High dimensional Extensions

(Regularized regression, sparse PCA)

In this section, we take a brief look at methods for high dimensional parameter space cases, and the applications of these methods in dealing with similar high dimension issues in multivariate methods such as Principle Component Analysis.

1 Issues of high dimensional parameter space in linear regressions

Review LS estimator for linear regression models

Recall, to fit the univariate linear regression model

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

based on n observations $(y_1, \cdots, y_n) = \boldsymbol{y}'$, where y_i is an observed response at given values $\{z_{i,k}, k = 1, \cdots, r\}$ of r explanatory variables, the Least Square method seeks solutions of a set of parameters $\boldsymbol{\beta} \in \mathbb{R}_{r+1}$ which minimize the sum of squared errors, that is,

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

Let p be the number of model parameters. In the above formulation, $p = r + 1$. In the case of centered $\boldsymbol{y}$ at its sample mean, then we may consider the intercept-less case of $p = r$.

When $n \geq p$, and when $\boldsymbol{Z}$ is of full rank, 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 estimate is unbiased,

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

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

Proof. Below we show the optimal variance-covariance property achieved by LS estimator.

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

$$\hat{\boldsymbol{\beta}} = A\boldsymbol{y} = \left[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' + B\right]\boldsymbol{y}, \qquad \mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta},$$

where B is a $p \times n$ matrix. Note that the unbiasedness

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

implies

$$B\boldsymbol{Z} = 0_{p\times n}$$

Then

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

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

Furthermore, the trace of the covariance matrix is

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

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

□

Problems with high dimensional parameter space

When the parameter space $\mathbb{R}^p$ is of dimensions higher than the number of observations, sample size $n < p$,

- 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.

  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.

  This situation does occur in practice when there is near collinearity among explanatory variables.

2 Regularization method in linear regression models

2.1 Ridge Regression

Consider the estimator of the type

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

The estimator will be biased. The hope is that the added *term* matrix is of entries (especially diagonal entries) large enough to make the matrix $(\boldsymbol{Z}'\boldsymbol{Z} + \; term)$ invertible, yet small enough to have a reasonable estimate not too biased, possibly with a smaller variance. Consider a solution with a diagonal matrix as the *term* matrix,

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

This leads to the Ridge regression.

Ridge regression 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 $\|\boldsymbol{\beta}\|_2^2 = \sum_i \beta_i^2$. Equivalently, Ridge regression can be written as solving the Lagrangian problem

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

which shows the constraint on the size of the components of $\boldsymbol{\beta}$ explicitly. The two formulations (1) and (2) are equivalent, and there is a one-to-one relationship between the λ in (1) and the s in (2).

By setting

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

we find that Ridge estimator of the coefficient parameters is

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

Ridge regression properties

- The tuning parameter $\lambda \in [0, \infty)$ in Ridge regression controls the penalty 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 shrikage. 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 $\hat{\boldsymbol{\beta}}_{Ridge}$ decreases as λ increases.
- The overall mean squared error $E\|\hat{\boldsymbol{\beta}}_{Ridge} - \boldsymbol{\beta}\|_2^2$ can be reduced (compared with LS estimator) for a range of λ, thus improve prediction accuracy.
- Ridge regression will include all variables by having $\hat{\beta}_i \neq 0$ for all coefficients.

Comments on Ridge regression

- When we encounter collinearity among explanatory variables in model fitting, Ridge Regression often provides better, more stable parameter estimation, reducing the undesirable situation when two correlated explanatory variables ended with large coefficients of opposite signs.

  Write the estimator of Ridge Regression as (note another common notation uses X in the place of Z),

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

  It is designed to deal with the situation when $Z^T Z$ is close to singularity.
- Ridge Regression estimation generally shrinks the estimated parameters from the ordinary least squares estimates.

  For example, if we assume orthogonal design $Z^T Z = I$ (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 generally smaller in magnitude than the original Least Squares estimates.

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

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

  However for suitable λ's, the mean squared error (or risk) of Ridge estimation can achieve smaller variance than the least squared estimation, which leads to possibly smaller Mean Squared Error (MSE)

  $$MSE(\hat{\beta}_i) = \mathbb{E}[(\hat{\beta}_i - \beta_i)^2] = \Big(\text{bias}(\hat{\beta}_i)\Big)^2 + \text{Var}(\hat{\beta}_i)$$
- Note that intercept is not penalized, which is reasonable.
- Scaling of the input variables will affect the model estimates. Often the inputs are normalized before fitting Ridge model, especially when the inputs have large variations in magnitude and spread.

2.2 LASSO Regression

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

In many applications with a large number of explanatory variables, there are often extraneous explanatory variables included in the data, however these variables actually play no role in predicting the response variable y.

This situation can be stated as the following: There is a group of coefficients β_i's with true value $= 0$.

In variable selection problem, it is often desirable to have a small subset of non-zero estimates of β_i for the purpose of model interpretation, especially when the number of all predictors p is large.

It turns out that replacing the 2-norm $\|\boldsymbol{\beta}\|_2$ in Ridge regression 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 regression optimizes

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

where $\|\boldsymbol{\beta}\|_1 = \sum_i |\beta_i|$. Equivalently, LASSO regression can be written as solving

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

Remarks on LASSO

- A characteristic of LASSO estimator $\hat{\boldsymbol{\beta}}_{Lasso}$ is its **sparsity**. That is, the parameter estimates by LASSO would have $\hat{\beta}_i = 0$ 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.
- Unlike Ridge, there is no closed form solution for $\hat{\boldsymbol{\beta}}_{Lasso}$, the coefficient estimator of LASSO.

  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 demo examples in class, we can see that the current algorithms are efficient in finding LASSO solutions for each λ in sufficiently dense set or grid.

2.3 Elastic Net Regression

A combination of Ridge and LASSO regressions, Elastic Net 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 $\lambda \geq 0$. Another common notation for 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 example demo in class, the methods of Ridge, LASSO and Elastic Net are compared using simulation and real data sets. The examples illustrate and compare basic characteristics of each method.

2.4 Vector norm and regularized linear regression models

The sparse property of LASSO come from the properties of the 1-norm penalty term.

A p-norm for a vector $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$ is defined as

$$\|x\|_p = \left(|x_1|^p + \cdots + |x_n|^p \right)^{1/p}, \qquad p \in (0, \infty)$$

- $p = 2$ is the most familiar Euclidean norm. The unit-norm set $\{x : \|x\|_2 = 1\}$ forms a unit circle.
  Ridge regression coefficient estimator shrinks to $\|\boldsymbol{\beta}\|_2^2 \leq s$.
- $p = 1$ is very useful. The unit-norm set $\{x : \|x\|_1 = 1\}$ forms a squared diamond.
- The limiting case $p = \infty$ has the unit-norm set $\{x : \|x\|_\infty = \max_i |x_i|\}$
- The choice of vector norm in the penalty term in Ridge regression and LASSO regression has made a huge difference in the input variable selection problem in each model, as we can see from the demo examples in class. The ensuing applications abound.
- Recent convention use the term "0-norm" (which is not a metric norm, strictly speaking)

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

3 Sparse PCA

In this section, we will introduction a method of Principal Component Analysis in the high dimensional case when the number of variables is large. Currently the algorithm used to implement the method is based on regularized regression methods discussed above.

Given an $n \times p$ data matrix X, 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 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 fewer non-zero components in $\boldsymbol{a}$, for the sake of interpretability. In other words, sparsity is needed. This leads to the development of sparse principal component analysis.

3.1 A natural formulation

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

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

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

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

Similar to the idea in LASSO regression, sparse PCA wishes to have fewer non-zero components in the principal direction vector $\boldsymbol{a}$. Assume that at most k components are allowed to be non-zero, $k \leq p$, often $k << p$ is desired. A natural formulation of finding the first sparse principal component can be stated as

$$\text{maximize } \boldsymbol{v}'\boldsymbol{S}\boldsymbol{v} \qquad \text{under the constraints} \quad \|\boldsymbol{v}\| = 1, \quad \|\boldsymbol{v}\|_0 \leq k. \tag{3}$$

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 that our earlier derivation gives

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

with

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

By the Spectral Theorem for symmetric matrix,

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

witch 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

$$\boldsymbol{S} - \lambda_1 \boldsymbol{a}_1 \boldsymbol{a}_1' = \boldsymbol{S} - (\boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1)\boldsymbol{a}_1\boldsymbol{a}_1', \qquad since \quad \boldsymbol{a}_1\boldsymbol{S}\boldsymbol{a}_1 = \lambda_1,$$

and carry out PCA with this matrix to find the second PC, and so on.

Consecutive sparse principal component solutions of (3) are obtained in a similar manner. Assume the optimal solution of (3) 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

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

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 $\boldsymbol{v}_i$'s are not necessarily orthogonal or uncorrelated to each other without imposing further conditions.

The formulation of (3) seems natural, but turns out to be very computationally demanding.

One of the popular, alternative approaches for sparse PCA is to devise PCA in a regression setting, then utilize the shrinkage method in linear regression discussed above. This approach is closely related to the Singular Value Decomposition interpretation of PCA.

3.2 PCA in regression form

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

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

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 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

$$\hat{\beta}_{ridge} = \hat{\boldsymbol{\beta}}^{(i)}_{ridge} = \arg\min_{\beta} \left\{ \|\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \right\} \tag{4}$$

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

$$\frac{\hat{\beta}_{ridge}}{\|\hat{\beta}_{ridge}\|} = \boldsymbol{v}_i \tag{5}$$

Proof. In the following we prove the relation (5).

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 (4) as

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

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 $\hat{\boldsymbol{\beta}}^{(i)}_{ridge} = \hat{\beta}_{ridge} \propto \boldsymbol{v}_i$. □

3.3 Sparse PCA in regression formulation

In PCA with the number of variable dimension p large, it is often desirable to have fewer original variables contributing to each principal component, that is, it is desirable to have sparse principal loadings.

Based on the Ridge regression representation of principle components (4) and the sparse property of LASSO regression, we may consider adding an L_1 norm term in (4) in order to reduce the number of non-zero loadings.

$$\hat{\beta}_{sparse} = \hat{\boldsymbol{\beta}}^{(i)}_{sparse} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2 \right\}, \qquad \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|.$$

A larger λ_1 will give fewer non-zero component of $\hat{\beta}$. Then the estimated coefficients would be an approximation of the principal component direction,

$$\frac{\hat{\beta}_{sparse}}{\|\hat{\beta}_{sparse}\|} \approx \boldsymbol{v}_i \tag{6}$$

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

The sparsity is obtained, at the expense of capturing less variations and losing uncorrelated-ness or orthogonality of the PC variables. Further conditions and formulations are needed, say, to save the orthogonality.

References
Sections 14.5 (on PCA and generalizations), 14.7 (on ICA) in Hastie, Tibshirani and Friedman.
Sections 10.1-10.4 (on ICA), 13.4 (on Sparse PCA) in Koch.
Chapter 9 (on Mixture models and EM) in *Pattern Recognition and Machine Learning* by Bishop.
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.