Methods for high dimensional data

— Regularized linear regression models, sparse PCA

Classical multivariate analysis based on multivariate normal theory assumes $n > p$ (at most $n \geq p$), where n is the sample size and p is the dimension of the variables, or the dimension of the parameter space.

The case $n < p$ is the so called high dimensional case, most multivariate methods developed in classical theory need to be revised to apply to high dimensional data.

In the following, we take a brief look at the challenges and resolutions when the parameter space is of high dimensions.

In particular, we introduce the regularization method to deal with high dimension issues in linear regression models, and its applications to multivariate data in sparse PCA.

1 Review: Classical univariate linear regression model

Linear regression models formulates relationship between response Y and explanatory variable $X = (X_1, \cdots, X_r)$, assuming that $\mathbb{E}(Y|X)$, the conditional expectation of Y given the values of X, is linear in the X_i's. The model form is

$$\mathbb{E}(Y|X) = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_r.$$

Linear models encompass variable transformations.

In this section, we review classical linear regression models with univariate response variables and employ their vector-matrix formulation. Model parameters are typically estimated by the method of least squares or likelihood methods. There are a set of standard inferences. Model checking and goodness of fit diagnostic methods are well developed.

The model

In classical linear regression model, the response variable is modeled as linear combination of r explanatory variables $z_1, \cdots, z_r$, plus a random variation.

$$Y = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r + \varepsilon, \qquad \mathbb{E}(\varepsilon) = 0,\ V(\varepsilon) = \sigma^2.$$

The parameters β_i are to be estimated, along with the variance σ^2 of the error term ε. The formula implies conditional expectation.

$$\mathbb{E}(Y) = \mathbb{E}(Y \mid Z_1 = z_1, \cdots, Z_r = z_r) = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r$$

If $\hat{\beta}_i$'s are estimates of β_i's, then the estimated Y is

$$\hat{Y} = \widehat{\mathbb{E}}(Y \mid z_1, \cdots, z_r) = \hat{\beta}_0 + \hat{\beta}_1 z_1 + \cdots + \hat{\beta}_r z_r$$

If sample data are collected, let Y_j denote the jth measurement of the response variable Y when the values of the r explanatory variables are set at $z_{j1}, \cdots, z_{jr}$. The regression model to be fitted by the data has the form

$$Y_j = \beta_0 + \beta_1 z_{j1} + \cdots + \beta_r z_{jr} + \varepsilon_j, \quad j = 1, \cdots, n.$$

Under the assumption that the n observations are independent, the error terms have the properties

$$\mathbb{E}(\varepsilon_j) = 0, \qquad V(\varepsilon_j) = \sigma^2, \qquad cov(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j.$$

The objective is to estimate the β_i's and σ^2 based on the observed data.

Model estimations with observed data

Assume there are n observations $(y_1, \cdots, y_n) = \boldsymbol{y}'$. For each $i = 1, \cdots, n$, y_i denotes an observed response at given values $\{z_{i,k}, k = 1, \cdots, r\}$ of r explanatory variables $\{z_1, \cdots, z_r\}$.

To model the effects of the of explanatory variables on the values of the response variable $y_i, i = 1, \cdots, n$, we may fit a univariate linear regression model.

The regression model to be fitted with n observed data points can be written in matrix form as

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_j \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & \cdots & z_{1r} \\ \vdots & \vdots & & \vdots \\ 1 & z_{j1} & \cdots & z_{jr} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & z_{n1} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_j \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

In matrix notations, the above n model equations can be written as

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{Z}_{n\times(r+1)}\boldsymbol{\beta}_{(r+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

or

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{Z}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

or simply

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

In the above expressions, we treat data $\boldsymbol{y}$ as a realization of the random vector $\boldsymbol{Y}$, p be the number of model coefficient parameters. When the model contains an intercept term β_0, we have $p = r + 1$. A common practice is to center $\boldsymbol{y}$ at its sample mean (for derivation convenience), so $\beta_0 = 0$, the model has no intercept, then we have $p = r$.

Matrix $\boldsymbol{Z}$ is called the **design matrix** of the model. The independence, mean zero, and common variance assumptions on the error terms of the observations can be written in vector-matrix form as

$$\mathbb{E}(\boldsymbol{\varepsilon}) = 0, \qquad Cov(\boldsymbol{\varepsilon}) = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{I}_n.$$

Least squares estimation of model parameters

When $n > r$,the parameters β_i's in the classical linear regression model parameters are commonly estimated by the method of least squares (LS).
The method of approach is to minimize the sum of squares of errors (error = observed - estimated), written as

$$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}) = \sum_{j=1}^{n}(Y_j - \beta_0 - \beta_1 z_{j1} - \cdots - \beta_r z_{jr})^2$$

When $\boldsymbol{Z}$ is of full rank with $rank(\boldsymbol{Z}) = r + 1$, the LS estimator of $\boldsymbol{\beta}$ that minimizes the sum of squares of errors $S(\boldsymbol{\beta})$ is

$$\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$$

(Notes: When $rank(\boldsymbol{Z}) < r + 1$, $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ is replace by a generalized inverse of $\boldsymbol{Z}'\boldsymbol{Z}$.)

Proof. In the following, we give a proof of $\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$ which uses the concept of orthogonal projection.

Denote $\boldsymbol{\beta}_* = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$.

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\boldsymbol{Y} - \boldsymbol{Z\beta})'(\boldsymbol{Y} - \boldsymbol{Z\beta}) \\ &= [\boldsymbol{Y} - \boldsymbol{Z\beta}_* + \boldsymbol{Z}(\boldsymbol{\beta}_* - \boldsymbol{\beta})]'[\boldsymbol{Y} - \boldsymbol{Z\beta}_* + \boldsymbol{Z}(\boldsymbol{\beta}_* - \boldsymbol{\beta})] \\ &= (\boldsymbol{Y} - \boldsymbol{Z\beta}_*)'(\boldsymbol{Y} - \boldsymbol{Z\beta}_*) + (\boldsymbol{\beta}_* - \boldsymbol{\beta})'\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{\beta}_* - \boldsymbol{\beta}) + 2(\boldsymbol{Y} - \boldsymbol{Z\beta}_*)'\boldsymbol{Z}(\boldsymbol{\beta}_* - \boldsymbol{\beta}) \end{aligned}$$

The last term is a zero vector, because

$$\boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{Z\beta}_*) = \boldsymbol{Z}'\boldsymbol{Y} - \boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} = \boldsymbol{Z}'\boldsymbol{Y} - \boldsymbol{Z}'\boldsymbol{Y} = \boldsymbol{0}_{(r+1)\times 1}$$

Thus

$$(\boldsymbol{Y} - \boldsymbol{Z\beta}_*)'\boldsymbol{Z} = \boldsymbol{0}_{1\times(r+1)} \qquad \Rightarrow \qquad (\boldsymbol{Y} - \boldsymbol{Z\beta}_*)'\boldsymbol{Z}(\boldsymbol{\beta}_* - \boldsymbol{\beta}) = 0$$

Therefore,

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\boldsymbol{Y} - \boldsymbol{Z\beta}_*)'(\boldsymbol{Y} - \boldsymbol{Z\beta}_*) + (\boldsymbol{\beta}_* - \boldsymbol{\beta})'\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{\beta}_* - \boldsymbol{\beta}) \\ &= \|\boldsymbol{Y} - \boldsymbol{Z\beta}_*\|^2 + \|\boldsymbol{Z}(\boldsymbol{\beta}_* - \boldsymbol{\beta})\|^2 \\ &\geq \|\boldsymbol{Y} - \boldsymbol{Z\beta}_*\|^2 \end{aligned}$$

The equality holds if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_* = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$. In other words,

$$S(\boldsymbol{\beta}^*) = \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$$

which means $\boldsymbol{\beta}_* = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{LS}$, the least squares estimator of β in the linear regression model.

□

Remarks on Least Squares Estimator (LSE)

- The above least squares estimation results can also be stated using matrix calculus, by taking derivative of $S(\boldsymbol{\beta})$ with respect to vector β, setting $\frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{\beta}) = -2\boldsymbol{Z}'\boldsymbol{Y} + \boldsymbol{Z}'\boldsymbol{Z\beta} = \boldsymbol{0}_{r+1}$ and solve for $\boldsymbol{\beta}$ (as we did in class).
- $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ is called the **hat matrix**, a symmetric projection matrix satisfying $\boldsymbol{H}^2 = \boldsymbol{H}$ (idempotent).
- For observed data $\boldsymbol{y}$, $\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{Hy}$ are the model fitted values of $\boldsymbol{y}$.
- $\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ are the residuals of the fitted model.
- Both $\hat{\beta}$ and $\hat{\varepsilon}$ can be obtained from the design matrix $\boldsymbol{Z}$ and observed response variable $\boldsymbol{Y}$.
- Both $\hat{\beta}$ and $\hat{\varepsilon}$ can be expressed as linear combinations of observed response variables $Y_1, \cdots, Y_n$.

Geometry of Least Squares

- Minimizing $S(\beta)$ is equivalent to finding minimal error norm in projecting $\boldsymbol{Y}$ to $C(\boldsymbol{Z})$, the column space of $\boldsymbol{Z}$. The optimal minimum is achieved by orthogonal projection.
- $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$ is the result of the orthogonal projection of $\boldsymbol{Y}$ to $C(\boldsymbol{Z})$.
- $\boldsymbol{Z}'\hat{\varepsilon} = 0$, the residual vector $\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ is orthogonal to $C(\boldsymbol{Z})$, a result of the orthogonal projection, can be derived via $\boldsymbol{Z}'\hat{\varepsilon} = \boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Z}'\boldsymbol{Y} - \boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} = \boldsymbol{Z}'\boldsymbol{Y} - \boldsymbol{Z}'\boldsymbol{Y} = 0$.

- $\hat{\boldsymbol{Y}}'\hat{\varepsilon} = 0$, the residual vector $\hat{\varepsilon}$ is orthogonal to the projected, fitted value $\hat{\boldsymbol{Y}}$, another result of the orthogonal projection, can be obtained from the derivation $\hat{\boldsymbol{Y}}'\hat{\varepsilon} = (\boldsymbol{HY})'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}'(\boldsymbol{H} - \boldsymbol{H}^2)\boldsymbol{Y} = 0$, by $\boldsymbol{H}^2 = \boldsymbol{H}$.

Sum of squares decomposition

The total response sum of squares

$$\sum_{i=1}^{n} Y_i^2 = \boldsymbol{Y}'\boldsymbol{Y}$$

By $\hat{\mathbf{Y}}'\hat{\varepsilon} = 0$, we obtain the decomposition

$$\boldsymbol{Y}'\boldsymbol{Y} = (\hat{\boldsymbol{Y}} + \hat{\varepsilon})'(\hat{\boldsymbol{Y}} + \hat{\varepsilon}) = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} + \hat{\varepsilon}'\hat{\varepsilon}$$

To consider the centralized sum of squares, let $\bar{Y} = \sum_{i=1}^{n} Y_i/n$,

$$\boldsymbol{Y}'\boldsymbol{Y} - n\bar{Y}^2 = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} - n\bar{Y}^2 + \hat{\varepsilon}'\hat{\varepsilon}.$$

Let $\bar{\hat{Y}} = \sum_{i=1}^{n} \hat{Y}_i/n$. Using the fact $\bar{Y} = \bar{\hat{Y}}$ (exercise), the above equation can be written in terms of sums of squares. which can be stated as

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

The decomposition can be described as

$$\sum (\text{total sum of squares about mean})^2 = \sum (\text{regression sum of squares})^2 + \sum (\text{residual sum of squares})^2$$

On simplified notations,

$$SS_{total} = SS_{reg} + SS_{err}$$

Coefficient of determination

The sum of squares decomposition provides a measure of the quality of the regression model, in terms of the closeness of the fitted model to the observed data.

The measure is called coefficient of determination, also commonly known as R-square.

$$R^2 = \frac{SS_{reg}}{SS_{total}} = 1 - \frac{SS_{err}}{SS_{total}}$$

The coefficient of determination measures the proportion of the total variation "explained" by the fitted regression model, or by the explanatory variables. The variation is in terms of sums of squares about the mean.

Checking normal assumptions

To have statistical inference on the model fitting and parameter estimation, normality assumption is needed.

To check if observations $Y_1, \cdot, Y_n$ are indeed from a normal distribution, Q-Q plot is a quick diagnostic tool.

Normal Q-Q plot is a graphical method to check the normality assumption of the ε_i's.

The method of normal Q-Q plot

Let $\varepsilon_{(1)} \leq \cdots \leq \varepsilon_{(n)}$ be the ordered values of observed $\varepsilon_1, \cdots, \varepsilon_n$, called order statistics of $\varepsilon_1, \cdots, \varepsilon_n$.

If ε_i's are from independent random variables $\sim N(\mu, \sigma^2)$, then there is a linear relationship

$$\varepsilon_{(i)} = \mu + \sigma Z_{(i)}$$

where

$$Z_{(1)} \le \cdots \le Z_{(n)}$$

denote the order statistics of n independent observations $Z_1, \cdots, Z_n$, with

$$Z_i \sim N(0,1), \qquad i = 1, \cdots, n.$$

The so called Q-Q plot (quantile-quantile plot) for the ε_i's is the plot of quantile points

$$\left(\Phi^{(-1)}\Big(\frac{i}{n+1}\Big),\ \varepsilon_{(i)}\right), \qquad i =, 1 \cdots, n,$$

where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

is the cumulative distribution function of $N(0,1)$, and (proof is based on order statistics, as in Mathematical Statistics Methods II)

$$\Phi^{(-1)}\left(\frac{i}{n+1}\right) \approx \mathbb{E}[Z_{(i)}]$$

The plot should resemble a straight line, if the ε_i's are indeed i.i.d. observations from a normal distribution.

In linear regression case, the true residual ε_i's are not observed. If the normality assumption on the Y_i's and $varepsilon_i$'s are appropriate, the fitted residual $\hat{\varepsilon}_i$'s, though not exactly, but are close to i.i.d. $\sim N(0, \hat{\sigma}^2)$. Common normality checks such as the Normal Q-Q plot are applied to the fitted residuals $\hat{\varepsilon}_i$'s.

Further optimal properties of LS-ML estimators for linear models

To obtain statistical inference such as confidence intervals of parameter estimation, commonly we impose normality assumption on the error term, thus the model becomes

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{Z}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \qquad \boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

Assume there are n observations $(y_1, \cdots, y_n) = \boldsymbol{y}'$. For each $i = 1, \cdots, n$, y_i denotes an observed response at given values $\{z_{i,k}, k = 1, \cdots, r\}$ of r explanatory variables $\{z_1, \cdots, z_r\}$.

To estimate the model parameter vector $\boldsymbol{\beta}$, with the normality distribution assumption, we may use the maximum likelihood (ML) method in addition to the the least squares (LS) method. In this case, the ML method and LS method yield the same estimator $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}_{LS} \in \mathbb{R}_p$ which minimizes the sum of squared errors. We may write the estimator as

$$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2$$

When $n \ge p$, and when $\boldsymbol{Z}$ is of full rank p, the LS estimate of $\boldsymbol{\beta}$ can be written explicitly as

$$\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

The Least Squares estimator (as a random variable based on sample data) is unbiased,

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\beta}}_{LS}) &= \mathbb{E}[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}] \\ &= \mathbb{E}\left[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\right] \\ &= \mathbb{E}[\boldsymbol{\beta} + (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{\varepsilon}] = \boldsymbol{\beta} + (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\,\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{\beta}\end{aligned}$$

where the last equation used $\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}_n$.

In addition, the LS estimator achieves optimal variance properties among all unbiased linear estimators.

Proof.

We are to show that the LS estimator achieves an optimal variance-covariance property.

Any unbiased linear estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be expressed as

$$\hat{\boldsymbol{\beta}} = A\boldsymbol{Y}, \qquad \mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$$

where A is a $p \times n$ scalar matrix. Define $p \times n$ matrix $B = A - (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$, we can write

$$\hat{\boldsymbol{\beta}} = A\boldsymbol{Y} = \left[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' + B\right]\boldsymbol{Y}$$

Since $\hat{\boldsymbol{\beta}}$ is unbiased,

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\left(\left[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' + B\right](\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\right) \\ &= \boldsymbol{\beta} + B\boldsymbol{Z}\boldsymbol{\beta} + \left[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' + B\right]\mathbb{E}(\varepsilon) \ = \ \boldsymbol{\beta} + B\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{\beta}\end{aligned}$$

The unbiasedness of $\hat{\boldsymbol{\beta}}$ implies $B\boldsymbol{Z} = 0_{p\times p}$. Then

$$\begin{aligned}Cov(\hat{\boldsymbol{\beta}}) &= A\,Cov(\boldsymbol{Y})A' = \sigma^2 AA' \\ &= \sigma^2\left[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' + B\right]\left[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' + B\right]' \\ &= \sigma^2\left[(\boldsymbol{Z}'\boldsymbol{Z})^{-1} + (\boldsymbol{Z}'\boldsymbol{Z})^{-1}(B\boldsymbol{Z})' + (B\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1} + BB'\right] \\ &= \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1} + \sigma^2 BB' \ = Cov(\hat{\boldsymbol{\beta}}_{LS}) + \sigma^2 BB'\end{aligned}$$

Thus, for any unbiased linear estimator $\hat{\boldsymbol{\beta}}$, its covariance matrix

$$Cov(\hat{\boldsymbol{\beta}}) \ = \ Cov(\hat{\boldsymbol{\beta}}_{LS}) + \sigma^2 BB'$$

where $\sigma^2 BB'$ is a symmetric, positive semi-definite matrix with non-negative eigenvalues.

In the case $p = 1$, we have $\sigma^2 BB' \ge 0$, then the above equation implies

$$var(\hat{\beta}) \ \ge \ var(\hat{\beta}_{LS})$$

For $p \ge 2$, the LS estimator achieves optimal covariance in the sense that the covariance matrix of any unbiased linear estimator is the covariance of LSE plus a positive semi-definite matrix.

In addition, we can obtain

$$Cov(\hat{\boldsymbol{\beta}}) \ = \ Cov(\hat{\boldsymbol{\beta}}_{LS}) + \sigma^2 BB' \qquad \Rightarrow \qquad trace\left[Cov(\hat{\boldsymbol{\beta}})\right] \ge trace\left[Cov(\hat{\boldsymbol{\beta}}_{LS})\right]$$

which means

$$\sum_{k=1}^{p} \mathbb{E}\left[(\hat{\beta}_k^{LS} - \beta_k)^2\right] = \sum_{k=1}^{p} var(\hat{\beta}_k^{LS}) = trace\left[Cov(\hat{\boldsymbol{\beta}}_{LS})\right] \le trace\left[Cov(\hat{\boldsymbol{\beta}})\right] = \sum_{k=1}^{p} var(\hat{\beta}_k)$$

That is, for any unbiased estimator $\hat{\boldsymbol{\beta}}$,

$$\sum_{k=1}^{p} var(\hat{\beta}_k^{LS}) \le \sum_{k=1}^{p} var(\hat{\beta}_k)$$

Thus LS estimators has the smallest total variance among all unbiased linear estimators.

□

2 Issues of high dimensional parameter space in linear models

The high dimensional problem refers to the situation when the number of variables p is large, which is a very common phenomenon in the large data era.

Problems with high dimensional parameter space

- $n < p$

  When $n < p$, that is, when the dimension of the parameter space (e.g., $\mathbb{R}^p$) is higher than the number of observations,

  - The linear model has infinitely many solutions, thus not well defined.
  - The $p \times p$ matrix $\boldsymbol{Z}'\boldsymbol{Z}$ has rank $\leq n < p$, thus does not have a proper inverse.

- Collinearity

  Similar problems could occur even when $n \geq p$, in the thorny case of $\boldsymbol{Z}'\boldsymbol{Z}$ having very small eigenvalues ≈ 0, thus of rank practically $< p$, that is, $\boldsymbol{Z}'\boldsymbol{Z}$ is practically not invertible, or $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ has very large eigenvalues, making the estimate $\hat{\boldsymbol{\beta}}_{LS}$ of very large variance and unstable.

  A typical such case in practice is when near collinearity exists among explanatory variables.

Related statistical inference problems

- As shown in the above, linear model coefficient estimators are unbiased, thus of small bias in practice. However, the variance can be large, which yields poor prediction power.

  We may consider improving the predictive accuracy by allowing small amount of bias in order to reduce the variance.

- In linear regression model parameter estimation, a non-zero estimate is assigned to the coefficient of every explanatory predictor variable. When the number of variables p is large, we may desire to focus on only a subset of explanatory variables of sizable, non-zero coefficients, and disregard or zero out other non-important variables of near-zero coefficients (after standardizing), therefore achieve better interpretive ability.

3 Regularization methods in high dimensional linear models

A common methods dealing with the above mentioned issues in linear regression is the regularization or penalization method. The regularization method adds a penalty term to the loss function, the sum of residuals $S(\boldsymbol{\beta})$, effectively reducing the magnitude or the number of non-zero coefficients in the model.

3.1 Ridge linear regression model

When proper $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ does not exist (as in the high dimensional case of $n < p$) or of very large eigenvalues (as when input variables are nearly linearly dependent, called collinearity), instead of

$$\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$$

we may consider a modified estimator

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z} + \ term\ )^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

The estimator which contains the nonzero $term$ will no longer be unbiased. The hope is that, with the added $term$, the inverse matrix $(\boldsymbol{Z}'\boldsymbol{Z} + \ term)^{-1}$ exists and have not-too-large eigenvalues, yet the $term$ is small enough to have a reasonable estimate of $\boldsymbol{\beta}$ not too biased, possibly with a smaller variance.

Consider a solution with adding a diagonal matrix, with

$$term = \lambda\boldsymbol{I}$$

then $\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{I}$ is invertible for large enough $\lambda > 0$, and the modified estimator is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{y} = \hat{\boldsymbol{\beta}}_{Ridge}$$

which is the estimator by the Ridge linear regression model.

Mathematically equivalently (see below), the Ridge regression estimator optimizes

$$\min_{\beta_0,\boldsymbol{\beta}} \left(\|\boldsymbol{y} - \boldsymbol{\beta}_o - \boldsymbol{Z}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \right),$$

where the intercept term $\boldsymbol{\beta}_o = \beta_o\boldsymbol{1}_n$, with $\boldsymbol{1}_n$ being the n-vector with every entry $= 1$. $\boldsymbol{Z}, \boldsymbol{\beta}$ are adjusted accordingly to exclude the intercept term.

Assuming the data are centered at sample means so the intercept term β_o will have estimate $\hat{\beta}_o = \bar{y}_{centered} = 0$.

Then the Ridge estimator for the linear regression has the simpler form

$$\hat{\boldsymbol{\beta}}_{Ridge} = \arg\min_{\boldsymbol{\beta}} \Big(\underbrace{\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2}_{loss} + \lambda \underbrace{\|\boldsymbol{\beta}\|_2^2}_{penalty} \Big) \tag{1}$$

where the 2-norm

$$\|\boldsymbol{\beta}\|_2^2 = \sum_i \beta_i^2$$

is the Euclidean norm.

Also equivalently, Ridge regression can be written as solving the Lagrangian problem

$$\min_{\beta} \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 \qquad \text{under the constraint} \qquad \|\boldsymbol{\beta}\|_2^2 \leq s. \tag{2}$$

This formation shows the constraint ($a.k.a.$ **regularization** or penalty) on the size of the components of β explicitly.

The two formulations (1) and (2) are equivalent, and there exists a one-to-one relationship between the λ in (1) and the s in (2).

By taking derivative in (1),

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right) = -2\boldsymbol{Z}'\boldsymbol{y} + 2\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} = \boldsymbol{0}_p$$

$$\Rightarrow \quad (\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{Z}'\boldsymbol{y}$$

For λ values making $(\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{I})$ invertible, we obtain the Ridge estimator of the coefficient parameters:

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{y} \tag{3}$$

Ridge regression properties

- The tuning parameter $\lambda \in [0, \infty)$ in the Ridge regression model controls the penalty or the regulatory term.
- $\lambda > 0$ puts a constraint on the size of $\|\boldsymbol{\beta}\|_2$, therefore induces a reduction, termed as a **shrinkage**, on the magnitude of the parameter estimates.
- When $\lambda = 0$, we get back to $\hat{\boldsymbol{\beta}}_{Ridge} = \hat{\boldsymbol{\beta}}_{LS}$.
- λ is a measure of the **shrinkage** of $\|\boldsymbol{\beta}\|_2$. The larger λ, that more constraint on the size of $\|\boldsymbol{\beta}\|_2$.

  As $\lambda \to \infty$, the norm $\|\hat{\boldsymbol{\beta}}_{Ridge}\|_2 \to 0$.

  The limit of $\lambda \to \infty$, denoted as $\lambda = \infty$, corresponds to $\hat{\boldsymbol{\beta}}_{Ridge} = \boldsymbol{0}_p$.
- In general, the bias $\|\hat{\boldsymbol{\beta}}_{Ridge} - \boldsymbol{\beta}\|$ increases as λ (the amount of shrinkage) increases.
- In general, the variance of each component $\hat{\beta}_i$ of $\hat{\boldsymbol{\beta}}_{Ridge}$ decreases as λ increases.
- The overall mean squared error $\mathbb{E}\|\hat{\boldsymbol{\beta}}_{Ridge} - \boldsymbol{\beta}\|_2^2$ can be reduced (compared with LS estimator) for a range of λ, thus improve prediction accuracy.
- Ridge estimator $\hat{\boldsymbol{\beta}}_{Ridge}$ will include each and all variables in the model by having $\hat{\beta}_i \neq 0$ for all coefficients.

Remarks on properties of Ridge regression

- When we encounter collinearity among explanatory variables in fitting a linear regression model, Ridge linear model often provides better, more stable parameter estimations.

  For example, Ridge linear models greatly reduce the undesirable situation when two highly correlated explanatory variables ended up with large coefficient estimates of opposite signs.
- There are other common notations, such as using letter X in the place of Z, and write

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

  As mentioned, the formulation indicates that it is designed to deal with the situation when $X^T X$ is close to singularity.

- The shrinkage property

  As mentioned above, the added tuning parameter λ in the Ridge estimator makes the estimated $\|\boldsymbol{\beta}\|_2$ smaller.

  Example:

  Using a very simple illustrative example, where we assume that the model has $Z^T Z = I$ (understandably LS works in this case so Ridge is not really needed), then the coefficient estimator for the ith component

  $$\hat{\beta}_i^{ridge} = \frac{\hat{\beta}_i^{LS}}{\lambda + 1} \qquad \Rightarrow \qquad \left|\hat{\beta}_i^{ridge}\right| < \left|\hat{\beta}_i^{LS}\right|$$

  In most cases, Ridge estimates of linear regression coefficients are smaller in magnitude than the unbiased Least Squares estimates.
- The bias-variance tradeoff

  The estimates of Ridge Regression is biased, as can be seen from

  $$\mathbb{E}(\hat{\boldsymbol{\beta}}^{Ridge}) \neq \mathbb{E}(\hat{\boldsymbol{\beta}}^{LS}) = \boldsymbol{\beta}$$

  However for a range of λ's, the components of the Ridge estimator can achieve smaller variance $var(\hat{\beta}_i)$ than that of the least squared estimations.

  For a suitable range of λ's, this bias-variance tradeoff can lead to a smaller **Mean Squared Error** (MSE)

  $$MSE(\hat{\beta}_i) = \mathbb{E}[(\hat{\beta}_i - \beta_i)^2] = \Big(\text{bias}(\hat{\beta}_i)\Big)^2 + \text{var}(\hat{\beta}_i)$$

  for each coefficient estimate. MSE is an important common measure of goodness of an estimator.
- Note that intercept term is not regularized thus not penalized, which is reasonable.
- Scaling of the input variables will affect the model estimates.

  Often, before fitting Ridge model, the input variables are normalized, e.g., with variance = 1, especially when there are large variations in magnitude and spread among the original input variables.

3.2 LASSO linear regression model

In general, Ridge estimators will have a non-zero estimate for every component β_i in $\boldsymbol{\beta} \in \mathbb{R}^p$.

In high-dimensional linear modeling situation with a large number of explanatory variables, among them often there are extraneous variables that actually play no role in predicting the response variable y.

The situation can be phrased as the following: There is a subgroup of coefficients β_i's with true value $= 0$.

We may state the problem in model variable selection terms: When it is reasonable to assume that many of the coefficient parameters with true value $\beta_i = 0$, it is desirable for the model variable selection to be **sparse**. That is, the model selects only a small subset of variables with their coefficient estimates $\hat{\beta}_i$ non-zero. In other words, the model should produce many estimates $\hat{\beta}_i = 0$.

The sparse property is desirable for model interpretation, especially when the total number of predictors p is large.

We know that Ridge estimator does not make the cut, since $\hat{\beta}_i^{Ridge} \neq 0$ for every β_i in $\boldsymbol{\beta} \in \mathbb{R}^p$.

It turns out that replacing the 2-norm $\|\boldsymbol{\beta}\|_2$ in the penalty term in the Ridge linear model by the 1-norm $\|\boldsymbol{\beta}\|_1$ does a good job in the desired selective variable selection.

LASSO stands for Least Absolute Shrinkage and Selection Operator. LASSO linear model estimator optimizes

$$\min_{\beta} \left\{ \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

where

$$\|\boldsymbol{\beta}\|_1 = \sum_i |\beta_i|$$

is the 1-norm or the absolute-value norm.

Equivalently, LASSO regression can be written as solving

$$\min_{\beta} \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 \qquad \text{under the constraint} \qquad \|\boldsymbol{\beta}\|_1 \leq s.$$

Remarks on LASSO

- A characteristic of LASSO estimator $\hat{\boldsymbol{\beta}}_{Lasso}$ is its **sparsity**. That is, the parameter estimates by the LASSO model result in $\hat{\beta}_i = 0$ for many components β_i of $\boldsymbol{\beta}$.
- While Ridge regression coefficient estimator is still linear in the y_i's, the LASSO estimator $\hat{\boldsymbol{\beta}}_{Lasso}$ is non-linear, thus does not have a form analogous to the Ridge estimator (3).
- Unlike the Ridge estimator, there is no closed form solution such as (3) for $\hat{\boldsymbol{\beta}}_{Lasso}$, the coefficient estimator of the LASSO estimator.
- Numerical methods have to be used to approximate LASSO estimators.

  Computing the LASSO solution $\hat{\boldsymbol{\beta}}_{Lasso}$ is a quadratic programming problem.
- In the examples show in the demo in class, we can see that the current algorithms are efficient in finding LASSO solutions for each λ in a sufficiently dense set or grid of the range of λ.

3.3 Elastic Net linear regression model

A combination of the Ridge and LASSO methods, the Elastic Net linear regression model optimizes

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\}$$

with $\lambda_1, \lambda_2 \geq 0$. A more common parametrization for the Elastic Net optimization is

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 + \lambda\Big((1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2^2\Big) \right\}$$

with $\lambda \geq 0$ and $\alpha \in [0, 1]$.

If a few covariates z_i's are correlated, Ridge tends to keep them similar sized, LASSO tends to keep one of them non-zero, Elastic Net at certain value, say, near $\alpha = 0.5$, tends to either keep them all in or to leave them all out. This property is only obvious in some data settings.

Optimization algorithms (e.g., coordinate descent) can obtain parameter estimates efficiently.

In the examples in the demo in class, the methods of Ridge, LASSO and Elastic Net are compared using simulation and real data (for an example using LASSO). The examples illustrate and compare basic characteristics of each method.

3.4 Vector norms and regularized linear regression models

The sparsity property of the LASSO method comes from mathematical properties of the 1-norm of vectors used to regulate the coefficient vector β.

A class of vector norms, called p-norm, is defined as

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}, \qquad p \in [1, \infty)$$

for vector $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$.

- The 2-norm $\|x\|_2$ is the most familiar Euclidean norm.

  The the unit-norm set $\{x : \|x\|_2 = 1\}$ forms a unit circle.

  Ridge regression coefficient estimator shrinks to a disk $\|\boldsymbol{\beta}\|_2^2 \leq s$.
- The 1-norm $\|x\|_1$ is very useful, as shown in the LASSO regression method.

  The the unit-norm set $\{x : \|x\|_1 = 1\}$ forms a squared diamond.

  The corners of the diamond shaped $\{x : \|x\|_1 = c\}$ are on the axes, which other components $= 0$.

  This geometric property gives the sparsity of the LASSO estimator.
- The limiting case $\|x\|_\infty$ has the unit-norm set $\{x : \|x\|_\infty = \max_i |x_i|\}$
- Sometimes the term "0-norm" is used (which is not a vector norm, mathematically speaking).

$$\|x\|_0 = \sum_i 1_{\{x_i \neq 0\}} = \text{count of number of non-zero component } x_i$$

The choice of vector norm in the regularization or penalty term in the Ridge and LASSO linear regression models has made a huge difference in the input variable selection outcome in each model, (as shown in the demo examples). The ensuing applications abound.

4 Sparse PCA

In this section, we introduce an extension of the Principal Component Analysis method to the high dimensional case when the number of variables is large. The method is termed as Spares PCA.

Currently the algorithm used to implement Spares PCA is based on regularized regression methods discussed above.

Given an $n \times p$ data matrix X, the task of finding the first principal component in PCA can be described as

$$\text{Maximize } \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a} \qquad \text{under the constraint} \quad \|\boldsymbol{a}\|^2 = 1,$$

where $\boldsymbol{S}$ is the sample covariance matrix of X, $\|\boldsymbol{a}\|^2 = \|\boldsymbol{a}\|_2^2$ is the sum of squares of the components of $\boldsymbol{a}$, the 2-norm or Euclidean norm.

In the standard PCA, all p components of the principal direction vector $\boldsymbol{a}$ can be and usually are non-zero. If the number of variables p is large, it is often desirable to have only a few non-zero components in $\boldsymbol{a}$, for the sake of interpretability. In other words, sparsity is desired. This leads to the development of sparse principal component analysis, or sparse PCA.

4.1 A natural formulation*

First, we discuss a natural formulation to obtain sparse principal components occurred in early development of the methodology of sparse PCA. The method ended up not a practical one due to its high computation demand. However we describe its train of thought below due to its interesting ideas.

Allowing all components in principal direction vector $\boldsymbol{a}$ to be non-zero can be stated as the condition

$$\|\boldsymbol{a}\|_0 \le p,$$

where

$$\|\boldsymbol{a}\|_0 = \text{ the number of non-zero components of } \boldsymbol{a}$$

is the "0-norm" of $\boldsymbol{a}$. Then finding the first principal component can be reformulated as

$$\text{Maximize } \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a} \qquad \text{under the constraints} \quad \|\boldsymbol{a}\| = 1, \quad \|\boldsymbol{a}\|_0 \le p.$$

Similar to the idea in LASSO regression, sparse PCA wishes to have fewer non-zero components in the principal direction vector $\boldsymbol{a}$. Suppose at most k components are allowed to be non-zero, $k \le p$, often $k << p$ is desired.

A natural formulation of finding the first sparse principal component can be stated as

$$\text{maximize } \boldsymbol{v}'\boldsymbol{S}\boldsymbol{v} \qquad \text{under the constraints} \quad \|\boldsymbol{v}\| = 1, \quad \|\boldsymbol{v}\|_0 \le k. \tag{4}$$

Let $\boldsymbol{a}_i$ be the ith principal direction vector for $i = 1, \cdots, p$. Because the covariance matrix $\boldsymbol{S}$ is symmetric positive semi-definite, recall from our earlier derivation,

$$\boldsymbol{S}\boldsymbol{a}_i = \lambda_i \boldsymbol{a}_i, \quad \lambda_1 \ge \cdots \ge \lambda_p \ge 0$$

with

$$\boldsymbol{a}_i\boldsymbol{S}\boldsymbol{a}_i = \lambda_i, \qquad \boldsymbol{a}_i'\boldsymbol{a}_j = \begin{cases} 1, & i = j, \\ 0, & i \ne j. \end{cases}$$

By the Spectral Theorem for symmetric matrix,

$$\boldsymbol{S} = \lambda_1 \boldsymbol{a}_1\boldsymbol{a}_1' + \cdots + \lambda_p \boldsymbol{a}_p\boldsymbol{a}_p'$$

which is a sum of p symmetric matrices. After obtaining $\boldsymbol{a}_1$ corresponding to the leading eigenvalue λ_1 for the first principal component, finding the second principal component direction vector $\boldsymbol{a}_2$ corresponding to eigenvalue λ_2 is equivalent to finding the principal component corresponding to the leading eigenvalue for the matrix

$$\boldsymbol{S} - \lambda_1 \boldsymbol{a}_1\boldsymbol{a}_1' = \boldsymbol{S} - (\boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1)\boldsymbol{a}_1\boldsymbol{a}_1'$$

since $\boldsymbol{a}_1\boldsymbol{S}\boldsymbol{a}_1 = \lambda_1$. and carry out PCA with this matrix to find the second PC, and so on.

Consecutive sparse principal component solutions of (4) are obtained in a similar manner. Assume the optimal solution of (4) is $\boldsymbol{v} = \boldsymbol{v}_1$, a sparse version of $\boldsymbol{a}_1$. Let

$$\boldsymbol{S}_1 = \boldsymbol{S} - (\boldsymbol{v}_1'\boldsymbol{S}\boldsymbol{v}_1)\boldsymbol{v}_1\boldsymbol{v}_1'$$

The second sparse principal component can be found via

$$\text{maximize } \boldsymbol{v}'\boldsymbol{S}_1\boldsymbol{v} \qquad \text{under the constraints} \quad \|\boldsymbol{v}\| = 1, \quad \|\boldsymbol{v}\|_0 \le k.$$

The rest sparse principal components $\boldsymbol{v}_2, \cdots$, can be found by iterating this process.

Unlike the original principal components $\boldsymbol{a}_i$'s, the sparse PC $\boldsymbol{v}_i$'s are not necessarily orthogonal or uncorrelated to each other (without imposing further conditions).

The formulation of (4) seems natural, but turns out to be very computationally demanding.

We need to explore alternative methods.

4.2 PCA in Ridge regression format

A popular alternative for sparse PCA is to place the problem in a linear model setting, then utilize sparsity methods developed in linear regression models such as LASSO and Elastic Net.

This approach is closely related to the Singular Value Decomposition interpretation of PCA.

Suppose the $n \times p$ data matrix $\boldsymbol{X}$ is centered with column sum zero, denoted as $\boldsymbol{X} = \boldsymbol{X}_c$. From the relation between principal components and the Singular Value Decomposition of the centered data matrix $\boldsymbol{X}_c = UDV'$, the $n \times p$ score matrix of the principal components for the centered data $\boldsymbol{X}_c$ is given by

$$\boldsymbol{Y}_c = \boldsymbol{X}_c V = UD$$

as stated in lecture notes Principal Component Analysis. Thus, the ith principal component $\boldsymbol{y}_i$, consisting of n scores (derived from data), can be written as

$$\boldsymbol{y}_i = X\boldsymbol{v}_i, \qquad \|\boldsymbol{v}_i\| = 1, \qquad i = 1, \cdots, p.$$

For a fix i, given the principal component $\boldsymbol{y}_i \in \mathbb{R}^n$, we may treat $\boldsymbol{y}_i$ as a function of X, that is, view $\boldsymbol{y}_i$ as a response variable with X and the input, and write a Ridge regression version of the above relation. For $\lambda > 0$, let

$$\hat{\beta}_{ridge} = \hat{\boldsymbol{\beta}}_{ridge}^{(i)} = \arg\min_{\beta} \left\{ \|\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \tag{5}$$

wehre

$$\boldsymbol{\beta} = [\beta_1 \ \cdots \ \beta_p]', \qquad \|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2.$$

Then we can relate the Ridge regression coefficients with the ith principal direction vector by

$$\frac{\hat{\beta}_{ridge}}{\|\hat{\beta}_{ridge}\|} = \boldsymbol{v}_i \tag{6}$$

Proof. In the following we prove the relation (6).

For fixed i, we need to show that $\hat{\beta}_{ridge}$ is a constant multiple of $\boldsymbol{v}_i$.

By taking derivative w.r.t. β, we can obtain the solution of (5) as

$$\hat{\beta}_{ridge} = (\boldsymbol{X}'\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}'Y$$

with $Y = \boldsymbol{y}_i$.

By the singular value decomposition $\boldsymbol{X} = UDV'$, $X'X = VD^2V'$, $Y = \boldsymbol{y}_i = X\boldsymbol{v}_i$, and $V'v_i = \boldsymbol{e}_i$, which is the p vector with ith component $= 1$ and 0 otherwise. Write the diagonal matrix of singular values $D = [d_{ij}]_{n\times p}$. Then

$$\begin{aligned}
\hat{\beta}_{ridge} &= (VD^2V' + \lambda I)^{-1}X'X\boldsymbol{v}_i \\
&= [V(D^2 + \lambda I)V']^{-1}VD^2V'\boldsymbol{v}_i \\
&= V'^{-1}(D^2 + \lambda I)^{-1}V^{-1}VD^2V'\boldsymbol{v}_i \\
&= V(D^2 + \lambda I)^{-1}D^2V'\boldsymbol{v}_i = V(D^2 + \lambda I)^{-1}D^2\boldsymbol{e}_i \\
&= V\frac{d_{ii}^2}{d_{ii}^2 + \lambda}\boldsymbol{e}_i \;=\; \frac{d_{ii}^2}{d_{ii}^2 + \lambda}\boldsymbol{v}_i
\end{aligned}$$

Therefore $\hat{\boldsymbol{\beta}}^{(i)}_{ridge} = \hat{\beta}_{ridge} \propto \boldsymbol{v}_i$. Since $\|\boldsymbol{v}_i\| = 1$, we have $\frac{\hat{\beta}_{ridge}}{\|\hat{\beta}_{ridge}\|} = \boldsymbol{v}_i$. □

4.3 Sparse PCA in sparse regression formulation

The above illustrate the formulation of PCA in Ridge regression setting.

When the number of variables p is large, it is often desirable to have fewer original variables contributing to each principal component. In other words, it is desirable to have sparse principal loadings.

Based on the Ridge regression representation of principle components stated in (5) and the sparse property of LASSO regression, we may consider adding an 1-norm term in (5) in order to reduce the number of non-zero loadings to achieve sparsity in the ith PC.

$$\hat{\boldsymbol{\beta}}_{sparse} = \hat{\boldsymbol{\beta}}^{(i)}_{sparse} = \arg\min_{\boldsymbol{\beta}} \left\{\|\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\right\}$$

where as before,

$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$$

A larger λ_1 will result in fewer non-zero components of $\hat{\beta}$. Then the direction of the estimated coefficient vector would be an approximation of the principal component direction,

$$\frac{\hat{\boldsymbol{\beta}}_{sparse}}{\|\hat{\boldsymbol{\beta}}_{sparse}\|} \approx \boldsymbol{v}_i \tag{7}$$

where only a few components in $\hat{\boldsymbol{\beta}}_{sparse}$ are non-zero, much like in LASSO regression.

Remarks

- The sparse PC selection process can be carried out for the ith PC, $i = 1, \cdots, p$, simultaneously.
- There is no free lunch: the sparsity is obtained at the cost of capturing less variations and losing uncorrelatedness or orthogonality of the PC variables.
- Further conditions and formulations are needed, say, to save the orthogonality.

References

Section 14.5 in Hastie, Tibshirani and Friedman.
Section 13.4 in Koch.

Article https://tibshirani.su.domains/ftp/lasso-retro.pdf (on Lasso regression) by Tibshirani.
Article http://users.stat.umn.edu/ zouxx019/Papers/elasticnet.pdf (on Elastic Net) by Zou and Hastie.
Article http://web.stanford.edu/ hastie/Papers/sparsepc.pdf (on Sparse PCA) by Zou, Hastie, and Tibshirani.