# Final Project

Benjamin Noland

## Introduction

A common problem in applied statistics is estimation of a vector $\beta^* \in \mathbb{R}^p$ of unknown but fixed parameters in the linear model

$$y = X\beta^* + \epsilon, \tag{1}$$

where $y \in \mathbb{R}^n$ is a vector of observed responses, $X \in \mathbb{R}^{n\times p}$ is the design matrix, and $\epsilon \in \mathbb{R}^n$ is a zero-mean random vector representing the uncertainty in the model.

In the classical setting, we assume that the number of parameters $p$ is small relative to the number of observations, specifically $p \leq n$. In this setting, assuming the design matrix $X$ has full row rank, straightforward linear algebra yields an explicit, unique least-squares estimator of $\beta^*$.

However, the situation when there are more parameters than observations, i.e., $p > n$, is not so well understood, and belongs to the active area of research known as *high-dimensional statistics*. One of the strategies commonly employed in high-dimensional statistics is to assume that the data is *truly low-dimensional* in some sense. In the context of our linear model (1), this means assuming that a large number of the entries of the true parameter vector $\beta^*$ are zero. To be precise, define the *support* of $\beta^*$ by

$$S(\beta^*) = \{i \in \{1, \dots, p\} : \beta_i^* \neq 0\},$$

and let $k = |S(\beta^*)|$ denote its cardinality, i.e., the number of non-zero entries of $\beta^*$. We assume that the vector $\beta^*$ is *sparse*, in the sense that $k \ll p$. Under this *sparsity assumption*, the problem reduces to that of computing the support $S(\beta^*)$, allowing us to identify which parameters in the vector $\beta^*$ are truly important. In this way, we have the potential to substantially reduce the dimensionality of the original problem.

A computational tractable method for computing estimates of the parameters $\beta^*$ in the high-dimensional setting is the *LASSO* [2] (Least Absolute Shrinkage And Selection Operator). The LASSO computes an estimate of $\beta^*$ as a solution $\beta \in \mathbb{R}^p$ to the following $l_1$-constrained quadratic program:

$$\begin{array}{ll} \text{minimize} & \|y - X\beta\|_2^2 \\ \text{subject to} & \|\beta\|_1 \leq C_n \end{array}, \tag{2}$$

or equivalently, as the solution to the unconstrained problem

$$\text{minimize } \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda_n\|\beta\|_1, \tag{3}$$

where $\lambda_n \geq 0$ is a *regularization parameter* that is in one-to-one correspondence with $C_n$ via Lagrangian duality [1].

# Project overview

This project will explore the contributions of the paper [1] to the problem of inferring the support $S(\beta^*)$ of $\beta^*$ (i.e., the problem of *support recovery*) in the linear model (1) using the LASSO as a means of estimating $\beta^*$.

## Overview of the paper

The paper [1] provides both necessary and sufficient conditions for the LASSO to recover the *signed support* $\mathbb{S}_\pm(\beta^*) \in \mathbb{R}^p$ of $\beta^*$ with high probability, where $\mathbb{S}_\pm(\beta)$ is defined as follows for any $\beta \in \mathbb{R}^p$:

$$\mathbb{S}_\pm(\beta)_i = \begin{cases} +1 & \text{if } \beta_i > 0 \\ -1 & \text{if } \beta_i < 0 \\ 0 & \text{if } \beta_i = 0 \end{cases} \quad (i = 1, \ldots, p).$$

Specifically, the authors consider the following two questions:

- What relationships between $n$, $p$, and $k$ yield a *unique* LASSO solution $\hat{\beta}$ satisfying $\mathbb{S}_\pm(\hat{\beta}) = \mathbb{S}_\pm(\beta^*)$?
- For what relationships between $n$, $p$, and $k$ does *no solution* of the LASSO yield the correct signed support?

These questions are analyzed for both deterministic designs and random designs in the linear model (1).

In addition to providing theoretical guarantees, the authors describe the results of simulations to investigate the success/failure of the LASSO in recovering the true signed support for random designs under each of the following sparsity regimes:

- *linear sparsity*: $k(p) = \lceil \gamma p \rceil$ for some $\gamma \in (0, 1)$;
- *sublinear sparsity*: $k(p) = \lceil \gamma p / \log(\gamma p) \rceil$ for some $\gamma \in (0, 1)$, and
- *fractional power sparsity*: $k(p) = \lceil \gamma p^\delta \rceil$ for some $\gamma, \delta \in (0, 1)$.

In each case, the authors take $\gamma = 0.40$ and $\delta = 0.75$, and the number of observations $n$ is taken to be proportional to $k \log(p - k)$. The true support of the parameter vector is chosen at random.

For each sparsity regime and for several values of $p$, the authors compute a sequence of values of the *rescaled sample size* (or *control parameter*) $\theta = n/(k \log(p - k))$ and for each such value, compute a sequence of corresponding LASSO solutions $\hat{\beta}$ in order to approximate the probability $\mathrm{P}\{\mathbb{S}_\pm(\hat{\beta}) = \mathbb{S}_\pm(\beta^*)\}$ of recovering the true signed support. This approximated probability is then plotted against the control parameter $\theta$.

The first round of experiments samples the design matrix $X \in \mathbb{R}^{n\times p}$ from a uniform Gaussian ensemble; that is, its rows are sampled independently from the distribution $N_p(0, I_p)$. A second round of experiments samples $X$ from a non-uniform Gaussian ensemble; specifically, one such that the rows are sampled independently from the distribution $N_p(0, \Sigma)$, where $\Sigma$ is a $p \times p$ Toeplitz matrix of the form

$$\Sigma = \begin{pmatrix} 1 & \mu & \mu^2 & \cdots & \mu^{p-2} & \mu^{p-1} \\ \mu & 1 & \mu & \mu^2 & \cdots & \mu^{p-2} \\ \mu^2 & \mu & 1 & \mu & \cdots & \mu^{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu^{p-1} & \cdots & \mu^3 & \mu^2 & \mu & 1 \end{pmatrix},$$

where $\mu = 0.10$. In both cases, the authors note good agreement with the their theoretical predictions.

## This project

In addition to duplicating the simulations from the paper [1], this project extends the simulations by considering the more general case of *elastic net* penalties [3], which extend the $l_1$ penalty in (3) to include an $l_2$ term as well. Specifically, we consider solutions $\beta \in \mathbb{R}^p$ to the problem

$$\text{minimize } \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda_n \left(\frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\right),$$

where $\alpha \in [0, 1]$ is the elastic net *mixing parameter*. We repeat the simulations described in [1] for the case of uniform Gaussian ensembles, but instead use elastic net solutions $\beta$ for each of $\alpha = 0.75$ and $\alpha = 0.50$ to estimate the probability of signed support recovery. We compare the results to those from the original simulations.

# Theoretical results

## Unconstrained form of the problem

As noted in the introduction, the $l_1$-constrained problem (2) is equivalent to the unconstrained problem (3) in the following sense: for every value of $C_n$ in (2) there exists a value

$\lambda_n \geq 0$ in (3) such that (3) is equivalent to (2), and vice versa (in fact, it can be shown that $C_n$ and $\lambda_n$ are in one-to-one correspondence). We now demonstrate this equivalence.

First, we need a lemma. We need to show that the constraint

$$\|\beta\|_1 \leq C_n \tag{4}$$

in (2) is equivalent to a finite collection of linear equality and inequality constraints on $\beta$. Consider the $l_1$-ball $B = \{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq C_n\}$. Let $\{e_1, \ldots, e_p\}$ denote the standard ordered basis for $\mathbb{R}^p$. We claim that $B = \operatorname{conv} S$, where

$$S = \operatorname{conv}\{C_n e_1, \ldots, C_n e_p, -C_n e_1, \ldots, -C_n e_p\}.$$

If we can show that $B = \operatorname{conv} S$, then $B$ is a polyhedron, so that the constraint (4) is equivalent to a finite collection of linear equalities and inequalities.

Note that $S \subseteq B$, and since $B$ is a convex set, and $\operatorname{conv} S$ is the smallest convex set containing $S$, we have $\operatorname{conv} S \subseteq B$. Conversely, let $\beta \in B$. Then there exist $a_1, \ldots, a_p \in \mathbb{R}$ with $\beta = \sum_{i=1}^p a_i e_i$. Assume without loss of generality that $a_1, \ldots, a_m \geq 0$ and $a_{m+1}, \ldots, a_p < 0$. Then we can write

$$\begin{aligned} \beta &= \sum_{i=1}^p a_i e_i \\ &= \sum_{i=1}^m \left(\frac{a_i}{C_n}\right)(C_n e_i) + \sum_{i=m+1}^p \left(-\frac{a_i}{C_n}\right)(-C_n e_i) \\ &= \sum_{i=1}^m \left|\frac{a_i}{C_n}\right|(C_n e_i) + \sum_{i=m+1}^p \left|\frac{a_i}{C_n}\right|(-C_n e_i). \end{aligned}$$

Then the coefficients $|a_i/C_n| \geq 0$ for every $1 \leq i \leq p$, and since $\|\beta\|_1 \leq C_n$, we have

$$\frac{\|\beta\|_1}{C_n} = \sum_{i=1}^p \left|\frac{a_i}{C_n}\right| \leq 1.$$

Let $K = \sum_{i=1}^p |a_i/C_n|$, so that $0 \leq K \leq 1$. Now, $0 \in \operatorname{conv} S$ since we can write $0 = (1/2)(C_n e_1) + (1/2)(-C_n e_1)$. Therefore,

$$\beta = \sum_{i=1}^m \left|\frac{a_i}{C_n}\right|(C_n e_i) + \sum_{i=m+1}^p \left|\frac{a_i}{C_n}\right|(-C_n e_i) + (1-K)\cdot 0 \in \operatorname{conv} S.$$

This shows that $B \subseteq \operatorname{conv} S$, and therefore $B = \operatorname{conv} S$.

Now for the main argument. Let $\hat{\beta} \in \mathbb{R}^p$ be a solution to the constrained problem (2). The Lagrangian of this problem is

$$L(\beta, \lambda) = \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda(\|\beta\|_1 - C_n),$$

where $\lambda \in \mathbb{R}$, and so the Lagrange dual function is given by

$$g(\lambda) = \inf_{\beta \in \mathbb{R}^p} L(\beta, \lambda) = \inf_{\beta \in \mathbb{R}^p} \left[ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda(\|\beta\|_1 - C_n) \right].$$

Note that for any $\lambda \geq 0$, $g(\lambda) > -\infty$, so that the dual problem

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \geq 0 \end{array} \tag{5}$$

is always feasible. Now, since $\hat{\beta}$ is a solution to the primal problem (2), it in particular satisfies $\|\beta\|_1 \leq C_n$ (i.e., $\hat{\beta}$ is feasible for the primal problem). By the lemma above, this $l_1$-constraint is equivalent to a finite number of linear equality and inequality constraints. Thus Slater's condition is satisfied for (2), so that strong duality holds. Since this problem is convex and its dual problem (5) is feasible, this also implies the existence of a dual solution $\lambda_n$. We therefore conclude that

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta \in \mathbb{R}^p} L(\beta, \lambda_n) \\ &= \arg\min_{\beta \in \mathbb{R}^p} \left[ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda_n(\|\beta\|_1 - C_n) \right] \\ &= \arg\min_{\beta \in \mathbb{R}^p} \left[ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1 \right], \end{aligned}$$

i.e., $\hat{\beta}$ is a solution to the unconstrained problem (3).

Conversely, let $\hat{\beta} \in \mathbb{R}^p$ be a solution to the unconstrained problem (3). We claim that $\hat{\beta}$ is a solution to the constrained problem (2) with $C_n = \|\hat{\beta}\|_1$. First, note that $\hat{\beta}_1$ is clearly feasible due to the choice of $C_n$. Suppose it is *not* optimal, i.e., there exists a feasible point $\tilde{\beta} \in \mathbb{R}^p$ with

$$\frac{1}{2n} \|y - X\tilde{\beta}\|_2^2 < \frac{1}{2n} \|y - X\hat{\beta}\|_2^2.$$

Since $\tilde{\beta}$ is feasible, $\|\tilde{\beta}\|_1 \leq C_n = \|\hat{\beta}\|_1$. Thus,

$$\frac{1}{2n} \|y - X\tilde{\beta}\|_2^2 + \lambda_n \|\tilde{\beta}\|_1 < \frac{1}{2n} \|y - X\hat{\beta}\|_2^2 + \lambda_n \|\tilde{\beta}\|_1 \leq \frac{1}{2n} \|y - X\hat{\beta}\|_2^2 + \lambda_n \|\hat{\beta}\|_1,$$

contradicting the assumption that $\hat{\beta}$ was optimal for (3). Hence $\hat{\beta}$ is a solution to (2) with this choice of $C_n$.

## Conditions for signed support recovery

The paper [1] provides necessary and sufficient conditions for the LASSO to recover the signed support of the true parameter $\beta^*$ in the model (1) for both deterministic and random designs. Here, we restrict our attention to the case of random design, and state the pertinent results from [1].

The design matrix $X \in \mathbb{R}^{n \times p}$ is in this case drawn from a random Gaussian ensemble; that is, its rows are sampled independently from the distribution $N_p(0, \Sigma)$ for some choice of covariance matrix $\Sigma$.

The results given below rely upon (subsets of) the following conditions on $\Sigma$ being satisfied:

$$\begin{aligned}
\|\Sigma_{S^cS}\Sigma_{SS}^{-1}\|_\infty &\leq 1 - \gamma \quad \text{for some } \gamma \in (0, 1] && \text{(6a)}\\
\Lambda_{\min}(\Sigma_{SS}) &\geq C_{\min} > 0 && \text{(6b)}\\
\Lambda_{\max}(\Sigma_{SS}) &\leq C_{\max} < \infty, && \text{(6c)}
\end{aligned}$$

where $\gamma$ is known as an *incoherence parameter*, and

$$\begin{aligned}
\Sigma_{SS} &= \mathrm{E}\left(\frac{1}{n} X_S^T X_S\right)\\
\Sigma_{S^cS} &= \mathrm{E}\left(\frac{1}{n} X_{S^c}^T X_S\right),
\end{aligned}$$

and similarly for $\Sigma_{SS^c}$ and $\Sigma_{S^cS^c}$. The expressions $\Lambda_{\min}(\Sigma_{SS})$ and $\Lambda_{\max}(\Sigma_{SS})$ denote the minimum and maximum eigenvalues of $\Sigma_{SS}$, respectively.

For a positive semidefinite matrix $A$, define

$$\begin{aligned}
\rho_l(A) &= \frac{1}{2} \min_{i \neq j}(A_{ii} + A_{jj} - 2A_{ij})\\
\rho_u(A) &= \max_i A_{ii}.
\end{aligned}$$

It is not difficult to show that

$$0 \leq \rho_l(A) \leq \rho_u(A).$$

Finally, using the conditional covariance matrix

$$\Sigma_{S^c|S} = \Sigma_{S^cS^c} - \Sigma_{S^cS}\Sigma_{SS}^{-1}\Sigma_{SS^c} \succeq 0$$

of $(X_{S^c}|X_S)$, we define

$$\begin{aligned}
\theta_l(\Sigma) &= \frac{\rho_l(\Sigma_{S^c|S})}{C_{\max}(2 - \gamma(\Sigma))}\\
\theta_u(\Sigma) &= \frac{\rho_u(\Sigma_{S^c|S})}{C_{\min}\gamma(\Sigma)^2},
\end{aligned}$$

where $\gamma(\Sigma) \in (0, 1]$ is the incoherence parameter in (6a). Moreover, we have the inequalities

$$0 \leq \theta_l(\Sigma) \leq \theta_u(\Sigma) < \infty.$$

With these definitions and conditions in place, we are now able to state the pertinent results from [1]. The first result, Theorem 3 from [1], provides sufficient conditions for the LASSO to recover the signed support of the true parameter $\beta^*$ in the linear model (1).

**Theorem 3** (Sufficiency)**.** *Consider the linear model* (1) *with random Gaussian design* $X \in \mathbb{R}^{n\times p}$ *and error term* $\epsilon \sim N_n(0, \sigma^2 I_n)$*. Assume that the covariance matrix* $\Sigma$ *satisfies conditions* (6a) *and* (6b)*. Consider the sequence* $(\lambda_n)$ *of regularization parameters given by*

$$\lambda_n = \lambda_n(\phi_p) = \sqrt{\frac{\phi_p \rho_u(\Sigma_{S^c|S})}{\gamma^2} \frac{2\sigma^2 \log p}{n}}, \tag{7}$$

*where* $\phi_p \geq 2$*. Suppose there exists* $\delta > 0$ *such that the sequences* $(n, p, k)$ *and* $(\lambda_n)$ *satisfy*

$$\frac{n}{2k\log(p-k)} > (1+\delta)\theta_u(\Sigma)\left(1 + \frac{\sigma^2 C_{\min}}{\lambda_n^2 k}\right).$$

*Then the following properties hold with probability* $> 1 - c_1 \exp(-c_2 \min\{k, \log(p-k)\})$*:*

*(i) The LASSO has a unique solution* $\hat{\beta} \in \mathbb{R}^p$ *with* $S(\hat{\beta}) \subseteq S(\beta^*)$ *(i.e., its support is contained in the true support).*

*(ii) Define*

$$g(\lambda_n) = c_3 \lambda_n \|\Sigma_{SS}^{-1/2}\|_\infty^2 + 20\sqrt{\frac{\sigma^2 \log k}{C_{\min} n}}.$$

*Then if* $\beta_{\min} = \min_{i\in S} |\beta_i^*|$ *satisfies* $\beta_{\min} > g(\lambda_n)$*, we have*

$$\mathbb{S}_\pm(\hat{\beta}) = \mathbb{S}_\pm(\beta^*)$$

*and*

$$\|\hat{\beta}_S - \beta_s^*\|_\infty \leq g(\lambda_n).$$

The second result, Theorem 4 from [1] provides *necessary* conditions for signed support recovery to be successful. Specifically, it provides sufficient conditions for signed support recovery to fail.

**Theorem 4** (Necessity)**.** *Consider the linear model* (1) *with random Gaussian design* $X \in \mathbb{R}^{n\times p}$ *and error term* $\epsilon \sim N_n(0, \sigma^2 I_n)$*. Assume that the covariance matrix* $\Sigma$ *satisfies conditions* (6a)*,* (6b)*, and* (6c)*. Consider the sequence* $(\lambda_n)$ *of regularization parameters* (7)*. Suppose there exists* $\delta > 0$ *such that the sequences* $(n, p, k)$ *and* $(\lambda_n)$ *satisfy*

$$\frac{n}{2k\log(p-k)} < (1-\delta)\theta_l(\Sigma)\left(1 + \frac{\sigma^2 C_{\max}}{\lambda_n^2 k}\right).$$

*Then with probability converging to one, no solution of the LASSO has the correct signed support.*

# Simulations

## Set up

## Simulations from the paper

## Custom simulations

# References


[1] Wainwright, M. (2006). *Sharp thresholds for high-dimensional and noisy sparsity recovery using $l_1$-constrained quadratic programming (Lasso)*. Technical Report 709, Dept. Statistics, Univ. California, Berkeley

[2] Tibshirani, R. (1996). *Regression shrinkage and selection via the Lasso*. J. Roy. Statist. Soc. Ser. B **58** 267–288

[3] Zou, H. and Hastie, T. (2005) *Regularization and variable selection via the elastic net* J. Roy. Statist. Soc. Ser. B **67** 301–320