High Dimensional Data and Statistical Learning

Lecture 3: Linear Regression

Weixing Song

Department of Statistics
Kansas State University

Outline

Outline

Classical Linear Regression

Suppose

- Y: output or outcome variable;
- $X = (X_1, X_2, \ldots, X_p)^T$: input or feature variables;
- Model: $E(Y|X) = f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$.

The β_j's are unknown parameters /coefficients.

The input variables X_j's may come from various sources:

- quantitative inputs;
- transformations of original inputs, e.g., log or square-root transformation of length, area and volume;
- bases expansions, e.g., X_1, $X_1^2, \ldots$; Fourier or wavelet basis functions;
- dummy coding of qualitative (categorical) inputs, e.g. $X_1 = I(\text{non-smoker})$, and $X_2 = I(\text{smoker})$;
- interactions between variables, e.g., $X_3 = X_1 X_2$.

More generally, we model $E(Y|X)$ using linear combinations of fixed linear/nonlinear functions of the input variables. The key is that the model is linear in the parameters.

We assume the a set of i.i.d. input-output pairs

$$\mathcal{T}_n = \{(\mathbf{x}_{(i)}, y_i) : 1 \leq i \leq n\}$$

where $\mathbf{x}_{(i)} = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, are drawn from the following linear regression model

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i, \quad 1 \leq i \leq n.$$

The following notation will be repeatedly used in the sequel.

- Response: $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$;
- Predictors: $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$ for $j = 1, 2, \ldots, p$;
- Residuals: $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$;
- Design matrix: $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p)$ with $\mathbf{x}_0 = 1$;
- Regression coefficients: $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$.

Thus the linear regression model can be written as

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

Least Squares Estimation (LSE)

The most popular estimation method is LSE, in which we pick the coefficient $\boldsymbol{\beta}$ to minimize the residual sum of squares (RSS):

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

In the supervised learning problems as regression, we usually do not seek to model the distribution of the input variables. Thus the predictors always appear in the set of conditioning variables.

The above criterion is reasonable as long as y_i's are conditionally independent given the inputs $\mathbf{x}_{(i)}$.

Well-known Results of LSE

Suppose that $\text{rank}(\mathbf{X}) = p + 1$. Then LSE of $\boldsymbol{\beta}$ is given by

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

The fitted value of $\mathbf{y}$ is

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

Notes:

- The quantity $\mathbf{X}^+ = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is known as the Moore-Penrose pseudo-inverse of the matrix $\mathbf{X}$. It can be regarded as a generalization of the notion of matrix inverse to nonsquare matrices. Indeed, if $\mathbf{X}$ is square and invertible, then $\mathbf{X}^+ = \mathbf{X}^{-1}$.
- The qunatity $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is known as the "hat" matrix. What is $\text{trace}(\mathbf{H})$?
- The resulting estimate $\hat{y}$ is the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$.
- If $\mathbf{X}$ is not full rank, e.g., $p > n$, then the LS solution is not unique.

Assume that y_i's are uncorrelated, have constant variance σ^2, and $\mathbf{x}_{(i)}$ are fixed.

Then we can show that

- $E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ denotes the true value of the regression coefficient.
- $\text{Cov}(\boldsymbol{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.
- *Gauss-Markov Theorem:* Among all linear unbiased estimators of $a^T\boldsymbol{\beta}$, the least squares estimator $a^T\hat{\boldsymbol{\beta}}$ has the smallest variance. (a is a $(p+1)$ real valued vector)

  Proof:

Mean Squared Error (MSE): For an estimator $\hat{\theta}$ of θ,

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2.$$

Gauss-Markov Theorem implies that the LSE has the smallest MSE among all linear estimators with no bias.

Question: Is it possible to achieve a smaller MSE or better prediction accuracy by considering biased estimators?

Estimator of σ^2:

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})}{n - p - 1} = MSE$$

$\hat{\sigma}^2$ is an unbiased estimator of σ^2.

Proof:

To draw inferences about the parameters and the model, we now assume ε_i i.i.d. $\sim N(0, \sigma^2)$.

Then we have

- Maximum likelihood estimates of $\boldsymbol{\beta}$ an σ^2.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \quad \hat{\sigma}^2 = RSS/n.$$

- Distribution of $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^T\mathbf{X})^{-1}\right), \quad \frac{RSS}{\sigma^2} \sim \chi^2_{n-p-1}.$$

- The test statistic for checking $H_0 : L'\boldsymbol{\beta} = c$:

$$\frac{L'\hat{\boldsymbol{\beta}} - c}{\sqrt{MSE \cdot L'(\mathbf{X}^T\mathbf{X})^{-1}L}} \sim t_{n-p-1}.$$

Outline

Matrix Decomposition

QR decomposition.

Suppose an $n \times r$ matrix A satisfying $\text{rank}(A) = r$, then $A = QR$, where Q is an $n \times r$ matrix such that $Q^T Q = I$, and R is an $r \times r$ upper triangular matrix with positive diagonal elements.

Gram-Schmidt Orthogonalization.

Suppose $A = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r)$, where $\mathbf{a}_j$'s are $n \times 1$ vectors.

- Set $\mathbf{z}_1 = \mathbf{a}_1$;
- Set $\mathbf{z}_2 = \mathbf{a}_2 - \frac{<\mathbf{a}_2, \mathbf{z}_1>}{<\mathbf{z}_1, \mathbf{z}_1>}\mathbf{z}_1$;
- Set $\mathbf{z}_3 = \mathbf{a}_3 - \frac{<\mathbf{a}_3, \mathbf{z}_2>}{<\mathbf{z}_2, \mathbf{z}_2>}\mathbf{z}_2 - \frac{<\mathbf{a}_3, \mathbf{z}_1>}{<\mathbf{z}_1, \mathbf{z}_1>}\mathbf{z}_1$;
- $\ldots\ldots\ldots$;
- Set $\mathbf{z}_r = \mathbf{a}_r - \frac{<\mathbf{a}_r, \mathbf{z}_{r-1}>}{<\mathbf{z}_{r-1}, \mathbf{z}_{r-1}>}\mathbf{z}_{r-1} - \cdots - \frac{<\mathbf{a}_r, \mathbf{z}_1>}{<\mathbf{z}_1, \mathbf{z}_1>}\mathbf{z}_1$;

Let

$$\mathbf{a}_1 = \mathbf{z}_1; \quad \mathbf{a}_2 = \frac{< \mathbf{a}_2, \mathbf{z}_1 >}{< \mathbf{z}_1, \mathbf{z}_1 >}\mathbf{z}_1 + \mathbf{z}_2;$$

$$\cdots\cdots\cdots$$

$$\mathbf{a}_r = \frac{< \mathbf{a}_r, \mathbf{z}_1 >}{< \mathbf{z}_1, \mathbf{z}_1 >}\mathbf{z}_1 + \cdots + \frac{< \mathbf{a}_r, \mathbf{z}_{r-1} >}{< \mathbf{z}_{r-1}, \mathbf{z}_{r-1} >}\mathbf{z}_{r-1} + \mathbf{z}_r.$$

Define $\mathbf{e}_i = \mathbf{z}_i / \|\mathbf{z}_i\|$. Note that $\|\mathbf{z}_i\| =< \mathbf{a}_i, \mathbf{e}_i >$, then

$$\mathbf{a}_1 =< \mathbf{a}_1, \mathbf{e}_1 > \mathbf{e}_1; \quad \mathbf{a}_2 =< \mathbf{a}_2, \mathbf{e}_1 > \mathbf{e}_1 + < \mathbf{a}_2, \mathbf{e}_2 > \mathbf{e}_2;$$

$$\cdots\cdots\cdots$$

$$\mathbf{a}_r =< \mathbf{a}_r, \mathbf{e}_1 > \mathbf{e}_1 + \cdots + < \mathbf{a}_r, \mathbf{e}_{r-1} > \mathbf{e}_{r-1} + < \mathbf{a}_r, \mathbf{e}_r > \mathbf{e}_r.$$

This leads to

$$[\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_r] = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_r] \begin{pmatrix} < \mathbf{a}_1, \mathbf{e}_1 > & < \mathbf{a}_2, \mathbf{e}_1 > & \cdots & < \mathbf{a}_r, \mathbf{e}_1 > \\ 0 & < \mathbf{a}_2, \mathbf{e}_2 > & \cdots & < \mathbf{a}_r, \mathbf{e}_2 > \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & < \mathbf{a}_r, \mathbf{e}_r > \end{pmatrix}$$

$$\Downarrow \qquad\qquad \Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

$$A_{n\times r} \quad = \quad Q_{n\times r} \qquad\qquad\qquad R_{r\times r}$$

Singular Value Decomposition (SVD).

Let A be an $m \times n$ real matrix. There exist two orthonormal matrices $U_{m\times m}$, $V_{n\times n}$ such that

$$A = UDV^T,$$

where

$$D = \begin{pmatrix} \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_r) & 0 \\ 0 & 0 \end{pmatrix}_{m\times n}, \quad r = \text{rank}(A).$$

How to find U, V, D?

- U: the orthonormal eigenvectors of AA^T;
- V: the orthonormal eigenvectors of A^TA;
- σ_i^2's: the nonzero-eigenvalues of AA^T or A^TA.

Usually, we write

$$A = U_1 D_1 V_1^T$$

where U_1 consists of the first r columns of U, and V_1 consists of the first r columns of V, $D_1 = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_r)$.

Regression by Successive Orthogonalization

- Initialization: $\mathbf{z}_0 = \mathbf{x}_0 = 1$;
- Successive Orthogonalization: For $j = 1, 2, \ldots, p$,
  Regression $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{j-1}$ to produce coefficient
  $\gamma_{jl} =< \mathbf{x}_j, \mathbf{z}_l > / < \mathbf{z}_l, \mathbf{z}_l >$ for $l = 0, 1, \ldots, j-1$ and the residual vector
  $\mathbf{z}_j = \mathbf{x}_j - \sum_{l=0}^{j-1} \hat{\gamma}_{lj} \mathbf{z}_l$.
- Regress $\mathbf{y}$ on the residual $\mathbf{z}_p$ gives the LSE $\hat{\beta}_p$.

This is known as the Gram-Schmidt procedure for regression.

Questions:

- Why does this work?
- What does it imply?

An illustration

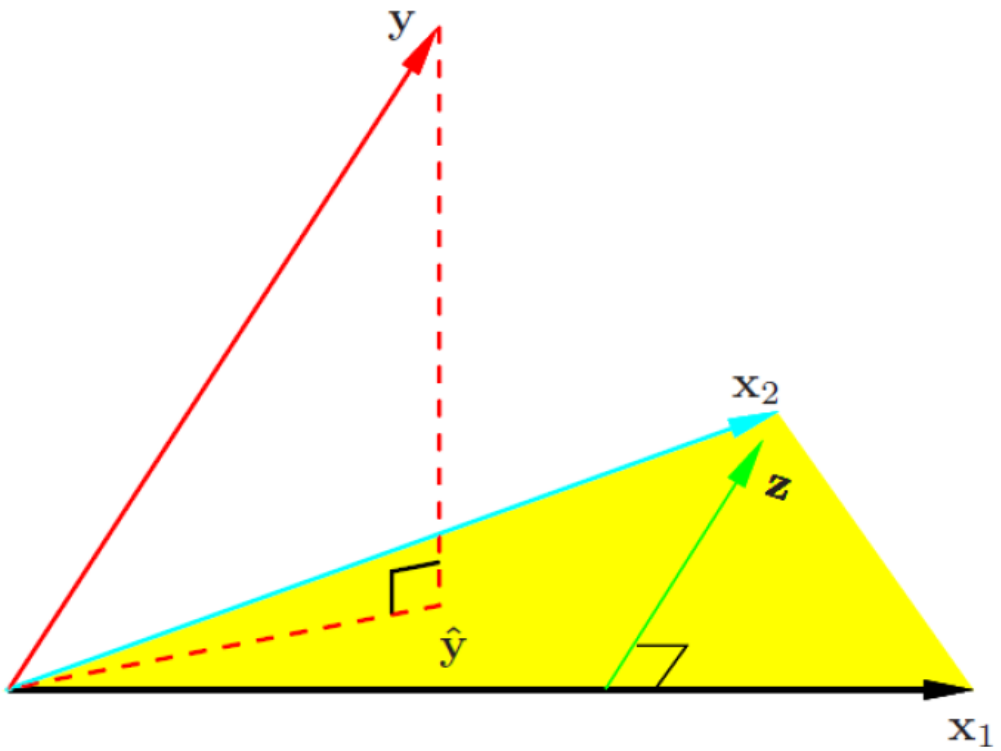


FIGURE 3.4. *Least squares regression by orthogonalization of the inputs. The vector* $\mathbf{x}_2$ *is regressed on the vector* $\mathbf{x}_1$*, leaving the residual vector* $\mathbf{z}$*. The regression of* $\mathbf{y}$ *on* $\mathbf{z}$ *gives the multiple regression coefficient of* $\mathbf{x}_2$*. Adding together the projections of* $\mathbf{y}$ *on each of* $\mathbf{x}_1$ *and* $\mathbf{z}$ *gives the least squares fit* $\hat{\mathbf{y}}$*.*

Some Remarks:

- The orthogonalization does not change the subspace spanned by the predictors;
- The $\mathbf{z}_j$ are all orthogonal, and they form a basis for the column space of $\mathbf{X}$;
- Regressing $\mathbf{y}$ on $\mathbf{z}_j$'s also produces the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$;
- Since $\mathbf{z}_p$ along involves $\mathbf{x}_p$, we see that regressing $\mathbf{y}$ on $\mathbf{z}_p$ produces the LSE of β_p;
- Moreover, this implies that the LSE $\hat{\beta}_j$ can be obtained by regressing $\mathbf{y}$ on $\mathbf{e}_j$, where $\mathbf{e}_j$ is the residual vector by regressing $\mathbf{x}_j$ on the other predictors;
- That is, the multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of $\mathbf{x}_j$ on $\mathbf{y}$, after $\mathbf{x}_j$ has been adjusted by the other predictors, including the intercept.

From the step of Successive Orthogonalization, it can be easily shown that $\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma}$, where

$$\mathbf{Z} = (\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_p), \quad \boldsymbol{\Gamma} = \begin{pmatrix} 1 & \hat{\gamma}_{10} & \hat{\gamma}_{20} & \cdots & \hat{\gamma}_{p-1,0} & \hat{\gamma}_{p,0} \\ 0 & 1 & \hat{\gamma}_{21} & \cdots & \hat{\gamma}_{p-1,1} & \hat{\gamma}_{p,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \hat{\gamma}_{p,p-1} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

Let $\mathbf{D}$ be the $(p+1) \times (p+1)$ diagonal matrix with $\|\mathbf{z}_j\|$ on the diagonal, $\mathbf{Q} = \mathbf{Z}\mathbf{D}^{-1}$ and $\mathbf{R} = \mathbf{D}\boldsymbol{\Gamma}$. Then $\mathbf{X} = \mathbf{Q}\mathbf{R}$ gives the QR-decomposition of $\mathbf{X}$.

We can show that

$$\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}, \quad \hat{\mathbf{y}} = \mathbf{Q}\mathbf{Q}^T\mathbf{y}.$$

Questions:

- What happens if some predictors are highly correlated?
- What if we consider the singular value decomposition (SVD) of $\mathbf{X}$?

The LSE has the smallest MSE among all unbiased linear estimators. However, it is possible to achieve a smaller MSE or better prediction accuracy by considering biased estimation. In other words, it is possible to trade a little bias for a large reduction in variance.

In fact, biased estimates are commonly used in practice. One popular method is called ridge regression. It is in fact closely related to the SVD method.

Outline

In this section, we will try to answer the following questions:

- What is shrinkage estimation?
- How does shrinkage estimation arise?
- What is the rational behind shrinkage estimation?
- Wiener filter and ridge regression.

The Wiener Filter

Wiener filtering is a classical estimation problem in the electrical engineering. The example introduced here will play a pedagogical role as it wonderfully presents some key ideas covered in this course.

We will see how a shrinkage method naturally arises in an estimation problem.

Problem Set-up: We wish to recover a Gaussian signal $Z = (Z_1, \ldots, Z_n)^T$ from noisy data $Y = (Y_1, \ldots, Y_n)^T$ of the form

$$Y_i = Z_i + E_i, \quad i = 1, 2, \ldots, n,$$

where Y is the observed process, Z is the signal, which is assumed to be a Gaussian process with mean zero and covariance matrix Σ, i.e. $Z \sim MVN(0, \Sigma)$, and $E = (E_1, \ldots, E_n)^T$ is Gaussian white noise, i.e., $E \sim MVN(0, \sigma^2 I)$, which is independent of the signal X.

One may view this as a Bayesian estimation problem where the prior on the unknown signal is Gaussian.

Goal: reconstruct the signal by producing an estimator $\hat{Z} = g(Y)$, which can be computed from data.

Suppose square loss function is adopted. That is, we are looking for $\hat{Z}$ minimizing the MSE:

$$MSE(Z, \hat{Z}) = E\|Z - \hat{Z}\|_2^2 = E\sum_{i=1}^{n}(Z_i - \hat{Z}_i)^2.$$

We can show that the estimator which achieves the minimum MSE is the conditional expectation of Z given the observed process Y: $\hat{Z} = E(Z|\,Y)$. In detail, the i-th component is given by

$$\hat{Z}_i = \int_{\mathbb{R}^n} z_i p_{Z|\,Y}(z)dz.$$

Can we analytically evaluate the conditional expectation?

The computation of the desired conditional expectation can be greatly simplified by means of the **principal component analysis (PCA)**.

The key is to use principal component analysis to decompose a Gaussian process as a superposition of its principal components, which are independent each other.

(Population) Principal Component Analysis

Suppose a random vector Z has covariance matrix Σ with its eigenvalue-eigenvector pairs $(d_i^2, \mathbf{u}_i)$, where $d_1^2 \geq \cdots \geq d_n^2$.

Define $\mathbf{D} = \text{diag}(d_i^2)$, $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$. Then $\Sigma = \mathbf{U}\mathbf{D}\mathbf{U}^T$.

The i-th principal component is defined as

$$Z_i^* = \mathbf{u}_i^T Z, \quad i = 1, 2, \ldots, n.$$

Thus,

$$Z^* = \mathbf{U}^T Z.$$

(Population) Principal Component Analysis

Suppose a random vector Z has covariance matrix Σ with its eigenvalue-eigenvector pairs $(d_i^2, \mathbf{u}_i)$, where $d_1^2 \geq \cdots \geq d_n^2$.

Define $\mathbf{D} = \text{diag}(d_i^2)$, $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$. Then $\Sigma = \mathbf{U}\mathbf{D}\mathbf{U}^T$.

The i-th principal component is defined as

$$Z_i^* = \mathbf{u}_i^T Z, \quad i = 1, 2, \ldots, n.$$

Thus,

$$Z^* = \mathbf{U}^T Z.$$

Questions:

(Population) Principal Component Analysis

Suppose a random vector Z has covariance matrix Σ with its eigenvalue-eigenvector pairs $(d_i^2, \mathbf{u}_i)$, where $d_1^2 \geq \cdots \geq d_n^2$.

Define $\mathbf{D} = \text{diag}(d_i^2)$, $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$. Then $\Sigma = \mathbf{U}\mathbf{D}\mathbf{U}^T$.

The i-th principal component is defined as

$$Z_i^* = \mathbf{u}_i^T Z, \quad i = 1, 2, \ldots, n.$$

Thus,

$$Z^* = \mathbf{U}^T Z.$$

Questions:

- $\text{Var}(Z_i^*) = ?$

(Population) Principal Component Analysis

Suppose a random vector Z has covariance matrix Σ with its eigenvalue-eigenvector pairs $(d_i^2, \mathbf{u}_i)$, where $d_1^2 \geq \cdots \geq d_n^2$.

Define $\mathbf{D} = \text{diag}(d_i^2)$, $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$. Then $\Sigma = \mathbf{U}\mathbf{D}\mathbf{U}^T$.

The i-th principal component is defined as

$$Z_i^* = \mathbf{u}_i^T Z, \quad i = 1, 2, \ldots, n.$$

Thus,

$$Z^* = \mathbf{U}^T Z.$$

Questions:

- $\text{Var}(Z_i^*) =?$ d_i^2

(Population) Principal Component Analysis

Suppose a random vector Z has covariance matrix Σ with its eigenvalue-eigenvector pairs $(d_i^2, \mathbf{u}_i)$, where $d_1^2 \geq \cdots \geq d_n^2$.

Define $\mathbf{D} = \text{diag}(d_i^2)$, $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$. Then $\Sigma = \mathbf{U}\mathbf{D}\mathbf{U}^T$.

The i-th principal component is defined as

$$Z_i^* = \mathbf{u}_i^T Z, \quad i = 1, 2, \ldots, n.$$

Thus,

$$Z^* = \mathbf{U}^T Z.$$

Questions:

- $\text{Var}(Z_i^*) = ?$ d_i^2
- $\text{Cov}(Z_i^*, Z_j^*) = ?$

(Population) Principal Component Analysis

Suppose a random vector Z has covariance matrix Σ with its eigenvalue-eigenvector pairs $(d_i^2, \mathbf{u}_i)$, where $d_1^2 \geq \cdots \geq d_n^2$.

Define $\mathbf{D} = \text{diag}(d_i^2)$, $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$. Then $\Sigma = \mathbf{U}\mathbf{D}\mathbf{U}^T$.

The i-th principal component is defined as

$$Z_i^* = \mathbf{u}_i^T Z, \quad i = 1, 2, \ldots, n.$$

Thus,

$$Z^* = \mathbf{U}^T Z.$$

Questions:

- $\text{Var}(Z_i^*) =?$ d_i^2
- $\text{Cov}(Z_i^*, Z_j^*) =?$ 0

(Population) Principal Component Analysis

Suppose a random vector Z has covariance matrix Σ with its eigenvalue-eigenvector pairs $(d_i^2, \mathbf{u}_i)$, where $d_1^2 \geq \cdots \geq d_n^2$.

Define $\mathbf{D} = \text{diag}(d_i^2)$, $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$. Then $\Sigma = \mathbf{U}\mathbf{D}\mathbf{U}^T$.

The i-th principal component is defined as

$$Z_i^* = \mathbf{u}_i^T Z, \quad i = 1, 2, \ldots, n.$$

Thus,

$$Z^* = \mathbf{U}^T Z.$$

Questions:

- $\text{Var}(Z_i^*) =?$ d_i^2
- $\text{Cov}(Z_i^*, Z_j^*) =?$ 0
- $\sum_{i=1}^n \text{Var}(Z_i^*) =?$

(Population) Principal Component Analysis

Suppose a random vector Z has covariance matrix Σ with its eigenvalue-eigenvector pairs $(d_i^2, \mathbf{u}_i)$, where $d_1^2 \geq \cdots \geq d_n^2$.

Define $\mathbf{D} = \text{diag}(d_i^2)$, $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$. Then $\Sigma = \mathbf{U}\mathbf{D}\mathbf{U}^T$.

The i-th principal component is defined as

$$Z_i^* = \mathbf{u}_i^T Z, \quad i = 1, 2, \ldots, n.$$

Thus,

$$Z^* = \mathbf{U}^T Z.$$

Questions:

- $\text{Var}(Z_i^*) =?$ d_i^2
- $\text{Cov}(Z_i^*, Z_j^*) =?$ 0
- $\sum_{i=1}^n \text{Var}(Z_i^*) =?$ $\sum_{i=1}^n d_i^2$

PCA is the action of decomposing a process Z as a superposition of its principal components. The analysis consists of two steps:

Analysis Step: Find the orthonormal eigenvectors $\mathbf{u}_i$'s and construct the principal components,

$$Z_i^* = \mathbf{u}_i^T Z, \quad Z^* = \mathbf{U}^T Z.$$

Synthesis Step: Reconstruct the process for the principal components using the orthonormal eigenvectors,

$$Z = \mathbf{U} Z^*, \quad Z_i = \sum_{k=1}^{n} Z_k^* u_k(i).$$

Now, let's get back to the estimation problem.

By the definition of principal components, we have $\mathrm{Cov}(Z^*) = \mathbf{D}$. Therefore, the Z_i^* are independent in the case where Z is Gaussian.

Now, we "rotate" the observation vector Y in the orthonormal basis of principal components by

$$\mathbf{U}^T Y = \mathbf{U}^T Z + \mathbf{U}^T E, \Longrightarrow Y^* = Z^* + E^*.$$

If we denote E_i^* the i-th entry of E^*. Then

$$Z_i^* \sim N(0, d_i^2), \quad E_i^* \sim N(0, \sigma^2).$$

Also, Z_i^* and E_i^* are independent.

Note that $\|Z - \hat{Z}\| = \|Z^* - \hat{Z}^*\|$, so to estimate Z, we may just to estimate Z^* by any estimator $\hat{Z}^*$. The synthesis step would provide the reconstruction $\hat{Z} = \mathbf{U}\hat{Z}^*$.

Obviously, the problem has not changed and we merely looking at it from a different perspective.

Can we compute $\mathbf{U}\hat{Z}^*$ or $\hat{Z}^*$ now?

Can we compute $\mathbf{U}\hat{Z}^*$ or $\hat{Z}^*$ now?

$$\hat{Z}_i^* = E(Z_i^* \,|\, Y^*) = E(Z_i^* \,|\, Y_i^*) = \frac{d_i^2}{d_i^2 + \sigma^2}\, Y_i^*.$$

Denote

$$\Lambda = \text{Diag}\left(\frac{d_i^2}{d_i^2 + \sigma^2}\right).$$

Then $\hat{Z}^* = \Lambda\, Y^* = \Lambda \mathbf{U}^T\, Y$.

Finally, we see that the Wiener estimator of Z is given by

$$\hat{Z} = \mathbf{U}\hat{Z}^* = \mathbf{U}\Lambda\mathbf{U}^T\, Y.$$

The Wiener filter transforms the data with respect to the ortho-basis of principal components, and downweights each coefficient as a function of the signal-to-noise ratio since one can think of the coordinates of the weights as the ratio between the expected signal power and the expected (signal+noise) power. So, Wiener filter performs shrinkage estimation.

Here, both the shrinkage and the estimation procedure are linear.

The Wiener filter is optimal for Gaussian. In the non-Gaussian case, however, the estimator is only guaranteed to have minimum MSE among all linear estimators.

This estimation scheme can be summarized as

Transformation — Shrinkage — Reconstruction.

Question: What if Z does not follow a multivariate normal distribution?

Suppose $Z \sim g(z)$, and $Y|Z = z \sim e^{z^T y - \psi(z)} f_0(y)$, where $f_0(y)$ is the density of $Y|Z = 0$. Then Bayes rule provides the posterior density of Z given Y,

$$p(z|y) = e^{z^T y - \lambda(y)}[g(z)e^{-\psi(z)}], \quad \text{where } \lambda(y) = \log\left(\frac{f(y)}{f_0(y)}\right)$$

which represents an exponential family with canonical parameter y and cumulant generating function $\lambda(y)$. So

$$E(Z|\,Y = y) = \lambda'(y), \quad \text{Var}(Z|\,Y = y) = \lambda''(y).$$

This is the so called *Tweedie's formula*.

For example, if $Z \sim N(0, \Sigma)$, $E \sim N(0, \sigma^2 I)$, then $Y|Z = z \sim N(Z, \sigma^2 I)$. Using Tweedie's formula, we have

$$E(Z|\,Y) = Y - \sigma^2(\sigma^2 I + \Sigma)^{-1}\,Y = \Sigma(\sigma^2 I + \Sigma)^{-1}\,Y.$$

Outline

Ridge Regression

Consider linear regression model:

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i, \quad 1 \leq i \leq n.$$

- Response: $\mathbf{y} = (y_1, \ldots, y_n)^T$
- Predictors: $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, 2, \ldots, p$.
  Design Matrix: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$
- Error: $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$
- Regression Coefficient: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$

The linear regression model in matrix form:

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

We know that the LSE is not satisfactory.

Let's now perform a special type of shrinkage estimation called ridge regression.

WLOG, we assume that the response and predictors are centered and the predictors are standardized as follows:

$$\sum_{i=1}^{n} y_i = 0, \quad \mathbf{1}^T \mathbf{y} = 0$$

$$\sum_{i=1}^{n} x_{ij} = 0, \quad \mathbf{1}^T \mathbf{x}_j = 0, \quad \mathbf{1}^T \mathbf{X} = 0$$

$$\sum_{i=1}^{n} x_{ij}^2 = n, \quad \mathbf{x}_j^T \mathbf{x}_j = n, \quad j = 1, 2, \ldots, p.$$

There is no intercept in the regression model.

Each predictor is standardized to have the same magnitude in L_2. So the corresponding regression coefficients are "comparable".

After model fitting, the results can be readily transformed back to the original scale.

Recall in decision theory, we introduced the idea of adding a regularization term to an error function in order to control over-fitting and achieve smaller predictor risk.

In regression, the traditional penalized method is the ridge regression that uses the L_2-norm of the coefficients as penalty.

Originally, it was proposed to regularize ill-conditioned design matrices in linear regression.

Ridge Citerion:

$$\hat{\beta}_{\text{ridge}}(\lambda) = \text{argmin}_{\beta}\{\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2\}.$$

Ridge Estimator:

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

Proof: (Matrix derivative and data augmentation)

With the choice of quadratic penalty $\boldsymbol{\beta}^T\boldsymbol{\beta}$, the ridge estimator is again a linear function of $\mathbf{y}$.

The solution adds a positive constant to the diagonal of $\mathbf{X}^T\mathbf{X}$ before inversion. This makes the problem nonsingular, even if $\mathbf{X}$ is not of full column rank, and was the main motivation when ridge regression was first introduced. See Hoerl and Kennard (1970).

For any $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, we have a solution path

$$\{\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) : \lambda \in [\lambda_{\min}, \lambda_{\max}]\}.$$

The optimal λ and hence the optimal solution has to be selected by some information criteria or cross validation methods, which will be discussed later.

Ridge Regression: Examples

Example 1: Consider the linear regression model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad 1 \le i \le n,$$

where $\varepsilon \sim N(0, 1.5^2)$, $p = 100$.

Generate z_{ij}'s and w independently from $N(0, 1)$, let

$$x_{ij} = z_{ij} + w, \quad 1 \le j \le 4, \quad x_{i5} = z_{i5} + 2w, \quad x_{i6} = z_{i6} + w,$$

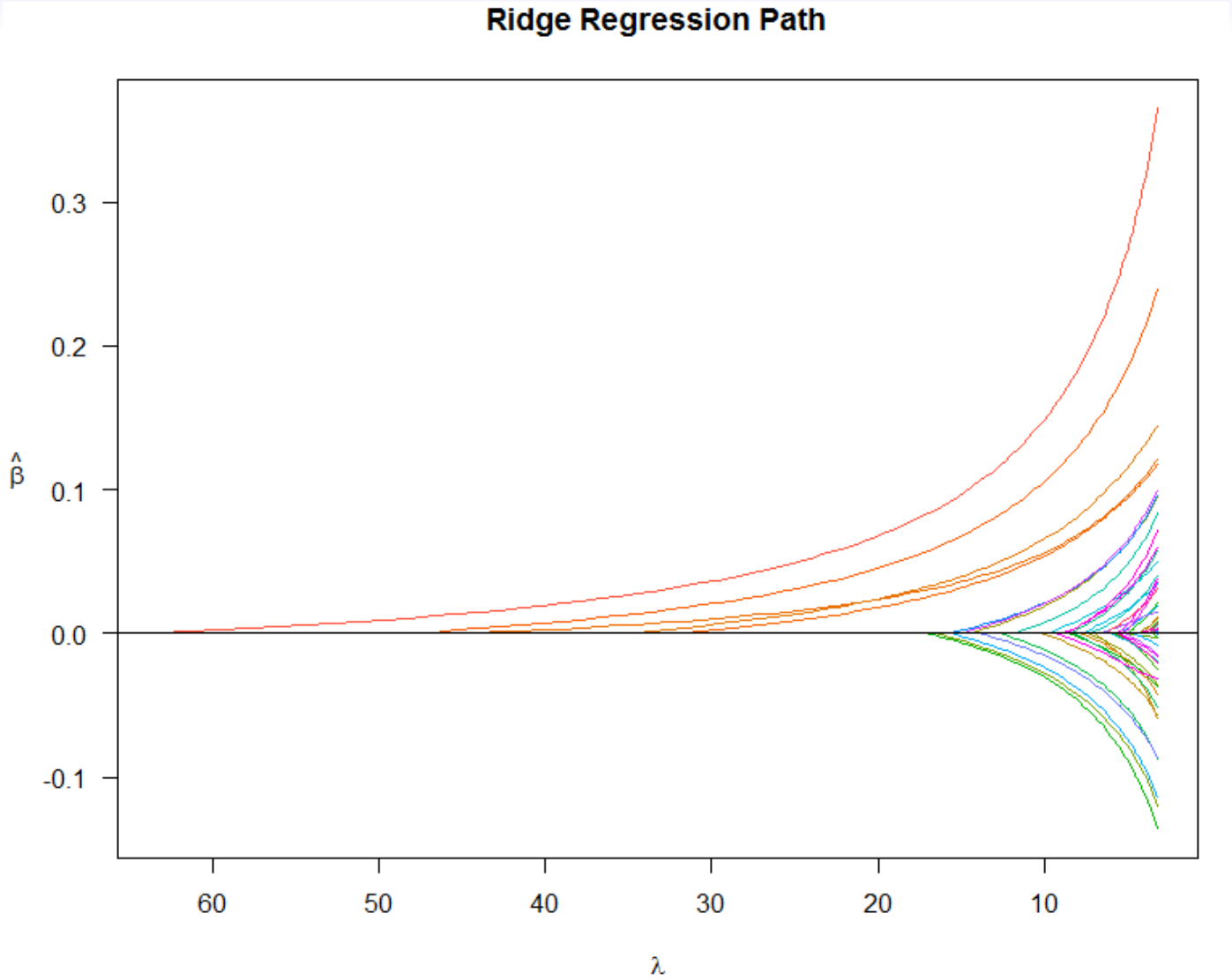
and $x_{ij} = z_{ij}$ for $j \ge 7$.

Let $(\beta_1, \beta_2, \beta_3) = (2, 1, -1)$ and $\beta_j = 0$ for $j \ge 4$.

To implement the ridge estimation for the above example, we use the R-package **ncvreg**. A sample R-code is

```
library(ncvreg);

n=100;
p=100;
sigma=1.5;
tau=1;

Z=matrix(rnorm(n*p,0,tau),nrow=n);
w=rnorm(n,0,tau);
X=matrix(0,nrow=n,ncol=p);

for(j in 1:6)
 {
   X[,j]=Z[,j]+ifelse(j==5,2,1)*w;
 }
X[,7:100]=Z[,7:100];

beta=as.vector(c(2,1,-1,rep(0,97)));
e=rnorm(n,0,sigma);
y=X%*%beta+e;

fit=ncvreg(X,y,gamma=1000,alpha=0.05)
plot(fit,main="Ridge Regression Path")
```

Remark: The ncvreg pacakge is built upon the work by Breheny and Huang (2011).

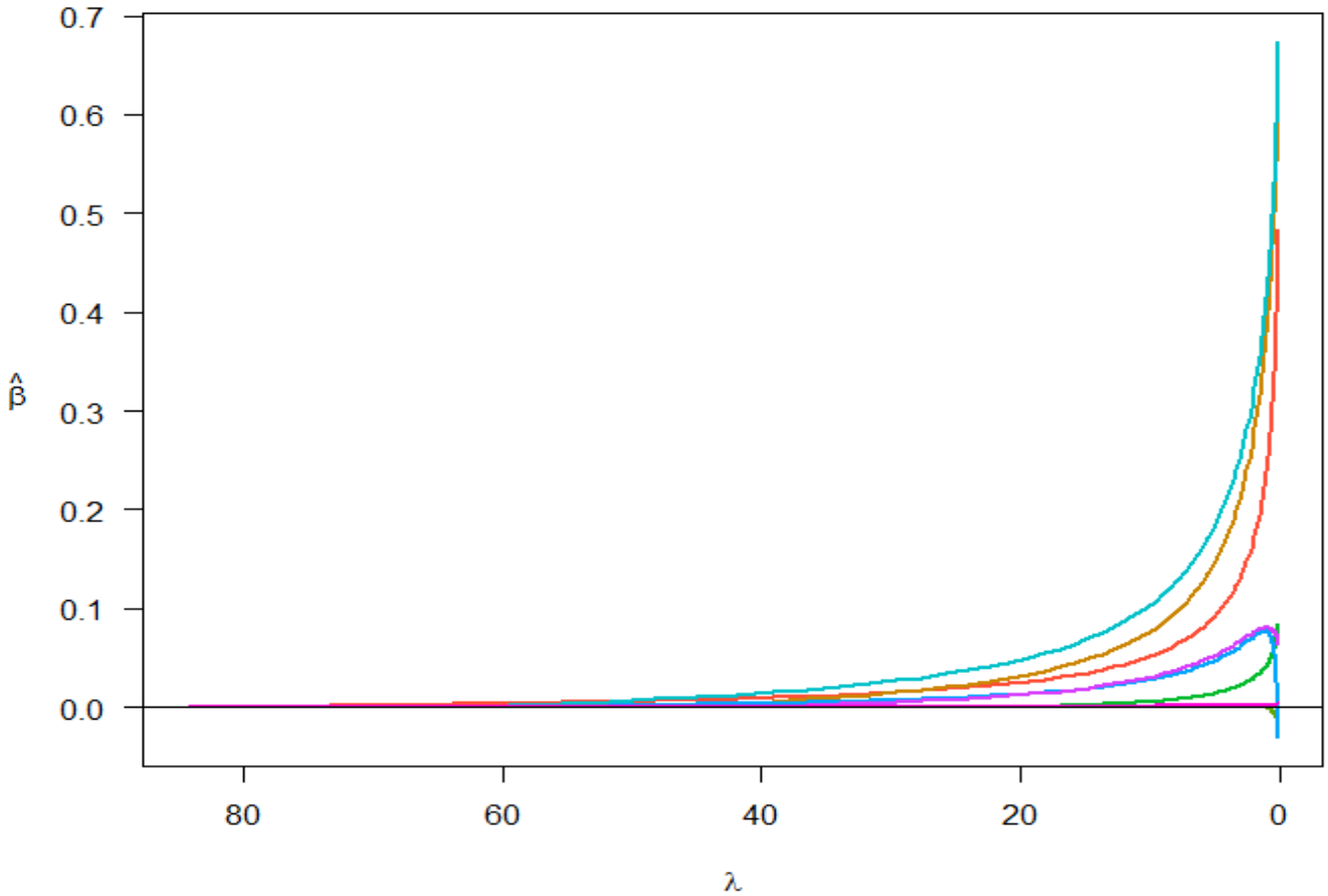

Ridge Regression Path
0.3
0.2
0.1
0.0
-0.1
β̂
60
50
40
30
20
10
λ

Example 2 (Prostate Data): The data is from a study by by Stamey et al. (1989) to examine the association between prostate specific antigen (PSA) and several clinical measures that are potentially associated with PSA in men who were about to receive a radical prostatectomy. The variables are as follows:

- lcavol: Log cancer volume
- lweight: Log prostate weight
- age: The man's age
- lbph: Log of the amount of benign hyperplasia
- svi: Seminal vesicle invasion; 1=Yes, 0=No
- lcp: Log of capsular penetration
- gleason: Gleason score
- pgg45: Percent of Gleason scores 4 or 5
- lpsa: Log PSA

To implement the ridge estimation for the above example, we use the R-package **ncvreg**. A sample R-code is

```
library(ncvreg)
data(prostate);
X=prostate[,1:8];
y=prostate$lpsa;
fit=ncvreg(X,y,gamma=1000,alpha=0.01)
plot(fit)
```



We can also use the R-package **glmnet** to do ridge regression. A sample R-code is

```
library(glmnet)
grid=seq(84.3427,0.0843,length=1000);
ridge.mod=glmnet(X,y,alpha=0,lambda=grid)
plot(ridge.mod)

# An improved one: Lambda in X-axis

beta=ridge.mod$beta;
plot(grid,beta[1,],type="l",xlim=rev(range(grid)),ylim=range(beta));
for(k in seq(2,8))
  {
    lines(grid,beta[k,]);
  }
```

The ridge regression method can be viewed as a Bayesian regression approach.

RR as Bayesian Regression

If $\boldsymbol{\beta} \sim N(0, \sigma^2 I/\lambda)$, and $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$, then

$$E(\boldsymbol{\beta}|\mathbf{y}) = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}.$$

Proof: Consider a general case. Suppose $\boldsymbol{\beta} \sim N(\mu, \Lambda^{-1})$, and $\mathbf{y}|\boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta} + b, L^{-1})$. Let $Z = (\boldsymbol{\beta}^T, \mathbf{y}^T)^T$, then the log-likelihood function of Z is given by

$$\begin{aligned}
\log f(z) &= -\frac{1}{2}(\boldsymbol{\beta} - \mu)^T\Lambda(\boldsymbol{\beta} - \mu) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - b)^T L(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - b) + \text{const.} \\
&= -\frac{1}{2}\boldsymbol{\beta}^T\Lambda\boldsymbol{\beta} + \boldsymbol{\beta}^T\Lambda\mu - \frac{1}{2}\mathbf{y}^T L\mathbf{y} - \frac{1}{2}\boldsymbol{\beta}^T\mathbf{X}^T L\mathbf{X}\boldsymbol{\beta} \\
&\quad + \mathbf{y}^T L\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^T Lb - \boldsymbol{\beta}^T\mathbf{X}^T Lb + \text{const.} \\
&= -\frac{1}{2}\left(\boldsymbol{\beta}^T\Lambda\boldsymbol{\beta} + \mathbf{y}^T L\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T L\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}^T L\mathbf{X}\boldsymbol{\beta}\right) \\
&\quad + \left(\boldsymbol{\beta}^T\Lambda\mu + \mathbf{y}^T Lb - \boldsymbol{\beta}^T\mathbf{X}^T Lb\right) + \text{const.}
\end{aligned}$$

Let

$$R = \begin{pmatrix} \Lambda + \mathbf{X}^T L \mathbf{X} & -\mathbf{X}^T L \\ -L\mathbf{X} & L \end{pmatrix}.$$

Then

$$\begin{aligned} \log f(z) &= -\frac{1}{2} Z^T R Z + Z^T \begin{pmatrix} \Lambda\mu - \mathbf{X}^T L b \\ Lb \end{pmatrix} + \text{const.} \\ &= -\frac{1}{2} Z^T R Z + Z^T R R^{-1} \begin{pmatrix} \Lambda\mu - \mathbf{X}^T L b \\ Lb \end{pmatrix} + \text{const.} \end{aligned}$$

Note that

$$R^{-1} \begin{pmatrix} \Lambda\mu - \mathbf{X}^T L b \\ Lb \end{pmatrix} = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}\mathbf{X}^T \\ \mathbf{X}\Lambda^{-1} & L^{-1} + \mathbf{X}\Lambda^{-1}\mathbf{X}^T \end{pmatrix} \begin{pmatrix} \Lambda\mu - \mathbf{X}^T L b \\ Lb \end{pmatrix} = \begin{pmatrix} \mu \\ \mathbf{X}\mu + b \end{pmatrix}.$$

This implies

$$Z = \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{y} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mathbf{X}\mu + b \end{pmatrix}, \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}\mathbf{X}^T \\ \mathbf{X}\Lambda^{-1} & L^{-1} + \mathbf{X}\Lambda^{-1}\mathbf{X}^T \end{pmatrix} \right)$$

Therefore,

$$E(\boldsymbol{\beta}|\mathbf{y}) = \mu + \Lambda^{-1}\mathbf{X}^T(L^{-1} + \mathbf{X}\Lambda^{-1}\mathbf{X}^T)^{-1}(\mathbf{y} - \mathbf{X}\mu - b).$$

Let $\Lambda = \lambda I/\sigma^2$, $L = I/\sigma^2$, $\mu = 0$, $b = 0$. After some matrix computation, one can prove the desired result.

Ridge Regression: Relation to SVD

Suppose the SVD of $\mathbf{X}$ is given by $\mathbf{X} = \mathbf{U}D\mathbf{V}^T$, where $\mathbf{U}^T\mathbf{U} = I$, $\mathbf{V}^T\mathbf{V} = I$, and $D = \text{diag}(d_i)$.

Then we can rewrite the ridge estimator as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) &= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{V}D^2\mathbf{V}^T + \lambda I)^{-1}\mathbf{V}D\mathbf{U}^T\mathbf{y} = \mathbf{V}(D^2 + \lambda I)^{-1}D\mathbf{U}^T\mathbf{y},\end{aligned}$$

and

$$\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = \mathbf{U}D(D^2 + \lambda I)^{-1}D\mathbf{U}^T\mathbf{y}.$$

Also, note that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) &= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}}_{\text{LS}} \\ &= (\mathbf{V}D^2\mathbf{V}^T + \lambda I)^{-1}\mathbf{V}D^2\mathbf{V}^T\hat{\boldsymbol{\beta}}_{\text{LS}} = \mathbf{V}(D^2 + \lambda I)^{-1}D^2\mathbf{V}^T\hat{\boldsymbol{\beta}}_{\text{LS}},\end{aligned}$$

So

$$\|\hat{\boldsymbol{\beta}}_{\text{ridge}}\| \leq \|\mathbf{V}^T\hat{\boldsymbol{\beta}}_{\text{LS}}\| = \|\hat{\boldsymbol{\beta}}_{\text{LS}}\|.$$

Ridge Regression: MSE, Bias and Variance Trade-off

MSE of Ridge Regression

Let $\Sigma_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda I)$. Then

$$E\|\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}\|^2 = \lambda^2\boldsymbol{\beta}^T\Sigma_\lambda^{-2}\boldsymbol{\beta} + \sigma^2\text{trace}(\Sigma_\lambda^{-1}\mathbf{X}^T\mathbf{X}\Sigma_\lambda^{-1}).$$

Note that $\mathbf{X}^T\mathbf{X} + \lambda I = \mathbf{V}(D^2 + \lambda I)\mathbf{V}^T$, so

$$(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} = \mathbf{V}(D^2 + \lambda I)^{-1}\mathbf{V}^T = \sum_{i=1}^{p} \frac{1}{d_i^2 + \lambda}\mathbf{v}_i\mathbf{v}_i^T.$$

Also

$$\begin{aligned}\text{trace}(\Sigma_\lambda^{-1}\mathbf{X}^T\mathbf{X}\Sigma_\lambda^{-1}) &= \text{trace}(\Sigma_\lambda^{-2}\mathbf{X}^T\mathbf{X}) = \text{trace}(\mathbf{V}(D^2 + \lambda I)^{-2}\mathbf{V}^T\mathbf{V}D^2\mathbf{V}^T) \\ &= \sum_{i=1}^{p} \frac{d_i^2}{(d_i^2 + \lambda)^2}.\end{aligned}$$

Hence,

$$E\|\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}\|^2 = \sum_{i=1}^{p} \frac{\lambda^2(\boldsymbol{\beta}^T\mathbf{v}_i)^2 + \sigma^2 d_i^2}{(d_i^2+\lambda)^2}.$$

Recall that

$$E\|\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta}\|^2 = \sigma^2 \text{trace}((\mathbf{X}^T\mathbf{X})^{-1}) = \sum_{i=1}^{p} \frac{\sigma^2}{d_i^2}.$$

Exercise: Show that if $0 < \lambda < \frac{2\sigma^2}{\boldsymbol{\beta}^T\boldsymbol{\beta}}$, then

$$E\|\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}\|^2 < E\|\hat{\boldsymbol{\beta}}_{\text{LS}} - \boldsymbol{\beta}\|^2.$$

Outline

Principal Component

Suppose each column vector in the design matrix $\mathbf{X}$ has already been centered. Therefore, the sample covariance matrix of p predictors is $\mathbf{X}^T\mathbf{X}/n$.

Based on the SVD of $\mathbf{X} = \mathbf{U}D\mathbf{V}^T$, we have

$$\mathbf{X}^T\mathbf{X} = (\mathbf{U}D\mathbf{V}^T)^T(\mathbf{U}D\mathbf{V}^T) = \mathbf{V}D^2\mathbf{V}^T.$$

The eigenvectors of $\mathbf{X}^T\mathbf{X}$, $\mathbf{v}_j$'s, are called **principal component direction** of $\mathbf{X}$.

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

$\mathbf{z}_j = \mathbf{X}\mathbf{v}_j = d_j\mathbf{u}_j$, $j = 1, 2, \ldots, p$, are called the **principal components** of $\mathbf{X}$.

The first principal component of $\mathbf{X}$ has the largest sample variance among all normalized linear combinations of the columns of $\mathbf{X}$. In fact, the sample variance of the first component is d_1^2/n; the second d^2/n; and so on. The covariance of different principal components is 0.

Principal Component Regression

Principal component regression (PCR) is the regression of $\mathbf{y}$ against $k(\leq p)$ principal components of $\mathbf{X}$. In other words, the design matrix in PCR is

$$[\mathbf{X}\mathbf{v}_1, \ldots, \mathbf{X}\mathbf{v}_k] = \mathbf{X}\mathbf{V}\Phi,$$

where $\Phi = [I_{k\times k}, \mathbf{0}_{k\times(p-k)}]^T$.

The PCR fitted response can be shown as

$$\hat{\mathbf{y}}_{\text{PCR}} = \mathbf{X}\mathbf{V}\Phi(\Phi^T\mathbf{V}^T\mathbf{X}^T\mathbf{X}\mathbf{V}\Phi)^{-1}\Phi^T\mathbf{V}^T\mathbf{X}^T\mathbf{y} = \sum_{i=1}^{k}(\mathbf{u}_i^T\mathbf{y})\mathbf{u}_i.$$

It is also easy to show that the fitted response based on LSE and ridge regression estimator are

$$\hat{\mathbf{y}}_{\text{LSE}} = \sum_{i=1}^{p}(\mathbf{u}_i^T\mathbf{y})\mathbf{u}_i, \quad \hat{\mathbf{y}}_{\text{ridge}} = \sum_{i=1}^{p}\frac{d_i^2}{d_i^2+\lambda}(\mathbf{u}_i^T\mathbf{y})\mathbf{u}_i.$$

Ridge Regression: Relation to PCR

Principal component regression forms the derived input variable by performing a PCA of the original variables. If the original variables are highly correlated, then only a few principal components are kept. Then the response y is regressed on the few derived variables.

Therefore, the principal components regression is very similar to ridge regression: both operate via the principal components of the input matrix.

Ridge regression shrinks the coefficients of the principal components, shrinking more depending on the size of the corresponding eigenvalue; principal components regression discards some principal component with smallest eigenvalues.

The difference is between shrinkage and hard-threshholding. There is another commonly used method called soft-threshholding, which is related to lasso methods we will dive into.

Some Further Remarks:

- Ridge regression protects against the potentially high variance of gradients estimated in the short directions. The implicit assumption is that the response will tend to vary most in the directions of high variance of the inputs. A related method is called principal component regression.
- We have demonstrated that using shrinkage estimation can alleviate overfitting and achieve better prediction performance.
- There are mainly two reasons why we are often not satisfied with the LSE:

  (a). Prediction: the LSE often have low bias but large variance;
  (b). Interpretation: The model is hard to interpret with a large number of predictors.
- Ridge regression addresses the first, but it does not perform variable selection.

Question: How to perform variable selection.

Outline

Inverse of Block Matrix

Let $\mathbf{A}$, $\mathbf{C}$, and $\mathbf{C}^{-1}+\mathbf{D}\mathbf{A}^{-1}\mathbf{B}$ be non-singular square matrices; then

$$(\mathbf{A}+\mathbf{B}\mathbf{C}\mathbf{D})^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1}+\mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}.$$

General Formula: Matrix Inversion in Block form

Let a $m \times n$ matrix $\mathbf{M}$ be partitioned into a block form: $\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{matrix} \}m \\ \}n \end{matrix}$

$\underbrace{}_{m}\ \underbrace{}_{n}$

where the $m \times m$ matrix $\mathbf{A}$ and $n \times n$ matrix $\mathbf{D}$ are invertible. Then we have

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & (\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} (\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix}$$

Outline

References

- Breheny, P. and Huang, J. (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection.*Ann. Appl. Statist.*, 5, 232-253.
- Hastie, Tibshirani and Friedman (2009). The Element of Statistical Learning. Springer.
- Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.
- Izenman, A. J. (2008). Modern Multivariate Statistical Techniques. Springer.
- Stamey, T., et al. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *Journal of Urology*, 16, 1076-1083.