Sparsity and the Lasso

Statistical Machine Learning, Spring 2015

Ryan Tibshirani (with Larry Wasserman)

1 Regularization and the lasso

1.1 A bit of background

- If ℓ_2 was the norm of the 20th century, then ℓ_1 is the norm of the 21st century ... OK, maybe that statement is a bit dramatic, but at least so far, there's been a frenzy of research involving the ℓ_1 norm and its sparsity-inducing properties. Why? Where did this come from?

- In short, the ℓ_1 norm provides a computationally attractive, theoretically well-justified form of *regularization*. Starting off in the regression setting, suppose that we observe an outcome vector $y \in \mathbb{R}^n$ and a predictor matrix $X \in \mathbb{R}^{n \times p}$, whose columns $X_1, \ldots X_p \in \mathbb{R}^n$ correspond to predictor variables. Linear regression solves the least squares problem

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j X_j)^2 \quad \Longleftrightarrow \quad \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2.$$

  When $\mathrm{rank}(X) = p$, i.e, the predictor variables $X_1, \ldots X_p$ are linearly independent, the above least squares criterion is strictly convex, and so there exists a unique solution $\hat{\beta}$. But when $\mathrm{rank}(X) < p$—e.g., this happens when $p > n$—there are infinitely many solutions; given any one solution $\hat{\beta}$, the quantity $\hat{\beta} + \eta$ is also a solution for any $\eta \in \mathrm{null}(X)$. Furthermore, this type of nonuniqueness makes interpretation of solutions meaningless: it follows that for at least one variable $j \in \{1, \ldots p\}$, we will have $\hat{\beta}_j > 0$ at one solution $\hat{\beta}$, but $\tilde{\beta}_j < 0$ at another solution $\tilde{\beta}$

- Moreover, suppose that $\mathrm{rank}(X) = p$, so that a unique linear regression estimate exists. If p is moderately close to n, then the linear regression estimate can have quite poor predictive accuracy. This is because its variance scales linearly with p

- How do we deal with such problems? One way is to constrain the least squares estimator, i.e., choose $\hat{\beta}$ to solve

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \ \text{ subject to } \ \beta \in C,$$

  for some set $C \subseteq \mathbb{R}^p$. We call the constraint $\beta \in C$ a form of regularization. Obviously, the choice of C is going to have a big effect on the accuracy of our estimate $\hat{\beta}$. It will also have an effect on the interpretability of our estimate

- Two well-known examples are given by choosing C to the ℓ_2 ball, and the ℓ_1 ball. This reveals the problems

$$\text{Ridge regression:} \quad \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \ \text{ subject to } \ \|\beta\|_2 \leq t, \tag{1}$$

$$\text{Lasso regression:} \quad \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \ \text{ subject to } \ \|\beta\|_1 \leq t. \tag{2}$$

Though it is likely the case that these ideas were around earlier in other contexts, in statistics we typically credit ridge regression to Hoerl & Kennard (1970), and the lasso to Tibshirani (1996), Chen et al. (1998). In each of the above problems, $t \geq 0$ serves as a parameter that controls the size of the constraint set. Also, in each problem, the constraint set is convex, so the problems are convex overall and this makes them computationally tractable

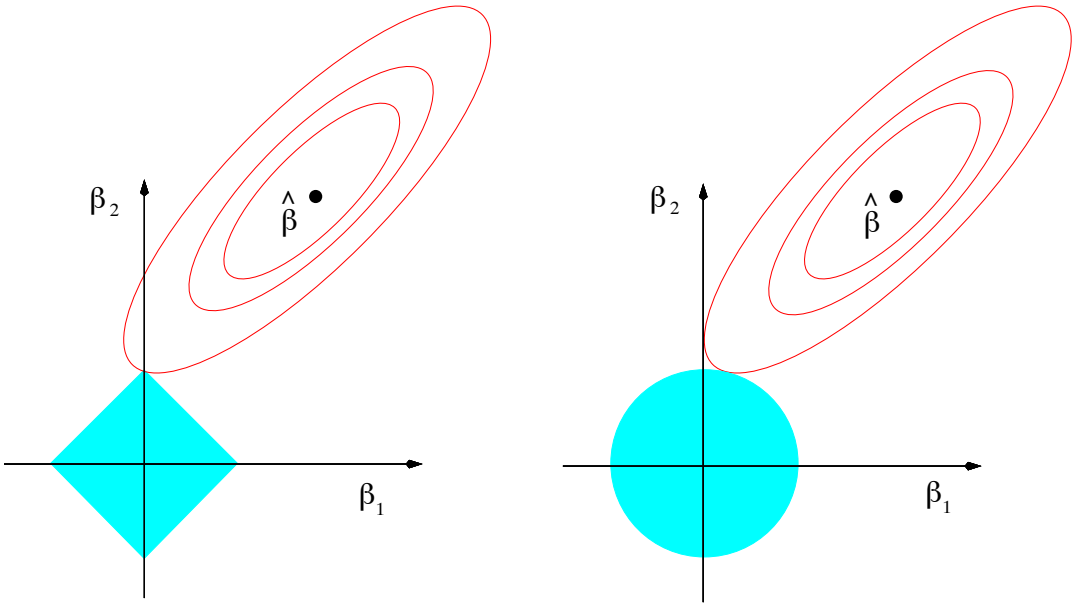

Figure 1: *The "classic" illustration comparing lasso and ridge constraints. From Chapter 3 of Hastie et al. (2009)*

- At a high level, what good do these norm constraints do? They restrict our estimate to lie in a ball around 0, with respect to some norm; i.e., they shrink the estimated coefficients toward 0, in a particular way; see Figure 1. The smaller the value of the tuning parameter t, the more shrinkage. Shrinkage reduces the variance of our estimate, at the expense of (likely) introducing bias. It is often the case that the reduction in variance (tuned at some level t) is well worth the price of bias, so that the tradeoff in overall accuracy is favorable. Think about what would happen if we centered the ball round some point $a \in \mathbb{R}^p$ (instead of 0)—what would happen to the variance? The bias?

- The ℓ_1 norm has a further, highly special property: constraining the coefficients by their ℓ_1 norm induces *sparsity* in the estimate. That is, at the lasso solution $\hat{\beta}$, we will have $\hat{\beta}_j = 0$ for many components $j \in \{1, \ldots p\}$, and the smaller the value of the tuning parameter t, the fewer the number of nonzero components in $\hat{\beta}$. This is not true with ridge regression. Note that such sparsity is desirable, because it corresponds to performing variable selection in the constructed model; this provides a level of interpretability (beyond sheer accuracy)

- From duality and the KKT conditions, we know we can rewrite problems (1), (2) as

$$\text{Ridge regression:} \quad \min_{\beta \in \mathbb{R}^p} \ \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2, \tag{3}$$

$$\text{Lasso regression:} \quad \min_{\beta \in \mathbb{R}^p} \ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1, \tag{4}$$

with $\lambda \geq 0$ the tuning parameter. These versions are called the *penalized* (or Lagrange) forms of the problems, whereas the ones above are called the *constrained* forms. I.e., for each of the constrained problems above, and every value of $t \geq 0$, there is a corresponding value of $\lambda \geq 0$ such that the analogous penalized problem has the same solutions

- From here on we will focus on the penalized form of the lasso problem, and we will use a combination of optimization tools and statistics tools to describe some of its favorable properties. Much of what is said will also apply to the more general problem

$$\min_{\beta \in \mathbb{R}^p} \; f(X\beta) + \lambda \|\beta\|_1, \tag{5}$$

for some smooth convex loss function f. Note that when $f(X\beta)$ is a negative log likelihood, this covers ℓ_1 penalized maximum likelihood estimation (e.g., ℓ_1 penalized logistic regression). The lasso literature has become immense, and so we'll only cover a handful of interesting theoretical properties. Come and ask one of us and you are interested in hearing more. Also, we won't talk much about computation, but you should keep it in mind that there are many fast algorithms for solving the lasso regression problem (and the more general ℓ_1 penalized problem (5))

1.2 Let's back up: what about the ℓ_0 norm?

- Yeah, yeah, the ℓ_1 norm and the lasso sounds great and all, but what about ℓ_0 norm? People have considered variable selection with the ℓ_0 norm, also called *best subset selection*, for a long time (long before the lasso). Written in a form to match the penalized lasso problem, this is

$$\text{Subset selection :} \qquad \min_{\beta \in \mathbb{R}^p} \; \|y - X\beta\|_2^2 + \lambda \|\beta\|_0, \tag{6}$$

for some tuning parameter $\lambda \geq 0$, where to remind you, $\|\beta\|_0 = \sum_{j=1}^p 1\{\beta_j \neq 0\}$

- This is arguably more natural as a variable selector, because we are directly penalizing the presence of nonzero components in β, and so this will induce sparsity in the subset selection solution $\hat{\beta}$. The trouble is that the above problem is *nonconvex*, which makes it difficult to understand theoretically, and especially, to solve computationally (it is known to be NP-hard). This is because of the presence of the ℓ_0 penalty

- We do, however, understand some theoretical properties of this estimator. A classic result is due to Foster & George (1994), which we will paraphrase here. First we raise a very simple point: if A denotes the support (also called the active set) of the subset selection solution $\hat{\beta}$ in (6)—meaning that $\hat{\beta}_j = 0$ for all $j \notin A$, and written $A = \text{supp}(\hat{\beta})$—then we have

$$\begin{aligned} \hat{\beta}_A &= (X_A^T X_A)^{-1} X_A^T y, \\ \hat{\beta}_{-A} &= 0. \end{aligned} \tag{7}$$

(Here and throughout we write X_A to denote the columns of matrix X in a set A, and x_A to denote the components of a vector x in A. We will also use X_{-A} and x_{-A} for the columns or components not in A.) This observation follows from the fact that, given the support set A, the ℓ_0 penalty term in the subset selection criterion doesn't depend on the actual magnitudes of the coefficients (it contributes a constant factor), so the problem reduces to least squares. Now for the Foster & George (1994) result. These authors consider a standard linear model

$$y = X\beta_0 + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I),$$

where $\beta_0 \in \mathbb{R}^p$ are some true unknown regression coefficients, and the predictors X are considered fixed. They examine the predictive risk of an estimate $\hat{\beta}$ of the regression coefficients,

$$\frac{1}{n}\mathbb{E}_{\beta_0}\|X\hat{\beta} - X\beta_0\|_2^2.$$

(The subscript β_0 on the expectation emphasizes the fact that the expectation here is taken with respect to the distribution of y in which the true regression coefficients are β_0.) Suppose that β_0 is supported on a subset $S \subseteq \{1, \dots p\}$, denoted by $S = \text{supp}(\beta_0)$, and let $s_0 = |S|$. Consider the estimator given by least squares on S:

$$\hat{\beta}_S^{\text{oracle}} = (X_S^T X_S)^{-1} X_S^T y,$$
$$\hat{\beta}_{-S}^{\text{oracle}} = 0.$$

We call $\hat{\beta}^{\text{oracle}}$ an *oracle estimator* because it is defined in terms of the unknown and unobtainable true support set S. Note that the risk of this oracle estimator is easily computed:

$$\frac{1}{n}\mathbb{E}_{\beta_0}\|X\hat{\beta}^{\text{oracle}} - X\beta_0\|_2^2 = \frac{1}{n}\mathbb{E}_{\beta_0}\|X_S(X_S^T X_S)^{-1}X_S\epsilon\|_2^2 = \frac{\sigma^2 s_0}{n}.$$

Foster & George (1994) compare the risk of the best subset selection estimator $\hat{\beta}$ to the oracle risk $\sigma^2 s_0$. They show that, if we choose λ on the order of $2\sigma^2 \log p$, then the best subset selection estimator satisfies

$$\max_{X,\,\beta_0} \frac{\mathbb{E}_{\beta_0}\|X\hat{\beta} - X\beta_0\|_2^2/n}{\sigma^2 s_0/n} \leq 4\log p + 2 + o(1),$$

as $p \to \infty$. I.e., its worst-case risk, over all predictor matrices X and underlying coefficients β_0, is no more a $\log p$ factor larger than the oracle risk. Further, they prove the lower bound

$$\min_{\hat{\beta}} \max_{X,\,\beta_0} \frac{\mathbb{E}_{\beta_0}\|X\hat{\beta} - X\beta_0\|_2^2/n}{\sigma^2 s_0/n} \geq 2\log p - o(\log p),$$

where the outer minimum is over *all estimators* $\hat{\beta}$. Therefore, in terms of rate, best subset selection achieves the optimal risk inflation from the oracle risk

- Returning to what was said above, the kicker is that we can't really compute the best subset selection estimator for even moderately-sized problems. As we will in the following, the lasso provides a similar risk inflation guarantee, though under stronger assumptions

- Finally, it is worth mentioning that even if we *could* compute the subset selection estimator at scale, it's not at all clear that we would even want to use this in place of the lasso. (Many people assume that we would.) We must remember that theory gives us an understanding of the performance of various estimators under typically idealized conditions, and it doesn't tell the complete story. It could be the case that the lack of shrinkage in the subset selection coefficients ends up being harmful in practical situations, where the signal-to-noise ratio is really low, and the lasso could still perform favorably in such settings

2 Basic properties

2.1 Sign patterns

- Let's start with some basics. For convenience, we'll rescale the lasso problem in (4), and from here on, we'll consider

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

A few basic observations:

1. There need not always be a unique solution $\hat{\beta}$ in (8), because the criterion is not always strictly convex in β (e.g., this happens when $p > n$).
2. There is however always a unique fitted value $X\hat{\beta}$ in (8), because the least squares loss is strictly convex in $X\beta$.

The first observation is potentially worrisome; it would be bad if we encountered the same problems with interpretation as we did with ordinary least squares regression. We will see shortly that there is nothing really to worry about. The second observation is standard (it is also true in least squares), but will be helpful

- Now we turn to the KKT conditions, i.e., subgradient optimality, for the lasso problem in (8). This tells us that any lasso solution $\hat{\beta}$ must satisfy

$$X^T(y - X\hat{\beta}) = \lambda s, \tag{9}$$

where $s \in \partial\|\hat{\beta}\|_1$, a subgradient of the ℓ_1 norm evaluated at $\hat{\beta}$. Recall that this means

$$s_j \in \begin{cases} \{+1\} & \hat{\beta}_j > 0 \\ \{-1\} & \hat{\beta}_j < 0 \\ [-1, 1] & \hat{\beta}_j = 0, \end{cases} \quad j = 1, \ldots p \tag{10}$$

- From (9) we can read off a straightforward but important fact: even though the solution $\hat{\beta}$ may not be uniquely determined, the optimal subgradient s is a function of the unique fitted value $X\hat{\beta}$ (assuming $\lambda > 0$), and hence is itself unique

- Now from (10), notice that uniqueness of s implies that any two lasso solutions must have the same signs on the overlap of their supports. That is, it cannot happen that we find two different lasso solutions $\hat{\beta}$ and $\tilde{\beta}$ with $\hat{\beta}_j > 0$ but $\tilde{\beta}_j < 0$ for some j, and hence we have no problem interpretating the signs of components of lasso solutions

2.2 Canonical representation

- Now we'll aim to represent the lasso estimate in an analogous form what we did for best subset selection in (7). Define the *equicorrelation set*

$$E = \big\{j \in \{1, \ldots p\} : |X_j^T(y - X\hat{\beta})| = \lambda\big\}.$$

This is the set of variables that achieves the maximum absolute inner product (i.e., correlation for standard predictors) with the lasso residual vector. Assuming that $\lambda > 0$, this is the same

$$E = \big\{j \in \{1, \ldots p\} : |s_j| = 1\big\}.$$

This is a uniquely determined set (since $X\hat{\beta}, s$ are unique). Also, we remark that E contains the active set $A = \text{supp}(\hat{\beta})$ of any lasso solution $\hat{\beta}$, because for $j \notin E$, we have $|s_j| < 1$, which implies that $\hat{\beta}_j = 0$

- We can hence write the KKT conditions (9) over the equicorrelation variables as:

$$X_E^T(y - X\hat{\beta}) = \lambda s_E.$$

As we said, any lasso solution must satisfy $\hat{\beta}_{-E} = 0$, since $|s_j| < 1$ whenever $j \notin E$, therefore $X\hat{\beta} = X_E\hat{\beta}_E$, and the above becomes

$$X_E^T(y - X_E\hat{\beta}_E) = \lambda s_E,$$

which is a linear system that we can solve for $\hat{\beta}_E$

- Doing so gives the representation

$$\begin{aligned} \hat{\beta}_E &= (X_E^T X_E)^+ \big(X_E^T y - \lambda s_E\big) + \eta, \\ \hat{\beta}_{-E} &= 0. \end{aligned} \tag{11}$$

  Here $(X_E^T X_E)^+$ denotes the (Moore-Penrose) pseudoinverse of $X_E^T X_E$, and η is a vector in the null space of X_E. To be clear: we have shown that any lasso solution $\hat{\beta}$ must satisfy (11), for some $\eta \in \text{null}(X_E)$ that results in $\text{sign}(\hat{\beta}_E) = s_E$

- We come to another straightforward but important fact: if X_E has column full rank (i.e., the equicorrelation variables are linearly independent), then $\text{null}(X_E) = \{0\}$, and hence there is only one choice for η, i.e., $\eta = 0$. In other words, when X_E has full column rank, there is a unique lasso solution $\hat{\beta}$ and it satisfies

$$\begin{aligned} \hat{\beta}_E &= (X_E^T X_E)^{-1} \big(X_E^T y - \lambda s_E\big), \\ \hat{\beta}_{-E} &= 0. \end{aligned} \tag{12}$$

  Examining the above: the nonzero lasso coefficients $\hat{\beta}_E$ are given by the least squares coefficients $(X_E^T X_E)^{-1} X_E^T y$ plus the shrinkage term $-(X_E^T X_E)^{-1} \lambda s_E$. This term precisely characterizes the shrinkage in the estimated lasso coefficients. Finally, it is not hard to see that, except for $y \in \mathbb{R}^n$ on a set of measure zero, the solution in (12) satisfies $\hat{\beta}_j \neq 0$ for all $j \in E$, so its support is indeed $A = \text{supp}(\hat{\beta}) = E$

2.3 Uniqueness of solution

- The question is now: when does X_E have linearly independent columns? If this is the case, then the lasso solution $\hat{\beta}$ is unique and given by (12). Clearly, this happens when X itself has linearly independent columns, i.e., $\text{rank}(X) = p$, in the $p \leq n$ setting

- However, it turns out that X_E has full column rank in a much broader setting than this one. One can show that X_E has full column rank, for any value of $\lambda > 0$, provided that X has columns in *general position* (Tibshirani 2013). This is a very weak condition on X and it can hold regardless of the (relative) sizes of n and p. Formally, we say that $X_1, \ldots X_p \in \mathbb{R}^n$ are in general position provided that for any $k < \min\{n, p\}$, indices $i_1, \ldots i_{k+1} \in \{1, \ldots p\}$, and signs $\sigma_1, \ldots \sigma_{k+1} \in \{-1, +1\}$, the affine span of $\sigma_1 X_{i_1}, \ldots \sigma_{k+1} X_{i_{k+1}}$ does not contain any element of $\{\pm X_i : i \neq i_1, \ldots i_{k+1}\}$. This is equivalent to the following statement: no k-dimensional subspace $L \subseteq \mathbb{R}^n$, for $k < \min\{n, p\}$, contains more that $k + 1$ points of $\{\pm X_1, \ldots \pm X_p\}$, excluding antipodal pairs (i.e., $+X_i$ and $-X_i$).

- In summary, general position of X implies that X_E must have full column rank at any value of $\lambda > 0$, which ensures uniqueness of the lasso solution $\hat{\beta}$. This is quite a favorable result for the lasso; essentially, we only get nonunique solutions when there are exact collinearities between small subsets of variables

- Actually, the exact same arguments carry over to the more general ℓ_1-penalized problem (5), assuming that f is strictly convex. The KKT conditions for this problem are now

$$-X^T \nabla f(X\hat{\beta}) = \lambda s,$$

  where $s \in \partial \|\hat{\beta}\|_1$, as before. It is not hard to see that the fitted value $X\hat{\beta}$ is always unique, because f is strictly convex. This makes the optimal subgradient s unique, and hence any two solutions of (5) must again have signs that agree on the overlap of their supports. Also, even

though we do not have a canonical representation for the solution like we developed for the lasso regression estimate in the last section, we can still define the equicorrelation set

$$E = \big\{j \in \{1, \dots p\} : |X_j^T \nabla f(X\hat{\beta})| = \lambda\big\} = \big\{j \in \{1, \dots p\} = 1\big\},$$

and we can show that when X_E has full column rank, there is a unique solution $\hat{\beta}$ in (5). This follows from noting that $\hat{\beta}_{-E} = 0$, and $\hat{\beta}_E$ solves the problem

$$\min_{\beta \in \mathbb{R}^{|E|}} \; f(X_E \beta_E) + \lambda \|\beta_E\|_1,$$

whose criterion is strictly convex. Finally, we if X has columns in general position, then X_E must have full column rank, and therefore there is a unique solution $\hat{\beta}$ in (5)[1]

3 Regularization path

- From the lasso KKT conditions (9), (10), it is possible to compute the lasso estimate in (8) as a function of λ, which we will denote by $\hat{\beta}(\lambda)$, for all values of the tuning parameter $\lambda \in [0, \infty]$. This is called the *regularization path* or *solution path* of the problem (8). Note that the path provides an exhaustive, explicit characterization of the tradeoff between the least squares loss term and the ℓ_1 penalty term

- Path algorithms such as the one we will describe are not always possible; the reason that this ends up being feasible for the lasso problem (8) is that the solution path $\hat{\beta}(\lambda)$, $\lambda \in [0, \infty]$ turns out to be a piecewise linear, continuous function of λ. Hence, we only need to compute and store the knots in this path, which we will denote by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0$, and the lasso solution at these knots. From this information, we can then compute the lasso solution at any value of λ by linear interpolation

- The knots $\lambda_1 \geq \ldots \geq \lambda_r$ in the lasso path correspond to values of λ at which the active set $A(\lambda) = \text{supp}(\hat{\beta}(\lambda))$ changes. As we decrease λ from ∞ to 0, each knot usually corresponds to the entry of a variable into the active set; this connects the lasso to an incremental variable selection procedure like forward stepwise selection. Interestingly though, as we decrease λ, a knot in the lasso path can also correspond to the deletion of a variable from the active set. See Figure 2

- The lasso solution path was described by Osborne et al. (2000*a*,*b*), Efron et al. (2004). Like the construction of all other solution paths that followed these seminal works, the lasso path is essentially given by an iterative or inductive verification of the KKT conditions; if we can maintain that the KKT conditions holds as we decrease λ, then we know we have a solution. The trick is to start at a value of λ at which the solution is trivial; for the lasso, this is $\lambda = \infty$, at which case we know the solution must be $\hat{\beta}(\infty) = 0$

- We will assume for simplicity that the columns of X are in general positon, so that a unique lasso solution exists. General position also implies that at every value of λ, the matrix $X_A^T X_A$ is invertible, where $A = A(\lambda)$ is the lasso active set. Note that this follows from our arguments above. This simplification is not necessary; we can also construct a solution path even when the solution is not unique, for a completely generic X

[1]To be careful, all statements for this general problem (5) should be interpreted conditional on the assumption of the existence of a solution in the first place. Just because a function is smooth and strictly convex doesn't mean that it achieves its minimum; e.g., the function $f(x) = e^{-x}$ is smooth and strictly convex but has no minimizer on $\mathbb{R}$.

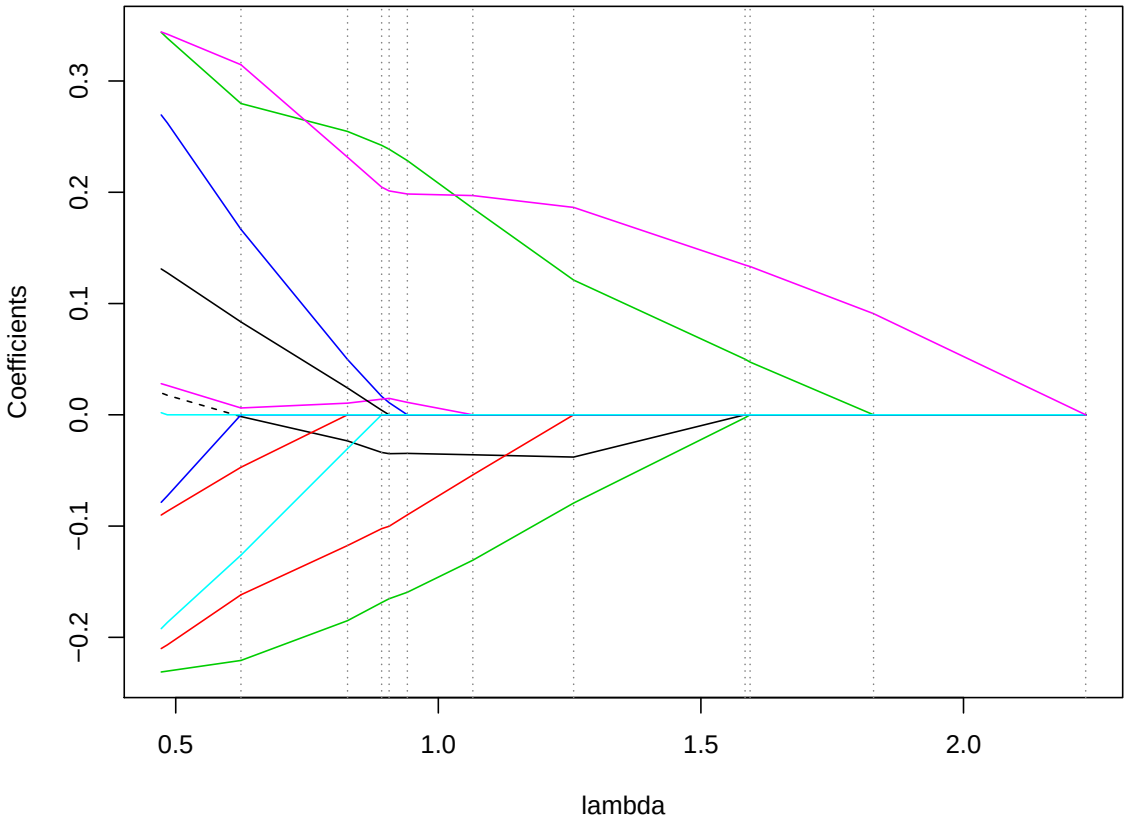

Figure 2: *An example of the lasso path. Each colored line denotes a component of the lasso solution* $\hat{\beta}_j(\lambda)$, $j = 1, \ldots p$ *as a function of* λ. *The gray dotted vertical lines mark the knots* $\lambda_1 \geq \lambda_2 \geq \ldots$.

- Now we will walk through the lasso path, starting at $\lambda = \infty$ and $\hat{\beta}(\infty) = 0$, as indicated above. Consider decreasing λ from ∞, and continuing to set $\hat{\beta}(\lambda) = 0$ as the lasso solution. The KKT conditions (9) read

$$X^T y = \lambda s,$$

where s is a subgradient of the ℓ_1 norm evaluated at 0, i.e., $s_j \in [-1, 1]$ for every $j = 1, \ldots p$. For large enough values of λ, this is satisfied, since we can choose $s = X^T y/\lambda$. But this ceases to be a valid subgradient if we decrease λ past the point at which $\lambda = |X_j^T y|$ for some variable $j = 1, \ldots p$. In short, $\hat{\beta}(\lambda) = 0$ is the lasso solution for all $\lambda \geq \lambda_1$, where

$$\lambda_1 = \max_{j=1,\ldots p} |X_j^T y|. \tag{13}$$

What happens next? As we decrease λ from λ_1, we know that we're going to have to change $\hat{\beta}(\lambda)$ from 0 so that the KKT conditions remain satisfied. Let j_1 denote the variable that achieves the maximum in (13). Since the subgradient was $|s_{j_1}| = 1$ at $\lambda = \lambda_1$, we see that we are "allowed" to make $\hat{\beta}_{j_1}(\lambda)$ nonzero. Consider setting

$$\begin{aligned} \hat{\beta}_{j_1}(\lambda) &= (X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}) \\ \hat{\beta}_j(\lambda) &= 0, \quad \text{for all } j \neq j_1, \end{aligned} \tag{14}$$

as λ decreases from λ_1, where $s_{j_1} = \text{sign}(X_{j_1}^T y)$. Note that this makes $\hat{\beta}(\lambda)$ a piecewise linear

and continuous function of λ, so far. The KKT conditions are then

$$X_{j_1}^T\Big(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1})\Big) = \lambda s_{j_1},$$

which can be checked with simple algebra, and

$$\Big|X_j^T\Big(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1})\Big)\Big| \leq \lambda,$$

for all $j \neq j_1$. Recall that the above held with strict inequality at $\lambda = \lambda_1$ for all $j \neq j_1$, and by continuity of the constructed solution $\hat{\beta}(\lambda)$, it should continue to hold as we decrease λ for at least a little while. In fact, it will hold until one of the piecewise linear paths $X_j^T(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}))$ becomes equal to $\pm\lambda$, at which point we have to modify the solution because otherwise the implicit subgradient $s_j = X_j^T(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}))/\lambda$ will cease to be in $[-1, 1]$. It helps to draw yourself a picture of this

Thanks to linearity, we can compute the critical "hitting time" explicitly; a short calculation shows that, the lasso solution continues to be given by (14) for all $\lambda_1 \geq \lambda \geq \lambda_2$, where

$$\lambda_2 = \max_{j \neq j_1,\, s_j \in \{-1,1\}}{}^{+} \frac{X_j^T(I - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}X_{j_1})y}{s_j - X_j^T X_{j_1}(X_{j_1}^T X_{j_1})^{-1}s_{j_1}}, \tag{15}$$

and $\max^+$ denotes the maximum over all of its arguments that are $< \lambda_1$

To keep going: let j_2, s_2 achieve the maximum in (15). Let $A = \{j_1, j_2\}$, $s_A = (s_{j_1}, s_{j_2})$, and consider setting

$$\begin{aligned} \hat{\beta}_A(\lambda) &= (X_A^T X_A)^{-1}(X_A^T y - \lambda s_A) \\ \hat{\beta}_j(\lambda) &= 0, \quad \text{for all } j \notin A, \end{aligned} \tag{16}$$

as λ decreases from λ_2. Again, we can verify the KKT conditions for a stretch of decreasing λ, but will have to stop when one of the implicit subgradients $s_j = X_j^T(y - X_A(X_A^T X_A)^{-1}(X_A^T y - \lambda s_A)$ becomes equal to $\pm\lambda$, for some $j \notin A$. From linearity, we can compute this next "hitting time" explicitly, as before. Furthermore, though, we will have to check whether the active components of the computed solution in (16) are going to cross through zero, because past such a point, s_A will no longer be a proper subgradient over the active components. We can again compute this next "crossing time" explicitly, due to linearity. Therefore, we maintain that (16) is the lasso solution for all $\lambda_2 \geq \lambda \geq \lambda_3$, where λ_3 is the maximum of the next hitting time and the next crossing time

For convenience, the lasso path algorithm is summarized below.

Algorithm 1 (**Lasso path algorithm**).

Given y and X.

- *Start with the iteration counter $k = 0$, regularization parameter $\lambda_0 = \infty$, active set $A = \emptyset$, and active signs $s_A = \emptyset$*
- *While $\lambda_k > 0$:*
  1. *Compute the lasso solution as λ decreases from λ_k by*

$$\begin{aligned} \hat{\beta}_A(\lambda) &= (X_A^T X_A)^{-1}(X_A^T y - \lambda s_A) \\ \hat{\beta}_j(\lambda) &= 0, \quad \textit{for all } j \notin A \end{aligned}$$

2. *Compute the next hitting time,*

$$\lambda_{k+1}^{\text{hit}} = \max_{j \notin A,\, s_j \in \{-1,1\}}{}^{+} \frac{X_j^T(I - X_A(X_A^TX_A)^{-1}X_A^T)y}{s_j - X_j^TX_A(X_A^TX_A)^{-1}s_A}$$

where $\max^+$ *denotes the maximum of its arguments* $< \lambda_k$

3. *Compute the next crossing time,*

$$\lambda_{k+1}^{\text{cross}} = \max_{j \in A}{}^{+} \frac{[(X_A^TX_A)^{-1}X_A^Ty]_j}{[(X_A^TX_A)^{-1}s_A]_j},$$

where $\max^+$ *denotes the maximum of its arguments* $< \lambda_k$

4. *Decrease* λ *until* λ_{k+1}*, defined by*

$$\lambda_{k+1} = \max\{\lambda_{k+1}^{\text{hit}}, \lambda_{k+1}^{\text{cross}}\}$$

5. *If* $\lambda_{k+1}^{\text{hit}} > \lambda_{k+1}^{\text{cross}}$*, then add the hitting variable to* A *and its sign to* s_A*; otherwise, remove the crossing variable from* A *and its sign from* s_A*. Update* $k = k + 1$

- One last point: as we decrease λ from a knot λ_k we can view the definition of the lasso path in Step 1 above as

$$\hat{\beta}_A(\lambda) = \hat{\beta}_A(\lambda_k) + (\lambda_k - \lambda)(X_A^TX_A)^{-1}s_A$$
$$\hat{\beta}_j(\lambda) = 0, \quad \text{for all } j \notin A.$$

  We can see that we are moving the active coefficients in the direction $(\lambda_k - \lambda)(X_A^TX_A)^{-1}s_A$ for decreasing λ. In other words, the lasso fitted values proceed as

$$X\hat{\beta}(\lambda) = X\hat{\beta}(\lambda_k) + (\lambda_k - \lambda)X_A(X_A^TX_A)^{-1}s_A,$$

  for decreasing λ. Efron et al. (2004) call $X_A(X_A^TX_A)^{-1}s_A$ the *equiangular direction*, because this direction, in $\mathbb{R}^n$, takes an equal angle with all $X_j \in \mathbb{R}^n$, $j \in A$

  For this reason, the lasso path algorithm above is also often referred to as the *least angle regression* path algorithm in "lasso mode", though we have not mentioned this yet to avoid confusion. Least angle regression is considered as another algorithm by itself, where Step 3 is skipped altogether. In words, Step 3 disallows any component path to cross through zero. The left side of the plot in Figure 2 visualizes the distinction between least angle regression and lasso estimates: the dotted black line displays the least angle regression component path, crossing through zero, while the lasso component path remains at zero

4 Theoretical results

4.1 Rates of convergence

- There has been an enormous amount of theoretical work on the lasso, and so we only briefly summarize it here. Some early references include Greenshtein & Ritov (2004), Fuchs (2005), Donoho (2006), Candes & Tao (2006); a helpful text with plenty of references for these results is Buhlmann & van de Geer (2011)

- This section surveys rates of convergence for the lasso estimator. Assume that

$$y = X\beta_0 + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I),$$

  for fixed X and β_0. Also assume that $\|X_j\|_2^2 = n$, for $j = 1, \ldots p$. For brevity, we will not go through any proofs, but we do note that the proof of the basic result is particularly simple

- Basic result: for $\lambda = O(\sigma\sqrt{n \log p})$, the lasso fit satisfies

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 = O_{\mathbb{P}}\Big(\sigma\sqrt{\frac{\log p}{n}} \cdot \|\beta_0\|_1\Big). \tag{17}$$

  Hence, if the ℓ_1 norm of the true coefficients $\|\beta_0\|_1$ grows slower that $\sqrt{n/\log p}$, then lasso fit is consistent. Note that there are no real assumptions on X here (except the norm condition on its columns)

- Compatibility results: assume that X satisfies the *compatibility condition* for the true support set $S = \mathrm{supp}(\beta_0)$, with compatibility constant ϕ_0; i.e., letting $s_0 = |S|$, this means that

$$\frac{1}{n}\|X\beta\|_2^2 \geq \frac{\phi_0^2}{s_0} \cdot \|\beta_S\|_1^2 \quad \text{for all } \beta \in \mathbb{R}^p \text{ such that } \|\beta_{-S}\|_1 \leq 3\|\beta_S\|_1.$$

  Then for $\lambda = O(\sigma\sqrt{n \log p})$, we have

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 = O_{\mathbb{P}}\Big(\sigma^2\frac{\log p}{n} \cdot \frac{s_0}{\phi_0^2}\Big), \tag{18}$$

  and

$$\|\hat{\beta} - \beta_0\|_1 = O_{\mathbb{P}}\Big(\sigma\sqrt{\frac{\log p}{n}} \cdot \frac{s_0}{\phi_0^2}\Big)$$

- Look at the result on the lasso fit in (18): this says that the risk inflation of the lasso (using the oracle risk of $\sigma^2 s_0/n$ as a reference point) is of the order $\log p$. Recall that we saw we could achieve the same with best subset selection, and that this was minimax optimal. However, with best subset selection, there were no conditions on X; here we have to assume compatibility, which roughly means that the columns of X over the true support set S cannot be too correlated

- Restricted eigenvalue result: assume that X satisfies the *restricted eigenvalue condition* with constant ϕ_0, i.e.,

$$\frac{1}{n}\|X\beta\|_2^2 \geq \phi_0^2 \cdot \|\beta\|_2^2 \quad \text{for all subsets } J \subseteq \{1, \ldots p\} \text{ such that } |J| = s_0 \\ \text{and all } \beta \in \mathbb{R}^p \text{ such that } \|\beta_{J^c}\|_1 \leq 3\|\beta_J\|_1.$$

  Then for $\lambda = O(\sigma\sqrt{n \log p})$, we have

$$\|\hat{\beta} - \beta_0\|_2 = O_{\mathbb{P}}\Big(\sigma\sqrt{\frac{\log p}{n}} \cdot \frac{\sqrt{s_0}}{\phi_0^2}\Big)$$

- Note the similarity between this and the compatibility condition. The restricted eigenvalue condition actually stronger, i.e., it implies the compatibility condition, since $\|\beta\|_2^2 \geq \|\beta_J\|_2^2 \geq \|\beta_J\|_1^2/s$

- A simple, but noteworthy point: it was actually important to assume throughout here that the columns of X grow in norm, i.e., $\|X_i\|_2^2 = n$ for all $i = 1, \ldots p$. If we instead assumed that $\|X_i\|_2^2 = 1$ for $i = 1, \ldots p$, then we would have to assume that the error variance was shrinking, i.e., we would have to replace σ^2 by σ^2/n to achieve the same results

4.2 Proof of basic result

- The basic result (17) is sometimes called a *slow rate bound* or *weak error bound* for the lasso. While this rate is not as fast as the result in, say, (18), it is desirable because there are no conditions on X. It is also simple to prove, and we will do so here. Start by arguing that, by optimality of the lasso solution $\hat{\beta}$ for its criterion in (8),

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|y - X\beta_0\|_2^2 + \lambda\|\beta_0\|_1.$$

  Rearranging this gives

$$\|X\hat{\beta}\|_2^2 - \|X\beta_0\|_2^2 \leq 2y^T(X\hat{\beta} - X\beta_0) + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1),$$

  and plugging in $y = X\beta_0 + \epsilon$,

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\epsilon^T(X\hat{\beta} - X\beta_0) + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1). \tag{19}$$

  This is referred to as the *basic inequality* for the lasso

- The strategy is to choose λ large enough so that the second term in (19) dominates the first term, the empirical process $2\epsilon^T(X\hat{\beta} - X\beta_0)$. But we don't want to choose λ too large, because that will make our bound poor. Using Holder's inequality,

$$\epsilon^T(X\hat{\beta} - X\beta_0) \leq \|X^T\epsilon\|_\infty\|\hat{\beta} - \beta_0\|_1,$$

  and therefore if $\lambda \geq \|X^T\epsilon\|_\infty$, then we have

$$\begin{aligned}\|X\hat{\beta} - X\beta_0\|_2^2 &\leq 2\lambda\|\hat{\beta} - \beta_0\|_1 + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1) \\ &\leq 4\lambda\|\beta_0\|_1,\end{aligned} \tag{20}$$

  the last step coming from the triangle inequality

- Now we must figure how large λ must be so that $\lambda \geq \|X^T\epsilon\|_\infty$, with high probability. Note that

$$\|X^T\epsilon\|_\infty = \max_{j=1,\ldots p} |X_j^T\epsilon| \leq 1.01(\sigma\sqrt{n})\sqrt{2\log p},$$

  with probability tending to 1. Why? A standard Gaussian concentration result says that for any $\delta > 0$, the maximum of p absolute Gaussians with mean 0 and common variance τ^2 is bounded by $(1+\delta)\tau\sqrt{2\log p}$, with probability tending to 1. The above hence applies this to the univariate Gaussians $X_j^T\epsilon$, $j = 1, \ldots p$, that have common variance $\tau^2 = \sigma^2 n$

- Plugging $\lambda = 1.01(\sigma\sqrt{n})\sqrt{2\log p}$ into (20), and dividing by n, we conclude that

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq 4.04\sqrt{2} \cdot \sigma\sqrt{\frac{\log p}{n}}\|\beta_0\|_1,$$

  with probability tending to 1, as desired

4.3 Support recovery

- Lastly, we discuss results on support recovery of the lasso estimator. There are a few versions of support recovery results and again Buhlmann & van de Geer (2011) is a good place to look for a thorough coverage. Here we describe a result due to Wainwright (2009), who introduced a proof technique called the *primal-dual witness method*; related techniques are described by Zhao & Yu (2006), Candes & Plan (2009)

- Again we assume a standard linear model

$$y = X\beta_0 + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I),$$

for fixed X and β_0, with the scaling $\|X_j\|_2^2 \leq n$, for $j = 1, \ldots p$. Let $S = \text{supp}(\beta_0)$ be the true support set, and assume that X_S has full column rank. We aim to show that, at some value of λ, the lasso solution $\hat{\beta}$ has an active set that exactly equals the true support set,

$$A = \text{supp}(\hat{\beta}) = S,$$

with high probability. We actually aim to show that the signs also match,

$$\text{sign}(\hat{\beta}_S) = \text{sign}(\beta_{0,S}),$$

with high probability. The primal-dual witness method basically plugs in the true support S into the KKT conditions for the lasso (9), (10), and checks whether they can be verified

- We start by breaking up (9) into two blocks, over S and S^c. Suppose that $\text{supp}(\hat{\beta}) = S$ at a solution $\hat{\beta}$. Then the KKT conditions become

$$X_S^T(y - X_S\hat{\beta}_S) = \lambda s_S \tag{21}$$

$$X_{-S}^T(y - X_S\hat{\beta}_S) = \lambda s_{-S}. \tag{22}$$

Hence, if we can satisfy the two conditions (21), (22) with a proper subgradient s, such that

$$s_S = \text{sign}(\beta_{0,S}) \quad \text{and} \quad \|s_{-S}\|_\infty = \max_{i \notin S} |s_i| < 1,$$

then we have met our goal: we have recovered a unique lasso solution whose active set is S, and whose active signs are $\text{sign}(\beta_{0,S})$. (Note that uniqueness follows from the fact that if s satisfies the above properties, then we have an equicorrelation set $E = S$, and $X_E = X_S$ has full column rank by assumption)

So, let's solve for $\hat{\beta}_S$ in the first block (21). Just as we did in the work on basic properties of the lasso estimator, this yields

$$\hat{\beta}_S = (X_S^T X_S)^{-1}\big(X_S^T y - \lambda \text{sign}(\beta_{0,S})\big), \tag{23}$$

where we have substituted $s_S = \text{sign}(\beta_{0,S})$. From (22), this implies that s_{-S} must satisfy

$$s_{-S} = \frac{1}{\lambda} X_{-S}^T\big(I - X_S(X_S^T X_S)^{-1} X_S^T\big)y + X_{-S}^T(X_S^T X_S)^{-1}\text{sign}(\beta_{0,S}). \tag{24}$$

To lay it out, for concreteness, the primal-dual witness method proceeds as follows:

1. Solve for the lasso solution over the S components, $\hat{\beta}_S$, as in (23), and set $\hat{\beta}_{-S} = 0$
2. Solve for the subgradient over the S^c components, s_{-S}, as in (24)
3. Check that $\text{sign}(\hat{\beta}_S) = \text{sign}(\beta_{0,S})$, and that $\|s_{-S}\|_\infty < 1$. If these two checks pass, then we have verified the existence of a unique lasso solution that exactly recovers the true support and signs

The success of the primal-dual witness method hinges on Step 3. Recalling that $y = X\beta_0 + \epsilon$, we can re-express the required conditions, $\text{sign}(\hat{\beta}_S) = \text{sign}(\beta_{0,S})$ and $\|s_{-S}\|_\infty < 1$, as

$$\text{sign}(\beta_{0,i} + \Delta_i) = \text{sign}(\beta_{0,i}), \text{ where}$$

$$\Delta_i = e_i^T(X_S^T X_S)^{-1}\big(X_S^T\epsilon - \lambda\text{sign}(\beta_{0,S})\big), \text{ for all } i \in S, \tag{25}$$

and

$$\Big\|\frac{1}{\lambda}X_{-S}^T\big(I - X_S(X_S^TX_S)^{-1}X_S^T\big)\epsilon + X_{-S}^T(X_S^TX_S)^{-1}\text{sign}(\beta_{0,S})\Big\|_\infty < 1. \tag{26}$$

Recalling that $\epsilon \sim N(0, \sigma^2 I)$, we see that the two required conditions reduce to statements about Gaussian processes, that can be analyzed with concentration of measure arguments. These arguments can actually made rather simple, and will only use tail bounds of Gaussian random variables, but we will need to make assumptions on X and β_0. These are:

- Mutual incoherence: for some $\gamma > 0$, we have

$$\|(X_S^TX_S)^{-1}X_S^TX_j\|_1 \leq 1 - \gamma, \quad \text{for } i \notin S,$$

- Minimum eigenvalue: for some $C > 0$, we have

$$\Lambda_{\min}\Big(\frac{1}{n}X_S^TX_S\Big) \geq C,$$

  where $\Lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix A

- Minimum signal:

$$|\beta_{0,i}| \geq \lambda\Big(\|(X_S^TX_S)^{-1}\|_\infty + \frac{4\sigma}{C}\Big), \quad \text{for } i \in S,$$

  where $\|A\|_\infty = \max_{i=1,\ldots m}\sum_{j=1}^q |A_{ij}|$ denotes the ℓ_∞ norm of an $m \times q$ matrix A

Under these assumptions on X and β_0, and with $\lambda \geq 2\sigma\sqrt{2n\log p}/\gamma$, one can check that (25) and (26) hold with high probability, and hence the primal-dual witness method passes—i.e., the lasso solution is unique and recovers the exact support and signs—with high probability

- The mutual incoherence and minimum eigenvalue conditions are restrictions on the amount of correlation present in the predictor variables. Mutual incoherence says that a variable in S^c cannot be too correlated with a variable in S; the minimum eigenvalue condition says that the variables in S cannot be too correlated with each other. A common theoretical pursuit is to show that these hold with high probability under a *random* choice of X with respect to some model, i.e., a Gaussian model for the entries of X

- The minimum signal condition ensures that the nonzero entries of the true coefficient vector β_0 are big enough to detect. Note that this is a restrictive condition and is not present in the results on the lasso convergence rates in the previous section

- Finally, to be perfectly clear, the primal-dual witness method is not a practical algorithm for finding a solution, because it requires knowledge of the true support and signs; it is instead a theoretical routine that we can study to verify support and sign recovery. Also, it may seem more natural to call it something like "primal-subgradient witness method". However, the subgradient s is related to the solution of the lasso dual problem

$$\hat{u} = \underset{u \in \mathbb{R}^n}{\text{argmin}} \ \|y - u\|_2^2 \ \text{ subject to } \ \|X^Tu\|_\infty \leq \lambda,$$

  in that $X^T\hat{u} = \lambda s$

References

Buhlmann, P. & van de Geer, S. (2011), *Statistics for High-Dimensional Data*, Springer, Berlin.

Candes, E. J. & Plan, Y. (2009), 'Near ideal model selection by ℓ_1 minimization', *Annals of Statistics* **37**(5), 2145–2177.

Candes, E. J. & Tao, T. (2006), 'Near optimal signal recovery from random projections: Universal encoding strategies?', *IEEE Transactions on Information Theory* **52**(12), 5406–5425.

Chen, S., Donoho, D. L. & Saunders, M. (1998), 'Atomic decomposition for basis pursuit', *SIAM Journal on Scientific Computing* **20**(1), 33–61.

Donoho, D. L. (2006), 'Compressed sensing', *IEEE Transactions on Information Theory* **52**(12), 1289–1306.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *Annals of Statistics* **32**(2), 407–499.

Foster, D. & George, E. (1994), 'The risk inflation criterion for multiple regression', *The Annals of Statistics* **22**(4), 1947–1975.

Fuchs, J. J. (2005), 'Recovery of exact sparse representations in the presense of bounded noise', *IEEE Transactions on Information Theory* **51**(10), 3601–3608.

Greenshtein, E. & Ritov, Y. (2004), 'Persistence in high-dimensional linear predictor selection and the virtue of overparametrization', *Bernoulli* **10**(6), 971–988.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer, New York. Second edition.

Hoerl, A. & Kennard, R. (1970), 'Ridge regression: biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Osborne, M., Presnell, B. & Turlach, B. (2000*a*), 'A new approach to variable selection in least squares problems', *IMA Journal of Numerical Analysis* **20**(3), 389–404.

Osborne, M., Presnell, B. & Turlach, B. (2000*b*), 'On the lasso and its dual', *Journal of Computational and Graphical Statistics* **9**(2), 319–337.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B* **58**(1), 267–288.

Tibshirani, R. J. (2013), 'The lasso problem and uniqueness', *Electronic Journal of Statistics* **7**, 1456–1490.

Wainwright, M. J. (2009), 'Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)', *IEEE Transactions on Information Theory* **55**(5), 2183–2202.

Zhao, P. & Yu, B. (2006), 'On model selection consistency of lasso', *Journal of Machine Learning Research* **7**, 2541–2564.