Advanced Methods in Biostatistics II

Lecture 8

November 16, 2017

Linear model

- Consider the linear model

  $$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

  where $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

- Today we will continue discussing regularization methods, and also cover data reduction techniques.

Penalized likelihood

- Last time we discussed ridge regression.
- It was originally proposed as a method to deal with collinearity.
- Today it is more commonly viewed as a form of penalized likelihood estimation:

  $$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda p(\beta)$$

  where $p(\beta)$ is a nonnegative penalty function.

Lasso regression

- A particularly popular penalized likelihood approach has been the least absolute shrinkage and selection operator or Lasso.
- The central idea of the Lasso is to create a penalty that rather than simply shrinking certain (unimportant) coefficients instead forces them to be zero.
- Hence, it tends to provide a sparse solution to the least-squares problem.

Lasso regression

- For centered and scaled $\mathbf{X}$ and $\mathbf{y}$, consider minimizing

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

  subject to

  $$\sum_{i=1}^{p} |\beta_i| < t.$$

- This is referred to as an L_1-penalty.

Lasso regression

- The Lasso constraint:

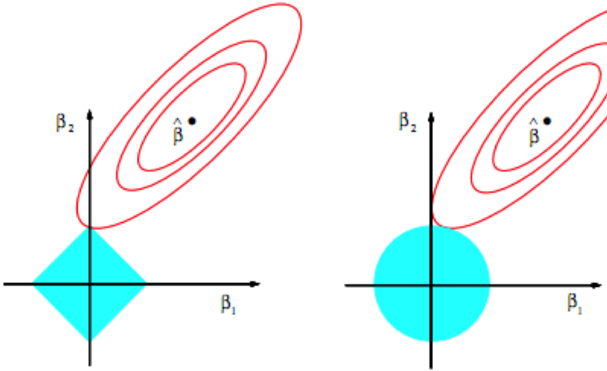
$$\sum_{i=1}^{p} |\beta_i| < t$$

  has sharp corners on the axes.

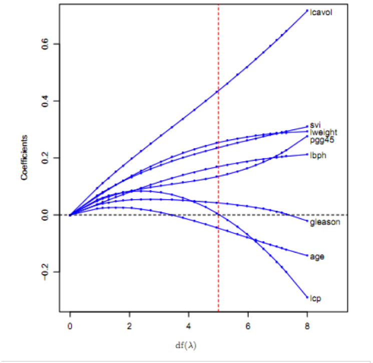
- Thus, it has a tendency to force certain coefficients to be exactly zero.

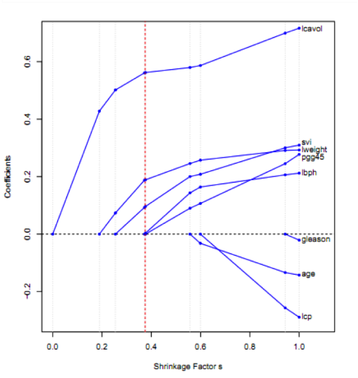
- Thus, it provides model selection along with penalization.

Lasso vs. Ridge



Lasso vs. Ridge





Lasso regression

- Using a Lagrange multiplier we can express the problem as follows:

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

- The term $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.

- There is a on-to-one correspondence between λ and s.

Lasso regression

- In contrast to ridge regression, the Lasso does not have a closed-form solution.
- It is a quadratic programming problem, whose solution can efficiently be estimated.
- Computation is performed using the LARS algorithm or coordinate descent.
- The penalty term is usually chosen using cross-validation.

Bayesian interpretation

- As with ridge regression, the Lasso has a Bayesian representation.
- Let the prior on β_i be iid Laplace$(0, \theta)$, which has density:

$$\frac{\theta}{2}\exp(-\theta|\beta_i|).$$

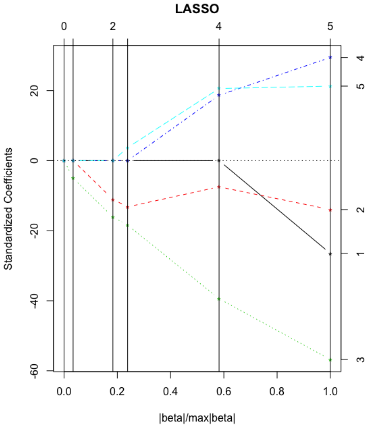
- Assuming this prior and the likelihood $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 I)$, the posterior mean is equivalent to the Lasso estimate.

Coding example

Use the Lars package to fit Lasso in R.

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

Coding example



```
> x[1,]
     Agriculture    Examination    Education    Catholic    Infant.Mortality
```

Penalized least-squares

- We often more generally specify the penalized least-squares criteria as

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

  for $q > 0$.

- We obtain as special cases ridge regression when $q = 2$ and the Lasso when $q = 1$.

Penalized least-squares

- 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 called ℓ_q regression.

- We could write the penalized least-squares estimate as

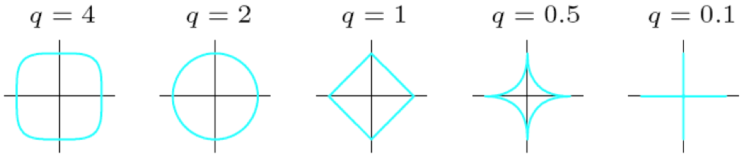
  $$||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda||\beta||_q^q$$

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

Penalized least-squares

- As the term q tends to zero, it tends to place all of the mass on the axes.
- The limit as q tends to 0 is called the ℓ_0 norm, which just penalizes the number of non-zero coefficients.
- As when q tends to infinity, it tends to a square.

More Penalties



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.

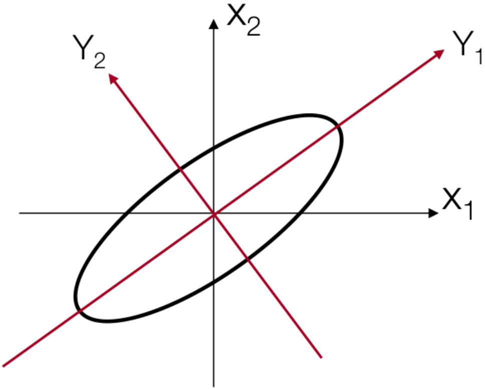
- We refer to these techniques broadly as dimension reduction methods.

- Here we focus on principal components regression.

Principal components 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.

Principal components analysis



Principal component analysis

- Let us express $\mathbf{X}$ in terms of its SVD, $\mathbf{X} = \mathbf{UDV}'$.
- Here
$$\mathbf{D} = \mathrm{diag}\{d_1, d_2, \ldots d_p\},$$
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 $\mathbf{X}'\mathbf{X}$.

Principal component analysis

- The principal components of the matrix **X** is a linear re-parameterization $\mathbf{Z} = \mathbf{XW}$ such that:

  1. The re-parameterized variables are uncorrelated with one another.
  2. The first component has the largest variance of all linear combinations of the the columns of **X**, the second component has the largest variance conditional on being uncorrelated withe the first, etc.

Principal component analysis

- 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

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

Principal component analysis

- 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.

Principal component analysis

- Hence, the principal components are given by:

$$\mathbf{Z} = \mathbf{X}\mathbf{V}.$$

- In addition, the following relationship holds:

$$\begin{aligned}\mathbf{Z} &= \mathbf{X}\mathbf{V} \\ &= \mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V} \\ &= \mathbf{U}\mathbf{D}\end{aligned}$$

- The principal components are the weighted columns of **U**.

Principal component analysis

- Note that

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

  for $i = 1, \ldots p$.

- We often quantify the proportion of the explained variance by the first m principal components as follows:

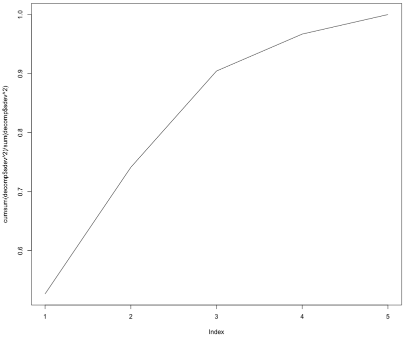
  $$\frac{d_1^2 + \cdots d_m^2}{d_1^2 + \cdots d_p^2}.$$

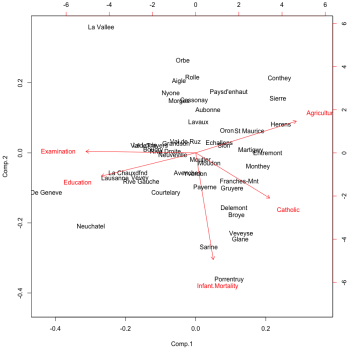
Coding example

```
> 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
```

Coding example





Principal component regression

- Principal components regression (PCR) uses $\mathbf{Z}$ instead of $\mathbf{X}$ as the explanatory variables.

- Importantly, the columns of $\mathbf{Z}$ are uncorrelated, so we can fit the model sequentially.

- In addition, only the variables $\mathbf{z}_1, \dots \mathbf{z}_m$ for some $m \leq p$ are typically used.

- It therefore disregards the $p - m$ smallest eigenvalue components.

- By manually setting the projection onto the principal component directions with small eigenvalues equal to 0, dimension reduction is achieved.

Principal component regression

- 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,

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

Principal component regression

- Hence, be can write:

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

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

Principal component regression

- 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:

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

Bias and variance

- The total variability can be written:

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

- Compare this to the OLS solution:

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

- Hence, it holds that $tr(\mathrm{var}(\hat{\beta}_{(m)}) \leq tr(\mathrm{var}(\hat{\beta}))$.

Bias and variance

- However, $\hat{\beta}_{(m)}$ will be biased.
- Thus, the mean square error is given by

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

Principal component bases

- As more principal components are used in the regression model, the bias decreases but the variance increases.

- PCR performs well in cases 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.

- In PCR, the number of principal components is typically chosen by cross-validation.

Principal component bases

- PCR identifies linear combinations, or directions, that best represents the predictors.
- These directions are identified is an unsupervised way, since the response **y** is not used to help determine the principal component directions.
- That is, the response does not influence the identification of the principal components.
- 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.

Ridge regression - revisited

- Recall that for ridge regression, we have:

$$\hat{\mathbf{y}}_r = \mathbf{H}_\lambda \mathbf{y}$$

where

$$\mathbf{H}_\lambda = \mathbf{UWU}'.$$

- Here **W** is a diagonal matrix whose elements are:

$$\frac{d_i^2}{d_i^2 + \lambda}$$

where d_i are the diagonal elements of **D** (i.e., the eigenvalues).

Ridge regression - revisited

- Thus, we can write:

$$\hat{\mathbf{y}}_r = \sum_{j=1}^{p} \frac{d_i^2}{d_i^2 + \lambda} (\mathbf{u}_j \mathbf{u}_j') \mathbf{y}.$$

- Ridge regression projects the vector **y** onto the principal component directions and then shrinks the projection on each direction.

- The amount of shrinkage depends on the variance of that principal component.

- In contrast, PCR sets directions with small variance equal to zero a priori.

High dimensional settings

- Most linear model techniques are designed for the low-dimensional setting (i.e., $n >> p$).
- However, increasingly we are faced with situations where $p > n$ (so-called small n, large p problems).
- Many traditional approaches are not appropriate in this setting, as they will overfit the data.

High dimensional settings

- Techniques such as ridge regression, the Lasso, and PCR are more appropriate.
- By either constraining the solution or performing data reduction, overfitting can be avoided.