STA 5820
Chapter 6
Linear Model Selection and Regularization

Kazuhiko Shinki

Wayne State University

Overview:

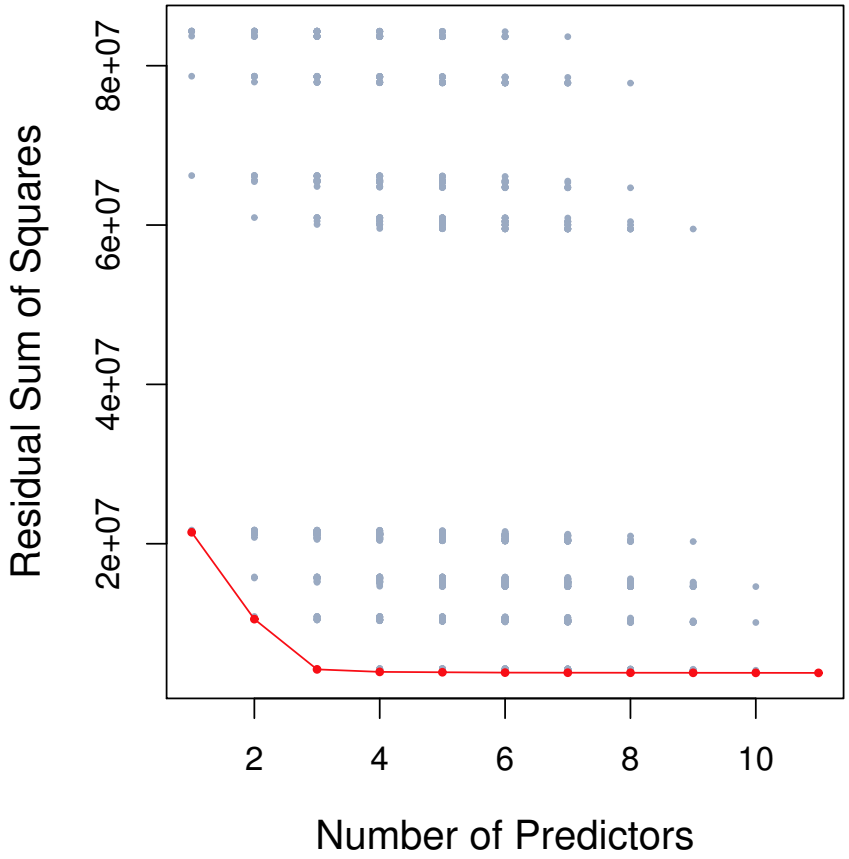
We discuss how to determine the best model among linear models:

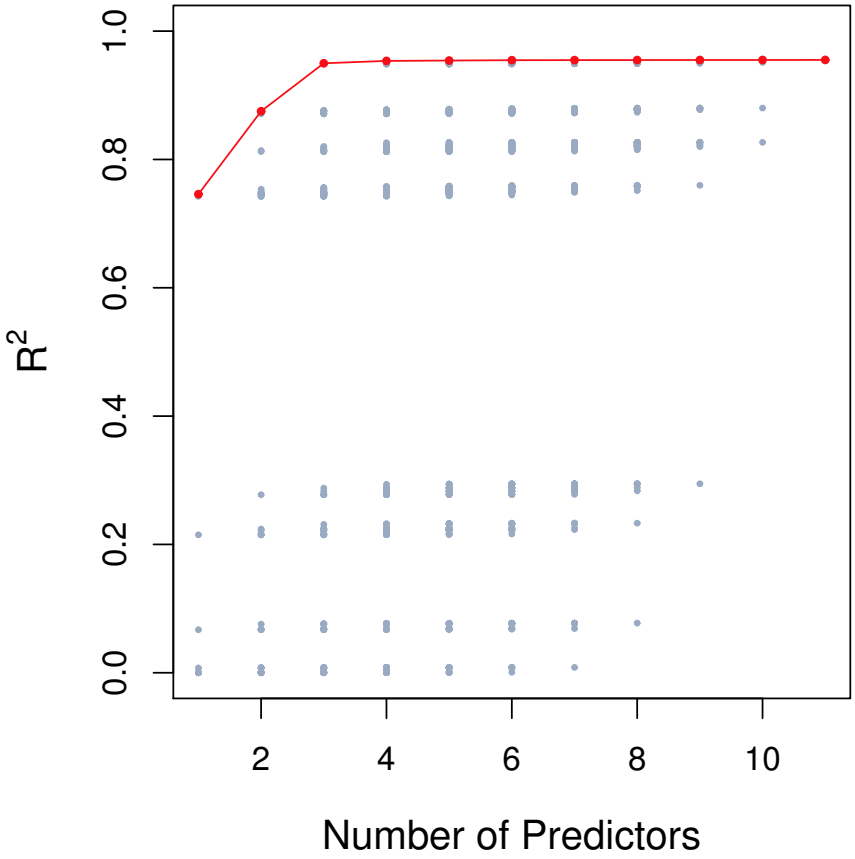
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

where $X_1, \cdots, X_p$ are some predictors.

Overview:

Figure 6-1: Training error is always smaller if the model is bigger.





Overview:

What is the best model?

There are a couple of objectives.

- The best prediction model:
  - Minimize the expected test error.
- The true model:
  - Assume that there is a correct model among our candidate model. We want to identify the correct model with an increasing probability as $n \to \infty$.

In general, the best prediction model is larger/more complex than the true model, because the former keeps variables which makes the prediction marginally better but are not significant.

The true model (and a model obtained by the method to search for it) is simpler and more interpretable than the best prediction model.

Overview:

Methods

We discuss three methods to find the best model.

- Subset (variable) selection
  - Choose a subset (of size d) of all p predictors, as we discussed in stepwise selection.
- Shrinkage
  - Use all p predictors, but make the coefficients smaller (shrink to 0).
- Dimension reduction
  - Projecting p-dimensional predictor vectors on to a lower dimensional space with keeping the maximum amount of information.

Shrinkage and dimension reduction are particularly useful when the number of preditors p is close to n or even larger than n, by computational, mathematical and empirical reasons.

6.1 Subset (variable) selection

There are two approaches:

- Best subset selection
  - Try all subsets of the entire set of predictors, and choose the model with minimum AIC.
  - Takes time (time order of 2^p). Realistic for $p \leq 10$.
  - Guaranteed to find the best model among all combinations of predictors.
- Stepwise selection
  - Start with a certain model, and add/delete one variable at a time to minimize AIC.
  - 3 types: forward selection, backward deletion, combination of the two.
  - Fast (time order of p^2 for backward or forward). Realistic for a large $p \leq 100$.
  - Not guaranteed to find the best model among all combinations of predictors.

6.1 Subset (variable) selection

What criterion can be used other than AIC?

AIC is defined by ($-2\log(\text{likelihood}) + 2d$) where d is the number of predictors. The idea is combining the error size (the 1st term) and a penalty for a larger model ($2d$). A smaller AIC means a better model.

6.1 Subset (variable) selection

There are a few other criteria.

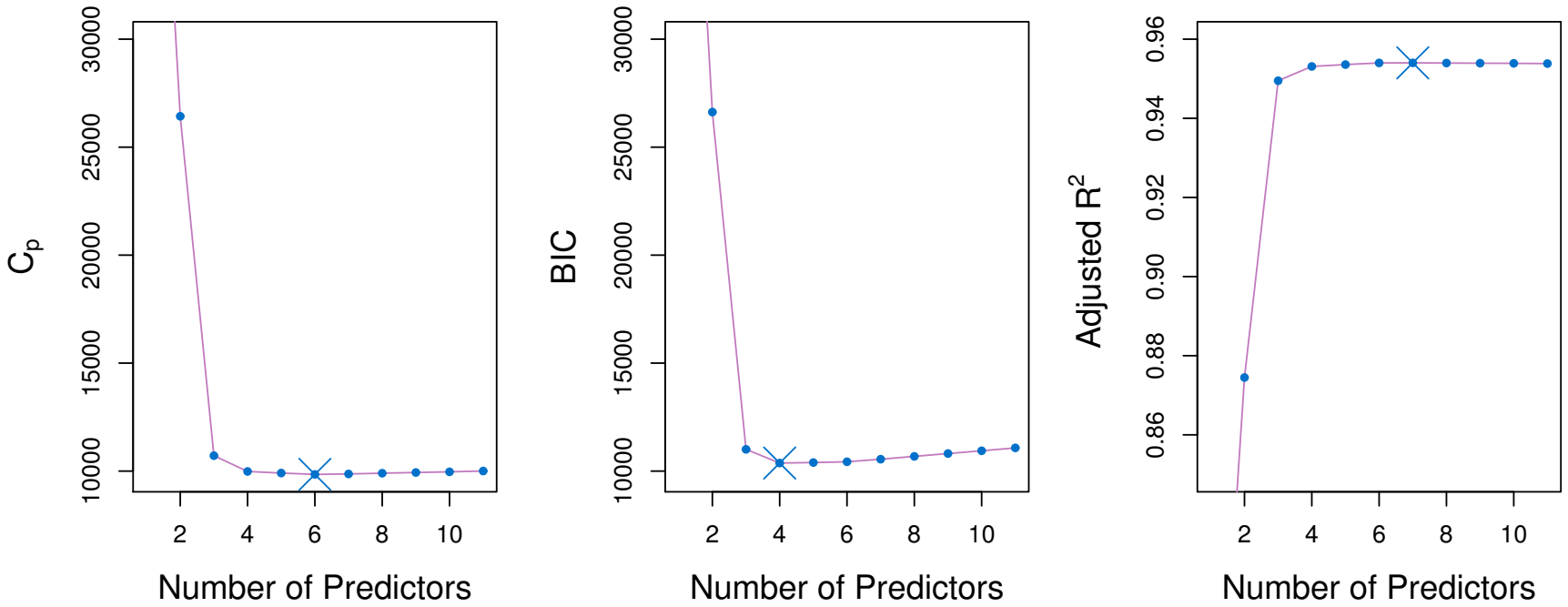
- Mallows' C_p:
  - Defined as

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

    where RSS is the sum of residual squares. d is the number of selected predictors $d \leq p$. $\hat{\sigma}^2$ is the estimated error size for the model including all p predictors.
  - A special case of AIC for least-square regression.
- Schwarz's Bayesian Information Criterion (BIC or SBIC)
  - Defined as $-\log(\text{likelihood}) + \log(n)d\hat{\sigma}^2$.
  - This is the criterion to choose the true model. It has a larger penalty on the number of predictors, so it tends to choose a smaller model than AIC does.

6.1 Subset (variable) selection

Figure 6-2: The model chosen by C_p (left), ***BIC*** (center) and adjusted R^2 (left).



AIC or cross-validation?

Both AIC and CV can be used to select the best prediction model. It is known that both choose the same model as the sample size $n \to \infty$.

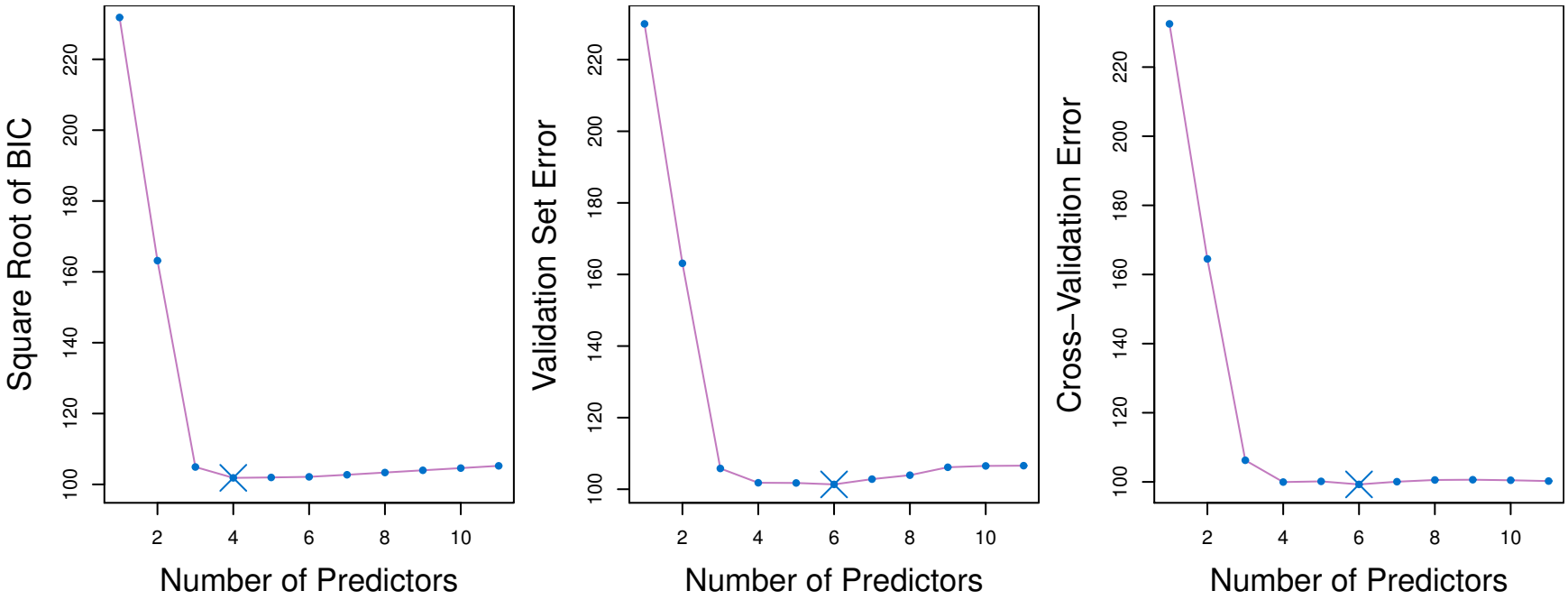
AIC is much faster, while CV is computer-intensive.

CV is applicable for any model, while AIC needs the likelihood function which is hard to obtain for complex models such as neural network.

As computation gets much faster with today's computer, CV is more common in many cases.

6.1 Subset (variable) selection

Figure 6-3: The models chosen by ***BIC*** (left), validation set (center) and CV (left) are similar. ***BIC*** chooses a slightly smaller model.



6.1 Subset (variable) selection

What if AIC (or CV error) is very close each other for many *d*'s?

When the number of predictor ***d*** is hard to decide, some people use a rule of thumb that choose the smallest ***d*** such that 'the test error **±1*s.d.*** of the test error' includes the smallest AIC (or CV error).

6.2.1 Ridge regression

Shrinkage method

Ridge regression

Recall that the least-square method minimizes the residual sum of squares:

$$\boldsymbol{RSS} = (\boldsymbol{Y} - \boldsymbol{X}\beta)'(\boldsymbol{Y} - \boldsymbol{X}\beta) \quad \text{(in a matrix form.)}$$

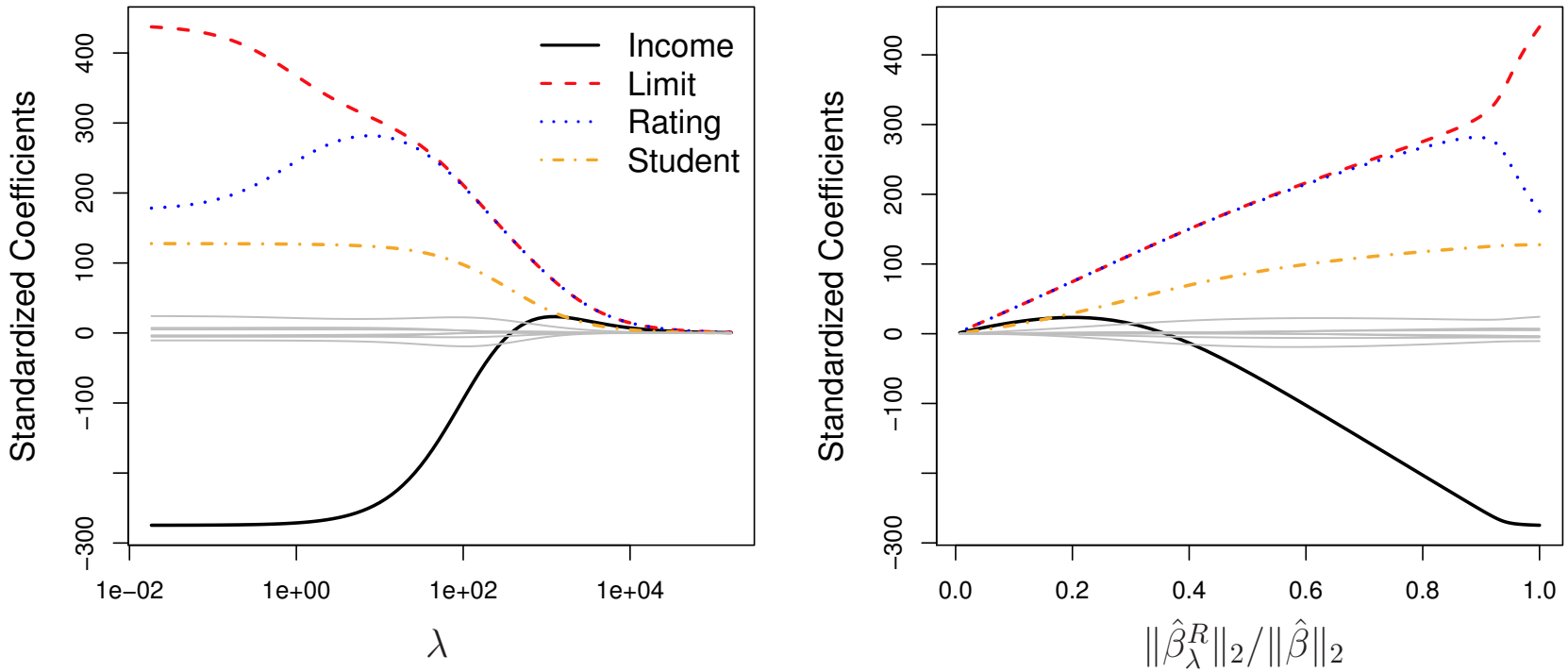
Theridge regression minimizes

$$\boldsymbol{loss} = \boldsymbol{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 = \boldsymbol{RSS} + \lambda ||\beta||^2$$

where λ is a so-called tuning parameter and $||\bullet||$ is the Euclidian norm.
lam is a constant, and we will discuss how to determine it later.
The idea is to penalize large parameter estimates to make the model more stable.

6.2.1 Ridge regression

Figure 6-4: An example of ridge regression. (Left:) A large λ shrinks the parameter estimates. (Right:) The same results but the horizontal axis is the ratio of ridge regression parameter vector size to linear regression parameter vector size (a small ratio implies a large λ).



6.2.1 Ridge regression

Pros of ridge against linear regression:

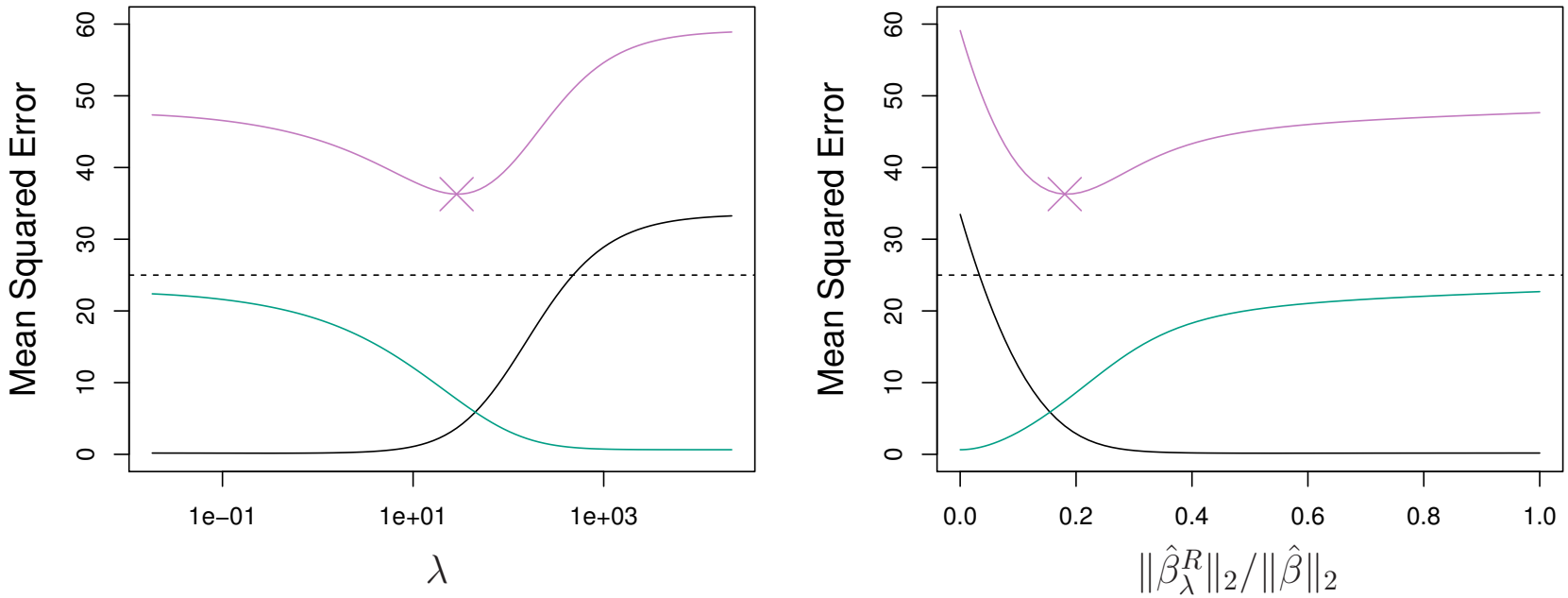
- It has a smaller variance, meaning the model is more stable when the sample changes.
- Perticularlly, it reduces a problem of multicollinearity.
- Prediction performance may be better for a large p or under severe multicollinearity.
- It is applicable when $p > n$, for which the least-square method can not uniquely identify the parameters (so shrinkage methods are also called regularization methods).
- Much faster than subset selection of all combinations.

Cons of ridge against linear regression:

- The parameter estimates are biased, so the prediction performance is worse when n is large and p is small.

6.2.1 Ridge regression

Figure 6-4: Due to bias-variance trade-off, a ridge regression with an optimal λ has smaller test error. Suquared bias (black); variance (green); test MSE (purple).



6.2.1 Ridge regression

Standardizing predictors

The penalty term in ridge regression changes when the scale of predictor $\boldsymbol{x}$ changes.

For example, if the unit of $\boldsymbol{x}$ changes from ***m*** to ***cm***, the value of $\boldsymbol{x}$ becomes 100 time larger, making the estimated coefficient becomes 100 times smaller. Therefore, the penalty on the coefficient of $\boldsymbol{x}$ becomes $\mathbf{100^2}$ times smaller!

To make the ridge regression result invariant in scale of predictors, all predictors should be standardized to have the sample variance of 1.

6.2.2 lasso

lasso

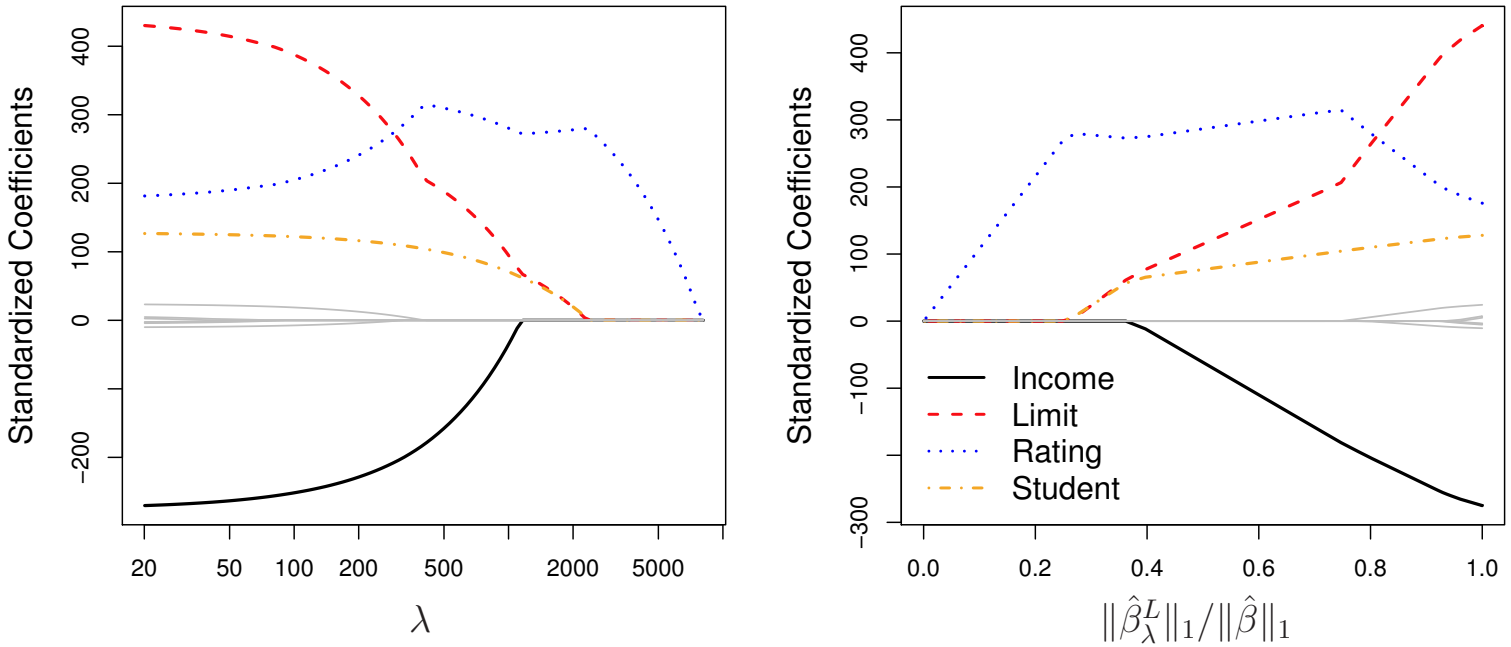
The lasso is similar to ridge regression, but with a different penalty on the size of parameter estimate of β. To be concrete, the lasso minimizes a loss function:

$$\textbf{\textit{loss}} = \textbf{\textit{RSS}} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

The lasso makes the parameter estimates smaller than the least-square regression does.

6.2.2 lasso

Figure 6-6: Example of lasso. Parameter estimates get smaller and often get to zero when λ gets larger.



6.2.2 lasso

Differences between ridge and lasso

Lasso has more parameter estimates equal to zero.

One explanation is that the absolute value $|\beta|$ is larger than $|\beta^2|$ when $|\beta|$ is small, so near-zero estimates of β shrink to zero in lasso.

6.2.2 lasso

Another formulation of lasso and ridge

The parameter estimates of the lasso can be obtained by minimizing ***RSS*** subject to

$$\sum_{j=1}^{p} |\beta_j| \leq s$$

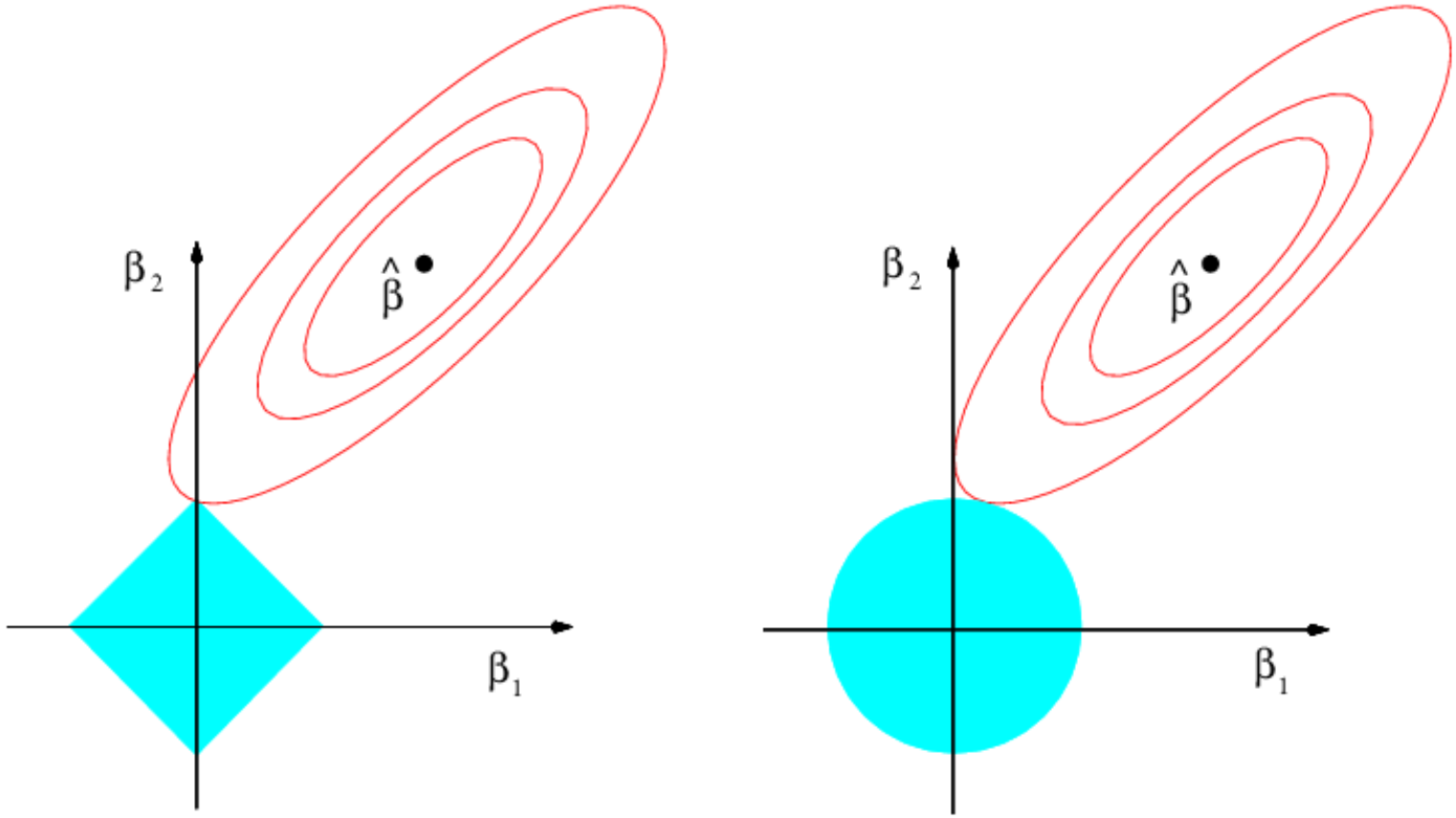
where s is a constant which depends on λ.

Similarly the parameter estimates of the ridge regression can be obtained by minimizing ***RSS*** subject to

$$\sum_{j=1}^{p} \beta_j^2 \leq s.$$

6.2.2 lasso

Figure 6-7: Another explanation on why lasso parameters tend to be zero, by using another formulation of lasso and ridge regression.



6.2.2 lasso

Comparing lasso and ridge regression

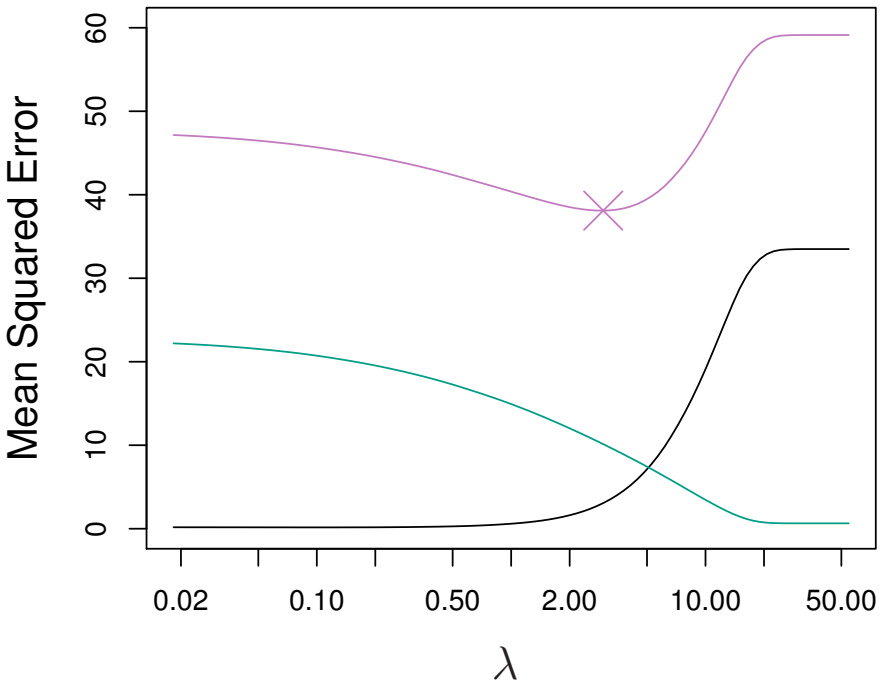
The lasso tends to give a simpler model since more parameter estimates become zero.

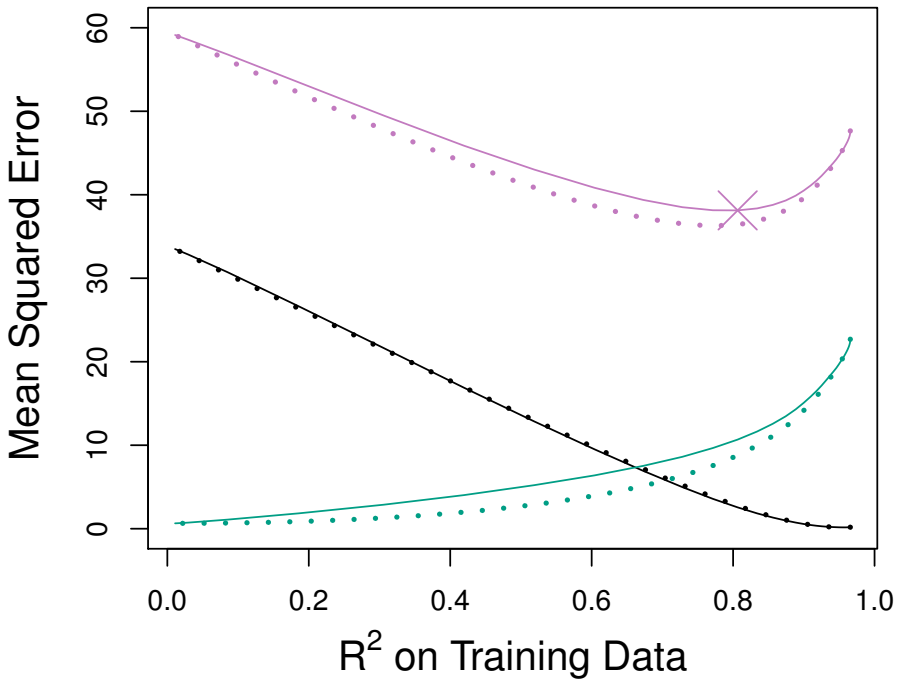
In general, we can not say whether lasso or ridge regression is better in prediction.

- The lasso works better if the true model has a few non-zero parameters.
- The ridge regression works better if the true model has many non-zero but small parameters.

6.2.2 lasso

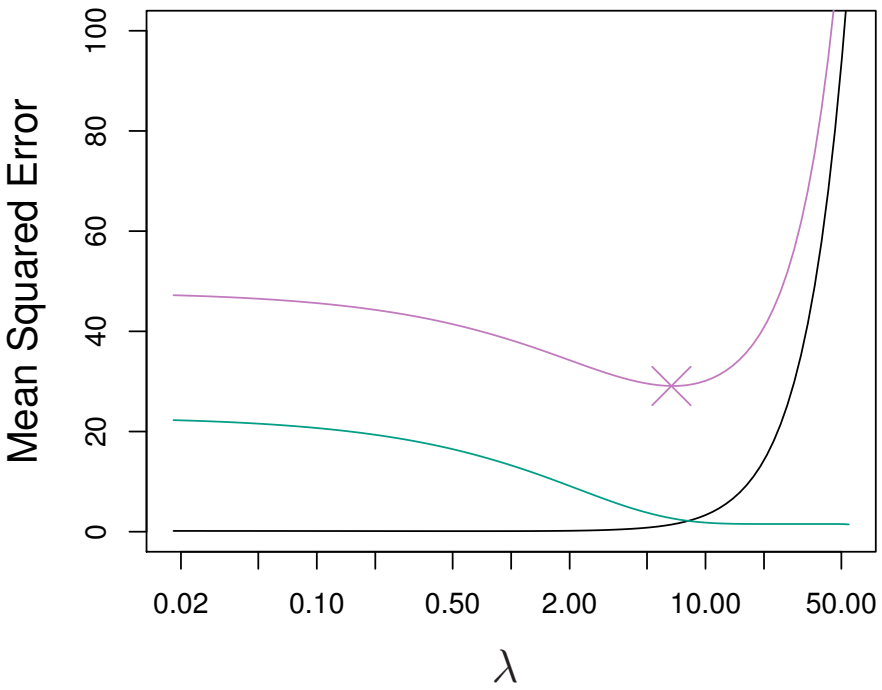
Figure 6-8: An example with many non-zero true parameters (Ridge is better). (Left:) squared bias (black), variance (green) and test MSE (purple). (Right:) MSE for lasso (solid) and for ridge regression (dotted).

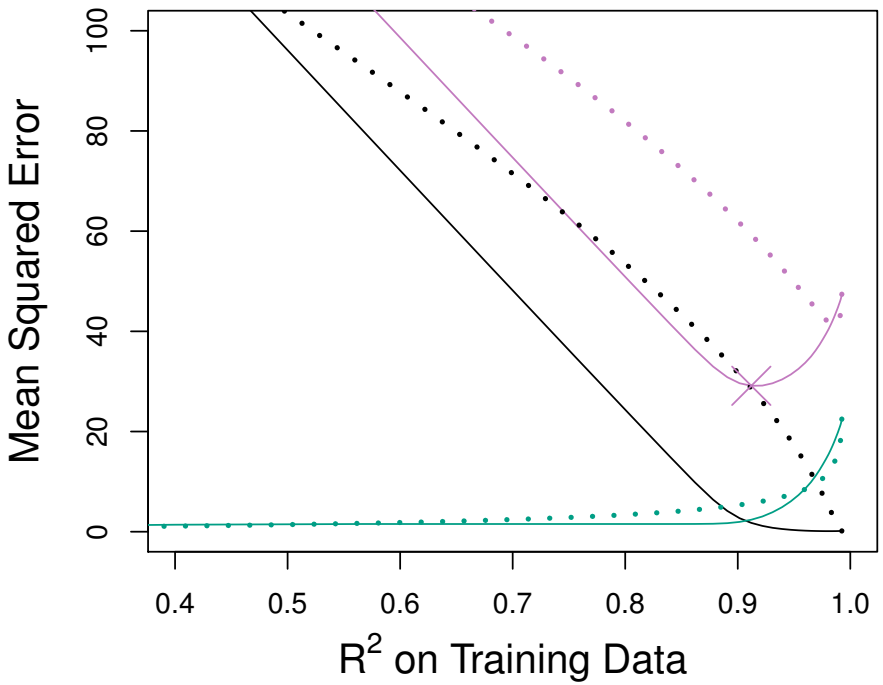




6.2.2 lasso

Figure 6-9: An example with a few non-zero parameters (lasso is better). (Left:) squared bias (black), variance (green) and test MSE (purple). (Right:) MSE for lasso (solid) and for ridge regression (dotted).





6.2.2 lasso

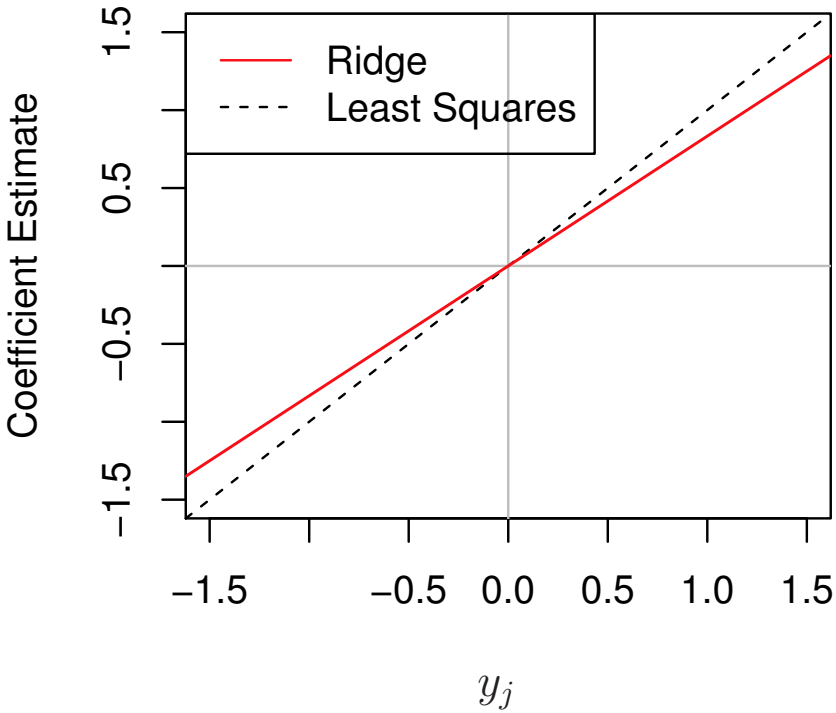
A simple example

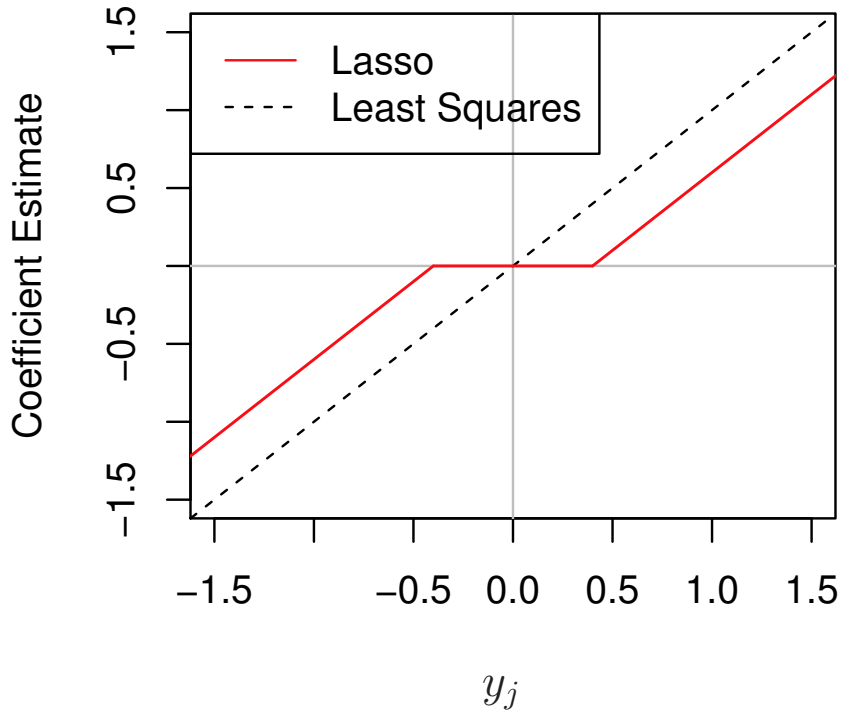
Consider a model $y_j = \beta_j + \epsilon_j$ ($1 \leq j \leq n$). There are n parameters, so $p = n$.

Obviously, $\hat{\beta}_j = y_j$ in least-square regression How about lasso and ridge?

6.2.2 lasso

Figure 6-10: Ridge regression shrinks estimates globally (left); lasso shrinks estimates to zero only when y_j is near zero (right).





6.2.2 lasso

Bayesian Interpretation of ridge and lasso

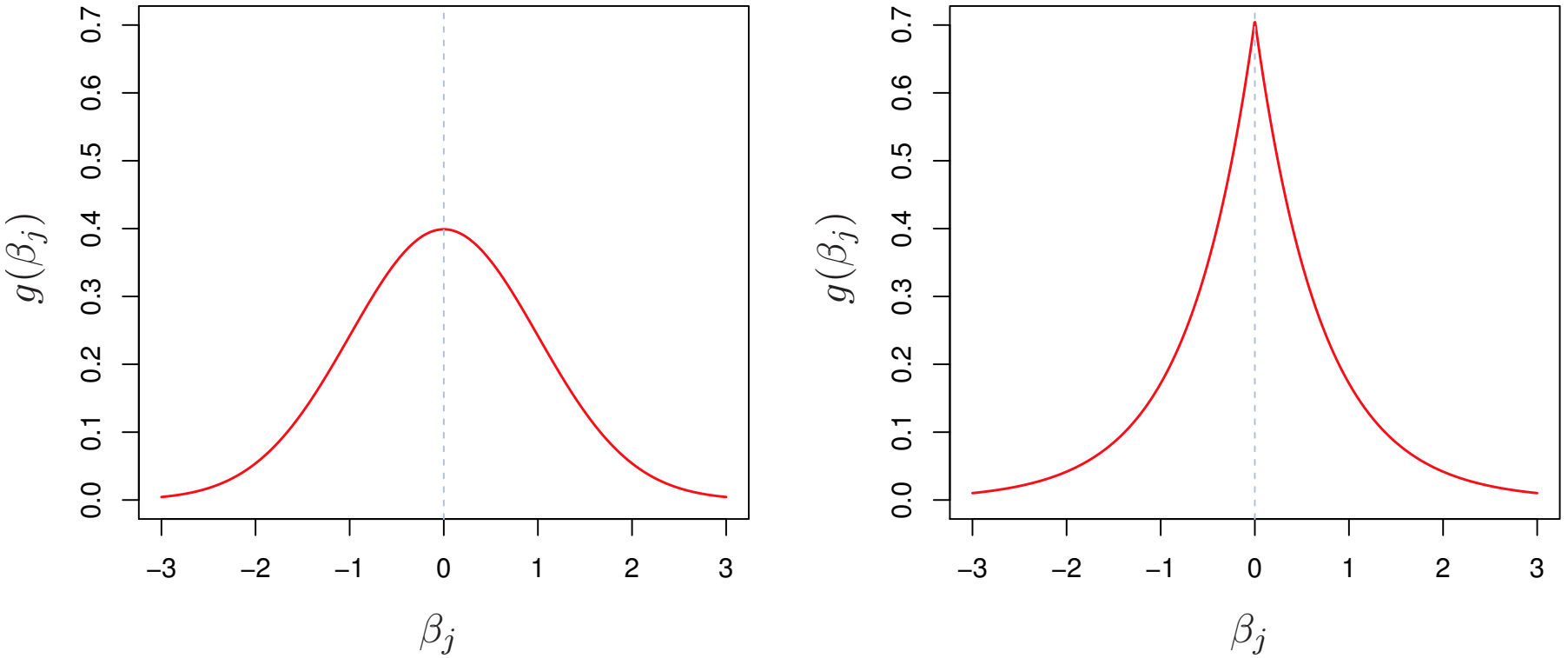
Bayesian statistics assumes that parameters has a prior distribution, which is an analyzer's subjective view on β prior to observing data.

In this context, the following interpretation is derived.

- Ridge regression assumes that β_j ($\mathbf{1 \leq j \leq p}$) is normally distributed with mean zero and variance inverse proportional to λ.
- The lasso assumes that β_j ($\mathbf{1 \leq j \leq p}$) is distributed as double-exponential (Laplace) with mean zero and standard deviation inverse proportional to λ.

6.2.2 lasso

Figure 6-11: Bayesian Interpretation of ridge (left) and lasso (right)



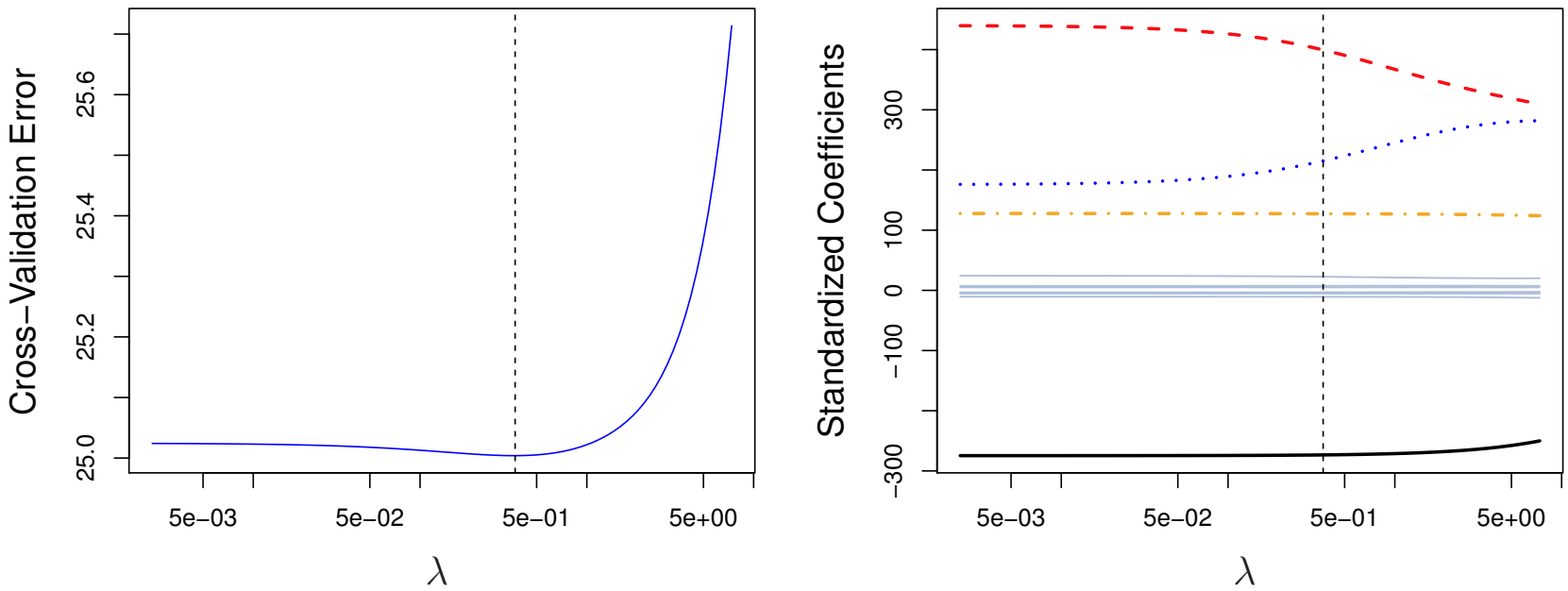
6.2.3 Selecting the tuning parameter

The tuning parameter λ is optimized by cross-validation.

Note: When test error is also evaluated by CV, you may need two nested CVs. Suppose that you have a training set and a test set during CV to evaluate test error. You may have to calculate λ within the training set, so you may have to cross-validate within the training set.

6.2.3 Selecting the tuning parameter

Figure 6-12: Cross-validation error by λ. Ridge (left) and lasso (right).



6.3 Dimension Reduction Methods

Suppose that there are many predictors $X_1, \cdots, X_p$.

To summarize the predictors and extract useful information, let $Z_1, \cdots, Z_M$ ($M < p$) linear combinations of the original predictors. Namely,

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

Then, we can fit the linear regression model:

$$y = \theta_0 + \sum_{i=1}^{M} \theta_m z_m + \epsilon.$$

We have better prediction performance if we choose $Z_1, \cdots, Z_M$ wisely.

This also reduces computational burden and avoids identification problem when $p > n$.

6.3.1 Principal Component Regression

Principal Component Analysis (PCA)

PCA is a technique for reducing the dimension of a $n \times p$ data matrix X. Since Y is not involved, this is a method for unsupervised learning.

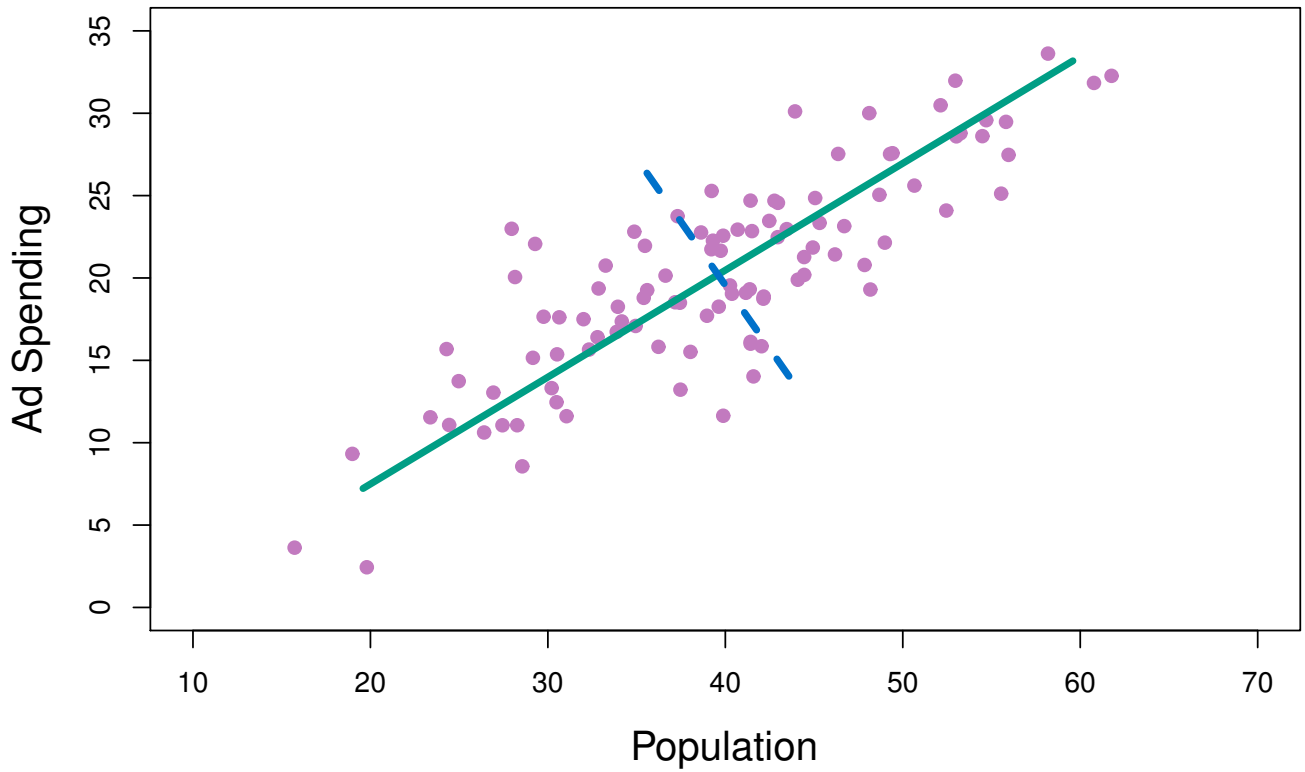
Each component $(Z_1, \cdots, Z_M)$ is defined as the projected position of each row vector of X on to a straight line.

The straight line for the 1st component (Z_1) is chosen so that the variance of the projected positions is maximized.

The straight line for the k-th $(k > 1)$ component (Z_k) is chosen so that the variance of the projected positions is maximized, subject to Z_k is orthogonal to $Z_1, \cdots, Z_{k-1}$.

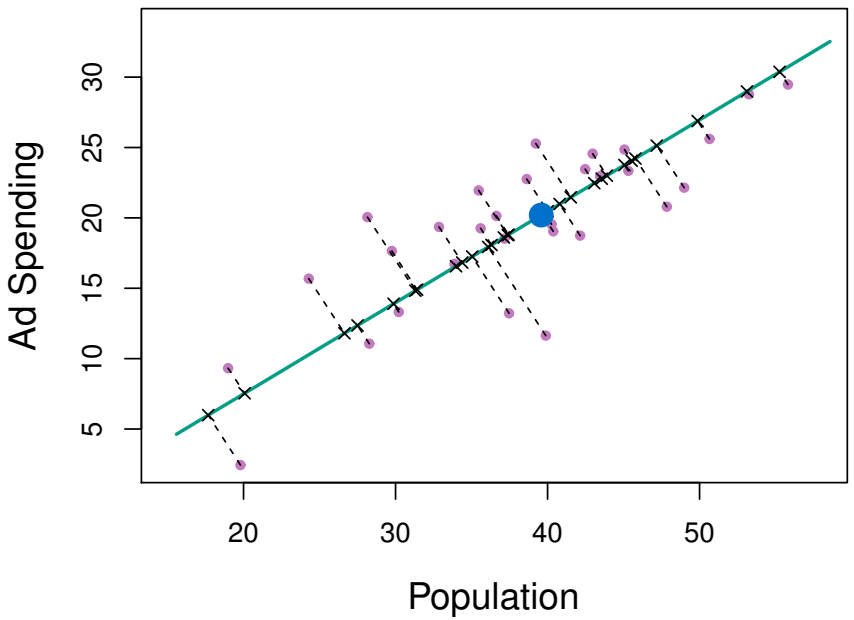
6.3.1 Principal Component Regression

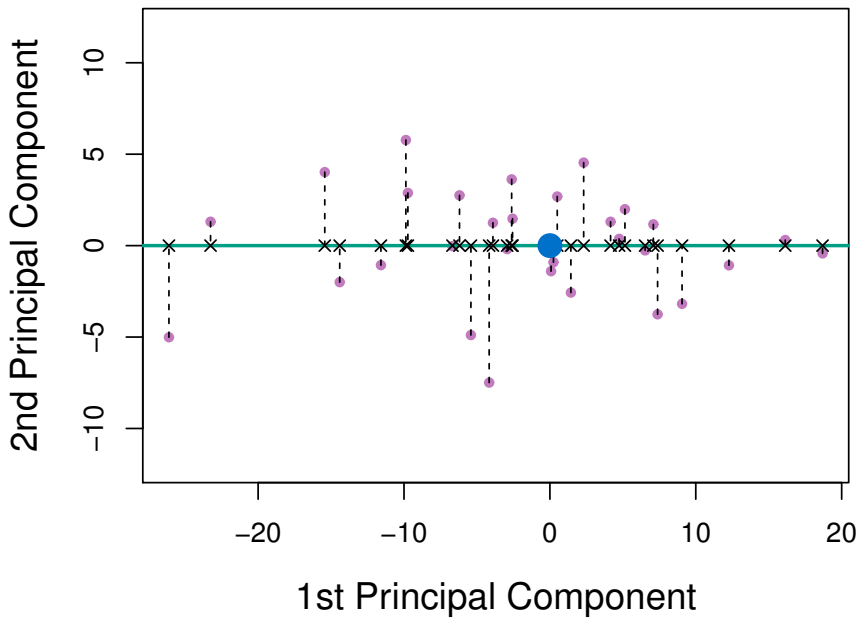
Figure 6-14: The line for the 1st component (green) and the 2nd component (blue dotted) of PCA for 2-dimensional data ($\boldsymbol{X}$ is $\boldsymbol{n} \times \mathbf{2}$).



6.3.1 Principal Component Regression

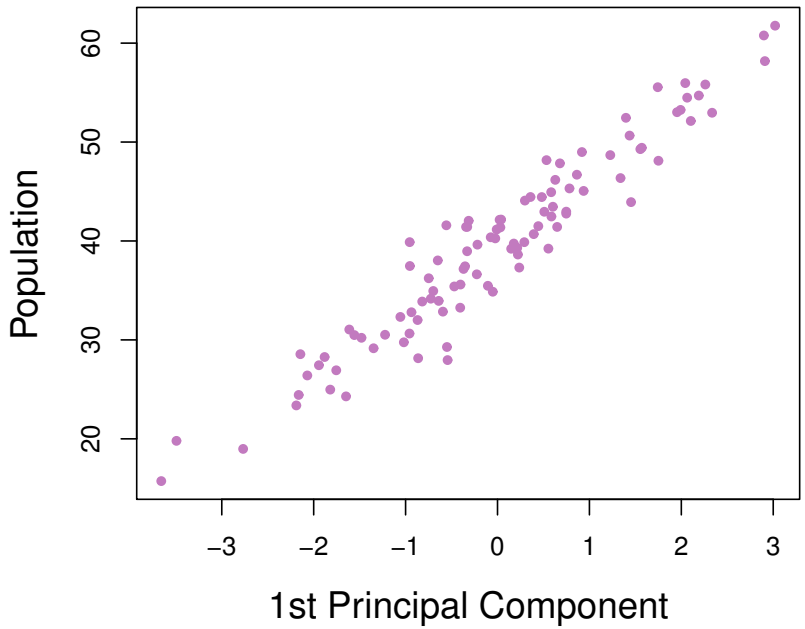
Figure 6-15: The 1st principal component (right) of PCA. The line for the 1st component is also interpreted as the line which makes closest distance to data points.

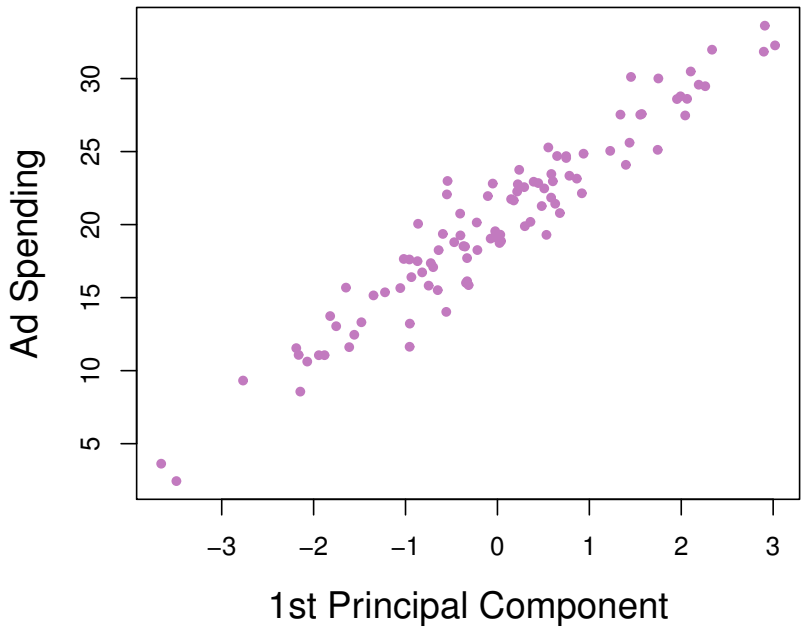




6.3.1 Principal Component Regression

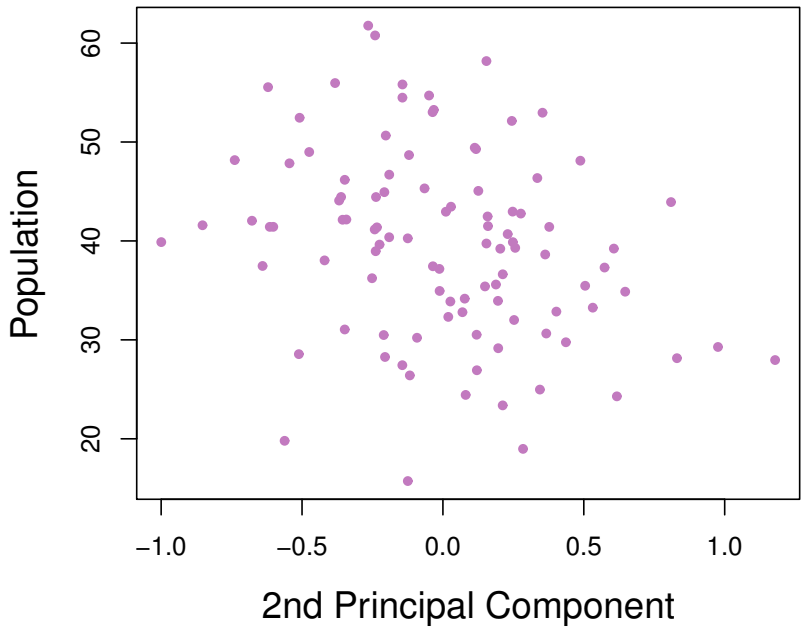
Fig 6-16: The 1st component Z_1 is usually highly correlated with most predictors. Z_1 is a sort of the representative of all variables.

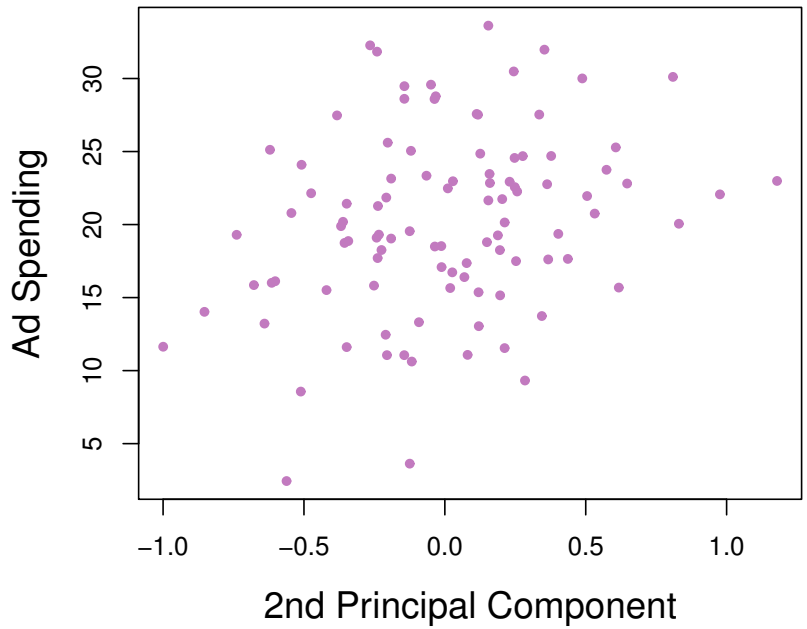




6.3.1 Principal Component Regression

Fig 6-17: The 2nd component Z_2 is not much correlated with predictors if $p = 2$.





6.3.1 Principal Component Regression

The principal components may and may not be interpreted in the context of data.

Example

Consider test scores for 1,000 students. There are five scores: English (E), Math (M), Science (S), Social Studies (SS) and French (F) for each students.

- The 1st component is often the overall performance of students.
  - E.g., $\mathbf{0.45}\boldsymbol{E} + \mathbf{0.45}\boldsymbol{M} + \mathbf{0.45}\boldsymbol{S} + \mathbf{0.45}\boldsymbol{SS} + \mathbf{0.45}\boldsymbol{F}$.
- The 2nd component may be the tendency towards scientific areas.
  - E.g., $\mathbf{-0.5}\boldsymbol{E} + \mathbf{0.5}\boldsymbol{M} + \mathbf{0.5}\boldsymbol{S} + \mathbf{0}\boldsymbol{SS} - \mathbf{0.5}\boldsymbol{F}$.

6.3.1 Principal Component Regression

A mathematical formulation

The projected position on to a straight line is better understood if inner product is used.

Let $\tilde{x}_1, \cdots, \tilde{x}_n$ be row vectors (observations) of X.

For the the 1st component of PCA, choose a p-dimensional unit vector $\mathbf{u}$ such that

$$\mathbf{Var}(\tilde{x} \cdot \mathbf{u})$$

is maximized.

6.3.1 Principal Component Regression

When each column of $\boldsymbol{X}$ has mean zero, this is equivalent to maximize

$$\boldsymbol{w}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{w}$$

and it is achieved when $\boldsymbol{w}$ is the eigenvector for the largest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$.

6.3.1 Principal Component Regression

Standardization

Usually the column vectors of $\boldsymbol{X}$ is standardized to have mean zero and variance one.

If standardization is not done,

- the principal components will be similar to a column vector with a larger variance, and
- the PCA is not scale-invariant.

6.3.1 Principal Component Regression

Principal Component Regression (PCR)

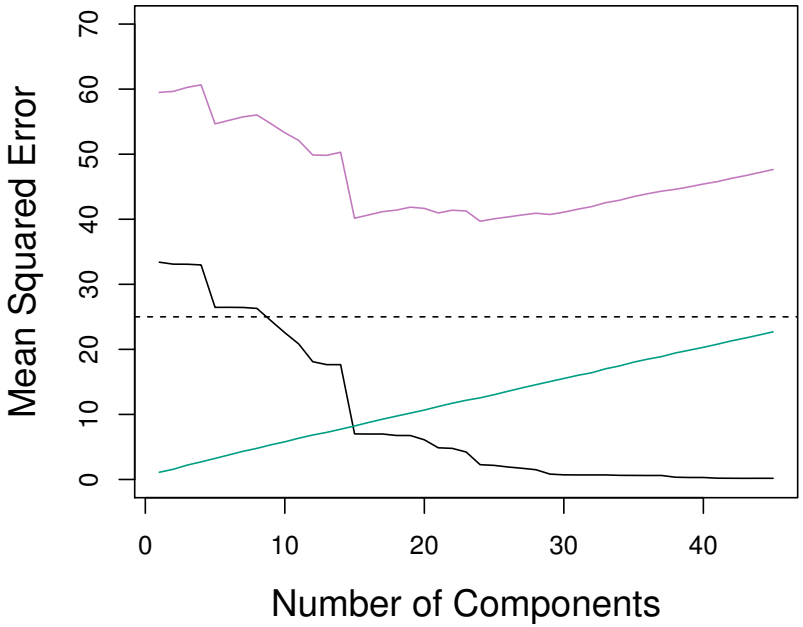
The PCR is the linear regression model:

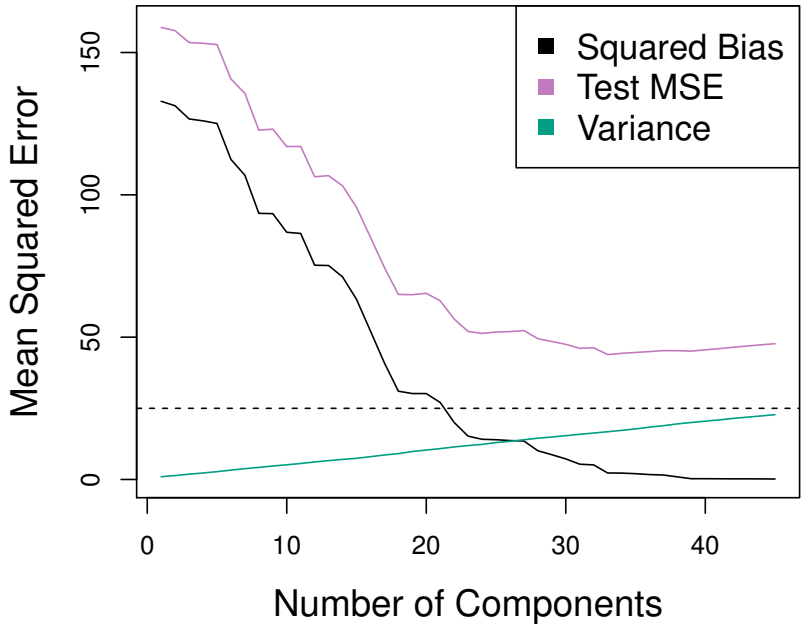
$$Y = \beta_0 + \beta_1 Z_1 + \cdots + \beta_M Z_M + \epsilon$$

- It often performs better than multiple linear regression since information is efficiently summarized.
- It can be combined with lasso or stepwise forward selection. That is, the principal components are used in lasso or stepwise regression.
- It avoids the parameter identification problem when $p > n$.

6.3.1 Principal Component Regression

Figure 6-18: Test MSE (purple) of PCR by the number of components used (***M***). (The green curve is variance, and the black is squared bias.)





6.3.2 Partial Least Squares

The PCA does not necessarily extract optimal information to predict $\boldsymbol{Y}$.

The partial least squares (PLS) an algorithm to extract useful information from $\boldsymbol{X}$ to predict $\boldsymbol{Y}$.
It can be used:

- when there are too many predictors so that multiple regression $\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}_1 + \cdots \beta_p \boldsymbol{X}_p + \epsilon$ does not work.
- when there are multiple response variables $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_q$ and we want to determine most useful information $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_M$ to predict all responses.

6.3.2 Partial Least Squares

Algorithm of PLS for one predictor Y

Suppose $X_1, \cdots X_p$ and Y are all standardized.

To determine Z_1,

- Regress Y on each X_j in simple linear regression: $Y = \phi_{j1} X_j + \epsilon$.
- Define $Z_1 := \sum_{j=1}^{p} \phi_{j1} X_j$.

To determine Z_2,

- Regress X_j on Z_1: $X_j = \alpha_j Z_1 + r_j$.
- Regress Y on each r_j: $Y = \phi_{j2} r_j + \epsilon$.
- Define $Z_2 := \sum_{j=1}^{p} \phi_{j2} r_j$.

Continue the same procedure (replace X_j by r_j, Z_1 by Z_{k-1}) to get Z_k ($k \geq 2$).

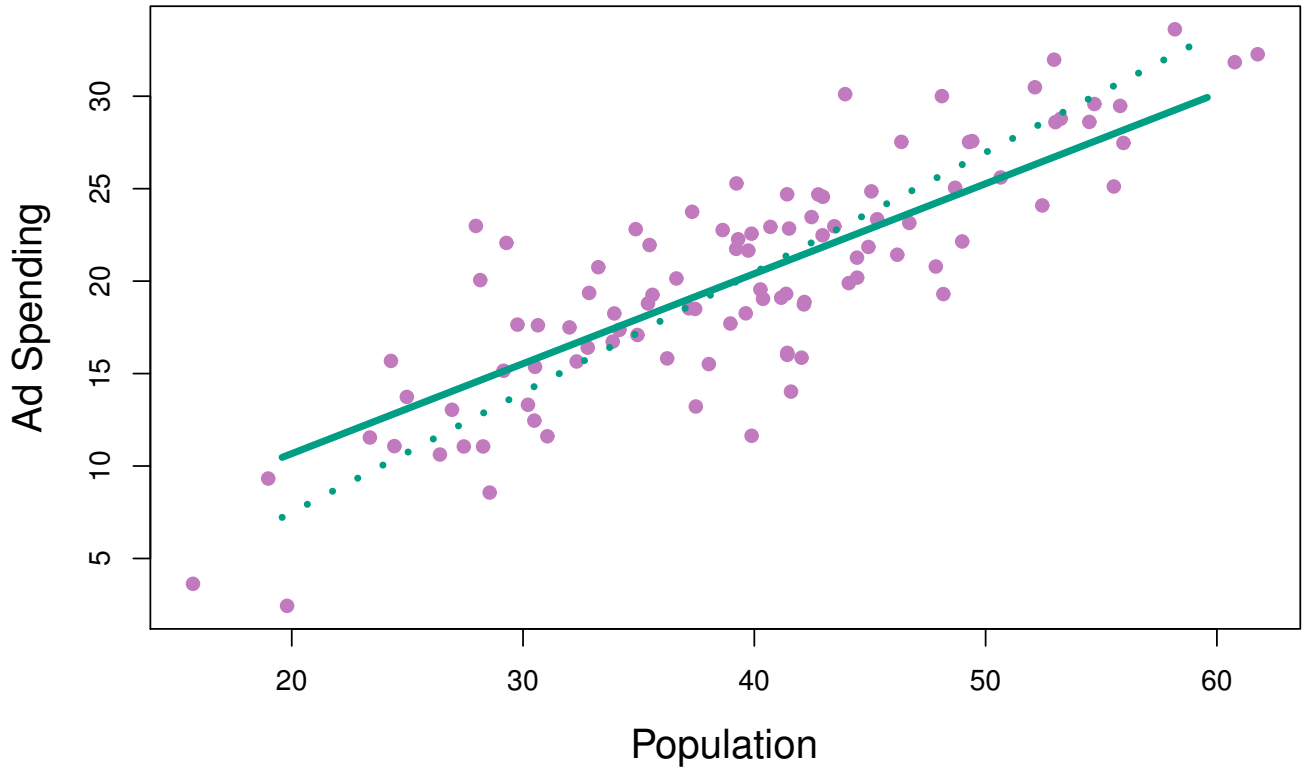
6.3.2 Partial Least Squares

Once we get $Z_1, \cdots, Z_M$, do multiple linear regression:

$$Y = \beta_0 + \beta_1 Z_1 + \cdots + \beta_M Z_M + \epsilon$$

6.3.2 Partial Least Squares

Figure 6-21: The 1st component of PLS (solid line) and the 1st compoent of PCA (dotted line). PLS put more weight on `population`.



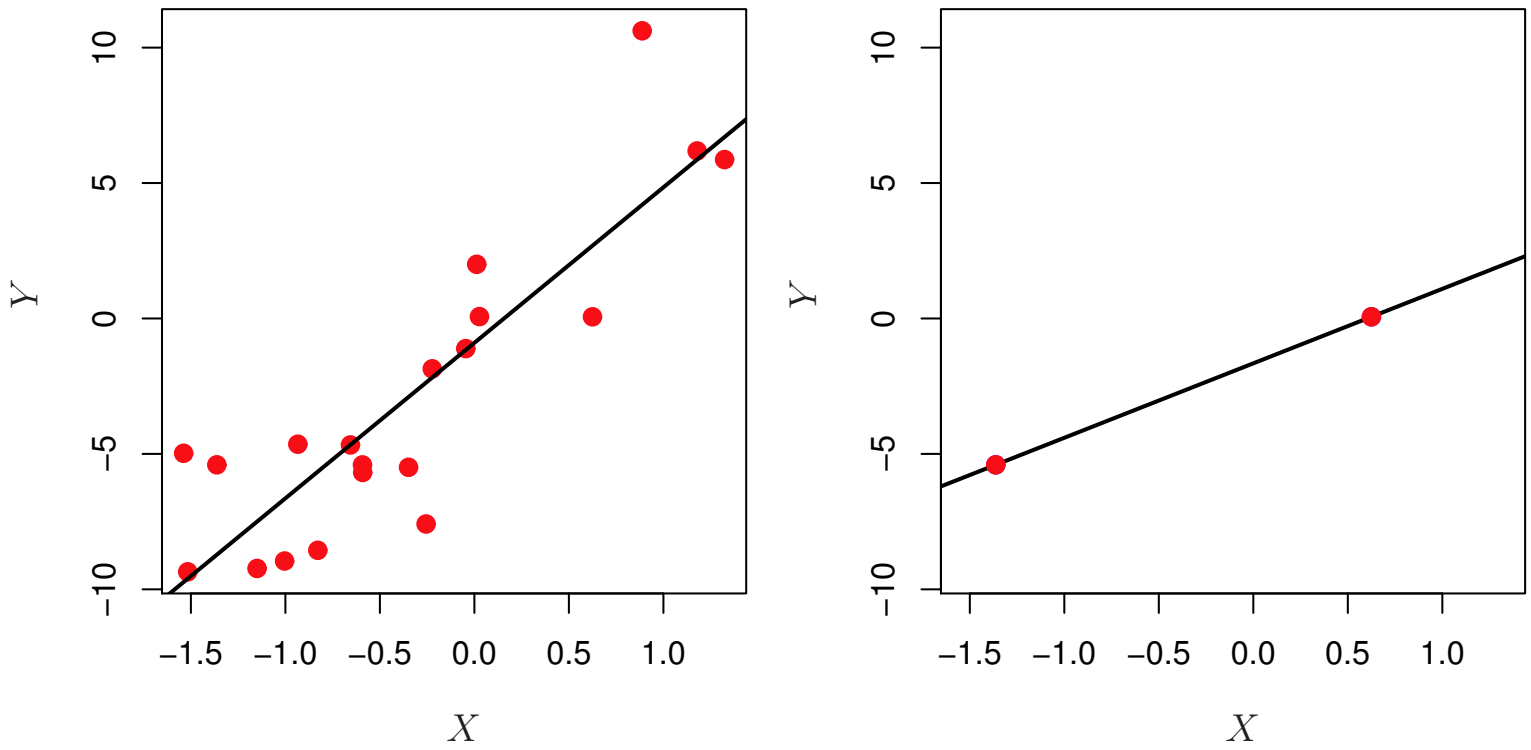
6.4.1 High-Dimensional Data

Recall that n is the sample size, and p is the number of predictors.

- Traditional statistical models deals with cases with $n >> p$.
  - E.g., Estimate blood pressure BP by age, gender and BMI (body mass index) with a sample of 1,000 people. $n = 1,000$ and $p = 3$.
  - Typically takes time order of $O(n)$.
  - Parameters have some asymptotic properties. As $n \to \infty$, we can approach to the truth. Parameter estimates $\hat{\beta}$ approaches to the true parameter β.
- High-dimensional data means $p >> n$.
  - E.g., Estimate blood pressure BP by 500,000 DNA mutations with 200 people. $n = 200$ and $p = 500,000$.
  - Parameters can not be identified correctly.
  - Increasingly computer-intensive. For linear regression time order of $O(p^3)$.

6.4.2 What Goes Wrong in High Dimensions?

Figure 6-22: Linear regression $Y = \beta_0 + \beta_1 X + \epsilon$ for $n = 20$ (left) and $n = 2$ (right). The error becomes zero if $p \geq n$.



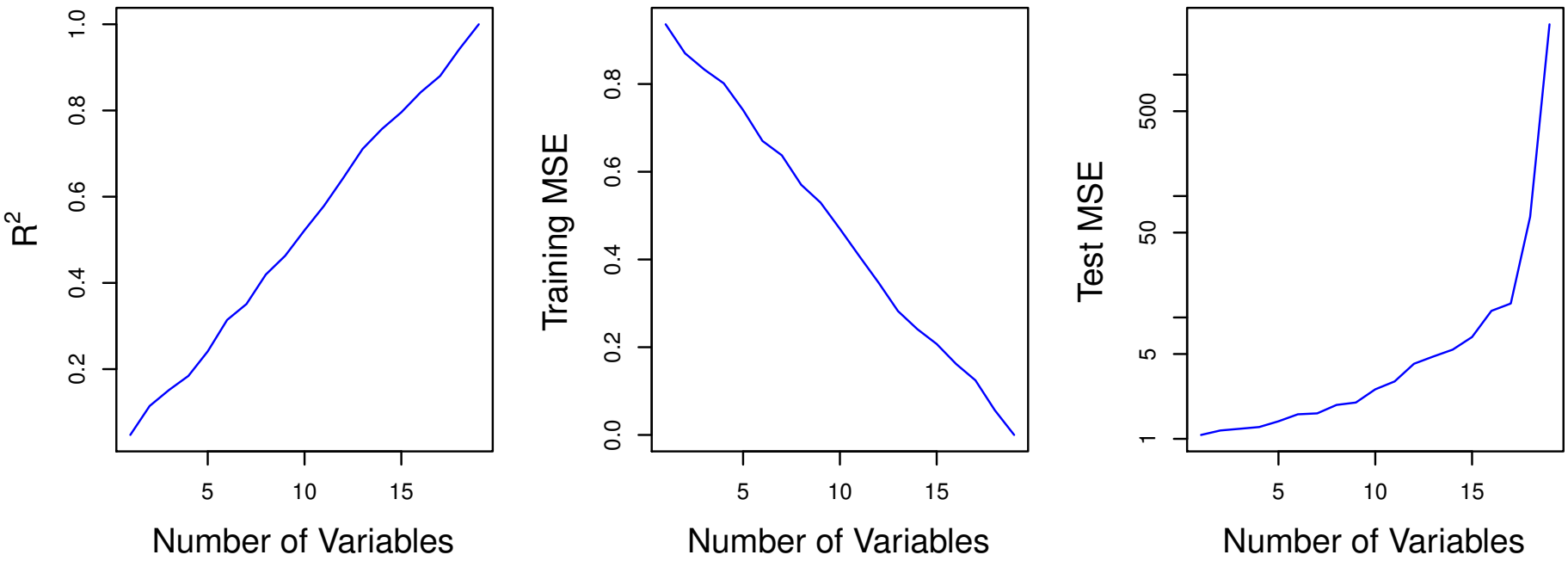
6.4.2 What Goes Wrong in High Dimensions?

Methods to select the best prediction model:

- Simulation-based methods such as CV and validation set approach works good.
- AIC, Mallow's C_p BIC are unstable, due to unreliable estimate of σ^2.
- Training error, adjusted R^2 and other traditional statistics for training set do not work. It always select a large model which makes zero training error.

6.4.2 What Goes Wrong in High Dimensions?

Fig 6-23: Fitting a regression model with $n = 20$. Predictors $X_1, \cdots, X_{20}$ are random numbers so that they are totally uncorrelated to Y. A larger p makes a larger R^2 and a smaller training MSE, but a larger test MSE.



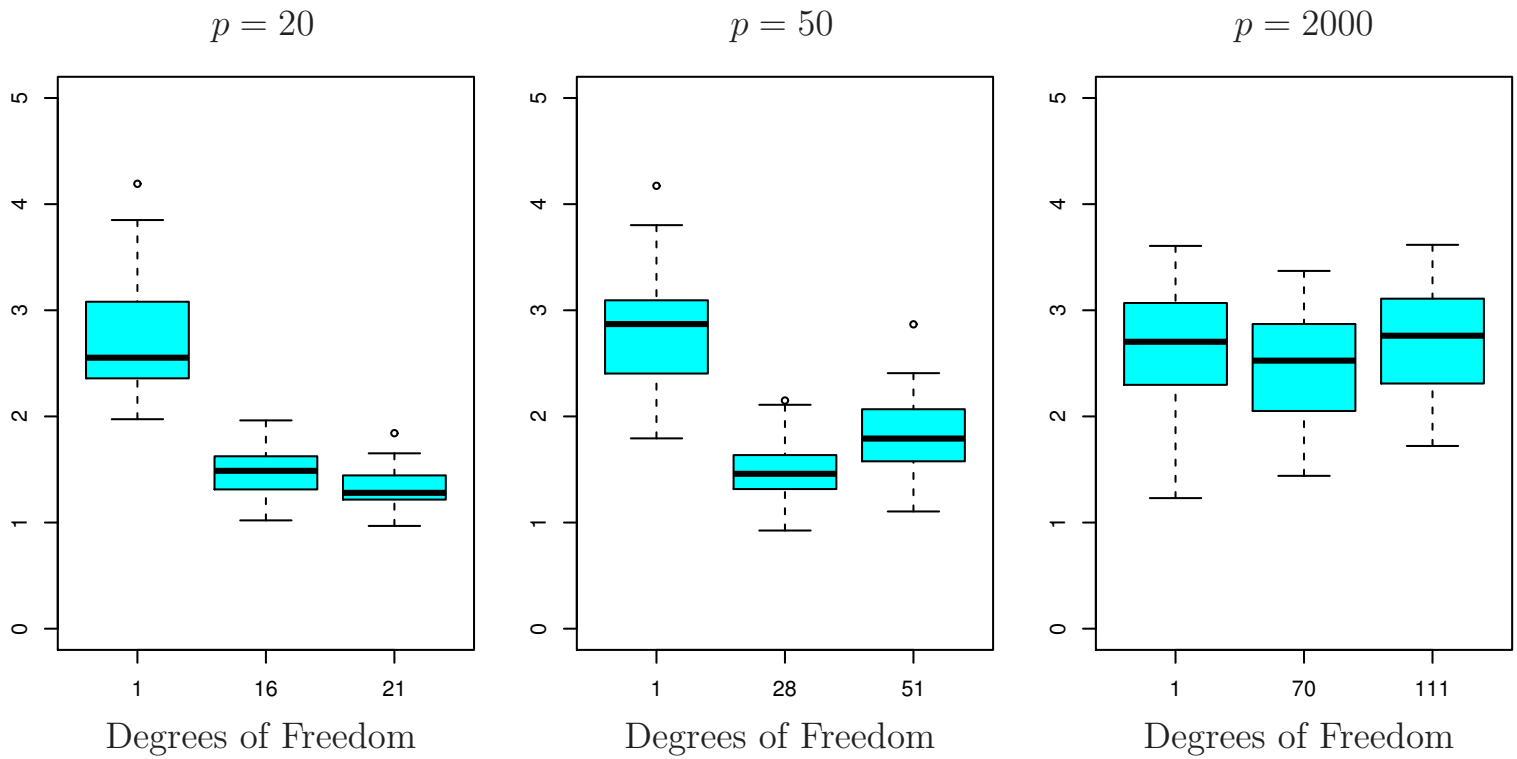
6.4.3 Regression in High Dimensions

Consider there are 20 true predictors to explain the response $\boldsymbol{Y}$, but they are included in 20, 50 or 2000 candidate variables of predictors ($\boldsymbol{p = 20, 50,}$ or $\boldsymbol{2000}$).

We do not know which variables are true predictors, so consider a variable selection problem for $\boldsymbol{p = 20, 50, 2000}$ by lasso.

6.4.3 Regression in High Dimensions

Figure 6-24: Test MSE by the number of predictors for $\boldsymbol{p = 20}$ (left), $\boldsymbol{p = 50}$ (center) and $\boldsymbol{p = 2000}$ (right). The number of predictors is the degrees of freedom minus 1. The optimal degrees of freedom to minimize test MSE is 21 for $\boldsymbol{p = 20}$ (correct), 28 for $\boldsymbol{p = 50}$ (7 larger than the true model) and 70 for $\boldsymbol{p = 2000}$ (49 larger than the true model).



6.4.3 Regression in High Dimensions

The lasso selected more variables than the true set of predictors. This is called the curse of dimensionality.

When the true predictors are hidden in a “forest of predictors”, it is harder to identify all of them. The lasso chooses more variables to make sure that all important information is included in the model.

6.4.4 Interpreting results in High Dimensions

For high-dimensional data, even with an appropriate variable selection procedure such as CV,

- multicollinearity is severe,
- the selected set of predictors may not be a unique best set.
  - For example, two set of predictors $\{X_1, X_5, X_{10}, X_{13}\}$ and $\{X_2, X_8, X_{11}, X_{15}\}$ may be as good for prediction.

Memo