Trace lasso: a trace norm regularization for correlated designs

Edouard Grave; Guillaume Obozinski;

Francis Bach

Presented by
Zhengming Xing

Outline

- Introduction
- Definition and properties of trace lasso
- Approximation around lasso
- Optimization algorithm
- Experiments

Introduction

Main idea: introduce the trace norm that interpolate between l1-norm and l2-norm depending on the correlation.

Linear model

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$$

Estimate $\mathbf{w}$

$$\hat{\mathbf{w}} \in \underset{\mathbf{w}}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathbf{w}^\top \mathbf{x}_i) + \lambda f(\mathbf{w})$$

Penalty term

- Lasso
- Ridge regression
- Group lasso
- Trace lasso

Introduction

Notations

$\mathbf{M}^{(i)}$	ith column of matrix **M**
$\mathbf{M}_i$	ith row of matrix **M**
$\mathrm{diag}(\mathbf{M})$	The diagonal of matrix **M**
$\mathrm{Diag}(\mathbf{u})$	diagonal matrix whose diagonal element is u_i
$\|\mathbf{M}\|_*$	the trace norm. the sum of the singular value of **M**
$\|\mathbf{M}\|_F$	The Frobenius norm

$$\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{p} \|\mathbf{M}^{(i)}\|_2$$

Definition of Trace lasso

Trace lasso

$$\Omega(\mathbf{w}) = \|\mathbf{X}\,\mathrm{Diag}(\mathbf{w})\|_*$$

Sum of the singular value of the matrix

Decomposition

$$\mathbf{X}\,\mathrm{Diag}(\mathbf{w}) = \sum_{i=1}^{p} \left(\|\mathbf{X}^{(i)}\|_2 w_i\right) \frac{\mathbf{X}^{(i)}}{\|\mathbf{X}^{(i)}\|_2} \mathbf{e}_i^\top$$

Case 1

If predictor $\mathbf{X}^{(i)}$ is orthogonal

$$\|\mathbf{X}\,\mathrm{Diag}(\mathbf{w})\|_* = \sum_{i=1}^{p} \|\mathbf{X}^{(i)}\|_2 |w_i| = \|\mathbf{X}\,\mathrm{Diag}(\mathbf{w})\|_{2,1}$$

Case 2

If all the predictor are equal to $\mathbf{X}^{(1)}$

$$\mathbf{X}\,\mathrm{Diag}(\mathbf{w}) = \mathbf{X}^{(1)}\mathbf{w}^\top$$

$$\|\mathbf{X}\,\mathrm{Diag}(\mathbf{w})\|_* = \|\mathbf{X}^{(1)}\|_2 \|\mathbf{w}\|_2 = \|\mathbf{X}\,\mathrm{Diag}(\mathbf{w})\|_F$$

Definition of Trace lasso

Family of penalty function

$$\Omega_{\mathbf{P}}(\mathbf{w}) = \|\mathbf{P}\,\mathrm{Diag}(\mathbf{w})\|_*$$

Special case of this family

Case 1 $\|\mathrm{Diag}(\mathbf{w})\|_* = \|\mathbf{w}\|_1$

Case 2 $||1^T Diag(\mathbf{w})||_* = ||\mathbf{w}^T||_* = ||\mathbf{w}||_2$

Case 3 Group lasso $\|\mathbf{w}\|_{GL} = \sum_{S_j} \|\mathbf{w}_{S_j}\|_2$ S_j is a group partition

Define $$\mathbf{P}^{GL}_{ij} = \begin{cases} 1/\sqrt{|S_k|} & \text{if } i \text{ and } j \text{ are in the same group } S_k \\ 0 & \text{otherwise.} \end{cases}$$

Decompose

$$\mathbf{P}^{GL}\,\mathrm{Diag}(\mathbf{w}) = \sum_{S_j} \frac{1_{S_j}}{\sqrt{|S_j|}} \mathbf{w}_{S_j}^{\top}$$

Result

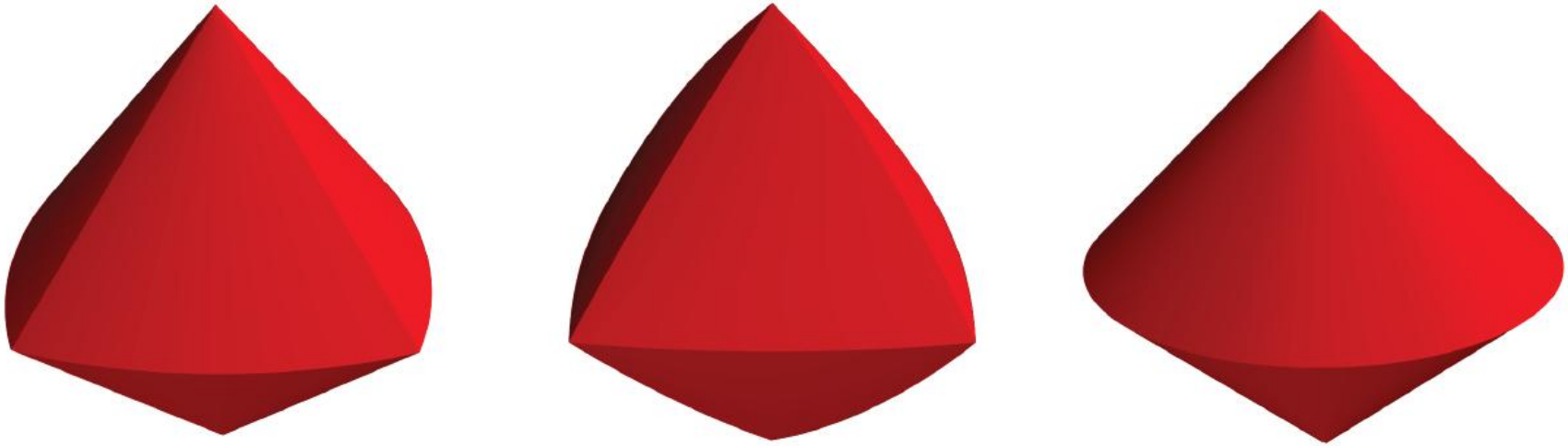
$$\|\mathbf{P}^{GL}\,\mathrm{Diag}(\mathbf{w})\|_* = \sum_{S_j} \|\mathbf{w}_{S_j}\|_2 = \|\mathbf{w}\|_{GL}$$

Properties of trace lasso

Proposition 1

$$\Omega_{\mathbf{P}}(\mathbf{w}) = \|\mathbf{P}\,\mathrm{Diag}(\mathbf{w})\|_* \quad \Rightarrow \quad \Omega_{\mathbf{P}}(\mathbf{w}) = \|(\mathbf{P}^\top\mathbf{P})^{1/2}\,\mathrm{Diag}(\mathbf{w})\|_*$$

Unit ball



Correlation matrix

$$\begin{pmatrix} 1 & 0.9 & 0.1 \\ 0.9 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0.7 & 0.49 \\ 0.7 & 1 & 0.7 \\ 0.49 & 0.7 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Properties of trace lasso

Proposition 2:

If the loss function is strongly convex with respect to its second argument, then the solution of the empirical risk minimization penalized by the trace lass is unique.

$$\hat{\mathbf{w}} \in \operatorname*{argmin}_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \mathbf{w}^\top \mathbf{x}_i) + \lambda f(\mathbf{w})$$

Proposition 3:

Let $\mathbf{P} \in \mathbb{R}^{k \times p}$, *all of its columns having unit norm. We have*

$$\|\mathbf{w}\|_2 \leq \Omega_{\mathbf{P}}(\mathbf{w}) \leq \|\mathbf{w}\|_1.$$

Approximate around lasso

Second order approximation

$$\|(\mathbf{I}+\Delta)\,\mathrm{Diag}(\mathbf{w})\|_* = \|\mathbf{w}\|_1 + \mathrm{diag}(\Delta)^\top|\mathbf{w}| + \sum_{i,j}\frac{\Delta_{ij}^2(|w_i|-|w_j|)^2}{4(|w_i|+|w_j|)} + o(\|\Delta\|^2)$$

When two covariates is correlated, shrink the corresponding coefficient toward each other

Optimization method

Objective function

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{X}\,\mathrm{Diag}(\mathbf{w})\|_*$$

Use theorem

Let $\mathbf{M} \in \mathbb{R}^{n\times p}$. *The trace norm of* $\mathbf{M}$ *is equal to:*

$$\|\mathbf{M}\|_* = \frac{1}{2}\inf_{\mathbf{S}\succeq 0} \mathrm{tr}\left(\mathbf{M}^\top \mathbf{S}^{-1}\mathbf{M}\right) + \mathrm{tr}\left(\mathbf{S}\right),$$

the infimum is attained for $\mathbf{S} = \left(\mathbf{M}\mathbf{M}^\top\right)^{1/2}$

Substitute

$$\min_{\mathbf{w}} \inf_{S\succeq 0} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathrm{Diag}\left(\mathrm{diag}(\mathbf{X}^\top\mathbf{S}^{-1}\mathbf{X})\right)\mathbf{w} + \frac{\lambda}{2}\mathrm{tr}(\mathbf{S})$$

Algorithm

ITERATIVE ALGORITHM FOR ESTIMATING $\mathbf{w}$

Input: the design matrix $\mathbf{X}$, the initial guess $\mathbf{w}^0$, number of iteration N, sequence μ_i.
For $i = 1...N$:

- Compute the eigenvalue decomposition $\mathbf{U}\,\mathrm{Diag}(s_k)\mathbf{U}^\top$ of $\mathbf{X}\,\mathrm{Diag}(\mathbf{w}^{i-1})^2\mathbf{X}^\top$.
- Set $\mathbf{D} = \mathrm{Diag}(\mathrm{diag}(\mathbf{X}^\top\mathbf{S}^{-1}\mathbf{X}))$, where $\mathbf{S}^{-1} = \mathbf{U}\,\mathrm{Diag}(1/\sqrt{s_k + \mu_i})\mathbf{U}^\top$.
- Set $\mathbf{w}^i$ by solving the system $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{D})\mathbf{w} = \mathbf{X}^\top\mathbf{y}$.

Experiment

$\mathbf{x}_i$ is draw from zeros mean Gaussian with different covariance matrix.

Experiment 1: identity

Experiment 2: block diagonal

Experiment 3: Toeplitz matrix