Notes for 751-752
Sections 19-20

Martin Lindquist*

November 15, 2017

*Thanks to Brian Caffo and Ingo Ruczinski for supplying their notes, upon which much of this document is based.

19 Regularization Methods

Regularization imposes an upper threshold on the value least-squares coefficients can take, potentially providing more parsimonious solutions. Regularization methods are particularly useful when variables are highly correlated with one another (i.e. when there is multicollinearity). But they also have utility as a variable selection tool.

19.1 Bias-variance Tradeoff

So far we have been using the ordinary least-squares estimate:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

We have previously shown this to be the BLUE. However, is it still possible to improve upon it?

If an estimate has only a small bias but is substantially more precise than an unbiased estimate it may in fact be preferable. The quality of an estimator can be quantified using the mean square error:

$$MSE(\hat{\boldsymbol{\beta}}) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})].$$

The MSE can be written as the sum of the total variation of the estimator and its squared bias, i.e.,

$$MSE(\hat{\boldsymbol{\beta}}) = tr(\mathrm{var}(\hat{\boldsymbol{\beta}})) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

If $\hat{\boldsymbol{\beta}}$ is unbiased, then $MSE(\hat{\boldsymbol{\beta}}) = tr(\mathrm{var}(\hat{\boldsymbol{\beta}}))$.

19.2 Ridge regression

Consider adding a quadratic constraints to the least squares equation:.

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \boldsymbol{\beta}'\boldsymbol{\Gamma}\boldsymbol{\beta}.$$

In this case we consider instances where $\mathbf{X}$ is not necessarily full rank. The addition of the penalty is called "Tikhonov regularization" for the mathematician of that name. The specific instance of this regularization in regression is called ridge regression. The matrix $\boldsymbol{\Gamma}$ is typically assumed known or alternatively set to $\gamma\mathbf{I}$.

Another way to envision ridge regression is to think in the terms of a posterior mode on a regression model. Specifically, $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Gamma}/\sigma^2$ and consider the model where $\mathbf{y} \mid \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Then one obtains the posterior for $\boldsymbol{\beta}$ and σ by multiplying the two densities. The posterior mode would be obtained by minimizing minus twice the log of this product

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2/\sigma^2 + \boldsymbol{\beta}'\boldsymbol{\Gamma}\boldsymbol{\beta}/\sigma^2.$$

which is equivalent to above in the terms of maximization for $\boldsymbol{\beta}$.

We'll leave it as an exercise to obtain that the estimate actually obtained is

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}\mathbf{X} + \boldsymbol{\Gamma})^{-1}\mathbf{X}'\mathbf{y}.$$

To see how this regularization helps with invertibility of $\mathbf{X}\mathbf{X}$, consider the case where $\boldsymbol{\Gamma} = \gamma\mathbf{I}$. If γ is very large then $\mathbf{X}\mathbf{X} + \gamma\mathbf{I}$ is simply small numbers added around an identity matrix, which is clearly invertible.

Consider the case where $\mathbf{X}$ is column centered and is of full column rank. Let $\mathbf{U}\mathbf{D}\mathbf{V}'$ be the SVD of $\mathbf{X}$ where $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$. Note then $\mathbf{X}\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}'$ and $(\mathbf{X}\mathbf{X})^{-1} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}'$ so that the ordinary least squares estimate satisfies

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V}\mathbf{D}^{-2}\mathbf{V}'\mathbf{V}\mathbf{D}\mathbf{U}\mathbf{y} = \mathbf{U}\mathbf{U}'\mathbf{y}.$$

Consider now the fitted values under ridge regression with $\boldsymbol{\Gamma} = \gamma\mathbf{I}$:

$$\begin{aligned}
\hat{\mathbf{y}}_{ridge} &= \mathbf{X}(\mathbf{X}\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}'(\mathbf{V}\mathbf{D}^2\mathbf{V}' + \gamma\mathbf{I})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}'(\mathbf{V}\mathbf{D}^2\mathbf{V}' + \gamma\mathbf{V}\mathbf{V}')^{-1}\mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}'\{\mathbf{V}(\mathbf{D}^2 + \gamma\mathbf{I})\mathbf{V}'\}^{-1}\mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V}(\mathbf{D}^2 + \gamma\mathbf{I})^{-1}\mathbf{V}'\mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{y} \\
&= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \gamma\mathbf{I})^{-1}\mathbf{D}\mathbf{U}'\mathbf{y} \\
&= \mathbf{U}\mathbf{W}\mathbf{U}'\mathbf{y}
\end{aligned}$$

where the third line follows since $\mathbf{X}$ is full column rank so that $\mathbf{V}$ is $p \times p$ of full rank and $\mathbf{V}^{-1} = \mathbf{V}'$ so that $\mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{I}$. Here $\mathbf{W}$ is a diagonal matrix with elements

$$\frac{D_i^2}{D_i^2 + \gamma}$$

where D_i^2 are the eigenvalues.

In the not full rank case, the same identity can be found, though it takes a bit more work. Now assume that $\mathbf{X}$ is of full row rank (i.e. that $n < p$ and there are no redundant subjects). Now note that $\mathbf{V}$ does not have an inverse, while $\mathbf{U}$ does (and $\mathbf{U}^{-1} = \mathbf{U}'$. Further note via the Woodbury theorem (where $\theta = 1/\gamma)d$:

$$\begin{aligned}
(\mathbf{X}\mathbf{X} + \gamma\mathbf{I})^{-1} &= \theta\mathbf{I} - \theta^2\mathbf{X}'(\mathbf{I} + \theta\mathbf{X}\mathbf{X}')^{-1}\mathbf{X} \\
&= \theta\mathbf{I} - \theta^2\mathbf{V}\mathbf{D}\mathbf{U}'(\mathbf{U}\mathbf{U}' + \theta\mathbf{U}\mathbf{D}^2\mathbf{U}')^{-1}\mathbf{U}\mathbf{D}\mathbf{V}' \\
&= \theta\mathbf{I} - \theta^2\mathbf{V}\mathbf{D}\mathbf{U}'\{\mathbf{U}(\mathbf{I} + \theta\mathbf{D}^2)\mathbf{U}')\}^{-1}\mathbf{U}\mathbf{D}\mathbf{V}' \\
&= \theta\mathbf{I} - \theta^2\mathbf{V}\mathbf{D}\mathbf{U}'\{\mathbf{U}(\mathbf{I} + \theta\mathbf{D}^2)^{-1}\mathbf{U}')\}\mathbf{U}\mathbf{D}\mathbf{V}' \\
&= \theta\mathbf{I} - \theta^2\mathbf{V}\mathbf{D}(\mathbf{I} + \theta\mathbf{D}^2)^{-1}\mathbf{D}\mathbf{V}' \\
&= \theta\mathbf{I} - \theta^2\mathbf{V}\tilde{\mathbf{D}}\mathbf{V}'
\end{aligned}$$

where $\tilde{\mathbf{D}}$ is diagonal with entries $D_i^2/(1+\theta D_i^2)$ where D_i are the diagonal entries of $\mathbf{D}$. Then:

$$\begin{aligned}\hat{\mathbf{y}}_{ridge} &= \mathbf{X}(\mathbf{X}\mathbf{X}+\gamma\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{U}\mathbf{D}\mathbf{V}'(\theta\mathbf{I}-\theta^2\mathbf{V}\tilde{\mathbf{D}}\mathbf{V}')\mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{y} \\ &= \mathbf{U}\mathbf{D}(\theta\mathbf{I}-\theta^2\tilde{\mathbf{D}})\mathbf{D}\mathbf{U}'\mathbf{y} \\ &= \mathbf{U}\mathbf{W}\mathbf{U}'\mathbf{y}\end{aligned}$$

Thus we've covered the full row and column rank cases. (Omitting the instance where $\mathbf{X}$ is neither full row nor column rank.)

Let us now study the bias and variance properties of the ridge estimator when $\Gamma = \gamma\mathbf{I}$:

$$\begin{aligned}E(\hat{\boldsymbol{\beta}}_{ridge}) &= (\mathbf{X}'\mathbf{X}+\gamma\mathbf{I})^{-1}\mathbf{X}'E(\mathbf{y}) \\ &= (\mathbf{X}'\mathbf{X}+\gamma\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= (\mathbf{X}'\mathbf{X}+\gamma\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X}+\gamma\mathbf{I}-\gamma\mathbf{I})\boldsymbol{\beta} \\ &= (\mathbf{I}-\gamma(\mathbf{X}'\mathbf{X}+\gamma\mathbf{I})^{-1})\boldsymbol{\beta} \\ &= \boldsymbol{\beta}-\gamma(\mathbf{X}'\mathbf{X}+\gamma\mathbf{I})^{-1}\boldsymbol{\beta}\end{aligned}$$

The bias increases as a function of γ. The variance-covariance matrix can be expressed as follows:

$$Var(\hat{\boldsymbol{\beta}}_{ridge}) = \sigma^2(\mathbf{X}'\mathbf{X}+\gamma\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X}+\gamma\mathbf{I})^{-1}$$

which can be simplified as follows:

$$Var(\hat{\boldsymbol{\beta}}_{ridge}) = \sigma^2\mathbf{V}(\mathbf{D}^2+\gamma\mathbf{I})^{-1}\mathbf{D}^2(\mathbf{D}^2+\gamma\mathbf{I})^{-1}\mathbf{V}'.$$

The total variability is represented by the trace of the variance-covariance matrix. Here we can write:

$$tr(Var(\hat{\boldsymbol{\beta}}_{ridge})) = \sigma^2\sum_{j=1}^{p}\frac{d_j^2}{(d_j^2+\gamma)^2}.$$

Compare this to the OLS solution:

$$tr(Var(\hat{\boldsymbol{\beta}})) = \sigma^2\sum_{j=1}^{p}\frac{1}{d_j^2}.$$

Thus, one can show that the ridge estimator has systematically smaller total variation, i.e.

$$tr(Var(\hat{\boldsymbol{\beta}}_{ridge})) \leq tr(Var(\hat{\boldsymbol{\beta}})).$$

19.3 Coding example

In the example below, we use the `swiss` data set to illustrate fitting ridge regression. In this example, penalization isn't really necessary, so the code is more used to simply show the fitting. Notice that `lm.ridge` and our code give slightly different answers. This is due to different scaling options for the design matrix.

```
data(swiss)
y = swiss[,1]
x = swiss[,-1]
y = y - mean(y)
x = apply(x, 2, function(z) (z - mean(z)) / sd(z))
n = length(y); p = ncol(x)
##get ridge regression estimates for varying lambda
lambdaSeq = seq(0, 100, by = .1)
betaSeq = sapply(lambdaSeq, function(l) solve(t(x) %*% x + l * diag(rep(1, p)), t(x
plot(range(lambdaSeq), range(betaSeq), type = "n", xlab = "- lambda", ylab = "Beta'
for (i in 1 : p) lines(lambdaSeq, betaSeq[i,])

##Use R's function for Ridge regression
library(MASS)
fit = lm.ridge(y ~ x, lambda = lambdaSeq)
plot(fit)
```

19.4 Lasso regression

The Lasso has been somewhat of a revolution in statistics and biostatistics of late. The central idea of the lasso is to create a penalty that forces certain coefficients to be zero. For centered $\mathbf{y}$ and centered and scaled $\mathbf{X}$, consider minimizing

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

subject to $\sum_{i=1}^{p} |\beta_i| < t$. The Lagrangian form of this minimization can be written as minimizing

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \gamma \sum_{i=1}^{n} |\beta_i|.$$

Here γ is a penalty parameter. As the Lasso constrains $\sum_{i=1}^{p} |\beta_i| < t$, which has sharp corners on the axes, it has a tendency to set parameters exactly to zero. Thus, it is thought of as doing model selection along with

penalization. Moreover, the Lasso handles the $p > n$ problem. Finally, it's a convex optimization problem, so that numerically solving for the Lasso is stable. We can more generally specify the parameter as

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \gamma \sum_{i=1}^{n} |\beta_i|^q.$$

for $q > 0$. We obtain a case of ridge regression when $q = 2$ and the Lasso when $q = 1$. Since $(\sum_{i=1}^{n} |\beta_i|^q)^{1/q}$ is a norm, usually called the ℓ_q norm, the various forms of regression are often calld ℓ_q regression. For example, ridge regression could be called ℓ_2 regression, the Lasso ℓ_1 regression and so on. We could write the penalized regression estimate as

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \gamma ||\boldsymbol{\beta}||_q^q$$

where $|| \cdot ||_q$ is the ℓ_q norm.

Notice that as q tends to zero, it tends to all of the mass on the axes. In contrast, as q tends to infinity, it tends to a square. The limit as q tends to 0 is called the l_0 norm, which just penalizes the number of non-zero coefficients.

Just like with ridge regression, the Lasso has a Bayesian representation. Let the prior on β_i be iid from a Laplacian distribution with mean 0, which has density $\frac{\theta}{2} \exp(-\theta|\beta_i|)$, and is denoted Lapplace$(0, \theta)$. Then, the Lasso estimate is the posterior mean assuming $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$ and $\beta_i \sim_{iid}$ Laplace$(0, \gamma/2\sigma^2)$. Then minus twice the log of the posterior for $\boldsymbol{\beta}$, conditioning on σ, is proportional to

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

The connection with Bayesian statistics is somewhat loose for Lasso regression. While the Lasso is the posterior mode under a specific prior, whether or not that prior makes sense from a Bayesian perspective is not clear. Furthermore, the full posterior for a parameter in the model is averaged over several sparse models, so is actually not sparse. Also, the posterior mode is conditioned on σ under these assumptions, Bayesian analysis usually take into account the full posterior.

19.5 Coding example

Let's give an example of coding the Lasso. Here, because the optimization problem isn't closed form, we'll rely on the `lars` package from Tibshirani and Efron. Also assume the code from the ridge regression example.

```
library(lars)
data(swiss)
y = swiss$Fertility
x = as.matrix(swiss[,-1])
fit2 = lars(x, y, type = c("lasso"))
plot(fit2)
```

20 Dimension Reduction Approaches

Next we explore a class of approaches that transform the predictors and then fit an OLS model using a subset of the transformed variables. We refer to these techniques as dimension reduction methods. Here we focus in particular on principal components regression. However, we begin by reviewing principal components analysis (PCA).

20.1 Principle component analysis

Principal components analysis (PCA) is a multivariate procedure concerned with explaining the variance-covariance structure of a random vector. In PCA, a set of correlated variables are transformed into a set of uncorrelated variables, ordered by the amount of variability in the data that they explain. The new variables are linear combinations of the original variables, and several of them can be ignored with a minimum loss of information. Thus, PCA provides a lower dimensional basis to represent the data.

Let us express $\mathbf{X}$ in terms of its SVD, $\mathbf{X} = \mathbf{UDV}'$. Here

$$\mathbf{D} = \text{diag}\{d_1, d_2, \ldots d_p\},$$

where $d_1 \geq d_2 \geq \ldots \geq d_p$. Note, we can write $\mathbf{X}'\mathbf{X} = \mathbf{VD}^2\mathbf{V}'$. Hence, the columns of $\mathbf{V}$ are the eigenvectors for $\mathbf{X}'\mathbf{X}$.

The principal components of the matrix $\mathbf{X}$ is a linear re-parameterization $\mathbf{ZW}$ such that: (i) the re-parameterized variables are uncorrelated with one another; and (ii) the first component has the largest variance of all linear combinations of the the columns of $\mathbf{X}$, the second component has the largest variance conditional on being uncorrelated withe the first, etc.

Here $\mathbf{W}$ is an orthogonal matrix called the loadings. Let $\mathbf{w}_1$ be the first column of the matrix $\mathbf{W}$. Then the first principal component is $\mathbf{z}_1 = \mathbf{X}\mathbf{w}_1$. We seek $\mathbf{w}_1$ so that

$$\max_{||\mathbf{w}_1||=1} \{\langle \mathbf{X}\mathbf{w}_1, \mathbf{X}\mathbf{w}_1 \rangle\}.$$

This is maximized when $\mathbf{w}_1$ is a multiple of the first right singular vector, i.e., the first column of $\mathbf{V}$ from the SVD. Similarly, the second column of $\mathbf{W}$ is the the second column of $\mathbf{V}$, etc.

Hence, the principal components are given by:

$$\mathbf{Z} = \mathbf{XV}.$$

In addition, the following relationship holds:

$$\begin{aligned} \mathbf{Z} &= \mathbf{XV} \\ &= \mathbf{UDV}'\mathbf{V} \\ &= \mathbf{UD} \end{aligned}$$

Thus, the principal components are the weighted columns of $\mathbf{U}$.

Note that

$$\text{var}(\mathbf{z}_i) = d_i^2$$

for $i = 1, \ldots p$. Therefore, we often quantify the proportion of the explained variance by the first m principal components as follows:

$$\frac{d_1^2 + \cdots d_m^2}{d_1^2 + \cdots d_p^2}.$$

20.2 Coding example

```
> data(swiss)
> y = swiss$Fertility
> x = as.matrix(swiss[,-1])
> n = nrow(x)
> decomp = princomp(x, cor = TRUE)
> decomp$loadings

Loadings:
                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Agriculture       0.524  0.258          0.809
Examination      -0.572                 0.422 -0.702
Education        -0.492 -0.190 -0.539  0.332  0.567
Catholic          0.385 -0.370 -0.726 -0.101 -0.422
Infant.Mortality        -0.872  0.425  0.215
```

20.3 Principal component regression

Principal components regression (PCR) uses $\mathbf{Z}$ instead of $\mathbf{X}$ as the explanatory variables in the linear model. Importantly, the columns of $\mathbf{Z}$ are uncorrelated, so we can fit the model sequentially. In addition, only the variables $\mathbf{z}_1, \ldots \mathbf{z}_m$ for some $m \leq p$ are typically used. It therefore disregards the $p - m$ components with smallest eigenvalues. By manually setting the projection onto the principal component directions with small eigenvalues equal to 0, dimension reduction is achieved.

If we use all p principal components, the linear model can be written:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}\mathbf{V}\mathbf{V}'\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{Z}\gamma + \boldsymbol{\varepsilon} \end{aligned}$$

where $\gamma = \mathbf{V}'\boldsymbol{\beta}$. Under this formulation,

$$\begin{aligned}\hat{\gamma} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \\ &= \mathbf{D}^{-2}\mathbf{Z}'\mathbf{y}.\end{aligned}$$

Hence, be can write:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathbf{V}\hat{\gamma} \\ &= \mathbf{V}\mathbf{D}^{-2}\mathbf{Z}'\mathbf{y}.\end{aligned}$$

Using all p principal components, this is equivalent to the OLS solution.

In practice, we only use $m < p$ principal components. Let $\mathbf{Z}_{(m)} = [\mathbf{z}_1, \ldots \mathbf{z}_m]$. Then, in a similar manner as above we can show that:

$$\hat{\boldsymbol{\beta}}_{(m)} = \mathbf{V}_{(m)} D_{(m)}^{-2} \mathbf{Z}'_{(m)}\mathbf{y}.$$

The total variability can be written:

$$tr(\text{var}(\hat{\boldsymbol{\beta}}_{(m)}) = \sigma^2 \sum_{i=1}^{m} \frac{1}{d_i^2}.$$

Compare this to the OLS solution:

$$tr(\text{var}(\hat{\boldsymbol{\beta}})) \;\; = \;\; \sigma^2 \sum_{j=1}^{p} \frac{1}{d_j^2}.$$

Hence, it holds that $tr(\text{var}(\hat{\boldsymbol{\beta}}_{(m)}) \leq tr(\text{var}(\hat{\boldsymbol{\beta}}))$. However, $\hat{\boldsymbol{\beta}}_{(m)}$ will be biased. Thus, the mean square error is given by

$$MSE(\hat{\boldsymbol{\beta}}_{(m)}) \;\; = \;\; \sigma^2 \sum_{j=1}^{m} \frac{1}{d_j^2} + \sum_{j=m+1}^{p} \gamma_j^2.$$

As more principal components are used in the regression model, the bias decreases but the variance increases. PCR tends to perform well in settings when the first few principal components capture most of the variation in the predictors as well as the relationship with the response. Note that even though PCR provides a simple way to perform regression using $m < p$ predictors, it is not a feature selection method. One can typically choose the number of principal components by cross-validation.

It should be noted that PCR identifies linear combinations, or directions, that best represents the predictors. These directions are identified is an unsupervised way, since the response $\mathbf{y}$ is not used to help determine the principal component directions. Thus, there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response. To achieve this, as an alternative, methods such as partial least squares can be used.