A Survey on High-Dimensional Regression
MATH 533 Final Project

Jiajun Zhang* Caroline Wang† James Liang‡

December 4, 2025

Abstract

Modern applications of statistical theory and methods may involve extreme datasets with huge number of measurements compared to relatively small sample sizes. High-dimensional data occurs when the number of covariates (p) is greater than the number of observations (n). In this article, we first address the drawbacks of the ordinary least squares estimate and investigate some datasets where high-dimensionality is present. We then introduce some popular regularization methods and a brief introduction to variable selection. We end this article by providing an overview of a high-dimensional classification problem and address its modern challenges.

Contents

*Department of Mathematics and Statistics, jiajun.zhang@mail.mcgill.ca, McGill ID: 261105766
†Department of Mathematics and Statistics, caroline.wang2@mail.mcgill.ca, McGill ID: 260548387
‡Department of Mathematics and Statistics, james.liang2@mail.mcgill.ca, McGill ID: 261180286

1 Introduction

1.1 Drawbacks of the Ordinary Least Square Estimator

Consider the dataset $\mathcal{D}_n := \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_i, y_i)\}$ where $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, we denote $\boldsymbol{x}_i$ as the covariates and y_i as the response. In linear regression we have the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)^\top$, $\boldsymbol{X} := (\boldsymbol{x}_i, \cdots, \boldsymbol{x}_n)^\top$, $\boldsymbol{\beta}$ is the coefficient vector and $\boldsymbol{\epsilon}$ is a noise term such that $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$. Under full rank, fixed design the ordinary least square estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{1.1}$$

In a Gaussian linear model, we assume $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \mathbb{I}_n)$ as well as fixed design model, then the mean of *prediction error* of the least square estimator is given by

$$\mathbb{E}\left[\frac{||\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||_2^2}{n}\right] = \frac{p}{n} \cdot \sigma^2. \tag{1.2}$$

It suggests that the least squares estimate only provides accurate prediction in the case where $p << n$ or the variance of the noise σ^2 is very small (Lederer [Led22]). In a high dimensional setting where $p > n$, the prediction risk will be large and yield inaccuracies. Another way to see this is to consider the *estimation error* of $\widehat{\boldsymbol{\beta}}$, which is the L^2 distance from $\boldsymbol{\beta}$ to $\widehat{\boldsymbol{\beta}}$ defined as $L = ||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}||_2$, then we have

$$\mathbb{E}[L^2] = \mathbb{E}\Big[(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\Big] = \sigma^2 \cdot \operatorname{trace}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \sigma^2 \cdot \sum_{i=1}^{p} \lambda_i^{-1},$$

where $\lambda_i, i \in [p]$ are the eigenvalues of the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ (the matrix is positive definite). Under a Gaussian linear model $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \mathbb{I}_n)$ we have

$$\mathbb{V}(L^2) = 2\sigma^4 \cdot \operatorname{trace}(\boldsymbol{X}^\top \boldsymbol{X})^{-2} = 2\sigma^2 \cdot \sum_{i=1}^{p} \lambda_i^{-2}.$$

Hence the least squares estimate will again yield inaccuracies if $\boldsymbol{X}^\top \boldsymbol{X}$ moves from a unit matrix to a ill-conditioned one with small eigenvalues (Hoerl and Kennard [HK70]).

1.2 High Dimensional Data and Simulations

The words "high dimensional" refers to the situations where the number of covariates is larger than the number of data points, i.e., $p > n$. Such situations happen in many fields nowadays, including genomics, text classification, image analysis, astronomy, atmospheric science, etc. (Sirimongkolkasem and Drikvandi [SD19]; Johnstone and Titterington [JT09]). In such case the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is not full (column) rank and hence the least squares estimate (1.1) will yield infinitely many solutions, leading to overfitting. Regularization methods such as Ridge regression (Hoerl and Kennard [HK70]), Lasso (Tibshirani [Tib96]), Elastic Net (Zhou and Hastie [ZH05]),

Bayesian Lasso (Park and Casella [PC08]), and many other were proposed for analysing high-dimensional data by imposing a penalty on the number of covariates. We briefly investigate a computer simulation model and two real dataset in this chapter.

1. A Computer Simulation: First consider a simulation of a Gaussian linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}),$$

with $p = 20$ and $n = 10$. For each $x_{ij} \in \boldsymbol{X}$ $(i \in [10], j \in [20])$, we set $x_{ij} \sim \mathcal{N}(0, (0.5)^{|i-j|})$ and the true coefficients vector as $\boldsymbol{\beta} = \begin{pmatrix} 1.5 & 2 & 3 & 2.5 & 1.8 & 0 & \cdots & 0 \end{pmatrix}$. We apply 3 different regularization methods: Ridge, Lasso, Elastic Net. Their corresponding estimates are

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}} := \arg\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2^2 \right\} \tag{1.3}$$

$$\widehat{\boldsymbol{\beta}}_{\text{Lasso}} := \arg\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right\} \tag{1.4}$$

$$\widehat{\boldsymbol{\beta}}_{\text{ENet}} := \arg\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 ||\boldsymbol{\beta}||_2^2 \right\}, \tag{1.5}$$

where the tuning parameters $\lambda, \lambda_1, \lambda_2$ are selected using Leave-One-Out Cross-Validation (LOOCV). We train the model on 9 data points and test it on the 1 remaining point, and repeat this process 10 times. The optimal candidate of the tuning parameter is chosen to minimize the average prediction error of all 10 tests. The performance is shown in figure 1 and table 1.

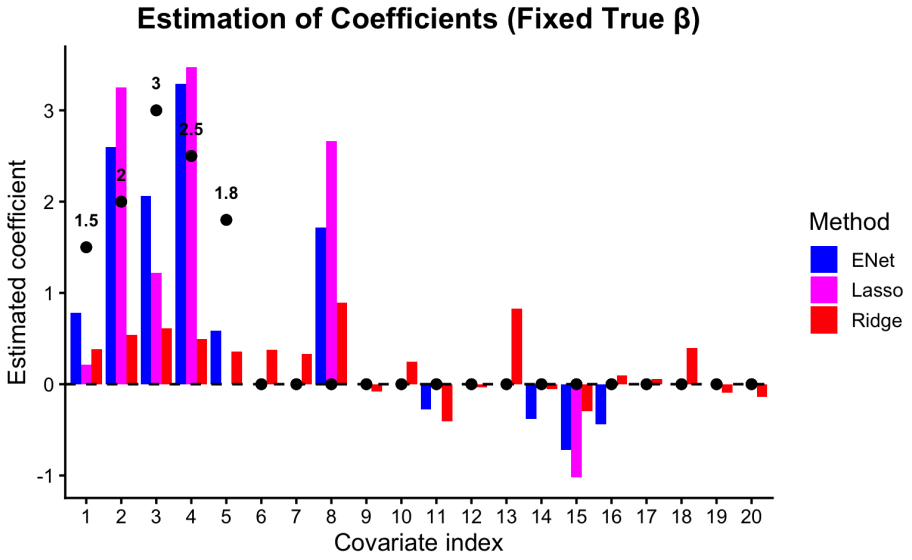


Figure 1: The estimation of each coefficient by different regularization methods where the black dots indicate the true value.

Method	Estimation Error	Prediction Error	Sparsity Index
Lasso	2.1835	1.0921	5
Ridge	3.9478	1.3586	20
Elastic Net	2.5456	1.1034	6

Table 1: Comparison of estimation error ($||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2^2$); prediction error ($||\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$) and sparsity index ($\sum_{j=1}^{p} \mathbf{1}\{\boldsymbol{\beta}_j \neq 0\}$) for Ridge, Lasso and Elastic net.

2. Riboflavin Production: We next investigate a real dataset about riboflavin (vitamin B_2) production by bacillus subtilis. The data contains $n = 71$ observations and $p = 4088$ gene-expression covariates measuring the logarithm of the expression level of 4088 genes. A previous analysis was conducted by Bühlmann and Van de Geer ([BdG11]). They applied FMRLasso (Finite Mixture of Regression with an L^1 Penalty) and then selected the top 20 most influential genes based on the sum of the absolute estimated coefficients across the three mixture components. In our analysis, we employ an enhanced version of stability selection, incorporating repeated subsampling, Elastic Net regularization, and random feature weighting. The top 9 stability-selected features using Ridge, Lasso and Elastic Net is shown in figure 2:

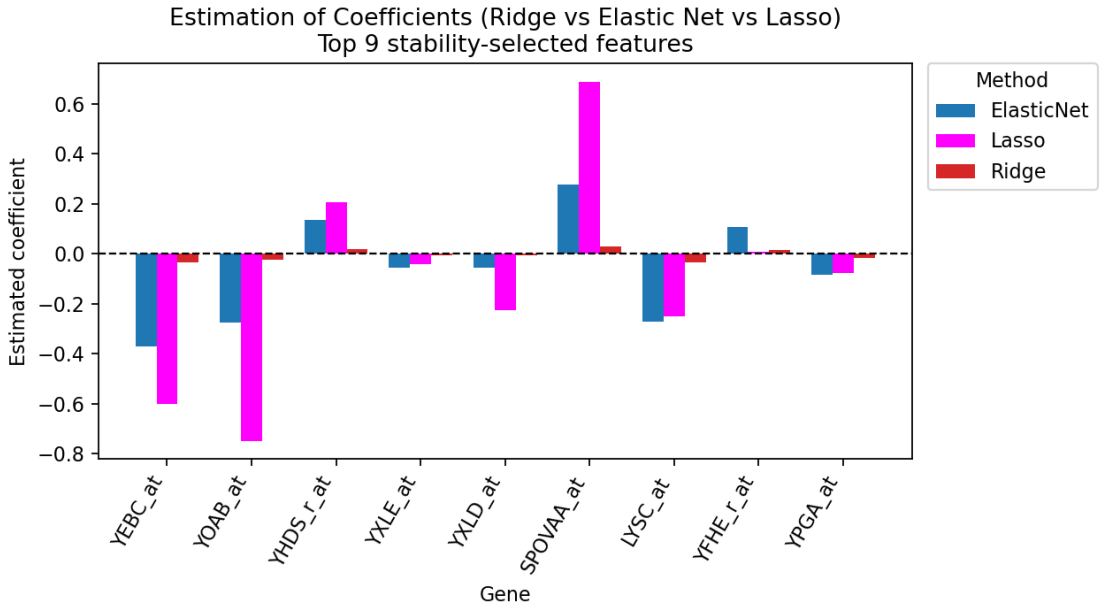


Figure 2: The estimated coefficient for the top 9 selected features (genes) using Ridge, Lasso and Elastic Net

3. IMDb Movie Review: We analyze movie reviews from IMDb using the Large Movie Review Dataset by Maas et al ([MDP+11]) from Stanford University to evaluate the effectiveness of regularization methods on real world high-dimensional datasets. The dataset contains 25,000 plain-text reviews $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{25000}$ and corresponding ratings y from 0 to 10. We focus on a random subset of 10,000 samples for training and evaluate models on the full 25,000 samples in the test set. Reviews are categorized as negative for ratings 1–4 and positive for ratings 7–10. For instance, `pos/154_8.txt` corresponds to the 154th positive review with rating 8, while `neg/7_3.txt` corresponds to the 7th negative review with rating 3. The reviews have an average length of 233 words with a standard deviation of 174 words.

We convert the text data into numerical covariates using TF–IDF (Term Frequency–Inverse Document Frequency). TF–IDF assigns each word w in a document d a score based on its term frequency (the number of times w appears in d divided by the total number of words in d) and its inverse document frequency (the total number of documents divided by the number of documents containing w). This preprocessing yields $p = 51,563$ covariates.

We consider four linear models: Unregularized Linear Regression, Lasso regression, Ridge regression, and ElasticNet. These four models are implemented using Scikit-Learn, a Python library implementing various machine learning models. Unregularized Linear Regression is implemented

by performing Ordinary Least Squares Regression using an SVD-based solver, which provides a stable pseudoinverse in high-dimensional settings. Lasso regression modifies this by introducing an l_1 penalty to encourage coefficient sparsity, Ridge regression introduces an l_2 penalty to shrink down coefficients smoothly, while ElasticNet uses both an l_1 and l_2 penalty to regularize the coefficients.

For Lasso, Ridge, and ElasticNet, we perform hyperparameter tuning to select the best α using 5-fold cross-validation, computing the root mean squared error (RMSE) as the evaluation metric (Table 2). In the case of ElasticNet, the l_1 ratio is fixed to 0.5 to prevent it from collapsing to Ridge regression. The hyperparameter search is performed in two steps: first, we explore $\alpha \in \{10, 1, 0.1, 0.01, 0.001, \dots\}$ to identify the order of magnitude with the lowest RMSE, then refine the search around that order of magnitude.

Model	**RMSE (avg ± std)**	**Selected α**
Lasso	2.275 ± 0.053	2e-4
Ridge	**2.204 ± 0.048**	0.9
Elastic Net	2.226 ± 0.047	1e-4
Unregularized Linear Regression	2.533 ± 0.047	

Table 2: RMSE of Lasso, Ridge, Elastic Net, and Unregularized Linear Regression on the ACL IMDb dataset (5-fold cross-validation)

From table 2, Ridge regression performs best on the IMDb dataset with an average RMSE of 2.20, followed closely by ElasticNet with 2.23. The worst-performing model is unregularized Linear Regression with an RMSE of 2.53, likely due to the high dimensionality ($\sim$ 50,000 features), correlated covariates, and many non-informative variables.

We can further analyze the most influential features by examining the absolute values of the model coefficients, $|\beta_i|$, using the best hyperparameter α for each model. For Unregularized Linear Regression, the top 10 features are listed in Table 3, while for Ridge regression, the top features are listed in Table 4. It can be seen that unregularized Linear Regression overfits to infrequent words such as "turkey," "minton," and "audition," while Ridge regression emphasizes words that frequently appear in reviews, such as "worst," "awful," "amazing," and "excellent." The difference in the top 10 features between regularized and unregularized linear regression highlights how regularization prevents the model from assigning disproportionately high coefficients to rare or non-informative words, thereby avoiding overfitting while still achieving a lower RMSE.

Index	Feature	Coefficient	Abs. Coefficient
0	units	11.215	11.215
1	worst	-11.171	11.171
2	minton	9.782	9.782
3	audition	-9.569	9.569
4	awful	-9.267	9.267
5	appreciated	9.052	9.052
6	disappointment	-8.988	8.988
7	accepting	8.841	8.841
8	precictable	-8.835	8.835
9	turkey	-8.805	8.805
10	17	-8.710	8.710

Table 3: Top 10 features by absolute coefficient for the unregularized linear regression model.

Index	Feature	Coefficient	Abs. Coefficient
0	worst	-11.732	11.732
1	awful	-8.300	8.300
2	great	7.539	7.539
3	favorite	7.032	7.032
4	amazing	6.907	6.907
5	best	6.777	6.777
6	nothing	-6.711	6.711
7	excellent	6.360	6.360
8	wonderful	6.156	6.156
9	bad	-5.974	5.974
10	boring	-5.856	5.856

Table 4: Top 10 features by absolute coefficient for the Ridge regression model.

Although RMSE evaluates regression performance, we also adapt the regression outputs for sentiment classification: predicted ratings above 5 are labeled positive and ratings below 5 negative and compute the classification accuracy.

Model	**Accuracy (%)**	**RMSE**
Lasso	84.640	2.293
Ridge	**85.448**	**2.263**
Elastic Net	85.328	2.270
Unregularized Linear Regression	78.836	2.741

Table 5: Accuracy and RMSE of Lasso, Ridge, ElasticNet and Unregularized Linear Regression on the ACL IMDb dataset (test split)

Additionally, we can determine the number of significant covariates used by each regularized model. In the case of Lasso, we select all non-zero coefficients, while in the case of both Ridge and ElasticNet, we deem a coefficient to be significant enough if its norm is more than a threshold (1e-5 for this experiment). Lasso has 1,987 covariates with non-zero coefficients, Ridge has 51,563 (all) features above the threshold, and ElasticNet only has 11,074 features above the threshold.

Lasso, ElasticNet, and Ridge all achieve comparable accuracy, but Lasso relies on far fewer features. Consequently, for this experiment, Lasso or ElasticNet is preferred, delivering near-best performance with 84.64% accuracy and an RMSE of 2.29 for Lasso, and 85.34% accuracy with an RMSE of 2.27 for ElasticNet, while using approximately 20 and 10 times fewer features than the full TF–IDF representation of 51,563 covariates.

2 Regularization

2.1 Ridge and Lasso Regression

As we have discussed in section 1.1, ordinary least squares estimate in general may not perform well under certain settings. In order to control the inflation and instability associated with the least squares estimates, Horel ([HK70]) suggests the following form

$$\widehat{\boldsymbol{\beta}}_R := \arg\min ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2 \quad \lambda > 0, \tag{2.1}$$

with a L^2 penalty. Since we have a convex combinations of functions, it can easily be shown that the solution to (2.1) is given by

$$\widehat{\boldsymbol{\beta}}_R := (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \mathbb{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

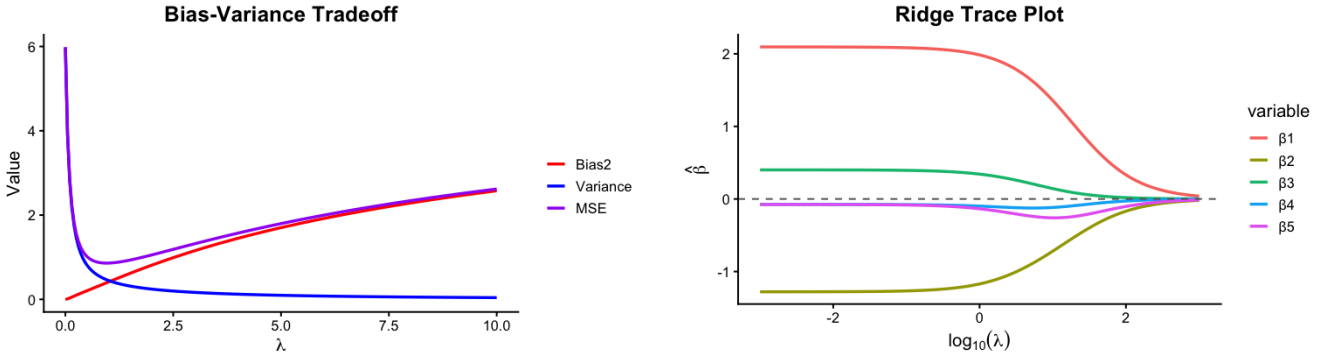
The matrix $\boldsymbol{Z} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \mathbb{I})^{-1}$ is always invertible for $\lambda > 0$ since all eigenvalues of $\boldsymbol{Z}$ are strictly positive and hence $\boldsymbol{Z}$ is positive definite. The mean and variance of the Ridge estimate are given by

$$\mathbb{E}\Big[\widehat{\boldsymbol{\beta}}_R \Big| \boldsymbol{X}\Big] = \boldsymbol{Z}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \quad \mathbb{V}\Big(\widehat{\boldsymbol{\beta}}_R \Big| \boldsymbol{X}\Big) = \sigma^2 \boldsymbol{Z}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{Z}^\top.$$

It suggests that Ridge estimate $\widehat{\boldsymbol{\beta}}_R$ is biased, but as λ increases, the bias increases and the variance decreases and hence there is a natural bias-variance tradeoff process (see figure 3a). It is not hard to see that if $\lambda = 0$ then it reduces to the least square estimate, and $\widehat{\boldsymbol{\beta}}_R \to 0$ as $\lambda \to \infty$ (see figure 3b). A rigorous treatment showed that the estimation error of $\widehat{\boldsymbol{\beta}}_R$ satisfies the form

$$\begin{aligned}\mathbb{E}\Big[(\widehat{\boldsymbol{\beta}}_R - \boldsymbol{\beta})^\top (\widehat{\boldsymbol{\beta}}_R - \boldsymbol{\beta})\Big] &= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + \lambda)^2} + \lambda^2 \boldsymbol{\beta}^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \mathbb{I})^{-2} \boldsymbol{\beta} \\ &= \gamma_1(\lambda) + \gamma_2(\lambda),\end{aligned}$$

where $\lambda_i, i \in [p]$ are the eigenvalues of $\boldsymbol{Z}$. The first term can be viewed as the sum of total variances of the parameter estimates and the last term can be viewed as the square of bias when $\widehat{\boldsymbol{\beta}}_R$ is used rather than the least square estimate $\widehat{\boldsymbol{\beta}}$. The variance function $\gamma_1(\lambda)$ is a continuous, monotonically de-creasing function of λ, and the square bias $\gamma_2(\lambda)$ is a continuous monotonically increasing function of λ (Hoerl and Kennard, [HK70]), which is also illustrated in figure 3a.



(a) Bias–variance tradeoff for Ridge regression. As λ increase the bias increase and variance decrease.

(b) Ridge trace plot showing coefficient shrinkage as λ increases.

One may also view (2.1) as the following optimization problem:

$$\widehat{\boldsymbol{\beta}}_R := \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{subject to } ||\boldsymbol{\beta}||_2^2 \leq \lambda.$$

Hence we may view Ridge as a Lagrange multiplier like problem where we minimize the least squares estimate under the constraint $||\boldsymbol{\beta}||_2^2 \leq \lambda$. Note that we have

$$\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 := (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}), \tag{2.2}$$

hence Ridge estimates can be viewed as the first point on the constraint that touches the contour, as shown in figure 4 for a special case when $p = 2$:

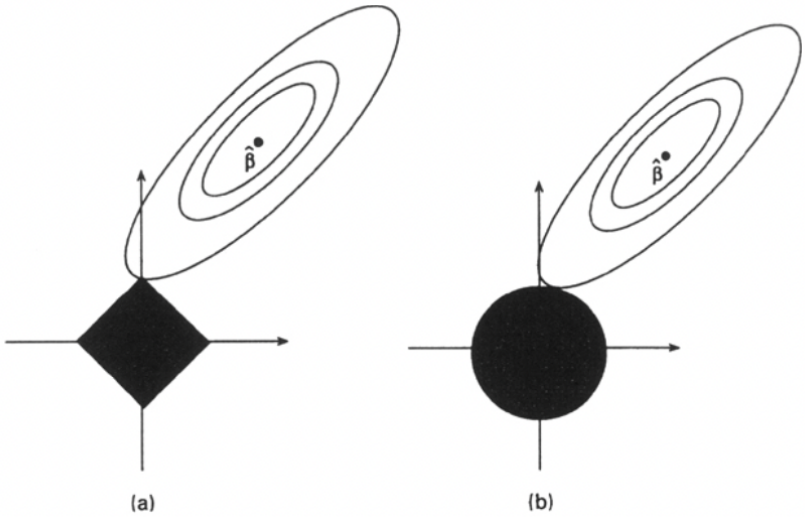


Figure 4: The celliptical contours of the function $(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})$ with Lasso (a) and Ridge (b) regression when $p = 2$. Figure is taken from Tibshirani [Tib96].

From the geometry of Ridge regression we notice that the contour will rarely hit the constraint where one covariate is zero, and hence Ridge estimates will result in non-zero estimates for all covariates with the feature that shrinks non-dominant covariates towards zero. This idea can be verified through Figure 1 of the simulation in section 1.2, where Ridge regression tends to shrink all non-dominant covariates with index > 5 towards zero, but never reach zero despite the fact that their true coefficients are zero.

Another issue with Ridge regression is that when p is large, $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \mathbb{I}$ is a $p \times p$ matrix which still potentially be very large. A dominant strategy is try to exploit sparsity where the response y is dominated by only a few covariates and many elements of $\boldsymbol{\beta}$ are zero. (Johnstone and Titterington [JT09]), also see example in section 2.1). Especially in the case when $p > n$, it is often scientifically plausible that only a small proportion are likely to be influential as predictors. With this and the shrinkage property of Ridge regression in mind, we may try different penalty functions so that from the geometric point of view, the OLS contour may touch the "corner" of the penalty constraint and result in zero coefficients solutions. Tibshirani ([Tib96]) proposed the Lasso (least absolute shrinkage and selection operator) estimates of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_L := \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{subject to } ||\boldsymbol{\beta}||_1 \leq \lambda, \tag{2.3}$$

with the L^1 norm penalty. In a special case of $p = 2$, the Lasso constraint is shown in figure 4, and we can see that now the OLS contour have a better chance to touch the corner of the contour, corresponding to a zero coefficient. This idea can also be verified in the simulation study in section 1.2, and from figure 1 we see that the coefficients for Lasso estimate are mostly zero for covariate index > 5.

2.2 Asymptotic Properties of Lasso

The Lasso solution $\widehat{\boldsymbol{\beta}}_L$ defined in (2.3) can also be written in the penalized form

$$\widehat{\boldsymbol{\beta}}_L := \arg\min_{\boldsymbol{\beta}} \left\{ \frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2}{n} + \lambda ||\boldsymbol{\beta}||_1 \right\}. \tag{2.4}$$

Assume a fixed design, Gaussian linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, also by rescaling $\boldsymbol{X}$ we assume $\sum_{j=1}^n \boldsymbol{X}_{(ij)}^2 = n, i \in [p]$, we get the *basic inequality* by direct derivation from (2.4) (Also see Chapter 6 of Bühlmann and Van de Geer [BdG11], Tibshirani [Tib23]):

Lemma 1 (Basic Inequality)**.**

$$\frac{||\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta})||_2^2}{n} + \lambda ||\widehat{\boldsymbol{\beta}}_L||_1 \leq \frac{2\boldsymbol{\epsilon}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta})}{n} + \lambda ||\boldsymbol{\beta}||_1.$$

Note that by Hölder's inequality we have

$$|\boldsymbol{\epsilon}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta})| \leq ||\boldsymbol{\epsilon}^\top \boldsymbol{X}||_\infty \cdot ||\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}||_1,$$

hence

$$\begin{aligned} \frac{1}{n}||\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta})||_2^2 &\leq \frac{2}{n}||\boldsymbol{\epsilon}^\top \boldsymbol{X}||_\infty \cdot ||\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}||_1 + \lambda\Big(||\boldsymbol{\beta}||_1 - ||\widehat{\boldsymbol{\beta}}_L||_1\Big) \\ &\leq 2\lambda ||\boldsymbol{\beta}||_1, \end{aligned}$$

provided that $\lambda \geq \frac{2}{n}||\boldsymbol{\epsilon}^\top \boldsymbol{X}||_\infty$. Note that under our assumption we also have

$$\frac{\boldsymbol{\epsilon}^\top \boldsymbol{X}^{(i)}}{\sigma\sqrt{n}} \sim \mathcal{N}(0, 1).$$

To give an upper bound of $||\boldsymbol{\epsilon}^\top \boldsymbol{X}||_\infty$, we will use a lemma from i.i.d Gaussian random variables:

Lemma 2. *Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$, then for any $t > 0$, we have*

$$\mathbb{P}\Big(\max_{i\in[n]} |X_i| \geq \sqrt{t^2 + 2\log n}\Big) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Proof. We have that

$$\begin{aligned} \mathbb{P}\Big(\max_{i\in[n]} |X_i| \geq \sqrt{t^2 + 2\log n}\Big) &:= \mathbb{P}\Big(\exists i \in [n], |X_i| \geq \sqrt{t^2 + 2\log n}\Big) \\ &= \bigcup_{i\in[n]} \mathbb{P}\Big(|X_i| \geq \sqrt{t^2 + 2\log n}\Big) \\ &\leq n\mathbb{P}\Big(|X_1| \geq \sqrt{t^2 + 2\log n}\Big) \quad \text{(Union bound)} \\ &= 2n\mathbb{P}\Big(X_1 \geq \sqrt{t^2 + 2\log n}\Big) \\ &\leq 2n\exp\left(-\frac{t^2 + 2\log n}{2\sigma^2}\right) \quad \text{(Chernoff bound)} \\ &= 2\exp\left(-\frac{t^2}{2\sigma^2}\right). \end{aligned} \tag{2.5}$$

□

Hence we have

$$\mathbb{P}\left(\max_{i\in[p]}\left|\boldsymbol{\epsilon}^\top \boldsymbol{X}^{(i)}\right| \geq \sigma\sqrt{2n(\log(2p)+t)}\right) \leq e^{-t},$$

for all $t > 0$. Taking $\lambda = \sigma\sqrt{2n(\log(2p)+u)}$, we have

$$\mathbb{P}\left(\frac{1}{n}||\boldsymbol{X}\widehat{\boldsymbol{\beta}}_L - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \leq 4\sigma||\boldsymbol{\beta}||_1 \cdot \sqrt{\frac{2(\log(2p)+t}{n}}\right) \geq 1 - e^{-t}.$$

This bound yields the "slow" rate for the Lasso estimator: the in-sample prediction risk scales as $||\boldsymbol{\beta}||_1\sqrt{\frac{\log p}{n}}$ (Tibishirani [Tib23]). Furthermore, if the regularization λ is taken of order $\sqrt{\frac{\log p}{n}}$, while assuming the L^1 norm of the true $\boldsymbol{\beta}$ is of smaller order than $\sqrt{\frac{n}{\log p}}$, the Lasso is consistent (Bühlmann and Van de Geer [BdG11]).

If we do not assume a linear model, that is, the true model is $\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, we can get a so called *oracle inequality*:

Lemma 3 (Oracle Inequality)**.** *Assume a model $\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{\epsilon}$, let $\widehat{\boldsymbol{\beta}}$ be the Lasso solution under the constraint $||\boldsymbol{\beta}||_1 \leq \lambda$, then*

$$\mathbb{P}\left(\frac{||\boldsymbol{X}\widehat{\boldsymbol{\beta}} - f(\boldsymbol{X})||_2^2}{n} \leq \inf_{||\boldsymbol{\beta}||\leq\lambda}\left[\frac{||\boldsymbol{X}\boldsymbol{\beta} - f(\boldsymbol{X})||_2^2}{n}\right] + 4\sigma\lambda\sqrt{\frac{2(\log(2p)+t)}{n}}\right) \geq 1 - e^{-t}.$$

Thus, in terms of in-sample risk, we can expect the Lasso to perform close as well as the best L^1 linear predictor to f (see Tibshirani [Tib23]).

2.3 Variants of Lasso

Despite Lasso has some nice properties and has shown success in many situations, Zhou and Hastie ([ZH05]) showed its limitations by the following:

- In the $p > n$ case, the Lasso selects at most n variables because of the nature of the convex optimization problem.
- The Lasso is not well defined unless the bound on the L^1 norm of the coefficients is smaller than a certain value.
- For usual $n > p$ case, if there are high correlations between predictors then the prediction performance of the Lasso is dominated by Ridge regression.
- If there is a group of variables among which the pairwise correlations are high, then Lasso tends to select only one variable from the group and does not care which one is selected.

Proposed by Zhou and Hastie, the naïve elastic net can be viewed as a linear combination of L^1 and L^2 penalty:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left\{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda_2||\boldsymbol{\beta}||_2^2 + \lambda_1||\boldsymbol{\beta}||_1\right\}, \tag{2.6}$$

for $\lambda_1, \lambda_2 > 0$. Let $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$, (2.6) can be written in the following penalized form:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{subject to } (1-\alpha)||\boldsymbol{\beta}||_1 + \alpha||\boldsymbol{\beta}||_2^2 \leq t.$$

Zhou and Hastie ([ZH05]) also showed that, one may transform the naïve elastic net problem into an equivalent lasso problem on augmented data, by defining $\boldsymbol{X}^* \in \mathbb{R}^{(n+p)\times p}, \boldsymbol{y}^* \in \mathbb{R}^{n+p}$, where

$$\boldsymbol{X}^* = \sqrt{1+\lambda_2} \cdot \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\mathbb{I} \end{pmatrix}; \quad \boldsymbol{y}^* = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix},$$

then let $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$ we have

$$L(\boldsymbol{\beta}, \lambda_1, \lambda_2) = ||\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*||_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}||\boldsymbol{\beta}^*||_1. \tag{2.7}$$

Now we note that $\boldsymbol{X}^*$ has rank p so the naïve elastic net can potentially select all p predictors in all situations, and it can perform variable selection in a fashion similar to the Lasso. By minimizing (2.7) over $\boldsymbol{\beta}^*$, we have the relation between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}^*$ given by

$$\widehat{\boldsymbol{\beta}} = \frac{1}{\sqrt{1+\lambda_2}}\widehat{\boldsymbol{\beta}}^*.$$

The naïve elastic net has the ability of selecting 'grouped' variables. Qualitatively speaking, a regression method exhibits the grouping effect if the regression coefficients of a group of highly correlated variables tend to be equal up to change of sign. Zou and Hastie ([ZH05]) showed the following:

Theorem 1. *Given data* $(\boldsymbol{y}, \boldsymbol{X})$ *where* $\boldsymbol{y}$ *is centered and* $\boldsymbol{X}$ *is standardized. Let* $\widehat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ *be the naïve elastic net estimate, and suppose* $\widehat{\beta}_1(\lambda_1, \lambda_2)\widehat{\beta}_j(\lambda_1, \lambda_2) > 0$*, then*

$$\frac{1}{||\boldsymbol{y}||_1}|\widehat{\beta}_i(\lambda_1, \lambda_2) - \widehat{\beta}_j(\lambda_1, \lambda_2)| \leq \frac{1}{\lambda_2}\sqrt{2(1-\boldsymbol{x}_i^\top \boldsymbol{x}_j)}.$$

The theorem says if $\boldsymbol{x}_i, \boldsymbol{x}_j$ are highly correlated, i.e $\rho \approx 1$, then the difference between coefficient paths of predictor i and predictor j is almost 0, which shows that elastic net exhibits the property of grouping effect that Lasso does not have.

2.4 Bayesian View of Regression

Another approach is to use Bayesian statistics, where we introduce *prior distribution* $\pi(\beta)$ for unknown parameters, and we observe $p(y|\beta)$ as the conditional distribution of y given the parameter β. By Bayes theorem, we may get the *posterior distribution* as

$$p(\beta|y) = \frac{p(y|\beta)\pi(\beta)}{\displaystyle\int_\Theta p(y|\beta)\pi(\beta)d\beta}.$$

Tishirani ([Tib23]) suggested that the Lasso can be iterpreted as a Bayesian posterior mode estimate when the parameters β_i have i.i.d Laplace priors. With

$$\pi(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{\lambda}{2} e^{-\lambda|\beta_j|}, \tag{2.8}$$

and an independent prior $\pi(\sigma^2)$ on the variance of ϵ, the posterior distribution can be expressed as

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto \pi(\sigma^2)(\sigma^2)^{-(n-1)/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - \lambda\sum_{j=1}^{p}|\beta_j|\right\},$$

where $\boldsymbol{y}$ is the mean-centered response vector. Instead of the prior defined in (2.8), Park and Casella ([PC08]) proposed Bayesian lasso, where they used the following prior:

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j-1}^{p}\frac{\lambda}{2\sqrt{\sigma^2}}e^{-\lambda|\beta_i|/\sqrt{\sigma^2}}, \tag{2.9}$$

with an additional inverse gamma prior on σ^2:

$$\pi(\sigma^2) = \frac{\gamma^a}{\Gamma(a)}(\sigma^2)^{-a-1}e^{-\gamma/\sigma^2} \quad (a, \gamma > 0). \tag{2.10}$$

Park and Casella ([PC08]) showed that the joint posterior distribution of $\boldsymbol{\beta}, \sigma^2$ under priors (2.9) and (2.10) is proportional to

$$(\sigma^2)^{-(n+p-1)/2-a-1}\exp\left\{-\frac{1}{\sigma^2}\Big((\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})/2+\gamma\Big) - \frac{\lambda}{\sqrt{\sigma^2}}\sum_{j=1}^{p}|\beta_j|\right\}.$$

To choose the Bayesian lasso parameter λ, Park and Casella ([PC08]) showed that, if assuming a parametric model, then λ has a likelihood function that may be maximized to obtain the empirical Bayes estimate.

In other contexts, the analysis is not so straightforward especially if we have latent variables z which is unobservable. In this case the key quantity of interest is the conditional distribution $p(z, \beta|y)$. Johnstone and Titterington ([JT09]) briefly summarized two modern approaches:

- *Stochastic method.* Generate a large number of realizations from $p(z, \beta|y)$ where

  $$p(z, \beta|y) \propto p(y, z, \beta) = p(y, z|\beta)\pi(\beta).$$

  The corresponding realizations of β are a sample from $p(\beta|y)$ and for instance their average would be a good estimator of the true posterior mean of β.

- *Variational method.* One may choose a deterministic approximation $q_y(z, \beta)$ as close to $p(z, \beta|y)$ as possible, measured by Kullback-Leibler divergence

  $$KL(q_y, p) = \int_z\int_\beta q_y(z, \beta)\log\frac{q_y(z, \beta)}{p(z, \beta|y)}d\beta dz.$$

3 Variable Selections with Lasso

Before presenting the theoretical results on variable selection with the Lasso, we will first introduce the key concepts and assumptions that underlie its selection properties.

Let us first assume the following linear model for the entirety of the subsection:

$$y_i = \sum_{j=1}^{p} \psi_j(\boldsymbol{x}_i)\beta_j^0 + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\{\psi_i\}_{j=1}^p$ is a given dictionary, i.e. a set of original covariates as linear functions, $X_i \in \mathcal{X}, i = 1, \ldots, n$ is fixed design and $\epsilon_1, \ldots, \epsilon_n$ are noise variables.

Now, we define the *active set* corresponding to the Lasso estimator and the *true active set* of the underlying data-generating model. These notions form the foundation for understanding how and when the Lasso can successfully identify relevant predictors.

Definition 1 (Active set)**.** *For a given estimated coefficient vector $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$, the **active set** is the set of variables whose estimated coefficients are nonzero:*

$$\widehat{A} = \{j \in \{1, \ldots, p\} : \widehat{\beta}_j \neq 0\}.$$

In other words, the active set consists of the predictors that the Lasso has selected after applying its L^1-penalization.

Having defined the estimated active set $\widehat{A}$, it is natural to contrast it with the set of variables that truly influence the response. This motivates the notion of the true active set, which provides the benchmark against which variable selection performance is evaluated.

Definition 2 (True active set)**.** *Let $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)$ denote the true regression coefficients in the linear model*

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}.$$

*The **true active set** (or **active set of truth**) is defined as*

$$A^0 = \{j \in \{1, \ldots, p\} : \beta_j^0 \neq 0\}.$$

This set contains the variables that genuinely contribute to the data-generating mechanism.

Introducing both active sets leads to the fundamental objective of variable selection with the Lasso: the method aims to recover the true active set, ideally achieving $\widehat{A} = A^0$, or at least obtaining an approximation of it. Understanding the relationship between these sets is central to the theoretical analysis of when and why the Lasso succeeds in identifying relevant predictors in high-dimensional settings.

3.1 KKT conditions

To understand how the Lasso performs variable selection, it will be useful to examine the Karush–Kuhn–Tucker (KKT) conditions for the Lasso estimator. These conditions characterize precisely when a coefficient is set to zero or remains nonzero, and therefore play a central role in understanding the structure of the active set $\widehat{A}$.

Consider $\boldsymbol{\beta}_{\text{init}}$ to be the noiseless Lasso with tuning parameter λ_{init} such that

$$\boldsymbol{\beta}_{\text{init}} := \arg\min_{\boldsymbol{\beta}} \{\|f_\beta - f^0\|^2 + \lambda_{\text{init}}\|\boldsymbol{\beta}\|_1\},$$

where $f_\beta = \sum_{j=1}^p \beta_j \psi_j = x^\intercal \beta$ is the candidate function corresponding to any β and $f^0 := \sum_{j=1}^p \beta_j^0 \psi_j = x^\intercal \beta^0$ is the true regression function.

Now, define $S_{\text{init}} := \{j : \boldsymbol{\beta}_{\text{init},j} \neq 0\}$ to be its true active set and let Q be some probability measure on $\mathcal{C}$ and let $\|\cdot\|$ be the $L_2(Q)$ norm. Define

$$\Sigma = \int \psi^\intercal \psi dQ$$

be the Gram matrix whose entries $\sigma_{j,k} = (\psi_j, \psi_k)$ and $(\cdot,\cdot)$ is the inner product of $L_2(Q)$. Note that for a given index set S, consider the following submatrices:

$$\Sigma_{1,1}(S) := (\sigma_{j,k})_{j,k\in S}, \quad \Sigma_{2,2}(S) := (\sigma_{j,k})_{j,k\in S},$$

and

$$\Sigma_{2,1}(S) := (\sigma_{j,k})_{j\notin S,k\in S}, \quad \Sigma_{2,1}(S) := (\sigma_{j,k})_{j\in S,k\notin S}.$$

Then, the KKT condition can be written as

$$2\Sigma(\boldsymbol{\beta}_{\text{init}} - \boldsymbol{\beta}^0) = -\lambda_{\text{init}}\boldsymbol{\tau}_{\text{init}},$$

where $\boldsymbol{\tau}_{\text{init}} \in \mathbb{R}^p$ satisfies

$$\|\boldsymbol{\tau}_{\text{init}}\|_\infty \leq 1 \quad \text{and} \quad \boldsymbol{\tau}_{\text{init},j} = \text{sign}(\beta_{\text{init},j}) \text{ whenever } \beta_{\text{init},j} \neq 0, \quad j = 1,\ldots,p.$$

3.2 Irrepresentable condition

While the Lasso produces sparse solutions, the recovery of the true active set A^0 requires additional assumptions on the design matrix. In particular, Bühlmann and van de Geer ([BdG11]) show that the *irrepresentable condition* is essentially necessary and sufficient for perfect variable selection. We briefly introduce this condition below.

Definition 3 (Irrepresentable condition)**.** *The* irrepresentable condition *is said to be met for a set S with a cardinality of s such that*

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\boldsymbol{\tau}_S\|_\infty < 1 \quad \forall \boldsymbol{\tau}_S \in R^S \text{ satisfying } \|\boldsymbol{\tau}_S\|_\infty \leq 1.$$

For a fixed $\boldsymbol{\tau}_s \in \mathbb{R}^s$ where $\|\boldsymbol{\tau}_s\| \leq 1$, the weak irrepresentable condition *holds for $\boldsymbol{\tau}_s$ if*

$$\|\Sigma_{2,1}(s)\Sigma_{1,1}^{-1}(s)\boldsymbol{\tau}_s\|_\infty \leq 1.$$

For a given $0 < \theta < 1$, the θ-uniform irrepresentable condition *is met for some set S if*

$$\max_{\|\tau_s\|_\infty \leq 1} \|\Sigma_{2,1}(s)\Sigma_{1,1}^{-1}(s)\boldsymbol{\tau}_s\| \leq \theta.$$

While the irrepresentable condition is essential for preventing false positives, it is not sufficient on its own to guarantee perfect variable selection. Even under the irrepresentable condition, the Lasso may still fail to recover the true active set A^0 if some of the nonzero coefficients β_j^0 are too small to be distinguished from noise after penalization.

This motivates the introduction of an additional requirement, often called the **β-min condition**, which ensures that all truly active coefficients are sufficiently large in magnitude for the Lasso to detect them. In combination with the irrepresentable condition, the β-min condition provides sign consistency and makes perfect support recovery possible.

3.3 β-min condition

The difficulty of variable selection depends strongly on the magnitude of the nonzero coefficients. We define

$$|\boldsymbol{\beta}^0|_{\min} := \min_{j \in S_0} |\beta_j^0|,$$

to be the smallest nonzero coefficient of the true model. A lower bound on $|\beta^0|_{\min}$ is known as a *β-min condition*.

Even under the irrepresentable condition, perfect support recovery is impossible when some true coefficients are too small to be distinguished from noise after penalization. The β-min condition ensures that every active coefficient is sufficiently large for the Lasso to detect it, and is therefore necessary for sign consistency. Combined together, the irrepresentable condition and the β-min condition capture the fundamental difficulties of model selection with Lasso: prevent inactive variables from entering the model and ensuring that the active coefficients are large enough to be detectable. Both conditions motivate not only the behavior of Lasso, but also turn out to characterize when the perfect variable selection is theoretically possible.

In particular, Bühlmann and van de Geer [BdG11] show that the irrepresentable condition is sufficient for sign consistency, and conversely, that the irrepresentable condition is essentially necessary for any estimator of Lasso type to recover the true active set S_0. We summarize these necessity and sufficiency results below.

3.4 Necessity & Sufficiency for variable selection

Let

$$S_0^{\text{relevant}} := \{j : |\boldsymbol{\beta}_j^0 > \frac{\lambda_{\text{init}} \sup_{\|\tau_{S_0}\|} \|\Sigma_{1,1}^{-1}(S_0)\tau_{S_0}\|_\infty}{2}.$$

Theorem 2. ***Part 1***. *Suppose the irrepresentable condition is satisfied for S_0. Then $S_0^{\text{relevant}} \subset S_{\text{init}} \subset S_0$, and*

$$\|(\boldsymbol{\beta}_{\text{init}})_{S_0} - \boldsymbol{\beta}_{S_0}^0\|_\infty \leq \frac{\lambda_{\text{init}} \sup_{\|\tau_{S_0}\| \leq 1} \|\Sigma_{1,1}^{-1}(S_0)\boldsymbol{\tau}_{S_0}\|_\infty}{2}.$$

In other words, this would mean that the $\beta_{\text{init},j}$ have the same sign as the β_j^0 for all $j \in S_0^{\text{relevant}}$.

Part 2. *Conversely, suppose that $S_0^{\text{relevant}} = S_0$ and that $S_{\text{init}} \subseteq S_0$. Under these assumptions, the weak irrepresentable condition is satisfied for the sign vector $\boldsymbol{\tau}_{S_0}^0 := \text{sign}(\beta_{S_0}^0)$.*

Proof. **Part 1**. Let S_0, S_0^C be the true active set and its complement respectively. Take the *KKT conditions* and split vectors and matrices according to S_0 and S_0^c, we get the following two equations:

$$2\Sigma_{1,1}(S_0)\left((\boldsymbol{\beta}_{\text{init}})_{S_0} - \boldsymbol{\beta}_{S_0}^0\right) + 2\Sigma_{1,2}(S_0)(\boldsymbol{\beta}_{\text{init}})_{S_0^c} = -\lambda_{\text{init}}(\boldsymbol{\tau}_{\text{init}})_{S_0},$$
$$2\Sigma_{2,1}(S_0)\left((\boldsymbol{\beta}_{\text{init}})_{S_0} - \boldsymbol{\beta}_{S_0}^0\right) + 2\Sigma_{2,2}(S_0)(\boldsymbol{\beta}_{\text{init}})_{S_0^c} = -\lambda_{\text{init}}(\boldsymbol{\tau}_{\text{init}})_{S_0^c}.$$

Now, assume that some coefficient outside of S_0 is nonzero, i.e. $\|(\boldsymbol{\beta}_{\text{init}})_{S_0^c}\|_1 \neq 0$.

After some algebraic operations on both equations, subtract the second equation from the first above, the term $\left((\boldsymbol{\beta}_{\text{init}})_{S_0} - \boldsymbol{\beta}_{S_0}^0\right)$ cancels giving

$$2(\boldsymbol{\beta}_{\text{init}})_{S_0^c}^\intercal \Sigma_{2,2}(S_0)(\boldsymbol{\beta}_{\text{init}})_{S_0^c} - 2(\boldsymbol{\beta}_{\text{init}})_{S_0^c}^\intercal \Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)\Sigma_{1,2}(S_0)(\boldsymbol{\beta}_{\text{init}})_{S_0^c}$$
$$= \lambda_{\text{init}}[(\boldsymbol{\beta}_{\text{init}})_{S_0^c}^\intercal \Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)(\tau_{\text{init}})_{S_0} - \|(\boldsymbol{\beta}_{\text{init}})_{S_0^c}\|],$$

where $(\boldsymbol{\beta}_{\text{init}})_j \boldsymbol{\tau}_{\text{init},j} = |\boldsymbol{\beta}_{\text{init},j}|$. Using the irrepresentable condition, we can show that the right-side of this equation is strictly negative when $\|(\boldsymbol{\beta}_{\text{init}})_{S_0^c}\|_1 \neq 0$. As for the left-hand side, after some rearranging the terms, we can observe that the matrix

$$\Sigma_{2,2}(S_0) - \Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)\Sigma_{1,2}(S_0),$$

is a Schur complement, which is known to be positive semi-definite, proving positivity for the left-hand side.

However, this is impossible as we have shown previously that the right-hand side of the equation is stricly negative. Thus the initial assumption was wrong, meaning we must have

$$\|(\boldsymbol{\beta}_{\text{init}})_{S_0^c}\|_1 = 0,$$

i.e. $(\boldsymbol{\beta}_{\text{init}})_{S_0^c} = 0$ implying that $S_{\text{init}} \subset S_0$.

Now, we plug this new result into the *KKT equations*:

$$2\Sigma_{1,1}(S_0)\left((\boldsymbol{\beta}_{\text{init}})_{S_0} - \boldsymbol{\beta}^0_{S_0}\right) = -\lambda_{\text{init}}(\boldsymbol{\tau}_{\text{init}})_{S_0}, \tag{3.1}$$

$$2\Sigma_{2,1}(S_0)\left((\boldsymbol{\beta}_{\text{init}})_{S_0} - \boldsymbol{\beta}^0_{S_0}\right) = -\lambda_{\text{init}}(\boldsymbol{\tau}_{\text{init}})_{S_0^c}. \tag{3.2}$$

By taking the $l_\infty-$norm of (3.1), we get

$$\|(\boldsymbol{\beta}_{\text{init}})_{S_0} - \boldsymbol{\beta}^0_{S_0}||_\infty \leq \lambda_{\text{init}} \frac{\|\Sigma_{1,1}^{-1}(S_0)\boldsymbol{\tau}_{S_0}\|_\infty}{2} \leq \lambda_{\text{init}} \sup_{\|\tau_{S_0}\|\leq 1} \frac{\|\Sigma_{1,1}^{-1}(S_0)\boldsymbol{\tau}_{S_0}\|_\infty}{2}.$$

Now, let $j \in S^{\text{relevant}}$. If $\beta_{\text{init}} = 0$, we show that

$$|\boldsymbol{\beta}^0_j| \leq \lambda_{\text{init}} \sup_{\|\boldsymbol{\tau}_{S_0}\|\leq 1} \frac{\|\Sigma_{1,1}^{-1}(S_0)\boldsymbol{\tau}_{S_0}\|_\infty}{2}.$$

However, this result contradicts the β-min condition defining S_0^{relevant}. So, for all $j \in S_0^{\text{relevant}}$, we must have $\beta_{\text{init},j} \neq 0$ which gives $S_0^{\text{relevant}} \subset S_{\text{init}}$. Combining with the conclusion of $S_{\text{init}} \subset S_0$, it follows that

$$S_0^{\text{relevant}} \subset S_{\text{init}} \subset S_0.$$

□

Proof. **Part 2.** We want to show that the weak irrepresentable condition for $\tau^0_{S_0}$ is a necessary condition for variable selection. In other words, we need to prove show:

$$\|\Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)\boldsymbol{\tau}^0_{S_0}\|_\infty \leq 1.$$

Assume that $S_{\text{init}} \subset S_0$. Thus $(\beta_{\text{init}})_{S_0^c} \equiv 0$. By taking the KKT conditions once again, we get the same equations (3.1) and (3.2) in the Part 1.

Once again, with some elementary algebraic operations on the first equation and plugging it into the second equation, we can take the l_∞-norm of both side to get

$$\|\Sigma_{2,1}(S_0)\Sigma_{1,1}^{-1}(S_0)\boldsymbol{\tau}^0_{S_0}\| = \|(\boldsymbol{\tau}_{\text{init}})_{S_0^c}\|_\infty \leq 1.$$

□

4 Final Remarks

In this article we study some popular regularization and variable selection methods for high-dimensional data. We also study one data simulation and two real datasets to see the performance of different regularization methods. However our study only focus on the concept of *regression*, but *classification*, on the other hand, is also worth exploring. We provide this final remark by giving a brief overview of a classification problem and its challenges in high-dimensional setting.

The main difference between regression and classification is that the response $y \in \mathcal{Y}$ of a classification problem is categorical. A *classifier* $f : \mathcal{X} \to \mathcal{Y}$ is a Borel measurable function where it maps each data point to a label. our task is to find the optimal $f \in \mathcal{C}$. where $\mathcal{C}$ is the class of all classifiers. For simplicity let us consider a binary classification problem where $\mathcal{Y} = \{0, 1\}$ with the dataset $\mathcal{D} := \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$. Misclassification occurs when the predicted label is different from the true label, i.e., $f(\boldsymbol{x}) \neq y$. With the theory from Vapnik and Chervonenkis ([VC71]), one may obtain a probabilistic bound for the misclassification rate as follows: Suppose $\widehat{f}$ is the binary classifier that is consistent on the dataset, denote $R(\widehat{f})$ as its misclassification rate, then for any $\epsilon > 0, n \geq d$, we have

$$\mathbb{P}(R(\widehat{f}) \geq \epsilon) \leq 2 \left(\frac{2en}{d}\right)^d e^{-\epsilon n/2}, \tag{4.1}$$

where d is the VC dimension of $\mathcal{C}$. A detailed proof of (4.1) is in Zhang and Petit ([ZP25]). By the nature of binary classification, we may also summarize the misclassification into type (I) and type (II) error:

$$\text{Type (I) error: } R_1(f) := \mathbb{P}(f(\boldsymbol{x}) = 1 | y = 0), \quad \text{Type (II) error: } R_2(f) := \mathbb{P}(f(\boldsymbol{x}) = 0 | y = 1).$$

The optimal classifier can also be obtained by minimizing a weighted sum of type (I), (II) error:

$$\widehat{f}_{\text{optimal}} := \arg\min_{f \in \mathcal{C}} \mathbb{P}(y = 0) R_1(f) + \mathbb{P}(y = 1) R_2(f).$$

Note that under many studies like cancer diagnosis, it is often important to lower one type of error as much as possible, for example we minimize $R_2(f)$ under the constraint $R_1(f) \leq 0.05$. Then one may apply Neyman-Pearson lemma if the distribution is known. Zhang and Petit ([ZP25]) summarized two algorithms for building a Neyman-Pearson classifier.

In general, regular classification problem when $n > p$ has been well studied. For example, the earliest linear discriminant analysis (LDA) proposed by Fisher in 1936; support vector machine (SVM) by Vapnik in 1996; random forests by Breiman in 2001, and so on. While in the high-dimensional setting where $p > n$, those classifiers mentioned above may face serious issues, and we end the article by addressing those challenges as follows:

- **The failure to identify the sparsity structure**. In many high-dimensional classification problems emerge from scientific studies such as cancer diagnosis, the primary interest is to identify informative features for the class label. With too many noise features, the overall classification performance may be severely deteriorated ([Zou18]).

- **The problem of class imbalance**. In many studies researchers often encounter a huge difference among the number of samples in each class, such as rare disease, spam detection, etc.. With huge class imbalance the label of the new data point will be favored by the label of the majority class and yield inaccuracies.

- **The problem of censored observations.** In life time data analysis, researchers often encounter censored observations, such as right censored data where we only observe $T = \min\{X, C\}$ where X is the true event time while C is the censoring time. Under censoring we only have partial observation, which makes classification difficult.

5 References

[BdG11] Peter Bühlmann and Sara Van de Geer. *Statistics for High-Dimensional Data*. Springer Series in Statistics, 2011.

[HK70] E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12:55–67, 1970.

[JT09] Iain M. Johnstone and D. Michael Titterington. *Statistical Challenges of HIgh-Dimensional Data*. Phil. Trans. R. Soc, 367 edition, 2009.

[Led22] Johannes Lederer. *Fundamentals of High Dimensional Statistics*. Springer Texts in Statistics, 2022.

[MDP+11] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. *Learning Word Vectors for Sentiment Analysis*. Association for Computational Linguistics, 2011.

[PC08] Trevor Park and George Casella. *The Bayesian Lasso*. 2008.

[SD19] Tanin Sirimongkolkasem and Reza Drikvandi. *On Regularisation Methods for Analysis of High Dimensional Data*. Annals of Data Science, 6(4) edition, 2019.

[Tib96] Robert Tibshirani. *Regression Shrinkage and Selection via the Lasso*. J.R Statist. Soc. B, 58 edition, 1996.

[Tib23] Ryan Tibshirani. *High-Dimensional Regression: Lasso*. Advanced Topics in Statistical Learning, 2023.

[VC71] V.N Vapnik and A. Ya. Chervonenkis. *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*. 1971.

[ZH05] Hui Zou and Trevor Hastie. *Regularization and variable selection via the elastic net*. J.R Statist. Soc. B, 67 edition, 2005.

[Zou18] Hui Zou. *Classification with high dimensional features*. 2018.

[ZP25] Jiajun Zhang and Emile Petit. *Vapnik Chervonenkis Dimension, Probabilistic Bounds and Neyman-Pearson Classification*. 2025.