Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

Ridge Regression

Ridge Regression

- A commonly used method to deal with multicollinearity.
- The idea is to constrain the fitted regression coefficients to achieve variance reduction at the expense of introducing bias.
- With suitably chosen *tuning parameter*, ridge regression can achieve good bias-variance trade-off.

Ridge Estimator

$$Y = X\beta + \epsilon, \quad E(\epsilon) = 0, \; \sigma^2\{\epsilon\} = \sigma^2\mathbf{I}$$

The ridge estimator is the minimizer of the ℓ_2 *penalized least-squares criterion*:

$$Q_\lambda(\mathbf{b}) = (Y - X\mathbf{b})^T(Y - X\mathbf{b}) + \lambda\mathbf{b}^T\mathbf{b}, \quad \mathbf{b} \in \mathbb{R}^p$$

- $\lambda \geq 0$: tuning parameter; $\lambda = 0 \Longrightarrow$ the least-squares criterion
- $\lambda\mathbf{b}^T\mathbf{b} = \lambda||\mathbf{b}|_2^2$: penalty on the size of the regression coefficients

- Setting the gradient of $Q_\lambda(\cdot)$ with respect to **b** to zero gives the normal equation:

$$\frac{\partial Q_\lambda(b)}{\partial \mathbf{b}} = 2X^T X\mathbf{b} + 2\lambda\mathbf{b} - 2X^T Y = 0$$

- The solution of the normal equation is the ridge estimator:

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

Standardization

Ridge regression is usually applied to standardized X variables (i.e., centered and scaled) and centered response variable to:

- make the amount of penalty comparable across different regression coefficients;
- make the regression intercept unaffected by the penalty: If both X and Y are centered, the estimated intercept $\hat{\beta}_{0,\lambda}$ will always be zero (and not dependent on λ).

Tuning Parameter

- $\lambda = 0 \Longrightarrow$ ordinary least-squares estimator $\hat{\beta}_{ols}$
- $\lambda > 0$: $||\hat{\beta}_{\lambda}||_2 < ||\hat{\beta}_{ols}||_2 \Longrightarrow$ shrinkage

Ridge Estimator: Bias

The ridge estimators are biased:

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

$$bias(\hat{\beta}_{k,\lambda}) := E(\hat{\beta}_{k,\lambda}) - \beta_k, \quad k = 1, \cdots, p-1$$

The amount of bias increases with the increase of λ.

Ridge Estimator: Variance

The variance of the ridge estimator:

$$\sigma^2\{\hat{\beta}_\lambda\} = \sigma^2(X^TX + \lambda\mathbf{I})^{-1}X^TX(X^TX + \lambda\mathbf{I})^{-1}$$

The variance decreases with the increase of λ.

Ridge Estimator: Bias-Variance Trade-off

There exists a $\lambda \geq 0$ that minimizes the (overall) *mean squared estimation error (msee)* of the regression coefficients:

$$\sum_{k=1}^{p-1} msee(\hat{\beta}_{k,\lambda}) = \sum_{k=1}^{p-1} var(\hat{\beta}_{k,\lambda}) + bias^2(\hat{\beta}_{k,\lambda})$$

- In theory, ridge regression will always beat the ordinary least-squares regression.
- In practice, we need to choose a good λ. tuning

Smoothing Operator S_λ

- Ridge regression conducts a linear smoothing of Y:

$$\hat{Y}_\lambda = X\hat{\beta}_\lambda = X(X^TX + \lambda\mathbf{I})^{-1}X^TY = S_\lambda Y$$

- $S_\lambda = X(X^TX + \lambda\mathbf{I})^{-1}X^T$ is referred to as a *smoothing operator/matrix*.

注意到：forall z, z'X'Xz=||Xz||2^2>=0
所以X'X半正定，tr(X'X)=Σi(λ_i)>=0
所以(X'X+λI)是正定的，特征值都为正，if λ>0
所以(X'X+λI)^-1也是正定的(特征值是上面的倒数)，if λ>0

- The trace of S_λ is referred to as the *effective number of parameters*.

tr(Sλ)=ΣSii,λ <p, if λ>0

注意:(X'X+λI)的λ的效果是
使得tr(*)=nλ+Σi(λ_i)>Σi(λ_i)>=0

tr[(X'X+λI)^-1X'X]=tr[(X'X+λI)^-1(X'X+λI-λI)]=tr(Ip)-λtr((X'X+λI)^-1)
=p-λtr((X'X+λI)^-1)>p, if λ>0

Deleted Residuals

The *deleted residuals* can be expressed through the corresponding (ordinary) residuals:

$$Y_i - \hat{Y}_{i(i),\lambda} = \frac{Y_i - \hat{Y}_{i,\lambda}}{1 - S_{ii,\lambda}}, \quad i = 1, \cdots, n,$$

where $S_{ii,\lambda}$ is the ith diagonal element of the smoothing matrix S_λ.

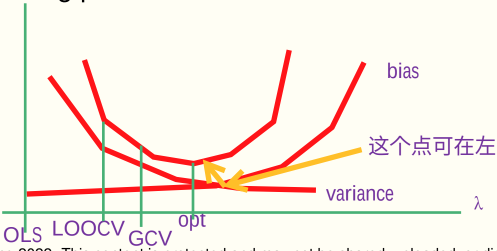
- Derived in the same way as under OLS.
- Hold for any linear smoothing of the form: $\hat{Y} = SY$ that arises from a least squares projection.

Leave-One-Out-Cross-Validation

(di,λ)^2

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i),\lambda})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_{i,\lambda}}{1 - S_{ii,\lambda}}\right)^2$$

The tuning parameter λ can then be chosen to minimize $CV(\lambda)$.



Generalized Cross-Validation (GCV)

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_{i,\lambda}}{1 - trace(S_\lambda)/n}\right)^2$$

- Replace individual diagonals $S_{ii,\lambda}$ by their average $trace(S_\lambda)/n$.
- Ease of computation
- Alleviates (to some degree) the tendency of LOOCV criterion favoring small λ (which leads to under-smoothing and overfitting).

Principal Component Regression (PCR)

PCR

- Another strategy to deal with multi-collinearity is to use a smaller number of linear combinations of the original variables that are orthogonal to each other, and then use these new variables in place of the original X variables in the regression.
- Two such methods are the *principal component regression (PCR)* and the *partial least squares regression (PLSR)*.
- PCR uses the first few *principal components* of the X variables in the model.

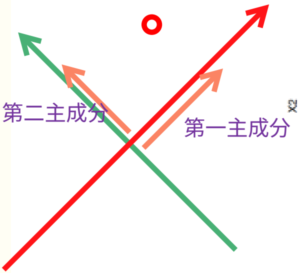
Standardization

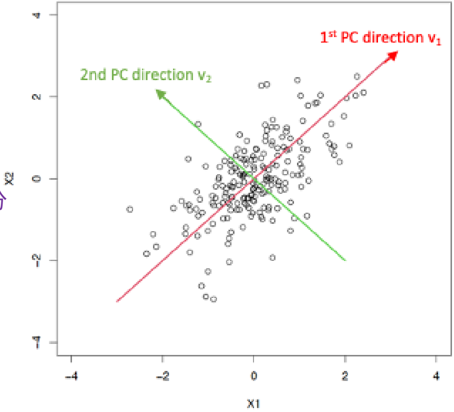
？与普通的standardized{scale()}是不是一样

As in ridge regression, in practice, the X variables are usually standardized before applying PCR. In the following, we assume the X variables have been standardized and the response variable Y has been centered, and the design matrix does not include the column of 1's.

Principal Component Analysis (PCA)

- Input data are projected to successive directions with maximum variation subject to orthogonality constraints with the previous directions.
- The main application of PCA is *dimension reduction*.
- The rationale of PCR is that the few leading principle components which explain the majority of variation in the X variables are more useful to explain the response variable (wishful thinking!).


第二主成分
第一主成分


1st PC direction v_1
2nd PC direction v_2
X1
X2
-4
-2
0
2
4

Singular Value Decomposition

Any $q \times r$ matrix $\mathbf{X}$ has a *singular value decomposition (SVD)* in the form:

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

- $\mathbf{U}$ and $\mathbf{V}$ are $q \times q$ and $r \times r$ orthogonal matrices, respectively:

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}_q, \mathbf{V}^T\mathbf{V} = \mathbf{I}_r$$

- $\mathbf{D}$ is a $q \times r$ diagonal matrix with non-negative entries $d_1 \geq d_2 \geq \cdots \geq d_{\min(q,r)} \geq 0$: *singular values* of $\mathbf{X}$.

- The rank of $\mathbf{X}$ equals the number of positive singular values, i.e, $d_1 \geq d_2 \geq \cdots \geq d_{rank(\mathbf{X})} > 0 = d_{rank(\mathbf{X})+1} = \cdots$
- $col(\mathbf{U}[, \mathbf{1} : \mathbf{rank}(\mathbf{X})]) = col\langle \mathbf{X} \rangle$, $col(\mathbf{V}[, \mathbf{1} : \mathbf{rank}(\mathbf{X})]) = row\langle \mathbf{X} \rangle$, i.e., the first $rank(\mathbf{X})$ columns of $\mathbf{U}$ generate the column space of $\mathbf{X}$ and the first $rank(\mathbf{X})$ columns of $\mathbf{V}$ generate the row space of $\mathbf{X}$.
- $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^T\mathbf{D}\mathbf{V}^T$: the columns of $\mathbf{V}$ are eigenvectors of $\mathbf{X}^T\mathbf{X}$ and the squared singular values d_j^2s are the eigenvalues.

Principal Components

Suppose $\mathbf{X}_{n \times p-1}$ is a data matrix (e.g., the design matrix). For $j = 1, \cdots, d_{rank(\mathbf{X})}$:

- The jth column of the matrix $\mathbf{V}$, v_j, is called the jth *principal component direction* of $\mathbf{X}$
- $z_j := Xv_j = d_j u_j$ is called the jth *principal component (PC)* of $\mathbf{X}$.

- The jth column of the matrix $\mathbf{V}$, $v_j \in \mathbb{R}^{p-1}$, consists of the linear combination coefficients (a.k.a. *loadings*) used to construct the jth PC, $z_j = \mathbf{X}v_j = d_j u_j \in \mathbb{R}^n$.
- $var(z_j) \propto d_j^2$ and is decreasing with the index j.
- The PCs are orthogonal to each other: $z_j^T z_{j'} = 0$ for $j \neq j'$.

Principal Component: Interpretation

It can be shown that, the jth PC direction v_j solves:

$$\max_{v \in \mathbb{R}^{p-1}} var(\mathbf{X}v), \quad \text{subject to } \|v\|_2 = 1,\ v^T(\mathbf{X}^T\mathbf{X})v_l = 0, \quad l = 1, \cdots, j\text{–}1$$

- The 1st PC $z_1 = \mathbf{X}v_1$ has the largest (sample) variance among all normalized linear combinations of the columns of $\mathbf{X}$;
- The 2nd PC $z_2 = \mathbf{X}v_2$ has the largest (sample) variance among all normalized linear combinations of the columns of $\mathbf{X}$ that are orthogonal to the 1st PC z_1;
- etc.

PCR vs. Ridge Regression vs. OLS

The hat matrix in OLS:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T = \sum_{j=1}^{rank(\mathbf{X})} u_j u_j^T$$

The OLS fitted values are:

$$\hat{Y}^{ols} = \mathbf{H}Y = \sum_{j=1}^{rank(\mathbf{X})} (u_j^T Y) u_j$$

$\{u_j^T Y : j = 1, \cdots, rank(\mathbf{X})\}$ are coordinates of Y with respect to the orthonormal basis $\{u_j : j = 1, \cdots, rank(\mathbf{X})\}$ of the column space of $\mathbf{X}$.

The smoothing matrix $\mathbf{S}_\lambda$ in ridge regression:

$$\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T = \sum_{j=1}^{rank(\mathbf{X})} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T$$

The Ridge fitted values are:

$$\hat{Y}_\lambda^{ridge} = \mathbf{S}_\lambda Y = \sum_{j=1}^{rank(\mathbf{X})} (\frac{d_j^2}{d_j^2 + \lambda} u_j^T Y) u_j$$

For $\lambda > 0$, $0 < \frac{d_j^2}{d_j^2+\lambda} < 1$, so ridge regression shrinks the coordinates of Y with respect to the orthonormal basis $\{u_j : j = 1, \cdots, rank(\mathbf{X})\}$. Moreover, greater amount of shrinkage is applied to the basis vectors with smaller d_j (i.e., increasing index j).

In PCR, the response Y is regressed to the first k ($1 \leq k \leq rank(\mathbf{X})$) PCs. Since the PCs are orthogonal, the estimated coefficient of the jth PC $z_j = d_j u_j$ is simply $\hat{\theta}_j = \frac{z_j^T Y}{z_j^T z_j} = \frac{d_j u_j^T Y}{d_j^2} = \frac{u_j^T Y}{d_j}$ and the PCR fitted values are:

$$\hat{Y}^{pcr,k} = \sum_{j=1}^{k} \hat{\theta}_j z_j = \sum_{j=1}^{k} (u_j^T Y) u_j$$

So PCR conducts a *hard thresholding* of the coordinates of Y with respect to the basis vectors $\{u_j : j > k\}$. The number of PCs, k, is a tuning parameter of PCR. If $k = rank(\mathbf{X})$, then PCR becomes OLS and no form of shrinkage occurs.

Ridge and PCR: When/How to Use?

- When there is high multicollinearity and prediction is the main goal

LASSO用1范数为penalty term

LASSO对variable selection很好，因为他会对有些variable取零为系数

- Not good for variable selection; Could be hard to interpret
- Common practice is to standardize the X variables and center the Y variable
- Tuning parameters (λ in Ridge and k in PCR) control bias-variance trade-off. In practice, these can be chosen through cross-validation.