimages (2)

Frontal lobe		Parietal lobe		Occipital lobe		Temporal lobe	
1	Frontal_Sup_L	13	Parietal_Sup_L	21	Occipital_Sup_L	27	Temporal_Sup_L
2	Frontal_Sup_R	14	Parietal_Sup_R	22	Occipital_Sup_R	28	Temporal_Sup_R
3	Frontal_Mid_L	15	Parietal_Inf_L	23	Occipital_Mid_L	29	Temporal_Pole_Sup_L
4	Frontal_Mid_R	16	Parietal_Inf_R	24	Occipital_Mid_R	30	Temporal_Pole_Sup_R
5	Frontal_Sup_Medial_L	17	Precuneus_L	25	Occipital_Inf_L	31	Temporal_Mid_L
6	Frontal_Sup_Medial_R	18	Precuneus_R	26	Occipital_Inf_R	32	Temporal_Mid_R
7	Frontal_Mid_Orb_L	19	Cingulum_Post_L			33	Temporal_Pole_Mid_L
8	Frontal_Mid_Orb_R	20	Cingulum_Post_R			34	Temporal_Pole_Mid_R
9	Rectus_L					35	Temporal_Inf_L 8301
10	Rectus_R					36	Temporal_Inf_R 8302
11	Cingulum_Ant_L					37	Fusiform_L
12	Cingulum_Ant_R					38	Fusiform_R
						39	Hippocampus_L
						40	Hippocampus_R
						41	ParaHippocampal_L
						42	ParaHippocampal_R

Brain Connectivity using Neuroimages (3)

frontal, parietal, occipital, and temporal lobes in order



AD MCI NC

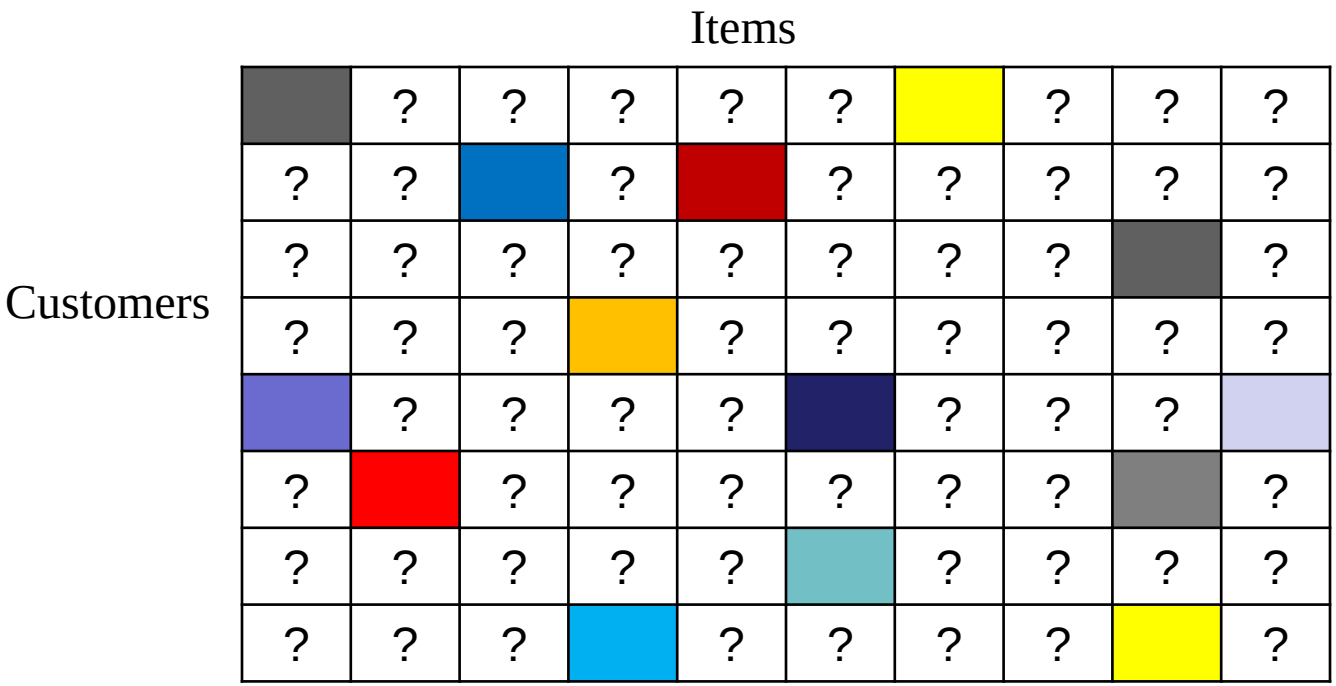
Brain Connectivity using Neuroimages (3)

- The temporal lobe of AD has significantly less connectivity than NC.
  - The decrease in connectivity in the temporal lobe of AD, especially between the Hippocampus and other regions, has been extensively reported in the literature.
- The temporal lobe of MCI does not show a significant decrease in connectivity, compared with NC.
- The frontal lobe of AD has significantly more connectivity than NC.
  - Because the regions in the frontal lobe are typically affected later in the course of AD, the increased connectivity in the frontal lobe may help preserve some cognitive functions in AD patients.

Outline

- Sparse Learning Models
  - 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

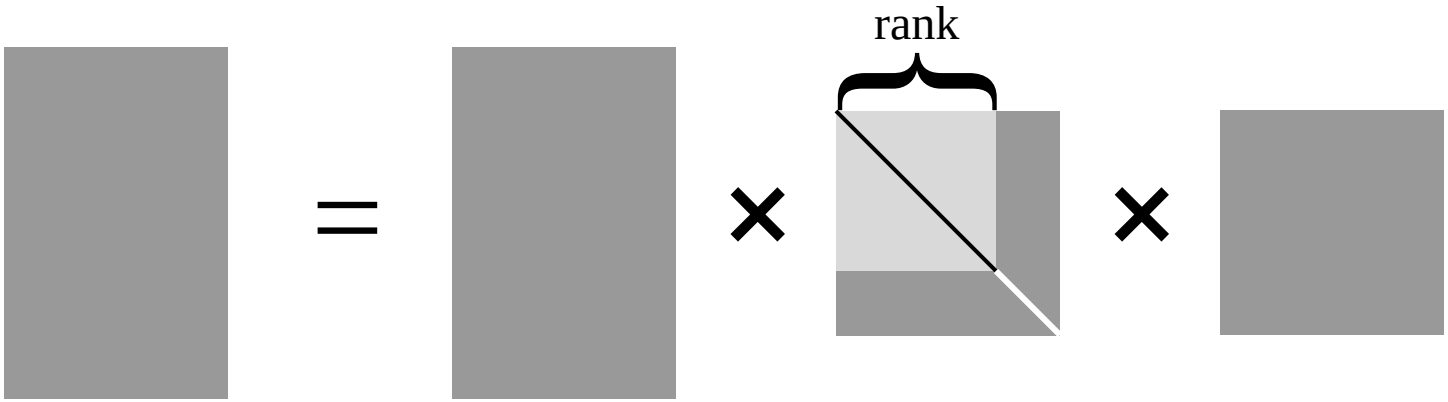
Users

- 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

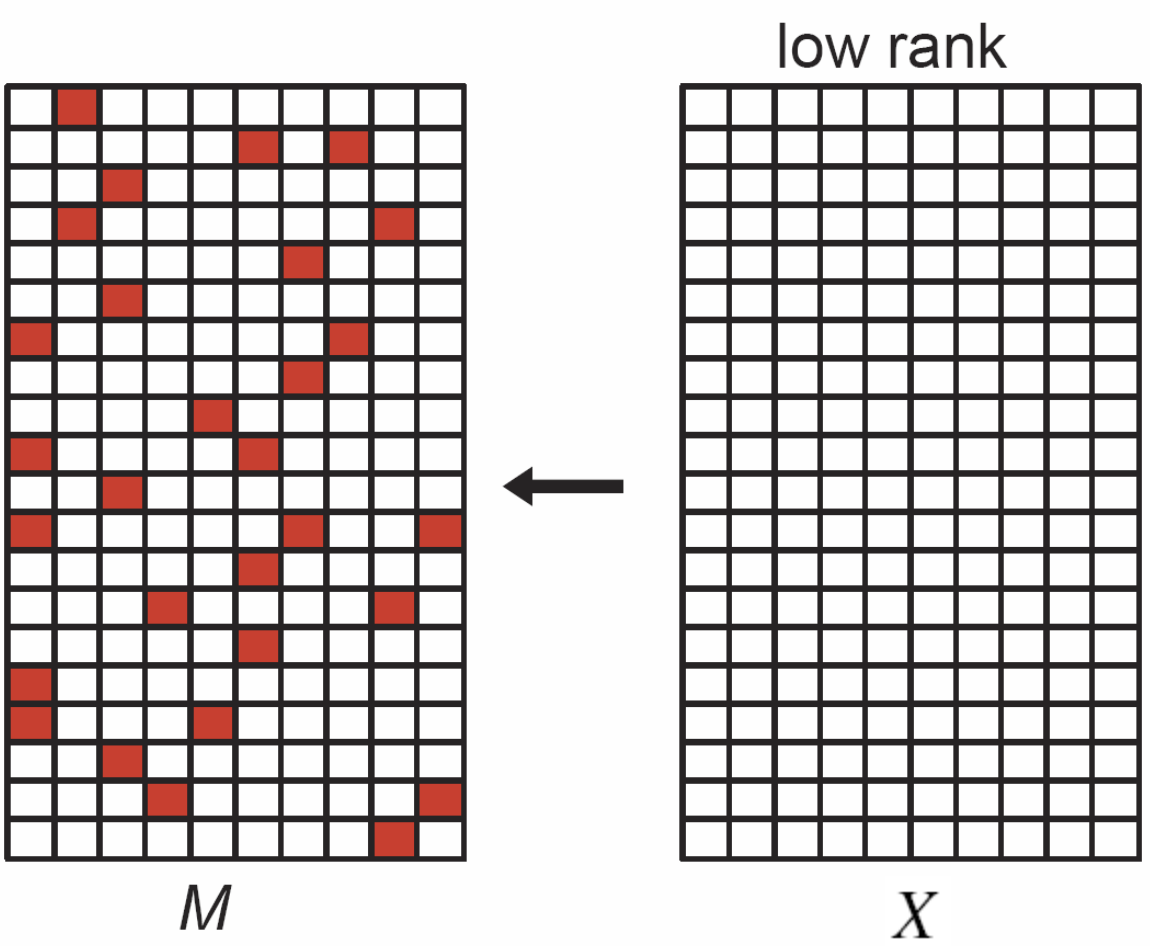
Matrix Rank

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



The Matrix Completion Problem

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



Other Low-Rank Problems

- Multi-Task/Class Learning
- Image compression
- System identification in control theory
- Structure-from-motion problem in computer vision
- Low rank metric learning in machine learning
- 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

$$\min \quad \text{loss}(X) + \lambda * \text{rank}(X)$$

$$\begin{aligned} \min \quad & \text{rank}(X) \\ \text{subject to} \quad & \text{loss}(X) \leq \varepsilon \end{aligned}$$

Rank minimization is NP-hard

Trace Norm (Nuclear Norm)

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

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

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

- **trace norm ⇔ 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 ⇒ a convex relaxation**

Two Formulations for Trace Norm

$$\min \quad \text{loss}(X) + \lambda \ \|X\|_*$$

$$\begin{aligned} \min \quad & \|X\|_* \\ \text{subject to} \quad & \text{loss}(X) \leq \varepsilon \end{aligned}$$

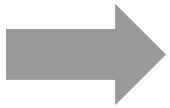
Trace norm minimization is convex

- Can be solved by
  - Semi-definite programming
  - Gradient-based methods

Semi-definite programming (SDP)

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

$$\text{s.t. } \text{loss}(X) \le \varepsilon$$



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

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

$$\text{loss}(X) \le \varepsilon$$

- 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)

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

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

Theory of Matrix Completion

(Candes and Recht, 2008)

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

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

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

- *random set of entries of size* m

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

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

$$m \geq C\, \mu^2 n r \log^6 n \quad \text{(Candes and Tao, 2010)}$$

Outline

- Sparse Learning Models
  - 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

Optimization Algorithms

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

Smooth and convex

- Least squares
- Logistic loss
- …

Convex but nonsmooth

- L_1
- L_1/L_q
- Fused Lasso
- Trace Norm
- …

Optimization Algorithms

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

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

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

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

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

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 Smooth 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 use of existing solvers
- Fast and high precision for small size problems

Disadvantages:

- Does not scale well for large size problems due to many additional variables and constraints introduced
- 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".

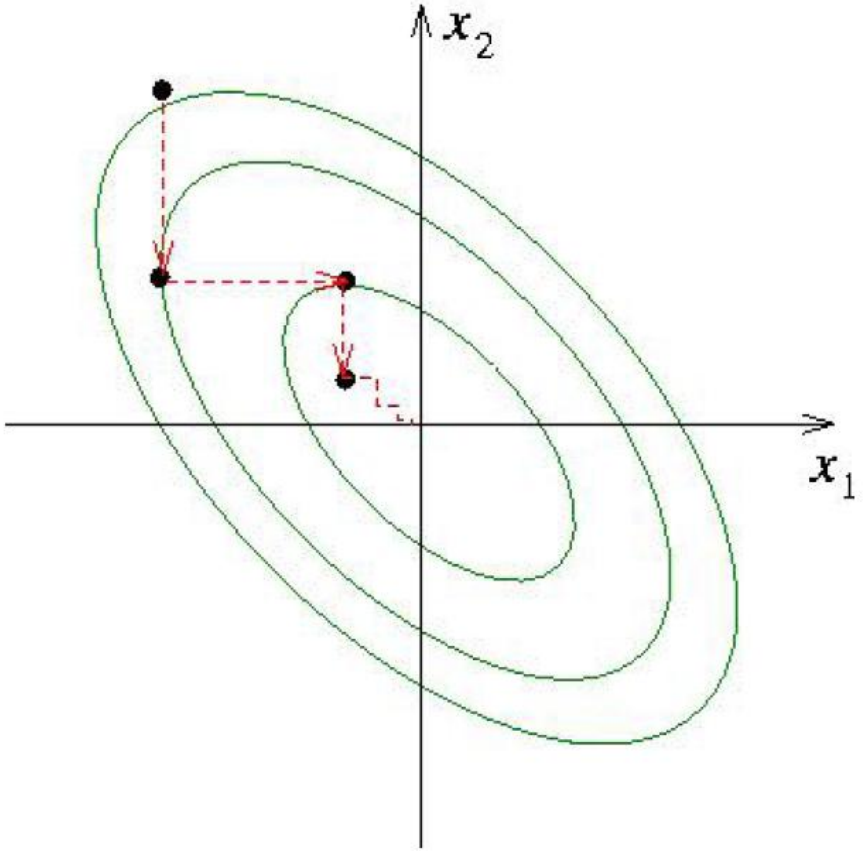
Gauss-Seidel: Choose i cyclically, 1, 2,..., n, 1, 2,...

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

Coordinate Descent: Example

(Tseng, 2002)

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

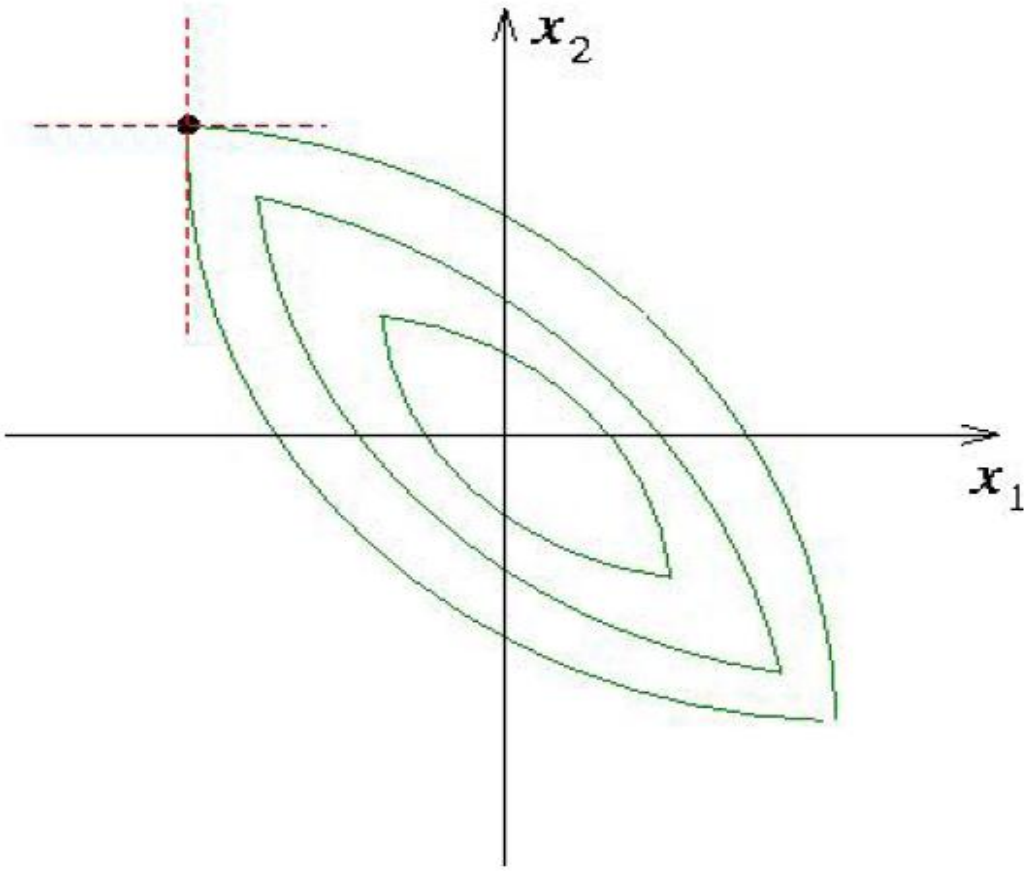


Coordinate Descent: Convergent?

(Tseng, 2002)

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 and convex, 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

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

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 inverse covariance estimation

  (Banerjee et al., 2008; Friedman et al., 2007)

- Fused Lasso and Fused Lasso Signal Approximator

  (Friedman et al., 2007; Hofling, 2010)

CD can get stuck for fused Lasso

Summary of CD

Advantages:

- Easy implementation, especially for the least squares loss
- Can be fast, especially when 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)

$$\min_x f(x)$$

Repeat

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

Until "convergence"

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

Subgradient: one element in the subdifferential set

Subgradient Descent: Convergent?

(Nemirovski, 1994; Nesterov, 2004)

Repeat

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

Until "convergence"

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

$$\gamma_i = Di^{-1/2}, i = 1, \cdots, N$$

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_∞ constrained optimization (Quattoni et al., 2009)

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

$$\min_x f(x)$$

Repeat

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

Until "convergence"

- How can we apply gradient descent to nonsmooth sparse learning problems?

$f(x)$ is continuously differentiable with Lipschitz continuous gradient L

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"

Equivalent

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

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

1st order Taylor expansion

Model

Regularization

Gradient Descent:

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

min $f(x)$= loss(x) + λ×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"

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

proximal operator (Moreau, 1965)

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

$$\rho = \gamma_i \lambda$$

Gradient Descent:

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

Model

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

- Can O(1/N) be further improved?
- The lower complexity bound shows that, the first-order methods can achieve a convergence rate no better than $O(1/N^2)$.
- Can we develop a method that can achieve the optimal convergence rate $O(1/N^2)$?

Accelerated Gradient Descent:

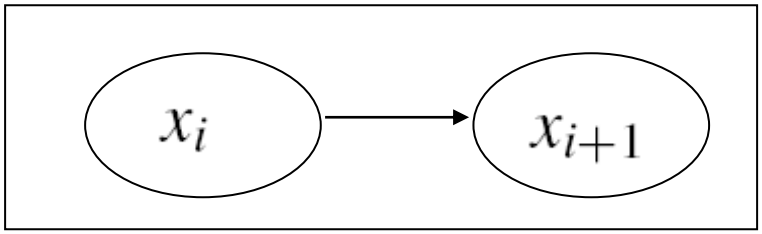
(Nesterov, 1983; Nemirovski, 1994; Nesterov, 2004)

GD $\min_x f(x)$

Repeat

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

Until "convergence"



$O(1/N)$

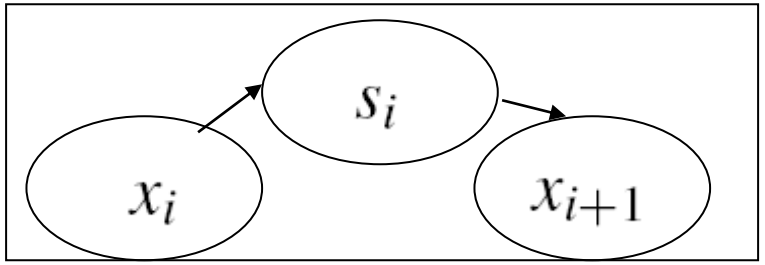
AGD

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:

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

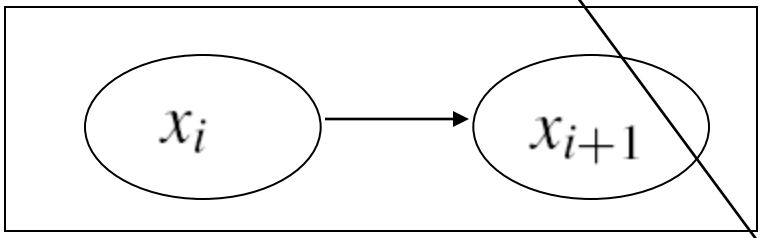
$$\mathcal{M}(x_i, \gamma_i) = [\text{loss}(x_i) + \langle \text{loss}'(x_i), x - x_i \rangle] + \frac{1}{2\gamma_i}\|x - x_i\|_2^2 + \lambda \times \text{penalty}(x)$$

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"



$O(1/N)$

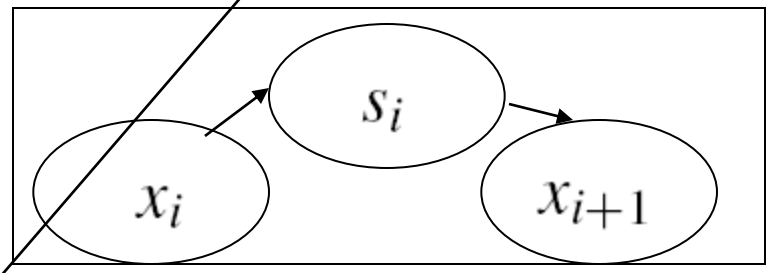
AGD

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)

- 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)

Accelerated Gradient Descent in Sparse Representations

Advantages:

- Easy implementation
- Optimal convergence rate
- Scalable to large-size problems

Key computational cost

- Gradient and functional value
- The associated proximal operator

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

L_1
Trace Norm
L_1/L_q
Fused Lasso

Proximal Operator Associated with L_1

Optimization problem

$$\min_x f(x) = \text{loss}(x) + \lambda \|x\|_1$$

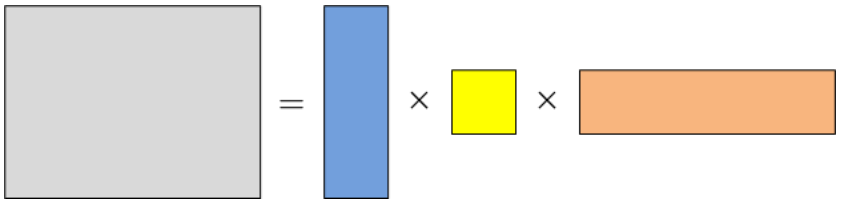
Associated proximal operator

$$x^* = \pi_{\ell_1}(v) = \arg\min_x \frac{1}{2}\|x - v\|_2^2 + \lambda \times \|x\|_1$$

Closed-form solution:

$$x_i^* = \begin{cases} v_i - \lambda & v_i > \lambda \\ v_i + \lambda & v_i < -\lambda \\ 0 & -\lambda \le v_i \le \lambda \end{cases}$$

Proximal Operator Associated with Trace Norm



Optimization problem

$$\min_X f(X) = \text{loss}(X) + \lambda \|X\|_*$$

Associated proximal operator

$$X^* = \pi_{tr}(V) = \arg\min_X \frac{1}{2}\|X - V\|_2^2 + \lambda \times \|X\|_*$$

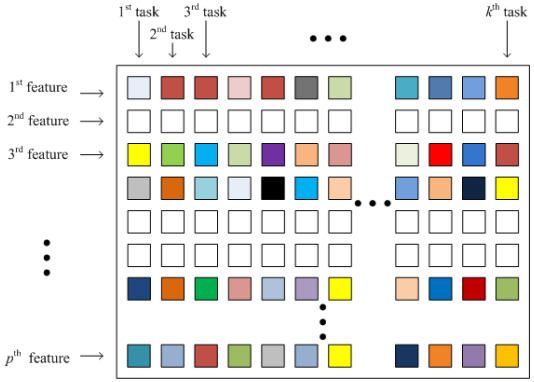
Closed-form solution:

$$X^* = P\text{diag}(\tilde{\sigma})Q^{\mathrm{T}},$$

where $V = P\text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)Q^{\mathrm{T}}$ is the SVD of $V \in \mathbb{R}^{m\times n}$, $k = \min(m, n)$, $P \in \mathbb{R}^{m\times k}$, $Q \in \mathbb{R}^{n\times k}$, and

$$\tilde{\sigma}_i = \begin{cases} v_i - \lambda & \sigma_i > \lambda \\ 0 & \sigma_i \leq \lambda \end{cases}$$

Proximal Operator Associated with L_1/L_q



Optimization problem:

$$\min_X f(X) = \text{loss}(X) + \lambda \sum_{i=1}^{p} \|\mathbf{x}_i\|_q$$

Associated proximal operator:

$$\pi_{1q}(V) = \arg\min_X \frac{1}{2}\|X - V\|^2 + \lambda \sum_{i=1}^{p} \|\mathbf{x}_i\|_q$$

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

$$\pi_q(\mathbf{v}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|^2 + \lambda\|\mathbf{x}\|_q$$

When $q = 1$ or $q = \infty$

When q=1, the problem admits a closed form solution

$$\pi_1(\mathbf{v}) = \mathrm{sgn}(\mathbf{v}) \odot \max(|\mathbf{v}| - \lambda, 0)$$

When $q = \infty$, we have

$$\pi_\infty(\mathbf{v}) = \mathbf{v} - \left(\arg \min_{\|\mathbf{y}\|_1 \leq \lambda} \frac{1}{2} \|\mathbf{y} - \mathbf{v}\|_2^2 \right)$$

Therefore, $\pi_\infty(\mathbf{v})$ can be solved via the Euclidean projection onto the 1-norm ball (Duchi et al., 2008; Liu and Ye, 2009).

The Euclidean projection onto the 1-norm ball can be solved in linear time (Liu and Ye, 2009) by converting it to a zero finding problem.

Proximal Operator Associated with L_1/L_q

$$\pi_q(\mathbf{v}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|^2 + \lambda\|\mathbf{x}\|_q$$

Method:

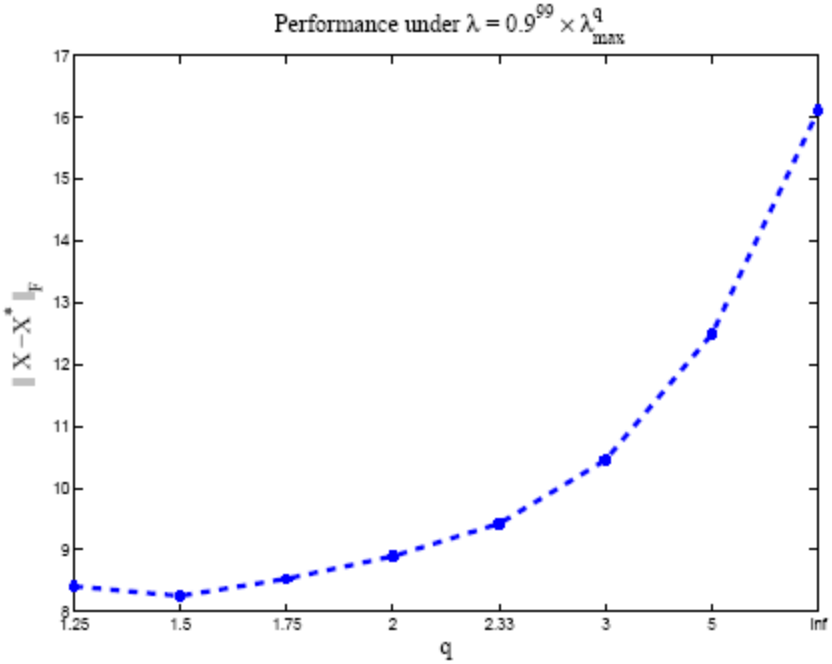
Convert it to two simple zero finding algorithms

Characteristics:

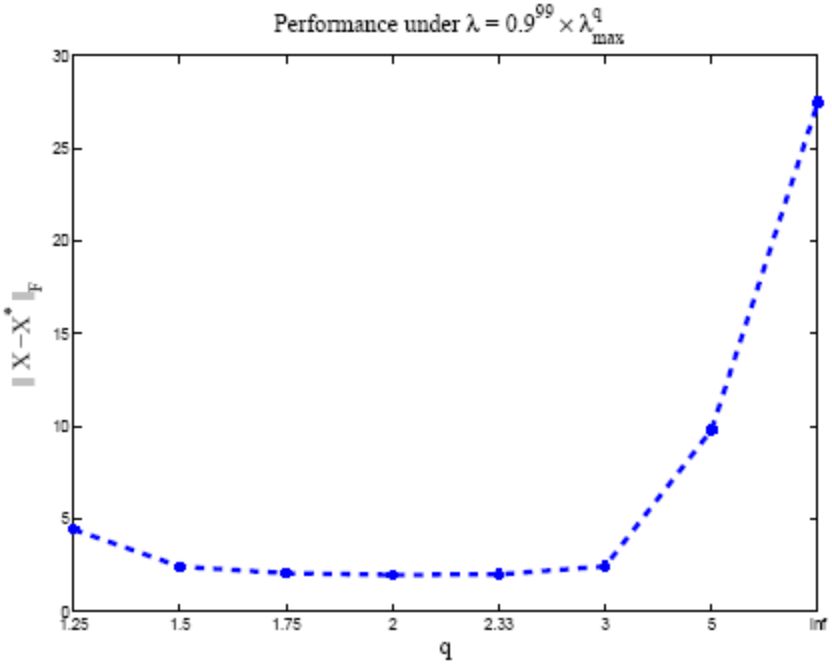
1. Suitable to any $q \geq 1$
2. The proximal plays a key building block in quite a few methods such as the accelerated gradient descent, coordinate gradient descent(Tseng, 2008), forward-looking subgradient (Duchi and Singer, 2009), and so on.

Effect of q in L_1/L_q

Multivariate linear regression



Truth X* is drawn from the random uniform distribution



Truth X* is drawn from the random Gaussian distribution

Proximal Operator Associated with Fused Lasso

Optimization problem:

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

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)

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

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

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

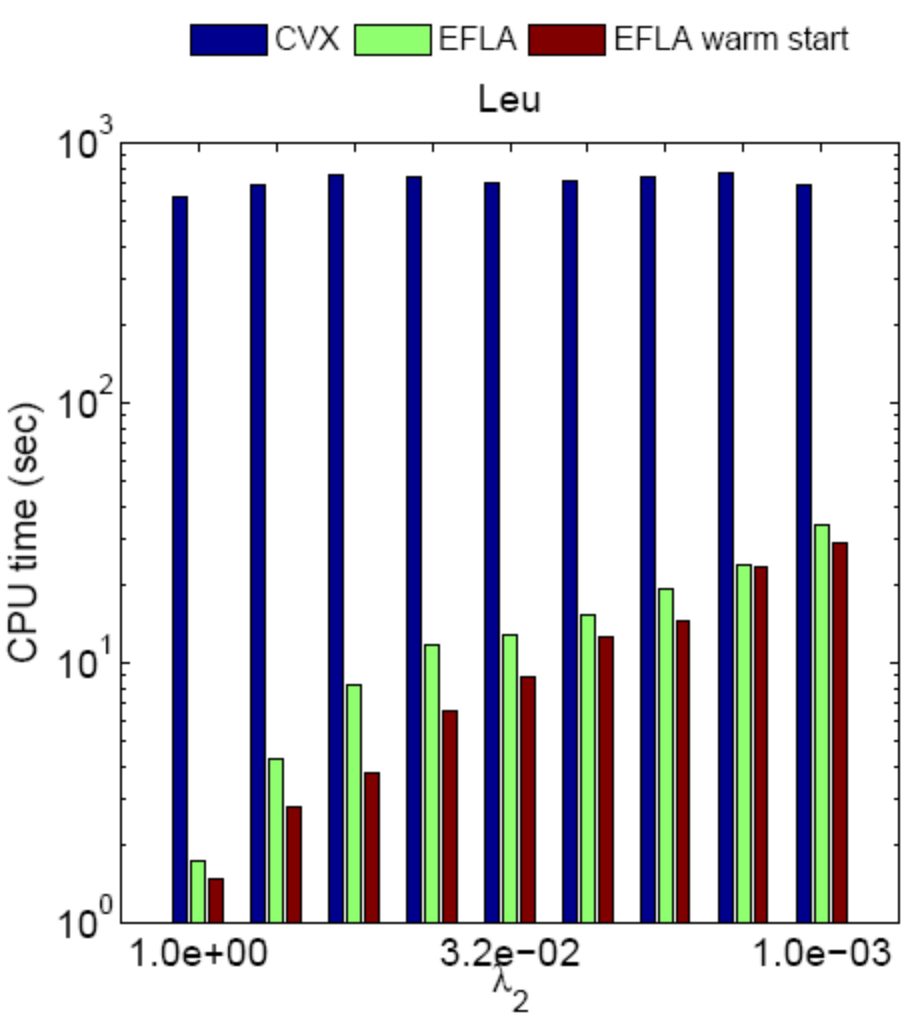
$$\min_{\|\mathbf{z}\|_\infty \le \lambda_2} \psi(\mathbf{z}) \equiv -\phi(\mathbf{v}-R^{\mathrm{T}}\mathbf{z},\mathbf{z}) = \frac{1}{2}\|R^{\mathrm{T}}\mathbf{z}\|^2 - \langle R^{\mathrm{T}}\mathbf{z},\mathbf{v}\rangle.$$

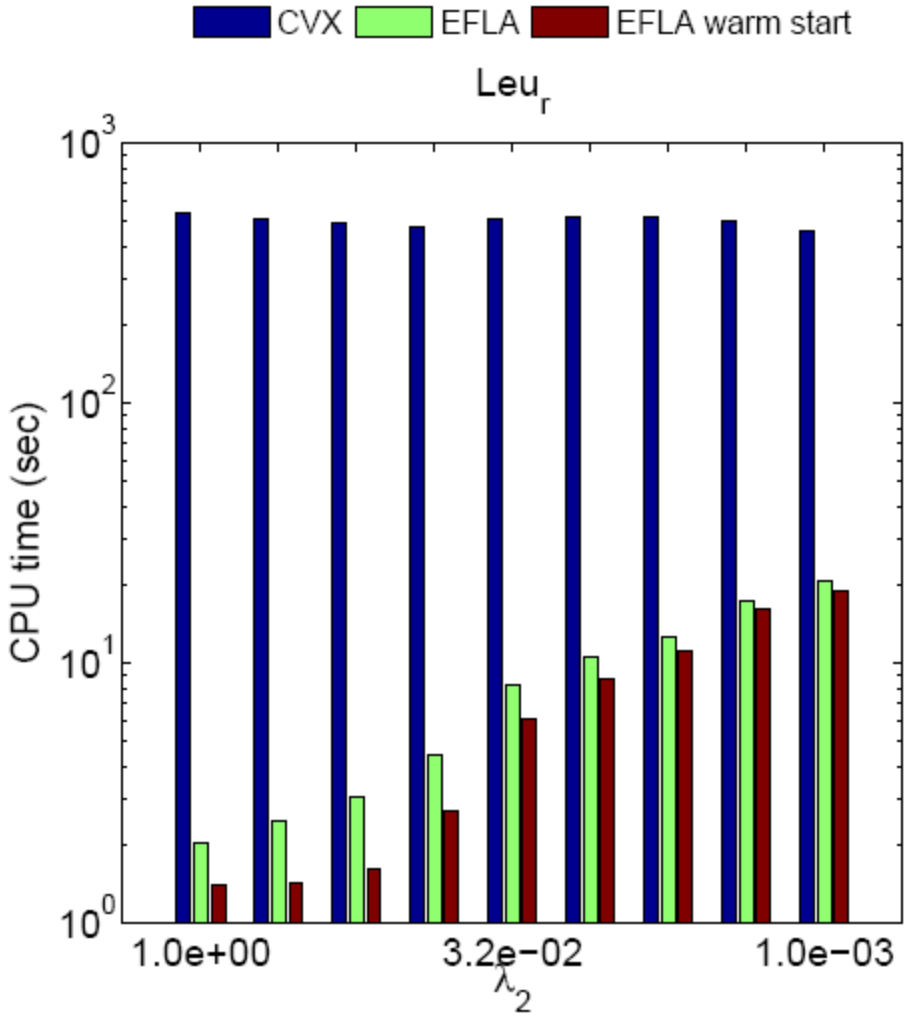
Method:

- Subgradient Finding Algorithm (SFA)---looking for an appropriate and unique subgradient
- Restart

Efficiency

(Comparison with the CVX solver)





Summary of Implementation

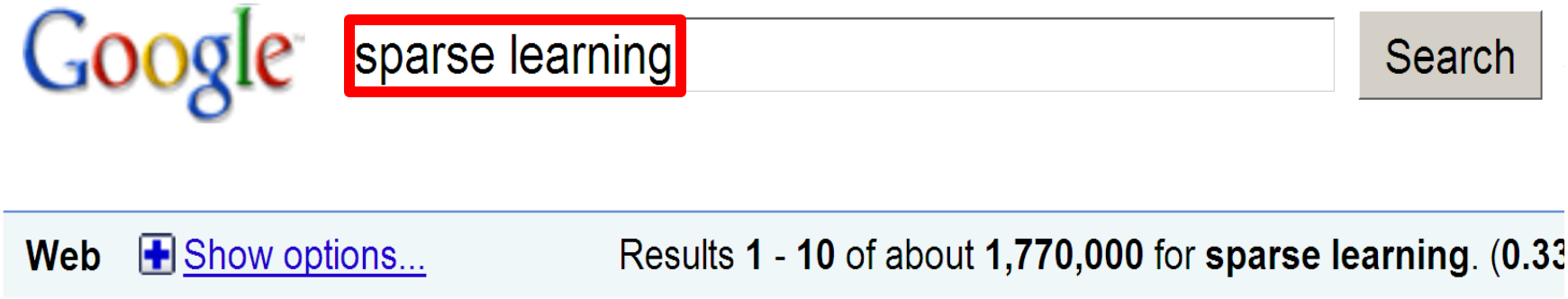
- Smooth reformulation
  - Easy to apply existing solves, but not scalable
- Subgradient descent
  - Easy implementation and guaranteed convergence rate, but slow and hard to achieve sparse solution in a limited number of iterations
- Coordinate descent
  - Easy implementation, but can get stuck for non-separable penalty
- Accelerated Gradient Descent
  - Optimal convergence rate, and the key is to design efficient algorithms for computing the associated proximal operator

SLEP Package

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

SLEP: Sparse Learning with Efficient Projections

Liu, Ji, and Ye (2009) SLEP: A Sparse Learning Package
http://www.public.asu.edu/~jye02/Software/SLEP/



Functions Provided in SLEP

- L_1

  Lasso, Logistic Regression

- Trace Norm

  Multi-task learning, primal-dual optimization

- L_1/L_q

  Group Lasso, mutli-task learning, multi-class classification

- Fused Lasso

  Fused Lasso, fused Lasso signal approximator

- Sparse Inverse Covariance Estimation

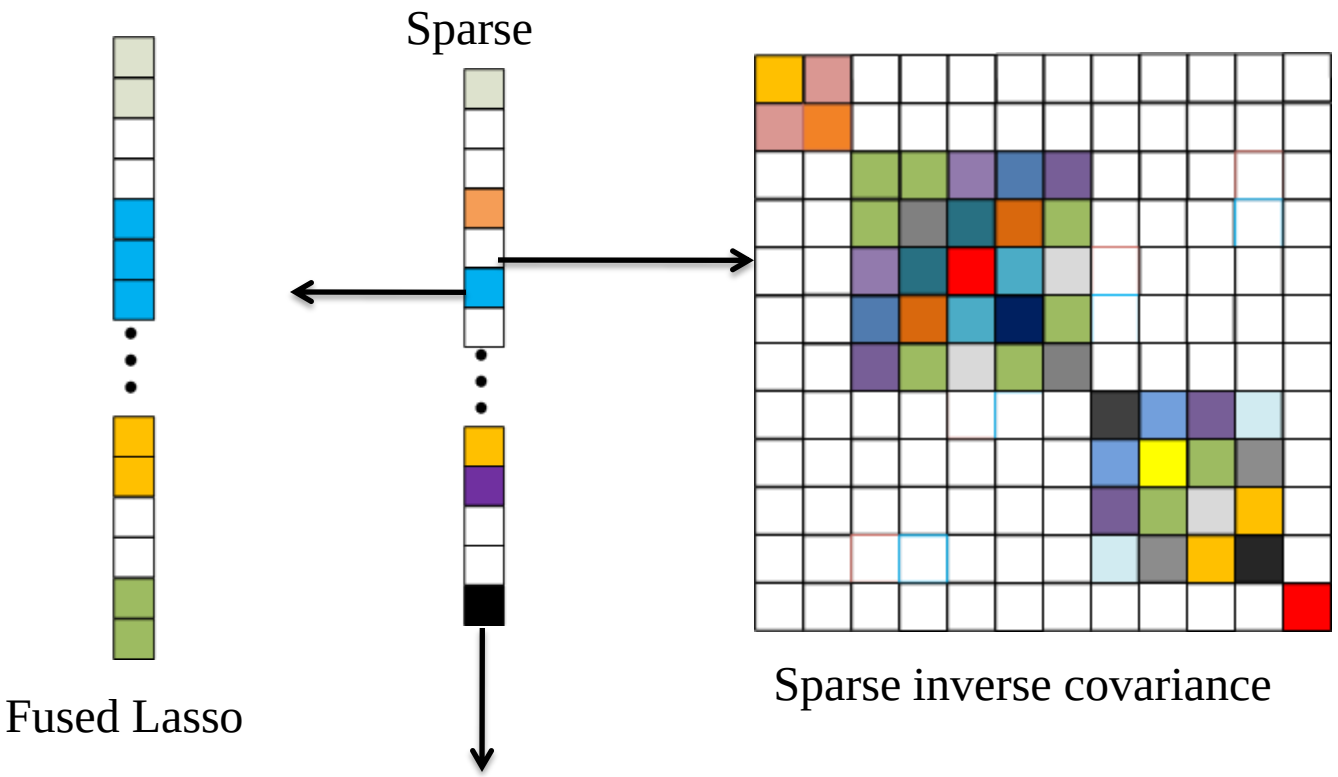
  L1 regularized inverse covariance estimation

Outline

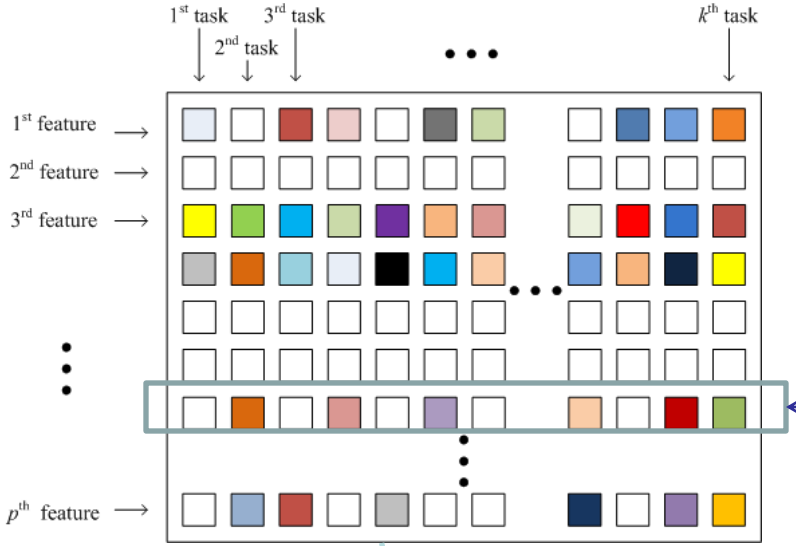
- Sparse Learning Models
  - Sparsity via L_1
  - Sparsity via L_1/L_q
  - Sparsity via Fused Lasso
  - Sparse Inverse Covariance Estimation
  - Sparsity via Trace Norm
- Implementations and the SLEP Package
- **Trends in Sparse Learning**

New Sparsity Inducing Penalties?

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



Sparse Group Lasso via L_1+ L_1/L_q



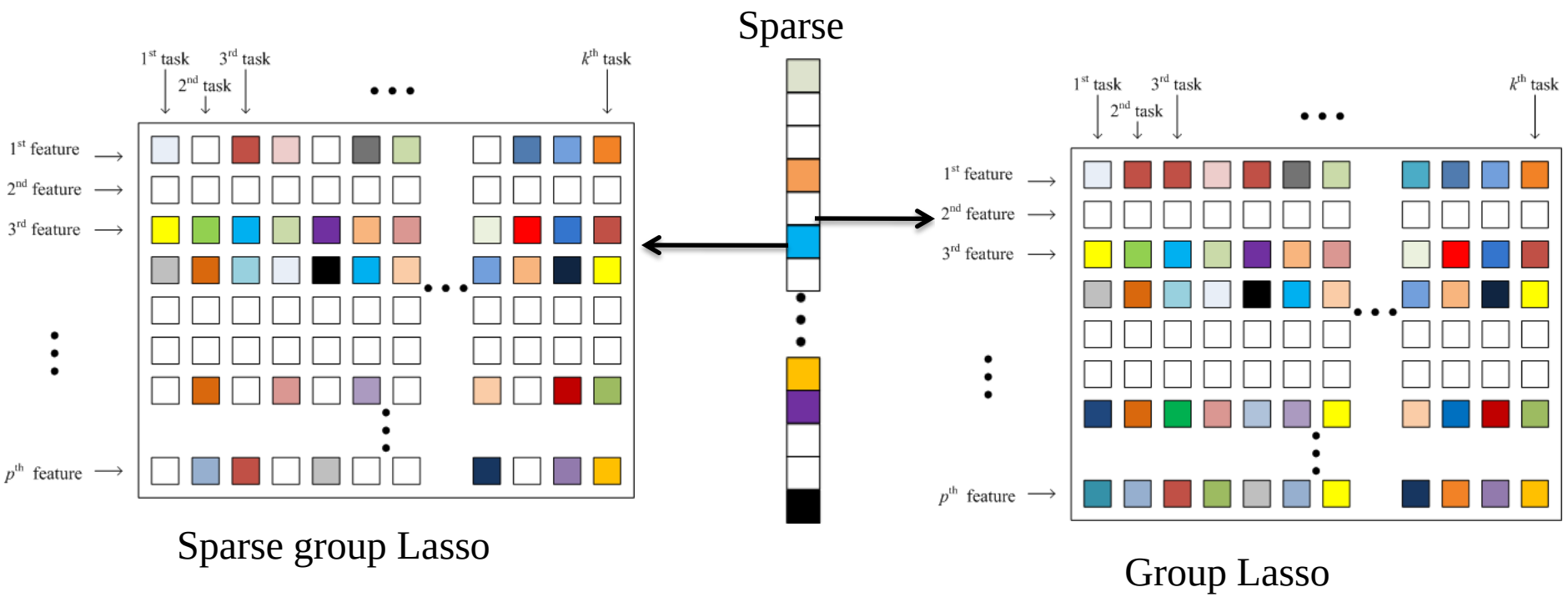
$$\lambda_1 \sum_{i=1}^{p} \|\mathbf{x}_i\|_1 + \lambda_q \sum_{i=1}^{p} \|\mathbf{x}_i\|_q$$

Sparse Group Lasso

Application: Multi-Task Learning, group feature selection

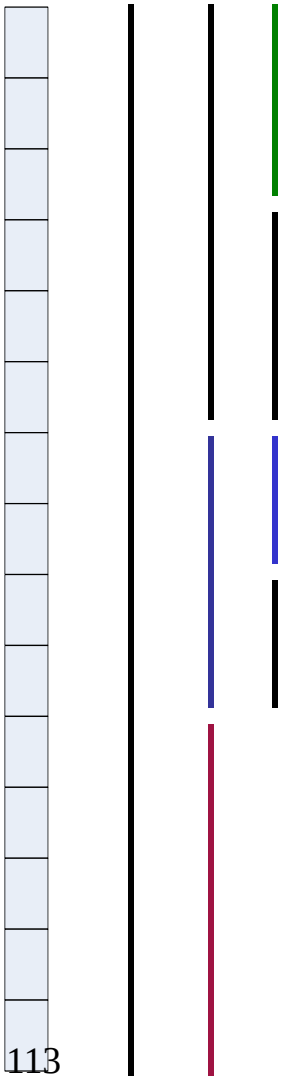
Overlapping Groups?

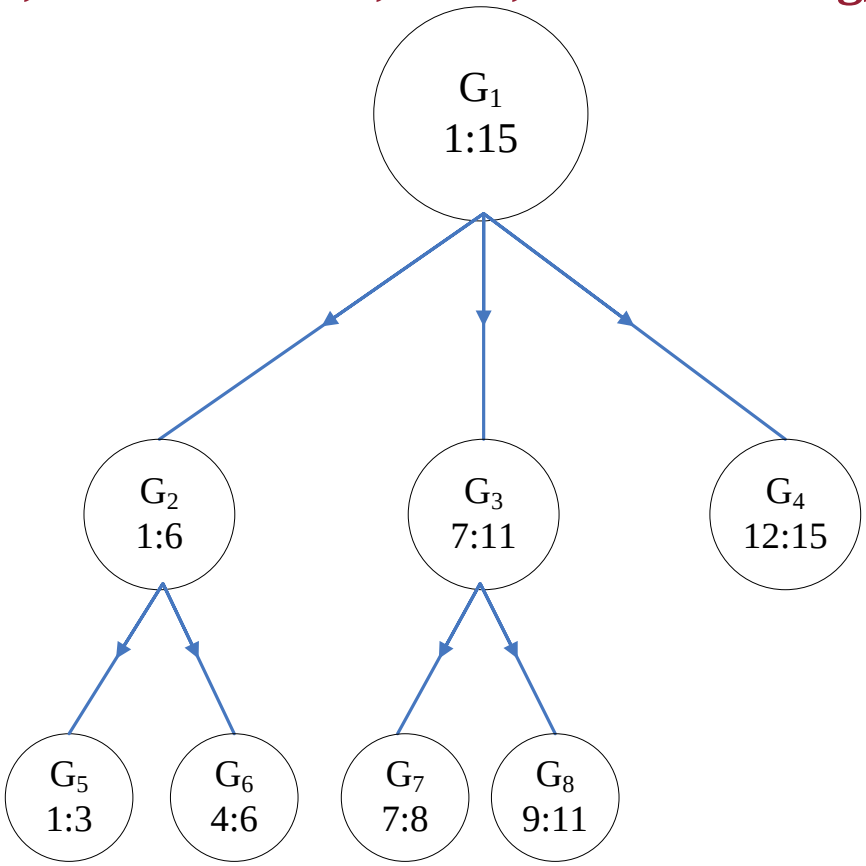
$$\min \; f(x) = \text{loss}(x) + \lambda \times \text{penalty}(x)$$



Group Sparsity with Tree Structure

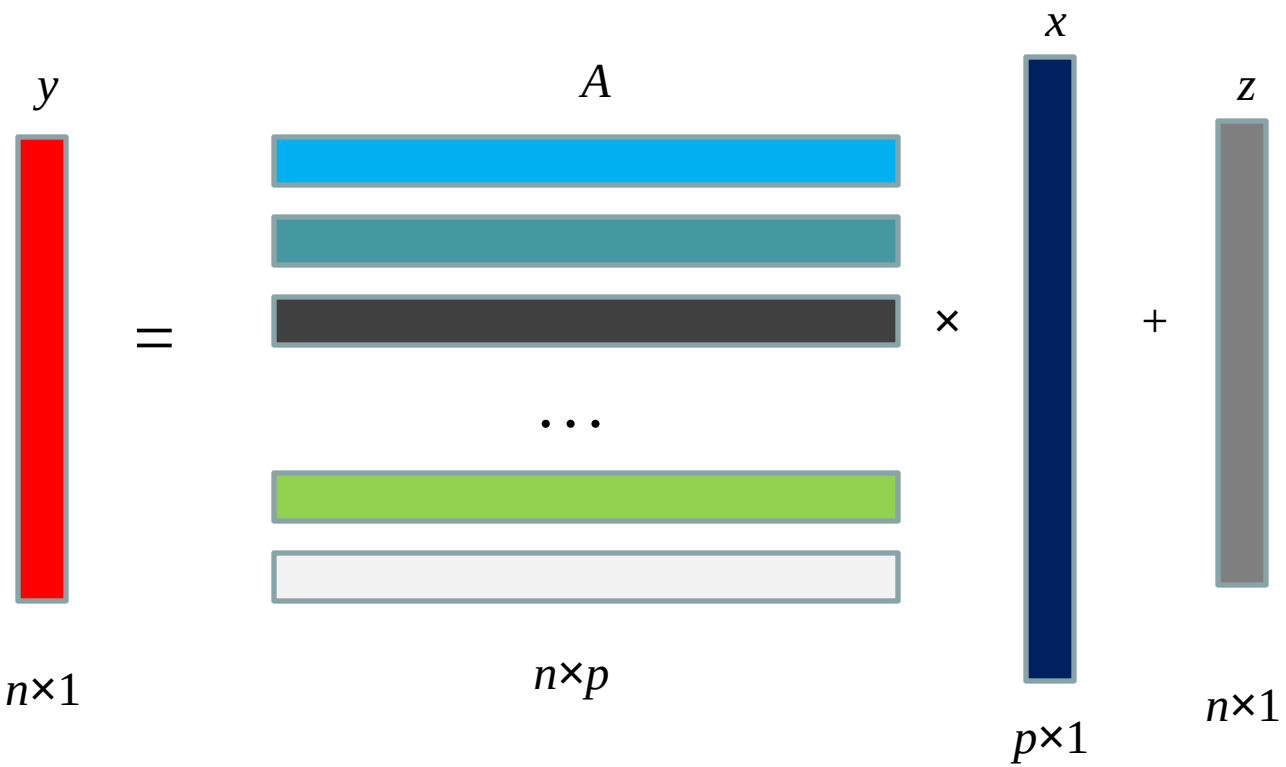
(Zhao et al., 2008; Janatton et al., 2009; Kim and Xing, 2010)





$$\phi(\mathbf{x}) = \sum_{i=1}^{g} w_i \|\mathbf{x}_{G_i}\|$$

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 ℓ_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 ℓ_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 ℓ_1-ℓ_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 ℓ_1/ℓ_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.

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.*

Acknowledgement

- National Science Foundation
- National Geospatial Agency
- Office of the Director of National Intelligence

Thank you!