Multivariate Analysis Lecture 10: Principal Component Analysis

Zhaoxia Yu
Professor, Department of Statistics

2023-05-04

Section 1

Intro to PCA

Introduction to PCA

- A component refers to a linear function of features/variables
- In English, “principal” means first or highest in rank/importance
- The first principal component refers to the linear function of the highest “rank/importance”
- The second principal component refers to the linear function of the second highest “rank/importance”
- The principal components in PCA are **red**{uncorrelated} **blue**{linear} combinations/functions of features

Origins of PCA

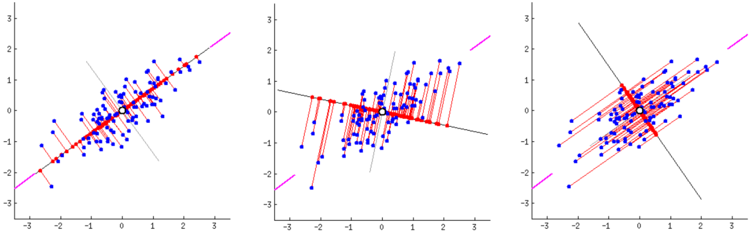
- PCA was first introduced by Karl Pearson in 1901 as a method to study the “lines and planes of closest fit” for high-dimensional data.
- Pearson developed PCA as a geometrical technique to find the direction that maximizes the variance in multivariate data.
- In 1933, Harold Hotelling extended PCA, establishing its statistical properties and mathematical foundation.
- Hotelling showed that PCA is equivalent to finding the eigenvectors and eigenvalues of the covariance matrix, thus connecting PCA to spectral decomposition.

Modern PCA

- Over time, PCA has become a widely used method in various disciplines, such as statistics, data science, finance, and engineering.
- PCA has had a significant impact on the field of multivariate analysis and dimensionality reduction.
- PCA has been extended and generalized, giving rise to many useful non-linear dimension reduction techniques like Kernel PCA, Sparse PCA, and UMAP.
- Its versatility and interpretability have made PCA a go-to technique for data visualization, noise reduction, and feature extraction.

A Visual Illustration of PCA

Click the link to see the animated version!



Revisit Example 2 in Lecture 08

```
n=1000
Sigma1=diag(c(4,1), 2, 2)
Sigma2=diag(c(1,4), 2, 2)
theta=pi/6
R1=matrix(c(cos(theta), sin(theta), -sin(theta), cos(theta)), 2,2)
theta=pi/4+pi/2
R2=matrix(c(cos(theta), sin(theta), -sin(theta), cos(theta)), 2,2)
Sigma3=R1%*%Sigma1%*%t(R1)
Sigma4=R2%*%Sigma1%*%t(R2)
set.seed(1)
X1=data.frame(mvrnorm(n, rep(0,2), Sigma1)); names(X1)=c("x","y")
X2=data.frame(mvrnorm(n, rep(0,2), Sigma2)); names(X2)=c("x","y")
X3=data.frame(mvrnorm(n, rep(0,2), Sigma3)); names(X3)=c("x","y")
X4=data.frame(mvrnorm(n, rep(0,2), Sigma4)); names(X4)=c("x","y")
```

Covariance Matrices

```
Sigma1
```

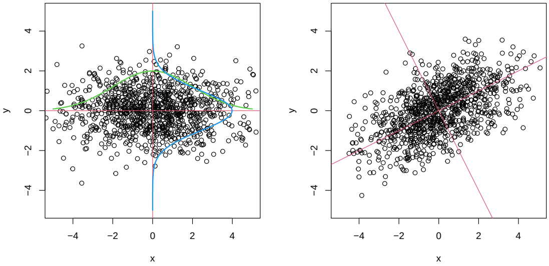
```
##      [,1] [,2]
## [1,]    4    0
## [2,]    0    1
```

```
Sigma3
```

```
##          [,1]     [,2]
## [1,] 3.250000 1.299038
## [2,] 1.299038 1.750000
```

Simulated Data

```
par(mfrow=c(1,2),pty="s")
plot(X1, xlim=c(-5,5), ylim=c(-5,5));
abline(0,0, col=2); abline(v=0, col=2)
lines(seq(-5,5,0.1), 10*dnorm(seq(-5,5,0.1), 0, 2), col=3, lwd=2)
lines(10*dnorm(seq(-5,5,0.1), 0, 1), seq(-5,5,0.1), col=4, lwd=2)
plot(X3, xlim=c(-5,5), ylim=c(-5,5));
abline(0,1/2, col=2); abline(0, -2, col=2)
```



Section 2

Theory and Spectral Decomposition

A Linear Combination of a Random Vector

- Consider a random vector $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$
- Suppose its covariance matrix is $\mathbf{\Sigma}$
- Let $a \in \mathbb{R}^p$ be a vector of length p.
- Consider a linear combination/function of $\mathbf{X}$, denoted by

$$Y = a^T\mathbf{X} = \sum_{i=1}^{p} a_i X_i = a_1X_1 + \cdots + a_pX_p,$$

- The variance of the random variable Y is

$$Var(Y) = Var(a^T\mathbf{X}) = a^T\mathbf{\Sigma}a \;\; (= \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j \sigma_{ij})$$

A Linear Combination of a Random Vector

- It is obvious that $Var(Y)$ depends on the scale of a, thus, it is not scale free.
- Let's put on a constraint: $||a|| = 1$, i.e., the norm of the vector a is fixed at 1. Note, alternatively, we can write

$$1 =< a, a >= a^T a = \sum_{i=1}^{p} a_i^2$$

- Can we maximize $Var(Y) = Var(a^T X)$ subject to $||a|| = 1$?
- We can. To do so, we will first introduce the spectral decomposition of a symmetric matrix and then apply this result to $\mathbf{\Sigma}$.

Subsection 1

Spectral Decomposition

Spectral Decomposition of A Symmetric Matrix A

- Spectral decomposition, also known as eigendecomposition, is a process by which a symmetric matrix is decomposed into a set of orthogonal eigenvectors and their corresponding eigenvalues.
- A symmetric matrix has real eigenvalues and orthogonal eigenvectors.
- For a symmetric matrix $A_{p\times p}$, the spectral decomposition is given by: $A = \Gamma\Lambda\Gamma^T$, where
  - Γ is an orthogonal matrix
  - Λ is a diagonal matrix

Spectral Decomposition of A Symmetric Matrix A

- Γ is an orthogonal matrix, i.e., Γ s.t.

$$\Gamma\Gamma^T = \Gamma^T\Gamma = \mathbf{I}_p$$

- The columns of Γ are the eigenvectors of A. Let γ_i denote the ith column, then $\Gamma = (\gamma_1, \cdots, \gamma_p)$ and each γ_i is a $p \times 1$ vector; in other words, $\gamma_i \in \mathbb{R}^p$.
- Γ is an orthogonal matrix. This implies that $\mathbf{I} = \Gamma\Gamma^T = (\gamma_1, \cdots, \gamma_p)(\gamma_1, \cdots, \gamma_p)^T = \sum_{i=1}^p \gamma_i\gamma_i^T$ and

$$\gamma_i^T\gamma_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

The Diagonal Matrix $\boldsymbol{\Lambda}$

- Λ is a diagonal matrix containing the eigenvalues of A, with $\Lambda_{ii} = \lambda_i$:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \cdots & 0 \\ 0 & \lambda_2 & 0 \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \lambda_p \end{pmatrix}$$

- Without of loss of generality, we often rank the eigenvalues from the largest to the smallest, i.e.,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

The Eigenvectors and Eigenvalues of a Symmetric Matrix A

- λ_i is the ith eigenvalue and γ is the corresponding eigenvector, i.e.,
$$A\gamma_i = \lambda_i\gamma_i$$
- The spectral decomposition $A = \Gamma\Lambda\Gamma^T$ implies that
$$A = \sum_{i=1}^{p} \lambda_i\gamma_i\gamma_i^T$$

The Spectral Decomposition of A Covariance Matrix

- Let $\mathbf{\Sigma}_{p\times p}$ be the covariance matrix of $\mathbf{X}$.
- A covariance matrix is positive definite or positive semi-definite, which means

$$\Sigma = \Gamma\Lambda\Gamma^T$$

where Λ is diagonal matrix with non-negative diagonal elements:

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$$

Subsection 2

Examples of Spectral Decomposition

Example 1 of Spectral Decomposition

```
Sigma1
```

```
##      [,1] [,2]
## [1,]    4    0
## [2,]    0    1
```

```
eigen(Sigma1)$value #\lambda's
```

```
## [1] 4 1
```

```
eigen(Sigma1)$vectors #\Gamma
```

```
##      [,1] [,2]
## [1,]   -1    0
## [2,]    0   -1
```

Example 1 of Spectral Decomposition

```
eigen(Sigma1)$vectors %*% t(eigen(Sigma1)$vectors)#\Lambda* \Lambda^T
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

```
t(eigen(Sigma1)$vectors) %*% eigen(Sigma1)$vectors
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

```
#\Gamma \Lambda \Gamma^T
eigen(Sigma1)$vectors %*% diag(eigen(Sigma1)$values) %*% eigen(Sigma1)$vectors
```

```
##      [,1] [,2]
## [1,]    4    0
## [2,]    0    1
```

Example 2 of Spectral Decomposition

```r
Sigma3
```

```
##           [,1]     [,2]
## [1,] 3.250000 1.299038
## [2,] 1.299038 1.750000
```

```r
eigen(Sigma3)$value #\lambda's
```

```
## [1] 4 1
```

```r
eigen(Sigma3)$vectors #\Gamma
```

```
##            [,1]       [,2]
## [1,] -0.8660254  0.5000000
## [2,] -0.5000000 -0.8660254
```

Example 2 of Spectral Decomposition

```
eigen(Sigma1)$vectors %*% t(eigen(Sigma1)$vectors) #\Lambda* \Lambda^T
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

```
t(eigen(Sigma1)$vectors) %*% eigen(Sigma1)$vectors #\Lambda^T* \Lambda
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

```
#\Gamma \Lambda \Gamma^T
eigen(Sigma3)$vectors %*% diag(eigen(Sigma3)$values) %*% t(eigen(Sigma3)$vectors)
```

```
##          [,1]     [,2]
## [1,] 3.250000 1.299038
## [2,] 1.299038 1.750000
```

Section 3

First PC

The Maximum Variance of $a^T\mathbf{X}$ S.B.T $||a|| = 1$

- Let $Y_1 = a^T\mathbf{X}$ denote the first principal component, which is defined as the linear combination reaches the maximum variance subject to $||a|| = 1$. Mathematically, we are looking for a s.t.

$$a = \arg\max_{a^Ta=1} a^T\mathbf{\Sigma}a$$

- The variance of Y_1 in terms of Γ and Λ

$$\begin{aligned} Var(Y_1) &= a^T\Sigma a \\ &= a^T\Gamma\Lambda\Gamma^T a = a^T(\sum_{i=1}^{p}\lambda_i\gamma_i\gamma_i^T)a \\ &= \sum_{i=1}^{p}\lambda_i a^T\gamma_i\gamma_i^T a \end{aligned}$$

The Maximum Variance of $a^T\mathbf{X}$ S.B.T $||a|| = 1$

- Let $z_i = a^T\gamma_i$. Note that

$$Var(Y) = \sum_{i=1}^{p} \lambda_i z_i^2$$

  This is because z_i is a scalar, thus $z_i^2 = z_i^T z_i = z_i z_i^T$

- You can also see that z_i is the inner product between a and γ_i:

$$z_i = a^T\gamma_i =< a, \gamma_i >=< \gamma_i, a >= \gamma_i^T a$$

The Maximum Variance of $a^T\mathbf{X}$ S.B.T $||a|| = 1$

- Also,

$$\begin{aligned}\sum_{i=1}^{p} z_i^2 &= \sum_{i=1}^{p} a^T\gamma_i\gamma_i^T a = a^T\left(\sum_{i=1}^{p}\gamma_i\gamma_i^T\right)a \\ &= a^T\Gamma\Gamma^T a \\ &\overset{1}{=} a^T\mathbf{I}a \\ &= a^T a \\ &\overset{2}{=} 1\end{aligned}$$

  - step 1: this is because Γ is an orthogonal matrix, which means $\Gamma\Gamma^T = \Gamma^T\Gamma = \mathbf{I}$.
  - step 2: this is due to the constraint that $a^T a = 1$.

The Maximum Variance of $a^T\mathbf{X}$ S.B.T $||a|| = 1$

- So far we have the following results
- $Var(a^TX) = \sum_{i=1}^{p} \lambda_i z_i^2$, where

$$\lambda_1 \geq \cdots \lambda_p \geq 0 \;\text{ and }\; \sum_{i=1}^{p} z_i^2 = 1$$

- Thus,

$$\begin{aligned} Var(a^T\mathbf{X}) &= \sum_{i=1}^{p} \lambda_i z_i^2 \overset{?}{\leq} \sum_{i=1}^{p} \lambda_1 z_i^2 \\ &= \lambda_1 \sum_{i=1}^{p} z_i^2 \\ &= \lambda_1 \end{aligned}$$

Thus, the maximum $Var(a^TX) = \lambda_1$ s.b.t. $||a|| = 1$.

The a (s.b.t$||a|| = 1$) Maximizes $Var(a^T X)$

- But how to find a?
- Which $z = (z_1, \cdots, z_p)^T$ makes the $=$ hold?
- This happens when $z_1 = 1, z_2 = 0, \cdots z_p = 0$
- Recall that $z_i = a^T \gamma_i$
- Thus, the following a satisfies all required conditions

$$a = \gamma_1$$

- Thus, we can conclude that

First Principal Component: Among all the linear combinations of $\mathbf{X}$, the one with the maximum variance is $\gamma_1 \mathbf{X}$ and the corresponding variance is λ_1.

- For notional clarity, let's denote the first PC by $Y_1 = \gamma_1^T \mathbf{X}$

Section 4

Understand 1st PC

Example

```
Sigma3
```

```
##          [,1]     [,2]
## [1,] 3.250000 1.299038
## [2,] 1.299038 1.750000
```

```
gamma1=eigen(Sigma3)$vectors[,1]
gamma1
```

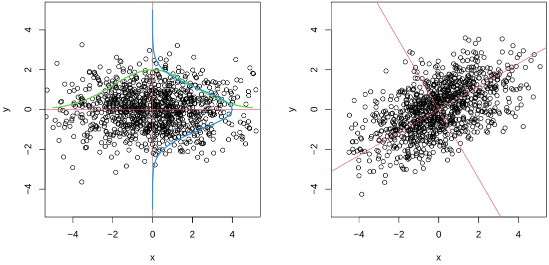
```
## [1] -0.8660254 -0.5000000
```

```
gamma2=eigen(Sigma3)$vectors[,2]
gamma2
```

```
## [1]  0.5000000 -0.8660254
```

Simulated Data

```
par(mfrow=c(1,2),pty="s")
plot(X1, xlim=c(-5,5), ylim=c(-5,5));
abline(0,0, col=2); abline(v=0, col=2)
lines(seq(-5,5,0.1), 10*dnorm(seq(-5,5,0.1), 0, 2), col=3, lwd=2)
lines(10*dnorm(seq(-5,5,0.1), 0, 1), seq(-5,5,0.1), col=4, lwd=2)
plot(X3, xlim=c(-5,5), ylim=c(-5,5));
abline(0,1/sqrt(3), col=2); abline(0, -sqrt(3), col=2)
```



Project One Vector on Another

- Let x and y be two vectors of the same length. Say both x and y are in $\mathbf{R}^k$.
- The direction of $proj_x(y)$ is the same as that of x.
- Let θ is the angle between x and y.

$$cos(\theta) = \frac{x^T y}{||x||||y||}$$

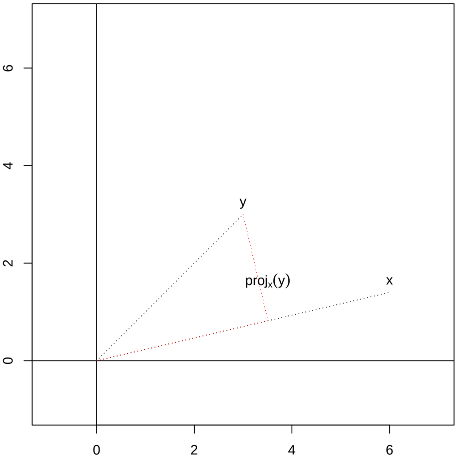
- The length of the projection is $||y||cos(\theta)$.
- The projection of y on x is

$$||y||cos(\theta)\frac{x}{||x||} = \frac{x^T y}{||x||^2}x$$

Example: Project One Vector on Another

```r
# Define the vectors
y <- c(3, 3)
x <- c(6, 1.4)
# Compute the projection of v1 onto v2
proj <- sum(y * x) / sum(x * x) * x
# Create a plot
par(pty="s")
plot(0, 0, xlim = c(-1, 7), ylim = c(-1, 7), type = "n", xlab = " ", ylab = " ")
abline(h = 0, v = 0)
text(y[1], y[2], "y", pos = 3)
text(x[1], x[2], "x", pos = 3)
text(proj[1], proj[2]+0.5, expression(proj[x](y)), pos = 3)
segments(0, 0, y[1], y[2], lty = "dotted")
segments(0, 0, x[1], x[2], lty = "dotted")
segments(y[1], y[2], proj[1], proj[2], lty = "dotted", col = "red")
segments(0, 0, proj[1], proj[2], lty = "dotted", col = "red")
```

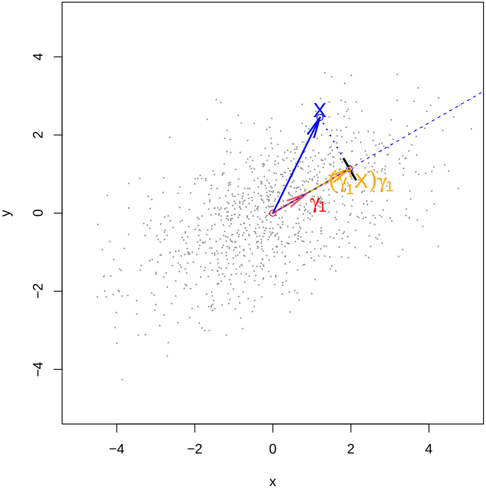
Example: Project One Vector on Another



Example: Project An Observation to 1st PC

```
par(mfrow=c(1,1),pty="s")
obs=unlist(X3[1,])
proj=c(matrix(gamma1,1,2)%*%matrix(obs,2,1)) *gamma1
plot(X3, xlim=c(-5,5), ylim=c(-5,5), col=8, pch=".");
points(x=0,y=0, col="red")
points(x=obs[1], y=obs[2], col="blue")
points(x=proj[1], y=proj[2], col="blue")
arrows(x0=0, y0=0,x1=obs[1], y1=obs[2], col="blue", lwd=2, angle=10)
arrows(x0=0, y0=0,x1=proj[1], y1=proj[2], col="orange", lwd=2, angle=10)
arrows(x0=0, y0=0,x1=-gamma1[1], y1=-gamma1[2], col=2, lwd=2, angle=10)
segments(0,0, -10*gamma1[1], -10*gamma1[2], col="blue", lty=2)
segments(obs[1], obs[2], proj[1], proj[2], lty = "dotted", col = "blue", lwd=2)
text(x=proj[1],y=proj[2], labels="|", col="black", srt=30, cex=2)
text(x=-gamma1[1]+0.3, y=-gamma1[2]-0.3, labels=expression(gamma[1]), col="red", cex=1.5)
text(x=obs[1], y=obs[2]+0.2, labels="X", col="blue", cex=1.5)
text(x=proj[1]+0.3, y=proj[2]-0.3, labels=expression( (gamma[1]^T * X)*gamma[1]), col="orange", cex=1.5)
```

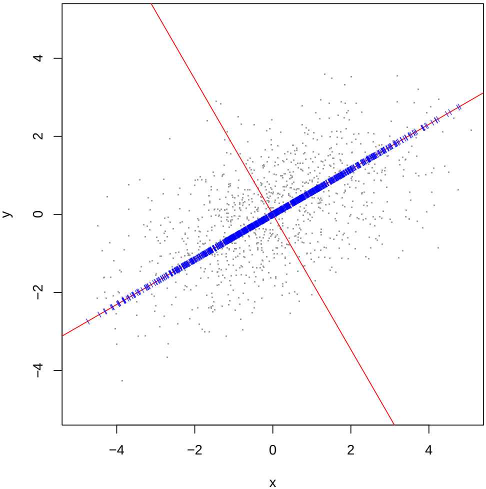
Example: Project An Observation to 1st PC



Example: Project All Observations to 1st PC

```
par(mfrow=c(1,1),pty="s")
plot(X3, xlim=c(-5,5), ylim=c(-5,5), col=8, pch=".");
#arrows(x0=0, y0=0,x1=-6*gamma1[1], y1=-6*gamma1[2], col=2, lwd=2, angle=10)
#arrows(x0=0, y0=0,x1=6*gamma1[1], y1=6*gamma1[2], col=2, lwd=2, angle=10)
abline(0,1/sqrt(3), col="red"); abline(0, -sqrt(3), col="red")
points(x=0,y=0, col="red")
for(i in 1:1000){
  proj=c(matrix(gamma1,1,2)%*% matrix(unlist(X3[i,]),2,1)) *gamma1
  #points(x=proj[1],y=proj[2], col=2, pch="|")
  text(x=proj[1],y=proj[2], labels="|", col="blue", srt=30, cex=0.5)
}
```

Example: Project All Observations to 1st PC



Section 5

2nd and ith PC

Subsection 1

2nd PC

Definition of the 2nd PC

- Let $Y_2 = a^T\mathbf{X}$ denote the second PC. It is defined as a linear combination of $\mathbf{X}$ such that
  1. It is uncorrelated to $Y_1 = \gamma_1^T\mathbf{X}$, i.e.,
  $$0 = cov(a^T\mathbf{X}, \gamma_1^T\mathbf{X}) = a^T\mathbf{\Sigma}\gamma_1$$
  2. The linear coefficients a has norm 1, i.e.,
  $$1 = ||a|| = ||a||^2 = a^T a$$
  3. It reaches the maximum variance among all the linear combinations satisfying the first two conditions

Mathematically, we are looking for a s.t.

$$a = \underset{a^T a=1, a^T\mathbf{\Sigma}\gamma_1=0}{\arg\max} \; a^T\mathbf{\Sigma}a$$

Identify the 2nd PC

- We would like to

$$max(a^T \mathbf{\Sigma} a) \text{ s.b.t. } a^T a = 1 \text{ and } a^T \mathbf{\Sigma} \gamma_1 = 0$$

Rewrite $Var(a^T \mathbf{X})$:

$$Var(a^T \mathbf{X}) = a^T \mathbf{\Sigma} a = a^T \left(\sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^T \right) a$$

$$= \sum_{i=1}^{p} \lambda_i a^T \gamma_i \gamma_i^T a$$

$$\text{Let } \overset{z_i = a^T \gamma_i}{=} \sum_{i=1}^{p} \lambda_i z_i^2$$

Identify the 2nd PC

- The first constraint of a is

$$a^T \mathbf{\Sigma} \gamma_1 = 0$$

- Recall that γ_1 is an eigenvector of $\mathbf{\Sigma}$ with the corresponding eigenvalue λ_1, we have

$$\mathbf{\Sigma} \gamma_1 = \lambda_1 \gamma_1$$

Thus the constraint $a^T \mathbf{\Sigma} \gamma_1 = 0$ implies that $a^T \gamma_1 = 0$, which further implies that $z_1 = 0$. As a result, we have

$$Var(a^T \mathbf{X}) = \sum_{i=2}^{p} \lambda_i z_i^2$$

Identify the 2nd PC

- The second constraint of a is

$$a^T a = 1$$

With this constraint, we have

$$\begin{aligned} 1 = a^T a = a^T \mathbf{I} a = a^T \Gamma \Gamma^T a \\ = (z_1, \cdots, z_p)(z_1, \cdots, z_p)^T \\ = \sum_{i=1}^{p} z_i^2 \\ = \sum_{i=2}^{p} z_i^2 \end{aligned}$$

Identify the 2nd PC

- From the previous slide

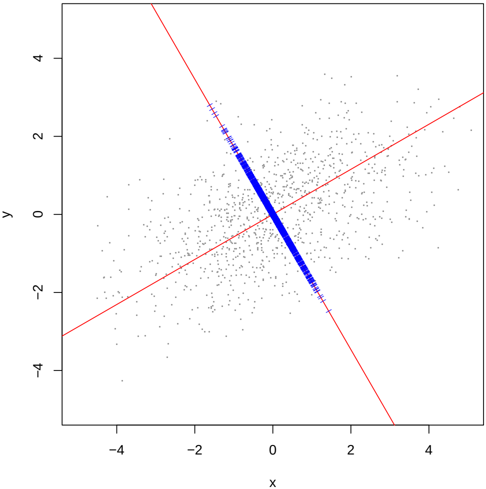
$$\begin{aligned} Var(a^T\mathbf{X}) &= \sum_{i=2}^{p} \lambda_i z_i^2 \\ &\leq \sum_{i=2}^{p} \lambda_2 z_i^2 \\ &= \lambda_2 \end{aligned}$$

- It can also be verified that $a = \gamma_2$ leads to this maximum.
- Therefore, the second PC is $Y_2 = \gamma_2^T\mathbf{X}$.

Example: Project All Observations to 2nd PC

```
par(mfrow=c(1,1),pty="s")
plot(X3, xlim=c(-5,5), ylim=c(-5,5), col=8, pch=".");
#arrows(x0=0, y0=0,x1=-6*gamma2[1], y1=-6*gamma2[2], col=2, lwd=2, angle=10)
#arrows(x0=0, y0=0,x1=6*gamma2[1], y1=6*gamma2[2], col=2, lwd=2, angle=10)
abline(0,1/sqrt(3), col="red"); abline(0, -sqrt(3), col="red")
points(x=0,y=0, col="red")
for(i in 1:1000){
  proj=c(matrix(gamma2,1,2)%*% matrix(unlist(X3[i,]),2,1)) *gamma2
  #points(x=proj[1],y=proj[2], col=2, pch="|")
  text(x=proj[1],y=proj[2], labels="|", col="blue", srt=120, cex=0.5)
}
```

Example: Project All Observations to 2nd PC



Subsection 2

ith PC

Identify the ith PC

- You probably can guess that the *i*th principal component is

$$Y_i = \gamma_i^T \mathbf{X}$$

- For the *i*th principal component, we are looking for a linear combination in terms of $a^T\mathbf{X}$ such as

$$a = \underset{a^Ta=1, a^T\gamma_1=0, \cdots, a^T\gamma_{i-1}=0}{\arg\max} \quad a^T\mathbf{\Sigma}a$$

- Note that $a^T\mathbf{\Sigma}\gamma_i = a^T\gamma_i$ because ...
- Use the same method, we will see that the *i*th principal component is

$$Y_i = \gamma_i^T \mathbf{X}$$