Principal Component Analysis

This version: September 27, 2023

1 Dimension Reduction: Encoding and Decoding

Suppose that we have p variables. Denote our data as $\boldsymbol{X}$, which is an $n \times p$ matrix, and each observation is a $p \times 1$ vector. In dimension reduction, we aim to find a function $f(\cdot)$

$$f : \mathbb{R}^p \to \mathbb{R}^q, \quad q < p,$$

that can "encode" an observation $\boldsymbol{x}_i \in \mathbb{R}^p$ into a lower-dimensional space $\boldsymbol{z}_i = f(\boldsymbol{x}_i) \in \mathbb{R}^q$. Ideally, we would also like to have a function $g(\cdot)$

$$g : \mathbb{R}^q \to \mathbb{R}^p,$$

that "decodes" the observation back to the original space $\hat{\boldsymbol{x}}_i = g(\boldsymbol{z}_i)$. Since the encoder maps observations into a lower-dimensional space, information loss in the process is inevitable. A good encoder/decoder pair should minimize the reconstruction error L

$$L = \frac{1}{n} \sum_{i=1}^{n} ||\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i||_2^2.$$

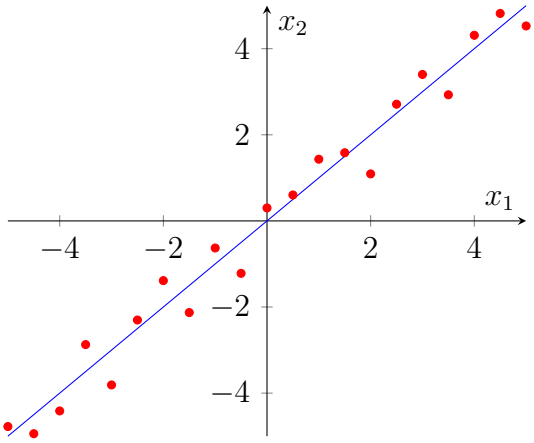
Remark 1. *For a vector $\boldsymbol{x} = (x_1, x_2, ..., x_p) \in \mathbb{R}^p$, the notation $||x_i||_l$ denotes its l-norm:*

$$\begin{aligned} ||\boldsymbol{x}||_l &= \left(\sum_{j=1}^{p} |x_j|^l \right)^{\frac{1}{l}} \\ &= \left(|x_1|^l + |x_2|^l + \cdots + |x_p|^l \right)^{\frac{1}{l}}. \end{aligned}$$

For example, when $l = 2$, it is also known as the L_2 norm, which corresponds to the Euclidean norm that we have been using since high school. Later in this course, we will also use the L_1 norm, denoted as $||x||_1$. However, unless otherwise specified, we usually consider the L_2 norm.

How do we find a encoder/ decoder? Consider the following scatter plot of our data $p = 2$:

Is our data two-dimensional or one-dimensional?



From the graph, we can see that most of our observations are close to the line 45-degree. So, why not project our data onto the line, which is an one dimensional space. Consider the following example.

Example 1 ("Eyeball" PCA)**.** *Suppose $p = 2$. From the scatter plot above, it appears that we can reduce the dimension of our data from $p = 2$ to $q = 1$ with minimal information loss by projecting the data onto the 45-degree line. Specifically, let*

$$\boldsymbol{A} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix},$$

and consider the function

$$f(\boldsymbol{x}) = \boldsymbol{A}^T \boldsymbol{x}$$

that maps from $\mathbb{R}^2$ to $\mathbb{R}^1$, and

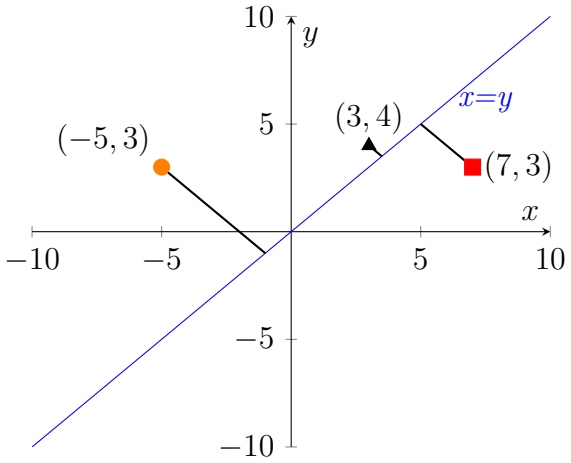
$$g(z) = \boldsymbol{A} z$$

that maps from $\mathbb{R}^1$ *to* $\mathbb{R}^2$.

You can verify that the encoder $f(\cdot)$ *is the function that projects observations onto the 45-degree line, and the compressed observation is given by* $\hat{\boldsymbol{x}} = g(z) = g(f(\boldsymbol{x}))$*, with the reconstruction error given by* $||\mathrm{x} - \widehat{\mathrm{x}}||$*. See the figures and table below for an illustration.*

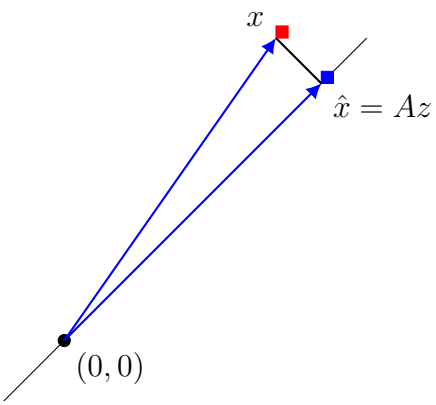
Quiz 1. *Can you find* $z \in \mathbb{R}$*,* $\hat{\boldsymbol{x}} \in \mathbb{R}^2$*, and the reconstruction error* $||\boldsymbol{x} - \hat{\boldsymbol{x}}||$ *for* $\boldsymbol{x} = (3,4)^T$*?*

Graph of projection



x	$A^T x$	$\hat{x} = Az$	$\|\|x - \hat{x}\|\|$
$(3,4)$	$\frac{7}{\sqrt{2}}$	$(\frac{7}{2}, \frac{7}{2})$	$\frac{1}{\sqrt{2}}$
$(7,3)$	$\frac{10}{\sqrt{2}}$	$(5,5)$	$2\sqrt{2}$
$(-5,3)$	$\frac{-2}{\sqrt{2}}$	$(-1,-1)$	$2\sqrt{5}$

Graph of $\boldsymbol{x} = \hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}$ *and* $\hat{\boldsymbol{x}} - \hat{\boldsymbol{z}}$



Example 1 essentially demonstrates how *principal component analysis (PCA)* operates in practice. PCA seeks to exploit the correlation between variables to capture the predominant variation in the data. In the extreme case, when the correlation coefficient is close to 1, two variables are nearly linearly dependent (i.e., $y \approx ax + b$ for some $a, b \in \mathbb{R}$), and they lie along a line, as illustrated in Example 1.

For $p = 2$ or $p = 3$, we have the luxury of visualizing the data to detect whether there exists a lower-dimensional space (a line or a plane) that adequately represents our data. However, this visual inspection, or "eyeballing," becomes unfeasible for dimensions beyond $p = 3$. What is our alternative? To address cases where $p > 3$, we must generalize our approach by abstracting the ideas of what we applied in lower dimensions.

2 Beyond $p = 3$: Deriving PCA from Scratch

As illustrated in Example 1, the core concept of PCA is to identify a lower-dimensional space onto which our observations can be projected. How can we formalize this concept for $p > 3$?

The key to generalization is abstraction. For $q = 1$, we may characterize PCA as

"finding (1) a line to (2) project onto, with the aim of minimizing the (3) reconstruction error." How might this statement be generalized for $q > 1$? Essentially, three questions need to be addressed:

1. What is the entity onto which we are projecting when $q > 1$?
2. How is projection defined for $q > 1$?
3. How is reconstruction defined for $q > 1$?

What is the entity onto which we are projecting when $q > 1$? For $q = 1$, we are looking for a line; for $q = 2$, we seek a plane, and so forth. Since one vector is required to span a line, and two vectors to define a plane, in general, we need q vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_q$ to construct a q-dimensional space. In PCA, our goal is to identify these vectors, termed as the "principal components", which form the lower-dimensional space. Here, $\boldsymbol{a}_1$ is the first principal component (PC), $\boldsymbol{a}_2$ is the second PC, and so on.

Project to	Find	Need vectors ("principal components")
$q = 1$	a line	$\boldsymbol{a}_1 \in \mathbb{R}^p$
$q = 2$	a plane	$\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^p$
$q = 3$	a 3D space	$\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3 \in \mathbb{R}^p$

Remark 2. *Note that if the original data is $p-$dimensional, we have at most p PC's.*

Given that the vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q \in \mathbb{R}^p$ fundamentally define a lower-dimensional subspace (of dimension q) in $\mathbb{R}^p$, we impose the conditions:

$$||\boldsymbol{a}_j||_2 = \sqrt{\boldsymbol{a}_j^T \boldsymbol{a}_j} = 1, \qquad \text{(unit length)}$$

$$\boldsymbol{a}_j^T \boldsymbol{a}_k = 0, \qquad \text{(orthogonality for } j \neq k\text{)}$$

The principal components $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q$ possess unit length as they fundamentally represent directions. Ensuring their orthogonality, as we will see below, simplifies the definition of the projection onto the space they span.

Quiz 2. *Which of the following vectors can possibly be the first two PC's for a data with three variables ($p = 3$)?*

1. $\boldsymbol{a}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \boldsymbol{a}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$

2. $\boldsymbol{a}_1 = \begin{bmatrix} 1 \\ \sqrt{2} \\ 0 \end{bmatrix}, \boldsymbol{a}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

3. $\boldsymbol{a}_1 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \boldsymbol{a}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

4. $\boldsymbol{a}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{a}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

Now, **how do we define project on** $q > 1$?[1] It turns out that if we define

$$\boldsymbol{A}_{p\times q} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_q \end{bmatrix}$$

then for an observation $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T \in \mathbb{R}^p$:

$$\boldsymbol{A}^T\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \\ \vdots \\ \boldsymbol{a}_q^T \end{bmatrix}_{q\times p} \cdot \boldsymbol{x}_{p\times 1} = \begin{bmatrix} \boldsymbol{a}_1^T\boldsymbol{x}_i \\ \boldsymbol{a}_2^T\boldsymbol{x}_i \\ \vdots \\ \boldsymbol{a}_q^T\boldsymbol{x}_i \end{bmatrix} \in \mathbb{R}^q.$$

So, the $j-$th component in $\boldsymbol{A}^T\boldsymbol{x}_i$ is simply the inner product of $\boldsymbol{a}_j$ and $\boldsymbol{x}_i$. Since $\boldsymbol{a}_j$ has unit length, then the projection of x_i on $\boldsymbol{a}_j$ is simply

$$\boldsymbol{a}_j^T\boldsymbol{x}_i \cdot \boldsymbol{a}_j.$$

So, the matrix multiplication $\boldsymbol{A}^T\boldsymbol{x}_i$ encodes $\boldsymbol{x}_i$ in terms of its coordinates on $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_q$. In matrix notation, the projection of $\boldsymbol{x}_i$ onto the space spanned by $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q$ is then

[1] Note that, in math, an understanding of the case for $q = 2$ would automatically generalize to any finite-dimensional case. So it is fine to assume $q = 2$ for the discussion.

given by

$$\begin{aligned}\hat{\boldsymbol{x}}_i &= \boldsymbol{a}_1^T\boldsymbol{x}_i \cdot \boldsymbol{a}_1 + \boldsymbol{a}_2^T\boldsymbol{x}_i \cdot \boldsymbol{a}_2 + \cdots + \boldsymbol{a}_q\boldsymbol{x}_i \cdot \boldsymbol{a}_q \\ &= \begin{bmatrix}\boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_q\end{bmatrix} \begin{bmatrix}\boldsymbol{a}_1^T\boldsymbol{x}_i \\ \boldsymbol{a}_2^T\boldsymbol{x}_i \\ \vdots \\ \boldsymbol{a}_q^T\boldsymbol{x}_i\end{bmatrix} \\ &= \boldsymbol{A}\boldsymbol{A}^T\boldsymbol{x}_i.\end{aligned}$$

Summary: to reduce the dimension from p to q, PCA aims to find q vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_q \in \mathbb{R}^p$ so that one can construct a $q-$ dimensional space to project our data onto. Our goal is find the optimal vectors (the principal components) that the reconstruction error is smallest. That is, we want to solve the minimization problem:

$$\begin{aligned}\min_{\boldsymbol{a}_1,\boldsymbol{a}_2,...,\boldsymbol{a}_p \in \mathbb{R}^p} & \frac{1}{n}\sum_{i=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{A}^T\boldsymbol{x}_i||_2^2 \\ s.t.\ \ & \boldsymbol{a}_j^T\boldsymbol{a}_j = 1,\ j = 1, 2, ..., q \\ & \boldsymbol{a}_j^T\boldsymbol{a}_k = 0,\ j \neq k\end{aligned}$$

where

$$\boldsymbol{A}_{p\times q} = \begin{bmatrix}\boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots \boldsymbol{a}_q\end{bmatrix}.$$

and $\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{x}_i \in \mathbb{R}^q$ is the projected (compressed) version of $\boldsymbol{x}_i \in \mathbb{R}^p$.

Quiz 3. *Suppose that $p = 3$. What is matrix $\boldsymbol{x}$ for the projection on the plane $x = y$?*

3 Solving PCA

So far, we have spent a lot of time defining PCA. But, how can we solve it, i.e., how do we find the PC's $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_p$?

PCA attempts to summarize variation in the random vector $\boldsymbol{X}$ with few principal components (PC). In below, we will define what are *principal components* and explain how

to derive them.

In fact, we can solve PCA iteratively, that is, we can first find the first PC, $\boldsymbol{a}_1$, then $\boldsymbol{a}_2$, ..., and $\boldsymbol{a}_q$. So, to find the first PC, we solve

$$\begin{aligned} \min_{\boldsymbol{a}_1 \in \mathbb{R}^p} & \frac{1}{n} \sum_{i=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{a}_1 \boldsymbol{a}_1^T \boldsymbol{x}_i||_2^2 \\ s.t. \;\; & \boldsymbol{a}_1^T \boldsymbol{a}_1 = 1. \end{aligned}$$

Given $\boldsymbol{a}_1$, we can then solve $\boldsymbol{a}_2$ by minimizing the remaining part $\boldsymbol{x}_i - \boldsymbol{a}_1 \boldsymbol{a}_1^T \boldsymbol{x}_i$

$$\begin{aligned} \min_{\boldsymbol{a}_2 \in \mathbb{R}^p} & \frac{1}{n} \sum_{i=1}^{n} ||(\boldsymbol{x}_i - \boldsymbol{a}_1 \boldsymbol{a}_1^T \boldsymbol{x}_i) - \boldsymbol{a}_2 \boldsymbol{a}_2^T \boldsymbol{x}_i||_2^2 \\ s.t. \;\; & \boldsymbol{a}_2^T \boldsymbol{a}_2 = 1, \\ & \boldsymbol{a}_2^T \boldsymbol{a}_1 = 0. \end{aligned}$$

While we motivated PCA by minimizing reconstruction errors, an alternative definition of PCA is by "maximizing variances". For example, the first PC can be found by solving the following problem:[2]

$$\max_{\boldsymbol{a}_1 \in \mathbb{R}^p} Var(\boldsymbol{a}_1' \boldsymbol{X}) \quad \text{s.t. } \boldsymbol{a}_1' \boldsymbol{a}_1 = 1.$$

Remark 3. *The restriction $\boldsymbol{a_i'} \boldsymbol{a_i} = 1$ is necessary, otherwise $Var(\boldsymbol{a}_i' X)$ can be made arbitrary large by multiplying the coefficient with a constant.*

The rest of the principal components can be defined iteratively. The second PC, $PC_2 = \boldsymbol{a}_2'$ is defined by the optimization problem:

$$\begin{aligned} \max_{\boldsymbol{a}_2 \in \mathbb{R}^p} & \quad Var(\boldsymbol{a}_2' \boldsymbol{X}) \\ \text{s.t. } & \boldsymbol{a}_2^T \boldsymbol{a}_2 = 1, \\ & \boldsymbol{a}_1^T \boldsymbol{a}_2 = 0. \end{aligned}$$

Notice that we require the second PC has to be uncorrelated with the first PC. We can think of it as we are trying to have the second PC explain the variation that is not explained by the first PC.

[2]For a rigor argument, see our textbook (Murphy 2022).

Similarly, the jth PC, $\boldsymbol{a}_j'$ is the solution to

$$\begin{aligned} \max_{\boldsymbol{a}_j \in \mathbb{R}^p} \quad & Var(\boldsymbol{a}_j' \boldsymbol{X}) \\ \text{s.t. } & \boldsymbol{a}_j^T \boldsymbol{a}_j = 1, \\ & \boldsymbol{a}_j^T \boldsymbol{a}_{j'} = 0 \text{ for } j' = 1, 2 \ldots, j-1. \end{aligned}$$

Remark 4. *Recall that*

$$\boldsymbol{a_j'x} = \langle \boldsymbol{a}_j, \boldsymbol{x} \rangle = \begin{pmatrix} a_{j1} & a_{j2} & \ldots & a_{jp} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jp}x_p,$$

so $\boldsymbol{a}'\boldsymbol{x}$ *is a scalar.*

3.1 Solve PCA when $\boldsymbol{\Sigma}$ Diagonal

Remark 5. *We can verify that*

$$Var(\boldsymbol{a^T X}) = \boldsymbol{a^T \Sigma a}.$$

See the following example.

Example 2. *Let* $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ *and* $\boldsymbol{a} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

$$\begin{aligned} Var(\boldsymbol{a^T X}) =& Var\left(\begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \right) \\ =& Var(X_1 + 2X_2) \\ =& Var(X_1) + 4Var(X_2) \\ =& 1 + 4 \cdot 2 = 9 \end{aligned}$$

Alternatively, we can calculate $Var(\boldsymbol{a}^T\boldsymbol{X})$ *by*

$$\boldsymbol{a^T\Sigma a} = \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \end{pmatrix} = 9.$$

Indeed, we end up with the same result, as the previous remarks implied.

By remark 5, the optimization problem can be written as

$$\max_{\boldsymbol{a}_1 \in \mathbb{R}^p} \quad \boldsymbol{a_1^T\Sigma a_1} \quad \text{s.t. } \boldsymbol{a_1^T a_1} = 1.$$

In his book "How to Solve It", mathematician and Probabilist George Pólya famously said

> "If you can't solve a problem, then there is an easier problem you can solve: find it."

Let's follow his suggestion by assuming the covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}_{p \times p},$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_p$. That is, we assume Σ is now a diagonal matrix with positive, decreasing diagonal elements.

Now we are ready to solve the optimization problem for PC_1:

$$\max_{\boldsymbol{a_1} \in \mathbb{R}^p} \quad \boldsymbol{a_1^T\Sigma a_1} \quad \text{s.t. } \boldsymbol{a^T a} = 1.$$

Notice the objective function is now

$$\begin{aligned}\boldsymbol{a_1^T \Sigma a_1} = &\begin{pmatrix} a_{11} & \dots & a_{1p} \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} \begin{pmatrix} a_{11} \\ \vdots \\ a_{1p} \end{pmatrix} \\ =&\lambda_1 a_{11}^2 + \lambda_2 a_{12}^2 + \cdots + \lambda_p a_{1p}^2.\end{aligned}$$

By redefining $b_{1i} = a_{1i}^2$, the maximization problem becomes

$$\max_{\boldsymbol{b}} \quad \lambda_1 b_{11} + \cdots + \lambda_p b_{1p} \quad \text{s.t. } b_{1i} + \cdots + b_{1p} = 1.$$

It is plain to see that $b_{11} = 1, b_{12} = \cdots = b_{1p} = 0$ attains the maximum, while subjecting to the constraint. Hence, $a_{11} = \pm 1, a_{12} = \cdots = a_{1p} = 0$ is the solution to the PC_1 problem.[3]

The first PC is hence

$$\boldsymbol{a}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

which points to the direction with largest variance.

Next we solve the PC_2 problem

$$\begin{aligned} \max_{\boldsymbol{a}_2 \in \mathbb{R}^p} \quad & Var(\boldsymbol{a}_2^T \boldsymbol{X}) \\ \text{s.t. } & \boldsymbol{a}_2^T \boldsymbol{a}_2 = 1 \\ & \boldsymbol{a}_1^T \boldsymbol{a}_2 = 0 \end{aligned}$$

[3] PCs are only uniquely defined up to reflection over the origin. It is not hard to see that if $\boldsymbol{a}^*$ is a solution to the PCA problem, then $-\boldsymbol{a}^*$ is also a solution.

The constraint $\boldsymbol{a}_1^T\boldsymbol{a}_2 = 0$ implies $a_{21} = 0$. So $a_2 = \begin{pmatrix} 0 \\ a_{22} \\ \vdots \\ a_{2p} \end{pmatrix}$, and the problem reduces to

$$\begin{aligned} \boldsymbol{a_T^T \Sigma a_2} &= \begin{pmatrix} 0 & a_{22} & \dots & a_{2p} \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} \begin{pmatrix} 0 \\ a_{22} \\ \vdots \\ a_{2p} \end{pmatrix} \\ &= \begin{pmatrix} a_{22} & \dots & a_{2p} \end{pmatrix} \begin{pmatrix} \lambda_2 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} \begin{pmatrix} a_{22} \\ \vdots \\ a_{2p} \end{pmatrix}. \end{aligned}$$

The problem is now the same as finding PC_1, with one less variable. Hence the solution to the second component is $a_{22} = \pm 1$, $a_{21} = a_{23} = \cdots = a_{2p} = 0$, and the second PC is $a_2 = (0, 1, 0, \cdots, 0)^T$. Similarly, we can see that $j-$th PC are just the unit vectors in $\mathbb{R}^p$ in which whose $j-$th component is 1 andzero otherwise.

Conclusion: When X_j's are uncorrelated, PCA is basically keeping X_j's with largest variances.

3.2 Solve PCA when $\boldsymbol{\Sigma}$ is not Diagonal

What if $\boldsymbol{\Sigma}$ is not diagonal? Luckily, we have the following theorem:

Theorem 1 (Real Spectral Theorem)**.** *If $\boldsymbol{\Sigma}$ is symmetric, then there exists a $p \times p$ matrix $\boldsymbol{P}$ such that*

$$\boldsymbol{\Sigma = PDP^{-1}},$$

where D is a diagonal matrix, and

$$\boldsymbol{P^T P = I}_{p\times p},$$

i.e., $\boldsymbol{P}^{-1} = \boldsymbol{P^T}$.

Example 3.

$$\begin{pmatrix} 34 & 12 \\ 12 & 41 \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & \frac{-4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{pmatrix} \begin{pmatrix} 50 & 0 \\ 0 & 25 \end{pmatrix} \begin{pmatrix} \frac{3}{5} & \frac{4}{5} \\ \frac{-4}{5} & \frac{3}{5} \end{pmatrix}.$$

Since $\boldsymbol{\Sigma}$ is symmetric, it can now be written as $\boldsymbol{\Sigma} = \boldsymbol{PDP^T}$, where the notation is followed by the Real Spectral theorem. Let $\boldsymbol{b}_i = \boldsymbol{P^T a}_i$,

$$\boldsymbol{a_i^T \Sigma a_i} = \boldsymbol{a_i^T PDP^T a_i} = \boldsymbol{(P^T a_i)^T D(P^T a_i)} = \boldsymbol{b_i^T D b_i}.$$

$$\boldsymbol{a_i^T a_i} = \boldsymbol{a_i^T (PP^{-1}) a_i} = \boldsymbol{a_i^T (PP^T) a_i} = \boldsymbol{(P^T a_i)^T (P^T a_i)} = \boldsymbol{b_i^T b_i}$$

Restate the maximization problem in terms of $\boldsymbol{b}_i$ gives

$$\max_{\boldsymbol{b}_i \in \mathbb{R}^p} \quad \boldsymbol{b_i' D b_i} \quad \text{s.t. } \boldsymbol{b_i' b_i} = 1,$$

and we know how to solve it since $\boldsymbol{D}$ is diagonal.

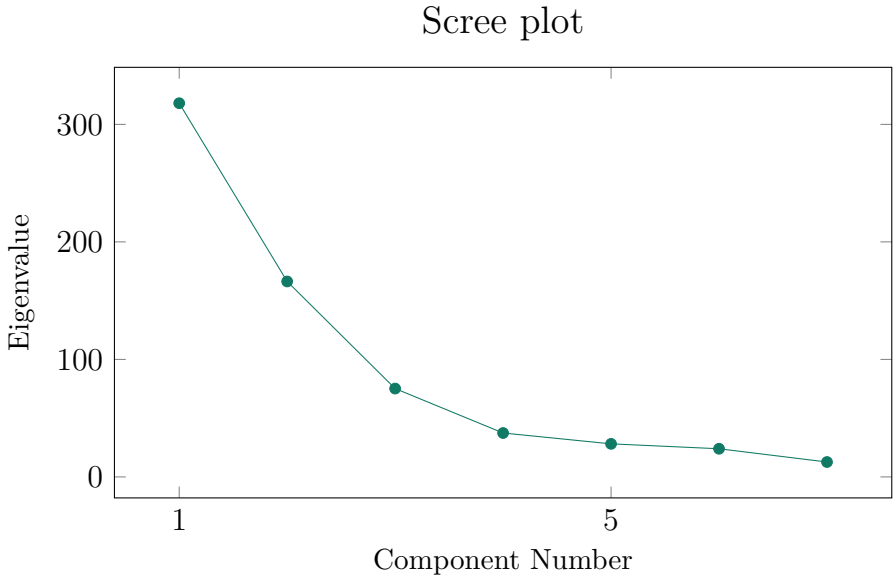
4 Applications of PCA

Here we list examples of applications of PCA.

1. Data compression
2. Summarize data and explorative analysis
3. Use PC in regression
4. Use PC in clustering
5. Factor analysis

5 Choosing q

One might ask how much variance is summarized in q, and how we choose such q. A scree plot might be helpful to answer the question.



The y-axis of a scree plot is the sorted eigenvalues of the covariance matrix, in a decreasing order, and the x-axis is the corresponding rank of the eigenvalue. As a rule of thumb, q is chosen at the elbow of the scree plot, so one might choose $q = 2$ for the example above.

Suppose that the covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}_{p \times p}$ is a diagonal matrix, then we have

$$
\begin{aligned}
Var(PC_1) = Var(X_1) &= \lambda_1 \\
Var(PC_1) = Var(X_2) &= \lambda_2 \\
&\vdots \\
Var(PC_p) = Var(X_p) &= \lambda_p
\end{aligned}
$$