Feature Selection and Sparse Learning

Jiayu Zhou

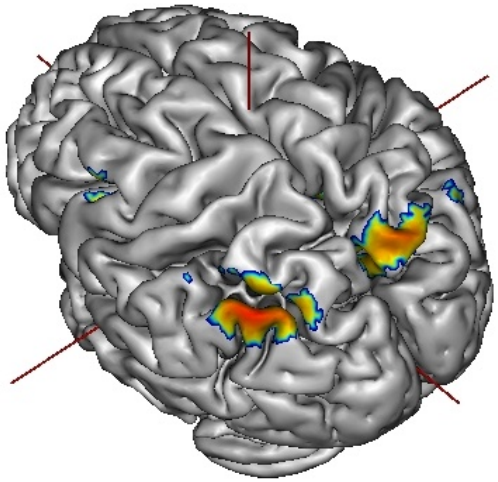
[1]Department of Computer Science and Engineering
Michigan State University
East Lansing, MI USA

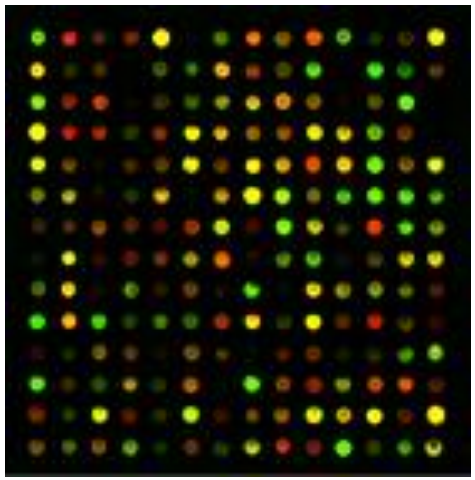
April 12, 2016

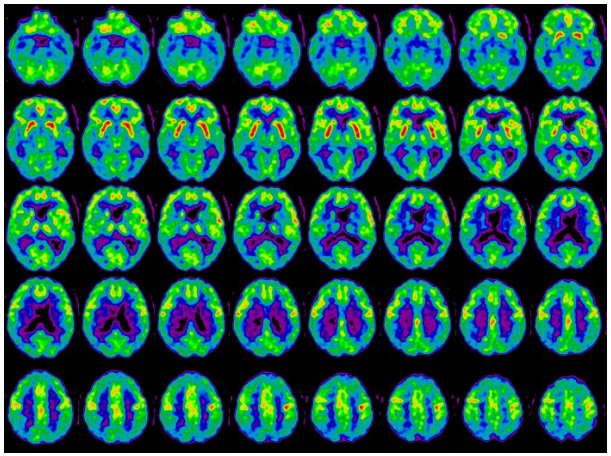
Table of contents

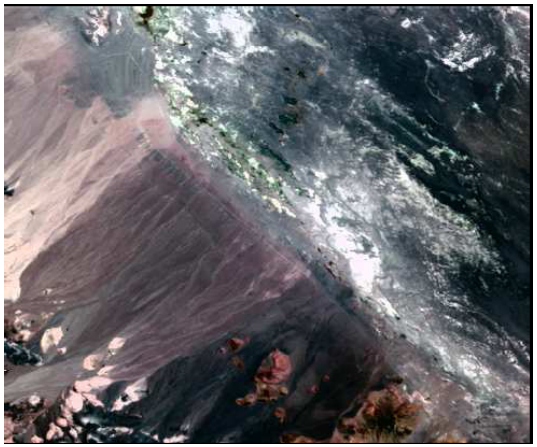
Feature Selection

Mining High-Dimensional Data

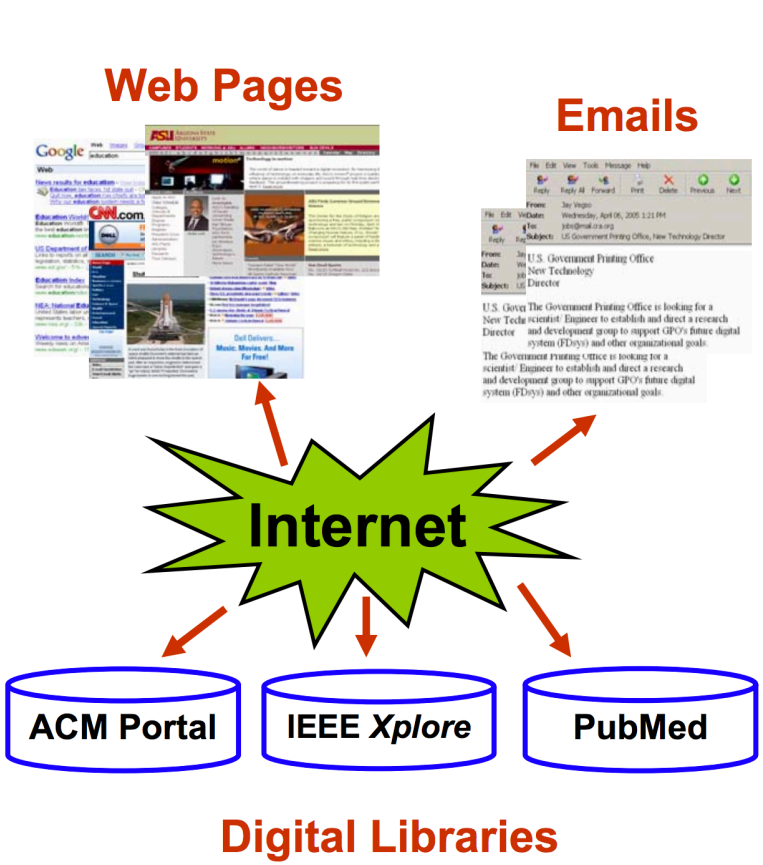








Document Classification

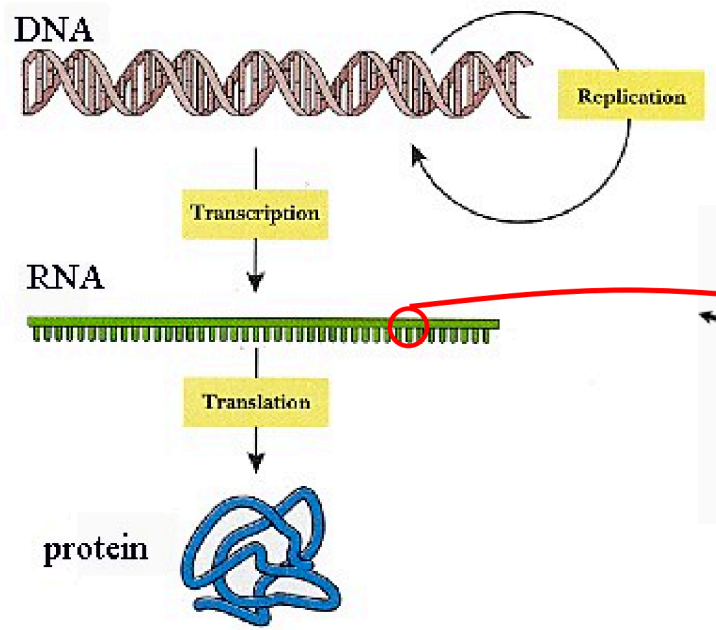


	Terms				
	T_1	T_2	.....	T_N	C
D_1	12	0	.....	6	Sports
D_2	3	10	.....	28	Travel
⋮	⋮			⋮	⋮
D_M	0	11	.....	16	Jobs

Documents

- **Task:** To classify unlabeled documents into categories
- **Challenge:** thousands of terms
- **Solution:** to apply dimensionality reduction

Gene Expression Microarray Analysis



Expression Microarray

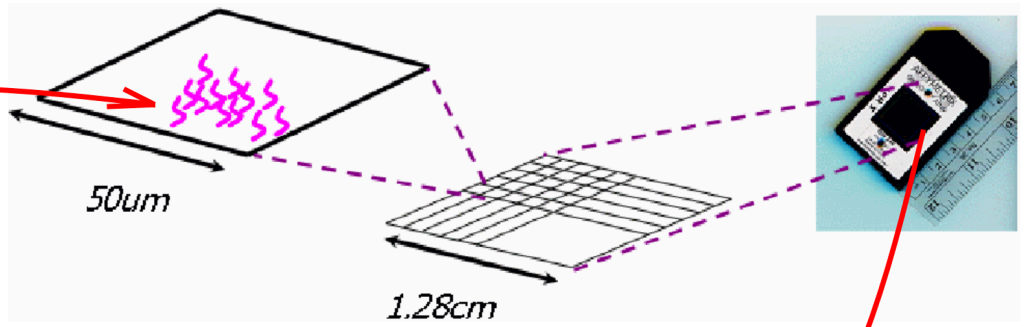


Image Courtesy of Affymetrix

- **Task:** To classify novel samples into known disease types (disease diagnosis)
- **Challenge:** thousands of genes, few samples
- **Solution:** to apply dimensionality reduction

Sample \ Gene	M23197_at	U66497_at	M92287_at	...	Class
Sample 1	261	88	4778	...	ALL
Sample 2	101	74	2700	...	ALL
Sample 3	1450	34	498	...	AML
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.

Expression Microarray Data Set

Major Techniques of Dimensionality Reduction

Feature Extraction

- Feature reduction refers to the mapping of the original high-dimensional data onto a lowerdimensional space
- Given a set of data points of p variables $\{x_1, \ldots, x_n\}$, Compute their low-dimensional representation:

$$x_i \in \mathbb{R}^d \rightarrow y_i \in \mathbb{R}^p (p \ll d)$$

- Criterion for feature reduction can be different based on different problem settings
  - Unsupervised setting: minimize the information loss
  - Supervised setting: maximize the class discrimination

Major Techniques of Dimensionality Reduction

Feature Selection

- A process that chooses an optimal subset of features according to a objective function
- Objectives
  - To reduce dimensionality and remove noise
  - To improve mining performance
    - Speed of learning
    - Predictive accuracy
    - Simplicity and comprehensibility of mined results

Feature Reduction vs Feature Selection

- Feature reduction
  - All original features are used
  - The transformed features are linear combinations of the original features
- Feature selection
  - Only a subset of the **original** features are selected
- Dimension Reduction Tutorial

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

  `http://www.cs.binghamton.edu/~lyu/SDM07/DR-SDM07.pdf`

Models of Feature Selection

Filter Model

- Separating feature selection from classifier learning
- Relying on general characteristics of data (information, distance, dependence, consistency)
- No bias toward any learning algorithm, fast

Models of Feature Selection

Filter Model

- Separating feature selection from classifier learning
- Relying on general characteristics of data (information, distance, dependence, consistency)
- No bias toward any learning algorithm, fast

Wrapper Model

- Relying on a predetermined classification algorithm
- Using predictive accuracy as goodness measure
- High accuracy, computationally expensive

Models of Feature Selection

Filter Model

- Separating feature selection from classifier learning
- Relying on general characteristics of data (information, distance, dependence, consistency)
- No bias toward any learning algorithm, fast

Wrapper Model

- Relying on a predetermined classification algorithm
- Using predictive accuracy as goodness measure
- High accuracy, computationally expensive

Embedded Method

- Takes advantage of its own variable selection algorithm
- Combine the advantages of both previous methods
- Relies on preliminary knowledge of a good selection.

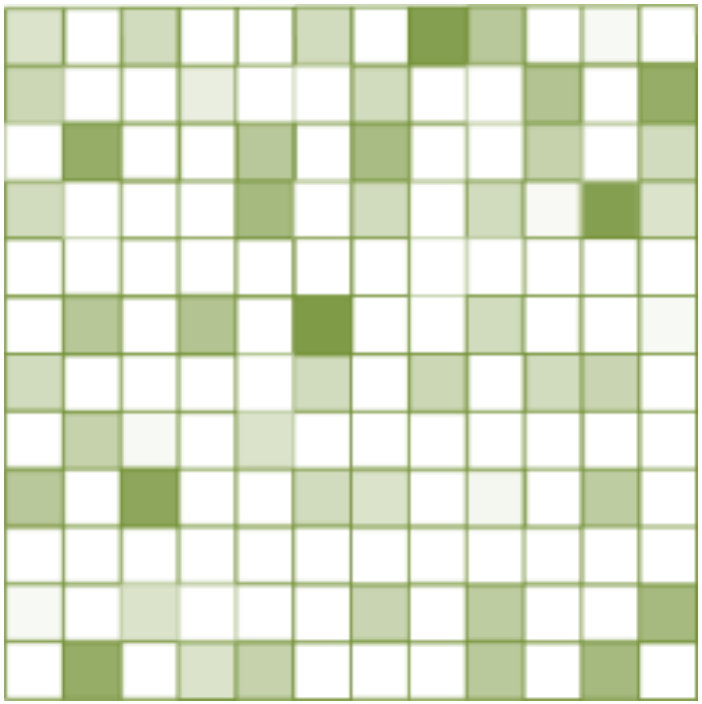
Sparse Learning

Why Sparse Learning

- **Embed** dimensionality reduction into data mining tasks
- Flexible models for complex feature **structures**
- Strong **theoretical** guarantee
- **Convex** formulations
- Empirical success in many **applications**
- Recent progress on **efficient** implementations

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.

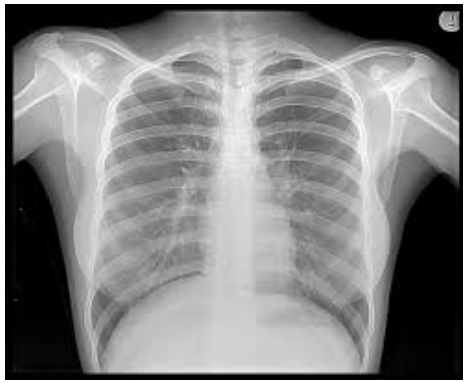


Human Anatomy

Anatomy Lesson of Dr. Nicolaes Tulp by Rembrandt van Rijn, 1632.

Biomedical Imaging

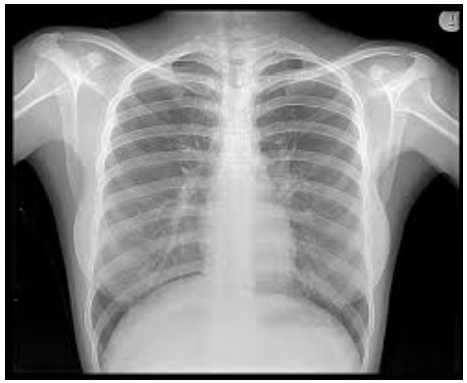
X-Ray,1895



1901 Nobel Prize in Physics
Wilhelm Röntgen's

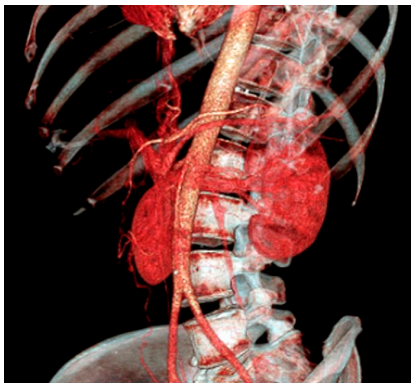
Biomedical Imaging

X-Ray,1895



1901 Nobel Prize in Physics
Wilhelm Röntgen's

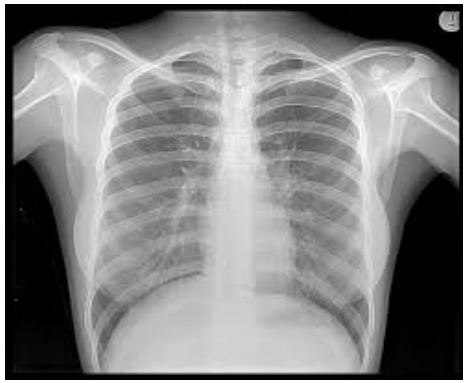
Computed Tomography (CT), 1967



1979 Nobel Prize in Physiology or Medicine
Allan M. Cormack and
Godfrey N. Hounsfield

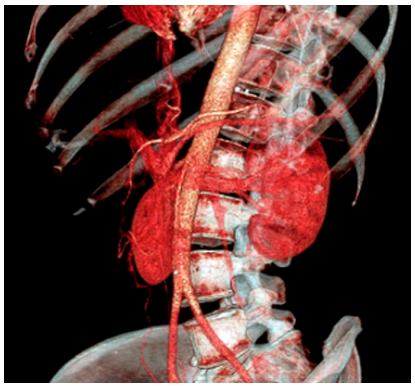
Biomedical Imaging

X-Ray,1895



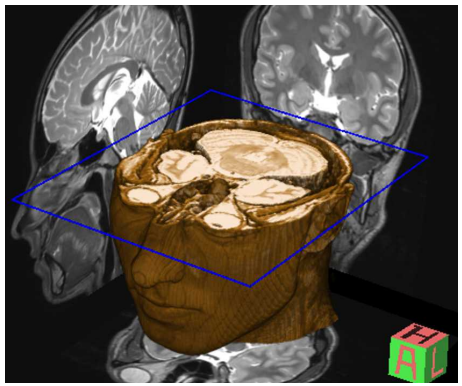
1901 Nobel Prize in Physics
Wilhelm Röntgen's

Computed Tomography (CT), 1967



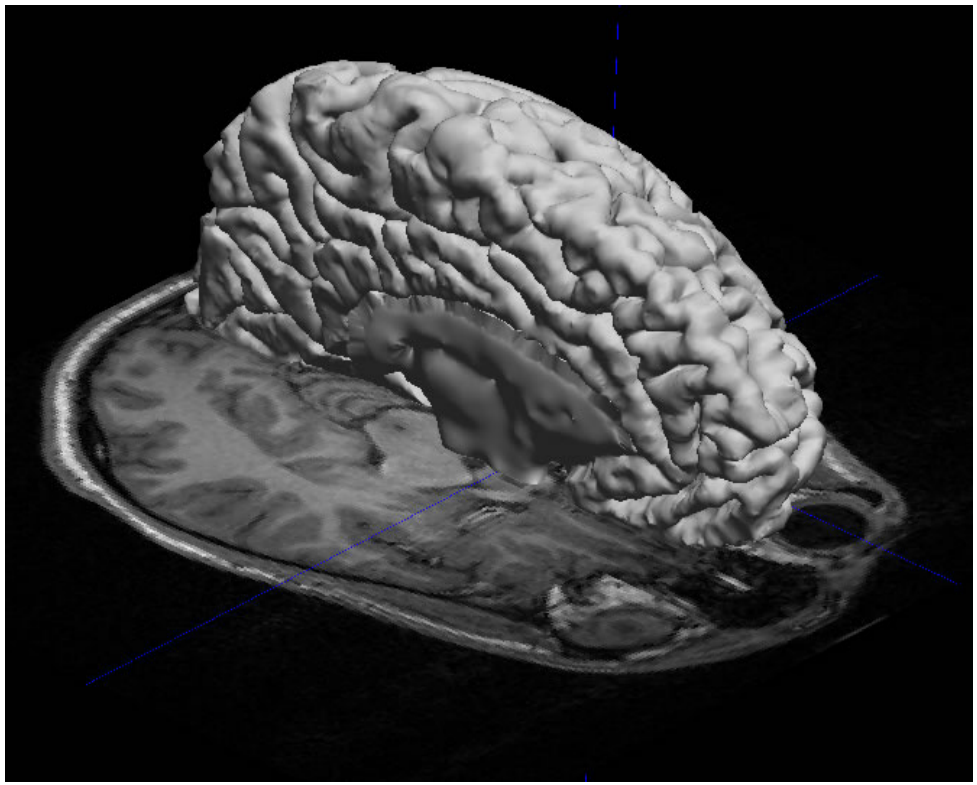
1979 Nobel Prize in Physiology or Medicine
Allan M. Cormack and Godfrey N. Hounsfield

Magnetic Resonance Imaging (MRI), 1971

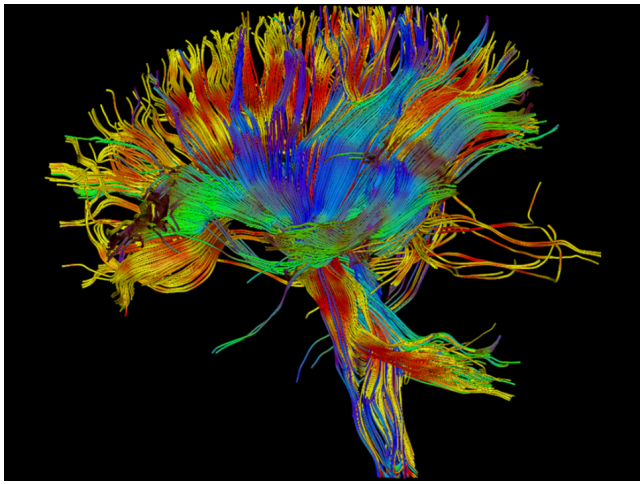


2003 Nobel Prize in Physiology or Medicine
Paul Lauterbur and Sir Peter Mansfield

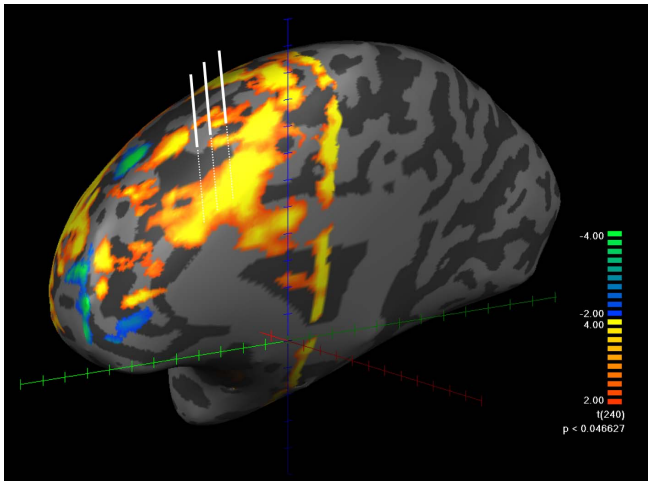
Magnetic Resonance Imaging



Structural



Diffusion



Functional

The Sensing Problem

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

$$y_i =< x, \varphi_i >, i = 1, 2, \ldots, n$$

  where waveforms φ_i are sinusoids, y is a vector of Fourier coefficients (e.g., measurements acquired by MRI)
- Recover the object from the measurements
  Solving a linear system of equations

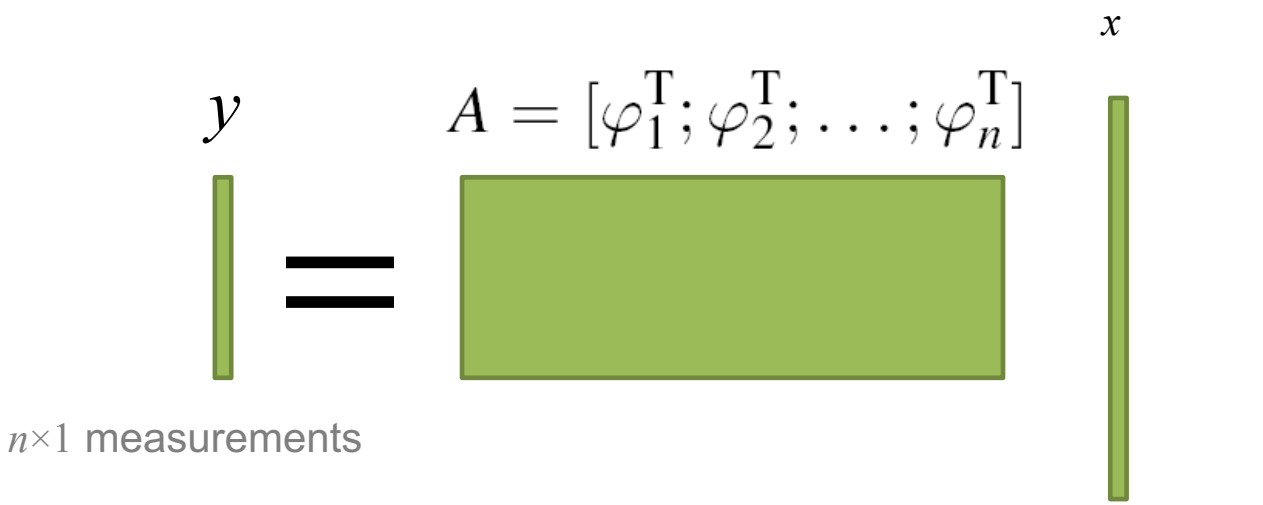
Magnetic Resonance Imaging (cont.)

Compressive Sensing?

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

Compressive Sensing?

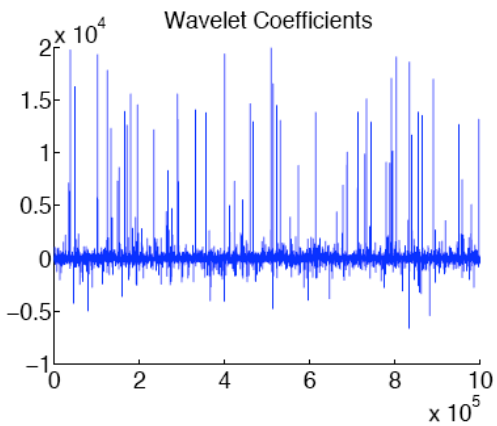
- Is accurate reconstruction possible from $n \ll p$ measurements only?
  - Few sensors
  - Measurements are very expensive
  - Sensing process is slow
  - Save lives
- Conventional wisdom: reconstruction is impossible
  - Number of measurements must match the number of unknowns



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

Compressive Sensing Works

- 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)



MRI by Compressive Sensing

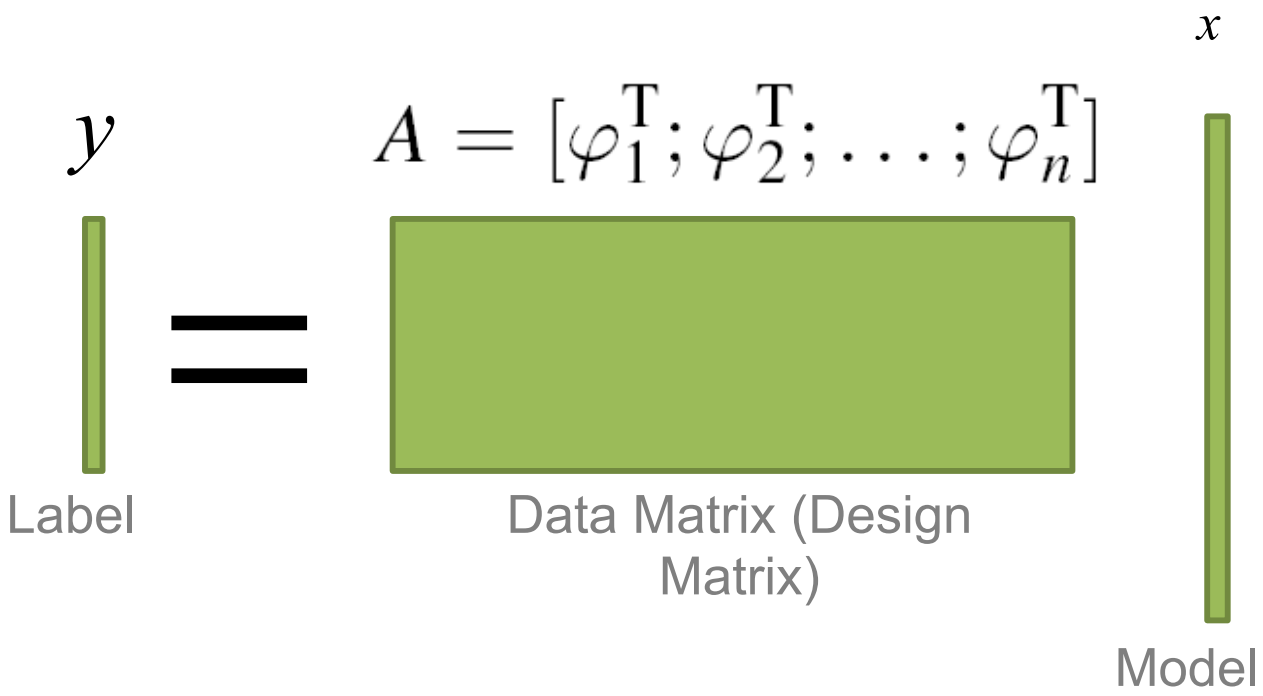
Sparsity

Dominant modeling tool in

- Genomics and Genetics
- Signal and audio processing
- Image processing
- Neuroscience (theory of sparse coding)
- **Machine learning and data mining**
- ...

Sparsity in Machine Learning

Regression, classification, collaborative filtering . . .



Convex Sparse Learning Models

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

$$\min_x \mathcal{L}(x) + \lambda \mathcal{R}(x)$$

- equivalent to the following model:

$$\min_x \mathcal{L}(x) \quad \text{s.t. } \mathcal{R}(x) \leq z$$

Theorem 1 in Efficient and Accurate ℓ_p-Norm Multiple Kernel Learning, NIPS 2009

Convex Sparse Learning Models

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

$$\min_x \mathcal{L}(x) + \lambda \mathcal{R}(x)$$

  - Sparsity via ℓ_1 and Eastic Net
  - Sparsity via ℓ_1/ℓ_q
  - Sparsity via Fused Lasso
  - Sparse Inverse Covariance Estimation
  - Sparsity via Trace Norm

Sparsity via Cardinality Minimization

- Many problems in scientific computing can be cast as

$$\min_x \text{Card}(x) \quad x \in \mathcal{P}$$

  where $\text{Card}(x)$. denotes the cardinality (nnz) of the vector x.
- Such problems seek a "sparse" solution, one with many zeroes in it.

Sparsity via Cardinality Minimization

- Many problems in scientific computing can be cast as

$$\min_x \text{Card}(x) \quad x \in \mathcal{P}$$

  where $\text{Card}(x)$. denotes the cardinality (nnz) of the vector x.
- Such problems seek a "sparse" solution, one with many zeroes in it.
- A related problem is a penalized version of the above, where we seek to trade-off an objective function against cardinality:

$$\min_x \mathcal{L}(x) + \lambda \text{Card}(x)$$

Sparsity via Cardinality Minimization

- Many problems in scientific computing can be cast as

$$\min_x \text{Card}(x) \quad x \in \mathcal{P}$$

  where $\text{Card}(x)$. denotes the cardinality (nnz) of the vector x.
- Such problems seek a "sparse" solution, one with many zeroes in it.
- A related problem is a penalized version of the above, where we seek to trade-off an objective function against cardinality:

$$\min_x \mathcal{L}(x) + \lambda \text{Card}(x)$$

- Cardinality minimization (aka ℓ_0 "norm") is a hard problem in general. It appears in many areas, such as classification.

Sparsity via ℓ_1-Norm Penalty

- We use the convex envelope of ℓ_0 norm: ℓ_1 norm:

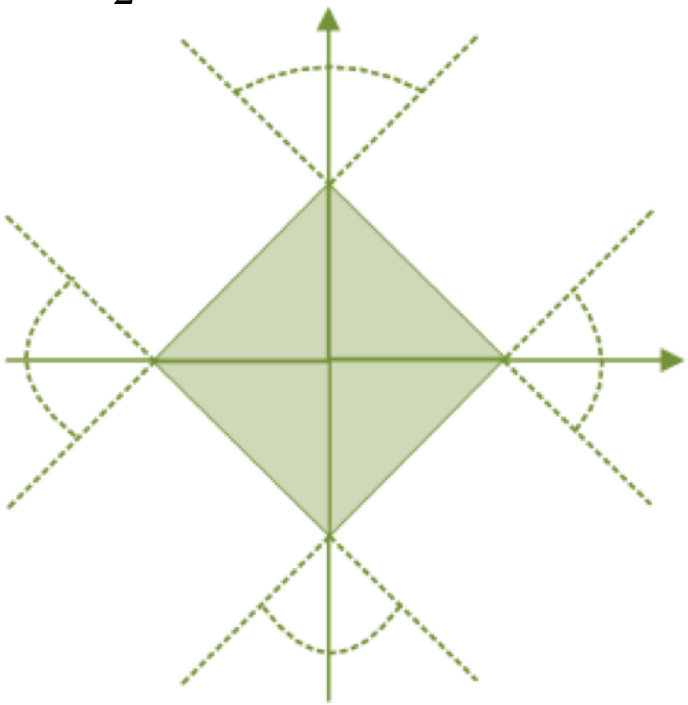
$$\min_x \mathcal{L}(x) + \lambda \|x\|_1$$

- Attracting properties
  - Sparsity induced norm
  - Valid norm
  - Convex
  - Computational tractable
  - Theoretical properties
  - Various extensions

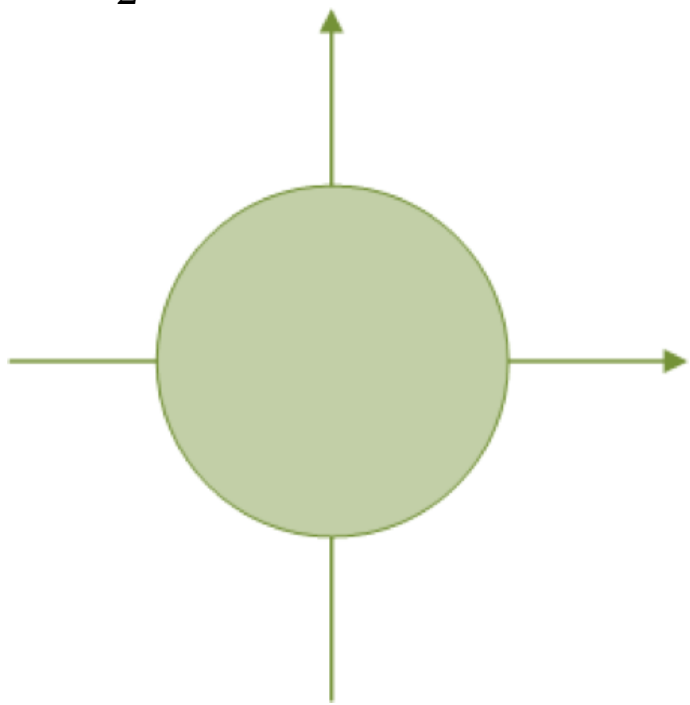
Why does ℓ_1-Norm Induce Sparsity?

Understand from projection

$\min_x \frac{1}{2}\|x - v\|^2$ s.t. $\|x\|_1 \leq 1$

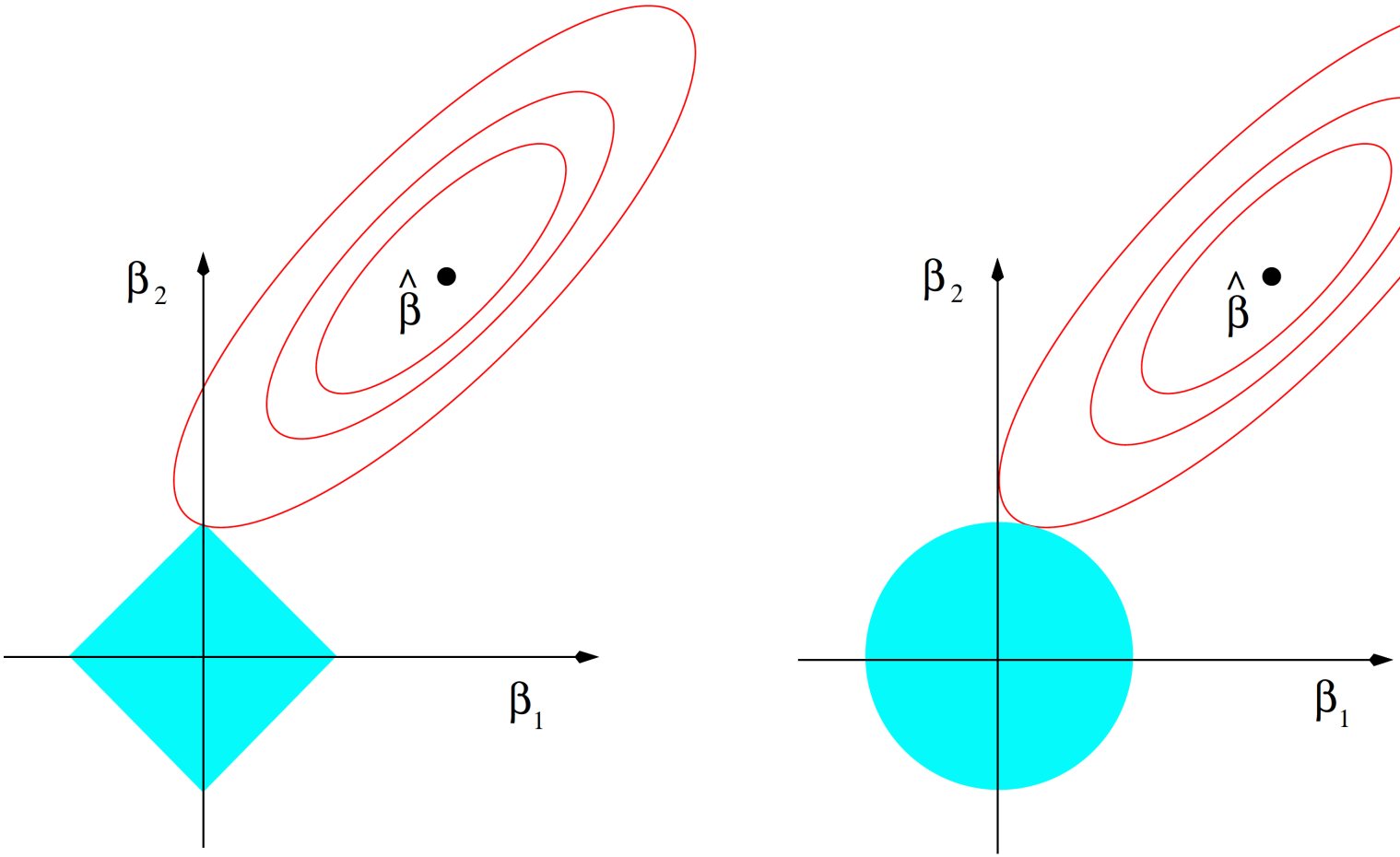


$\min_x \frac{1}{2}\|x - v\|^2$ s.t. $\|x\|_2 \leq 1$



Why does ℓ_1-Norm Induce Sparsity?

Understand from constrained optimization



Least Squares, Ridge and Lasso Regression

Under the orthogonal design ($A^T A = I$)

- Least Squares $\min_x \frac{1}{2}\|Ax - y\|_2^2$

Least Squares, Ridge and Lasso Regression

Under the orthogonal design ($A^T A = I$)

- Least Squares $\min_x \frac{1}{2}\|Ax - y\|_2^2$

$$x_{LS} = (A^T A)^{-1} A^T y = A^T y$$

Least Squares, Ridge and Lasso Regression

Under the orthogonal design ($A^T A = I$)

- Least Squares $\min_x \frac{1}{2}\|Ax - y\|_2^2$

$$x_{LS} = (A^T A)^{-1} A^T y = A^T y$$

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

Least Squares, Ridge and Lasso Regression

Under the orthogonal design ($A^TA = I$)

- Least Squares $\min_x \frac{1}{2}\|Ax - y\|_2^2$

$$x_{LS} = (A^TA)^{-1}A^Ty = A^Ty$$

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

$$x_{RR} = (A^TA + \lambda I)^{-1}A^Ty = A^Ty/(1+\lambda) = x_{LS}/(1+\lambda)$$

Least Squares, Ridge and Lasso Regression

Under the orthogonal design ($A^TA = I$)

- Least Squares $\min_x \frac{1}{2}\|Ax - y\|_2^2$

$$x_{LS} = (A^TA)^{-1}A^Ty = A^Ty$$

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

$$x_{RR} = (A^TA + \lambda I)^{-1}A^Ty = A^Ty/(1+\lambda) = x_{LS}/(1+\lambda)$$

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

Least Squares, Ridge and Lasso Regression

Under the orthogonal design ($A^TA = I$)

- Least Squares $\min_x \frac{1}{2}\|Ax - y\|_2^2$

$$x_{LS} = (A^TA)^{-1}A^Ty = A^Ty$$

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

$$x_{RR} = (A^TA + \lambda I)^{-1}A^Ty = A^Ty/(1+\lambda) = x_{LS}/(1+\lambda)$$

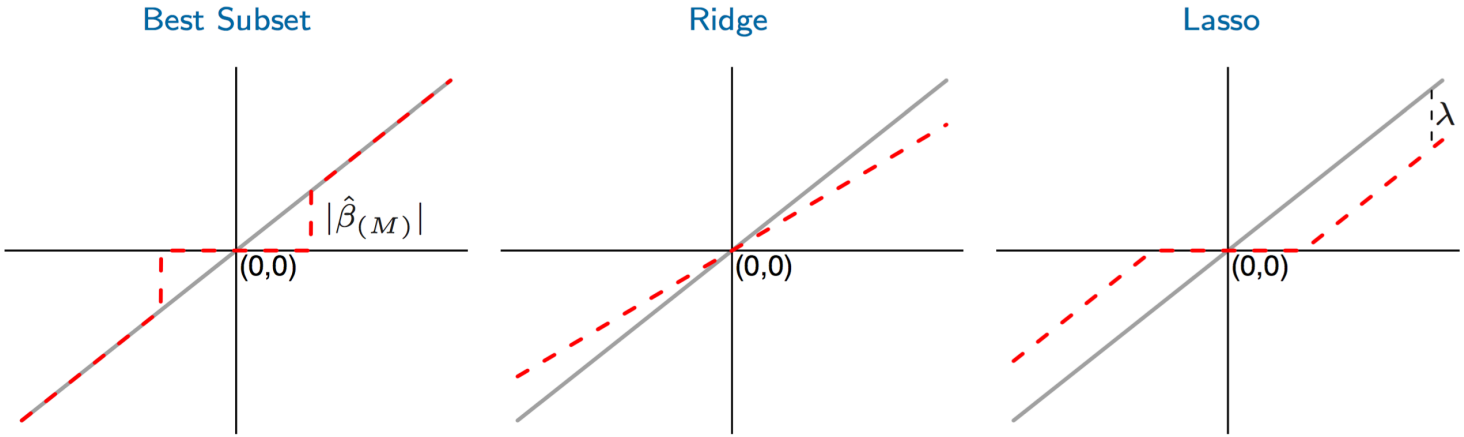
- Lasso Regression $\min_x \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_1$

$$x_{LA} = \frac{1}{2}\|x - A^Ty\|_F^2 + \lambda\|x\|_1 = \mathsf{sign}(x_{LS})\max(0, |x_{LS}| - \lambda)$$

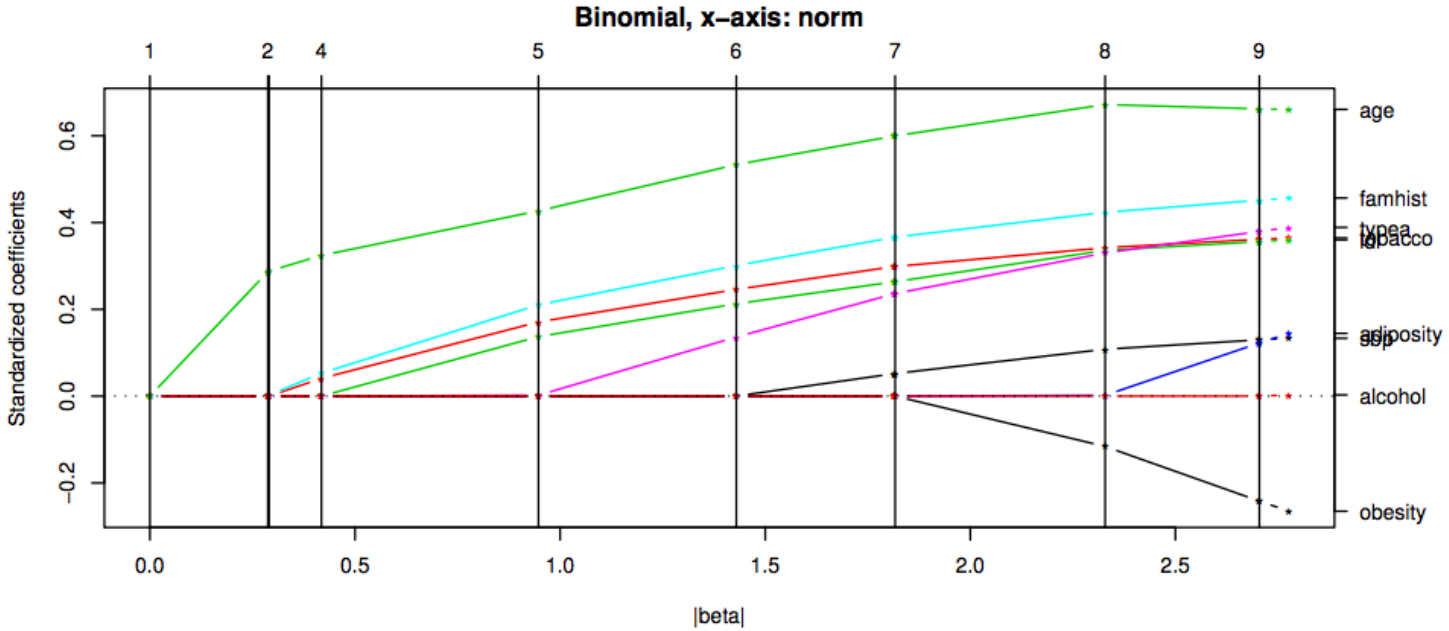
Least Squares, Ridge and Lasso Regression

Under the orthogonal design ($A^T A = I$)

- Best Subset Regression $x_{BS} = x_{LS} \cdot I(|x_{LS}| \geq |x_{(M)}|)$
- Ridge Regression $x_{RR} = x_{LS}/(1+\lambda)$
- Lasso Regression $x_{LA} = \text{sign}(x_{LS}) \max(0, |x_{LS}| - \lambda)$



Lasso Path



ℓ_1 regularization path in a constrained Lasso problem.

Elastic Net

Limitation of Lasso

- When $p > n$, the lasso selects at most n features before it saturates
- If there is a group of highly correlated features, then the lasso tends to select one features from a group and ignore the others
- Not stable in correlated design.

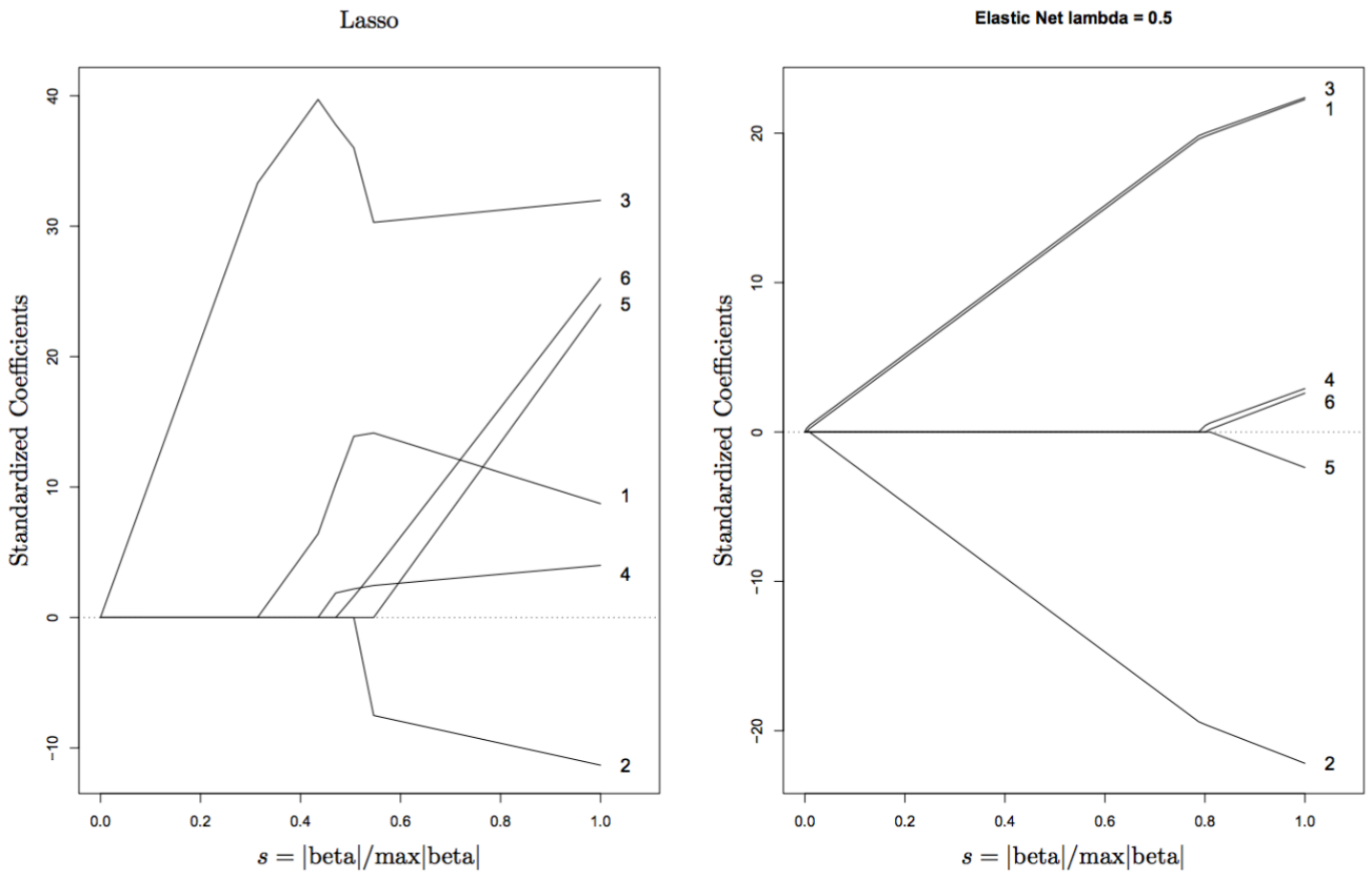
Elastic Net

- Linearly combines ℓ_1 and ℓ_2 penalties of the lasso and ridge

$$\min_x \|Ax - y\|_2^2 + \lambda_1 \|x\|_1 + \lambda_2 \|x\|_2^2$$

  - Removes the limitation on the number of selected variables
  - Encourages *grouping* effect
  - Stabilizes the ℓ_1 regularization path.

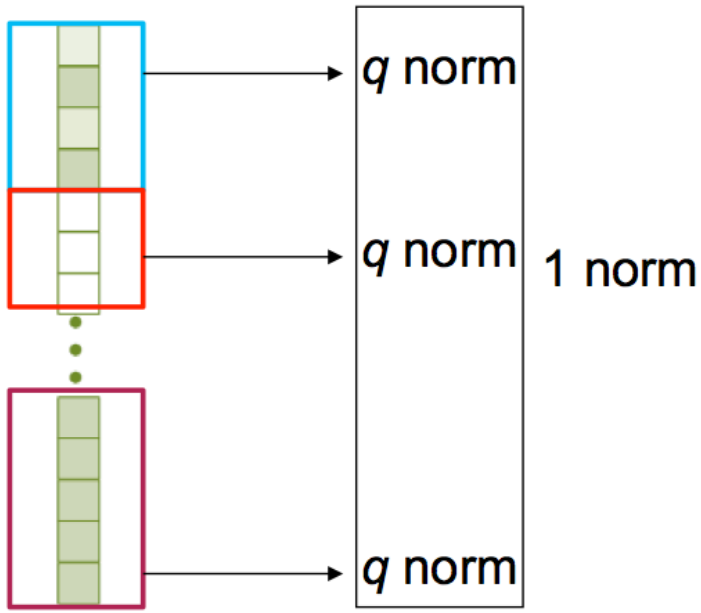
Elastic Net Regularization Path



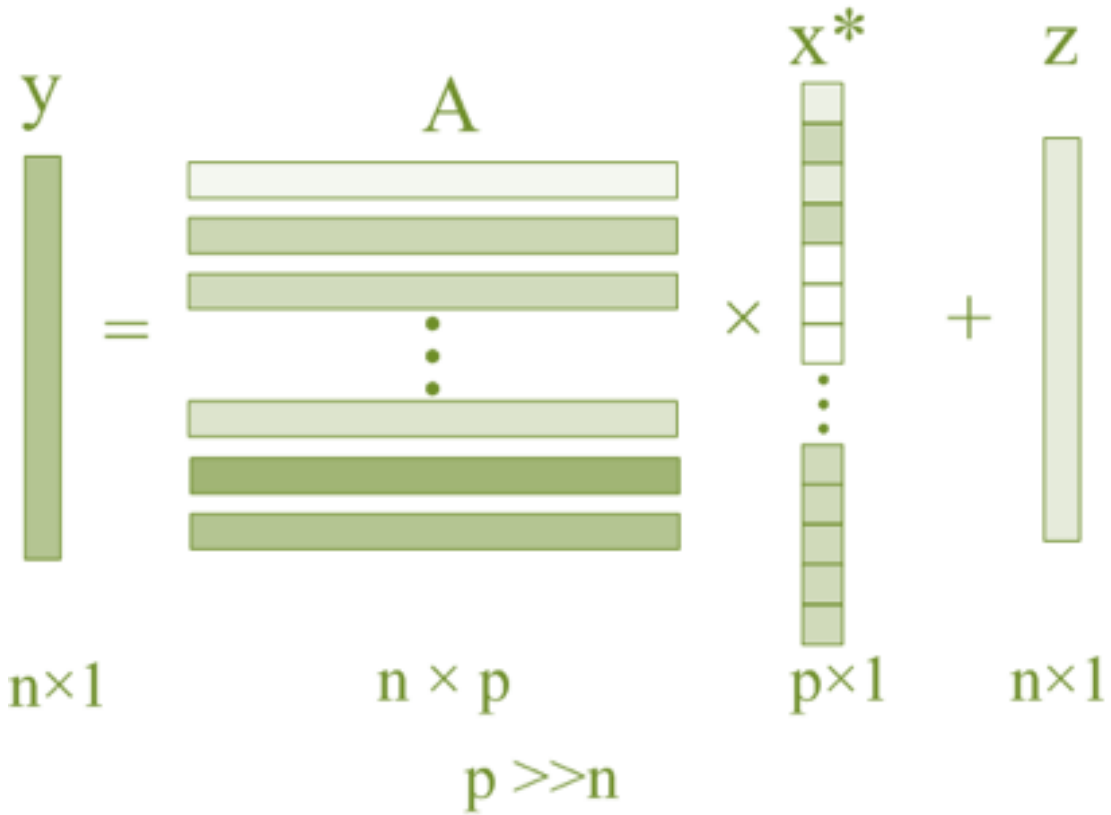
Advanced Topics on Sparsity

From ℓ_1 to ℓ_1/ℓ_q

- Pre-defined groups within the features
  - Genes that belong to the same biological pathway
  - Collections of indicator (dummy) variables
- Let $\mathcal{G}$ collectively denote a set of group assignments $\{G_1, \ldots, G_L\}$
- Define ℓ_1/ℓ_q norm as $\|x\|_{q,1} = \sum_{l=1}^{L} \|x_{G_i}\|_q$
  - Most existing work focus on $q = 2, \infty$
  - What about $q = 1$?

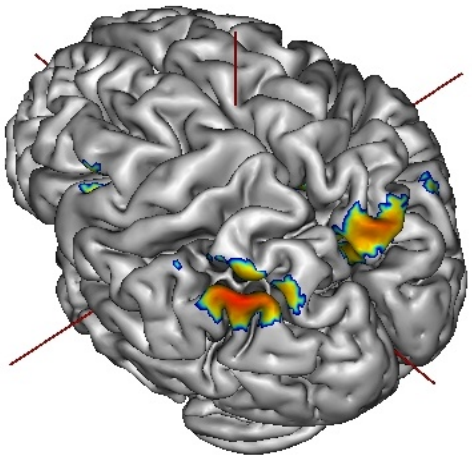


Group Lasso

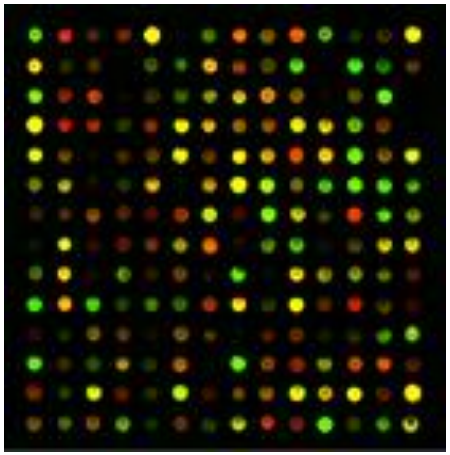


$$\min_x \frac{1}{2}\|Ax - y\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{|G_i|}\|x_{G_i}\|_2$$

Applications of Group Feature Selection



brain region



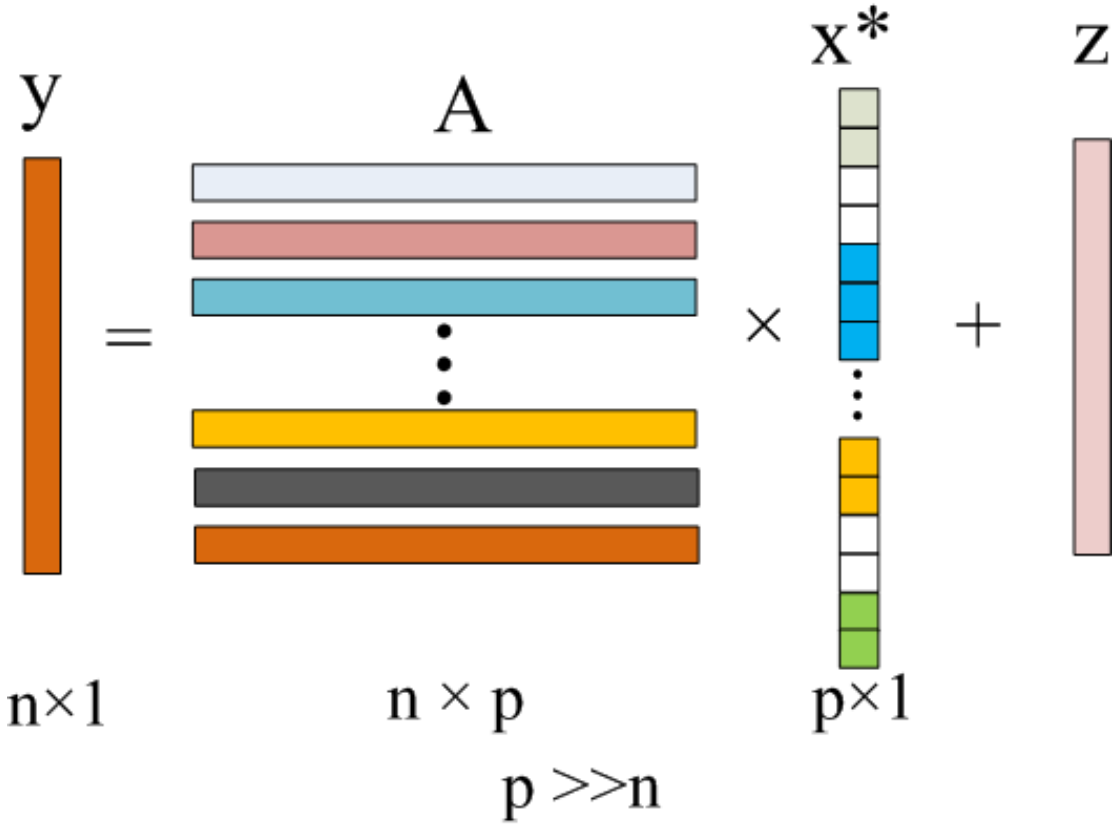
functional group

group

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

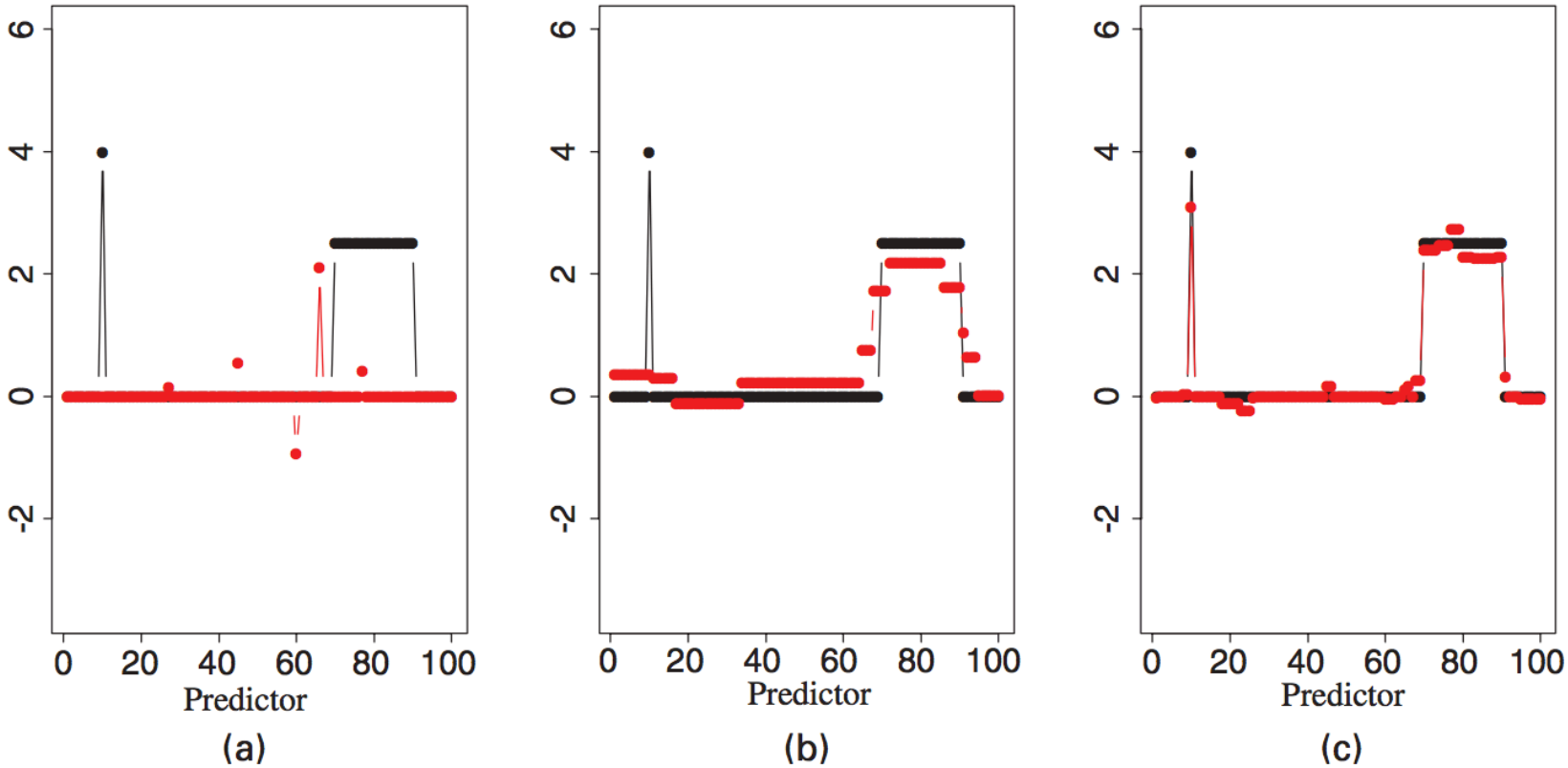
categorical variable

Fused Lasso



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

Fused Lasso Experiment



Many other extensions of sparsity

- Network Construction
- Matrix Completion
- Multi-Task Learning
- Sparse inverse covariance
- Complicated feature structures (graph, tree, etc.)
- ...

Unified Optimization Framework

$$\min_{x\in\mathbb{R}^d}\left\{f(x)=\mathcal{L}(x)+\mathcal{R}(x)=\frac{1}{n}\sum_{i=1}^{n}\ell_i(x)+\mathcal{R}(x)\right\}$$

Name	Loss function $\ell_i(x)$
Least squares	$\frac{1}{2}(y_i-a_i^Tx)^2$
Logistic regression	$\log(1+\exp(-y_ia_i^Tx))$
Squared Hinge Loss	$\max(0,1-y_ia_i^Tx)^2$

Name	Regularizer (penalty) $\mathcal{R}(x)$
Lasso	$\lambda\sum_{j=1}^{d}\lvert x_j\rvert$
Fused Lasso	$\lambda_1\sum_{j=1}^{d}\lvert x_j\rvert+\lambda_2\sum_{j=1}^{d-1}\lvert x_j-x_{j+1}\rvert$
Graph Fused Lasso	$\lambda_1\sum_{j=1}^{d}\lvert x_j\rvert+\lambda_2\sum_{(j,k)\in\mathcal{G}}\lvert x_j-x_k\rvert$
Group Lasso	$\lambda\sum_{k=1}^{K}\lVert x_{\mathcal{G}_k}\rVert$
Sparse Group Lasso	$\lambda_1\sum_{j=1}^{d}\lvert x_j\rvert+\lambda_2\sum_{k=1}^{K}\lVert x_{\mathcal{G}_k}\rVert$
Tree Lasso	$\sum_{j=1}^{J}\sum_{k=1}^{K_j}\lambda_k^j\lVert x_{\mathcal{G}_k^j}\rVert$

Gradient Descent for the Composite Model

(Nesterov, 2007; Beck and Teboulle, 2009)

- Optimization objective

$$\min_x f(x) = \mathcal{L}(x) + \lambda \mathcal{R}(x)$$

Gradient Descent for the Composite Model

(Nesterov, 2007; Beck and Teboulle, 2009)

- Optimization objective

$$\min_x f(x) = \mathcal{L}(x) + \lambda \mathcal{R}(x)$$

- At each iteration we construct a model

$$\mathcal{M}(x_i, \gamma_i) = [\mathcal{L}(x_i) + < \nabla \mathcal{L}(x_i), x - x_i >] + \frac{1}{2\gamma_i} \|x - x_i\|_2^2 + \lambda \mathcal{R}(x)$$

Terms: 1st order Taylor expansion, regularization, non-smooth part.
γ_i is obtained via a certain line search algorithm.

Gradient Descent for the Composite Model

(Nesterov, 2007; Beck and Teboulle, 2009)

- Optimization objective

$$\min_x f(x) = \mathcal{L}(x) + \lambda \mathcal{R}(x)$$

- At each iteration we construct a model

$$\mathcal{M}(x_i, \gamma_i) = [\mathcal{L}(x_i) + < \nabla \mathcal{L}(x_i), x - x_i >] + \frac{1}{2\gamma_i} \|x - x_i\|_2^2 + \lambda \mathcal{R}(x)$$

  Terms: 1st order Taylor expansion, regularization, non-smooth part.
  γ_i is obtained via a certain line search algorithm.
- Optimization algorithm
  - Repeat
  - $x_{i+1} = \arg\min \mathcal{M}(x_i, \gamma_i)$
  - Until convergence

  Convergence rate $O(1/N)$.

First Order Optimization

Proximal Gradient

$$
\begin{aligned}
x_{i+1} &= \arg\min_x \mathcal{L}(x_i) + \nabla\mathcal{L}(x_i)^T(x - x_i) + \frac{1}{2\gamma_i}\|x - x_i\|_2^2 + \lambda\mathcal{R}(x) \\
&= \arg\min_x \left\{ \frac{1}{2}\|x - (x_i - \gamma_i\nabla\mathcal{L}(x_i))\|^2 + \gamma_i\lambda\mathcal{R}(x) \right\} \\
&\equiv \text{Prox}_{\gamma_i}^{\lambda\mathcal{R}}(x_i - \gamma_i\nabla\mathcal{L}(x_i))
\end{aligned}
$$

First Order Optimization

Proximal Gradient

$$
\begin{aligned}
x_{i+1} &= \arg\min_x \mathcal{L}(x_i) + \nabla\mathcal{L}(x_i)^T(x - x_i) + \frac{1}{2\gamma_i}\|x - x_i\|_2^2 + \lambda\mathcal{R}(x) \\
&= \arg\min_x \left\{ \frac{1}{2}\|x - (x_i - \gamma_i\nabla\mathcal{L}(x_i))\|^2 + \gamma_i\lambda\mathcal{R}(x) \right\} \\
&\equiv \text{Prox}_{\gamma_i}^{\lambda\mathcal{R}}(x_i - \gamma_i\nabla\mathcal{L}(x_i))
\end{aligned}
$$

- Can be further accelerated with little extra costs
  - FISTA, SpaRSA
  - Convergence rate $O(1/N^2)$

First Order Optimization

Proximal Gradient

$$
\begin{aligned}
x_{i+1} &= \arg\min_x \mathcal{L}(x_i) + \nabla\mathcal{L}(x_i)^T(x - x_i) + \frac{1}{2\gamma_i}\|x - x_i\|_2^2 + \lambda\mathcal{R}(x) \\
&= \arg\min_x \left\{ \frac{1}{2}\|x - (x_i - \gamma_i\nabla\mathcal{L}(x_i))\|^2 + \gamma_i\lambda\mathcal{R}(x) \right\} \\
&\equiv \text{Prox}_{\gamma_i}^{\lambda\mathcal{R}}(x_i - \gamma_i\nabla\mathcal{L}(x_i))
\end{aligned}
$$

- Can be further accelerated with little extra costs
  - FISTA, SpaRSA
  - Convergence rate $O(1/N^2)$
- How to efficiently solve the proximal operator problem?
- Closed-form solution for ℓ_1, ℓ_1/ℓ_2, analytical form for trace norm

First Order Optimization

Proximal Gradient

$$
\begin{aligned}
x_{i+1} &= \arg\min_x \mathcal{L}(x_i) + \nabla\mathcal{L}(x_i)^T(x - x_i) + \frac{1}{2\gamma_i}\|x - x_i\|_2^2 + \lambda\mathcal{R}(x) \\
&= \arg\min_x \left\{ \frac{1}{2}\left\|x - (x_i - \gamma_i \nabla\mathcal{L}(x_i))\right\|^2 + \gamma_i\lambda\mathcal{R}(x) \right\} \\
&\equiv \text{Prox}_{\gamma_i}^{\lambda\mathcal{R}}(x_i - \gamma_i\nabla\mathcal{L}(x_i))
\end{aligned}
$$

- Can be further accelerated with little extra costs
  - FISTA, SpaRSA
  - Convergence rate $O(1/N^2)$
- How to efficiently solve the proximal operator problem?
- Closed-form solution for ℓ_1, ℓ_1/ℓ_2, analytical form for trace norm
- Can be extended to second order proximal optimization
  - Subproblem would require FISTA/SpaRSA

Stochastic Optimization

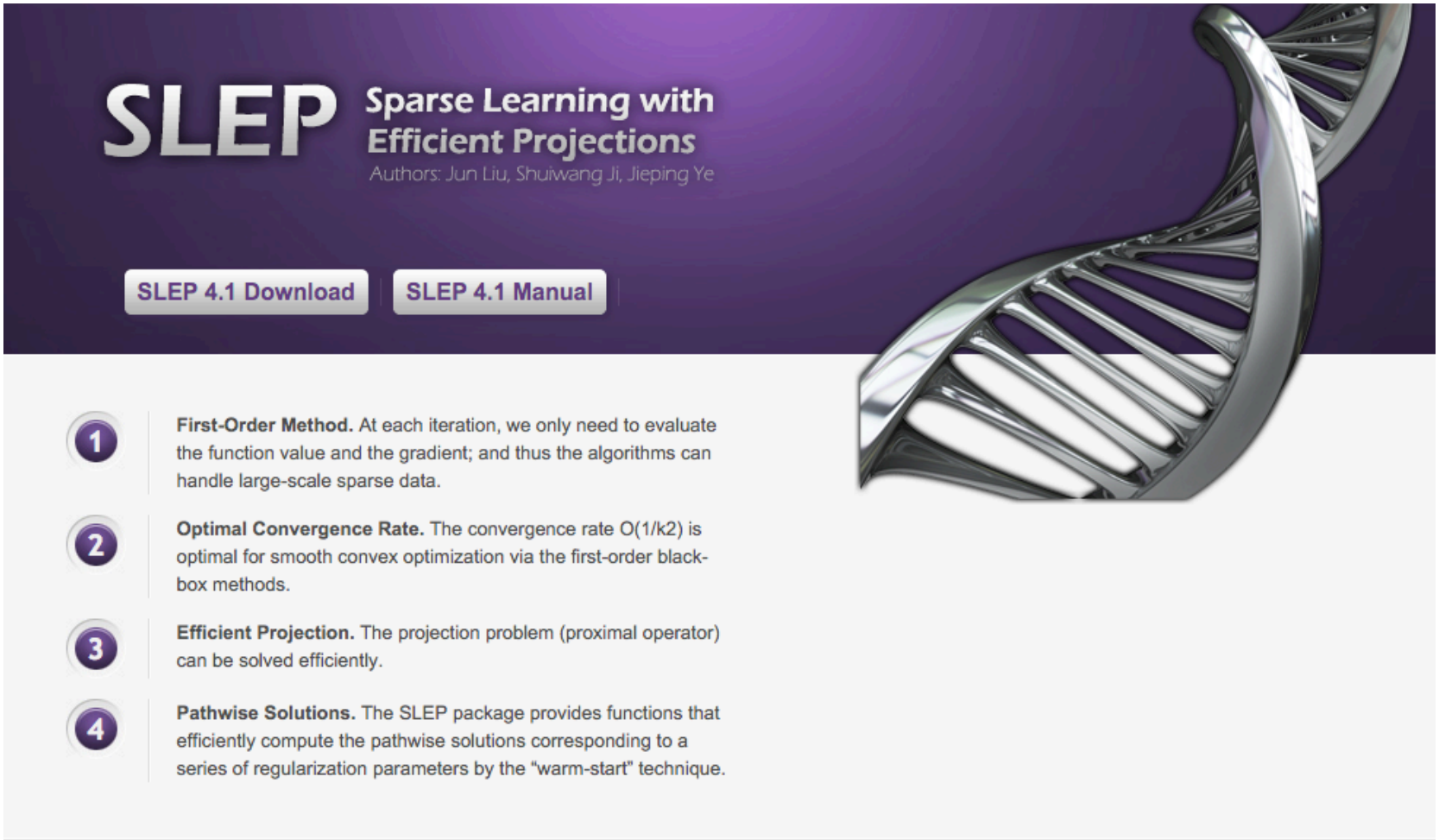
Stochastic Proximal Gradient

- Randomly pick a sample $j \in \{1, \ldots, n\}$
- Evaluate the gradient on the j-th sample $\nabla \ell_j(x_i)$ and generate a sequence $\{x_i\}$ via

$$\begin{aligned} x_{i+1} &= \arg\min_x \mathcal{L}(x_i) + \nabla \ell_j(x_i)^T (x - x_i) + \frac{1}{2\gamma_i} \|x - x_i\|_2^2 + \lambda \mathcal{R}(x) \\ &= \arg\min_x \left\{ \frac{1}{2} \|x - (x_i - \gamma_i \nabla \ell_j(x_i))\|^2 + \gamma_i \lambda \mathcal{R}(x) \right\} \\ &\equiv \text{Prox}_{\gamma_i}^{\lambda \mathcal{R}}(x_i - \gamma_i \nabla \ell_j(x_i)) \end{aligned}$$

- Can be extended to using mini-batch
- The use of an expected gradient may introduce variance in each iteration

Sparse Learning Package



Website: `http://www.yelab.net/software/SLEP/`

Github Version: `https://github.com/jiayuzhou/SLEP/`