Session 1: Lasso

Fall 2025

Luis Sierra Muntané
Reading Group in Applied Statistics

There once was a model, quite brash,
That gave bloated betas a slash.
With an ℓ_1 embrace,
It trimmed down the space,
And only left those with some flash.

Contents

1 Motivation

The lasso, short for least absolute shrinkage and selection operator as named in (Tibshirani 1996), minimizes the squared residuals while limiting the l_1-norm of the parameters. Because of the nature of this constraint it tends to produce coefficients that are exactly 0, hence giving interpretable models and selecting the most relevant predictors. It also enjoys favourable properties with respect to subset selction and ridge regression, two other alternatives to the limitations of the OLS estimator to be subsequently explored.

We begin by considering the problem of linear regression over p variables, which involves a DGP given by

$$y = X\beta^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n),$$

where it is often only necessary for the components of the noise ε to be sub-Gaussian with proxy variance σ^2. Setting this linear regression as the optimization problem of OLS gives us the cost function to be optimized

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

whose solution is readily calculated as the projection onto the column space of X given by $\hat{\beta} = (X^\top X)^{-1} X^\top y$. From this we infer the distribution $\hat{\beta} \sim \mathcal{N}_p\left(\beta^*, \sigma^2 (X^\top X)^{-1}\right)$. The covariance matrix can be ill-conditioned when $\det(X^\top X)$ is small, leading to the two following important considerations:

- Size of the variance of estimators in OLS make predictions unstable.
- Multi-collinearity (X is rank deficient) and so not invertible.

Both of these aspects are properties related to the span of the columns of X. If they are close to linearly dependent (or in particular close to zero) then these two phenomena will manifest themselves. In particular, linear regression becomes more numerically unstable as the number of covariates p increases and becomes outright infeasible for $p > n$.

Assuming an i.i.d. sample (x_i, y_i) where $X^\top X$ is almost surely invertible, the out-of-sample prediction risk of the OLS estimator $\hat{\beta}$ is given by

$$\mathbb{E}\left[\left(x_0^\top \hat{\beta} - x_0^\top \beta^*\right)^2\right] = \sigma^2 \operatorname{tr}\left(\mathbb{E}[x_0 x_0^\top]\mathbb{E}\left[(X^\top X)^{-1}\right]\right)$$
$$\approx \sigma^2 \frac{p}{n-p} \in O(n^{-1}),$$

where we use the fact that $X^\top X$ follows a Wishart distribution and so we have $n - p - 1$ in the denominator, giving us the standard parametric rate.

For $p > n$, the OLS problem admits any solution of the form

$$\hat{\beta} = (X^\top X)^\dagger X^\top y + \eta, \quad \text{for } \eta \in \ker(X),$$

where $A^\dagger$ denotes the pseudo-inverse of A. This implies that we can't even consistently estimate the *sign* of a coefficient β_j (assuming e_j is not perpedicular to $\ker(X)$).

2 Solving the problems through an l_1 penalty

To solve the previous problems, a possible approach is that of applying shrinkage/regularization in order to achieve the following:

- Constraining β to be inside some desirable set C.
- Penalizing the cost with some nonnegative function $h(\beta)$.

Note that under the right conditions, these two are equivalent by convex duality, so we will mainly focus on the second one for ease of exposition and notation. In fact, two predecessors of the lasso that aimed to do just that were ridge regression and subset selection. Ridges applies a shrinkage that helps in reducing variance for ill-conditioned $X^\top X$ while subset selection restricts the number of variables to be included in the regression with the same objective, while also imposing sparsity.

However, subset selection suffers from being non-convex and so a numerically inferior approach. It does provide interpretable models but it can still be unstable because of it being a discrete process—regressors are either retained or dropped from the model in a discontinuous manner, so small changes in the data can result in very different models being selected, reducing its prediction accuracy. Moreover, the models generated are not *nested*, meaning that the variables selected in a model of size k are not a subset of the variables selected in a model of size $k + 1$. Ridge regression is a continuous process that shrinks coefficients and hence is more stable although it does not promote sparsity. In fact, the form of ridge regression shrinkage depends on the correlation structure of the predictors, not just with the response. Moreover, it favours setting variables equal to each other, as this minimizes the squared norm.

Another predecessor is the nonnegative garotte. A drawback of the garotte is that its solution depends on both the sign and the magnitude of the OLS estimates. In settings with high multicollinearity or others where the OLS estimates behave poorly, the garotte will suffer too **(Tibshirani 1996)**.

Let us define the lasso as the optimization problem given by the cost function

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

which combines best fet with variable selection. Regularising with any other power is also possible, and a popular alternative is the l_2-norm, often referred to in statistics as *ridge* regression, although this option has dense solutions instead of sparse ones.

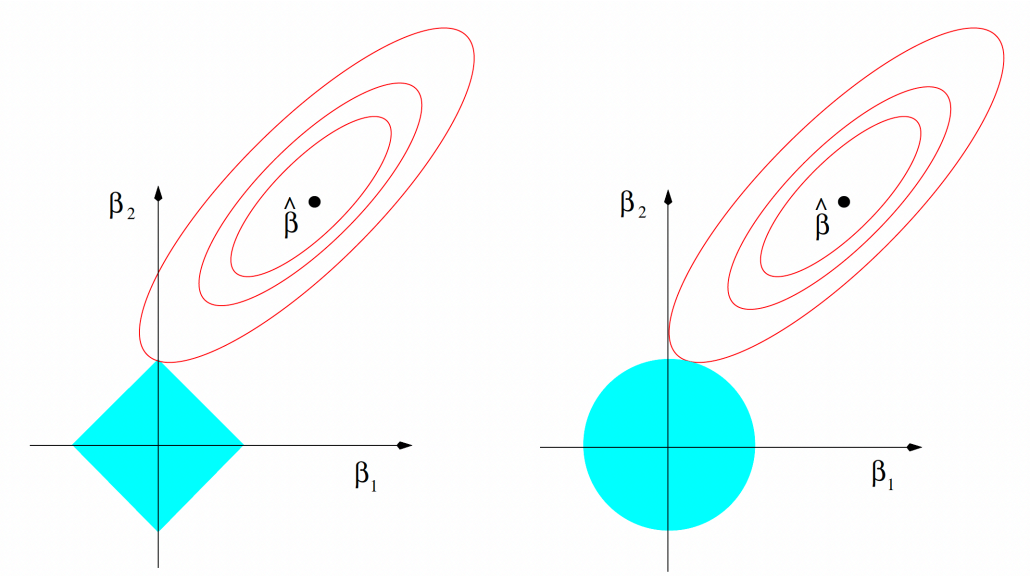

Figure 1: Level sets for lasso and ridge regression, from (Hastie et al. 2009).

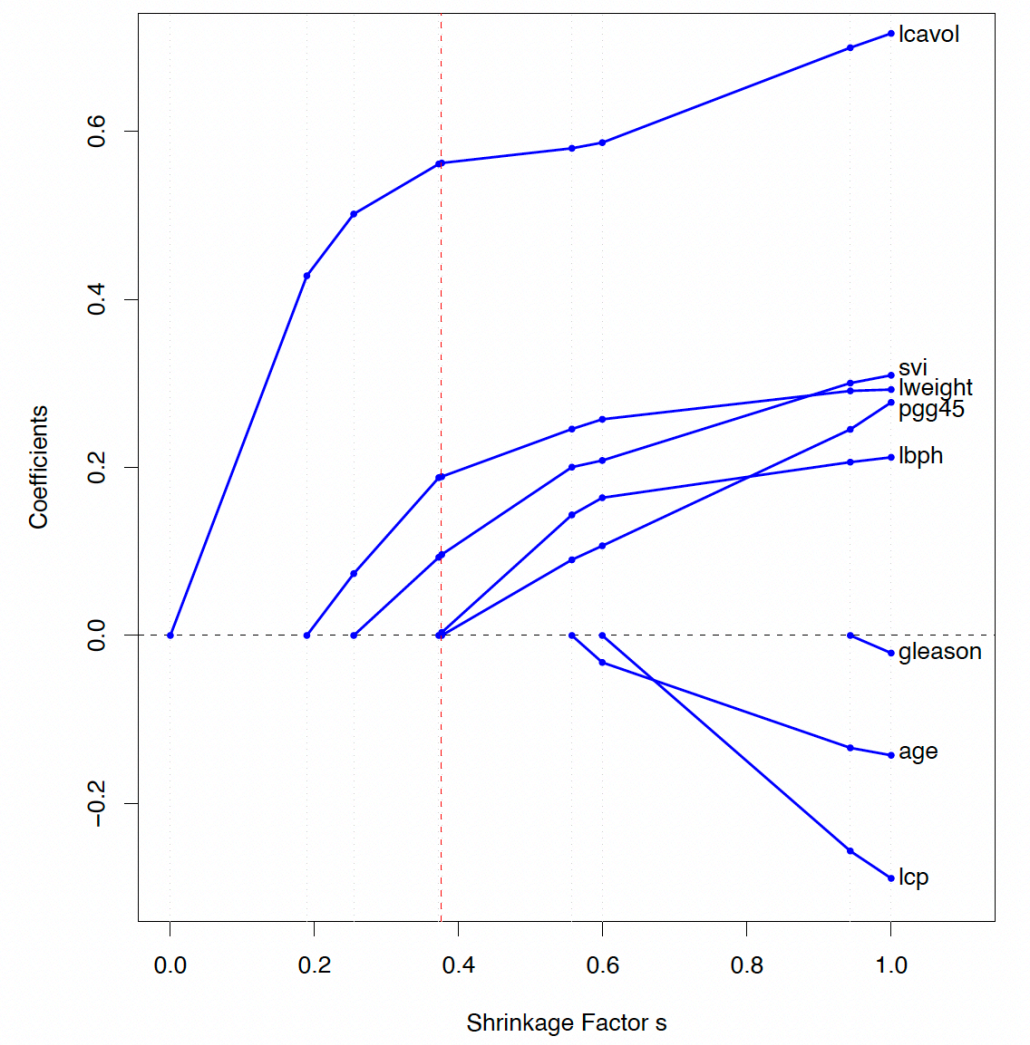

Figure 2: Graph of path of parameters can be found in (Zou et al. 2007) and (Hastie et al. 2009). We shall see later in the section on implementation how to get such a path.

Observation 2.1: If the data matrix X is orthogonal then

- $\hat{\beta}_{\text{ridge}} = (X^\top y)/(1+\lambda)$,
- $\hat{\beta}_{\text{lasso}} = S_\lambda(X^\top y)$, quad $S_\lambda(a) = \text{sgn}(a)(|a| - \lambda)_+$.

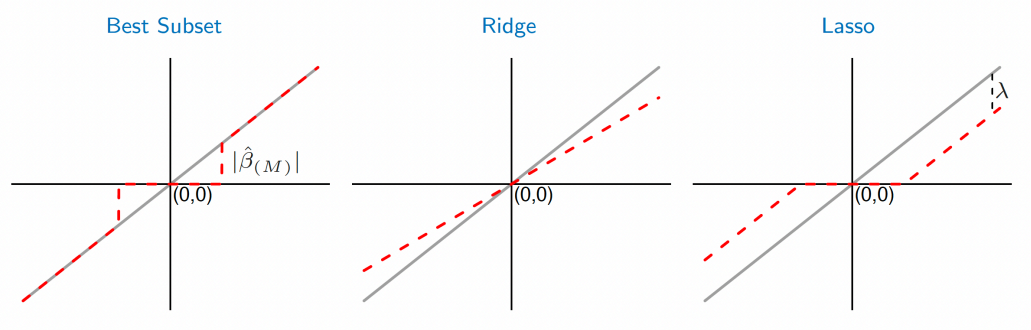

Figure 3: Level sets for lasso and ridge regression, from (Hastie et al. 2009).

Remark 2.2: For those interested in Bayesian statistics, the lasso also arises from considering a Laplace prior on β: $f(\beta_j) = \frac{1}{2\tau} e^{-\frac{|\beta_j|}{\tau}}$. Note that the laplace distribution has a higher density at its mean and heavier tails than the Gaussian, which is the corresponding prior for ridge regression.

Remark 2.3: The lasso is not strictly convex for $p > n$, meaning that unlike for ridge regression, there may not be a unique solution to Equation 1. That being said, the lasso fit $X\hat{\beta}$ is always unique (as it is for OLS), since the least squares loss is strictly convex.

Remark 2.4: The lasso is closely related to *Basis Pursuit Denoising*, which refers to the problem

$$\min_{\beta} \quad \|\beta\|_1, \quad \text{s.t. } X\beta = y,$$

where $p > n$. These two problems are equivalent by lagrangian duality but this way of expressing it is more common in signal processing. See (Chen et al. 1998) for more information.

3 Structure of solutions

3.1 KKT conditions

The problem in Equation 1 is not differentiable due to the l_1 norm term having a point of non-differentiability, but we can readily calculate its subgradient as

$$X^\top\left(y - X\hat{\beta}\right) = \lambda s, \quad s \in \partial \|\hat{\beta}_1\| \tag{2}$$

where

$$s_i \in \begin{cases} \{1\}, & \hat{\beta}_i > 0 \\ [-1,1], & \hat{\beta}_i = 0 \\ \{-1\}, & \hat{\beta}_i < 0. \end{cases} \tag{3}$$

From Equation 2 and Equation 3 we can see that the optimal subgradient s is always unique for any $\lambda > 0$. This is because it is determined by the unique fitted value $X\hat{\beta}$, so we don't have the previous problem of differently signed coefficients being equally valid solutions.

Definition 3.1.1: The *equicorrelation set* is the set of covariates

$$\mathcal{E}_\lambda := \left\{j \in [p] : \left|X_j^\top\left(y - X\hat{\beta}\right)\right| = \lambda\right\}.$$

Intuitively this corresponds to the variables that achieve the maximum inner product with the lasso residual for a specified value of λ. The definition is inspired by considering the KKT conditions from Equation 2 that are active in an optimization sense, more on this in Remark 3.1.3.

This tells us that the variables can be cleanly divided into

- Active variables: $\hat{\beta}_i \neq 0$ then the correlation between the predictor X_i and the residual vector $r = y - X\hat{\beta}$ is given by

$$X_i^\top r = \lambda \operatorname{sgn}\left(\hat{\beta}_i\right).$$

- Inactive variables: $\hat{\beta}_i = 0$ so

$$|X_i^\top r| \leq \lambda.$$

Note that the correlation for inactive variables is *not* strictly smaller than λ. Also observe that this set is unique, because $X\hat{\beta}, s$ are unique and contains the active set $\mathcal{A} = \operatorname{supp}\left(\hat{\beta}\right)$.

Observation 3.1.2: The equicorrelation set contains the active set of covariates $\mathcal{A}$ since if $j \notin \mathcal{E}$ then $\left|s_j\right| < 1 \implies \beta_j = 0$. As such, we can write the lasso fit as $X\hat{\beta} = X_{\mathcal{E}}\hat{\beta}_{\mathcal{E}}$.

Remark 3.1.3: The terminology "active" and "inactive" may cause confusion as in optimization we say that an inequality constraint $g(x) \leq 0$ is *active* if $g(x) = 0$, to express the fact that it is *actively* restricting the current solution. Instead, literature related to the lasso **(Hastie et al. 2009, Zou et al. 2007)** often refer to active set as the entries of $\hat{\beta}_\lambda$ that are different from zero. Critically, the difference is that the equicorrelation set and the support of $\hat{\beta}_\lambda$ are not necessarily the same.

Proposition 3.1.4: If $\operatorname{rank}(X_{\mathcal{E}}) = |\mathcal{E}|$ (the equicorrelated predictors are linearly independent) then the lasso solution is unique.

Proof: From the previous observation and the KKT condition in **Equation 2** we have that

$$X_{\mathcal{E}}^\top \left(y - X_{\mathcal{E}}\hat{\beta}_{\mathcal{E}}\right) = \lambda s_{\mathcal{E}},$$

where solving for $\hat{\beta}_{\mathcal{E}}$ gives

$$\begin{aligned} & X_{\mathcal{E}}^\top y - X_{\mathcal{E}}^\top X_{\mathcal{E}}\hat{\beta}_{\mathcal{E}} = \lambda s_{\mathcal{E}} \\ \iff\ & \hat{\beta}_{\mathcal{E}} = \left(X_{\mathcal{E}}^\top X_{\mathcal{E}}\right)^+ \left(X_{\mathcal{E}}^\top y - \lambda s_{\mathcal{E}}\right) + \eta, \quad \eta \in \ker(X_{\mathcal{E}}), \end{aligned}$$

and when $\operatorname{rank}(X_{\mathcal{E}}) = |E|$ the matrix $\left(X_{\mathcal{E}}^\top X_{\mathcal{E}}\right)$ is invertible thus giving us

$$\hat{\beta}_{\mathcal{E}} = \left(X_{\mathcal{E}}^\top X_{\mathcal{E}}\right)^{-1} \left(X_{\mathcal{E}}^\top y - \lambda s_{\mathcal{E}}\right), \quad \hat{\beta}_{-\mathcal{E}} = 0. \tag{4}$$

□

Observation 3.1.5: The expression in **Equation 4** also holds for the active set $\mathcal{A}$.

Corollary 3.1.6: Whenever the columns of X are in general position, the lasso solution in unique.

3.2 Saturation and elasticnet

Another interesting property of the lasso occurs in the high-dimensional regime where $p > n$. In this case, the lasso selects at most n covariates. This can be seen as a consequence of a result from convex geometry.

Proposition 3.2.1: For $p > n$, the lasso selects at most n variables in the active set.

Lemma 3.2.1 (Conic Carathéodory): Consider the set of points $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$. Every point x in the conic hull $\operatorname{conc}(X)$ can be expressed as the conic combination of d or less points from X.

Proof: Let x be a point in the conic hull of X. Consider its expression as a minimal conic combination

$$x = \sum_{i=1}^m \lambda_i x_i, \quad \lambda_i \geq 0,$$

for appropriately sorted indices. If $m \leq d$ then we are done. If $m > d$, then the points used are linearly dependent and we can find coefficients $\{\mu_i\}$ not all zero such that $\sum_{i=1}^{m} \mu_i x_i = 0$. Define $t := \min_{\mu_i<0} \frac{\lambda_i}{-\mu_i}$ and note that

$$\sum_{i=1}^{n} (\lambda_i + t\mu_i) x_i = x,$$

while $\lambda_i + t\mu_i \geq 0$ and one such coefficient is equal to zero, contradicting the minimality of m. □

Proposition 3.2.2: For any $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$ and $\lambda > 0$, there exists a solution to Equation 1 whose active set $\mathcal{A}$ has at most $\min(n, p)$ elements.

Proof: Let $\hat{\beta}$ be an optimal solution to the lasso and $\hat{y} = X\hat{\beta}$. Let $\hat{\beta}_i = s_i\gamma_i$ where s is defined as in Equation 3 with $s_i \in \{-1, 0, 1\}$, $\gamma_i = |\hat{\beta}_i|$. Then, we can write $\hat{y}$ in terms of $X_i \in \mathbb{R}^n$, the columns of X, as

$$\hat{y} = \sum_{i=1}^{p} \gamma_i s_i X_i = \sum_{i=1}^{p} \gamma_i v_i, \quad v_i \in \{\pm X_i\},$$

so $\hat{y}$ is in the conic hull of $\{\pm X_1, ..., \pm X_p\}$ and $\|\hat{\beta}\|_1 = \sum_{i=1}^{p} \gamma_i$. By Lemma 3.2.1 with $\hat{y} \in \text{conc}(X)$ we have that we can write

$$\hat{y} = \sum_{i=1}^{n} \tilde{\gamma}_i \tilde{v}_i,$$

where $\sum \tilde{\gamma}_i \leq \sum \gamma_i = \|\hat{\beta}\|_1$, so in fact $\text{supp}\left(\hat{\beta}\right) \leq n$. □

Note that this is not necessarily a good thing since it limits the fitting power of the lasso to only consider as many variables as available samples. This fact motivates the use of the elasticnet, a procedure combining penalty terms for both the l_1 and l_2 norms (Zou & Hastie 2005).

4 Choosing the right λ

The risk of the lasso estimator (and any penalization estimators for that fact) depends on the choice of λ. How to select the best λ? One approach is to use boostrap methods—cross-validation—to obtain estimates for the risk and selecting the λ which minimizes them in a meaningfull way (e.g. the 1 s.d. rule), with all of the advantages and disadvantages that this entails (Efron 2004). Another approach is to use use the fact that the lasso is almost differentiable and so admits a risk estimator in the form of SURE. This involves the expression

$$\text{SURE}_\lambda\left(\hat{\beta}_\lambda\right) = -p\sigma^2 + \mathbb{E}\|y - X\hat{\beta}_\lambda\|_2^2 + 2\sigma^2 \,\text{df}\left(\hat{\beta}_\lambda\right),$$

which begs the question of how to calculate the third term—the effective degrees of freedom—if that's possible. It turns out that indeed the degrees of freedom of the lasso can be expressed in closed form, and in fact a lot more can be said about these degrees of freedom.

For a given response vector y, there is a *finite* sequence of λ's,

$$\lambda_0 > \lambda_1 > \lambda_2 > ... > \lambda_K = 0,$$

such that:

- For $\lambda > \lambda_0$, $\hat{\beta}_\lambda = 0$.
- In the interior of the interval $(\lambda_{m+1}, \lambda_m)$, the active set $\mathcal{A}(\lambda)$ and the signs of the coefficients are constant with respect to λ.

The active set changes at each *knot* or *transition point* λ_m, where some variables get added into the active set as we decrease λ_m, while as we increase towards λ_{m-1}, there are possibly some predictors in $\mathcal{A}_m$ whose coefficients will reach zero. This gives us a solution path that is piece-wise linear on account of Equation 4 being linear in λ for a fixed active set.

Theorem 4.1: (Zou et al. 2007) For every λ, the lasso fit is a uniformly Lipschitz function on y. The degrees of freedom of $\hat{\mu}_\lambda(y)$ equal the expectation of the effective set $\mathcal{A}_\lambda$, that is

$$\mathrm{df}(\lambda) = \mathbb{E}|\mathcal{A}_\lambda|,$$

which holds as long as X is full rank.

A simple proof sketch proceeds as follows:

1. The theorem is trivially trur for $\lambda = 0$
2. Consider Equation 4 with the active set version at an interval $(\lambda_{m+1}, \lambda_m)$. Show that the equation is constant for a sufficiently small ε-ball inside a single interval. This gives us that the lasso fit is Lipschitz.
3. By continuity of the lasso fit in y we then extend this to all intervals and get that $X_m^\top \hat{\beta}_{\lambda_m}$ is uniformly Lipschitz, and so differentiable almost everywhere.
4. Use the trace formula from linear regression $\mathrm{tr}(\mathrm{H}_m)$ where H_m is the hat matrix in the corresponding interval.
5. Apply Stein's lemma and the divergence formula.

Corollary 4.2: $\hat{\mathrm{df}}(\lambda) = |\mathcal{A}_\lambda|$ is an unbiased estimate for $\mathrm{df}(\lambda)$, and therefore provides an unbiased estimator of the risk.

The estimator of the degrees of freedom can also be shown to be consistent using the set-up in (Knight & Fu 2000), and in fact we get the following theorem.

Theorem 4.3: (Zou et al. 2007) If $\lambda_n^* \to \lambda^* > 0$, where λ^* is a nontransition point such that $\lambda^* \neq \lambda_{*m}$ for all m, then $\hat{\mathrm{df}}(\lambda_n^*) - \mathrm{df}(\lambda_n^*) \to 0$ in probability.

In this way, to find the optimal λ, we can seek to minimize the expected risk as a function of λ. This can be combined with a parsimony term as in AIC or BIC. Since the LARS (see Section 6.2) efficiently solves the lasso solution for all λ, and in fact (Zou et al. 2007) shows that the optimal λ is at one of the transition points, it is only necessary to solve

$$m^* = \arg\min_m \frac{\|y - X_m^\top \hat{\beta}_m\|_2^2}{n\sigma^2} + \frac{w_n}{n}\hat{\mathrm{df}}(\lambda_m),$$

where $w_n = 2$ for AIC, and $w_n = \log(n)$ for BIC.

In practice, it is still common to resort to cross-validation and the subsequent rules around it. This is because the computational expense of sampling and fitting can be lower than calculating SURE when $p \gg n$, when estimating the noise variance is too hard or when prediction is not as relevant as variable selection, given that SURE only considers prediction risk. That is, unless we use LARS, to appear in Section 6.2.

5 Rates of convergence, fast and slow

(From Ryan Tibshirani's notes: Advanced Topics in Statistical Learning, Spring 2023)

In this section we derive the convergence rate for the excess risk of the lasso. We will first explore what is often referred to as the "slow" rate and then make additional assumptions on our data matrix that allow for faster rates. For the slow rate, we shall first require that the error term is sub-Gaussian and that $\|X_j\|_2 \leq \sqrt{n}$, which can be achieved simply by rescaling the data appropriately. Note that for an optimal fit $\hat{\beta}$ we have that

$$\begin{aligned}
&\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\,\|\hat{\beta}\|_1 \leq \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \\
\iff\ &\frac{1}{2}\|y - X\hat{\beta}\|_2^2 - \frac{1}{2}\|y - X\beta\|_2^2 \leq \lambda\Big(\|\beta\|_1 - \|\hat{\beta}\|_1\Big) \\
\iff\ &\frac{1}{2}\|X\hat{\beta} - X\beta\|_2^2 \leq \langle y - X\beta, X\hat{\beta} - X\beta\rangle + \lambda\Big(\|\beta\|_1 - \|\hat{\beta}\|_1\Big).
\end{aligned}$$

Taking $\beta = \beta^*$ we obtain

$$\frac{1}{2}\|X\hat{\beta} - X\beta^*\|_2^2 \leq \langle \varepsilon, X\hat{\beta} - X\beta^*\rangle + \Big(\|\beta\|_1 - \|\hat{\beta}\|_1\Big),$$

where applying Hölder's and the triangle inequality we get

$$\begin{aligned}\frac{1}{2}\|X\hat{\beta} - X\beta^*\|_2^2 &\leq \|X^\top \varepsilon\|_\infty \left(\|\hat{\beta}\|_1 + \|\beta^*\|_1\right) + \lambda\left(\|\beta^*\|_1 - \|\hat{\beta}\|_1\right) \\ &\leq 2\lambda\|\beta^*\|_1, \quad \text{for} \quad \lambda > \|X^\top\varepsilon\|_\infty.\end{aligned} \tag{5}$$

Up to here everything holds deterministically, but now using the fact that $X^\top\varepsilon$ has sub-Gaussian entries with variance proxy $\max\|X_j\|_2^2\ \sigma^2 \leq n\sigma^2$ we can see that

$$\begin{aligned}&\mathbb{P}\left(\|X^\top\varepsilon\|_\infty \geq \sigma\sqrt{2n(u+\log 2p)}\right) \leq e^{-u} \\ \implies &\frac{1}{n}\|X\hat{\beta} - X\beta^*\|_2^2 \leq 4\sigma\ \|\beta^*\|_1\ \sqrt{\frac{2(u+\log 2p)}{n}}, \quad \text{w.h.p. } (\geq 1 - e^{-u}),\end{aligned} \tag{6}$$

where we used $\lambda = \sigma\sqrt{2n(u+\log 2p)}$ in Equation 5 to obtain the second inequality. From this we may conclude that the in-sample prediction error scales as $\asymp \|\beta^*\|_1\sqrt{\frac{\log p}{n}} \in O\left(n^{-\frac{1}{2}}\right)$.

Remark 5.1: The relevance of this rate is that it is inferior to that of both OLS and subset selection, which also accounts for sparsity in the data. In fact for a level of sparsity $s_0 = \|\beta^*\|_0$ then:

- Best subset selection leads to an-sample prediction risk on the order of $s_0\frac{\log p}{n}$. Some authors like to say that the factor of $\log p$ is the "price to pay" for searching over which of the p variables are relevant for prediction.
- From the result in Equation 6, we see that the lasso leads to an in-sample prediction risk on the order of $\|\beta^*\|_1\ p\frac{\log p}{n}$. If each nonzero entry of β^* is of constant order, then this will be $s_0 p\frac{\log p}{n}$ which is a still "full square root factor slower" than the subset selection rate.

Since the lasso is the convex relaxation of subset selection we may consider whether there is a way to attain the same parametric rate, meaning that we could aim to have our cake and eat it too. This is in fact possible by imposing some extra assumptions on our data.

Again let $s_0 = |S| = |\mathcal{A}^*|$, the number of variables in the true active set.

Theorem 5.2: Either of the following two conditions on X lead to a parametric "fast" rate.

1. Compatibility:

   $$\frac{1}{n}\ \|Xv\|_2^2 \geq \frac{\varphi_0}{s_0}\|v_S\|_1^2, \quad \text{for all } v \in \mathbb{R}^d \text{ for which } \|v_{-S}\|_{-S} \leq 3\|v_S\|_1,$$

   which roughly means the true active predictors can't be too correlated.

2. Spectral:

   $$\frac{1}{n}\ \|Xv\|_2^2 \geq \varphi_0^2\ \|v\|_2^2, \quad \text{for all } v \in \mathbb{R}^d \text{ such that } \|v_{-J}\|_1 \leq 3\|v_J\|_1,$$

   where $J \subset [p]$ is any subset such that $|J| \leq s_0$.

The spectral condition implies the compatibility condition, and it roughly says that φ_0^2 is a lower bound on the eigenvalues of $\frac{1}{n}X^\top X$ for vectors that are s_0-sparse.

6 Computational considerations

In this section, we detail a couple of ways in which the lasso solution is calculated. Least squares fitting is usually done via either Cholesky decomposition of $X^\top X$ or the QR decomposition. The number of operations of these approaches are $p^3 + n\frac{p^2}{2}$ for Cholesky and np^2 for QR. The computational complexity of LARS is the same as that of least squares, whereas coordinate descent is an interative method. That being said, when the lasso problem is strongly convex ($X^\top X$ is full rank) then coordinate descent enjoys a linear rate of convergence.

6.1 Coordinate Descent

The main software package used to solve the lasso is `glmnet` developed by the lasso paper authors. In it, the main procedure to solve Equation 1 is that of using coordinate descent. It works by cyclically solving the lasso problem in each of the variables individually. By assuming a coefficient vector β and separating the k-th entry from the rest; let $\tilde{\beta}_j$ be the current estimate for β_j and re-write Equation 1 to isolate β_k as

$$\min_{\beta_i} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j\neq k} x_{ij}\tilde{\beta}_j - x_{ik}\beta_k\right)^2 + \lambda\sum_{j\neq k}|\tilde{\beta}_j| + \lambda|\beta_k|,$$

where we are also assuming the absence of an intercept. This "univariate lasso" has as its explicit solution

$$\tilde{\beta}_k \leftarrow S_\lambda\left(\sum_{i=1}^{n} x_{ik}\left(y_i - \sum_{j\neq k} x_{ij}\tilde{\beta}_j\right)\right), \tag{7}$$

where S_λ is the soft-thresholding operator $\operatorname{sgn}(|\cdot| - \lambda)_+$. We can write this more succinctly by noting that the univariate case is

$$\hat{\beta} \longleftarrow \begin{cases} \frac{1}{n}\langle X, y\rangle - \lambda\ , & \text{if } \frac{1}{n}\langle X,\ y\rangle > \lambda \\ 0, & \text{if } \frac{1}{n}\ |\langle X,\ y\rangle| \leq \lambda \\ \frac{1}{n}\langle X, y\rangle + \lambda\ , & \text{if } \frac{1}{n}\langle X,\ y\rangle < \lambda. \end{cases}$$

Thus, the updates for the multivariate case can be expressed as

$$\hat{\beta}_k \longleftarrow S_\lambda\left(\hat{\beta}_k + \frac{1}{n}\langle X_k, r\rangle\right),$$

where r is the full residual as in Equation 7.

Algorithm 1: Coordinate Descent

```
procedure COORDINATE-DESCENT(X, y, λ)
    ▷ Inputs: objective f : ℝ^p → ℝ, start x_0, iterations, tolerance
    x ← x_0
    k ← 0
    while k < max-iter do
        x^old ← x
        for j ∈ [p] do
            ▷ Minimize f along coordinate j with others fixed
            x_j ← arg min_{t∈ℝ} f(x_1, ..., x_{j−1}, t, x_{j+1}, ..., x_p)
        end
        if |x − x^old|_∞ ≤ tol then
            return x
        k ← k + 1
    end
    return x
end
```

However, this is not the only option, and (Efron et al. 2004) developed a simple procedure to find the whole solution path, inspired as a "democratic" version of forward stepwise regression.

6.2 LARS

The Least Angle Regression algorithm was developed by (Efron et al. 2004) as a flexible and efficient way of calulateing the lasso solution over the entire path of λ. It works by increasing the entries of the parameter β for which the covariates are most correlated with the residuals un steps. It adds variables one by one sequentially

Observation 6.2.1: Setting $\lambda = 0$ corresponds to the OLS solution, but from the subgradient condition in Equation 2 evaluated at $\beta = 0$ we have that

$$X^\top y \in \lambda\partial\|\beta\|_1 \quad \Leftrightarrow \quad |(X^\top y)_i| \leq \lambda, \forall i,$$

so if $\lambda \geq \|X^\top y\|_\infty = \max_i\{|X^\top y|_i\}$ then the solution $\beta = 0$ is optimal. In fact, the equicorrelation set only changes whenever the value of λ coincides with one of the correlations $|X^\top y|_i$.

This gives us endpoints to the values of λ to be analyzed. Thanks to the solution path being piece-wise linear, this "homotopy" method has be calculated exactly in finitely-many steps. Together they ensure the KKT conditions are preserved at every knot and produce the piecewise-linear lasso path.

Algorithm 2: Least Angle Regression

procedure LARS(A, n, v)
  ▷ Standardize the predictors to have mean zero and unit variance. Define residual and parameters.
  $r \leftarrow y - \bar{y}$
  $\beta_1, ..., \beta_p \leftarrow 0$

  ▷ Find the predictor X_j most correlated with r
  $j \leftarrow \arg\max_j |X_j^\top r|$

  ▷ Move β_j from 0 towards its least squares coefficient $X_j^\top r$ until some other component X_k has as much correlation as X_j.
  ▷ Move β_j, β_k in the equiangular direction defined by their joint least squares coefficient of the current residual.
  $u_{\mathcal{A}} \leftarrow u_{\mathcal{A}} = X_{\mathcal{A}}\left(X_{\mathcal{A}}^\top X_{\mathcal{A}}\right)^{-1} s_{\mathcal{A}}$

  ▷ Calculate step size until either another variable becomes more correlated or becomes zero.

  $\gamma \leftarrow \min\left\{\min_{j \notin \mathcal{A}} \frac{C - c_j}{A - X_j^\top u},\ \min_{k \in \mathcal{A}} \frac{-\beta_k}{u_k}\right\}$

  ▷ If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the equiangular direction.

  $\beta_{\mathcal{A}} \leftarrow \beta_{\mathcal{A}} + \gamma u_{\mathcal{A}}$
  $\beta_{\mathcal{A}^C} \leftarrow 0$

  ▷ Continue until all variables are included in the model.
end

7 Extensions

The lasso can be readily extended to generalised linear models through the ensuing loss functions. In fact, for any model that relies on minimizing a loss function $l(\beta)$—the log-likelihood for GLMs—we can modify this to promote sparsity by solving

$$\min_{\beta} \ l(\beta) + \|\beta\|_1,$$

in an IRLS scheme with Newton-Raphson with a lasso component. **(Tibshirani 1996)** provides an example of logistic regression applied on kyphosis data.

Another important generalization is that of using the l_1 norm to enforce certain structural constraints—instead of pure sparsity—on the coefficients in a linear regression. These problems are nicely encapsulated by the formulation from the *generalized lasso*, given by

$$\min_{\beta \in \mathbb{R}^p} \ \frac{1}{2}\|y - X\beta\|_2^2 + \|D\beta\|_1, \tag{8}$$

where $D \in \mathbb{R}^{m \times p}$ is a specified penalty matrix. Depending on the application, one should choose D so that the sparsity of $D\beta$ corresponds to some other desired behavior for β, typically one that is structural or geometric in nature. The solution to these problems are implemented in the `genlasso` library in `R`. As a prime example, consider a sequence of observations—such as those arising from a time series—and we wish to model the underlying process. As a signal approximation problem, **Equation 8** would look like

$$\min_{\beta \in \mathbb{R}^p} \ \frac{1}{2}\|y - \beta\|_2^2 + \lambda \ \|D\beta\|_1,$$

where if we wanted to impose sparsity on the *jumps* in the predicted sequence, the matrix D would look like

$$D = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ 0 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}, \quad \Rightarrow \quad \|D\beta\|_1 = \sum_{i=1}^{p} |\beta_{i+1} - \beta_i|.$$

This is known as the fused lasso, which aims to promote sparsity in the sense of local constancy of the coefficient profile **(Tibshirani et al. 2005)**.

If D is $p \times p$ and invertible, then we may apply the change of variables $\theta = D\beta$ in **Equation 8** to obtain

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{2}\|y - XD^{-1}\theta\|_2^2 + \lambda\|\theta\|_1,$$

which is a regular lasso problem in $\theta \in \mathbb{R}^p$. In fact, so long as D is $m \times p$ of rank m we can construct a matrix $\tilde{D} = \binom{D}{A}$ of rank p by setting the other rows to be a $(p-m) \times p$ matrix A whose rows are orthogonal to those in D. Then the change of variables $\theta = (\theta_1, \theta_2)^\top = \tilde{D}\beta$ almost produces a lasso problem since only the θ_1 are covered by an l_1 penalty, so the coeffecients θ_2 are fitted by a linear regression and the remaining terms are a pure lasso problem.

To learn more about the generalized lasso such as what happens when $m > p$ or its solution path, check out **(Tibshirani & Taylor 2011)**.

Bibliography

Chen SS, Donoho DL, Saunders MA. 1998. Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*. 20(1):33–61

Efron B. 2004. The Estimation of Prediction Error. *Journal of the American Statistical Association*. 99(467):619–32

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression. *The Annals of Statistics*. 32(2):407–99

Hastie T, Tibshirani R, Friedman JH. 2009. *The Elements of Statistical Learning*. New York: Springer New York. 2nd ed.

Knight K, Fu W. 2000. Asymptotics for Lasso-Type Estimators. *The Annals of Statistics*. 28(5):1356–78

Tibshirani R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 58(1):267–88

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. 2005. Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 67(1):91–108

Tibshirani RJ, Taylor J. 2011. The solution path of the generalized lasso. *The Annals of Statistics*. 39(3):1335–71

Zou H, Hastie T. 2005. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 67(2):301–20

Zou H, Hastie T, Tibshirani R. 2007. On the "degrees of freedom" of the lasso. *The Annals of Statistics*. 35(5):2173–92