Genome-Transcriptome-Phenome Structured Association

02-715 Advanced Topics in Computational Genomics
Regression with Regularization

- **Group lasso** (Yuan and Lin, 2006)

\[
L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda \sum_{j=1}^{J} ||\beta_j||_{L1/L2}
\]

\[
||\beta_j||_{L1/L2} = \sqrt{\sum_k \beta_{jk}^2}
\]
Regression with Regularization (Group Lasso Penalty)

Lasso penalty

Group lasso penalty

$L_2$ penalty
Lasso (Tibshirani, 1996)

\[
\arg\min_k \sum_k (y_k - X\beta_k)' \cdot (y_k - X\beta_k) + \lambda \sum_k \sum_j |\beta_{jk}|
\]
$L_1/L_2$-regularized Multi-task Regression
(Obozinski et al., 2008)

$$\arg\min_k \sum_k (y_k - X\beta_k)' \cdot (y_k - X\beta_k) + \lambda \sum_j \sqrt{\sum_k \beta_{jk}^2}$$

Regression Coefficients

Inputs

Outputs
Tree-Guided Group Lasso

\[
\text{argmin} \sum_k (y_k - X \beta_k)' \cdot (y_k - X \beta_k) + \lambda \cdot \text{Tree}(\beta_1, \ldots, \beta_K)
\]

**Key idea:** use overlapping groups in group lasso

Tree-guided group lasso penalty
Tree-Guided Group Lasso

• In a simple case of two outputs

In a simple case of two outputs, a tree-guided group lasso approach is illustrated with two outputs $Y_1$ and $Y_2$. The figure shows different scenarios based on the height of the tree:

- **Low height**:
  - Tight correlation
  - Joint selection

- **Large height**:
  - Weak correlation
  - Separate selection

The diagram conveys that the height of the tree influences the correlation and selection strategy between the inputs and outputs.
Tree-Guided Group Lasso

- In a simple case of two outputs

\[ C_1 = \{\beta_{j1}, \beta_{j2}\} \]

Select the child nodes **jointly** or separately?

\[ \text{argmin } \sum_k (y_k - X\beta_k)' \cdot (y_k - X\beta_k) \]

\[ + \lambda \sum_j \left[ h(|\beta_{j1}| + |\beta_{j2}|) + (1 - h)\left(\sqrt{\beta_{j1}^2 + \beta_{j2}^2}\right) \right] \]

**L₁ penalty**
- Lasso penalty
- **Separate** selection

**L₂ penalty**
- Group lasso
- **Joint** selection

Elastic net
Tree-Guided Group Lasso

• For a general tree

\[ \text{argmin} \sum_k (y_k - X\beta_k)' \cdot (y_k - X\beta_k) \]

\[ + \lambda \sum \left[(1 - h_2)(\sqrt{\beta_{j1}^2 + \beta_{j2}^2 + \beta_{j3}^2}) + h_2(|C_1| + |\beta_{j3}|)\right] \]

\[ (1 - h_1)(\sqrt{\beta_{j1}^2 + \beta_{j2}^2}) + h_1(|\beta_{j1}| + |\beta_{j2}|) \]

**Select the child nodes jointly or separately?**

Note that the groups overlap!
Overlapping Groups in Tree-guided Group Lasso

Proposition 1. For each of the $k$th output, the sum of the weights $w_v$ for all nodes $v \in V$ in $T$ whose group $G_v$ contains the $k$th output as a member equals one. In other words, the following holds:

$$\sum_{v : k \in G_v} w_v = \prod_{m \in \text{Ancestors}(v_k)} s_m + \sum_{l \in \text{Ancestors}(v_k)} g_l \prod_{m \in \text{Ancestors}(v_l)} s_m = 1.$$
Overlapping Groups

• Previously
  • Arbitrarily overlapping groups (Jenatton, Audibert, Bach, 2009)
  • Overlapping groups over tree-structured inputs (Zhao, Roach, Yu, 2008)

Unbalanced penalization
Unit Contour Surface for Various Penalty Functions

\[ C_2 = \{ \beta_{j1}, \beta_{j2}, \beta_{j3} \} \]

\[ C_1 = \{ \beta_{j1}, \beta_{j2} \} \]

- Lasso
- \( L_1/L_2 \)

Tree
- \( g_1 = 0.5, g_2 = 0.5 \)
- \( g_1 = 0.2, g_2 = 0.7 \)
- \( g_1 = 0.7, g_2 = 0.2 \)
Illustration with Simulated Data

True regression coefficients

Lasso

$L_1/L_2$-regularized multi-task regression

Tree-guided group lasso

Outputs (Genes)

Inputs (SNPs)

High association

No association
Analysis of Yeast Data

Hierarchical clustering tree for genes (outputs)

Lasso

$L_1/L_2$-regularized multi-task regression

Tree-guided group lasso
Dynamic Trait (d-trait) Association

Stationary Trait Association

- Stationary Trait
- Genotypes: ATCGATTTCCATA
- Association Strengths: Red dot
- One-time genetic effect

Dynamic Trait Association

- Dynamic Trait (e.g., drug response)
- Genotypes: ATCGATTTCCATA
- Association Strengths: Red dots with varying sizes
- Genetic effects active over time with possibly varying effect size over time

<table>
<thead>
<tr>
<th>t = 1</th>
<th>t = 2</th>
<th>t = 3</th>
<th>t = T</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Temporally-Smoothed Lasso

- **Step 1: Autoregressive Model**
  - Captures the shape of the temporal trend in the d-trait data
  - Estimates the model parameters based on the d-trait data only

- **Step 2: Temporally-Smoothed Lasso**
  - Penalized regression framework
  - Incorporates the estimated d-trait shape parameters from Step 1
  - Detects time-varying genetic effects on the d-trait
Step 1: Autoregressive Model

Autoregressive Model:

\[ y_{k,t+1} = \alpha_{k,t} y_{k,t} + \alpha_{k,t}^0 \mathbf{1} + \epsilon \]

Estimating Model Parameters:

\[ \hat{\alpha}_{k,t} = \arg\min (y_{k,t+1} - \alpha_{k,t} y_{k,t} - \alpha_{k,t}^0)^T \cdot (y_{k,t+1} - \alpha_{k,t} y_{k,t} - \alpha_{k,t}^0) \]

Estimates of the Model Parameters:

\[ \hat{\alpha}_{k,t} = \frac{y_{k,t}^T \cdot y_{k,t+1}}{y_{k,t}^T \cdot y_{k,t}} \]
Step 2: Temporally-Smoothed Lasso

\[ \hat{B}^{\text{dyn}} = \arg \min_{\beta} \sum_{k} \sum_{t} (y_{k,t} - X_i \beta_{i,t})^T \cdot (y_{k,t} - X_i \beta_{i,t}) + \lambda \sum_{k} \sum_{t} \sum_{j} \beta_{k,t}^j + \gamma \sum_{j} \sum_{k} \sum_{t=1}^{T-1} |\beta_{k,t+1}^j - \alpha_{k,t}^j \beta_{k,t}^j| \]
Simulation Study – Linear Dynamic

Trait data for all individuals

Trait data for individuals with no association SNPs

Trait data for individuals with 1-2 association SNPs

Trait data for individuals with >3 association SNPs

True association strength

Estimated association strength (single SNP analysis)

Estimated association strength (lasso)

Estimated association strength (temporally-smoothed lasso)
Simulation Study – Cyclic Dynamic

Trait data for all individuals

Trait data for individuals with no association SNPs

Trait data for individuals with 1-2 association SNPs

Trait data for individuals with >3 association SNPs

True association strength

Estimated association strength (single SNP analysis)

Estimated association strength (lasso)

Estimated association strength (temporally-smoothed lasso)
**Genome-Transcriptome-Phenome Structured Association**

- Various regularization functions can be used within the sparse regression framework to enforce a sparsity pattern that reflects prior knowledge
  - Tree-guided group lasso
    - Hierarchical clustering tree – prior knowledge on gene clusters
  - Temporally smoothed lasso
    - Dynamic traits – prior knowledge on temporal correlation