Notes on Statistical and Machine Learning

Principal Component Analysis

Chapter: *23* **Prepared by:** *Chenxi Zhou*

This note is prepared based on

- *Chapter 7, Linear Dimensionality Reduction* in Izenman (2009),
- *Chapter 16, Nonlinear Dimensionality Reduction and Manifold Learning* in Izenman (2009),
- *Chapter 14, Unsupervised Learning* in Hastie, Tibshirani, and Friedman (2009), and
- *Chapter 8, Sparse Multivariate Methods* in Hastie, Tibshirani, and Wainwright (2015).

I. Introduction

1. **Motivation:** Faced with high-dimensional data, there are two ways of projection the data onto a lower-dimensional subspace without losing important information regarding some characteristics of the original variables:
   (a) *feature selection*, also known as variable selection;
   (b) *feature extraction*, creating a reduced set of linear or nonlinear transformations of the input variables.
2. **Introduction:** Principal component analysis (PCA), initially proposed by Hotelling in 1933, is to derive a reduced set of *orthogonal* linear projections of a single collection of correlated variables, $X = (X_1, \cdots, X_p)^\top$, where the projections are ordered by decreasing variances.
3. **Main Usages of PCA:** It has two main usages:
   (a) to reduce dimensionality: for example, the reduced set of linear transformation of input variables can be used in principal components regression;
   (b) to discover important features of the data:
      - the *first* few principal component scores can be used to identify outliers, distribution peculiarities, and clusters of points;
      - the *last* few principal component scores show linear projections of X that have smallest variance: the component with zero or near-zero variance is virtually constant and can be used to detect collinearity and outliers and alter the perceived dimensionality of data.

II. Population Principal Components

1. **Setup:** Assume that a p-dimensional random vector

$$X = (X_1, \cdots, X_p)^\top \in \mathbb{R}^p,$$

has the mean $\mathbb{E}[X] = \boldsymbol{\mu}_X \in \mathbb{R}^p$ and the covariance matrix $\text{Var}[X] = \boldsymbol{\Sigma}_{XX} \in \mathbb{R}^{p\times p}$. Principal component analysis is to replace the set of p (unordered and correlated) input variables, $X_1, \cdots, X_p$, by a set of t (ordered and uncorrelated) linear projections, $\xi_1, \cdots, \xi_t$, where $t \leq p$, of the input variables,

$$\xi_j = \mathbf{b}_j^\top X = b_{j,1}X_1 + \cdots + b_{j,p}X_p, \qquad \text{for all } j = 1, \cdots, t. \tag{1}$$

We attempt to minimize the loss of information by such a replacement.

2. **Total Variation:** In PCA, the "information" is interpreted as the *total variation* of the original input variables,

$$\sum_{j=1}^{p} \text{Var}[X_j] = \text{trace}(\boldsymbol{\Sigma}_{XX}).$$

Since $\boldsymbol{\Sigma}_{XX}$ is positive semi-definite, according to the spectral decomposition theorem, we have

$$\boldsymbol{\Sigma}_{XX} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top,$$

where $\boldsymbol{\Lambda} := \text{diag}(\lambda_1, \cdots, \lambda_p) \in \mathbb{R}^{p\times p}$ contains all the eigenvalues of $\boldsymbol{\Sigma}_{XX}$ on its diagonal and $\mathbf{U} \in \mathbb{R}^{p\times p}$ is an orthogonal matrix with eigenvectors of $\boldsymbol{\Sigma}_{XX}$ as its columns. Hence, the total variation is

$$\text{trace}(\boldsymbol{\Sigma}_{XX}) = \text{trace}(\boldsymbol{\Lambda}) = \sum_{j=1}^{p} \lambda_j.$$

3. **Construction of the Principal Components:** The j-th coefficient vector, $\mathbf{b}_j = (b_{j,1}, \cdots, b_{j,p})^\top \in \mathbb{R}^p$, is chosen so that:

   - The first t linear projections ξ_j, $j = 1, 2, \cdots, t$, of X are ranked in terms of their variances $\text{Var}[\xi_j]$, which are listed in the *decreasing* order of magnitude

$$\text{Var}[\xi_1] \geq \text{Var}[\xi_2] \geq \cdots \geq \text{Var}[\xi_t] \geq 0.$$

   - ξ_j is uncorrelated with all ξ_k, for all $k < j$.

III. Derivation of PCA Using the Least Squares Method

1. **Least-Squares Optimality of PCA:** Let

$$\mathbf{B} := (\mathbf{b}_1, \cdots, \mathbf{b}_t)^\top \in \mathbb{R}^{t \times p}$$

be a matrix of weights, where $t \leq p$. Then, the linear projections (1) can be written as a t-vector

$$\boldsymbol{\xi} = \mathbf{B}X,$$

where $\boldsymbol{\xi} := (\xi_1, \cdots, \xi_t)^\top \in \mathbb{R}^t$. We want to find a p-vector $\boldsymbol{\mu}$ and an $p \times t$ matrix $\mathbf{A}$ such that the projections $\boldsymbol{\xi}$ have the property

$$X \approx \boldsymbol{\mu} + \mathbf{A}\boldsymbol{\xi}$$

in the least squares sense:

$$\mathbb{E}\Big[\big(X - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\xi}\big)^\top \big(X - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\xi}\big)\Big]. \tag{2}$$

Since $\boldsymbol{\xi} = \mathbf{B}X$, the problem now becomes to choose μ, $\mathbf{A}$ and $\mathbf{B}$ to minimize

$$f(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu}) := \mathbb{E}\Big[\big(X - \boldsymbol{\mu} - \mathbf{AB}X\big)^\top \big(X - \boldsymbol{\mu} - \mathbf{AB}X\big)\Big].$$

First note that f is convex in all $\mathbf{A}$, $\mathbf{B}$ and $\boldsymbol{\mu}$. Thus, any stationary point is a global minimizer. Differentiating f with respect to $\boldsymbol{\mu}$ and setting the result to zero, we have

$$\widehat{\boldsymbol{\mu}} := \arg\min_{\boldsymbol{\mu}} f(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu}) = (\mathbf{I} - \mathbf{AB})\boldsymbol{\mu}_X.$$

Plugging $\widehat{\boldsymbol{\mu}}$ back to f, we obtain

$$\begin{aligned} f(\mathbf{A}, \mathbf{B}, \widehat{\boldsymbol{\mu}}) &= \mathbb{E}\Big[(X - \boldsymbol{\mu}_X)^\top (\mathbf{I} - \mathbf{AB})^\top (\mathbf{I} - \mathbf{AB})(X - \boldsymbol{\mu}_X)\Big] \\ &= \mathbb{E}\Big[\operatorname{trace}\big((\mathbf{I} - \mathbf{AB})^\top (\mathbf{I} - \mathbf{AB})(X - \boldsymbol{\mu}_X)(X - \boldsymbol{\mu}_X)^\top\big)\Big] \\ &= \operatorname{trace}\big((\mathbf{I} - \mathbf{AB})\boldsymbol{\Sigma}_{XX}(\mathbf{I} - \mathbf{AB})^\top\big). \end{aligned}$$

Let $\mathbf{C} := \mathbf{AB}$ and $\boldsymbol{\Sigma}_{XX}^{\frac{1}{2}} = \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}^\top$ so that $\boldsymbol{\Sigma}_{XX}^{\frac{1}{2}}\boldsymbol{\Sigma}_{XX}^{\frac{1}{2}\top} = \boldsymbol{\Sigma}_{XX}$, where $\boldsymbol{\Sigma}_{XX} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ is the spectral decomposition of $\boldsymbol{\Sigma}_{XX}$. We can then rewrite f as

$$\tilde{f}(\mathbf{C}) = \operatorname{trace}\big(((\mathbf{I} - \mathbf{C})\boldsymbol{\Sigma}_{XX}^{\frac{1}{2}})((\mathbf{I} - \mathbf{C})\boldsymbol{\Sigma}_{XX}^{\frac{1}{2}})^\top\big).$$

By the Eckart-Young's inequality, $\tilde{f}$ is minimized when

$$\mathbf{C}\boldsymbol{\Sigma}_{XX}^{\frac{1}{2}} = \sum_{j=1}^{t} \lambda_j^{\frac{1}{2}} \mathbf{u}_j \mathbf{u}_j^\top = \mathbf{U}^{(t)}\boldsymbol{\Lambda}^{(t)\,\frac{1}{2}}\mathbf{U}^{(t)\top},$$

where $\mathbf{u}_j$ is the eigenvector of $\mathbf{\Sigma}_{XX}$ associated with the j-th largest eigenvalue of $\mathbf{\Sigma}_{XX}$, $\mathbf{U}^{(t)} = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_t) \in \mathbb{R}^{p\times t}$, and $\mathbf{\Lambda}^{(t)} \in \mathbb{R}^{t\times t}$ is the diagonal matrix with the largest t eigenvalues on its diagonal. It follows that

$$\widehat{\mathbf{C}} := \arg\min_{\mathbf{C}} \tilde{f}(\mathbf{C}) = \mathbf{U}^{(t)}\mathbf{U}^{(t)\top}.$$

We can then let

$$\widehat{\mathbf{A}} = \mathbf{U}^{(t)}, \qquad \text{and} \qquad \widehat{\mathbf{B}} = \mathbf{U}^{(t)\top},$$

where $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) := \arg\min_{\mathbf{A},\mathbf{B}} f(\mathbf{A}, \mathbf{B})$. In addition, we have

$$\widehat{\boldsymbol{\mu}} = (\mathbf{I} - \widehat{\mathbf{A}}\widehat{\mathbf{B}})\boldsymbol{\mu}_X.$$

Thus, the best rank-t approximation to the original X is given by

$$X^{(t)} = \widehat{\boldsymbol{\mu}} + \widehat{\mathbf{C}}X = \boldsymbol{\mu}_X + \widehat{\mathbf{C}}(X - \boldsymbol{\mu}_X).$$

It also follows that the minimum of f at $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mu})$ is

$$f(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\mu}}) = \sum_{j=t+1}^{p} \lambda_j,$$

the sum of the smallest $p - t$ eigenvalues of $\mathbf{\Sigma}_{XX}$.

2. **Principal Components:** The first t principal components are given by the linear projections $\xi_1, \cdots, \xi_t$, where

$$\xi_j = \mathbf{u}_j^\top X, \qquad \text{for all } j = 1, \cdots, t,$$

and $\mathbf{u}_j$ is the eigenvector of $\mathbf{\Sigma}_{XX}$ associated with the j-th largest eigenvalue.

3. **Covariance between ξ_i and ξ_j:** The covariance between ξ_i and ξ_j is

$$\text{Cov}(\xi_i, \xi_j) = \text{Cov}(\mathbf{u}_i^\top X, \mathbf{u}_j^\top X) = \mathbf{u}_i^\top \mathbf{\Sigma}_{XX}\mathbf{u}_j = \delta_{ij}\lambda_j,$$

where δ_{ij} is the Kronecker delta. In particular,

   (a) $\text{Var}[\xi_1]$ is the largest eigenvalue of $\mathbf{\Sigma}_{XX}$, λ_1, and
   (b) all pairs of derived variables are uncorrelated, $\text{Cov}(\xi_i, \xi_j) = 0$ for $i \neq j$.

4. **Goodness-of-fit Measurement for the First t Principal Components:** A *goodness-of-fit* measure of how well the first t principal components represent the p original variables is given by the ratio

$$R := \frac{\lambda_{t+1} + \cdots + \lambda_p}{\lambda_1 + \cdots + \lambda_p}, \tag{3}$$

the proportion of the total variation in the input variables that is explained by the last $p - t$ principal components. The larger proportion of the total variation in X that the first t principal components explains, the smaller value R is.

IV. Derivation of PCA Using a Variance-Maximization Technique

1. **Problem Formulation:** Choose the coefficient vectors

$$\mathbf{b}_j = (b_{j,1}, b_{j,2}, \cdots, b_{j,p})^\top \in \mathbb{R}^p, \qquad \text{for all } j = 1, \cdots, t,$$

in a sequential manner so that

   (a) the variances of the derived variables, $\text{Var}[\xi_j] = \mathbf{b}_j^\top \mathbf{\Sigma}_{XX} \mathbf{b}_j$, are arranged in the *descending* order subject to the normalizations $\mathbf{b}_j^\top \mathbf{b}_j = 1$ for all $j = 1, 2, \cdots, t$, and

   (b) they are uncorrelated with previously chosen derived variables, that is,

$$\text{Cov}(\xi_i, \xi_j) = \mathbf{b}_i^\top \mathbf{\Sigma}_{XX} \mathbf{b}_j = 0, \qquad \text{for all } i < j.$$

2. **Derivation of the First Principal Component:** The first principal component, ξ_1, is obtained by choosing the p coefficients, $\mathbf{b}_1$, so that the variance of ξ_1 is maximized, under the constraint that $\mathbf{b}_1^\top \mathbf{b}_1 = 1$. This amounts to solving the following constrained optimization problem

$$\begin{aligned} &\underset{\mathbf{b}}{\text{maximize}}\, \mathbf{b}_1^\top \mathbf{\Sigma}_{XX} \mathbf{b}_1 \\ &\text{subject to } \mathbf{b}_1^\top \mathbf{b}_1 = 1. \end{aligned}$$

   The Lagrangian function is

$$f_1(\mathbf{b}_1) := \mathbf{b}_1^\top \mathbf{\Sigma}_{XX} \mathbf{b}_1 - \lambda_1 (\mathbf{b}_1^\top \mathbf{b}_1 - 1), \tag{4}$$

   where $\lambda_1 \geq 0$ is the Lagrangian multiplier.

   Differentiating f_1 with respect to $\mathbf{b}_1$ and setting the result equal to zero yield

$$\frac{\partial f_1(\mathbf{b}_1)}{\partial \mathbf{b}_1} = 2(\mathbf{\Sigma}_{XX} - \lambda \mathbf{I}_p)\mathbf{b}_1 = \mathbf{0}_p, \tag{5}$$

   which is a set of p simultaneous equations. Since $\mathbf{b}_1 \neq 0$, due to the constraint $\mathbf{b}_1^\top \mathbf{b}_1 = 1$, then λ_1 must be chosen to satisfy the following equation

$$|\mathbf{\Sigma}_{XX} - \lambda_1 \mathbf{I}_p| = 0.$$

   Thus, λ_1 has to be the largest eigenvalue of $\mathbf{\Sigma}_{XX}$, and $\mathbf{b}_1$ the eigenvector, $\mathbf{u}_1$, associated with λ_1.

3. **Derivation of the Second Principal Component:** The second principal component, ξ_2, is obtained by choosing a second set of coefficients, $\mathbf{b}_2$, so that

   (a) the variance of ξ_2 is the largest among all linear projections of X,

   (b) $\mathbf{b}_2^\top \mathbf{b}_2 = 1$, and

(c) ξ_2 is uncorrelated with ξ_1.

The resulting maximization problem is

$$\begin{aligned} \underset{\mathbf{b}_2}{\text{maximize}}\, & \mathbf{b}_2^\top \boldsymbol{\Sigma}_{XX} \mathbf{b}_2 \\ \text{subject to } & \mathbf{b}_2^\top \mathbf{b}_2 = 1, \\ & \mathbf{b}_2^\top \boldsymbol{\Sigma}_{XX} \mathbf{b}_1 = 0. \end{aligned}$$

The Lagrangian function is

$$f_2(\mathbf{b}_2) := \mathbf{b}_2^\top \boldsymbol{\Sigma}_{XX} \mathbf{b}_2 - \lambda_2(\mathbf{b}_2^\top \mathbf{b}_2 - 1) - \nu(\mathbf{b}_2^\top \boldsymbol{\Sigma}_{XX} \mathbf{b}_1),$$

where $\lambda_2 \geq 0$ and $\nu \geq 0$ are the Lagrangian multipliers.

Differentiating f_2 with respect to $\mathbf{b}_2$ and setting the result equal to zero yields

$$\frac{\partial f_2(\mathbf{b}_2)}{\partial \mathbf{b}_2} = 2(\boldsymbol{\Sigma}_{XX} - \lambda_2 \mathbf{I}_p)\mathbf{b}_2 - \nu \boldsymbol{\Sigma}_{XX} \mathbf{b}_1 = \mathbf{0}_p. \tag{6}$$

Premultiplying (6) by $\mathbf{b}_1^\top$ yields

$$2\mathbf{b}_1^\top \boldsymbol{\Sigma}_{XX} \mathbf{b}_2 - 2\lambda_2 \mathbf{b}_1^\top \mathbf{b}_2 - \nu \mathbf{b}_1^\top \boldsymbol{\Sigma}_{XX} \mathbf{b}_1 = 0,$$

which is equivalent to

$$-2\lambda_2 \mathbf{b}_1^\top \mathbf{b}_2 - \nu \mathbf{b}_1^\top \boldsymbol{\Sigma}_{XX} \mathbf{b}_1 = 0. \tag{7}$$

On the other hand, premultiplying (6) by $\mathbf{b}_2^\top$ yields

$$2\mathbf{b}_2^\top \boldsymbol{\Sigma}_{XX} \mathbf{b}_1 - 2\lambda_1 \mathbf{b}_2^\top \mathbf{b}_1 = 0,$$

which is equivalent to

$$-2\lambda_1 \mathbf{b}_2^\top \mathbf{b}_1 = 0.$$

From this preceding equation, we obtain $\mathbf{b}_2^\top \mathbf{b}_1 = 0$, since $\lambda_1 \neq 0$. Thus, plugging this result into (7), we have

$$0 = -\nu \mathbf{b}_1^\top \boldsymbol{\Sigma}_{XX} \mathbf{b}_1 = -\nu \lambda_1,$$

implying $\nu = 0$. Thus, λ_2 and $\mathbf{b}_2$ have to satisfy

$$(\boldsymbol{\Sigma}_{XX} - \lambda_2 \mathbf{I}_p)\mathbf{b}_2 = \mathbf{0}_p.$$

This means that λ_2 is the second largest eigenvalue of $\boldsymbol{\Sigma}_{XX}$, and the coefficient vector $\mathbf{b}_2$ for the second principal component is the eigenvector, $\mathbf{u}_2$, associated with λ_2.

4. **Further Principal Components:** In this sequential manner, we obtain the remaining sets of coefficients for the principal components $\xi_3, \xi_4, \cdots, \xi_p$, where ξ_j is obtained by choosing the set of coefficients, $\mathbf{b}_j$, for ξ_j so that

   (a) ξ_j has the largest variance among all linear projections of X, and

   (b) ξ_j is uncorrelated with $\xi_1, \xi_2, \cdots, \xi_{j-1}$.

   The coefficients of these linear projections are given by the ordered sequence of eigenvectors $\{\mathbf{u}_j\}$, where $\mathbf{u}_j$ is associated with the j-th largest eigenvalue, λ_j, of $\boldsymbol{\Sigma}_{XX}$.

V. Sample Principal Components

1. **Setup:** Suppose we have n i.i.d observations from the random vector X, denoted by $\mathbf{x}_1, \cdots, \mathbf{x}_n$, where $\mathbf{x}_i \in \mathbb{R}^p$ for all $i = 1, \cdots, n$.

2. **Estimation of μ_X:** We estimate $\boldsymbol{\mu}_X$ by the sample mean

$$\bar{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i.$$

3. **Estimation of $\boldsymbol{\Sigma}_X$:** We estimate $\boldsymbol{\Sigma}_X$ by the sample covariance matrix

$$\widehat{\boldsymbol{\Sigma}}_{XX} := \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$$

   The ordered eigenvalues of $\widehat{\boldsymbol{\Sigma}}_{XX}$ are denoted by $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$, and the eigenvector associated with the j-th largest sample eigenvalue $\hat{\lambda}_j$ is the j-th sample eigenvector $\hat{\mathbf{u}}_j$, for all $j = 1, 2, \cdots, p$.

4. **Estimation of $\mathbf{A}^{(t)}$, $\mathbf{B}^{(t)}$ and $\mathbf{C}^{(t)}$:** We estimate $\mathbf{A}^{(t)}$ and $\mathbf{B}^{(t)}$ by

$$\widehat{\mathbf{A}}^{(t)} = (\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \cdots, \hat{\mathbf{u}}_t) = \widehat{\mathbf{B}}^{(t)\top}.$$

   The best rank-t reconstruction of $X = \mathbf{x}$ is given by

$$\hat{\mathbf{x}}^{(t)} = \bar{\mathbf{x}} + \widehat{\mathbf{C}}^{(t)}(\mathbf{x} - \bar{\mathbf{x}}),$$

   where

$$\widehat{\mathbf{C}}^{(t)} = \widehat{\mathbf{A}}^{(t)}\widehat{\mathbf{B}}^{(t)} = \sum_{j=1}^{t} \hat{\mathbf{u}}_j \hat{\mathbf{u}}_j^\top.$$

5. **Sample Principal Component:** The j-th sample principal component of $X = \mathbf{x}$ is given by

$$\hat{\xi}_j = \hat{\mathbf{u}}_j^\top (\mathbf{x} - \bar{\mathbf{x}}).$$

6. **Estimation of R in (3):** The variance, λ_j, of the j-th principal component is estimated by the sample variance $\hat{\lambda}_j$, for all $j = 1, 2, \cdots, t$. A sample estimate of the measure (3) of how well the first t principal components represent the p original variables is given by the statistic

$$\frac{\hat{\lambda}_{t+1} + \hat{\lambda}_{t+2} + \cdots + \hat{\lambda}_p}{\hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p},$$

   which is the proportion of the total sample variation that is explained by the last $p - t$ sample principal components.

7. **Distribution of the Eigenvalues of $\mathbf{X}\mathbf{X}^\top$ When $n > p$:** Let $n > p$ and

$$\mathbf{X} := \big[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\big] \in \mathbb{R}^{p\times n}.$$

where $\mathbf{x}_i \overset{\text{i.i.d}}{\sim} \text{Normal}_p(\mathbf{0}_p, \mathbf{I}_p)$. Then, $\mathbf{X}\mathbf{X}^\top \sim \text{Wishart}_p(n, \mathbf{I}_p)$.

The density function of the eigenvalues of $\mathbf{X}\mathbf{X}^\top$ is

$$f(\lambda_1, \cdots \lambda_p) = c_{p,n} \prod_{j=1}^{p} \sqrt{w(\lambda_j)} \prod_{j<k} (\lambda_j - \lambda_k),$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are the ordered eigenvalues of $\mathbf{X}\mathbf{X}^\top$, $w(x) := x^{n-p-1}e^{-x}$ is the weight function for the Laguerre family of orthogonal polynomials, and $c_{p,n}$ is a normalizing constant dependent upon p and n.

8. **Distribution of the Eigenvalues of $\mathbf{X}\mathbf{X}^\top$ When $p > n$:** Let $p > n$ and

$$\mathbf{X} = \big[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\big] \in \mathbb{R}^{p\times n},$$

where $\mathbf{x}_i \overset{\text{i.i.d}}{\sim} \text{Normal}_p(\mathbf{0}_p, \mathbf{I}_p)$. Then, $\mathbf{X}\mathbf{X}^\top \sim \text{Wishart}_p(n, \mathbf{I}_p)$.

The empirical distribution function computes the proportion of sample eigenvalues that are less than a given value of k,

$$G_p(k) = \frac{1}{p}\big|\{j \mid \hat{\lambda}_j \leq k\}\big|.$$

It can be shown that if $\frac{p}{n} \to \gamma \in (0, \infty)$, then $G_p(k) \to G(k)$ almost surely, where the limiting distribution G has density function

$$g(k) = \frac{\sqrt{(b_+ - k)(k - b_-)}}{2\pi\gamma k}, \qquad \text{where } b_\pm = (1 \pm \sqrt{\gamma})^2.$$

VI. How Many Principal Components to Retain?

1. **General Principle:** We should *only* retain those principal components with large variances.

2. **Scree Plot:** The sample eigenvalues from a PCA are ordered from largest to smallest. The *scree plot* plots the ordered sample eigenvalues against their order number.

   If the largest few sample eigenvalues dominate in magnitude, with the remaining sample eigenvalues very small, then the scree plot will exhibit an "elbow" in the plot corresponding to the division into "large" and "small" values of the sample eigenvalues.

   It is usually recommended to retain those principal components up to the elbow and also the first PC following the elbow.

3. **PC Rank Trace:** The problem of deciding how many principal components to retain is equivalent to obtaining an estimate of the rank of the regression coefficient matrix $\mathbf{C}$ in the principal components case.

   The rank trace plots the loss of information when approximating the full-rank regression by a sequence of reduced-rank regressions having increasing ranks. When the true rank of the regression, denoted by t_0, is reached, the points in the rank trace plot following that rank should cease to change significantly from both the point for t_0 and the full-rank point (rank p).

   In the principal components case, we plot

$$\Delta\widehat{\mathbf{C}}^{(t)} = \left(1 - \frac{t}{p}\right)^{\frac{1}{2}}, \qquad \text{and} \qquad \Delta\widehat{\mathbf{\Sigma}}_{XX}^{(t)} = \left(\frac{\hat{\lambda}_{t+1}^2 + \hat{\lambda}_{t+2}^2 + \cdots + \hat{\lambda}_p^2}{\hat{\lambda}_1^2 + \hat{\lambda}_2^2 + \cdots + \hat{\lambda}_p^2}\right)^{\frac{1}{2}},$$

   for all $t = 1, \cdots, p$. A plot of $\Delta\widehat{\mathbf{\Sigma}}_{XX}^{(t)}$ against $\Delta\widehat{\mathbf{C}}^{(t)}$ for different values of t is called a *PC rank trace plot*.

   We assess the rank t of $\mathbf{C}$ by $\hat{t}$, the smallest integer value between 1 and p at which an "elbow" can be detected in the PC rank trace plot.

VII. Invariance and Scaling

1. **Disadvantage of PCA:** A shortcoming of PCA is that the principal components are *not* invariant under re-scalings of the initial variables. In other words, PCA is sensitive to the units of measurement of the different input variables.

2. **Standardizing Variables:** Standardizing (centering and then scaling) the X-variables,

$$Z \longleftarrow (\text{diag}(\mathbf{\Sigma}_{XX}))^{-\frac{1}{2}}(X - \boldsymbol{\mu}_X),$$

   is equivalent to carrying out PCA using the correlation (rather than the covariance) matrix.

   When using the correlation matrix, the total variation of the standardized variables is p, the trace of the correlation matrix.

   Remark. The lack of scale invariance implies that a PCA using the correlation matrix may be *very different* from a similar analysis using the corresponding covariance matrix, and no simple relationship exists between the two sets of results.

VIII. Kernel PCA

1. **General Idea:** Standard linear PCA we have seen so far are obtained from the eigenvectors of the covariance matrix, and give directions in which the data have maximal variances.

   Kernel PCA expands the scope of the standard linear PCA by expanding the features using non-linear transformations.

2. **Setup:** Let $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$ be i.i.d observations from a p-dimensional random vector.

3. **General Procedures:** Kernel PCA has the following two-step process:

   (a) Nonlinearly transform $\mathbf{x}_i \in \mathbb{R}^p$ into a point $\Phi(\mathbf{x}_i)$ in a high-dimensional feature space $\mathcal{H}$ over which an inner product is defined, for all $i = 1, 2, \cdots, n$. The map $\Phi : \mathbb{R}^p \to \mathcal{H}$ is called a *feature map*;

   (b) Given $\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \cdots, \Phi(\mathbf{x}_n) \in \mathcal{H}$ satisfying $\sum_{i=1}^n \Phi(\mathbf{x}_i) = \mathbf{0}$, solve a *standard linear PCA problem* in the feature space $\mathcal{H}$.

   Remark. We do *not* need to define the feature map Φ explicitly. Instead, as we will see, we only work on the inner product $\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_{\mathcal{H}} =: K(\mathbf{x}, \mathbf{y})$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product in the feature space $\mathcal{H}$.

4. **PCA in Feature Space:** In order to carry out linear PCA in feature space so that it mimics the standard treatment of PCA, we have to find eigenvalues $\lambda \geq 0$ and nonzero eigenvectors $\mathbf{u} \in \mathcal{H}$ of the estimated covariance matrix,

$$\mathbf{C} = \frac{1}{n} \sum_{i=1}^n \Phi(\mathbf{x}_i) \Phi(\mathbf{x}_i)^\top,$$

   under the constraint $\sum_{i=1}^n \Phi(\mathbf{x}_i) = \mathbf{0}$.

   (a) *Characterizing the Solution:* We consider the eigen-problem

$$\mathbf{C}\mathbf{u} = \lambda \mathbf{u}, \tag{8}$$

   where $\mathbf{u}$ is the eigenvector corresponding to the eigenvalue $\lambda \geq 0$ of $\mathbf{C}$. Equivalently, we can rewrite (8) as

$$\langle \Phi(\mathbf{x}_i), \mathbf{C}\mathbf{u} \rangle_{\mathcal{H}} = \lambda \langle \Phi(\mathbf{x}_i), \mathbf{u} \rangle_{\mathcal{H}}, \qquad \text{for all } i = 1, 2, \cdots, n. \tag{9}$$

   Since

$$\mathbf{C}\mathbf{u} = \frac{1}{n} \sum_{i=1}^n \langle \Phi(\mathbf{x}_i), \mathbf{u} \rangle_{\mathcal{H}} \Phi(\mathbf{x}_i) = \lambda \mathbf{u},$$

   all solutions $\mathbf{u}$ with nonzero eigenvalue λ must reside in

$$\text{Span}\Big(\big\{ \Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \cdots, \Phi(\mathbf{x}_n) \big\} \Big).$$

   Hence, there exist coefficients $\alpha_1, \alpha_2, \cdots, \alpha_n$ such that

$$\mathbf{u} = \sum_{j=1}^n \alpha_j \Phi(\mathbf{x}_j). \tag{10}$$

With $\mathbf{u}$ in the form of (10), we have

$$\begin{aligned}\mathbf{Cu} &= \frac{1}{n}\sum_{\ell=1}^{n}\left\langle \Phi(\mathbf{x}_\ell), \sum_{j=1}^{n}\alpha_j\Phi(\mathbf{x}_j)\right\rangle_{\mathcal{H}}\Phi(\mathbf{x}_\ell)\\ &= \frac{1}{n}\sum_{\ell=1}^{n}\sum_{j=1}^{n}\alpha_j\langle\Phi(\mathbf{x}_\ell),\Phi(\mathbf{x}_j)\rangle_{\mathcal{H}}\Phi(\mathbf{x}_\ell),\end{aligned}$$

and, for all $i = 1, 2, \cdots, n$,

$$\langle\Phi(\mathbf{x}_i), \mathbf{Cu}\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{\ell=1}^{n}\sum_{j=1}^{n}\alpha_j\langle\Phi(\mathbf{x}_\ell),\Phi(\mathbf{x}_j)\rangle_{\mathcal{H}}\langle\Phi(\mathbf{x}_\ell),\Phi(\mathbf{x}_i)\rangle_{\mathcal{H}}.$$

Hence, (9) can be written as

$$\frac{1}{n}\sum_{\ell=1}^{n}\sum_{j=1}^{n}\alpha_j\langle\Phi(\mathbf{x}_\ell),\Phi(\mathbf{x}_j)\rangle_{\mathcal{H}}\langle\Phi(\mathbf{x}_\ell),\Phi(\mathbf{x}_i)\rangle_{\mathcal{H}} = \sum_{j=1}^{n}\lambda\alpha_j\langle\Phi(\mathbf{x}_j),\Phi(\mathbf{x}_i)\rangle_{\mathcal{H}}. \tag{11}$$

Let $\mathbf{K}$ be an $n\times n$ real matrix with the (i,j)-th entry being

$$\langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x}_j)\rangle_{\mathcal{H}},$$

and $\boldsymbol{\alpha} := (\alpha_1, \alpha_2, \cdots, \alpha_n)^\top \in \mathbb{R}^n$. Then, we can write (11) as

$$\mathbf{K}^2\boldsymbol{\alpha} = n\lambda\mathbf{K}\boldsymbol{\alpha}.$$

Assuming $\mathbf{K}$ is invertible and letting $\tilde{\lambda} := n\lambda$, we obtain

$$\mathbf{K}\boldsymbol{\alpha} = \tilde{\lambda}\boldsymbol{\alpha}. \tag{12}$$

In other words, $\tilde{\lambda}$ is the eigenvalue of $\mathbf{K}$ and $\boldsymbol{\alpha}$ is the corresponding eigenvector.

(b) *Solution:* Let $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_n \geq 0$ be the eigenvalues of the matrix $\mathbf{K}$ in the decreasing order, and $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,n})^\top \in \mathbb{R}^n$ be the eigenvector associated with $\tilde{\lambda}_i$, for all $i = 1, 2, \cdots, n$.
We have the solution to (8), $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n$, given by

$$\mathbf{u}_i = \sum_{j=1}^{n}\alpha_{i,j}\Phi(\mathbf{x}_j), \qquad \text{for all } i = 1, 2, \cdots, n.$$

If we require $\|\mathbf{u}_i\|_{\mathcal{H}} = 1$, we note

$$\begin{aligned}\|\mathbf{u}_i\|_{\mathcal{H}}^2 &= \sum_{j=1}^{n}\sum_{\ell=1}^{n}\alpha_{i,j}\alpha_{i,\ell}\langle\Phi(\mathbf{x}_j),\Phi(\mathbf{x}_\ell)\rangle_{\mathcal{H}}\\ &= \boldsymbol{\alpha}_i^\top\mathbf{K}\boldsymbol{\alpha}_i\\ &= \tilde{\lambda}_i\boldsymbol{\alpha}_i^\top\boldsymbol{\alpha}_i\\ &= \tilde{\lambda}_i,\end{aligned}$$

and the desired $\mathbf{u}_i$ is

$$\tilde{\mathbf{u}}_i := \frac{1}{\tilde{\lambda}_i^{1/2}} \sum_{j=1}^{n} \alpha_{i,j} \Phi(\mathbf{x}_j).$$

5. **Kernel Principal Component Score of a New Point x:** Let $\mathbf{x} \in \mathbb{R}^p$ be a point possibly different from $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$. In order to obtain the kernel principal component score of $\mathbf{x}$ corresponding to Φ, we project $\Phi(\mathbf{x}) \in \mathcal{H}$ onto the eigenvectors $\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2, \cdots, \tilde{\mathbf{u}}_n$ with the following coefficients

$$\langle \tilde{\mathbf{u}}_i, \Phi(\mathbf{x}) \rangle_{\mathcal{H}} = \frac{1}{\tilde{\lambda}_i^{1/2}} \sum_{j=1}^{n} \alpha_{i,j} \langle \Phi(\mathbf{x}_j), \Phi(\mathbf{x}) \rangle_{\mathcal{H}} = \frac{1}{\tilde{\lambda}_i^{1/2}} \sum_{j=1}^{n} \alpha_{i,j} K(\mathbf{x}_i, \mathbf{x}),$$

for all $i = 1, 2, \cdots, n$.

6. **On the Centering of the Feature Map:** All developments above assumes

$$\sum_{i=1}^{n} \Phi(\mathbf{x}_i) = \mathbf{0}.$$

This essentially assumes each column and each row of $\mathbf{K}$ sum to 0.

In practice, when we have chosen a kernel function and constructed $\mathbf{K}$ using real data, each row and each column do not necessarily sum to 0. We apply the following adjustment to the un-centered $\mathbf{K}$

$$\widetilde{\mathbf{K}} := \mathbf{HKH},$$

where $\mathbf{H} := \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$ is the centering matrix, and $\mathbf{J}_n$ is an $n \times n$ matrix with all entries being 1. The resulting matrix

$$\widetilde{\mathbf{K}} = \mathbf{K} - \mathbf{K}\left(\frac{1}{n}\mathbf{J}_n\right) - \left(\frac{1}{n}\mathbf{J}_n\right)\mathbf{K} + \left(\frac{1}{n}\mathbf{J}_n\right)\mathbf{K}\left(\frac{1}{n}\mathbf{J}_n\right)$$

corresponds to starting with a centered Φ as required.

IX. Sparse PCA

1. **Motivation:** We often interpret principal components by examining the direction vectors $\mathbf{u}_j$'s, referred to as *loadings*, to see which variables play a significant role. Often this interpretation is made easier if the loadings are *sparse*.

2. **Simplified Component Technique-LASSO (SCoTLASS):** SCoTLASS procedures consider the following optimization problem

$$\begin{aligned} &\text{maximize } \mathbf{u}^\top \boldsymbol{\Sigma}_{XX} \mathbf{u}, \\ &\text{subject to } \|\mathbf{u}\|_1 \le t, \|\mathbf{u}\|_2 = 1. \end{aligned} \tag{13}$$

The $\|\cdot\|_1$ constraint encourages some of the loadings to be zero and hence $\mathbf{u}$ to be sparse. Further sparse principal components are found in the same way, by forcing the k-th component to be orthogonal to the first $k-1$ components.

Remarks.

(a) The problem (13) is *not* convex and the computations are difficult;

(b) When $\mathbf{\Sigma}_{XX}$ is unknown, we can replace it by the sample covariance matrix constructed using the data;

(c) The problem (13) can be equivalently expressed as

$$\begin{aligned} &\underset{\mathbf{M}\succeq\mathbf{0}_{p\times p}}{\text{maximize}} \ \ \text{trace}(\mathbf{\Sigma}_{XX}\mathbf{M}) \\ &\text{subject to} \ \ \ \text{trace}(\mathbf{M})=1, \text{trace}(|\mathbf{M}|\mathbf{J}_p)\le t^2, \ \text{and} \ \ \text{rank}(\mathbf{M})=1, \end{aligned} \tag{14}$$

where $\mathbf{M}:=\mathbf{u}\mathbf{u}^\top\in\mathbb{R}^{p\times p}$ is a rank one matrix, $|\mathbf{M}|$ is obtained by taking absolute values entry-wise, and $\mathbf{J}_p$ is the $p\times p$ matrix with all entries being 1. In particular, note that $\|\mathbf{u}\|_2=1$ is equivalent to $\text{trace}(\mathbf{M})=1$, and $\|\mathbf{u}\|_1\le t$ is equivalent to $\text{trace}(|\mathbf{M}|\mathbf{J}_p)\le t^2$.

3. **Convex Relaxation of SCoTLASS (14):** The formulation (14) is non-convex, due to the presence of the constraint $\text{rank}(\mathbf{M})=1$. A relaxation is to drop this constraint and solve the following semi-definite program (SDP)

$$\begin{aligned} &\underset{\mathbf{M}\succeq\mathbf{0}_{p\times p}}{\text{maximize}} \ \ \text{trace}(\mathbf{\Sigma}_{XX}\mathbf{M}) \\ &\text{subject to} \ \ \ \text{trace}(\mathbf{M})=1, \ \text{and} \ \ \text{trace}(|\mathbf{M}|\mathbf{J}_p)\le t^2. \end{aligned}$$

Remarks.

(a) This relaxed problem is convex, it has no local optima, and a global optimum can be obtained by various standard methods.

(b) If we solve the SDP and *do* obtain a rank one solution, then we have in fact obtained the global optimum of the non-convex SCoTLASS criterion.

4. **Approach by Minimizing the Construction Error (Single Component):** For a single component, with $\mathbf{x}_1,\cdots,\mathbf{x}_n$ i.i.d samples from X, minimize

$$\begin{aligned} &\text{minimize} \ \ \frac{1}{n}\sum_{i=1}^n\left\|\mathbf{x}_i-\boldsymbol{\theta}\mathbf{u}^\top\mathbf{x}_i\right\|_2^2+\lambda\|\mathbf{u}\|_2^2+\nu\|\mathbf{u}\|_1, \\ &\text{subject to} \ \ \|\boldsymbol{\theta}\|_2=1. \end{aligned}$$

Note that

(a) If both λ and ν are zero and $n>p$, we have the minimizer of $\mathbf{u}$ and that of $\boldsymbol{\theta}$ are both the eigenvector associated with the largest eigenvalue of the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{XX}:=\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top$;

(b) When $p \gg n$, the solution is *not* necessarily unique unless $\lambda > 0$. For any $\lambda > 0$ and $\nu = 0$, the solution for $\mathbf{u}$ is proportional to the eigenvector associated with the largest eigenvalue of $\widehat{\mathbf{\Sigma}}_{XX}$;

(c) The second penalty on $\mathbf{u}$ encourages sparseness of the loadings.

5. **Approach by Minimizing the Construction Error (Multiple Components):** For $t > 1$ components, with $\mathbf{x}_1, \cdots, \mathbf{x}_n$ i.i.d samples from X, minimize

$$\begin{aligned} &\text{minimize } \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i - \mathbf{\Theta}\mathbf{U}^\top\mathbf{x}_i\right\|_2^2 + \lambda\sum_{k=1}^{t}\|\mathbf{u}_k\|_2^2 + \sum_{k=1}^{t}\nu_k\|\mathbf{u}_k\|_1, \\ &\text{subject to } \mathbf{\Theta}^\top\mathbf{\Theta} = \mathbf{I}_t. \end{aligned} \tag{15}$$

Here, $\mathbf{U} \in \mathbb{R}^{p\times t}$ is a matrix with columns $\mathbf{u}_k$ and $\mathbf{\Theta} \in \mathbb{R}^{p\times t}$.

Remarks.

(a) The optimization problem in (15) is *not* jointly convex in $\mathbf{U}$ and $\mathbf{\Theta}$, but it is convex in each parameter with the other parameter fixed;

(b) Minimization over $\mathbf{U}$ with $\mathbf{\Theta}$ fixed is equivalent to t elastic net problems and can be done efficiently;

(c) Minimization over $\mathbf{\Theta}$ with $\mathbf{U}$ fixed can be solved by a simple SVD calculation.

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Vol. 1. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC. ISBN: 1498712169.

Izenman, Alan J (Mar. 2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. en. Springer Science & Business Media.