Graph-constrained overlapping group lasso for case-control studies

Kipoong Kim[a], Hokeun Sun[b,*]

[a] *Department of Statistics, Changwon National University, 51140 Changwon, Korea*
[b] *Department of Statistics, Pusan National University, 46241 Busan, Korea*

Abstract

Keywords:

---

1. Introduction

2. Statistical Methods and Estimation

2.1. Problem Formulation

Model:

- Network structure among predictors:

  $\{\mathbf{X}_j : j = 1, 2, \ldots, p\} \sim \mathcal{G}$, where $\mathcal{G}$ is a network among genetic pathways.

- if y_i is gaussian-valued,

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

- if y_i is binary-valued,

  $y_i \sim Bernoulli(p_i)$ and $f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i}$.

  Link function: $g(\mathbb{E}Y_i) = g(p_i) = \mathbf{x}_i^T\boldsymbol{\beta}$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $\boldsymbol{\beta} \in \mathbb{R}^p$.

---

*Corresponding author

Email addresses: statpng@changwon.ac.kr (Kipoong Kim), hsun@pusan.ac.kr (Hokeun Sun)

2.2. Overlapping group lasso

2.3. Network-constrained regularization

Objective function

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\; L(\mathbf{y},\mathbf{X}\boldsymbol{\beta}) + \lambda_1 \sum_{k=1}^{K} \|\boldsymbol{\beta}^{(k)}\|_2 + \frac{\lambda_2}{2}\sum_{u\sim v}\left\{\frac{\|\boldsymbol{\beta}^{(u)}\|_2}{\sqrt{d_u}} - \frac{\|\boldsymbol{\beta}^{(v)}\|_2}{\sqrt{d_v}}\right\}^2$$

- For gaussian-valued response,

$$L(\mathbf{y},\mathbf{X}\boldsymbol{\beta}) = \frac{1}{n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$$

$\nabla L(\tilde{\boldsymbol{\beta}}|\mathbf{D}) = -\frac{1}{n}\mathbf{X}^T(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})$

- For binary-valued response,

$$L(\mathbf{y},\mathbf{X}\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\{y_i \log\frac{p_i}{1-p_i} + \log(1-p_i)\} = \frac{1}{n}\sum_{i=1}^{n}\{y_i\mathbf{x}_i^T\boldsymbol{\beta} - \log(1+\exp(\mathbf{x}_i^T\boldsymbol{\beta}))\}$$

We can solve the above optimization problem by using the group-lasso algorithm proposed by Yang and Zou (2015).

2.4. Groupwise majorization descent

Let $\mathbf{D} = (\mathbf{X},\mathbf{y})$ and the objective function is written by

$$Q(\boldsymbol{\beta}|\mathbf{D}) = L(\boldsymbol{\beta};\mathbf{D}) + \lambda_1 \sum_{k=1}^{K}\|\boldsymbol{\beta}^{(k)}\|_2 + \lambda_2 \sum_{u\sim v}\left\{\frac{\|\boldsymbol{\beta}^{(u)}\|_2}{\sqrt{d_u}} - \frac{\|\boldsymbol{\beta}^{(v)}\|_2}{\sqrt{d_v}}\right\}^2$$

By quadratic majorization,

$$L(\boldsymbol{\beta}|\mathbf{D}) \le L(\tilde{\boldsymbol{\beta}}|\mathbf{D}) + (\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T\nabla L(\tilde{\boldsymbol{\beta}}|\mathbf{D}) + \frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T\tilde{\mathbf{H}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) := L^*(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}|\mathbf{D})$$

where $\nabla L(\tilde{\boldsymbol{\beta}}|\mathbf{D}) = \left.\frac{\partial L(\boldsymbol{\beta}|\mathbf{D})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}$ and $\tilde{\mathbf{H}} = \left.\frac{\partial^2 L(\boldsymbol{\beta}|\mathbf{D})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}$

Here, we consider the u-th group only:

$$L^*(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}|\mathbf{D}) = L(\tilde{\boldsymbol{\beta}}^{(u)}|\mathbf{D}) + (\boldsymbol{\beta}^{(u)}-\tilde{\boldsymbol{\beta}}^{(u)})^T\nabla L(\tilde{\boldsymbol{\beta}}|\mathbf{D}) + \frac{1}{2}(\boldsymbol{\beta}^{(u)}-\tilde{\boldsymbol{\beta}}^{(u)})^T\tilde{H}^{(u)}(\boldsymbol{\beta}^{(u)}-\tilde{\boldsymbol{\beta}}^{(u)})$$

where $\tilde{H}^{(u)} = \frac{1}{n}\mathbf{X}^{(u)T}\mathbf{X}^{(u)} \in \mathbb{R}^{d_u \times d_u}$ is the submatrix of $H = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ for the u-th group.

We minimize $L^*(\boldsymbol{\beta}^{(u)} - \tilde{\boldsymbol{\beta}}^{(u)}|\mathbf{D}) + \lambda_1\|\boldsymbol{\beta}^{(u)}\|_2 + \frac{\lambda_2}{2}$

2.5. Overlapping Group Lasso

Model:

$$\begin{aligned} y_i &= \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i. \\ &= \beta_0 + \sum_{g=1}^{G}\sum_{k\in\mathcal{A}_g} x_{ij}\gamma_j \end{aligned}$$

where $\sum_{g=1}^{G}\gamma_j^{(g)} = \beta_j$, $\boldsymbol{\beta} = \sum_{g=1}^{G}\boldsymbol{\gamma}^{(g)}$, and $\boldsymbol{\Gamma} = \left[\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \cdots, \boldsymbol{\gamma}^{(G)}\right] \in \mathbb{R}^{p\times G}$ for $\boldsymbol{\beta}, \boldsymbol{\gamma}^{(g)} \in \mathbb{R}^p$.

2.6. Tuning parameter selection

3. Theoretical Results

4. Simulation

4.1. Comparing Methods

1. GroupLasso + Overlapping

$$\arg\max_{\boldsymbol{\Gamma}} L(\boldsymbol{\Gamma}|\mathbf{D}) + \lambda\sum_{g=1}^{G}\sqrt{d_g}\|\boldsymbol{\gamma}^{(g)}\|_2$$

2. GroupLasso + Network

$$\arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}|\mathbf{D}) + \lambda_1\sum_{g=1}^{G}\sqrt{d_g}\left\|\boldsymbol{\beta}_g\right\|_2 + \lambda_2\boldsymbol{\beta}^T\mathbf{L}\boldsymbol{\beta}$$

3. GroupLasso + Overlapping + Network

$$\arg\max_{\boldsymbol{\Gamma}} \mathcal{L}(\boldsymbol{\Gamma}|\mathbf{D}) + \lambda_1\sum_{g=1}^{G}\sqrt{d_g}\left\|\boldsymbol{\gamma}^{(g)}\right\|_2 + \lambda_2\sum_{u\sim v}\left\{\frac{\left\|\gamma_{(u)}\right\|_2}{\sqrt{d_u}} - \frac{\left\|\gamma_{(v)}\right\|_2}{\sqrt{d_v}}\right\}^2$$

5. Real Data Applications

6. Discussion

Acknowledgments

This work was supported by Samsung Science and Technology Foundation under Project Number SSTF-BA2002-03 as well as by the Bio-Synergy Research Project (NRF-2017M3A9C4065964) of the Ministry of Science, ICT and Future Planning through the National Research Foundation.

Appendix A. Conditions for the consistency of generalized Fréchet means

Recall that in Section ??, $x \in T$, $M \in (M, d)$, $M_0 \in \mathcal{H}(M)$, $M_n : (\Omega, \mathcal{A}, \mathcal{P}) \to \mathcal{H}(M)$ is a random sequence of closed subsets of M, $\mathcal{H}(M)$ is the collection of all nonempty closed subsets of M, and $\mathfrak{c} : T \times M \to \mathbb{R}$ is a loss function.

Appendix B. Supplementary material

Supplementary material related to this article can be found online.

Declarations

Competing interests

The authors declare that they have no competing interests or personal relationships that could have appeared to influence the work reported in this paper.

References