Shrinkage

November 29, 2018

Large *p* problems

More and more statistical datasets have *p* large, sometimes much larger than *n*.

- What variations in genome are associated with disease?

$p > n$

- If we regress our response on all the variables, what is the problem? There is not a unique to the ols problem if $p > n$.

$p > n$

- If we regress our response on all the variables, what is the problem? There is not a unique to the ols problem if $p > n$.

- Even if $p < n$ we will get a lot of variability in our estimates of $\hat{\beta}$ if both p and n are large. Model selection can help – i.e. reduce the number of variables – but the methods we've discussed so far are not going to be very compelling in the face of hundreds of variables, as 2^p can be very large.

Multiple Testing

- With a large number of tests, one for each variable, we are likely to get a great number of false positives. If we have a hundred variables, we get, on average, 5 significant variables, even if there is absolutely no relationship between the variables and the response, and that number grows with the number of variables.

Multiple Testing

- With a large number of tests, one for each variable, we are likely to get a great number of false positives. If we have a hundred variables, we get, on average, 5 significant variables, even if there is absolutely no relationship between the variables and the response, and that number grows with the number of variables.

- If we have a set of p-values (e.g. 1 per variable) that are each individually valid, then there are multiple testing corrections that we can do on the p-values to be able to make joint decisions that as a group will control the number of false positives we make. We will not cover this in detail, other than to note an obvious one: if we make *k* tests, then if each p-value is multiplied by *k* and compared to our level α, the total probability of a single false positive remains $\leq \alpha$. This is called the Bonferroni method.

Example: Marginal Testing

- When there are a large number of variables, some people focus on individual relationships rather than a joint analysis. This is very common in genomics, for example. In this way, they run a separate regression on each variable and test each variable separately.

Example: Marginal Testing

- When there are a large number of variables, some people focus on individual relationships rather than a joint analysis. This is very common in genomics, for example. In this way, they run a separate regression on each variable and test each variable separately.
- This can be done even if $p >> n$. In this setting the joint relationship is less important than finding single variables of interest (e.g. genes). In this case, the number of variables could be in the thousands, while the number of observations in the tens. Often in this case we control for the false positive rate – the proportion of discoveries that are FP, rather than the absolute number. Otherwise we would never find anything.

Shrinkage Methods

- Another approach to model selection/improvement in the face of large number of variables are “shrinkage” methods that adjust the $\hat{\beta}$ by making some of closer to zero.

Shrinkage Methods

- Another approach to model selection/improvement in the face of large number of variables are “shrinkage” methods that adjust the $\hat{\beta}$ by making some of closer to zero.
- First, let’s get all the variables on the same scale by subtracting off the mean and dividing by the standard deviation (per variable). Similarly, let’s center the response (which means we can omit intercept β_0).

Shrinkage Methods

- Another approach to model selection/improvement in the face of large number of variables are “shrinkage” methods that adjust the $\hat{\beta}$ by making some of closer to zero.
- First, let’s get all the variables on the same scale by subtracting off the mean and dividing by the standard deviation (per variable). Similarly, let’s center the response (which means we can omit intercept β_0).
- We have the least squares criteria

$$\min_{\beta} ||y - \mathbf{X}\beta||^2$$

Shrinkage Methods

- Another approach to model selection/improvement in the face of large number of variables are "shrinkage" methods that adjust the $\hat{\beta}$ by making some of closer to zero.
- First, let's get all the variables on the same scale by subtracting off the mean and dividing by the standard deviation (per variable). Similarly, let's center the response (which means we can omit intercept β_0).
- We have the least squares criteria

$$\min_{\beta} ||y - \mathbf{X}\beta||^2$$

- If we want to limit the contributions of variable $\boldsymbol{X}_j$ to the model, we can think of wanting $|\hat{\beta}_j|$ to be small. We don't know which variables we want to limit, though, so we want to write down some condition that is global on the vector β and then algorithmically let the data tell me which variables should get to contribute the most.

- One can think about these methods as smoother versions of variable selection that don’t require 0/1 choices or as much user choices.

- One can think about these methods as smoother versions of variable selection that don't require 0/1 choices or as much user choices.
- Mathematically, we can write this as

$$\min_{\mathcal{S}(\boldsymbol{\beta})\leq c} ||y - \mathbf{X}\beta||^2$$

We can make different choices as to what is the allowable 'size' of β, e.g.

$$\mathcal{S}(\beta) = \sum_j \beta_j^2, \text{ or }, \mathcal{S}(\beta) = \sum_j |\beta_j|$$

- One can think about these methods as smoother versions of variable selection that don't require 0/1 choices or as much user choices.
- Mathematically, we can write this as

  $$\min_{\mathcal{S}(\boldsymbol{\beta})\leq c} ||y - \mathbf{X}\beta||^2$$

  We can make different choices as to what is the allowable 'size' of β, e.g.

  $$\mathcal{S}(\beta) = \sum_j \beta_j^2, \text{ or } , \mathcal{S}(\beta) = \sum_j |\beta_j|$$

- As $c \rightarrow \infty$ we are putting on less constraint, so we get closer to the standard least squares model. Another way we can formulate this problem

  $$\min_{\beta} ||y - \mathbf{X}\beta||^2 + \lambda\mathcal{S}(\beta)$$

- In this way, we are adding a penalty for the size of $\boldsymbol{\beta}$. λ controls how much weight we assign to minimizing the coefficient magnitude versus minimizing the error.

- In this way, we are adding a penalty for the size of $\boldsymbol{\beta}$. λ controls how much weight we assign to minimizing the coefficient magnitude versus minimizing the error.
- There is a one-to-one relationship between *c* and λ so in theory these are equivalent.

- In this way, we are adding a penalty for the size of $\boldsymbol{\beta}$. λ controls how much weight we assign to minimizing the coefficient magnitude versus minimizing the error.
- There is a one-to-one relationship between *c* and λ so in theory these are equivalent.
- Note that we are NOT specifying a specific model (i.e. choice of variables). We are only controlling how much contribution they can make to the prediction.

- In this way, we are adding a penalty for the size of $\boldsymbol{\beta}$. λ controls how much weight we assign to minimizing the coefficient magnitude versus minimizing the error.
- There is a one-to-one relationship between *c* and λ so in theory these are equivalent.
- Note that we are NOT specifying a specific model (i.e. choice of variables). We are only controlling how much contribution they can make to the prediction.
- **Relationship to Model Subsets** If we wanted to make 0/1 choices about β, then we could write $\mathcal{S}(\beta) = \sum_j I(\beta_j > 0)$ and $\mathcal{S}(\beta) \leq c$ would limit us to *c* variables in our model. This is clearly a choice of $\mathcal{S}$ that will be untractable.

- In this way, we are adding a penalty for the size of $\boldsymbol{\beta}$. λ controls how much weight we assign to minimizing the coefficient magnitude versus minimizing the error.
- There is a one-to-one relationship between c and λ so in theory these are equivalent.
- Note that we are NOT specifying a specific model (i.e. choice of variables). We are only controlling how much contribution they can make to the prediction.
- **Relationship to Model Subsets** If we wanted to make 0/1 choices about β, then we could write $\mathcal{S}(\beta) = \sum_j I(\beta_j > 0)$ and $\mathcal{S}(\beta) \leq c$ would limit us to c variables in our model. This is clearly a choice of $\mathcal{S}$ that will be untractable.
- Before we saw model selection criteria as the error of a model (RSS) plus a penalty that depended on $p(m)$ of the model. This makes sense if we consider our 0/1 size evaluation. But now, we can allow for smoother versions.

Solving the penalized criteria

- The smoother size constraints like those above *are* tractable for a fixed λ. They can be solved for hundreds or thousands of variables.

Solving the penalized criteria

- The smoother size constraints like those above *are* tractable for a fixed λ. They can be solved for hundreds or thousands of variables.
- The reason is that for those constraints, this is a convex problem – no local optimum – for which there are algorithms that efficiently search the space of β and can guarantee that they will find the global optimum.

Solving the penalized criteria

- The smoother size constraints like those above *are* tractable for a fixed λ. They can be solved for hundreds or thousands of variables.
- The reason is that for those constraints, this is a convex problem – no local optimum – for which there are algorithms that efficiently search the space of β and can guarantee that they will find the global optimum.
- $p > n$

  By putting an appropriate constraint on the space of β we consider, we now can solve this even when $p > n$.

Ridge Regression

- Choosing $\mathcal{S}(\beta) = \sum_j \beta_j^2$ is called **Ridge Regression**,

$$\min_{\beta} \; ||y - \mathbf{X}\beta||^2 + \lambda||\beta||_2^2.$$

Ridge Regression

- Choosing $\mathcal{S}(\beta) = \sum_j \beta_j^2$ is called **Ridge Regression**,

$$\min_{\beta} \; ||y - \mathbf{X}\beta||^2 + \lambda||\beta||_2^2.$$

- For this penalty, we can find a closed-form solution for $\hat{\beta}_{RR}$,

$$\hat{\beta}_{RR} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}^T y$$

Ridge Regression

- Choosing $\mathcal{S}(\beta) = \sum_j \beta_j^2$ is called **Ridge Regression**,

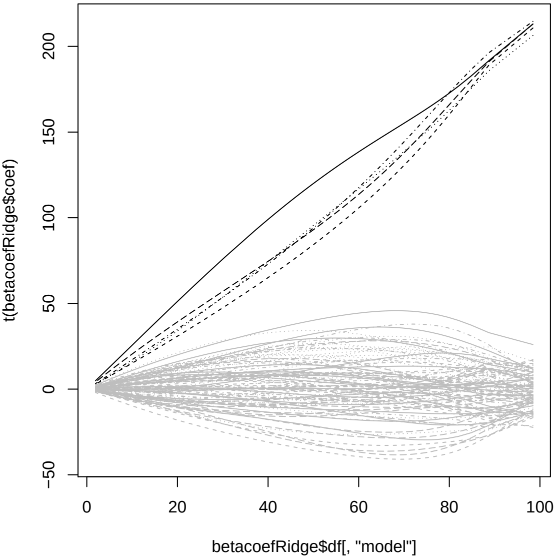
$$\min_{\beta} \; ||y - \mathbf{X}\beta||^2 + \lambda||\beta||_2^2.$$

- For this penalty, we can find a closed-form solution for $\hat{\beta}_{RR}$,

$$\hat{\beta}_{RR} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T y$$

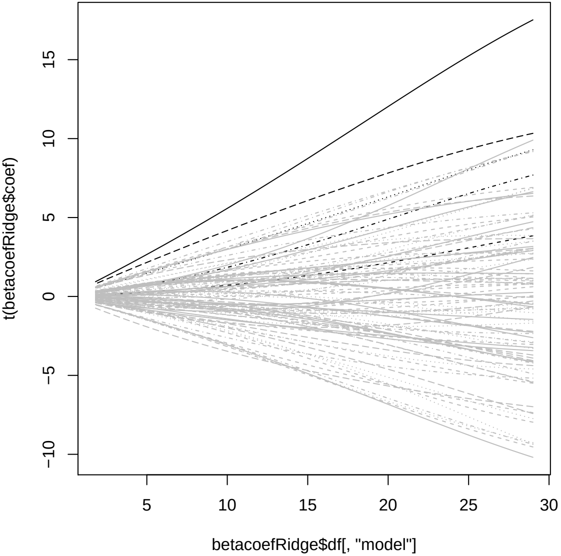
- Notice that $\hat{\beta}_{RR}$ is biased,

$$E(\hat{\beta}_{RR}) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\beta$$

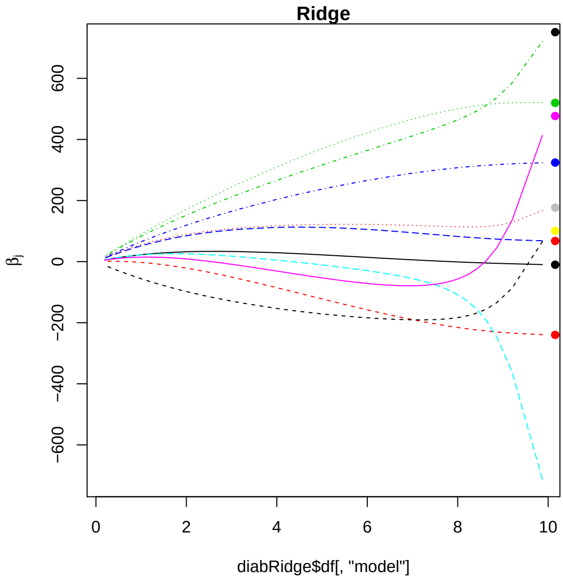

t(betacoefRidge$coef)
200
150
100
50
0
−50
0
20
40
60
80
100
betacoefRidge$df[, "model"]

- The lines separated from the rest are the five first variables. What is happening here? How what is the penalty doing

- The lines separated from the rest are the five first variables. What is happening here? How what is the penalty doing
- You can fit it with less observations than variables.



Do model selection after this, like cross-validation


Ridge
600
400
200
0
−200
−400
−600
β_j
0
2
4
6
8
10
diabRidge$df[, "model"]

- Note that our prediction of y is still linear combination of the y,

  $$y = \mathbf{H}_\lambda y = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}^T y$$

  Instead of calling it the hat matrix, it is usually called the smoothing matrix, it is no longer a projection matrix.

- Note that our prediction of y is still linear combination of the y,

$$y = \mathbf{H}_\lambda y = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T y$$

  Instead of calling it the hat matrix, it is usually called the smoothing matrix, it is no longer a projection matrix.

- **Stabilizing** $(\mathbf{X}^T\mathbf{X})^{-1}$ Notice that even if $(\mathbf{X}^T\mathbf{X})^{-1}$ is not invertible (e.g. $p > n$),

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$$

  is always invertible if $\lambda > 0$. Why?

- Choosing $\mathcal{S}(\beta) = \sum_j |\beta_j|$ is called **Lasso** (Least Absolute Shrinkage and Selection Operator). Our measure of size $(\mathcal{S})$ is also a norm on $\mathbb{R}^p$ and is called the L_1 norm,

$$\min_{\beta} ||\boldsymbol{y} - \mathbf{X}\beta||^2 + \lambda||\beta||_1.$$

- Choosing $\mathcal{S}(\beta) = \sum_j |\beta_j|$ is called **Lasso** (Least Absolute Shrinkage and Selection Operator). Our measure of size ($\mathcal{S}$) is also a norm on $\mathbb{R}^p$ and is called the L_1 norm,

$$\min_{\beta} ||\boldsymbol{y} - \mathbf{X}\beta||^2 + \lambda||\beta||_1.$$

- Lasso is particularly popular for model selection because it tends to zero-out values of β_j rather than just make them small. This has the effect of defining a subset model.

- Choosing $\mathcal{S}(\beta) = \sum_j |\beta_j|$ is called **Lasso** (Least Absolute Shrinkage and Selection Operator). Our measure of size ($\mathcal{S}$) is also a norm on $\mathbb{R}^p$ and is called the L_1 norm,

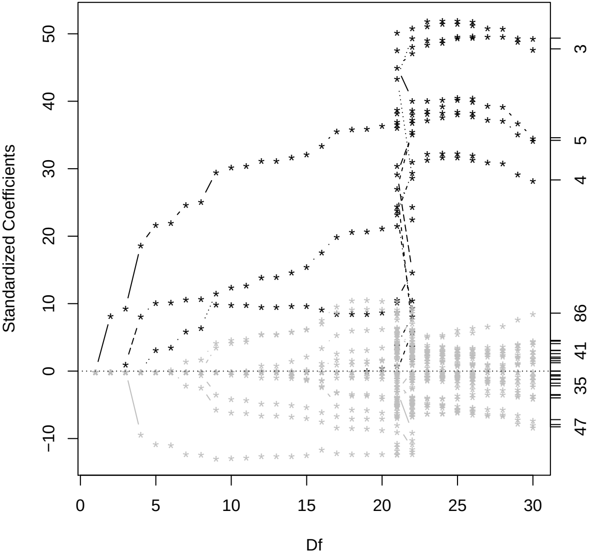
$$\min_{\beta} ||\boldsymbol{y} - \mathbf{X}\beta||^2 + \lambda||\beta||_1.$$

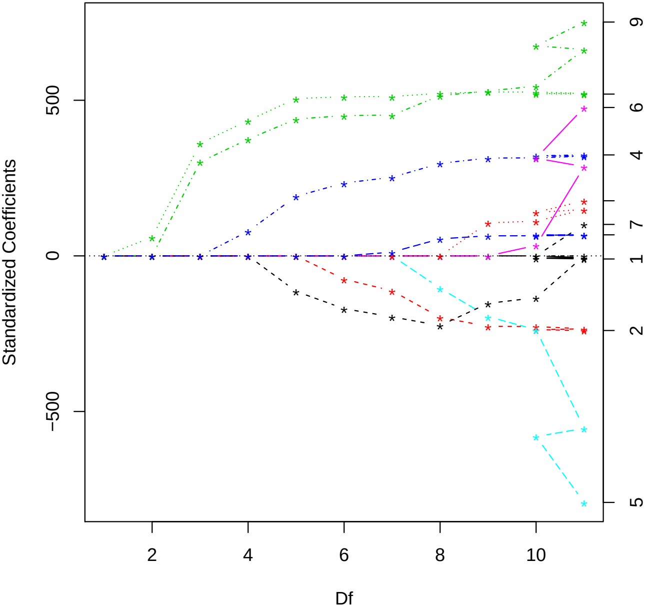
- Lasso is particularly popular for model selection because it tends to zero-out values of β_j rather than just make them small. This has the effect of defining a subset model.
- **Subset Selection** There is a lot of theoretical work on Lasso, and Lasso (i.e. L_1 penalties) actually can be shown to often mimic all subset selection.

- Choosing $\mathcal{S}(\beta) = \sum_j |\beta_j|$ is called **Lasso** (Least Absolute Shrinkage and Selection Operator). Our measure of size ($\mathcal{S}$) is also a norm on $\mathbb{R}^p$ and is called the L_1 norm,

$$\min_{\beta} ||\boldsymbol{y} - \mathbf{X}\beta||^2 + \lambda||\beta||_1.$$

- Lasso is particularly popular for model selection because it tends to zero-out values of β_j rather than just make them small. This has the effect of defining a subset model.
- **Subset Selection** There is a lot of theoretical work on Lasso, and Lasso (i.e. L_1 penalties) actually can be shown to often mimic all subset selection.
- And you can have less observations than variable.


Standardized Coefficients
Df
−10
0
10
20
30
40
50
0
5
10
15
20
25
30
3
5
4
86
41
35
47



Standardized Coefficients
500
0
−500
2
4
6
8
10
Df
9
6
4
7
1
2
5

Computation

- There is no closed-form solution to $\hat{\beta}$. Here are some facts for a fixed design matrix **X** and tuning parameter $\lambda \geq 0$:

Computation

- There is no closed-form solution to $\hat{\beta}$. Here are some facts for a fixed design matrix **X** and tuning parameter $\lambda \geq 0$:
- There is either a unique lasso solution or an (uncountably) infinite number of solutions.

Computation

- There is no closed-form solution to $\hat{\beta}$. Here are some facts for a fixed design matrix **X** and tuning parameter $\lambda \geq 0$:
- There is either a unique lasso solution or an (uncountably) infinite number of solutions.
- Every lasso solution $\hat{\beta}$ gives the same fitted value $\mathbf{X}\hat{\beta}$.

Computation

- There is no closed-form solution to $\hat{\beta}$. Here are some facts for a fixed design matrix $\mathbf{X}$ and tuning parameter $\lambda \geq 0$:
- There is either a unique lasso solution or an (uncountably) infinite number of solutions.
- Every lasso solution $\hat{\beta}$ gives the same fitted value $\mathbf{X}\hat{\beta}$.
- If $\lambda > 0$, then every solution $\hat{\beta}$, has the same ℓ_1 norm, $\|\hat{\beta}\|_1$.

Computation

- There is no closed-form solution to $\hat{\beta}$. Here are some facts for a fixed design matrix $\mathbf{X}$ and tuning parameter $\lambda \geq 0$:
- There is either a unique lasso solution or an (uncountably) infinite number of solutions.
- Every lasso solution $\hat{\beta}$ gives the same fitted value $\mathbf{X}\hat{\beta}$.
- If $\lambda > 0$, then every solution $\hat{\beta}$, has the same ℓ_1 norm, $\|\hat{\beta}\|_1$.
- Lasso is a nonlinear estimator, meaning we can't express $\hat{y} = \mathbf{A}y$.

- How would you compare lasso and ridge regression?

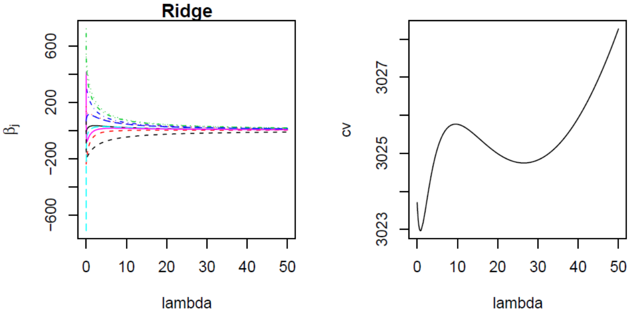
- How would you compare lasso and ridge regression?
- Lasso is setting some coefficients to zero, ridge is just decreasing them.

Choosing λ

- There are ways to do Mallow's cp or BIC model selection for LASSO, but this require using an estimate the number of parameters.

Choosing λ

- There are ways to do Mallow's cp or BIC model selection for LASSO, but this require using an estimate the number of parameters.
- Cross-Validation can also be used. We have parameterized our problem so that for each λ we have a model defined by $\hat{\boldsymbol{\beta}}_\lambda$ so we just need to pick between λ values. Specifically, we can choose a grid of λ values, and for each λ value do cross-validation.


Ridge
600
200
−200
−600
βj
0
10
20
30
40
50
lambda
3027
3025
3023
cv
0
10
20
30
40
50
lambda