Large-Scale Sparse Learning

Jieping Ye
Computer Science and Engineering
Center for Evolutionary Medicine and Informatics
Arizona State University
Sparsity has become an important modeling tool in **genomics, genetics, signal and audio processing, image processing, neuroscience (theory of sparse coding), machine learning, statistics** …
- Strong theoretical guarantees
- Empirical success in many applications
- Recent progress on efficient implementations of sparse learning models
- Flexible models for incorporating complex feature structures
Outline

Sparse Linear Regression (Lasso)

Structured Sparse Learning Models
  • Fused Lasso, Group Lasso, Tree Lasso, Graph Lasso

Multi-Source Sparse Learning

Sparse Gaussian Graphical Model

Model Selection
  • Stability Selection

The SLEP Package
  • Sparse Screening
Lasso/Basis Pursuit

(Tibshirani, 1996, Chen, Donoho, and Saunders, 1999)

\[ y = \begin{bmatrix} A \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} y \end{bmatrix} \]

\[ \min_{\|x\|_1} \|y - Ax\|_2^2 + \lambda \|x\|_1 \]

s.t. \( \|Ax - y\|_2 \leq \epsilon \)
Lasso/Basis Pursuit
(Tibshirani, 1996, Chen, Donoho, and Saunders, 1999)

\[ y = \begin{pmatrix} A \\ \vdots \\ \end{pmatrix} x + z \]

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

\[ \min \|x\|_1 \]

\[ \text{s.t.} \quad \|Ax - y\|_2 \leq \epsilon \]

Simultaneous feature selection and regression
Elucidate a Magnetic Resonance Imaging-Based Neuroanatomic Biomarker for Psychosis
Let $x$ be the model parameter to be estimated. A commonly employed model for estimating $x$ is

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

- Least squares loss
- Logistic loss
- Negative log likelihood

- $L_1$ Extensions
  - Convex, non-smooth
  - Induce desired structured sparsity
Structured Sparse Learning

- Group Lasso
- Fused Lasso
- Graph Lasso
- Tree Lasso
Graph Lasso

FGFS: Feature Grouping and Feature Selection Over an Undirected Graph

S. Yang, L. Yuan, X. Shen, V. Wonka, and J. Ye (2012)
Tree Lasso

Kim and Xing, 2010; Jenatton et al., 2010; Liu and Ye, 2010; Liu et al., 2012.
Tree Lasso (Cont’d)
Multi-Source Sparse Learning

Existing sparse models: Lasso, group Lasso, sparse group Lasso
Incomplete Multi-Source Fusion

PET
\[ P_{1}, \ldots, P_{116} \]

MRI
\[ M_{1}, M_{2}, \ldots, M_{305} \]

CSF
\[ C_{1}, \ldots, C_{5} \]

Subject_1
Subject_2
Subject_3
Subject_4
Subject_5
Subject_6
Subject_7
Subject_8
Subject_9
Subject_10
Subject_11
Subject_12
Subject_13
Subject_14
Subject_15
Subject_16
Subject_17
Subject_18
Subject_19
Subject_20
Subject_21
Subject_22
Subject_23
Subject_24
Subject_25
Subject_26
Subject_27
Subject_28
Subject_29
Subject_30
Subject_31
Subject_32
Subject_33
Subject_34
Subject_35
Subject_36
Subject_37
Subject_38
Subject_39
Subject_40
Subject_41
Subject_42
Subject_43
Subject_44
Subject_45
Subject_46
Subject_47
Subject_48
Subject_49
Subject_50
Subject_51
Subject_52
Subject_53
Subject_54
Subject_55
Subject_56
Subject_57
Subject_58
Subject_59
Subject_60
Subject_61
Subject_62
Subject_63
Subject_64
Subject_65
Subject_66
Subject_67
Subject_68
Subject_69
Subject_70
Subject_71
Subject_72
Subject_73
Subject_74
Subject_75
Subject_76
Subject_77
Subject_78
Subject_79
Subject_80
Subject_81
Subject_82
Subject_83
Subject_84
Subject_85
Subject_86
Subject_87
Subject_88
Subject_89
Subject_90
Subject_91
Subject_92
Subject_93
Subject_94
Subject_95
Subject_96
Subject_97
Subject_98
Subject_99
Subject_100
Subject_101
Subject_102
Subject_103
Subject_104
Subject_105
Subject_106
Subject_107
Subject_108
Subject_109
Subject_110
Subject_111
Subject_112
Subject_113
Subject_114
Subject_115
Subject_116
Subject_117
Subject_118
Subject_119
Subject_120
Subject_121
Subject_122
Subject_123
Subject_124
Subject_125
Subject_126
Subject_127
Subject_128
Subject_129
Subject_130
Subject_131
Subject_132
Subject_133
Subject_134
Subject_135
Subject_136
Subject_137
Subject_138
Subject_139
Subject_140
Subject_141
Subject_142
Subject_143
Subject_144
Subject_145
Subject_146
Subject_147
Subject_148
Subject_149
Subject_150
Subject_151
Subject_152
Subject_153
Subject_154
Subject_155
Subject_156
Subject_157
Subject_158
Subject_159
Subject_160
Subject_161
Subject_162
Subject_163
Subject_164
Subject_165
Subject_166
Subject_167
Subject_168
Subject_169
Subject_170
Subject_171
Subject_172
Subject_173
Subject_174
Subject_175
Subject_176
Subject_177
Subject_178
Subject_179
Subject_180
Subject_181
Subject_182
Subject_183
Subject_184
Subject_185
Subject_186
Subject_187
Subject_188
Subject_189
Subject_190
Subject_191
Subject_192
Subject_193
Subject_194
Subject_195
Subject_196
Subject_197
Subject_198
Subject_199
Subject_200
Subject_201
Subject_202
Subject_203
Subject_204
Subject_205
Subject_206
Subject_207
Subject_208
Subject_209
Subject_210
Subject_211
Subject_212
Subject_213
Subject_214
Subject_215
Subject_216
Subject_217
Subject_218
Subject_219
Subject_220
Subject_221
Subject_222
Subject_223
Subject_224
Subject_225
Subject_226
Subject_227
Subject_228
Subject_229
Subject_230
Subject_231
Subject_232
Subject_233
Subject_234
Subject_235
Subject_236
Subject_237
Subject_238
Subject_239
Subject_240
Subject_241
Subject_242
Subject_243
Subject_244
Subject_245
iMSF: Incomplete Multi-Source Fusion

Sparse Gaussian Graphical Model

When $S$ is invertible, directly maximizing the likelihood gives $X = S^{-1}$.

Log-likelihood

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

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

frontal, parietal, occipital, and temporal lobes in order

Model Selection

• How to choose a good $\lambda$?
  – $\min \ loss(x) + \lambda \times \text{penalty}(x)$

• Cross-validation is commonly applied for model selection.
  – Tend to select too many features
  – The correct value may not be in the candidate set

• **Stability selection** [Meinshausen, Bühlmann, 2010]
  – Subsampling/bootstrapping in the context of feature selection yields better and stable results
Stability Selection

\[ \Lambda = \{ \lambda_1, \lambda_2, \ldots, \lambda_M \} \]

N bootstrap samples
Stability Selection (Cont’d)

• Stability selection has strong theoretical guarantees
  – Weaker conditions for consistent feature selection

• Challenge:
  – Need to solve the sparse learning model NM times
    • N: bootstrap samples
    • M: number of parameter values

• Efficient algorithms for sparse learning are needed
  – The SLEP package supports pathwise solutions
## SLEP: A Sparse Learning Package

<table>
<thead>
<tr>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
</tr>
<tr>
<td>Fused Lasso</td>
</tr>
<tr>
<td>Group Lasso</td>
</tr>
<tr>
<td>Sparse Group Lasso</td>
</tr>
<tr>
<td>Tree Structured Group Lasso</td>
</tr>
<tr>
<td>Overlapping Group Lasso</td>
</tr>
<tr>
<td>Sparse Inverse Covariance Estimation</td>
</tr>
<tr>
<td>Trace Norm Minimization</td>
</tr>
</tbody>
</table>

[http://www.public.asu.edu/~jye02/Software/SLEP/]
More Efficiency?

- Very high dimensional data
- Non-smooth sparsity-induced norms
- Multiple runs in model selection
- A large number of runs in permutation test
Large-Scale Sparse Screening

- Standard Lasso

\[ \beta^* = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{2} \| y - X\beta \|_2^2 + \lambda \| \beta \|_1 \]

- Data Matrix: \( X \in \mathbb{R}^{n \times p} \). \( n \) is the number of observations. \( p \) is the number of features.
- Response Vector: \( y \in \mathbb{R}^n \).
- To fit the model, \( \beta^* \) needs to be estimated.
- Sparsity of \( \beta^* \) is equivalent to feature selection.

Screening Rule: Motivation

\[ y \in \mathbb{R}^n \quad x \in \mathbb{R}^{n \times p} \quad \beta^* \in \mathbb{R}^p \]

\[ \begin{array}{c}
\approx \\
\cdots \\
\vdots \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
\end{array} \]

\[ = \]

\[ \tilde{x} \in \mathbb{R}^{n \times (p-p_0)} \quad \tilde{\beta}^* \in \mathbb{R}^{(p-p_0)} \]

- \( p_0 \): the number of 0 components in \( \beta^* \).
- If \( \beta^* \) is sparse and we know the 0 components in advance, the size of the optimization problem can be significantly reduced.

Large-Scale Sparse Screening (Cont’d)

• Dual Formulation of Lasso

$$\sup_{\theta \in \mathbb{R}^n} \frac{1}{2} \| y \|_2^2 - \frac{\lambda^2}{2} \left\| \theta - \frac{y}{\lambda} \right\|_2^2$$

s.t. \hspace{1em} |x_i^T \theta| \leq 1, \hspace{1em} i = 1, 2, \cdots, p

- Feasible Set: \( \mathcal{F} = \{ \theta : |x_i^T \theta| \leq 1, i = 1, 2, \cdots, p \} \) is a polytope.

- Optimal Dual Solution: \( \theta_\lambda^* \) is the point inside \( \mathcal{F} \) which is closest to \( \frac{y}{\lambda} \), i.e., \( \theta_\lambda^* \) is the projection of \( \frac{y}{\lambda} \) onto \( \mathcal{F} \).
Large-Scale Sparse Screening (Cont’d)

- Geometric Intuition
  - KKT condition:
    \[
    (\theta_\lambda^*)^T x_i \in \begin{cases} 
    \text{sign}([\beta_\lambda^*]_i) & \text{if } [\beta_\lambda^*]_i \neq 0 \\
    [-1, 1] & \text{if } [\beta_\lambda^*]_i = 0
    \end{cases}
    \]
  - Testing Rule:
    \[|(\theta_\lambda^*)^T x_i| < 1 \Rightarrow [\beta_\lambda^*]_i = 0\]

Example:
\[\theta_\lambda^* = P_F \left( \frac{y}{\lambda} \right) \text{ is the point inside } F \]
which is closest to \( \frac{y}{\lambda} \).

\[|(\theta_\lambda^*)^T x_1| < 1 \Rightarrow [\beta_\lambda^*]_1 = 0\]

However, \( \theta_\lambda^* \) is unknown in general but can be estimated to be in a small region.
Large-Scale Sparse Screening (Cont’d)

Geometric Intuition (Cont.)

Non-expansiveness:

$$\| \theta_{\lambda''}^* - \theta_{\lambda'}^* \|_2 = \| P_{\mathcal{F}} \left( \frac{y}{\lambda''} \right) - P_{\mathcal{F}} \left( \frac{y}{\lambda'} \right) \|_2 \leq \| \frac{y}{\lambda''} - \frac{y}{\lambda'} \|_2 \equiv \varphi(\lambda', \lambda'') = r$$

**Theorem 1.** For the Lasso problem, assume we are given the solution of its dual problem $\theta^*_\lambda$ for a specific $\lambda'$. Let $\lambda''$ be a nonnegative value different from $\lambda'$, then the following holds:

$$|x_i^T \theta^*_\lambda| < 1 - \|x_i\|_2 \phi(\lambda', \lambda'') \Rightarrow [\beta^*_{\lambda''}]_i = 0$$

**Corollary 2. DPP:** For the Lasso problem, let $\lambda_{max} = \max_i |x_i^T y|$. If $\lambda > \lambda_{max}$, $\beta^*_\lambda = 0$. Otherwise,

$$|x_i^T \frac{y}{\lambda_{max}}| < 1 - \|x_i\|_2 \phi(\lambda_{max}, \lambda) \Rightarrow [\beta^*_\lambda]_i = 0$$

**Corollary 3. SDPP:** For the Lasso problem, suppose we are given a sequence of parameter values $\lambda_{max} = \lambda_0 > \lambda_1 > \cdots > \lambda_m$. Then for any integer $0 \leq k < m$, if $\beta^*_{\lambda_k}$ is known, the following holds:

$$|x_i^T \frac{y-x\beta^*_{\lambda_k}}{\lambda_k}| < 1 - \|x_i\|_2 \phi(\lambda_k, \lambda_{k+1}) \Rightarrow [\beta^*_{\lambda_{k+1}}]_i = 0$$
Evaluation

• 3D MRI-GM density map
  – 101 normal controls, 94 AD patients
  – Dimensionality: 262144

• Run pathwise Lasso using 81 parameter values between 0.2 to 1. [normalized scale]

• SLEP: 221.03 (s)
• SDPP+ SLEP: 2.89 (s)
Summary

• Structured sparse learning models
• Multi-source sparse learning
• Sparse Gaussian graphical model
• Model selection
  – Stability selection
• The SLEP package
• Sparse screening

• Sparse screening for other sparse models
• Sparse learning for big data