BIG DATA ANALYTICS
Linear Model Selection and Regularization

Olga Klopp
klopp@essec.edu

ESSEC BUSINESS SCHOOL

CentraleSupélec

Linear Regression Model

- The linear model has important advantages in terms of inference
- On real-world problems, it is often surprisingly competitive in relation to non-linear methods
- One typically fits linear regression model using least squares

Alternative Fitting Procedures

- Linear model can be improved by replacing plain least squares fitting with some alternative fitting procedures

- **Alternative fitting procedures can yield better prediction accuracy and model interpretability**

Prediction Accuracy

- If the true relationship is approximately linear, the least squares estimates will have low bias
- If $n \gg p$, the least squares estimate also has low variance
- If n is not much larger than p: a lot of variability in the least squares fit $\implies$ overfitting and poor predictions
- If $p > n$, then there is no a unique least squares coefficient estimate and least squares estimates can not be used

By constraining the estimated coefficients, we can reduce the variance $\implies$ improvements in the accuracy

Model Interpretability

- Often some (or many) of the variables are not associated with the response
- By removing these variables we can obtain a model that is more easily interpreted
- The least squares is unlikely to yield any coefficient estimates that are exactly zero
- **Approaches for automatically performing variable selection**

Alternatives approaches

There are many alternatives (classical and modern) to using least squares:

- Subset Selection
- Shrinkage
- Dimension Reduction

Subset Selection

- Identifies a subset of the p predictors that are related to the response
- Fits a model using least squares on the reduced set of variables
- Examples:
  - Forward Stepwise Selection
  - Backward Stepwise Selection

Shrinkage

- Fits a model involving all p predictors
- The estimated coefficients are shrunken towards zero relative to the least squares estimates
- This shrinkage (or regularization) has the effect of reducing variance
- Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero
- Shrinkage methods can also perform variable selection
- Examples:
    - Lasso
    - Ridge Regression

Dimension Reduction

- Projects the p predictors into a M-dimensional subspace, where $M < p$
- These M projections are used to fit a linear regression model by least squares
- Examples:
  - Principal Components Regression
  - Partial Least Squares

Outline

Subset Selection

- Best Subset Selection
- Forward Selection
- Backward Selection
- Mixed selection

Shrinkage Methods

- Ridge Regression
- Lasso

Dimension Reduction Methods

High-Dimensional Data

Outline

Subset Selection

- Best Subset Selection
- Forward Selection
- Backward Selection
- Mixed selection

Shrinkage Methods

- Ridge Regression
- Lasso

Dimension Reduction Methods

High-Dimensional Data

Best Subset Selection

- We fit a separate least squares regression for each possible combination of the p predictors:
  - we fit all p models that contain exactly one predictor
  - all $\binom{p}{2} = p(p-1)/2$ models that contain exactly two predictors
  - ...
- We look at all of the resulting models, with the goal of identifying the best one

Best Subset Selection

Algorithm: Best subset selection

1. Let M_0 denote the null model which contains no predictors. This model simply predicts the sample mean for each observation
2. For $k = 1, 2, \dots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors
   (b) Pick the best among these $\binom{p}{k}$ models, and call it M_k (here best is defined as having the smallest RSS, or equivalently largest R^2)
3. Select a single best model from among $M_0, \dots, M_p$ using cross-validated prediction error, Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2...

Best Subset Selection

- In this Algorithm, Step 2 identifies the best model (on the training data) for each subset size
  - reduce the problem from one of 2^p possible models to one of $p+1$ possible models
- In order to select a single best model, we must choose among these $p+1$ options

Best Subset Selection

- This task must be performed with care:
  - the RSS of these $p + 1$ models decreases monotonically and the R^2 increases monotonically as the number of features included in the models increases
  - **If we use these statistics to select the best model, we will always end up with a model involving all of the variables**
- In Step 3, we use cross-validated prediction error, C_p , BIC, or adjusted R^2 in order to select among $M_0, M_1, \ldots, M_p$
- Same ideas apply to other types of models, such as logistic regression:
  - instead of ordering models by RSS in Step 2, we use the deviance

Computational Limitations

- Best subset selection is a simple and conceptually appealing approach
- Computational limitations:
  - The number of possible models that must be considered grows rapidly as p increases
  - if $p = 10 \implies \approx 1,000$ possible models
  - if $p = 20 \implies$ over one million possibilities!
- Computationally infeasible for values of $p > 40$
- Computationally efficient alternatives to the best subset selection: stepwise methods

Forward Selection

- Forward stepwise selection begins with a model containing no predictors
- It adds predictors to the model, one-at-a-time
- At each step the variable that gives the greatest additional improvement to the fit is added to the model

Forward Selection

Algorithm: Forward Stepwise Selection

1. Let M_0 denote the null model which contains no predictors
2. For $k = 1, 2, \dots p$:
   (a) Consider all $p - k$ models that augment the predictors in M_k with one additional predictor

   (b) Choose the best among these $p - k$ models, and call it $M_{(k+1)}$. Here best is defined as having the smallest RSS, or equivalently largest R^2
3. Select a single best model from among $M_0, \dots, M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward Selection

- Best subset selection involves fitting 2^p models
- Forward stepwise selection involves fitting one null model, along with $p - k$ models in the kth iteration:

$$1 + \sum_{k=0}^{(p-1)} (p - k) = 1 + p(p+1)/2 \quad \text{models}$$

- A substantial difference:
  - when $p = 20$, best subset selection requires fitting 1,048,576 models
  - forward stepwise selection requires fitting 211 models

Forward Selection: advantages

- Computational advantage over the best subset selection
- It tends to do well in practice
- Can be applied even if $n < p$:
    - it is possible to construct submodels $M_0, \dots, M_{(n-1)}$

Forward Selection: drawbacks

- It is not guaranteed to find the best possible model:
  - Forward selection is a greedy approach
  - Might include variables early that later become redundant:
    - Example: $p = 3$
    - the best possible one-variable model contains X_1
    - the best possible two-variable model contains X_2 and X_3
    - forward stepwise selection will fail to select the best possible two-variable model
    - because M_1 will contain X_1, so M_2 must also contain X_1 together with one additional variable

Backward Selection

- Provides an efficient alternative to best subset selection
- It begins with the full least squares model containing all p predictors
- Iteratively removes the least useful predictor, one-at-a-time

Backward Selection

Algorithm: Backward stepwise selection

1. Let M_p denote the full model, which contains all p predictors.
2. For $k = p, p-1, \dots, 1$:
   (a) Consider all k models that contain all but one of the predictors in M_k, for a total of $k-1$ predictors

   (b) Choose the best among these k models, and call it $M_{(k-1)}$ (best is defined as having the smallest RSS, or equivalently largest R^2)
3. Select a single best model from among $M_0, \dots, M_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Backward Selection

- The backward selection approach searches through only $1 + p(p+1)/2$ models
- Can be applied in settings where p is too large to apply best subset selection
- Backward stepwise selection is not guaranteed to yield the best model containing a subset of the p predictors
- Requires that the number of samples n is larger than the number of variables p

Mixed selection

- A combination of forward and backward selection
- Start with no variables in the model and we add the variable that provides the best fit
- We continue to add variables one-by-one
- The p-values for variables can become larger as new predictors are added to the model
- If at any point the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model

Mixed selection

- Continue to perform these forward and backward steps until
  - all variables in the model have a sufficiently low p-value
  - all variables outside the model would have a large p-value if added to the model

Choosing the Optimal Model

- Best subset selection, forward selection, and backward selection result in the creation of a set of models with different numbers of predictors
- **RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors**
- In order to select the best model with respect to test error, we need to estimate it
- Two common approaches:
  1. Indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting
  2. Directly estimate the test error, using either a validation set approach or a cross-validation approach.

Adjusting the training error for the model size

- *C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC) and adjusted R^2*
- For a fitted least squares model containing d predictors the Mallow's C_p estimate:

  $$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

  where $\hat{\sigma}^2$ is an estimate of the variance of the error
  - C_p statistic adds a penalty of $2d\hat{\sigma}^2$ to the training RSS
- One can show that if $\hat{\sigma}^2$ is an unbiased estimate of σ^2, then C_p is an unbiased estimate of test MSE
- AIC and BIC look similar to Mallow's C_p

Adjusted R^2

- The adjusted R^2 statistic is another popular approach for selecting among a set of models with different numbers of variables
- Recall: $R^2 = 1 - RSS/TSS$, where $TSS = \sum(y_i - \bar{y})^2$ is the total sum of squares for the response
- R^2 always increases as more variables are added
- For a least squares model with d variables:

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

Adjusted R^2

- Unlike C_p, AIC and BIC, for which a small value indicates a model with a low test error, a large value of adjusted R^2 indicates a model with a small test error
- Intuition:
  - once all of the correct variables have been included in the model, adding additional variables will lead to only a very small decrease in RSS
  - adding variables leads to an increase in d and consequently a decrease in the adjusted R^2
  - the model with the largest adjusted R^2 will have only correct variables and no noise variables

Choosing the Optimal Model

- AIC, BIC, and C_p can also be defined for more general types of models
- Alternative: directly estimate the test error using the *validation set and cross-validation methods*
- Advantage: provides a direct estimate of the test error and makes fewer assumptions about the true underlying model

Outline

Subset Selection
- Best Subset Selection
- Forward Selection
- Backward Selection
- Mixed selection

Shrinkage Methods
- Ridge Regression
- Lasso

Dimension Reduction Methods

High-Dimensional Data

Shrinkage Methods

Shrinkage Methods: a technique that constrains the coefficient estimates (or equivalently, that shrinks the coefficient estimates) towards zero

- Shrinking the coefficient estimates can significantly reduce their variance
- The two best-known techniques:
  - Ridge Regression
  - Lasso

Ridge Regression

- The least squares estimates: the values that minimize

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2$$

- Ridge regression:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \|\beta\|_{l_2}^2$$

where $\lambda \geq 0$ is a tuning parameter, to be determined separately.

Ridge Regression

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

- This equation trades off two different criteria:
  - it seeks coefficient estimates that fit the data well by making the RSS small
  - the second term is small when $\beta_1, \ldots, \beta_p$ are close to zero
  - it has the effect of shrinking the estimates of β_j towards zero

Tuning parameter

- The tuning parameter λ controls the relative impact of these two terms on the regression coefficient estimates:
  - When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates
  - As $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero
- Ridge regression will produce a different set of coefficient estimates, $\hat{\beta}^R_\lambda$, for each value of λ
- **Selecting a good value for λ is critical!** (using cross-validation)

Ridge Regression

- The shrinkage penalty is applied to $\beta_1, \ldots, \beta_p$, but not to the intercept β_0:
    - shrinks the estimated association of each variable with the response
    - do not shrink the intercept $=$ a measure of the mean value of the response when all $x_i = 0$

Ridge Regression

- The least squares coefficient estimates are scale invariant:
  - multiplying X_j by a constant c leads to a scaling of the least squares coefficient estimates by a factor of $1/c$ $\implies$ $X_j\hat{\beta}_j$ will remain the same
- The ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant
  - e.g. we can measure the income variable in dollars or in thousands of dollars $\implies$ reduction in the observed values of income by a factor of $1,000$
  - $\hat{\beta}_j^\lambda$ will depend not only on the value of λ, but also on the scaling of the jth predictor

Ridge Regression

- It is best to apply ridge regression after standardizing the predictors:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

  - the denominator is the estimated standard deviation of the jth predictor
  - all of the standardized predictors will have a standard deviation of one

Application to the Credit data set

- Credit data set:
  - Output: balance
  - Quantitative predictors:
    - age
    - cards (number of credit cards)
    - education (years of education)
    - income (in thousands of dollars)
    - limit (credit limit)
    - rating (credit rating)
    - ...

Example: Credit data set

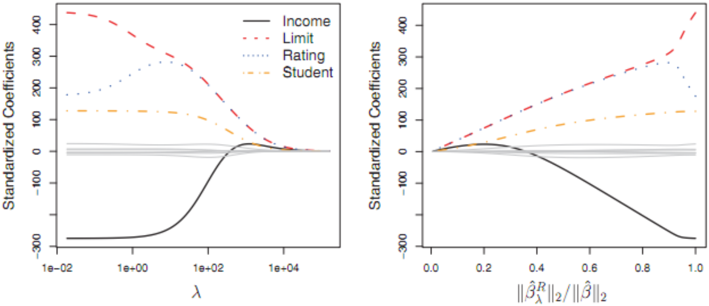


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

- 4 variables: the income, limit, rating and student variables have by far the largest coefficient estimates
- As λ increases, the ridge coefficient estimates shrink towards zero

Why Does Ridge Regression Improve Over Least Squares?

- Ridge regression's advantage over least squares is rooted in the **bias-variance trade-off**:
  - As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias

Why Does Ridge Regression Improve Over Least Squares?

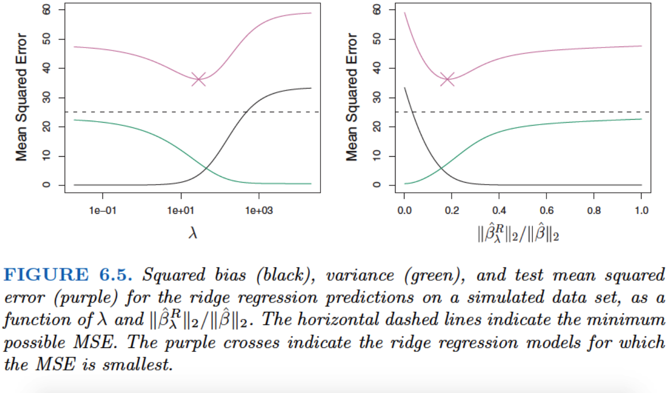


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

- At the least squares coefficient estimates $(\lambda = 0)$ the variance is high but there is no bias

Why Does Ridge Regression Improve Over Least Squares?

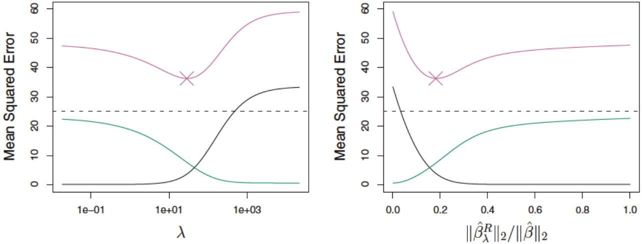


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

- For values of λ up to about 10, the variance (green) decreases rapidly with very little increase in bias (black)

Why Does Ridge Regression Improve Over Least Squares?

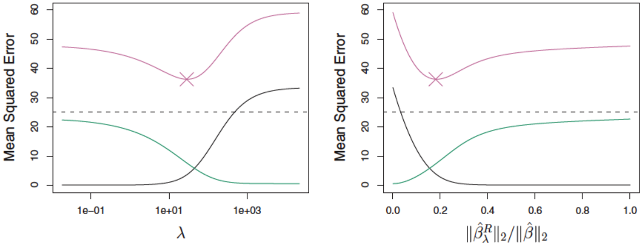


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

- Beyond this point, the decrease in variance slows, and the shrinkage on the coefficients results in a large increase in the bias

Lasso

- Ridge regression: will include all p predictors in the final model
- The penalty will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero
- Not a problem for prediction accuracy
- A challenge for model interpretation

The lasso is a relatively recent alternative to the ridge regression that overcomes this disadvantage

Lasso

- The lasso coefficients, β_λ^L minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \|\beta\|_{l_1}$$

- The lasso and ridge regression have similar formulations:
  - The only difference is that the β_j^2 term in the ridge regression penalty has been replaced by $|\beta_j|$
  - The lasso uses an l_1 penalty instead of an l_2 penalty:

$$\|\beta\|_{l_1} = \sum|\beta_i|, \quad \|\beta\|_{l_2} = \sqrt{\sum\beta_i^2}$$

Lasso

- The lasso shrinks the coefficient estimates towards zero
- The l_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large
  - Selecting a good value of λ for the lasso is critical! (using cross-validation)
- Like best subset selection, the lasso performs variable selection
- Models generated from the lasso are generally much easier to interpret than those produced by ridge regression

The lasso yields sparse models - that is, models that involve only a subset of the variables

Example: Credit data set

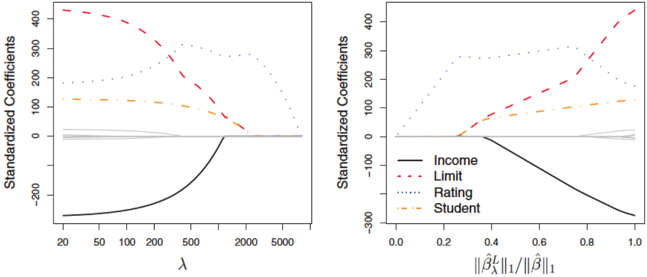


FIGURE 6.6. *The standardized lasso coefficients on the* Credit *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

- when $\lambda = 0$, then the lasso simply gives the least squares fit
- when λ becomes sufficiently large, the lasso gives the null model

Example: Credit data set

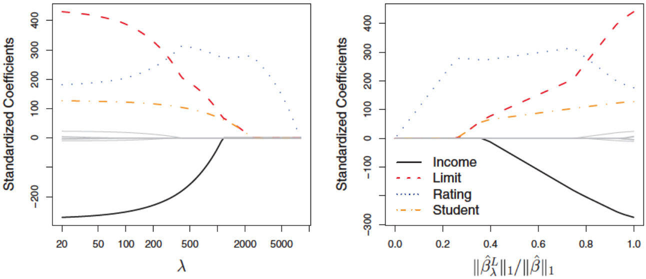


FIGURE 6.6. *The standardized lasso coefficients on the* Credit *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

- In between these two extremes, the ridge regression and lasso models are quite different from each other:
  - the rating predictor → student and limit → income → the remaining variables enter the model

Lasso vs Ridge

- Depending on the value of λ, the lasso can produce a model involving any number of variables
- In contrast, ridge regression will always include all of the variables in the model

Another Formulation for Ridge Regression and the Lasso

One can show that the lasso and ridge solve the problems

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

Connection between the lasso and best subset selection

$$\underset{\beta}{\text{minimize}} \quad \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 \tag{1}$$

$$\text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s$$

- $I(\beta_j \neq 0)$ is an indicator variable:
  - $I(\beta_j \neq 0) = 1$ if $\beta_j \neq 0$
  - $I(\beta_j \neq 0) = 0$ otherwise
- (1) is equivalent finding a set of coefficient estimates such that RSS is as small as possible, subject to the constraint that no more than s coefficients can be nonzero

Connection between the lasso and best subset selection

- This problem is equivalent to best subset selection
- Solving (1) is computationally infeasible when p is large
- We can interpret the lasso as a computationally feasible alternative to the best subset selection:
  - it replaces the intractable form of the budget in (1) with a form that is much easier to solve

The Variable Selection Property of the Lasso

Why does the lasso, unlike the ridge regression, result in coefficient estimates that are exactly equal to zero?

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 \quad \text{subject to} \quad \sum_{\mathbf{j=1}}^{\mathbf{p}} |\beta_{\mathbf{j}}| \leq \mathbf{s}$$

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 \quad \text{subject to} \quad \sum_{\mathbf{j=1}}^{\mathbf{p}} \beta_{\mathbf{j}}^{\mathbf{2}} \leq \mathbf{s}$$

- For $p = 2$ we have:

$$|\beta_{\mathbf{1}}| + |\beta_{\mathbf{2}}| \leq \mathbf{s} \quad \text{and} \quad \beta_{\mathbf{1}}^{\mathbf{2}} + \beta_{\mathbf{2}}^{\mathbf{2}} \leq \mathbf{s}$$

The Variable Selection Property of the Lasso

1996] REGRESSION SHRINKAGE AND SELECTION 271

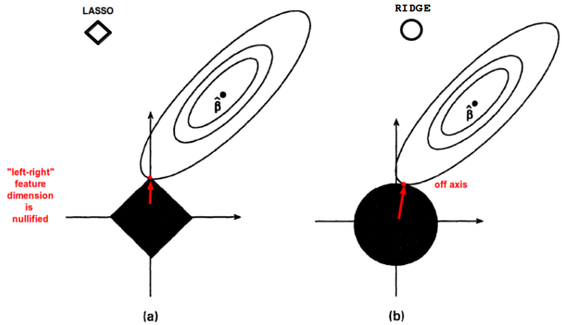
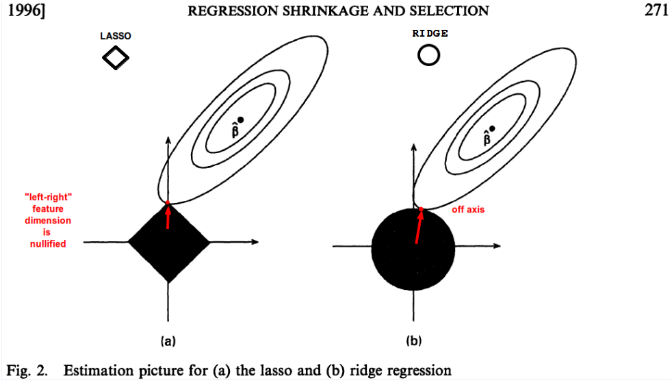


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

- $\hat{\beta}$: least squares solution
- the diamond and circle represent the lasso and ridge regression constraints

The Variable Selection Property of the Lasso



1996] REGRESSION SHRINKAGE AND SELECTION 271

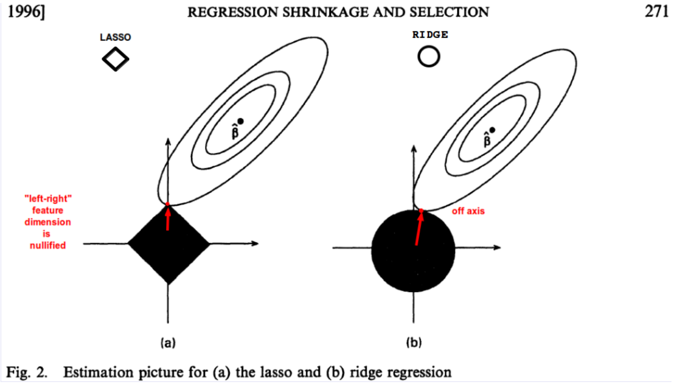
(a)

(b)

Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

- If s is sufficiently large, then the constraint regions will contain $\hat{\beta} \implies$ the ridge regression and lasso estimates will be the same as the least squares estimates

The Variable Selection Property of the Lasso

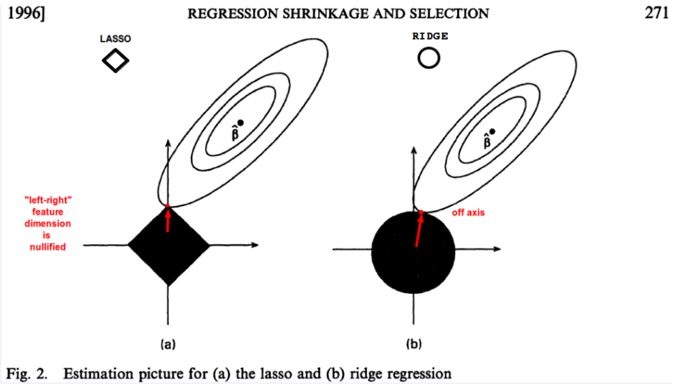


1996] REGRESSION SHRINKAGE AND SELECTION 271

Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

- The ellipses that are centered around $\hat{\beta}$ represent regions of constant RSS
- As the ellipses expand away from the least squares coefficient estimates, the RSS increases

The Variable Selection Property of the Lasso

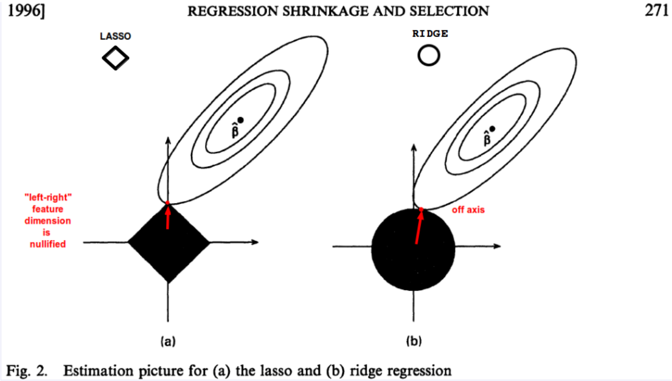


1996] REGRESSION SHRINKAGE AND SELECTION 271

Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

- lasso and ridge regression coefficient estimates are given by the first point at which an ellipse contacts the constraint region

The Variable Selection Property of the Lasso

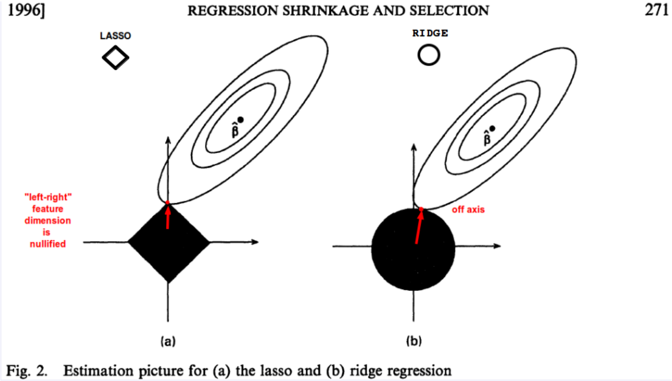


1996] REGRESSION SHRINKAGE AND SELECTION 271

Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

- ridge regression has a circular constraint; the intersection will not generally occur on an axis $\implies$ the ridge regression coefficient estimates will be non-zero

The Variable Selection Property of the Lasso



1996] REGRESSION SHRINKAGE AND SELECTION 271

Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

- the lasso constraint has corners at each of the axes, $\Longrightarrow$ the ellipse will often intersect the constraint region at an axis

Comparing the Lasso and Ridge Regression

- Lasso produces simpler and more interpretable models that involve only a subset of the predictors
- Which method leads to better prediction accuracy?
  - one expects the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients
  - ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size
  - the number of predictors that is related to the response is never known a priori for real data sets
  - a technique such as cross-validation can be used in order to determine which approach is better on a particular data set

Selecting the Tuning Parameter

- Selecting a value for the tuning parameter λ ?
- Cross-validation provides a simple way to tackle this problem:
  - choose a grid of λ values
  - compute the cross-validation error for each value of λ
  - select the tuning parameter value for which the cross-validation error is smallest
  - the model is refit using all of the available observations and the selected value of the tuning parameter.

Outline

Subset Selection

- Best Subset Selection
- Forward Selection
- Backward Selection
- Mixed selection

Shrinkage Methods

- Ridge Regression
- Lasso

Dimension Reduction Methods

High-Dimensional Data

Dimension Reduction Methods

- Transform the predictors and then fit a least squares model using the transformed variables
- For $M < p$ let $Z_1, Z_2, ..., Z_M$ be linear combinations of original p predictors:

  $$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

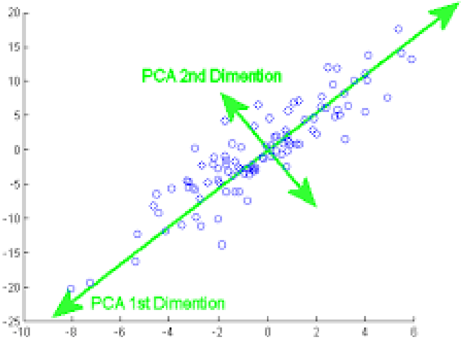
  for some constants ϕ_{jm}
- We fit the linear regression model using least squares and Z_m as predictors
- Can often outperform least squares regression
- The choice of $Z_1, Z_2, \ldots, Z_M$ can be achieved in different ways

Principal Components Regression

- *Principal components analysis (PCA) is an approach for deriving a low-dimensional set of features from a large set of variables*
- PCA is a popular tool for unsupervised learning
- Here we will see its use as a dimension reduction technique for regression

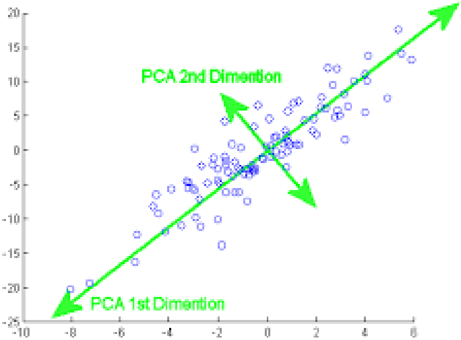
An Overview of Principal Components Analysis

The first principal component direction of the data is the direction along which the observations vary the most



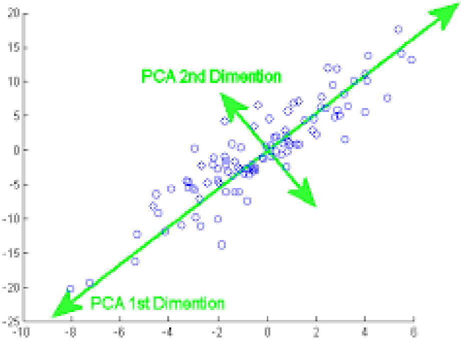
- The green line: the direction along which there is the greatest variability in the data

An Overview of Principal Components Analysis



- If we projected observations onto the first principal component direction then the resulting projected observations would have the largest possible variance

An Overview of Principal Components Analysis



- Another interpretation: the first principal component vector defines the line that is as close as possible to the data

An Overview of Principal Components Analysis

- Let $p = 2$: two predictors X_1 and X_2
- Assume that the first principal component direction is defined by (ϕ_1, ϕ_2) - the principal component loadings
- Out of every possible linear combination of X_1 and X_2 such that $\phi_1^2 + \phi_2^2 = 1$, this particular linear combination yields the highest variance:

$$\mathrm{Var}(\phi_1 \times (X_1 - \bar{X}_1) + \phi_2 \times (X_2 - \bar{x}_2)) \text{ is maximized}$$

An Overview of Principal Components Analysis

- It is necessary to consider only linear combinations of the form $\phi_1^2 + \phi_2^2 = 1$:
    - otherwise we could increase ϕ_1 and ϕ_2 arbitrarily in order to blow up the variance
- We set $Z_1 = \phi_1 \times (X_1 - \bar{X}_1) + \phi_2 \times (X_2 - \bar{X}_2)$

An Overview of Principal Components Analysis

- One can construct up to p distinct principal components:
  - The second principal component Z_2 is a linear combination of the variables that is uncorrelated with Z_1, and has largest variance subject to this constraint
  - The zero correlation condition of Z_1 with Z_2 is equivalent to the condition that the direction must be perpendicular to the first principal component direction.

Principal Components Regression

- Constructing the first M principal components, $Z_1, \ldots, Z_M$
- Using these components as the predictors in a linear regression model that is fit using least squares
- Key idea: we assume that the directions in which $X_1, \ldots, X_p$ show the most variation are the directions that are associated with Y

Principal Components Regression

- PCR will do well when the first few principal components are sufficient to capture most of the variation in the predictors as well as the relationship with the response
- Performing PCR with an appropriate choice of M can result in a substantial improvement over least squares:
  - by estimating only $M \ll p$ coefficients we can mitigate overfitting

Principal Components Regression

- PCR provides a simple way to perform regression using $M < p$
- It is not a feature selection method:
  - each of the M principal components used in the regression is a linear combination of all p of the original features
- The number of principal components is typically chosen by cross-validation.

Principal Components Regression

- It is better to standardize each predictor prior to generating the principal components:
    - it ensures that all variables are on the same scale
    - in the absence of standardization, the high-variance variables will tend to play a larger role in the principal components
- If the variables are all measured in the same units, then one might not need to standardize them.

Outline

Subset Selection

- Best Subset Selection
- Forward Selection
- Backward Selection
- Mixed selection

Shrinkage Methods

- Ridge Regression
- Lasso

Dimension Reduction Methods

High-Dimensional Data

High-Dimensional Data

- Most traditional statistical techniques for regression and classification are intended for the low-dimensional setting
- **The low-dimensional setting: n, the number of observations, is much greater than p, the number of features**
- Example: the task of developing a model to predict a patient's blood pressure on the basis of age, gender, and body mass index (BMI)
    - three predictors (or four if an intercept is included in the model)
    - several thousand patients for whom blood pressure, age, gender, and BMI are available
    - $n \gg p \implies$ the problem is low-dimensional

High-Dimensional Data

- New technologies have changed the way that data are collected
- In finance, marketing or medicine, it is now commonplace to collect a very large number of feature measurements
- While p can be extremely large, the number of observations n is often limited (due to cost or sample availability)

High-Dimensional Data: examples

- Rather than predicting blood pressure on the basis of just age, gender, and BMI, one might also collect measurements for half a million single nucleotide polymorphisms (SNP)
    - SNPs are individual DNA mutations that are relatively common in the population
- Then $n \approx 200$ and $p \approx 500,000$

High-Dimensional Data: examples

- Understanding peoples online shopping patterns:
    - features: all of the search terms entered by users of a search engine ("bag-of-words" model)
    - search histories of only a few hundred or a few thousand users who have consented to share their information
    - for a given user, each of the p search terms is scored present (0) or absent (1)
- Then $n \approx 1,000$ and p is much larger

High-Dimensional Data

Data sets containing more features than observations are referred to as high-dimensional

- Classical approaches such as least squares linear regression are not appropriate in this setting
- Many of the issues that we have discussed earlier (the bias-variance trade-off and the risk of overfitting) become particularly important in this setting
- These considerations also apply if p is slightly smaller than n

High-Dimensional Data

Need for extra care and specialized techniques for regression and classification when $p > n$!

What Goes Wrong in High Dimensions?

- When $p > n$ or $p \approx n$ a simple least squares regression line results in a perfect fit to the data, such that the residuals are zero
- This perfect fit will almost certainly lead to overfitting of the data
- **It is possible to perfectly fit the training data in the high-dimensional setting but the resulting linear model will perform extremely poorly on an independent test set, and does not constitute a useful model**
- Same applies to logistic regression, linear discriminant analysis, and other classical statistical approaches.

Regression in High Dimensions

- Methods, such as ridge regression, the lasso, and principal components regression, are particularly useful for performing regression in the high-dimensional setting
- They avoid overfitting by using a less flexible fitting approach than least squares

Curse of dimensionality

- The test error tends to increase as the number of predictors increases:
    - In general, adding additional features that are truly associated with the response will improve the fitted model
    - Adding noise features that are not associated with the response will lead to a deterioration in the fitted model and an increased test set error

Measurements for thousands or millions of features can lead to improved predictive models if these features are in fact relevant to the problem at hand, but will lead to worse results if the features are not relevant

Interpreting Results in High Dimensions

- When performing regression procedures in the high dimensional setting, we must be quite cautious in the way that we report the obtained results:
    - The variables might be correlated with each other (multicollinearity)
    - In the high-dimensional setting, the multicollinearity problem is extreme
    - We can never know exactly which variables (if any) truly are predictive of the outcome
    - At most, we can hope to assign large regression coefficients to variables that are correlated with the variables that truly are predictive of the outcome

Interpreting Results in High Dimensions: example

- Suppose that we are trying to predict blood pressure on the basis of half a million SNPs (single nucleotide polymorphisms)
- Forward stepwise selection indicates that 17 of those SNPs lead to a good predictive model on the training data
- It would be incorrect to conclude that these 17 SNPs predict blood pressure more effectively than the other SNPs not included in the model
- There are likely to be many sets of 17 SNPs that would predict blood pressure just as well as the selected model

Interpreting Results in High Dimensions: example

- An independent data: we would likely obtain a model containing a different, and perhaps even non-overlapping, set of SNPs!
- This does not detract from the value of the model obtained:
  - the model may be very effective in predicting blood pressure on an independent set of patients and clinically useful for physicians
- Do not to overstate the results obtained:
  - make it clear that what we have identified is simply one of many possible models for predicting blood pressure
  - it must be further validated on independent data sets

Interpreting Results in High Dimensions

- To be particularly careful in reporting errors and measures of model fit in the high-dimensional setting:
    - when $p > n$, it is easy to obtain a useless model that has zero residuals
    - one should never use sum of squared errors, p-values, R^2 statistics, or other traditional measures of model fit on the training data as evidence of a good model fit in the high-dimensional setting
- Instead report results on an independent test set, or cross-validation errors:
    - MSE or R^2 on an independent test set is a valid measure of model fit

References

- Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The elements of statistical learning - Data Mining, inference, and prediction.* Springer.
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning with applications in R.* Springer.