22 Lecture 22: March 17

Last time

- Collinearity (JF chapter 13, RD 8.3.2)
- Principal component analysis (JF 13.1.1, RD 8.3.4)

Today

- Midterm exam review
- Biased estimation (JF 13.2.3, CG's notes)
  - Ridge regression
  - Lasso regression

Additional reference

Lecture notes by Cedric Ginestet

Ridge Regression

Ridge regression and the Lasso regression are two forms of regularized regression. These methods can be used 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.

Constrained optimization

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

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2$$

$$\text{subject to } ||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2 \leqslant t$$

for $t \geqslant 0$.

Use a Lagrange multiplier, we can rewrite the formula as

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \}$$

for $\lambda \geqslant 0$ and where there is a one-to-one correspondence between t and λ. λ is an arbitrary constant usually referred to as the "ridge constant".

Analytical solutions

The ridge-regression estimator has analytical solution

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$$

This is obtained by differentiating the objective function with respect to β and set it to 0:

$$\begin{aligned}
&\frac{\partial}{\partial\beta}\{(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^T\beta\} \\
=&2(\mathbf{X}^T\mathbf{X})\beta - 2\mathbf{X}^T\mathbf{Y} + 2\lambda\beta \\
=&0
\end{aligned}$$

Therefore,

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta = \mathbf{X}^T\mathbf{Y}$$

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.

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

$$\hat{\beta}^{ridge} = \left[\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}\right]^{-1}\hat{\beta}^{OLS},$$

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

$$\begin{aligned}
\hat{\beta}^{ridge} &= \left[\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}\right]^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}
\end{aligned}$$

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

$$\hat{\beta}^{ridge} = \frac{1}{1+\lambda}\hat{\beta}^{OLS}$$

Bias and variance of ridge estimator

Ridge estimation produces a biased estimator of the true parameter β. With the definition of $\hat{\beta}^{ridge}$ and the model assumption $\mathbf{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$, we obtain,

$$\begin{aligned}
\mathbf{E}\left(\hat{\beta}^{ridge}|\mathbf{X}\right) &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\beta \\
&= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} - \lambda\mathbf{I})\beta \\
&= \beta - \lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\beta
\end{aligned}$$

where the bias of the ridge estimator is proportional to λ. The variance of the ridge estimator is

$$\mathbf{Var}\left(\hat{\beta}^{ridge}|\mathbf{X}\right) = \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}.$$

When λ increases, the inverted term $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}$ is increasingly dominated by $\lambda\mathbf{I}$. The variance of the ridge estimator, therefore, is a decreasing function of λ. This result is intuitively reasonable because the estimator itself is driven toward $\mathbf{0}$.

Variance-bias tradeoff

The mean-squared error of an estimator can be decomposed into the sum of its squared bias and sampling variance.

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbf{E}\left((\hat{\theta} - \theta)^2\right) = \mathbf{E}(\hat{\theta}^2) + \theta^2 - 2\theta\mathbf{E}(\hat{\theta}) \\ Bias^2(\hat{\theta}) &= \left[\mathbf{E}(\hat{\theta}) - \theta\right]^2 = \mathbf{E}^2(\hat{\theta}) + \theta^2 - 2\theta\mathbf{E}(\hat{\theta}) \\ \mathrm{Var}(\hat{\theta}) &= \mathbf{E}(\hat{\theta}^2) - \mathbf{E}^2(\hat{\theta}) \end{aligned}$$

Therefore

$$MSE(\hat{\theta}) = Bias^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta})$$

The essential idea here is to trade a small amount of bias in the coefficient estimates for a large reduction in coefficient sampling variance. Hoerl and Kennard (1970) prove that it is always possible to choose a positive value of the ridge constant λ so that the mean-squared error of the ridge estimator is less than the mean-squared error of the least-squares estimator. These ideas are illustrated heuristically in Figure 22.1

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.

Constrained optimization

Different from the L_2 penalty for ridge regression, the Lasso regression employs L_1-penalty.

$$\begin{aligned} \hat{\beta}^{lasso} &= \underset{\beta\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2 \\ \text{subject to } ||\beta||_1 &= \sum_{j=1}^{p}|\beta_j| \leqslant t \end{aligned}$$

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

$$\hat{\beta}^{lasso} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\}$$

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

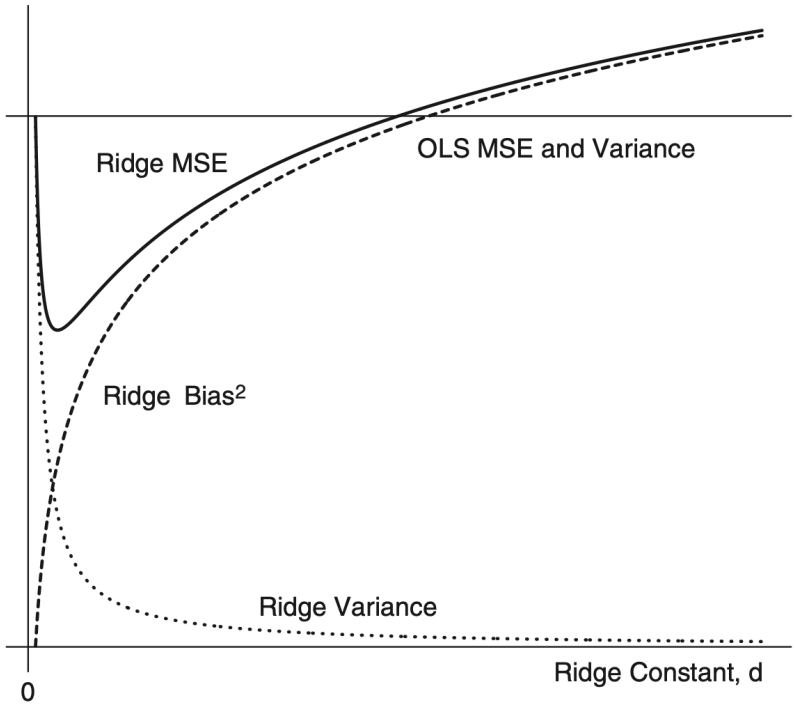

Figure 22.1: Trade-off of bias and against variance for the ridge-regression estimator. The horizontal line gives the variance of the least-squares (OLS) estimator; because the OLS estimator is unbiased, its variance and mean-squared error are the same. The broken line shows the squared bias of the ridge estimator as an increasing function of the ridge constant d (i.e. λ in our notes). The dotted line shows the variance of the ridge estimator. The mean-squared error (MSE) of the ridge estimator, given by the heavier solid line, is the sum of its variance and squared bias. For some values of d, the MSE error of the ridge estimator is below the variance of the OLS estimator. JF Figure 13.9.

Parameter estimation

Contrary to ridge regression, the Lasso does not have a closed-form solution. The L_1-penalty makes the solution non-linear in y_i's. The above constrained minimization is a quadratic programming problem, for which many solvers exist.

Choice of Hyperparameters

Regularization parameter

The choice of λ in both ridge and lasso regressions is more of an art than a science. This parameter can be constructed as a complexity parameter, since as λ increases, less and less effective parameters are likely to be included in both ridge and lasso regressions. 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 (as in cross-validation), for

instance, which is well approximated by AIC. We will discuss model selection in more detail later.

Bayesian perspective

The penalty terms in ridge and lasso regression can also be justified, using a Bayesian framework, whereby these terms arise as aresult 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.

$$y_i \overset{iid}{\sim} \mathcal{N}(\beta_0 + \mathbf{x}_i^T\beta, \sigma^2), \quad i = 1, \ldots, n$$
$$\beta_j \overset{iid}{\sim} \mathcal{N}(0, \tau^2), \quad j = 1, \ldots, p$$

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

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

In both cases, the value of the hyperparameter, τ^2, will be inversely proportional to the choice of the particular value for λ. For ridge regression, λ is exactly equal to the shrinkage parameter of the hierarchical model, $\lambda = \sigma^2/\tau^2$.