Multicollinearity Diagnostics and Remedies

March 1-5, 2010

Tiejun (Ty) Tong
Department of Applied Mathematics

Multiple linear regression

Recall that the multiple linear regression is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ are the observed response values,
$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^T$ are the regression coefficients,
$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ are the random errors,
$\mathbf{X}$ is an $n \times p$ matrix of predictor values with structure

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}_{n \times p}$$

Multicollinearity

- Multicollinearity, or near-linear dependence, is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated.
- If there is no linear relationship between the predictor variables, they are said to be orthogonal.
- In most applications of regression, the predictor variables are usually not orthogonal.

Multicollinearity

- Let the jth column of the $\mathbf{X}$ matrix be $\mathbf{X}_j$, so that

$$\mathbf{X} = [\mathbf{X}_1 \mid \mathbf{X}_2 \mid \ \ldots \ \mid \mathbf{X}_p].$$

  Note that $\mathbf{X}_j$ corresponds to the $(j-1)$th predictor variable.

- The vectors $\mathbf{X}_1, \ldots, \mathbf{X}_p$ are linearly dependent if there is a set of non-zero constants $c_1, \ldots, c_p$ such that

$$\sum_{j=1}^{p} c_j \mathbf{X}_j = \mathbf{0}. \tag{1}$$

- If Eq. (1) holds exactly for a subset of the columns of $\mathbf{X}$, the rank of $\mathbf{X}$ is less than p and the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible.

Multicollinearity

- If the dependence among the predictor variables is not strong, the impacts/effects caused by non-orthogonality are usually not serious.
- If there is some near-linear the dependence among the predictor variables, the problem of multicollinearity is said to exit which will make the $\mathbf{X}^T\mathbf{X}$ matrix ill-conditioned.
- The presence of multicollinearity can cause serious problems with the estimation of $\boldsymbol{\beta}$ and associated quantities, as well as interpretation.

Indications of possible multicollinearity

- Large correlation coefficients in the correlation matrix of predictor variables (visualized in the scatter plots).
- Large changes in the estimated regression coefficients when a predictor variable is added or deleted.
- Nonsignificant/opposite results in individual tests on the regression coefficients for important predictor variables.
- Wide CIs for the regression coefficients representing important predictor variables.

Variance Inflation Factor (VIF)

- The Variance Inflation Factor (VIF) quantifies the severity of multicollinearity in an ordinary least squares regression analysis.
- Let R_i^2 denote the coefficient of determination when x_i is regressed on all other predictor variables in the model. Let

$$\text{VIF}_i = \frac{1}{1 - R_i^2}, \qquad i = 1, 2, \ldots, p - 1.$$

- $\text{VIF}_i = 1$ when $R_i^2 = 0$, i.e., when x_k is not linearly related to the other predictor variables.
- $\text{VIF}_i \to \infty$ when $R_i^2 \to 1$, i.e., when x_k tends to have a perfect linear association with other predictor variables.

Variance Inflation Factor (VIF)

- The VIF provides an index that measures how much the variance of an estimated regression coefficient is increased because of the multicollinearity.
- Example: If the VIF of a predictor variable x_i is 8, this means that the variance of the estimated β_i is 8 times as large as it would be if x_i is uncorrelated with other predictor variables.
- **Rule of Thumb**: If any of the VIF values exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

Eigensystem analysis of $\mathbf{X}^T\mathbf{X}$

- The eigenvalues can also be used to measure the presence of multicollinearity. If there are one or more near-linear dependences in the predictor variables, then one or more of the eigenvalues will be small.
- Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of $\mathbf{X}^T\mathbf{X}$. The **condition number** of $\mathbf{X}^T\mathbf{X}$ is defined as

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}.$$

A matrix with a low condition number is said to be well-conditioned, while a matrix with a high condition number is said to be ill-conditioned.

Eigensystem analysis of $\mathbf{X}^T\mathbf{X}$

- The **condition indices** of $\mathbf{X}^T\mathbf{X}$ are defined as

$$\kappa_j = \frac{\lambda_{max}}{\lambda_j}, \qquad j = 1, 2, \ldots, p.$$

- The number of condition indices that are large is a useful measure of the number of near-linear dependences in $\mathbf{X}^T\mathbf{X}$.

- **Rule of Thumb**:
  (1) If the condition number is less than 100, there is no serious problem with multicollinearity;
  (2) A condition numbers between 100 and 1000 implies a moderate to strong multicollinearity;
  (3) If the condition number exceeds 1000, severe multicollinearity is indicated.

Example

```
> library(MASS) #load the "mvrnorm" function
> library(car)        #load the "vif" function
> set.seed(1000)
> mu <- c(0,0)
> Sigma <- matrix(c(1,0.9999,0.9999,1),2,2)
> x <- mvrnorm(n=10,mu,Sigma)
> x1 <- x[,1]
> x2 <- x[,2]
> x3 <- rnorm(10,0,1)
> X <- cbind(x1,x2,x3)
> cor(X)
           x1        x2        x3
x1 1.0000000 0.9998753 0.1701740
x2 0.9998753 1.0000000 0.1716474
x3 0.1701740 0.1716474 1.0000000
```

When $\rho = 0.9999$

```
> y <- x1+x2+x3+rnorm(10,0,0.1)
> fit1 <- lm(y~x1+x2+x3)
> summary(fit1)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0070	0.0381	-0.18	0.8611
x1	3.0924	2.8535	1.08	0.3201
x2	-1.0495	2.8568	-0.37	0.7260
x3	0.9911	0.0251	39.46	0.0000

```
> vif(fit1)
        x1         x2         x3
4044.660240 4046.759079    1.039410
> max(e)/min(e)
[1] 23823.28
```

When $\rho = 0$

```
> Sigma <- matrix(c(1,0,0,1),2,2)
> summary(fit2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0070	0.0381	-0.18	0.8611
x1	1.0429	0.0457	22.83	0.0000
x2	1.0293	0.0404	25.49	0.0000
x3	0.9911	0.0251	39.46	0.0000

```
> vif(fit2)
      x1       x2       x3
1.035435 1.012250 1.039410
> max(e)/min(e)
[1] 3.735222
```

When $\rho = 0.8$

```
> Sigma <- matrix(c(1,0.8,0.8,1),2,2)
> summary(fit3)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0070	0.0381	-0.18	0.8611
x1	1.0689	0.0666	16.05	0.0000
x2	0.9763	0.0698	13.98	0.0000
x3	0.9911	0.0251	39.46	0.0000

```
> vif(fit3)
      x1       x2       x3
2.339088 2.389312 1.039410
> max(e)/min(e)
[1] 12.15263
```

Remedial Measures

- In polynomial regression models, it is suggested to use the centered data for the predictor variables. Centering the data often reduces the multicollinearity.
- To drop one or several predictor variables in order to lesson the multicollinearity and thereby improve the estimation precision of the remaining regression coefficients.
- If none of the predictor variables can be dropped, we should consider alternative (biased) methods of estimation, e.g.,
  (1) Ridge Regression;
  (2) LASSO Regression;
  (3) Principal Component Regression (PCR).

Ridge Regression

- Recall that in the ordinary linear regression,

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

- Hoerl and Kennard (1970) proposed to estimate $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y},$$

  where $\lambda \geq 0$ is the ridge parameter. When $\lambda = 0$, the ridge estimation reduces to the ordinary LS estimation.

- The mean and variance of $\hat{\boldsymbol{\beta}}(\lambda)$ are

$$E[\hat{\boldsymbol{\beta}}(\lambda)] = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta},$$
$$\text{Var}[\hat{\boldsymbol{\beta}}(\lambda)] = \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}.$$

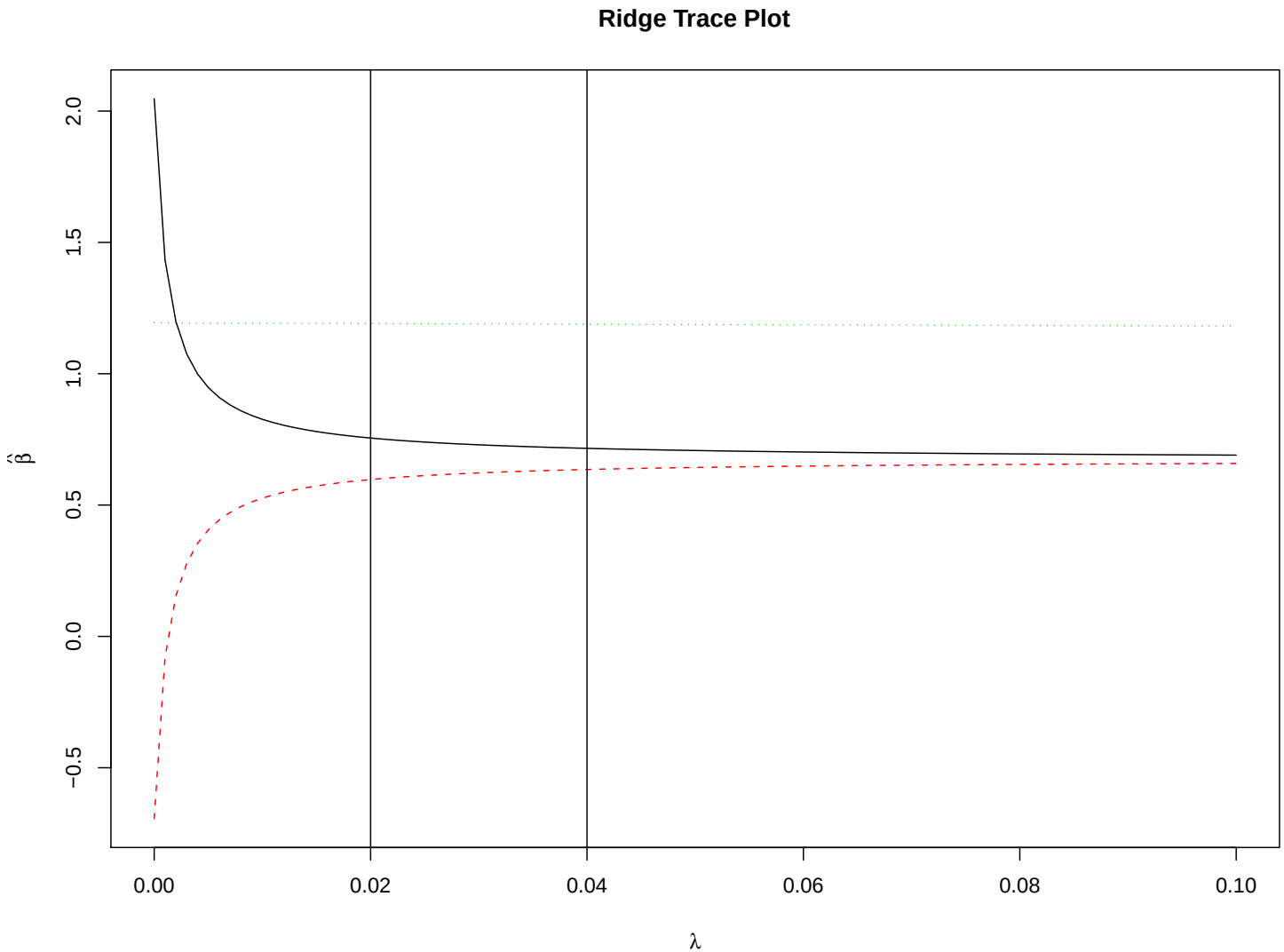
Choosing λ by Ridge Trace

- The λ is also referred to as the **biasing parameter**. When λ increases, the bias in $\hat{\boldsymbol{\beta}}(\lambda)$ will increase but the variance will decrease.
- For example, when the predictor variables are orthonormal so that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, the ridge estimator is

$$\hat{\boldsymbol{\beta}}(\lambda) = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}.$$

- The λ value can be estimated from the data by inspection of the **ridge trace plot**, which is a plot of $\hat{\beta}_i(\lambda)$'s versus λ with λ usually in the interval of $[0, 1]$.
- **Rule of Thumb**: Pick the smallest value of λ that produces a stable estimate of $\boldsymbol{\beta}$.

plot(lm.ridge(y∼x1+x2+x3,lambda=0:100/1000))



Choosing λ by automatic methods

In addition to the ridge trace plot, the λ value can also be chosen by automatic methods. Popular methods include:

(1) Cross-Validation (CV) estimator;

(2) Generalized Cross-Validation (GCV) estimator;

(3) Hoerl-Kennard-Baldwin (1975)'s HKB estimator;

(4) Lawless and Wang (1976)'s L-W estimator.

CV and GCV

- Let the prediction error of ridge regression be

$$\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}(\lambda)\right]^2, \tag{2}$$

  where $\hat{\boldsymbol{\beta}}_{(i)}(\lambda)$ is the ridge estimate calculated using ridge parameter λ but without the ith observation.

- This method is also called the "leave-one-out" cross-validation. The value of λ is chosen to minimize the PE in (2).

- When n is very large, the "leave-one-out" cross-validation can be computationally infeasible. One alternative method is to use the "leave-k-out" cross-validation or the N-fold cross-validation.

CV and GCV

- To avoid repeatedly calculating the ridge estimates $\hat{\boldsymbol{\beta}}_{(i)}(\lambda)$ for each i, we rewrite the above prediction error as

$$\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}(\lambda)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\left[Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}(\lambda)\right]^2}{[1-a_{ii}(\lambda)]^2},$$

  where $a_{ii}(\lambda)$ is the ith diagonal element of

$$\mathbf{A}(\lambda) = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T.$$

- The generalized cross-validation (GCV) is defined if we replace each $a_{ii}(\lambda)$ by their average $\mathrm{tr}[\mathbf{A}(\lambda)]/n$. In general, GCV is quite robust to the violation of model assumptions.

Ridge Regression

```
> fit.ridge1 <- lm.ridge(y~x1+x2+x3,lambda=0)
> fit.ridge1
                x1        x2       x3
-0.0070   3.0924  -1.0495   0.9911

> select(lm.ridge(y~x1+x2+x3,lambda=0:100/1000))
modified HKB estimator is 0.001446143
modified L-W estimator is 0.00231748
smallest value of GCV  at 0.029

> fit.ridge2 <- lm.ridge(y~x1+x2+x3,lambda=0.029)
> fit.ridge2
                x1        x2       x3
-0.0178   1.1046   0.9385   0.9868
```

More on Ridge regression

- Recall that the ridge estimator shrinks the least-squares estimator toward the origin. For example, when $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$,

$$\hat{\boldsymbol{\beta}}(\lambda) = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}.$$

  For this reason, the ridge estimators are also called the **shrinkage** estimators.

- $\hat{\boldsymbol{\beta}}(\lambda)$ is also the solution to

$$\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \\ \text{subject to} \quad & \sum_j \beta_j^2 \leq \text{Constant}. \end{aligned}$$

  By Lagrangian equations, we have $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^T\mathbf{Y} + \lambda\boldsymbol{\beta} = \mathbf{0}$.

LASSO regression

- LASSO (Tibshirani, 1996) stands for **Least Absolute Shrinkage and Selection Operator**. It aims to

$$\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \\ \text{subject to} \quad & \sum_j |\beta_j| \leq \text{Constant}. \end{aligned}$$

- LASSO regularizes the least square regression with an L_1-norm regularizer rather than an L_2-norm used in ridge regression.
- LASSO can lead to sparse solutions by shrinking the coefficients of the irrelevant/redundant variables to zero, and thus provide an effective technique for **variable selection**.

Variable selection

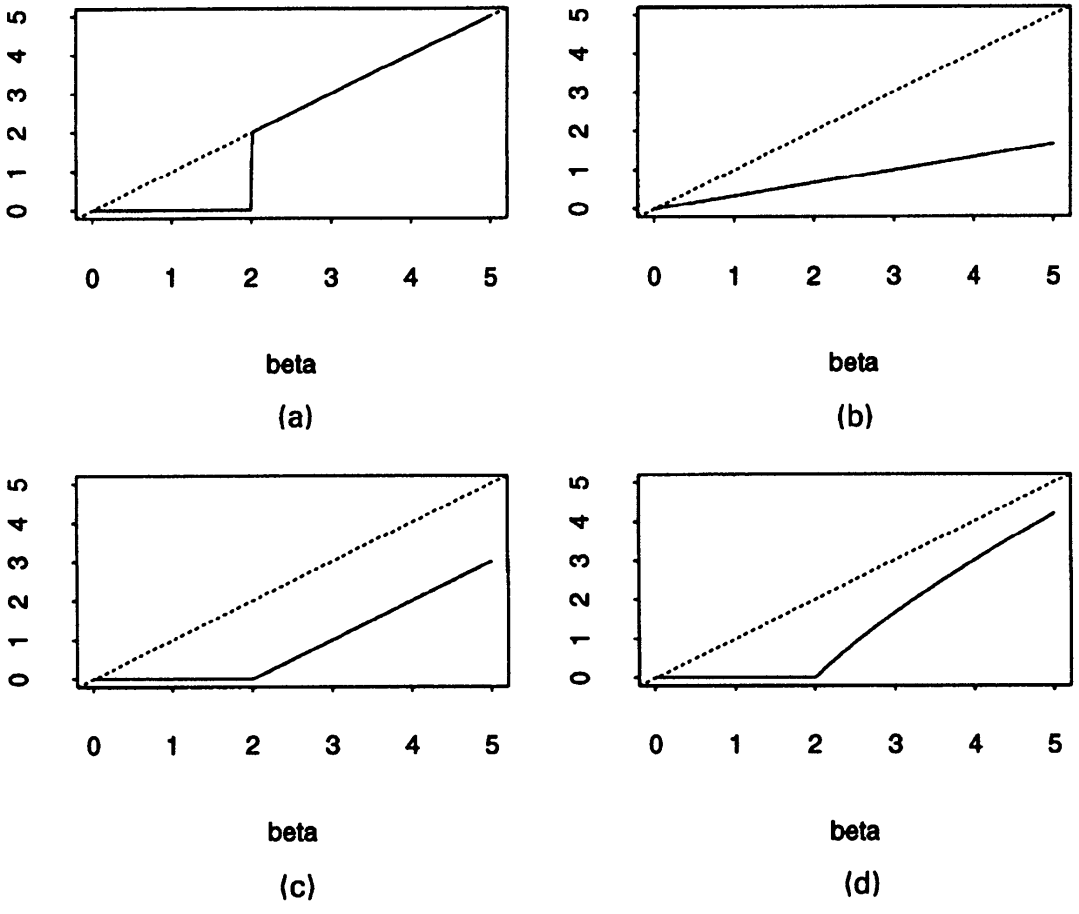


Fig. 1. (a) Subset regression, (b) ridge regression, (c) the lasso and (d) the garotte: ———, form of coefficient shrinkage in the orthonormal design case; ·····, 45°-line for reference

Variable selection

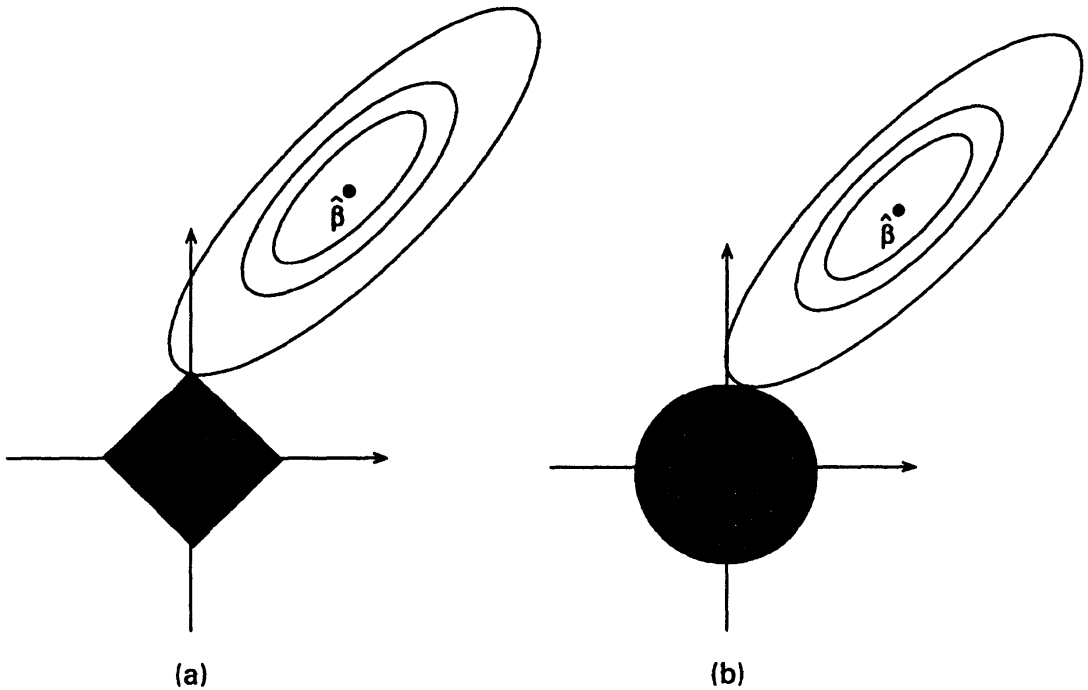


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression