Chapter 7. Biased Estimation

7.1 James - Stein Shrinkage Method

(1) Motivation

Let $\boldsymbol{z} \sim N_p(\boldsymbol{\mu}, \boldsymbol{I})$, then $\boldsymbol{z}$ is a good estimator for $\boldsymbol{\mu}$ because it is MLE and UMVUE (uniformly minimum variance unbiased estimator). Even though $\boldsymbol{z}$ is a good estimator for $\boldsymbol{\mu}$ in the sense of 1st moment (i.e., $E(\boldsymbol{z}) = \boldsymbol{\mu}$), however, James and Stein (1961), considered the 2nd moment (i.e., MSE : mean squared error);

$$E(\boldsymbol{z}'\boldsymbol{z}) = \sum_{i=1}^{p} E(z_i^2) = \sum(1 + \mu_i^2) = p + \boldsymbol{\mu}'\boldsymbol{\mu} > \boldsymbol{\mu}'\boldsymbol{\mu}.$$

Hence, $\boldsymbol{z}'\boldsymbol{z}$ is not an unbiased estimator of $\boldsymbol{\mu}'\boldsymbol{\mu}$. Based on this idea, they considered

$$\tilde{\boldsymbol{\mu}} = c\boldsymbol{z},\ 0 < c < 1$$

where c is called *shrinkage* constant.

(2) James-Stein estimator

James and Stein showed that $\exists\ 0 < c < 1$ s.t. MSE$(\tilde{\boldsymbol{\mu}}) \leq$ MSE$(\boldsymbol{z})$. We will prove it. Note that

$$\begin{aligned}
\text{MSE}(\tilde{\boldsymbol{\mu}}) &= E[(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})] \\
&= E[(\tilde{\boldsymbol{\mu}} - E(\tilde{\boldsymbol{\mu}}) + E(\tilde{\boldsymbol{\mu}}) - \boldsymbol{\mu})'(\tilde{\boldsymbol{\mu}} - E(\tilde{\boldsymbol{\mu}}) + E(\tilde{\boldsymbol{\mu}}) - \boldsymbol{\mu})] \\
&= E[(\tilde{\boldsymbol{\mu}} - E(\tilde{\boldsymbol{\mu}}))'(\tilde{\boldsymbol{\mu}} - E(\tilde{\boldsymbol{\mu}}))] + [E(\tilde{\boldsymbol{\mu}}) - \boldsymbol{\mu}]'[E(\tilde{\boldsymbol{\mu}}) - \boldsymbol{\mu}] \\
&= \text{tr}[Cov(\tilde{\boldsymbol{\mu}})] + ||Bias(\tilde{\boldsymbol{\mu}})||^2
\end{aligned}$$

and therefore,

$$\begin{aligned}
\text{MSE}(\tilde{\boldsymbol{\mu}}) &= \text{tr}(c^2\boldsymbol{I}) + (\boldsymbol{\mu} - E(\tilde{\boldsymbol{\mu}}))'(\boldsymbol{\mu} - E(\tilde{\boldsymbol{\mu}})) \\
&= pc^2 + (1-c)^2\boldsymbol{\mu}'\boldsymbol{\mu}.
\end{aligned}$$

On the other hand,

$$\text{MSE}(\boldsymbol{z}) = p$$

To find c minimizing $\text{MSE}(\tilde{\boldsymbol{\mu}})$, we take 1st derivative w.r.t. c, i.e.,

$$0 = \frac{\partial \text{MSE}(\tilde{\boldsymbol{\mu}})}{\partial c} = 2cp - 2(1-c)\boldsymbol{\mu}'\boldsymbol{\mu} = 2c(p + \boldsymbol{\mu}^T\boldsymbol{\mu}) - 2\boldsymbol{\mu}'\boldsymbol{\mu}$$

which gives

$$c = \frac{\boldsymbol{\mu}'\boldsymbol{\mu}}{p + \boldsymbol{\mu}'\boldsymbol{\mu}}$$

and then, we have

$$\tilde{\boldsymbol{\mu}} = \left(1 - \frac{p}{p + \boldsymbol{\mu}'\boldsymbol{\mu}}\right)\boldsymbol{z}$$

But, this estimator contains unknown parameter $\boldsymbol{\mu}'\boldsymbol{\mu}$, we replace it by its unbiased estimator $\boldsymbol{z}'\boldsymbol{z} - p$, i.e.,

$$\tilde{\boldsymbol{\mu}} = \left(1 - \frac{p}{\boldsymbol{z}^T\boldsymbol{z}}\right)\boldsymbol{z}$$

Finally, James and Stein showed that

$$\tilde{\boldsymbol{\mu}}_s = \left[1 - \frac{(p-2)}{\boldsymbol{z}^T\boldsymbol{z}}\right]\boldsymbol{z}$$

gives the minimum MSE, and it is called *Stein shrinkage* estimator.

7.2 Ridge Regression

7.2.1 Inference on the ridge regression

(1) Motivation and definition

The LSE $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ will be very unstable if there exists multicollinearity. Hoerl and Kennard (1970) suggested the *ridge estimator* defined as

$$\hat{\boldsymbol{\beta}}(\theta) = (\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

where θ is called *biasing parameter* or *shrinkage parameter*

(2) Estimation of θ

(a) ridge trace

First, we draw a figure of $\hat{\beta}_j(\theta),\ j = 0,\ \cdots,\ p-1$ for $\theta > 0$, and we estimate θ where $\hat{\beta}_j(\theta)$ is stabilized. This method is called *ridge trace*.

Fig. 7.1 (p. 280)

(b) GCV_θ

Wahba, Golub, and Heath (1979) suggested minimizing the GCV (generalized cross validation), defined as

$$GCV_\theta = \sum e_{i,\theta}^2 \ / \ \left[1 - \frac{1}{n}\mathrm{tr}(\boldsymbol{H}_\theta)\right]^2,$$

where $e_{i,\theta}$ is residual and $\boldsymbol{H}_\theta = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})^{-1}\boldsymbol{X}'$ is a hat matrix.

Ex.7.1 (p.280)

7.2.2 Properties of ridge estimator

(1) Regularized LSE

We can show that

$$\hat{\boldsymbol{\beta}}(\theta) = Arg_\beta min\ (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})\ s.t.\ \boldsymbol{\beta}'\boldsymbol{\beta} = c^2$$

i.e.,

$$\hat{\boldsymbol{\beta}}(\theta) = Arg_{\beta}min\ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda(\boldsymbol{\beta}'\boldsymbol{\beta} - c^2),$$

where λ is a lagrangian multiplier.

Fig. 7.2 (p. 283)

(2) MSE

Theorem 7.1. $\text{MSE}(\hat{\boldsymbol{\beta}}) - \text{MSE}(\hat{\boldsymbol{\beta}}(\theta))$ is positive definite for some $\theta > 0$.

(Proof) Since

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\beta}}(\theta)) &= E[(\hat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta})'] \\ &= Cov[(\hat{\boldsymbol{\beta}}(\theta))] + [E(\hat{\boldsymbol{\beta}}(\theta)) - \boldsymbol{\beta}][E(\hat{\boldsymbol{\beta}}(\theta)) - \boldsymbol{\beta}]' \end{aligned}$$

and let $\boldsymbol{K} = (\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})^{-1}$, then

$$\begin{aligned} Cov[\hat{\boldsymbol{\beta}}(\theta)] &= Cov[(\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}] \\ &= (\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})^{-1}\sigma^2 \\ &= \boldsymbol{K}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{K}\sigma^2 \end{aligned}$$

and

$$E(\hat{\boldsymbol{\beta}}(\theta)) - \boldsymbol{\beta} = E(\boldsymbol{K}\boldsymbol{X}'\boldsymbol{y}) - \boldsymbol{\beta} = (\boldsymbol{K}\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{K}\boldsymbol{K}^{-1})\boldsymbol{\beta} = -\theta\boldsymbol{K}\boldsymbol{\beta}$$

we have

$$\text{MSE}(\hat{\boldsymbol{\beta}}(\theta)) = \boldsymbol{K}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{K}\sigma^2 + \theta^2\boldsymbol{K}\boldsymbol{\beta}\boldsymbol{\beta}'\boldsymbol{K}$$

Here, $Cov(\hat{\boldsymbol{\beta}}) = \text{MSE}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$, so that

$$\begin{aligned}
\text{MSE} \quad & (\hat{\boldsymbol{\beta}}) - \text{MSE}(\hat{\boldsymbol{\beta}}(\theta)) \\
& = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2 - \boldsymbol{K}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{K}\sigma^2 - \theta^2\boldsymbol{K}\boldsymbol{\beta}\boldsymbol{\beta}'\boldsymbol{K} \\
& = \boldsymbol{K}\boldsymbol{K}^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}^{-1}\boldsymbol{K}\sigma^2 - \boldsymbol{K}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{K}\sigma^2 - \theta^2\boldsymbol{K}\boldsymbol{\beta}\boldsymbol{\beta}'\boldsymbol{K} \\
& = \boldsymbol{K}[\boldsymbol{K}^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}^{-1}\sigma^2 - \boldsymbol{X}'\boldsymbol{X}\sigma^2 - \theta^2\boldsymbol{\beta}\boldsymbol{\beta}']\boldsymbol{K}
\end{aligned}$$

Now, to show $\text{MSE}(\hat{\boldsymbol{\beta}}) - \text{MSE}(\hat{\boldsymbol{\beta}}(\theta))$ is p.d., we are only to show

$$\Delta = \boldsymbol{K}^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}^{-1}\sigma^2 - \boldsymbol{X}'\boldsymbol{X}\sigma^2 - \theta^2\boldsymbol{\beta}\boldsymbol{\beta}'$$

is p.d. because $\boldsymbol{K}$ is p.d. Note that

$$\begin{aligned}
\boldsymbol{K}^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}^{-1} & = (\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I}) \\
& = \boldsymbol{X}'\boldsymbol{X} + 2\theta\boldsymbol{I} + \theta^2(\boldsymbol{X}'\boldsymbol{X})^{-1}
\end{aligned}$$

implies that

$$\Delta = \theta[\{2\boldsymbol{I} + \theta(\boldsymbol{X}'\boldsymbol{X})^{-1}\}\sigma^2 - \theta\boldsymbol{\beta}\boldsymbol{\beta}']$$

Now, to show Δ is p.d., we need to show $2\sigma^2\boldsymbol{I} - \theta\boldsymbol{\beta}\boldsymbol{\beta}'$ is p.d., because $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is p.d and $\theta > 0$. Let $\boldsymbol{a} \neq \boldsymbol{0}$, then

$$\boldsymbol{a}'(2\sigma^2\boldsymbol{I} - \theta\boldsymbol{\beta}\boldsymbol{\beta}')\boldsymbol{a} \quad = 2\sigma^2\boldsymbol{a}'\boldsymbol{a} - \theta(\boldsymbol{a}'\boldsymbol{\beta})^2 \geq 2\sigma^2\boldsymbol{a}'\boldsymbol{a} - \theta(\boldsymbol{a}'\boldsymbol{a})(\boldsymbol{\beta}'\boldsymbol{\beta}) = (2\sigma^2 - \theta\boldsymbol{\beta}'\boldsymbol{\beta})\boldsymbol{a}'\boldsymbol{a}$$

by the Cauchy-Schwarz inequality. Hence, if $2\sigma^2 - \theta\boldsymbol{\beta}'\boldsymbol{\beta} > 0$, then $2\sigma^2\boldsymbol{I} - \theta\boldsymbol{\beta}\boldsymbol{\beta}'$ is p.d., i.e., For $0 < \theta < 2\sigma^2/\boldsymbol{\beta}'\boldsymbol{\beta}$, $\text{MSE}(\hat{\boldsymbol{\beta}}) - \text{MSE}(\hat{\boldsymbol{\beta}}(\theta))$ is p.d..

(3) Bayes estimator

Consider

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0,\, \sigma^2\boldsymbol{I})$$

and let

$$\boldsymbol{\beta} \sim N\left(\boldsymbol{0},\, \frac{\sigma^2}{\theta}\boldsymbol{I}\right)$$

be the prior distribution of $\boldsymbol{\beta}$. Then, we can show that the posterior distribution of $\boldsymbol{\beta}$ is

$$N_p(\,(\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y},\; (\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})^{-1}\sigma^2)$$

Therefore, under the squared error loss, $\hat{\boldsymbol{\beta}}(\theta) = (\boldsymbol{X}'\boldsymbol{X} + \theta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is the Bayes estimator of $\boldsymbol{\beta}$.

7.3 Principal Component Regression

(1) Motivation

When there are many covariates the PCR (principal component regression) uses a few pricipal components instead of the original covariates.

(1) Derivation of PCR

By the SVD (singular values decomposition) of the $n \times p$ matrix $\boldsymbol{X}$, we have

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}',$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $n \times p$ and $p \times p$ orthogonal matrix, respectively, and $\boldsymbol{D}$ is $p \times p$ diagonal matrix with diagonal elements $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ which are singular values of $\boldsymbol{X}$. Therefore,

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{V}'$$

which is just the spectral decomposition of $\boldsymbol{X}'\boldsymbol{X}$. Here, $d_1^2 \geq d_2^2 \geq \cdots \geq d_p^2 \geq 0$ are eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ and the corresponding eigenvectors are $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_p$ which are column vectors of $\boldsymbol{V}$, and they are orthogonal to each other. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)$, where $\boldsymbol{x}_i$ is $n-$vector for the ith covariate, and let

$$\boldsymbol{z}_i = \boldsymbol{X}\boldsymbol{v}_i = \boldsymbol{u}_i d_i, \ i = 1, 2, \cdots, p,$$

where $\boldsymbol{z}_i$ is called the ith principal component, and $\boldsymbol{u}_i$ is the ith column vector of $\boldsymbol{U}$. Hence, the sample variance of $\boldsymbol{z}_i$ is

$$Var(\boldsymbol{z}_i) = Var(\boldsymbol{X}\boldsymbol{v}_i) = d_i^2/n, \ i = 1, 2, \cdots, p.$$

Here, $\boldsymbol{z}_1$ is called the 1st principal component, and the i-th principal component is $\boldsymbol{z}_i = \boldsymbol{X}\boldsymbol{v}_i$. Note that $\boldsymbol{v}_i$ is called the ith principal component direction, and it satisfies the following conditions;

$$\underset{\substack{||\boldsymbol{\alpha}||=1 \\ \boldsymbol{v}_l'\boldsymbol{S}\boldsymbol{\alpha}=0 \\ l=1,2,\cdots,i-1}}{m} \quad ax\ Var(X\boldsymbol{\alpha}),$$

where

$$\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}/n$$

is the sample variance-covariance matrix of covariates. Also, since $\boldsymbol{v}_l'\mathbf{S}\boldsymbol{\alpha} = 0$, $\boldsymbol{z}_i = \boldsymbol{X}\boldsymbol{\alpha}$ and $\boldsymbol{z}_l = \boldsymbol{X}\boldsymbol{v}_l$, $l = 1, 2, \cdots, i-1$ are orthogonal to each other. The PCR uses just few principals $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_M,\ M << p$ out of $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_p$. If the response and covariates are centered, we may write the fitted vector of the PCR as

$$\hat{\boldsymbol{y}} = \sum_{m=1}^{M} \hat{\theta}_m \boldsymbol{z}_m,$$

where $\hat{\theta}_m = \boldsymbol{z}_m'\boldsymbol{y}/\boldsymbol{z}_m'\boldsymbol{z}_m$ which is just the estimate of the slope for the simple linear regression of $\boldsymbol{y}$ on $\boldsymbol{z}_m$. To estimate M, the number of principals to be used, is often estimated by CV.

7.4 PLS : Partial Least Squares

PLS is quite similar to the PCR, however, it uses the information of $\boldsymbol{y}$ in addition to $\boldsymbol{X}$. Assume that the ith PLS direction vector $\boldsymbol{v}_i$ satisfies the following condition;

$$\max_{\substack{||\boldsymbol{\alpha}||=1 \\ \boldsymbol{v}_l'\mathbf{S}\alpha=0 \\ l=1,2,\cdots,i-1}} Corr^2(\boldsymbol{y}, X\boldsymbol{\alpha})Var(X\boldsymbol{\alpha})$$

The PLS algorithm can be summarized as follows;

(algorithm)

(i) By regressing $\boldsymbol{y}$ on $\boldsymbol{x}_j, j = 1, 2, \cdots, p$, compute the estimate of regression coefficients $\hat{v}_{1j} =< \boldsymbol{x}_j, \boldsymbol{y} >$.

(ii) Compute the 1st PLS direction vector $\boldsymbol{z}_1 = \sum \hat{v}_{1j}X_j$.

(iii) Regress $\boldsymbol{y}$ on $\boldsymbol{z}_1$, and compute the estimate of coefficient $\hat{\theta}_1 =< \boldsymbol{z}_1, \boldsymbol{y} >$.

(iv) Orthogonalize $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p$ w.r.t. $\boldsymbol{z}_1$, i.e, $\boldsymbol{x}_j - \frac{<\boldsymbol{z}_1,\boldsymbol{x}_j>}{<\boldsymbol{z}_1,\boldsymbol{z}_1>}\boldsymbol{z}_1,\ j = 1, \cdots, p$.

(v) Repeat this process to the Mth direction.

7.5 LASSO

7.5.1 Estimation of LASSO

(1) Definition

LASSO (least absolute shrinkage and selection operator) is suggested by Tibshirani (1996). The ridge regression is obtained by assigning L_2 restriction on $\boldsymbol{\beta}$, however, the LASSO is obtained by assigning L_1 restriction on $\boldsymbol{\beta}$, i.e.,

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_L &= \underset{\boldsymbol{\beta}}{arg\,min}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \ s.t. \ \sum_{j=1}^{p} |\beta_j| \le s \\
&= \underset{\boldsymbol{\beta}}{arg\,min}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|,
\end{aligned}
$$

where s and λ are parameters to be estimated. The solution of the above equation does not have a closed form solution, so that we are only to compute $\boldsymbol{\beta}$ numerically by packages like LARS (least angle regression).

We can obtain an explicit solution when we have one covariate. For simplicity, assume that $\sum y_i = 0$, $\sum x_i = 0$ and $\sum x_i^2 = 1$, and consider $y_i = \beta x_i + \epsilon_i$. Then, the LASSO estimate is given by

$$
\hat{\boldsymbol{\beta}}_L(\lambda) = sgn(\boldsymbol{\beta})(|\boldsymbol{\beta}| - \lambda)_+,
$$

where $(x)_+ = xI(x > 0)$. We can prove as follows; Note that

$$
\begin{aligned}
f(\beta) &= \frac{1}{2}\sum(y_i - x_i\beta)^2 + \lambda|\beta| \\
&= \frac{1}{2}\sum(y_i - x_i\hat{\beta} + x_i\hat{\beta} - x_i\beta)^2 + \lambda|\beta| \\
&= \frac{1}{2}\sum(y_i - x_i\hat{\beta})^2 + \frac{1}{2}(\hat{\beta} - \beta)^2 + \lambda|\beta|
\end{aligned}
$$

Hence,

$$\underset{\beta}{min} f(\beta) \Leftrightarrow \underset{\beta}{min} g(\beta) = \frac{1}{2}(\hat{\beta} - \beta)^2 + \lambda|\beta|$$

so that $\partial g(\beta)/\partial\beta = 0$ gives the LASSO estimate.

(2) Estimation of s.

Note that $\sum|\beta_j| \leq s \sum \beta_j^2/|\beta_j| \leq s$, which is similar form to the ridge regression case. Hence, under this restriction, the ridge regression estimator is given by

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{W}^-)^{-1}\boldsymbol{X}\boldsymbol{y},$$

where $\boldsymbol{W} = diag(|\tilde{\beta}_1|, \cdots, |\tilde{\beta}_p|)$ and λ satisfies $\sum|\tilde{\beta}_j| = s$. Here, $\boldsymbol{W}^-$ denotes the generalized inverse of $\boldsymbol{W}$ which might contain $\tilde{\beta}_j = 0$. Now, the hat matrix is

$$\boldsymbol{H}_s = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{W}^-)^{-1}\boldsymbol{X}'$$

and to estimate s, we minimize the generalized cross validation, i.e.,

$$GCV(s) = \frac{1}{n}\sum(y_i - \hat{y}_i)^2/\{1 - \frac{1}{n}\mathrm{tr}(\boldsymbol{H}_s)\}^2$$

7.5.2 Properties of LASSO

(1) Geometric interpretation

Fig. 7.3 (p. 290)

(2) Bayesian interpretation

ridge regression estimator = posterior mean when the prior for $\boldsymbol{\beta}$ is Gaussian.

LASSO estimator = posterior mode when the prior for $\boldsymbol{\beta}$ is double exponential (Laplace).

(3) Comparisons

Under the orthonormal design, i.e., $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$,

(i) variable selection : $\hat{\beta}_{S,j} = \hat{\beta}_j I(|\hat{\beta}_j| > \lambda)$

(ii) ridge regression : $\hat{\beta}_{R,j} = \hat{\beta}_j/(1+\lambda)$

(iii) LASSO : $\hat{\beta}_{L,j} = sgn(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$

Fig. 7.4 (p. 290)

7.6 Example (Prostate Cancer Data)

(1) Data description

response : lpsa (log of the level of prostate specific antigen)

covariates : $X_1, \cdots, X_8$

LSE, variable selection, ridge reg., LASSO, PCR, PLS : Table 7.4 (p. 293)

Fig. 7.7 and 7.8 (p. 294)

R code (Fig. 7.9, p. 294)

7.7 Bayes Estimator and Biased Estimators

7.7.1 Bayes Estimator

(1) Posterior distribution

Let $f(x;\theta)$ be the pdf of the r.v. X, i.e.,

$$X \sim f(x;\theta),$$

where θ is a unknown parameter. In the non-Bayesian point of view, θ is regarded as a unknown constant. On the other hand, in the Bayesian point of view, θ is regarded as a r.v. with pdf. Hence, in the Bayesian point of view, $f(x;\theta)$ is regarded as a conditional pdf of X when θ is given, i.e.,

$$X \sim f(x|\theta)$$

Now, let the pdf of θ be $g(\theta;\gamma)$, i.e.,

$$\theta \sim g(\theta;\gamma),$$

where γ is also a parameter. We call $g(\theta;\gamma)$, the pdf of θ, as a *prior distribution*. Also, the product of the conditional pdf $f(x|\theta)$ and $g(\theta;\gamma)$, the prior pdf of θ, is called a *posterior distribution*, i.e.,

$$g(\theta|x;\gamma) = \frac{f(x|\theta)g(\theta;\gamma)}{\int f(x|\theta)g(\theta;\gamma)d\theta}$$

(2) Bayes estimator

Let $\delta \equiv \delta(\boldsymbol{x})$ be an estimator of θ, and let $L(\theta,\delta)$ be a loss function. Then, $L(\theta,\delta) = (\theta-\delta)^2$ is called a squared error loss function, and $L(\theta,\delta) = |\theta-\delta|$ is called an absolute error loss function. Further, the expectation of the loss function $R(\theta,\delta) = E[L(\theta,\delta)]$ is called a risk function, i.e.,

$$R(\theta,\delta) = \int_{\chi} L(\theta,\delta)f(x|\theta)dx$$

Now, the Bayes risk is defined as

$$r(\theta,\delta) = \int_{\Theta} R(\theta,\delta)g(\theta;\gamma)d\theta$$

which can be rewritten as

$$\begin{aligned}
r(\theta,\delta) &= \int_{\Theta} R(\theta,\delta)g(\theta;\gamma)d\theta \\
&= \int_{\Theta}\int_{\chi} L(\theta,\delta)f(x|\theta)g(\theta;\gamma)dxd\theta \\
&= \int_{\chi}\int_{\Theta} L(\theta,\delta)f(x|\theta)g(\theta;\gamma)d\theta dx \\
&= \int_{\chi} E[L(\Theta,\delta)|X=x]dx,
\end{aligned}$$

where $E[L(\Theta, \delta)|X = x]$ is called *posterior risk*. The *Bayes estimator* minimizes the Bayes risk or the posterior risk, i.e.,

$$\hat{\delta}_{Bayes} = argmin_{\delta}\ r(\theta, \delta).$$

We can show that the Bayes estimator is a posterior mean under squared error loss, and a posterior mode under absolute error loss.

7.7.2 Stein shrinkage estimator and the empirical Bayes estimator

If the parameter γ in the prior pdf $g(\theta; \gamma)$ is unknown, it should be estimated. To estimate γ, we use the marginal pdf of the posterior pdf by integrating it w.r.t. θ. If we replace the estimate of γ in the Bayes estimator, the resulting estimator is called *empirical Bayes estimator*.

Theorem 7.2 Assume that $\boldsymbol{z} \sim N_p(\boldsymbol{\mu}, \boldsymbol{I})$ and the prior for $\boldsymbol{\mu}$ is $\boldsymbol{\mu} \sim N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Then, the Stein shrinkage estimator is the empirical bayes estimator under the squared error loss.

(Proof) First, the posterior becomes

$$
\begin{aligned}
& f(\boldsymbol{z}|\boldsymbol{\mu})g(\boldsymbol{\mu}) \\
& \propto \exp\left[-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})'(\boldsymbol{z}-\boldsymbol{\mu})\right]\exp\left[-\frac{1}{2\sigma^2}\boldsymbol{\mu}'\boldsymbol{\mu}\right] \\
& \propto \exp\left[-\frac{1}{2}(\boldsymbol{z}'\boldsymbol{z}-2\boldsymbol{\mu}'\boldsymbol{z}+\boldsymbol{\mu}'\boldsymbol{\mu}+\frac{1}{\sigma^2}\boldsymbol{\mu}'\boldsymbol{\mu})\right] \\
& \propto \exp\left[-\frac{1}{2}\left\{(1+\frac{1}{\sigma^2})[\boldsymbol{\mu}'\boldsymbol{\mu}-2\frac{\sigma^2}{1+\sigma^2}\boldsymbol{\mu}'\boldsymbol{z}+\frac{\sigma_2}{1+\sigma^2}\boldsymbol{z}'\boldsymbol{z}]\right\}\right].
\end{aligned}
$$

If we let

$$
w = \frac{\sigma^2}{1+\sigma^2}
$$

then, the posterior can be written as

$$
\begin{aligned}
& \propto \exp\left[-\frac{1}{2w}(\boldsymbol{\mu}'\boldsymbol{\mu}-2w\boldsymbol{\mu}'\boldsymbol{z}+w^2\boldsymbol{z}'\boldsymbol{z})\right]\exp\left[-\frac{1}{2}(1-w)\boldsymbol{z}'\boldsymbol{z}\right] \\
& \propto \exp\left[-\frac{1}{2w}(\boldsymbol{\mu}-w\boldsymbol{z})'(\boldsymbol{\mu}-w\boldsymbol{z})\right]\exp\left[-\frac{1}{2}(1-w)\boldsymbol{z}'\boldsymbol{z}\right].
\end{aligned}
$$

Therefore, the posterior becomes

$$
\boldsymbol{\mu}|\boldsymbol{z} \sim N_p(w\boldsymbol{z}, w\boldsymbol{I}).
$$

Since the Bayes estimator is the posterior mean under squared error loss, we have

$$
w\boldsymbol{z} = \frac{\sigma^2}{1+\sigma^2}\boldsymbol{z} = \left(1-\frac{1}{1+\sigma^2}\right)\boldsymbol{z},
$$

however, it contains unknown parameter σ^2, we obtain the marginal distribution of $\boldsymbol{z}$ to compute the estimator of σ^2.

$$\int f(\boldsymbol{z}|\boldsymbol{\mu})g(\boldsymbol{\mu})d\boldsymbol{\mu} \propto \exp[-\frac{1}{2}(1-w)\boldsymbol{z}'\boldsymbol{z}],\ \boldsymbol{z} \sim N_p(\boldsymbol{0}, (1+\sigma^2)\boldsymbol{I})$$

Now, consider

$$V \equiv \frac{\boldsymbol{z}'\boldsymbol{z}}{1+\sigma^2} \sim \chi^2(p)$$

then

$$E(\frac{1}{V}) \;=\; \int_0^\infty \frac{1}{v}\frac{v^{\frac{p}{2}-1}e^{v/2}}{\Gamma(\frac{p}{2})2^{\frac{p}{2}}}dv = \frac{\Gamma(\frac{p}{2}-1)2^{\frac{p}{2}-1}}{\Gamma(\frac{p}{2})2^{\frac{p}{2}}} = \frac{\Gamma(\frac{p}{2}-1)}{2(\frac{p}{2}-1)\Gamma(\frac{p}{2}-1)} = \frac{1}{p-2}$$

hence,

$$E\left[\frac{1+\sigma^2}{\boldsymbol{z}'\boldsymbol{z}}\right] = p-2$$

Therefore, the unbiased estimator of $(1+\sigma^2)^{-1}$ is $(p-2)/\boldsymbol{z}'\boldsymbol{z}$, and finally the empirical Bayes estimator is

$$(1-\frac{p-2}{\boldsymbol{z}'\boldsymbol{z}})\boldsymbol{z}$$

which is just the Stein shrinkage estimator.

7.7.3 Bayes estimator in regression

Consider a multiple linear regression model

$$\boldsymbol{y} \;=\; \boldsymbol{X\beta} \;+\; \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

which is equivalent to $\boldsymbol{y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, and assume that the prior to $\boldsymbol{\beta}$ be

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{m}, \sigma^2\boldsymbol{V})$$

Now, the posterior of $\boldsymbol{\beta}$ becomes

$$\begin{aligned}
&\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right]\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right]\\
&\propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}+\boldsymbol{\beta}'\boldsymbol{V}^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}-2\boldsymbol{\beta}'\boldsymbol{V}^{-1}\boldsymbol{m}+\ldots)\right]\\
&\propto \exp\left[-\frac{1}{2\sigma^2}\{\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{V}^{-1})\boldsymbol{\beta}-2\boldsymbol{\beta}'(\boldsymbol{V}^{-1}\boldsymbol{m}+\boldsymbol{X}'\boldsymbol{y})+\ldots\}\right]\\
&\equiv \exp\left[-\frac{1}{2\sigma^2}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)'(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{V}^{-1})(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)\}\right],
\end{aligned}$$

where $\boldsymbol{\mu}_\beta$ is the mean of the posterior, then we let

$$-2\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{V}^{-1})\boldsymbol{\mu}_\beta \equiv -2\boldsymbol{\beta}'(\boldsymbol{V}^{-1}\boldsymbol{m}+\boldsymbol{X}'\boldsymbol{y})$$

Hence,

$$\boldsymbol{\mu}_\beta = (\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{V}^{-1})^{-1}(\boldsymbol{V}^{-1}\boldsymbol{m}+\boldsymbol{X}'\boldsymbol{y}) \equiv \tilde{\boldsymbol{\beta}}_{Bayes}$$

which is the Bayes estimator under the squared error loss.

As a special case, let $\boldsymbol{V} = \lambda^{-1}\boldsymbol{I}_p$ for some $\lambda > 0$ and $\boldsymbol{m} = \boldsymbol{0}$, then

$$\tilde{\boldsymbol{\beta}}_{Bayes} = (\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

which is just the ridge regression estimator. Next, if $\boldsymbol{V} = \lambda^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\boldsymbol{m} = \mathbf{0}$, then

$$\tilde{\boldsymbol{\beta}}_{Bayes} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (1+\lambda)^{-1}\hat{\boldsymbol{\beta}}$$

which is called $James - Stein\ regression\ estimator$.