Linear, Ridge Regression, and Principal Component Analysis

Jia Li

Department of Statistics
The Pennsylvania State University

Email: jiali@stat.psu.edu
http://www.stat.psu.edu/∼jiali

Introduction to Regression

- Input vector: $X = (X_1, X_2, ..., X_p)$.
- Output Y is real-valued.
- Predict Y from X by $f(X)$ so that the expected loss function

$$E(L(Y, f(X)))$$

  is minimized.
- Square loss:

$$L(Y, f(X)) = (Y - f(X))^2 .$$

- The optimal predictor

$$\begin{aligned} f^*(X) &= argmin_{f(X)} E(Y - f(X))^2 \\ &= E(Y \mid X) . \end{aligned}$$

- The function $E(Y \mid X)$ is the *regression function*.

Example

The number of active physicians in a Standard Metropolitan Statistical Area (SMSA), denoted by Y, is expected to be related to total population (X_1, measured in thousands), land area (X_2, measured in square miles), and total personal income (X_3, measured in millions of dollars). Data are collected for 141 SMSAs, as shown in the following table.

i :	1	2	3	...	139	140	141
X_1	9387	7031	7017	...	233	232	231
X_2	1348	4069	3719	...	1011	813	654
X_3	72100	52737	54542	...	1337	1589	1148
Y	25627	15389	13326	...	264	371	140

Goal: Predict Y from X_1, X_2, and X_3.

Linear Methods

- The linear regression model

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \ .$$

- What if the model is not true?
  - It is a good approximation
  - Because of the lack of training data/or smarter algorithms, it is the most we can extract robustly from the data.
- Comments on X_j:
  - Quantitative inputs
  - Transformations of quantitative inputs, e.g., $\log(\cdot)$, $\sqrt{(\cdot)}$.
  - Basis expansions: $X_2 = X_1^2$, $X_3 = X_1^3$, $X_3 = X_1 \cdot X_2$.

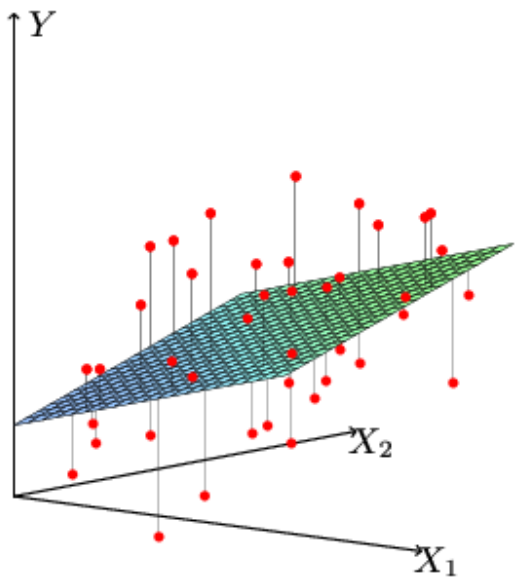


Figure 3.1: *Linear least squares fitting with* $X \in \mathbb{R}^2$. *We seek the linear function of* X *that minimizes the sum of squared residuals from* Y.

Estimation

- The issue of finding the regression function $E(Y \mid X)$ is converted to estimating β_j, $j = 0, 1, ..., p$.
- Training data:

$$\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\},$$
$$\text{where } x_i = (x_{i1}, x_{i2}, ..., x_{ip}) .$$

- Denote $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$.
- The loss function $E(Y - f(X))^2$ is approximated by the empirical loss $RSS(\beta)/N$:

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - f(x_i))^2 = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 .$$

Notation

- The input matrix $\mathbf{X}$ of dimension $N \times (p+1)$:

$$\begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,p} \end{pmatrix}$$

- Output vector $\mathbf{y}$:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}$$

- The estimated β is $\hat{\beta}$.
- The fitted values at the training inputs:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$$

and

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_N \end{pmatrix}$$

Point Estimate

- The *least square estimation* of $\hat{\beta}$ is

$$\boxed{\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

- The fitted value vector is

$$\boxed{\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

- Hat matrix:

$$\boxed{\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}$$

Geometric Interpretation

- Each column of $\mathbf{X}$ is a vector in an N-dimensional space (NOT the p-dimensional feature vector space).

$$\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_p)$$

- The fitted output vector $\hat{\mathbf{y}}$ is a linear combination of the column vectors $\mathbf{x}_j$, $j = 0, 1, ..., p$.
- $\hat{\mathbf{y}}$ lies in the subspace spanned by $\mathbf{x}_j$, $j = 0, 1, ..., p$.
- $RSS(\hat{\beta}) = \| \mathbf{y} - \hat{\mathbf{y}} \|^2$.
- $\mathbf{y} - \hat{\mathbf{y}}$ is perpendicular to the subspace, i.e., $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ on the subspace.
- The geometric interpretation is very helpful for understanding coefficient shrinkage and subset selection.

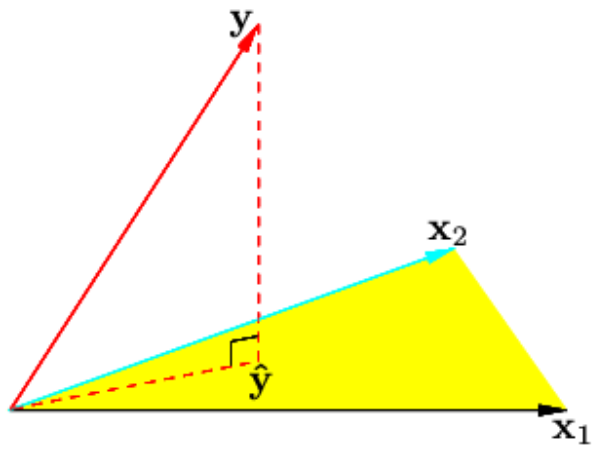
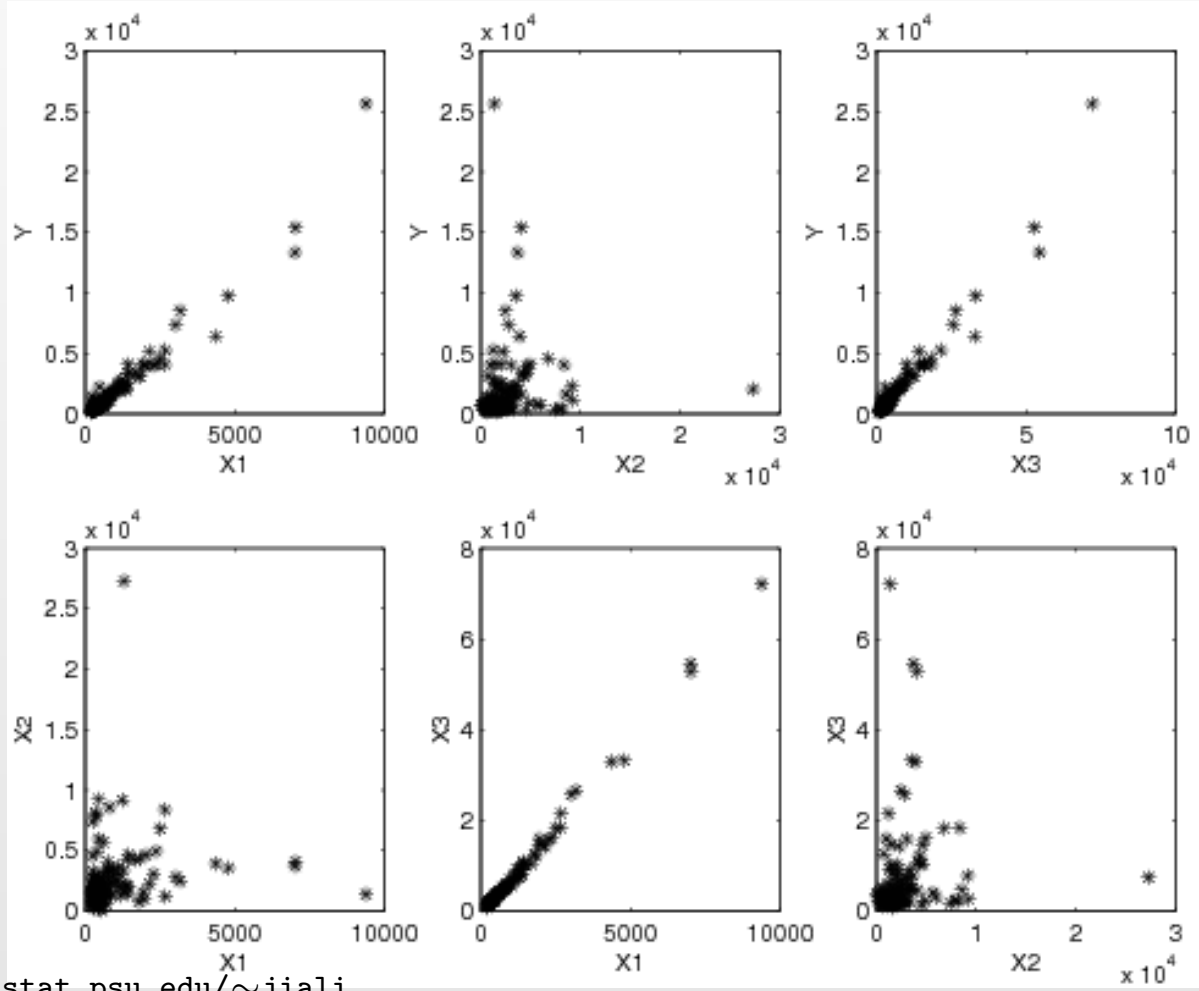


Figure 3.2: *The N-dimensional geometry of least squares regression with two predictors. The outcome vector* $\mathbf{y}$ *is orthogonally projected onto the hyperplane spanned by the input vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. The projection* $\hat{\mathbf{y}}$ *represents the vector of the least squares predictions*

Example Results for the SMSA Problem

- $\hat{Y}_i = -143.89 + 0.341X_{i1} - 0.0193X_{i2} + 0.255X_{i3}$.
- $RSS(\hat{\beta}) = 52942336$.



If the Linear Model Is True

- $E(Y \mid X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$
- The least square estimation of β is unbiased,

$$E(\hat{\beta}_j) = \beta_j \quad j = 0, 1, ..., p \ .$$

- To draw inferences about β, further assume:

$$Y = E(Y \mid X) + \epsilon$$

  where $\epsilon \sim N(0, \sigma^2)$ and is independent of X.
- X_{ij} are regarded as fixed, Y_i are random due to ϵ.
- Estimation accuracy: $Var(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$.
- Under the assumption, $\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$.
- Confidence intervals can be computed and significant tests can be done.

Gauss-Markov Theorem

- Assume the linear model is true.
- For any linear combination of the parameters $\beta_0,\ldots,\beta_p$, denoted by $\theta = a^T\beta$, $a^T\hat{\beta}$ is an unbiased estimation since $\hat{\beta}$ is unbiased.
- The least squares estimate of θ is

$$\begin{aligned}\hat{\theta} &= a^T\hat{\beta} \\ &= a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y} \\ &\triangleq \tilde{a}^T\mathbf{y}\,,\end{aligned}$$

which is linear in $\mathbf{y}$.

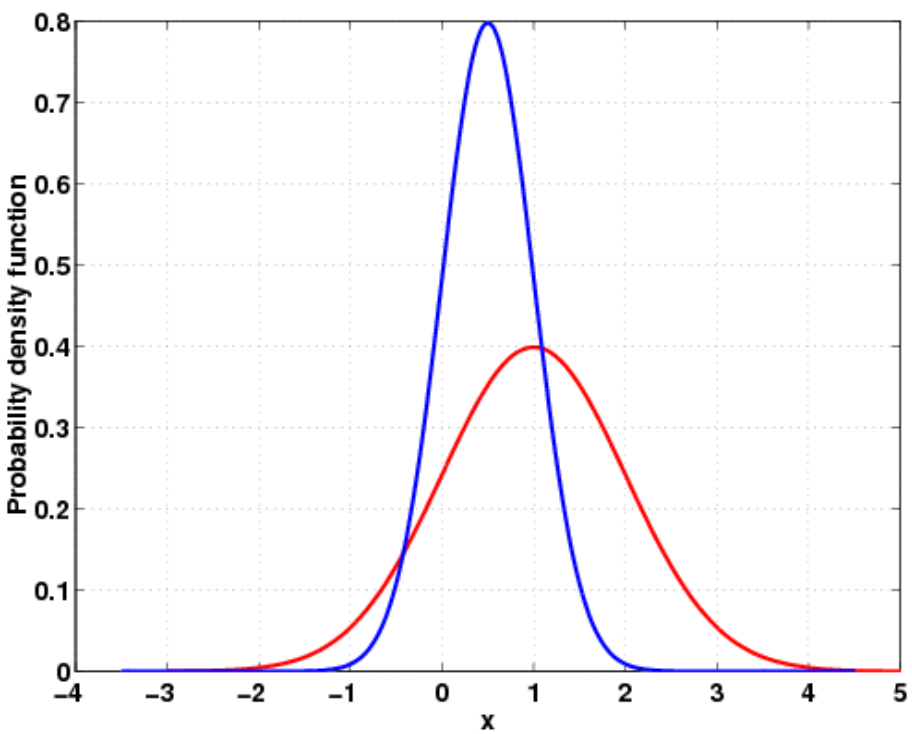
- Suppose $c^T\mathbf{y}$ is another `unbiased` linear estimate of θ, i.e., $E(c^T\mathbf{y}) = \theta$.
- The least square estimate yields the minimum variance among all `linear unbiased` estimate.

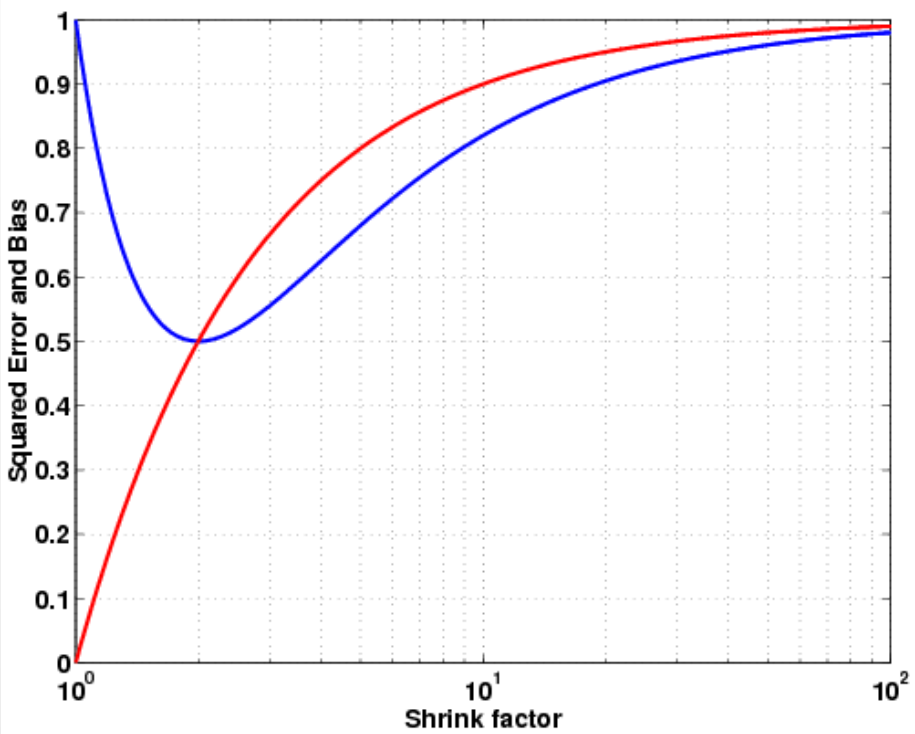
$$Var(\tilde{a}^T\mathbf{y}) \leq Var(c^T\mathbf{y}) \ .$$

- β_j, $j = 0, 1, ..., p$ are special cases of $a^T\beta$, where a^T only has one non-zero element that equals 1.

Subset Selection and Coefficient Shrinkage

- Biased estimation may yield better prediction accuracy.
- Squared loss: $E(\hat{\beta}-1)^2 = Var(\hat{\beta})$. For $\tilde{\beta} = \frac{\hat{\beta}}{a}$, $a \geq 1$, $E(\tilde{\beta}-1)^2 = Var(\tilde{\beta}) + (E(\tilde{\beta})-1)^2 = \frac{1}{a^2} + (\frac{1}{a}-1)^2$.
- Practical consideration: interpretation. Sometimes, we are not satisfied with a "black box".





Assume $\hat{\beta} \sim N(1, 1)$. The squared error loss is reduced by shrinking the estimation.

Subset Selection

- To choose k predicting variables from the total of p variables, search for the subset yielding minimum $RSS(\hat{\beta})$.
- *Forward stepwise selection*: start with the intercept, then sequentially adds into the model the predictor that most improves the fit.
- *Backward stepwise selection*: start with the full model, and sequentially deletes predictors.
- How to choose k: stop forward or backward stepwise selection when no predictor produces the F-ratio statistic greater than a threshold.

Ridge Regression

Centered inputs

- Suppose $\mathbf{x}_j$, $j = 1, ..., p$, are mean removed.
- $\hat{\beta}_0 = \bar{y} = \sum_{i=1}^{N} y_i / N$.
- If we remove the mean of y_i, we can assume

$$E(Y \mid X) = \sum_{j=1}^{p} \beta_j X_j$$

- Input matrix $\mathbf{X}$ has p (rather than $p + 1$) columns.
- $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$
- $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

Singular Value Decomposition (SVD)

- If the column vectors of $\mathbf{X}$ are orthonormal, i.e., the variables X_j, $j = 1, 2, ..., p$, are uncorrelated and have unit norm.
  - $\hat{\beta}_j$ are the coordinates of $\mathbf{y}$ on the orthonormal basis $\mathbf{X}$.
- In general

$$\boxed{\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T}.$$

  - $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_p)$ is an $N \times p$ orthogonal matrix. $\mathbf{u}_j$, $j = 1, ..., p$ form an orthonormal basis for the space spanned by the column vectors of $\mathbf{X}$.
  - $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p)$ is an $p \times p$ orthogonal matrix. $\mathbf{v}_j$, $j = 1, ..., p$ form an orthonormal basis for the space spanned by the row vectors of $\mathbf{X}$.
  - $\mathbf{D} = diag(d_1, d_2, ..., d_p)$, $d_1 \geq d_2 \geq ... \geq d_p \geq 0$ are the singular values of $\mathbf{X}$.

Principal Components

- The sample covariance matrix of $\mathbf{X}$ is

$$\mathbf{S} = \mathbf{X}^T\mathbf{X}/N\,.$$

- Eigen decomposition of $\mathbf{X}^T\mathbf{X}$:

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= (\mathbf{U}\mathbf{D}\mathbf{V}^T)^T(\mathbf{U}\mathbf{D}\mathbf{V}^T)\\ &= \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T\\ &= \mathbf{V}\mathbf{D}^2\mathbf{V}^T\end{aligned}$$

- The eigenvectors of $\mathbf{X}^T\mathbf{X}$, $\mathbf{v}_j$, are called *principal component direction* of $\mathbf{X}$.

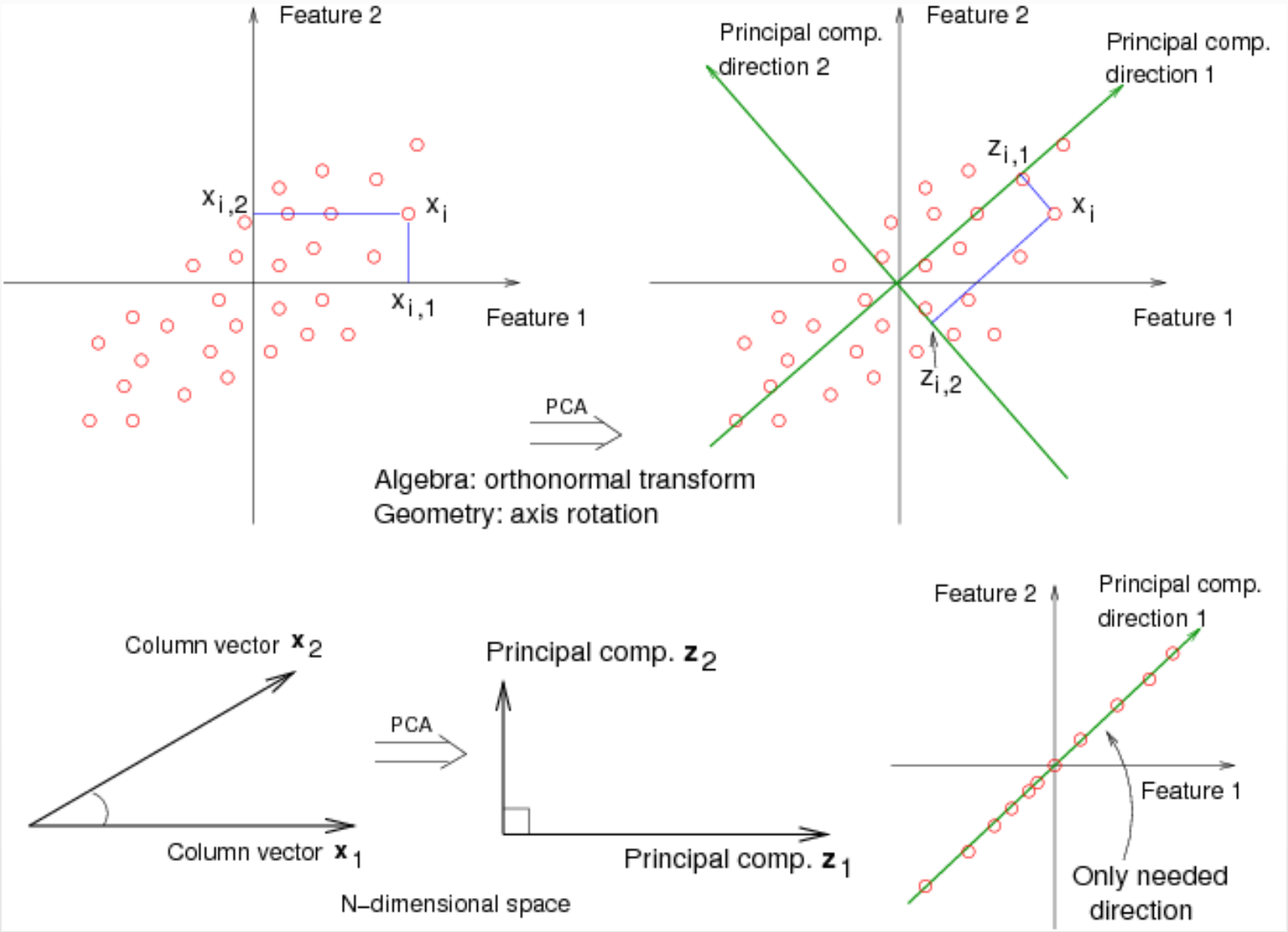
- It's easy to see that $\mathbf{z}_j = \mathbf{X}\mathbf{v}_j = \mathbf{u}_j d_j$. Hence $\mathbf{u}_j$, is simply the projection of the row vectors of $\mathbf{X}$, i.e., the input predictor vectors, on the direction $\mathbf{v}_j$, scaled by d_j. For example

$$\mathbf{z}_1 = \begin{pmatrix} X_{1,1}v_{1,1} + X_{1,2}v_{1,2} + \cdots + X_{1,p}v_{1,p} \\ X_{2,1}v_{1,1} + X_{2,2}v_{1,2} + \cdots + X_{2,p}v_{1,p} \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ X_{N,1}v_{1,1} + X_{N,2}v_{1,2} + \cdots + X_{N,p}v_{1,p} \end{pmatrix}$$

- The *principal components* of $\mathbf{X}$ are $\mathbf{z}_j = d_j\mathbf{u}_j$, $j = 1, ..., p$.
- The first principal component of $\mathbf{X}$, $\mathbf{z}_1$, has the largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{X}$.

$$Var(\mathbf{z}_1) = d_1^2/N \ .$$

- Subsequent principal components $\mathbf{z}_j$ have maximum variance d_j^2/N, subject to being orthogonal to the earlier ones.


Feature 2
$x_{i,2}$
x_i
$x_{i,1}$
Feature 1
PCA
Principal comp. direction 2
Principal comp. direction 1
$z_{i,1}$
$z_{i,2}$
Algebra: orthonormal transform
Geometry: axis rotation
Column vector $\mathbf{x}_2$
Column vector $\mathbf{x}_1$
Principal comp. $\mathbf{z}_2$
Principal comp. $\mathbf{z}_1$
N–dimensional space
Only needed direction

Ridge Regression

- Minimize a penalized residual sum of squares

$$\hat{\beta}^{ridge} = argmin_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- Equivalently

$$\hat{\beta}^{ridge} = argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq s \, .$$

- λ or s controls the model complexity.

Solution

- With centered inputs,

$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta \ ,$$

and

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

- Solution exists even when $\mathbf{X}^T\mathbf{X}$ is singular, i.e., has zero eigen values.
- When $\mathbf{X}^T\mathbf{X}$ is ill-conditioned (nearly singular), the ridge regression solution is more robust.

Geometric Interpretation

- Center inputs.
- Consider the fitted response

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\beta}^{ridge} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y} \\ &= \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda}\mathbf{u}_j^T\mathbf{y} \ ,\end{aligned}$$

  where $\mathbf{u}_j$ are the normalized principal components of $\mathbf{X}$.
- Ridge regression shrinks the coordinates with respect to the orthonormal basis formed by the principal components.
- Coordinate with respect to the principal component with a smaller variance is shrunk more.

- Instead of using $X = (X_1, X_2, ..., X_p)$ as predicting variables, use the transformed variables

$$(X\mathbf{v}_1, X\mathbf{v}_2, ..., X\mathbf{v}_p)$$

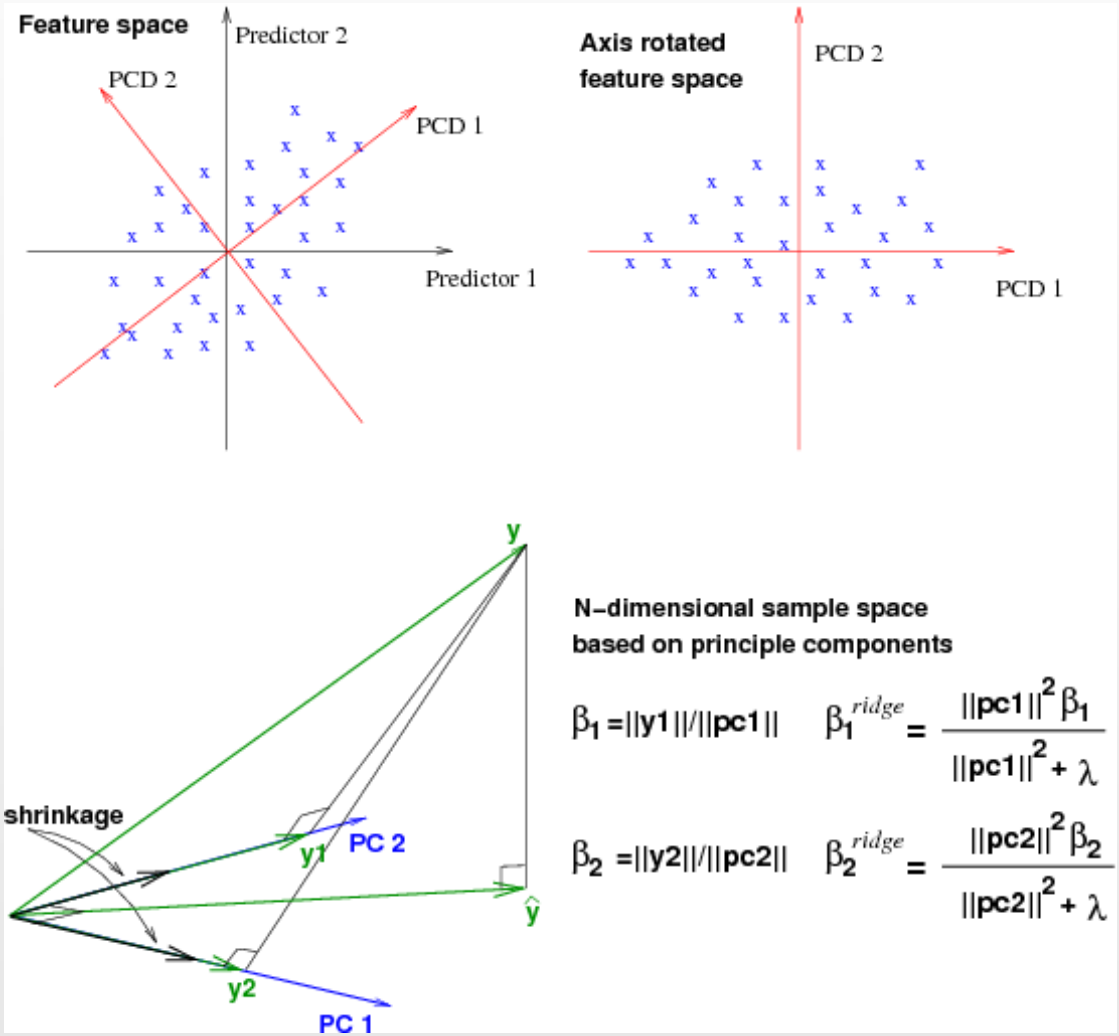
as predictors.
- The input matrix is $\tilde{\mathbf{X}} = \mathbf{UD}$ (Note $\mathbf{X} = \mathbf{UDV}^T$).
- Then for the new inputs

$$\hat{\beta}_j^{ridge} = \frac{d_j}{d_j^2 + \lambda}\mathbf{u}_j^T\mathbf{y}\ , \quad Var(\hat{\beta}_j) = \frac{\sigma^2}{d_j^2}$$

where σ^2 is the variance of the error term ϵ in the linear model.
- The factor of shrinkage given by ridge regression is

$$\frac{d_j^2}{d_j^2 + \lambda}\ .$$

The Geometric interpretation of principal components and shrinkage by ridge regression.



Compare squared loss $E(\beta_j - \hat{\beta}_j)^2$

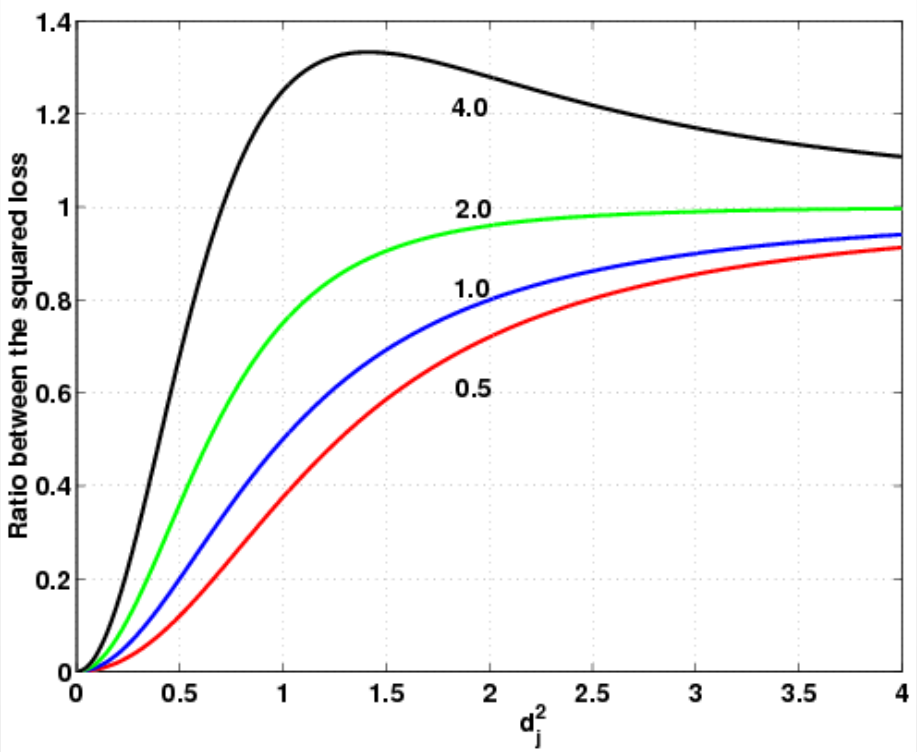
- Without shrinkage: σ^2/d_j^2.
- With shrinkage: $Bias^2 + Variance$.

$$(\beta_j - \beta_j \cdot \frac{d_j^2}{d_j^2+\lambda})^2 + \frac{\sigma^2}{d_j^2} \cdot (\frac{d_j^2}{d_j^2+\lambda})^2$$

$$= \frac{\sigma^2}{d_j^2} \cdot \frac{d_j^2(d_j^2 + \lambda^2 \frac{\beta_j^2}{\sigma^2})}{(d_j^2+\lambda)^2}$$

- Consider the ratio between squared loss

$$\frac{d_j^2(d_j^2 + \lambda^2 \frac{\beta_j^2}{\sigma^2})}{(d_j^2+\lambda)^2} \cdot$$



The ratio between the squared loss with and without shrinkage. The amount of shrinkage is set by $\lambda = 1.0$. The four curves correspond to $\beta^2/\sigma^2 = 0.5, 1.0, 2.0, 4.0$. When $\beta^2/\sigma^2 = 0.5, 1.0, 2.0$, shrinkage always leads to lower squared loss. When $\beta^2/\sigma^2 = 4.0$, shrinkage leads to lower squared loss when $d_j^2 \leq 0.71$. Shrinkage is more beneficial when d_j^2 is small.

Principal Components Regression (PCR)

- In stead of smoothly shrinking the coordinates on the principal components, PCR either does not shrink a coordinate at all or shrinks it to zero.
- Principal component regression forms the derived input columns $\mathbf{z}_m = \mathbf{X}\mathbf{v}_m$, and then regresses $\mathbf{y}$ on $\mathbf{z}_1$, $\mathbf{z}_2$, ..., $\mathbf{z}_M$ for some $M \leq p$.
- Principal components regression discards the $p - M$ smallest eigenvalue components.

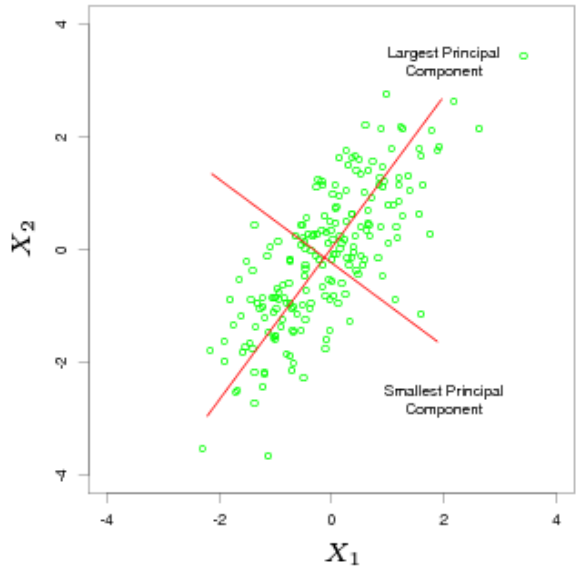


Figure 3.8: *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low-variance components more than the high-variance components.*

The Lasso

- The lasso estimate is defined by

$$\hat{\beta}^{lasso} = argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

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

- Comparison with ridge regression: L_2 penalty $\sum_{j=1}^{p} \beta_j^2$ is replaced by the L_1 lasso penalty $\sum_{j=1}^{p} |\beta_j|$.
- Some of the coefficients may be shrunk to exactly zero.
- Orthonormal columns in $\mathbf{X}$ are assumed in the following figure.

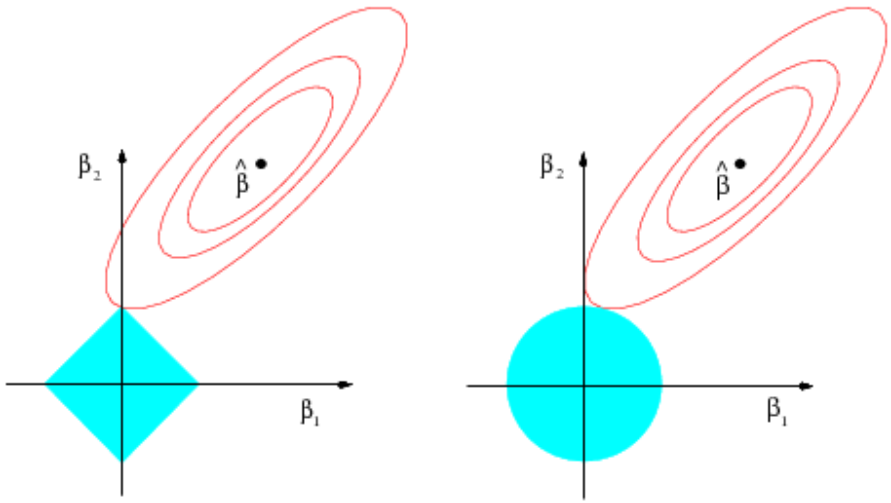


Figure 3.12: *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectively, while the red ellipses are the contours of the least squares error function.*