Introduction to High-dimensional Statistics:

LASSO & Beyond

Dr. Abhik Ghosh

Indian Statistical Institute, Kolkata, India.

2021

Table of contents

Outline

High-Dimensional Data around us ...

Set-up

Number of covariates $p \gg n$, number of observations.

- Polynomial Order: $p = O(n^{\alpha})$
- Non-Polynomial (NP) Order: $\log p = O(n^{\alpha})$

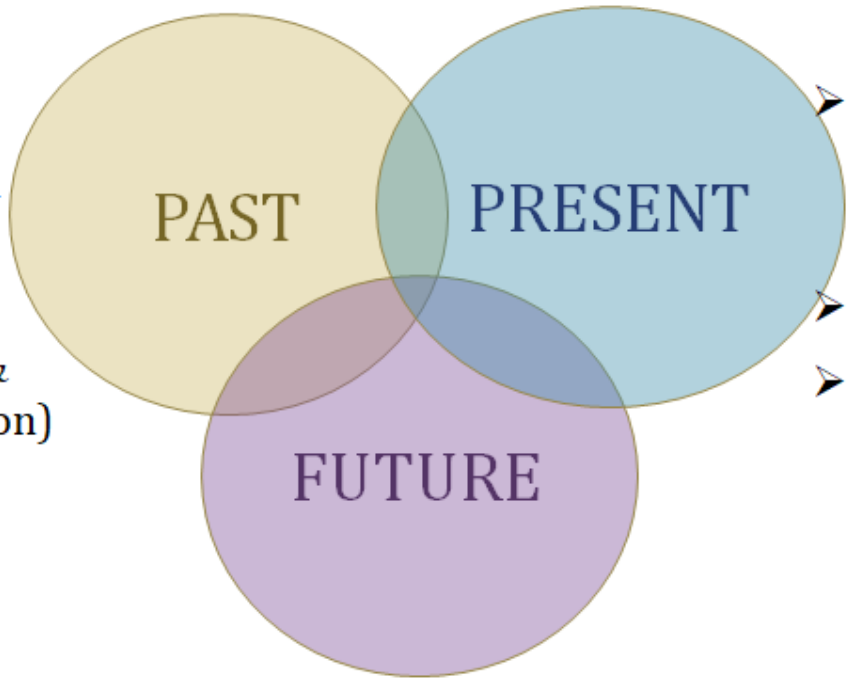
Plenty of high-dimensional data

- OMICS data on few patients in a medical study.
- Social media data, Text or image mining.
- Astronomy or climate research.

Challenges

- Classical estimators (e.g., OLS in regression) are not uniquely defined.
- Valid statistical inference with classical techniques is not possible.
- Variable selection is important (Difficult to interpret so many predictors).
- THE computational complexity.

Statistical Procedures for High-dimensional Data



- **LASSO** for Regression
- Extensions in penalty &/or loss function

(Estimation & variable selection)

- Beyond regression

(Estimation & variable selection)

- Significance testing
- Roust Extensions
  - **M-estimator**
  - **Minimum distance,**

Robust Significance testing & other advanced issues beyond regression

Outline

The LASSO for linear regression

Standard **linear regression model** (LRM):

$$\boldsymbol{y} \quad = \quad \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

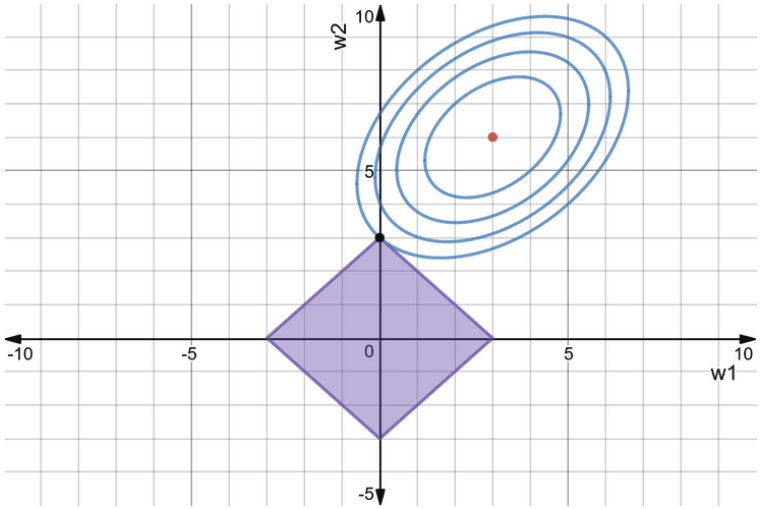
where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ are responses,
$\mathbf{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)^T$ is the design matrix,
and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ are the random error components.

LASSO (Tibshirani, 1996) of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, is defined as the minimizer of:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p}|\beta_j|$$

$$= \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\lambda_n \equiv \lambda$ is the regularization parameter depending on n.



So, LASSO works for both the cases $p < n$ and $p \gg n$!

Sparsity

The basic requirement for valid statistical analysis of high-dimensional data!

What is Sparsity Assumption?

- True regression coefficient β_0 has only a few ($s \ll n$) non-zero coefficients, i.e., among $p \gg n$ covarites, only a few (s) are actually significant for explaining the response.
- Mathematically, the true active set $S_0 = \{j : \beta_{0j} \neq 0\}$ has size $s \ll n$.

The **challenge** remains in that we do not know which s covarites are indeed important!

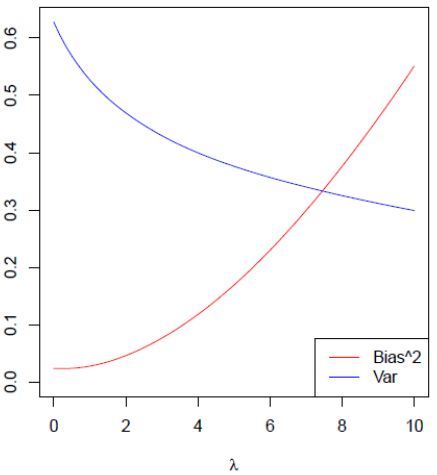
How is LASSO useful?

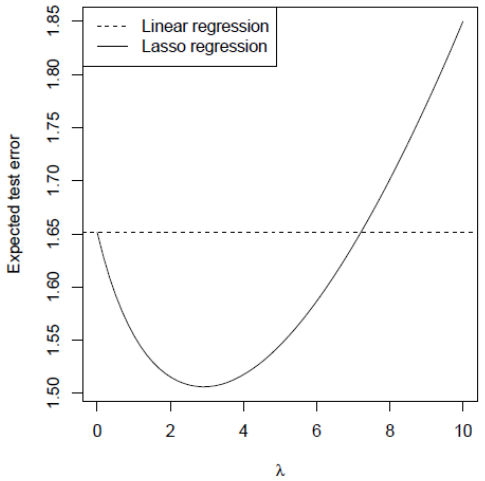
- LASSO estimates some regression coefficient exactly as zero, leading to the simultaneous selection of important variables via the non-zero estimated coefficients.
- Estimated active set $\widehat{S} = \left\{j : \widehat{\beta}_j \neq 0\right\}$.
- The important trick here, for such a sparse estimation and simultaneous variable section, is the incorporation of the ℓ_1-penalty (along with squared error loss)!

The Regularization parameter λ

The regularization parameter λ controls the amount of penalty!

- LASSO not only set coefficients to zero exactly, but it also shrinks the nonzero coefficients.
- λ controls the **bias-variance trade-off**.
- At $\lambda = 0$, LASSO coincides with the (unbiased) OLS.
- Larger the value of λ, more the shrinkage is (and also more zero coefficients) leading to higher bias.





Selection of λ in practice

- **Cross-Validation**: Split the sample in training and test samples, and minimize test-accuracy.
- **Extended BIC** (Chen and Chen, 2008, 2012): Put $\widehat{\sigma}^2 = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2$ and minimize

$$\text{EBIC}(\lambda) = \log(\widehat{\sigma}^2) + \left(\frac{\log n}{n} + \gamma \frac{\log p}{n}\right)|\widehat{S}|, \quad 0 \leq \gamma \leq 1.$$

- **High-dimensional BIC** (Kim et al., 2012): Put $\widehat{\sigma}^2 = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2$ and minimize

$$\text{HBIC}(\lambda) = \log(\widehat{\sigma}^2) + \frac{\log\log(n)\log p}{n}|\widehat{S}|$$

Properties of the LASSO: Prediction Consistency

Assume: The truth is linear with $\beta = \beta_0$ and $\widehat{\Sigma} = \frac{1}{n}(\mathbf{X}^T\mathbf{X})$ has all diagonal elements one.

Consider $\lambda = 4\widehat{\sigma}\sqrt{\frac{t^2+2\log p}{n}}$ for some $t > 0$, where $\widehat{\sigma}^2$ is any estimate of error variance σ^2.

Put $p_t = 1 - 2e^{-t^2/2} - P(\widehat{\sigma} \leq \sigma)$.

Prediction error bound for LASSO estimator $\widehat{\beta}$ (Buhlmann and Van de Geer, 2011)

$$\frac{2}{n}\left|\left|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta_0\right|\right|_2^2 \leq 3\lambda||\beta_0||_1, \quad \text{with propbability at least } p_t. \tag{2}$$

Implication

If $||\beta_0||_1 \ll \sqrt{\frac{n}{\log p}}$, then LASSO solution is **consistent in prediction error** for $\lambda = O\left(\sqrt{\frac{\log p}{n}}\right)$ (ensuring $p_t \to 1$).

Properties of the LASSO: Estimation and Prediction Consistency

Assume: The truth is linear with $\beta = \beta_0$ and $\widehat{\Sigma} = \frac{1}{n}(\mathbf{X}^T\mathbf{X})$ has all diagonal elements one.

Consider $\lambda = 4\widehat{\sigma}\sqrt{\frac{t^2+2\log p}{n}}$ for some $t > 0$, where $\widehat{\sigma}^2$ is any estimate of error variance σ^2.

Put $p_t = 1 - 2e^{-t^2/2} - P(\widehat{\sigma} \leq \sigma)$ and s is the size of $S_0 = \{j : \beta_{0j} \neq 0\}$.

Theorem (Buhlmann and Van de Geer, 2011)

Under ***Compatibility Condition***, the LASSO estimator $\widehat{\beta}$ satisfies

$$\frac{1}{n}\left|\left|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta_0\right|\right|_2^2 + \lambda\left|\left|\widehat{\beta} - \beta_0\right|\right|_1 \leq \frac{4\lambda^2 s}{\phi^2}, \quad \text{with propbability at least } p_t. \tag{3}$$

Implications

- Bound on estimation error: $\left|\left|\widehat{\beta} - \beta_0\right|\right|_1 \leq \frac{4\lambda s}{\phi^2}$.
- Bound on average prediction error: $\frac{1}{n}\left|\left|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta_0\right|\right|_2^2 \leq \frac{4\lambda^2 s}{\phi^2}$.
- Compatibility condition makes the convergence rate faster for the average prediction error.

Properties of the LASSO: Sure Variable Screening

Assume: The truth is linear with $\beta = \beta_0$ and $\widehat{\Sigma} = \frac{1}{n}(\mathbf{X}^T\mathbf{X})$ has all diagonal elements one.

Consider $\lambda = 4\widehat{\sigma}\sqrt{\frac{t^2+2\log p}{n}}$ for some $t > 0$, where $\widehat{\sigma}^2$ is any estimate of error variance σ^2.

Put $S_0 = \{j : \beta_{0j} \neq 0\}$ and $\widehat{S} = \left\{j : \widehat{\beta}_j \neq 0\right\}$.

Beta-Min Condition (Meinshausen and Buhlmann, 2006)

$$\min_{j \in S_0} |\beta_{0j}| \gg \frac{1}{\phi^2}\sqrt{\frac{s\log p}{n}}.$$

Larger the minimum non-zero coefficient, easier to select the active set.

Theorem (Buhlmann and Van de Geer, 2011)

Under the ***Restricted Eigenvalue Condition*** and ***beta-min Condition***, we have

$$P(S_0 \subseteq \widehat{S}) \to 1, \quad \text{as } p \geq n \to \infty. \tag{4}$$

Implication

- All truly significant variables will be selected by LASSO (having non-zero coefficient estimates) with high probability, tending to one for $\lambda = O\left(\sqrt{\frac{\log p}{n}}\right)$ (along with several other non-relevant covariates).

Properties of the LASSO: Variable Selection Consistency

Assume: The truth is linear with $\beta = \beta_0$ and $\widehat{\Sigma} = \frac{1}{n}(\mathbf{X}^T\mathbf{X})$ has all diagonal elements one.

Consider $\lambda = 4\widehat{\sigma}\sqrt{\frac{t^2+2\log p}{n}}$ for some $t > 0$, where $\widehat{\sigma}^2$ is any estimate of error variance σ^2.

Put $S_0 = \{j : \beta_{0j} \neq 0\}$ and $\widehat{S} = \left\{j : \widehat{\beta}_j \neq 0\right\}$.

Theorem (Buhlmann and Van de Geer, 2011)

Under the ***Irrepresentable Condition*** and ***beta-min Condition***, the LASSO estimator $\widehat{\beta}$ satisfies

$$P(S_0 = \widehat{S}) \to 1, \quad \text{as } p \geq n \to \infty. \tag{5}$$

Implication

- All and only the truly significant variables will be be selected by LASSO (having non-zero coefficient estimates) with probability tending to one for $\lambda = O\left(\sqrt{\frac{\log p}{n}}\right)$, with no non-relevant covariates being selected.

Assumption: Compatibility Condition

Denote: $\widehat{\Sigma} = \frac{1}{n}(\mathbf{X}^T\mathbf{X})$ and $S_0 = \{j : \beta_{0j} \neq 0\}$.

Compatibility Condition (van de Geer, 2008)

There exists $\phi > 0$ such that, for all β satisfying $\left|\left|\beta_{S_0^c}\right|\right|_1 \leq 3\left|\left|\beta_{S_0}\right|\right|_1$, we must have

$$\left|\left|\beta_{S_0}\right|\right|_1^2 \leq \frac{s}{\phi^2}(\beta^T\widehat{\Sigma}\beta).$$

Interpretation

- Replacing $\left|\left|\beta_{S_0}\right|\right|_1^2$ by its upper bound $s\left|\left|\beta_{S_0}\right|\right|_2^2$, we get

$$\phi^2 \leq \frac{(\beta^T\widehat{\Sigma}\beta)}{\left|\left|\beta_{S_0}\right|\right|_2^2} = \frac{(\beta_{S_0}^T\widehat{\Sigma}_{S_0,S_0}\beta_{S_0})}{\left|\left|\beta_{S_0}\right|\right|_2^2}.$$

- ϕ_0^2 looks like a minimum bound for eigenvalues of $\widehat{\Sigma}_{S_0,S_0}$ (although much relaxed - Why?)

Assumption: Restricted Eigenvalue

Denote: $\widehat{\Sigma} = \frac{1}{n}(\mathbf{X}^T\mathbf{X})$, $S_0 = \{j : \beta_{0j} \neq 0\}$ and $|S_0| = s$.

Restricted Eigenvalue Condition (Bickel et al., 2009)

There exists $\phi > 0$ such that, for all β satisfying $\left|\left|\beta_{S_0^c}\right|\right|_1 \leq 3\left|\left|\beta_{S_0}\right|\right|_1$, we must have

$$||\beta_S||_2^2 \leq \frac{1}{\phi^2}(\beta^T\widehat{\Sigma}\beta), \qquad \text{for all } S \text{ with } |S| = s.$$

Interpretation

- It implies Compatibility Condition!
- It is a stronger assumption that the compatibility condition, but still weaker than imposing a lower bound for eigenvalues!
- ϕ_0^2 is called a restricted-eigenvalue or compatibility constant!

Assumption: Irrepresentable Condition

Denote: $\widehat{\Sigma} = \frac{1}{n}(\mathbf{X}^T\mathbf{X})$ and $S_0 = \{j : \beta_{0j} \neq 0\}$.

Irrepresentable Condition: (Zou, 2006; Zhao and Yu, 2006)

$$\left|\left|\widehat{\Sigma}_{S_0^c,S_0}\widehat{\Sigma}_{S_0,S_0}sign(\beta_{S_0})\right|\right|_\infty \leq \theta, \quad \text{for some } 0 < \theta < 1. \tag{6}$$

It implies Compatibility Condition!

Interpretation

It holds if **X** is not too much “ill-posed" or does not exhibits strong linear dependence!
Examples:

- $\Sigma = (1-\rho)\mathbb{I} + \rho\mathbb{J}$ with $0 \leq \rho \leq \frac{\theta}{s(1-\theta)+\theta} < 1$.
- $\Sigma = ((\rho^{|j-k|}))_{j,k}$ with $|\rho| \leq \theta < 1$.
- Pairwise correlations (among important and unimportant covariates) are suitably bounded.

Necessary Condition for LASSO to perform consistent variable selection is (6) with $\theta = 1$!!

In practice, LASSO often select too many unimportant covariates (false discovery)!!!

Variants of LASSO

To reduce false discoveries in LASSO.

Adaptive LASSO (Zou, 2006)

$$\widehat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\widehat{\beta}_{j,init}|} \right\}$$

Advantage: Variable Selection Consistency under weaker assumptions (reduces bias).

Other variants

- Weighted LASSO - Generalization of Adaptive LASSO.
- Multi-Step Adaptive LASSO (Buhlmann and Meier, 2008).
- Relaxed LASSO (Meinshausen, 2007) - Performs similar to Adaptive LASSO.
- Thresholding LASSO (Zhou, 2010) - Make smaller LASSO coefficients zero.
- Considering penalty functions beyond LASSO penalty!

Group LASSO (Yuan and Lin, 2006)

To Achieve group sparsity in specific applications (e.g., genomics).

$$\text{Group LASSO penalty } = \lambda \sum_{j} m_j \|\beta_{\mathcal{G}_j}\|.$$

Outline

LASSO for Generalized Linear Models

Generalized linear model (GLM): Given a covariate value $\mathbf{X} = \boldsymbol{x}$, the response variable Y has density

$$f(y; \boldsymbol{x}^T\boldsymbol{\beta}) = \exp\{y\theta - b(\theta) + c(y)\}, \quad \text{with} \quad E[Y|\boldsymbol{x}] = b'(\theta) = g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta}), \tag{7}$$

where $b(\cdot)$ and $c(\cdot)$ are some appropriate known functions,
g is a known monotone differentiable link function,
and the canonical parameter θ is defined via the linear predictor $\eta = \boldsymbol{x}^t\boldsymbol{\beta}$.

LASSO estimate of $\boldsymbol{\beta}$, based on a sample $(y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$, is defined as

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log f(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \tag{8}$$

The loss-function is the negative log-likelihood!

Generalize for loss functions of the form $\rho_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i)$, which is **convex in $\boldsymbol{\beta}$**:

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{n}\sum_{i=1}^{n} \rho_{\boldsymbol{\beta}}(y_i, \mathbf{x}_i) + \lambda \|\boldsymbol{\beta}\|_1 \right), \tag{9}$$

Theoretical properties are studied in Loubes and van de Geer (2002) and van de Geer (2008).

General Non-Convex Loss Functions

Suppose that the distribution of a given sample depends on the unknown regression coefficient $\boldsymbol{\beta}$ and an additional parameter $\boldsymbol{\eta}$ (e.g., error variance, mixing proportions, random-effect variances).

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}}) = \arg \min_{(\boldsymbol{\beta}, \boldsymbol{\eta})} \left\{ L_n(\boldsymbol{\beta}, \boldsymbol{\eta}) + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \tag{10}$$

The useful loss-function L_n may often be **non-convex**!!

Examples

Negative log-likelihood loss function for the following models:

- Finite Mixture of Regression (Stadler et al., 2010)
- Linear Mixed-Effect Models (Schelldorfer et al., 2011)
- Generalized Linear Mixed-Effect Models (Schelldorfer et al., 2014)

Outline

Penalties beyond LASSO

Standard **linear regression model** (LRM): $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$,
with $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ are the random error components.

Replace the LASSO (ℓ_1) penalty by a **General Penalty** p_λ.

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}.$$

Examples of General penalty

- ℓ_q **penalty**: $p_\lambda\left(|\beta_j|\right) = \lambda |\beta_j|^q$ with $q \in [0, 2]$;
- For $q = 0$, we have hard thresholding penalty
- For $q = 2$, we have ridge estimator – not sparse although less biased in low-dimension!!

- **Elastic-net penalty**: $p_\lambda\left(|\beta_j|\right) = \lambda \left((1-\alpha)|\beta_j| + \alpha |\beta_j|^2\right)$ with $\alpha \in [0, 1]$.

Smoothly Clipped Absolute Deviation (SCAD) Penalty (Fan and Li, 2001)

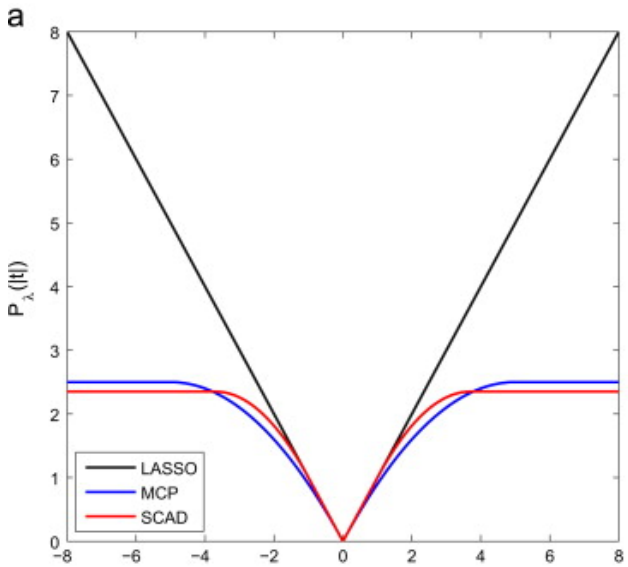
$$p_\lambda(|s|) = \begin{cases} \lambda|s| & \text{if} \quad |s| \le \lambda \\ \frac{2a\lambda|s| - |s|^2 - \lambda^2}{2(a-1)} & \text{if} \quad \lambda < |s| \le a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if} \quad |s| > a\lambda \end{cases}$$

with $a > 2$. Suggested choice: $a = 3.7$.

Minimax Concave Penalty (MCP) (Zhang, 2010)

$$p_\lambda(|s|) = \begin{cases} \lambda|s| - \frac{|s|^2}{2a} & \text{if} \quad |s| \le a\lambda \\ \frac{a\lambda^2}{2} & \text{if} \quad |s| > a\lambda \end{cases}$$

with $a > 1$.



Desired properties of a general Penalty function

A “good" penalty function p_λ should satisfy the following three properties (Fan and Li, 2001):

i) **Unbiasedness**: The resulting estimator is nearly unbiased if the true parameter value is large (to avoid unnecessary modeling bias), which holds if $p'_\lambda(|s|) = 0$ for large s.

ii) **Sparsity**: The resulting estimator is a thresholding rule, automatically setting some estimated coefficients to zero, which holds if $\min\left(|s| + p'_\lambda(|s|)\right) > 0$.

iii) **Continuity**: The resulting estimator is continuous in data, for which $\min\left(|s| + p'_\lambda(|s|)\right)$ should attain at $s = 0$.

Examples

- ℓ_q penalty does not satisfy the unbiasedness, sparsity or the continuity condition, respectively, if $q = 1$, $q > 1$ or $0 < q < 1$.
- SCAD satisfies all the desired properties (i)–(iii).
- MCP satisfies (i) and (ii), but not (iii).

A More General Assumption on Penalty Function

Assumption (P)

$p_\lambda(s)$ is continuously differentiable, increasing, and concave in $s \in [0, \infty)$.
Also $p'_\lambda(s)/\lambda$ is increasing in λ and $\rho(p_\lambda) := p'_\lambda(0+)/\lambda > 0$ is independent of λ.

Examples (Lv and Fan, 2009; Fan and Lv, 2011)

LASSO (ℓ_1), SCAD and MCP all satisfies Assumption (P)!

Relations with Fan and Li Criterion

- The unbiasedness and sparsity properties hold for penalties satisfying Assumption (P), if additionally $\lim_{t \to \infty} p'(s) = 0$.
- The continuity property does not hold in general for all penalties satisfying (P). [e.g., MCP]

Another Bridge Penlaty Family (Lv and Fan, 2009)

- $p_\lambda(t) = \lambda \left[\left(\frac{t}{a+t} \right) I(t \neq 0) + \left(\frac{a}{a+t} \right) t \right]$, a convex combination of ℓ_0 and ℓ_1 penalties, for $a \in [0, \infty]$.
- It satisfies Assumption (P) and the unbiasedness and sparsity properties for all $a > 0$.
- It satisfies the continuity property only for all $a \geq \lambda + \sqrt{\lambda^2 + 2\lambda}$.

Outline

Fan and Liao (2014)

Theory for general penalized estimator

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ L_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda \left(|\beta_j| \right) \right\}, \tag{11}$$

with convex loss $L_n(\boldsymbol{\beta})$ and noncocave penalty p_λ satisfying (P).

Ghosh and Thoresen (2018)

Extended the theory for more general non-convex loss $L_n(\boldsymbol{\beta}, \boldsymbol{\eta})$;

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}}) = \arg\min_{(\boldsymbol{\beta}, \boldsymbol{\eta})} \left\{ L_n(\boldsymbol{\beta}, \boldsymbol{\eta}) + \sum_{j=1}^{p} p_\lambda \left(|\beta_j| \right) \right\}. \tag{12}$$

Assumptions

Assumptions on the loss function (L1)

(i) $L_n((\boldsymbol{\beta}_S, \mathbf{0}); \boldsymbol{\eta})$ is twice differentiable with respect to $\boldsymbol{\beta}_S$ and $\boldsymbol{\eta}$ around true values $(\boldsymbol{\beta}_{S0}, \mathbf{0}; \boldsymbol{\eta}_0)$.

(ii) For some positive sequences $a_n = o(d_n)$, with $d_n = \frac{1}{2}\min\{|\beta_{0j}| : \beta_{0j} \neq 0\}$, and $c_n = o(1)$,

$$||\nabla_S L_n(\boldsymbol{\beta}_{S0}, \mathbf{0}; \boldsymbol{\eta}_0)|| = O_p(a_n), \quad \text{and} \quad ||\nabla_\eta L_n(\boldsymbol{\beta}_{S0}, \mathbf{0}; \boldsymbol{\eta}_0)|| = O_p(c_n).$$

(iii) For any $\epsilon > 0$, there exists some positive constant C_ϵ such that

$$P\left(\lambda_{min}(\nabla^2 L_n(\boldsymbol{\beta}_{S0}, \mathbf{0}; \boldsymbol{\eta}_0)) > C_\epsilon\right) > 1 - \epsilon, \quad \text{for all large } n$$

(iv) For any given $\epsilon > 0$, $\delta > 0$ and non-negative sequences $\alpha_n = o(d_n)$ and $\gamma_n = o(1)$, there exist a large N^* such that, for all $n > N^*$,

$$P\left(\sup_{||\boldsymbol{\beta}_S - \boldsymbol{\beta}_{S0}|| \leq \alpha_n, ||\boldsymbol{\eta} - \boldsymbol{\eta}_0|| \leq \gamma_n} ||\nabla^2 L_n(\boldsymbol{\beta}_S, \mathbf{0}; \boldsymbol{\eta}) - \nabla^2 L_n(\boldsymbol{\beta}_{S0}, \mathbf{0}; \boldsymbol{\eta}_0)|| \leq \delta\right) > 1 - \epsilon.$$

Additional assumption for penalty (P*)

(i) $\sqrt{s}p'_\lambda(d_n) = o(d_n)$.

(ii) There exists a constant $c > 0$ such that $\sup_{\boldsymbol{\beta} \in B(\boldsymbol{\beta}_{S_0}, cd_n)} \max_{j \leq p}\left[-p''_\lambda(|\boldsymbol{\beta}|)\right] = o(1)$.

Consistency

Theorem (Oracle consistency)

Under Assumptions (P), (P) and (L1), there exists a local minimum* $\left(\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_S^T, \mathbf{0})^T,\ \widehat{\boldsymbol{\eta}}\right)$ *of*

$$\left\{ L_n((\boldsymbol{\beta}_S^T, \mathbf{0})^T, \boldsymbol{\eta}) + \sum_{j=1}^{p} p_\lambda\left(|\beta_j|\right) \right\}$$

satisfying

$$||\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{S0}|| = O_p(a_n + \sqrt{s} p'_\lambda(d_n)), \quad \text{and} \quad ||\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0|| = O_p(c_n).$$

In addition, for any given $\epsilon > 0$*, the local minimizer* $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}})$ *is strict with probability at least* $1 - \epsilon$ *for sufficiently large n.*

Variable Selection Optimality

Assumptions on the loss (L2)

For the local minimizer $(\widehat{\boldsymbol{\beta}}_S, \widehat{\boldsymbol{\eta}})$ obtained in Theorem 1, there exists a neighborhood $\mathcal{H} \subset \mathbb{R}^{p+d}$ of $(\widehat{\boldsymbol{\beta}}_S^T, 0, \widehat{\boldsymbol{\eta}}^T)^T$ such that, with probability tending to one, we have

$$L_n(T\boldsymbol{\beta}, \boldsymbol{\eta}) - L_n(\boldsymbol{\beta}, \boldsymbol{\eta}) < \sum_{j \notin S} p_\lambda(|\beta_j|), \quad \text{for all} \quad \left((\boldsymbol{\beta}_S^T, \boldsymbol{\beta}_N^T), \boldsymbol{\eta}^T\right) \in \mathcal{H}, \quad \text{with } \boldsymbol{\beta}_N \neq 0.$$

Here, $T\boldsymbol{\beta}$ denote the projection of $\boldsymbol{\beta}$ onto the space generated by S, i.e., $T\boldsymbol{\beta} = (\beta_1', \ldots, \beta_p')^T$ with $\beta_j' = \beta_j I(j \in S)$.

Theorem (Variable selection optimality)

Under Assumptions (P), (P), (L1) and (L2), we have the followings:*

(i) $(\widehat{\boldsymbol{\beta}}_S, \mathbf{0}, \widehat{\boldsymbol{\eta}})$ *obtained in Theorem 1 is a local minimizer in* $\mathbb{R}^{p+d}$ *of the general objective function* $\left\{ L_n(\boldsymbol{\beta}, \boldsymbol{\eta}) + \sum_{j=1}^{p} p_\lambda\left(|\beta_j|\right) \right\}$, *with probability tending to one.*

(ii) *For any given* $\epsilon > 0$, *the local minimizer* $(\widehat{\boldsymbol{\beta}}_S, \mathbf{0}, \widehat{\boldsymbol{\eta}})$ *is strict with probability at least* $1 - \epsilon$ *for all sufficiently large* n.

Outline

Model and Estimator

Model

There are n_i observations in the i-th group, for $i = 1, \ldots, I$, with total number of observations $n = \sum_{i=1}^{I} n_i$. For each group, we consider the model (Pinheiro and Bates, 2000)

$$\boldsymbol{y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, ..., I, \tag{13}$$

where β is p-dimensional vector of fixed effect coefficients,
$\boldsymbol{b}_i \sim N_q(0, \boldsymbol{\Psi}_{\boldsymbol{\theta}})$ are the random effects with $\boldsymbol{\theta} \in \mathbb{R}^{q^*}$ being the variance parameters,
and the random error $\boldsymbol{\epsilon}_i \sim N_{n_i}(0, \sigma^2 \boldsymbol{I}_{n_i})$, independent of the random effects $\boldsymbol{b}_i$ and $\boldsymbol{X}_i$s.

For each i, given $\boldsymbol{X}_i$ (and $\boldsymbol{Z}_i$), $\boldsymbol{y}_i \sim N_{n_i}(\boldsymbol{X}_i\beta, \boldsymbol{V}_i(\boldsymbol{\eta}))$, where $\boldsymbol{V}_i(\boldsymbol{\eta}) = \boldsymbol{V}_i(\boldsymbol{\theta}, \sigma) = \boldsymbol{Z}_i\boldsymbol{\Psi}_{\boldsymbol{\theta}}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{I}_{n_i}$.

General Regularized Estimator

- Minimize $\{L_n(\beta, \boldsymbol{\eta}) + \sum_{j=1}^{p} p_\lambda\left(|\beta_j|\right)\}$, with the likelihood-based loss

$$L_n(\beta, \boldsymbol{\eta}) \quad = \quad \frac{1}{2}\sum_{i=1}^{I}\left[n_i \log(2\pi) + \log|\boldsymbol{V}_i(\boldsymbol{\eta})| + (\boldsymbol{y}_i - \boldsymbol{X}_i\beta)^T \boldsymbol{V}_i(\boldsymbol{\eta})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\beta)\right].$$

- Schelldorfer et al. (2011) studied this problem for LASSO penalty.
- Ghosh and Thoresen (2018) studied it for general non-concave penalties [including **SCAD**].

Theorem (Ghosh and Thoresen, 2018)

Consider the problem of regularized estimation in the high-dimensional linear mixed model (13) with the SCAD penalty and regularization parameter λ_n, where

$$s^3 \log p = o(n) \quad \text{and} \quad s\sqrt{\log p/n} + s^3 \log s/n \ll \lambda_n \ll d_n.$$

Assume that the observations $\boldsymbol{V}_k = (y_k, \boldsymbol{X}_k, D_k)$, $k = 1, \ldots, n$, are IID with suitably chosen covariate distribution and mixed effects structure. Then, there exists a local minimizer $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}}) = ((\widehat{\boldsymbol{\beta}}_S^T, \widehat{\boldsymbol{\beta}}_{S^c}^T)^T, \widehat{\boldsymbol{\eta}})$ of the (SCAD) penalized likelihood-based objective function that satisfies the following:

1. $\lim_{n\to\infty} P(\widehat{\boldsymbol{\beta}}_{S^c} = \mathbf{0}) = 1$. *In addition, the local minimizer is strict with probability arbitrarily close to one for all sufficiently large n.*
2. *Assuming that $\widehat{S} = \{j \leq p : \widehat{\beta}_j \neq 0\}$ denotes the estimated active set, $\lim_{n\to\infty} P(\widehat{S} = S) = 1$.*
3. *For any unit vector $\boldsymbol{\alpha} \in \mathbb{R}^s$,*

$$\sqrt{n}\boldsymbol{\alpha}^t \boldsymbol{I}_{S,S}^{1/2} \left(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{S0}\right) \xrightarrow{\mathcal{D}} N(0,1), \quad \sqrt{n}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{\mathcal{D}} N_d\left(\mathbf{0}, \boldsymbol{I}_{S^c,S^c}^{-1}\right).$$

Simulation Comparison

		$\lvert S(\hat{\beta})\rvert$	TP	PE	β_1	β_2	β_3	β_4	β_5	β_N	σ^2	θ^2
L_1 Penalty												
						$p = 300$						
$\rho = 0$	Mean	11.23	5.00	0.15	1.02	2.02	3.94	2.95	2.96	0.00	0.21	0.39
	SD	4.24	0.00	0.03	0.34	0.32	0.05	0.06	0.06	0.01	0.04	0.22
	MSE				0.1149	0.1019	0.0057	0.0061	0.0050	0.0001	0.0029	0.0772
$\rho = 0.5$	Mean	9.37	5.00	0.16	1.01	1.97	3.97	2.98	2.96	0.00	0.22	0.44
	SD	3.79	0.00	0.03	0.35	0.40	0.07	0.07	0.07	0.01	0.04	0.25
	MSE				0.1227	0.1594	0.0059	0.0049	0.0059	0.0001	0.0024	0.0757
						$p = 500$						
$\rho = 0$	Mean	10.85	5.00	0.15	0.97	1.95	3.93	2.94	2.94	0.00	0.22	0.42
	SD	4.12	0.00	0.03	0.36	0.33	0.05	0.06	0.06	0.01	0.04	0.24
	MSE				0.1278	0.1096	0.0078	0.0066	0.0075	0.0001	0.0032	0.0751
$\rho = 0.5$	Mean	10.53	5.00	0.16	1.05	2.05	3.97	2.98	2.96	0.00	0.22	0.36
	SD	3.89	0.00	0.03	0.40	0.38	0.08	0.07	0.07	0.01	0.04	0.23
	MSE				0.1625	0.1475	0.0065	0.0051	0.0058	0.0001	0.0026	0.0893
SCAD Penalty												
						$p = 300$						
$\rho = 0$	Mean	7.29	5.00	0.16	1.06	2.00	3.99	3.00	3.00	0.00	0.22	0.42
	SD	3.57	0.00	0.03	0.37	0.36	0.05	0.05	0.05	0.01	0.04	0.24
	MSE				0.1403	0.1279	0.0029	0.0027	0.0026	0.0001	0.0022	0.0745
$\rho = 0.5$	Mean	7.22	5.00	0.16	0.98	2.00	4.01	3.00	2.99	0.00	0.22	0.43
	SD	3.58	0.00	0.03	0.37	0.33	0.06	0.06	0.06	0.01	0.04	0.24
	MSE				0.1368	0.1078	0.0042	0.0037	0.0038	0.0001	0.0021	0.0735
						$p = 500$						
$\rho = 0$	Mean	8.30	5.00	0.15	1.04	1.94	3.99	2.99	3.00	0.00	0.21	0.42
	SD	4.16	0.00	0.03	0.33	0.34	0.05	0.05	0.06	0.00	0.04	0.29
	MSE				0.1093	0.1189	0.0024	0.0025	0.0033	0.0001	0.0031	0.1032
$\rho = 0.5$	Mean	7.62	5.00	0.16	1.01	2.01	4.00	3.00	2.99	0.00	0.23	0.42
	SD	3.52	0.00	0.03	0.34	0.37	0.08	0.07	0.07	0.00	0.04	0.22
	MSE				0.1130	0.1350	0.0056	0.0051	0.0044	0.0000	0.0025	0.0696

* Taken from Ghosh and Thoresen (2018).

A Real Data Example

- Ottestad et al. (2012) investigated the effects of intake of oxidized and non-oxidized fish oil on inflammatory markers in a randomized study of 52 subjects. Inflammatory markers were measured at baseline and after three and seven weeks.
- Available sample observations $n = 150$ (after removing missing values).

- **Our objective**: To investigate whether there are any associations between gene expressions measured at baseline and level of the inflammatory marker ICAM-1 throughout the study.

- **Model**:
  Fixed effects are treatment (3 groups), time and their interaction "Treatment $\times$ Time", along with 506 gene expression measurements. (having absolute correlation greater than or equal to 0.2 with the response at any time point) so that $p = 512 (\gg n)$.

  Random effects are the random intercept (b_I) and a random slope for "*Time*" (b_{Time}), and assume that $(b_I, b_{Time})^T \sim N_2(0, Diag\{\theta_I^2, \theta_{Time}^2\})$.

Data Results (Ghosh and Thoresen, 2018)

Penalty	Mixed Model SCAD	Mixed Model L_1	Regression Model SCAD	Regression Model L_1
Number of Genes Selected				
	29	30	32	37
Coefficients of Selected Genes				
DOCK10 (*)	3.03 (1)	5.71 (1)	3.04 (1)	4.94 (1)
CAST (*)	2.73 (2)	3.02 (2)	2.83 (2)	3.18 (2)
GZMK (*)	2.43 (3)	0.26 (14)	1.84 (3)	1.43 (5)
NA	2.08 (4)	1.68 (3)	1.82 (4)	2.41 (3)
HLA-H (*)	1.56 (5)	0.88 (8)	1.47 (6)	1.39 (6)
SLC22A16	1.52 (6)	– (15)	0.58 (11)	0.85 (10)
GSTM1 (*)	1.38 (7)	0.91 (7)	1.55 (5)	1.35 (7)
NA	1.13 (8)	0.31 (13)	0.86 (7)	0.71 (11)
SNX29	0.96 (9)	1.41 (4)	0.63 (9)	1.90 (4)
UTS2 (*)	0.73 (10)	0.48 (12)	0.59 (10)	0.45 (13)
FAM45A	0.34 (11)	1.09 (6)	0.15 (13)	0.96 (9)
LOC554223	0.26 (12)	0.59 (10)	0.56 (12)	0.68 (12)
ACCS	– (13)	1.38 (5)	0.78 (8)	1.16 (8)
PJA2	– (13)	0.63 (9)	– (15)	0.30 (15)
NFIB	– (13)	0.49 (11)	– (15)	0.42 (14)
IRF5 (*)	– (13)	– (15)	0.05 (14)	0.10 (16)
LOC100170939	– (13)	– (15)	– (15)	-0.02 (17)
MYL4	– (13)	– (15)	-0.17 (21)	-0.06 (19)
PKIA	– (13)	-0.57 (25)	– (15)	-0.86 (25)
FGD2	– (13)	-0.69 (26)	– (15)	-0.05 (18)
MX1 (*)	-0.12 (21)	-0.52 (24)	-0.40 (22)	-0.47 (21)
HSH2D (*)	-0.80 (22)	-0.40 (23)	-0.52 (23)	-0.86 (24)
LOC644936	-1.02 (23)	-1.34 (30)	-1.18 (27)	-1.07 (26)
PPAT	-1.21 (24)	-0.97 (28)	-1.03 (24)	-0.81 (23)
NA	-1.23 (25)	-0.77 (27)	-1.08 (25)	-0.80 (22)
NAPRT1	-1.36 (26)	-1.60 (31)	-1.60 (30)	-1.59 (29)
N4BP2L2	-1.49 (27)	-1.92 (34)	-1.82 (32)	-1.75 (32)
GYPC (*)	-1.63 (28)	-0.07 (22)	-1.61 (31)	-1.29 (27)
CENPK	-1.66 (29)	-1.66 (32)	-1.51 (29)	-1.69 (30)
COL18A1	-1.95 (30)	-1.16 (29)	-1.39 (28)	-1.44 (28)
C1orf85 (*)	-1.98 (31)	– (15)	-0.10 (20)	-0.15 (20)
ZNF266	-2.09 (32)	– (15)	-2.51 (35)	-2.60 (35)
COMMD2 (*)	-2.26 (33)	-2.42 (35)	-2.28 (34)	-2.28 (34)
ANPEP	-2.27 (34)	-1.70 (33)	-1.94 (33)	-2.01 (33)
PRUNE2	-2.91 (35)	– (15)	-1.14 (26)	-1.72 (31)
NAIP (*)	-2.96 (36)	-2.77 (36)	-3.20 (36)	-3.58 (36)
PKIA	-4.07 (37)	-4.72 (37)	-4.19 (37)	-4.68 (37)

Penalty	$\hat{\sigma}$	$\hat{\theta}_I$	$\hat{\theta}_{Time}$
SCAD	3.134	0	0.520
L_1	3.435	0	0.571

Findings!

- The mixed models based on the SCAD and the LASSO penalty select about the same number of genes (29 vs. 30).
- Active set becomes significantly smaller in the mixed model set-up compared to the ordinary linear regression models.
- Among the ten largest estimated β's (in absolute value), **six are known to be associated with inflammation for the SCAD penalty**, while only four known genes are picked up by the L_1 penalty.
- The estimated random intercept variation is zero in the presence of the gene expressions.
- The error variance σ^2 is also reduced slightly for the SCAD penalty as compared to the L_1 penalty.

Outline

Summary

- We Discussed the LASSO and its properties, along with the required assumptions
- We discussed different possible generalization of the LASSO both in terms of loss function and the penalty.
- We discussed the theoretical properties of the general penalized estimators, for non-convex losses and nonconcave penalties.
- An application in the context of high-dimensional linear mixed-model has been illustrated.

References

- Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Stat.*, **37**, 1705–1732.
- Bühlmann, P. and Meier, L. (2008). Discussion of “One-step sparse estimates in nonconcave penalized likelihood models" (auths H. Zou and R. Li), *Ann. Stat.*, **36**, 1534–1541.
- Bühlmann, P., and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.
- Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–71.
- Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-p sparse GLM. *Statist. Sinica*, **22**, 555–74.
- Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Amer. Statist. Assoc.*, **96**, 1348–1360.
- Fan J. and Liao Y. (2014). Endogeneity in high dimensions. *Ann. Stat.*, **42(3)**, 872–917.
- Fan, J., and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Info. Theory*, **57(8)**, 5467–5484.
- Ghosh, A. and Thoresen, M. (2018). Non-concave penalization in linear mixed-effect models and regularized selection of fixed effects. *AStA Adv. Stat. Anal.*, **102(2)**, 179–210.
- Kim, Y. , Kwon, S. , and Choi, H. (2012). Consistent Model Selection Criteria on High Dimensions. *J. Mach. Learn. Res.*, **13**, 1037–1057.
- Loubes, J. M. and van de Geer (2002). Adaptive estimation in regression, using soft thresholding type penalties. *Statistica Neerlandica*, **56**, 453–478.
- Meinshausen, N. (2007). Relaxed Lasso. *Comput. Stat. Data Anal.*, **52**, 374–393.
- Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Stat.*, **34(3)**, 1436–1462.
- Ottestad I., RetterstÃ¸l K., Myhrstad M.C., Andersen L.F., Vogt G., Nilsson A., et al. (2013). Intake of oxidised fish oil does not affect circulating levels of oxidised LDL or inflammatory markers in healthy subjects. *Nutrition, Metabolism and Cardiovascular Diseases*, **23(1)**, 3–4.

References

- Pinheiro J.C. and Bates D.M. (2000). *Mixed-effects models in S and S-plus.* Springer-Verlag, New York.
- Schelldorfer, J., Bühlmann, P. and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using l_1 −penalization. *Scand. J. Stat.*, **32**, 2, 197-214.
- Schelldorfer. J. , Meier, L. and Bühlmann, P. (2014). GLMM Lasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using l_1 −penalization. *J. Comput. Graphical Stat.*, **23**, 460–477.
- Städler, N.; Bühlmann, P. and van de Geer, S. (2010). l_1 −penalization for mixture regression models (with discussion). *Test*, **19**, 209–285.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Royal Stat. Soc. B*, **58**(1), 267-288.
- van de Geer, S.A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Stat.*, **36**, 614–645.
- van de Geer, S.A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, **3,** 1360–1392.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Royal Stat. Soc. B*, **68**, 1, 49–67.
- Zhang, C.H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, **38**, 894–942.
- Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Machine Learn. Res.*, **7**, 2541–2563.
- Zhou, S. (2010). Thresholded Lasso for high dimensional variable selection and statistical estimation. *arXiv preprint*, arXiv:1002.1583.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Stat. Assoc.*, **101**, 1418–1429.

THANK YOU!

Contact me at:
abhik.ghosh@isical.ac.in