MA 575 Linear Models:

Cedric E. Ginestet, Boston University

Regularization: Ridge Regression and Lasso

Week 14, Lecture 2

1 Ridge Regression

Ridge regression and the Lasso are two forms of **regularized** regression. These methods are seeking to alleviate the consequences of multicollinearity.

1. When variables are *highly correlated*, a large coefficient in one variable may be alleviated by a large coefficient in another variable, which is negatively correlated to the former.
2. *Regularization* imposes an upper threshold on the values taken by the coefficients, thereby producing a more parsimonious solution, and a set of coefficients with smaller variance.

1.1 Convex Optimization

Ridge regression is motivated by a **constrained minimization** problem, which can be formulated as follows,

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{\text{ridge}} &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2, \\ &\text{subject to } \sum_{j=1}^{p}\beta_j^2 \leq t, \end{aligned} \tag{1}$$

for $t \geq 0$. The **feasible set** for this minimization problem is therefore constrained to be

$$S(t) := \big\{\boldsymbol{\beta} \in \mathbb{R}^p : ||\boldsymbol{\beta}||_2^2 \leq t\big\},$$

where $\boldsymbol{\beta}$ does not include the intercept β_0. The ridge estimator are not **equivariant** under a re-scaling of the $\mathbf{x}_j$'s, because of the L_2-penalty. This difficulty is circumvented by *centering* the predictors. In this lecture, therefore, the design matrix $\mathbf{X}$ will, in fact, stand for the *centered matrix*, $\mathcal{X}$. In addition, we exclude the intercept, β_0, from the penalty term; as otherwise, a change in the centering of the y_i's would result in a change in the choice of λ.

The use of an L_2-penalty in least-squares problem is sometimes referred to as **Tikhonov regularization**. Using a **Lagrange multiplier**, this can be alternatively formulated as

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 \;+\; \lambda\sum_{j=1}^{p}\beta_j^2 \right\}, \tag{2}$$

for $\lambda \geq 0$; and where there is a *one-to-one correspondence* between t in equation (1) and λ in equation (2).

A constrained optimization problem is said to be a **convex optimization**, if both the objective function and the constraints are convex functions. In our case, this particular constrained optimization problem can be seen to be a *convex minimization* in $\boldsymbol{\beta}$. Indeed, we have already verified that $\mathrm{RSS}(\boldsymbol{\beta})$ is convex in $\boldsymbol{\beta}$, provided that $\mathbf{X}^T\mathbf{X}$ is *positive definite*, which follows from the fact that $\mathbf{X}$ is assumed to be *full-rank*. Moreover, one can also show that the sum of squared β_j's is also convex. Using matrix notation, the constraint on $\boldsymbol{\beta}$ may be expressed as the function,

$$f(\boldsymbol{\beta}) := \boldsymbol{\beta}^T\boldsymbol{\beta}.$$

As described in lecture 4.2., recall that the formula for the differentiation of a **quadratic form** states that for any square matrix $\mathbf{A}$,

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x} = \left(\mathbf{A} + \mathbf{A}^T\right)\mathbf{x}.$$

In our case, we have $\boldsymbol{\beta}^T\mathbf{I}\boldsymbol{\beta}$, and therefore

$$\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{\beta}^T\mathbf{I}\boldsymbol{\beta} = (\mathbf{I}+\mathbf{I})\boldsymbol{\beta} = 2\boldsymbol{\beta}.$$

Next, taking the second derivative, we obtain $\partial^2/\partial\boldsymbol{\beta}^2(\boldsymbol{\beta}^T\mathbf{I}\boldsymbol{\beta}) = 2 > 0$, which proves that $f(\boldsymbol{\beta})$ is **strictly convex** in $\boldsymbol{\beta}$.

1.2 Analytical Minimization

The criterion to be minimized in equation (2) can be reformulated using matrix algebra in order to obtain a *closed-form* solution. The RSS for ridge regression is expressed as

$$\mathrm{RSS}(\boldsymbol{\beta};\lambda) := (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}.$$

One can minimize this criterion using straightforward applications of matrix calculus, as was conducted for the classical OLS criterion for multiple regression, and described in lecture 4.2. That is, setting to zero and taking the first derivative, we obtain

$$\frac{\partial}{\partial \boldsymbol{\beta}}\mathrm{RSS}(\boldsymbol{\beta};\lambda) = 2(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} - 2\mathbf{X}^T\mathbf{y} + 2\lambda\boldsymbol{\beta} = 0.$$

This expression can be further simplified as follows,

$$\begin{aligned} 2(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} &= 2\mathbf{X}^T\mathbf{y} \\ (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} &= \mathbf{X}^T\mathbf{y}, \end{aligned}$$

and therefore the ridge estimators are

$$\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.$$

Since we are adding a **positive constant** to the diagonal of $\mathbf{X}^T\mathbf{X}$, we are, in general, producing an invertible matrix, $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$, even if $\mathbf{X}^T\mathbf{X}$ is *singular*. Historically, this particular aspect of ridge regression was the main motivation behind the adoption of this particular extension of OLS theory. In addition, this also shows that $\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}}$ is still a **linear** function of the observed values, $\mathbf{y}$.

The ridge regression estimator is related to the classical OLS estimator, $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$, in the following manner,

$$\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}} = [\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}},$$

assuming that $\mathbf{X}^T\mathbf{X}$ is non-singular. This relationship can be verified by simply applying the definition of $\widehat{\boldsymbol{\beta}}^{\text{OLS}}$,

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = [\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.$$

using the fact $\mathbf{B}^{-1}\mathbf{A}^{-1} = (\mathbf{AB})^{-1}$.

Moreover, when $\mathbf{X}$ is composed of orthonormal variables, such that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, it then follows that

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{I} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = ((1+\lambda)\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{1+\lambda}\widehat{\boldsymbol{\beta}}.$$

This is instructive, as it shows that, in this simple case, the ridge estimator is simply a **downweighted** version of the OLS estimator.

1.3 Ridge Regression as Perturbation

Assume that we can find an $n \times p$ matrix $\mathbf{A}$, whose columns are **orthogonal** to the ones of $\mathbf{X}$ and to $\mathbf{y}$, such that we have

$$\mathbf{A}^T\mathbf{X} = \mathbf{0}, \qquad \text{and} \qquad \mathbf{A}^T\mathbf{y} = \mathbf{0};$$

where the former is a $p \times p$ matrix of zeros, whereas the latter is a $p \times 1$ vector of zeros. We construct the following matrix

$$\mathbf{W}^T := \lambda^{1/2}(\mathbf{A}^T\mathbf{A})^{-1/2}\mathbf{A}^T.$$

It immediately follows that we also have $\mathbf{W}^T\mathbf{X} = \mathbf{0}$ and $\mathbf{W}^T\mathbf{y} = \mathbf{0}$, as when pre-multiplying by $\mathbf{A}^T$. Next, we can compute the OLS estimator for a criterion based on the perturbed version of the design matrix, such that the mean function for this model is now,

$$\mathbb{E}[\mathbf{y}|\mathbf{X}, \mathbf{W}] = (\mathbf{X} + \mathbf{W})\boldsymbol{\beta}.$$

The OLS estimator, denoted $\widehat{\boldsymbol{\beta}}$, can be simplified in the following fashion,

$$\widehat{\boldsymbol{\beta}} = [(\mathbf{X} + \mathbf{W})^T(\mathbf{X} + \mathbf{W})]^{-1}(\mathbf{X} + \mathbf{W})^T\mathbf{y} = \widehat{\boldsymbol{\beta}}^{\text{ridge}},$$

since $\mathbf{W}^T\mathbf{X} = \mathbf{0}$ and $\mathbf{W}^T\mathbf{y} = \mathbf{0}$, and by definition,

$$\mathbf{W}^T\mathbf{W} = \lambda\,(\mathbf{A}^T\mathbf{A})^{-1/2}\mathbf{A}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1/2} = \lambda^2\mathbf{I},$$

since every positive semi-definite matrix has a *unique square root*, and we know that $\mathbf{A}$ is positive semi-definite because it is a Gram matrix. This particular derivation of ridge regression highlights a particular geometric interpretation of its effect. By adding perturbation to $\mathbf{X}$ in directions, which are orthogonal to the columns of $\mathbf{X}$, we are 'enlarging the strip' on which the data is situated.

1.4 Effective Number of Parameters

The number of degrees of freedom in standard OLS regression was obtained by taking the trace of $\mathbf{I}-\mathbf{H}$. This can be also done in ridge regression. However, we will require to study more closely the spectral properties of $\mathbf{X}$. The design matrix can be expressed using a **singular value decomposition (SVD)**, such that

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices of order $n \times p$ and $p \times p$, respectively; and $\mathbf{D}$ is a diagonal matrix of order $p \times p$ containing the **singular values** of $\mathbf{X}$. The columns of $\mathbf{U}$, which are called the **left singular** vectors, span the *column space* of $\mathbf{X}$; whereas the columns of $\mathbf{V}$, which are called the **right singular** vectors, span the *row space* of $\mathbf{X}$. In particular, $\mathbf{U}$ is a set of eigenvectors for $\mathbf{X}\mathbf{X}^T$, and $\mathbf{V}$ is a set of eigenvectors for $\mathbf{X}^T\mathbf{X}$. The *non-zero singular values* of $\mathbf{X}$ are the square roots of the eigenvalues of both $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$.

Now, we can use the SVD of $\mathbf{X}$ for unveiling the properties of the hat matrix obtained, when performing ridge regression,

$$\mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

where we will define

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

Using the SVD of $\mathbf{X}$, observe that the Gram matrix $\mathbf{X}^T\mathbf{X}$ can be decomposed as follows,

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}^2\mathbf{V}^T,$$

which is the *eigendecomposition* of $\mathbf{X}^T\mathbf{X}$. Next, we can apply this to the hat matrix,

$$\begin{aligned}\mathbf{H}_\lambda &= \mathbf{U}\mathbf{D}\mathbf{V}^T(\mathbf{V}\mathbf{D}^2\mathbf{V}^T + \lambda\mathbf{I})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T \\ &= \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^T\mathbf{V}\mathbf{D}\mathbf{U}^T \\ &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T.\end{aligned}$$

since $\mathbf{V}\mathbf{D}^2\mathbf{V}$ and $\lambda\mathbf{I}$ commute and are hence **simultaneously diagonalizable**. Therefore, it also follows that $\mathbf{H}_\lambda$ is diagonalizable, with respect to $\mathbf{U}$, and with eigenvalues given by $\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}$, which is a diagonal matrix of order $p \times p$. Thus, since the trace a matrix is equal to the sum of its eigenvalues, it readily follows that

$$\operatorname{tr}(\mathbf{H}_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda},$$

where d_j is the j^{th} diagonal entry of $\mathbf{D}$. Thus, the diagonal entries of $\mathbf{H}_\lambda$ are the re-scaled eigenvalues of $\mathbf{X}^T\mathbf{X}$. Observe that as $\lambda \to 0$, we recover $\operatorname{tr}(\mathbf{H}_\lambda) = p$ and no regularization is performed. By contrast, if $\lambda \to \infty$, the *effective number* of parameters is shrunk to zero. Thus, regularization leads to a reduction in the effective number of parameters.

1.5 Bias and Variance of Ridge Estimator

Ridge estimation produces a biased estimator of the true parameter $\boldsymbol{\beta}$. Using the definition of $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$ and our modeling assumption on the mean function $\mathbb{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$, we obtain,

$$\begin{aligned}\mathbb{E}[\widehat{\boldsymbol{\beta}}^{\text{ridge}}|\mathbf{X}] &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} - \lambda\mathbf{I})\boldsymbol{\beta} \\ &= [\mathbf{I} - \lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}]\boldsymbol{\beta} \\ &= \boldsymbol{\beta} - \lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}.\end{aligned}$$

As one may have expected, the bias of the ridge estimator is proportional to λ. That is, the larger is λ, the larger is the bias of the ridge estimator with respect to $\boldsymbol{\beta}$.

Although the vector of ridge estimators incur a greater bias, it possesses a smaller variance than the vector of OLS estimators. One may compare these two quantities by taking the trace of the variance matrices of the two methods. The trace of the variance matrix of a random vector is sometimes referred to as the **total variance**. For the OLS, using the SVD decomposition presented earlier, we have

$$\operatorname{tr}(\mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}^{\text{OLS}}|\mathbf{X}]) = \operatorname{tr}(\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{d_j^2}.$$

By contrast, for the vector of ridge estimators, we obtain

$$\mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}^{\text{ridge}}|\mathbf{X}] = \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}.$$

Here, note that since $\mathbf{X}^T\mathbf{X}$ and $\lambda\mathbf{I}$ are *simultaneously diagonalizable*, we have as before,

$$(\mathbf{V}\mathbf{D}^2\mathbf{V}^T + \lambda\mathbf{I})^{-1} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^T.$$

Thus, it suffices to apply this formula twice in the above expression in order to produce

$$\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^T\mathbf{V}\mathbf{D}^2\mathbf{V}^T\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^T,$$

which simplifies to

$$\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}^2(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^T.$$

This is a diagonalizable matrix, and therefore we can simply take the trace of this quantity as the sum of the eigenvalues, such that

$$\text{tr}(\mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}^{\text{ridge}}|\mathbf{X}]) = \sigma^2\sum_{j=1}^{p}\frac{d_j^2}{(d_j^2 + \lambda)^2}.$$

It is then an exercise to verify that the ridge estimator has indeed systematically less *total variance* than the one of the OLS estimator,

$$\text{tr}(\mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}^{\text{OLS}}|\mathbf{X}]) \geq \text{tr}(\mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}^{\text{ridge}}|\mathbf{X}]).$$

2 Lasso Regression

We have seen that ridge regression essentially re-scales the OLS estimates. The lasso, by contrast, tries to produce a **sparse** solution, in the sense that several of the slope parameters will be set to zero. One may therefore refers to ridge regression as *soft* thresholding, whereas the lasso is *soft/hard*, and subset selection is a *hard* thresholding; since, in the latter, only a subset of the variables are included in the final model.

2.1 Constrained Minimization

As for ridge regression, the Lasso is also formulated with respect to the centered matrix, $\mathcal{X}$, here denoted $\mathbf{X}$. Moreover, the L_1-penalty is solely applied to the slope coefficients, and thus the intercept, β_0, is excluded from the penalty term. As for ridge regression, the Lasso can be expressed as a **constrained minimization** problem,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{lasso}} &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2, \\ \text{subject to } &\sum_{j=1}^{p}|\beta_j| \leq t,\end{aligned} \tag{3}$$

for $t \geq 0$; which can again be re-formulated using the Lagrangian for the L_1-penalty, as follows,

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 \;+\; \lambda\sum_{j=1}^{p}|\beta_j|\right\},$$

where $\lambda \geq 0$ and, as before, there exists a one-to-one correspondence between t and λ.

2.2 Parameter Estimation

Contrary to ridge regression, the Lasso does not admit a closed-form solution. The L_1-penalty makes the solution *non-linear* in the y_i's. The above constrained minimization is a **quadratic programming** problem, whose solution can be efficiently approximated.

3 Choice of Hyperparameters

3.1 Regularization Parameter

The choice of λ in both ridge regression and in the lasso is more of an art than a science. This parameter can be construed as a **complexity parameter**, since as λ increases, less and less effective parameters are likely to be included in both ridge and lasso regression. Therefore, one can adopt a model selection perspective and compare different choices of λ using cross-validation or an information criterion. That is, the value of λ should be chosen adaptively, in order to minimize an estimate of the **expected prediction error**, for instance, which is well approximated by the AIC.

3.2 Bayesian Perspectives

The penalty terms in ridge and lasso regression can also be justified, using a Bayesian framework, whereby these terms arise as a result of the specification of a particular prior distribution on the vector of slope parameters.

1. The use of an L_2-penalty in multiple regression is analogous to the choice of a *Normal prior* on the β_j's, in Bayesian statistics.

$$\begin{aligned} y_i &\overset{\text{ind}}{\sim} N(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2), \qquad i = 1, \ldots, n \\ \beta_j &\overset{\text{iid}}{\sim} N(0, \tau^2), \qquad j = 1, \ldots, p. \end{aligned}$$

2. Similarly, the use of an L_1-penalty in multiple regression is analogous to the choice of a *Laplace prior* on the β_j's, such that

$$\beta_j \overset{\text{iid}}{\sim} \text{Laplace}(0, \tau^2), \qquad \forall\ j = 1, \ldots, p.$$

In both cases, the value of the **hyperparameter**, τ^2, will be *inversely proportional* to the choice of a particular value for λ in both ridge and lasso regression. For ridge regression, λ is exactly equal to the shrinkage parameter of this hierarchical model, σ^2/τ^2, such that

$$\lambda = \frac{\sigma^2}{\tau^2}.$$

That is, when the β_j's have a large variance, we obtain a small λ. As with other problems in Bayesian statistics, uncertainty about this particular choice can be formally evaluated by specifying a (hyper)prior distribution on τ^2.