Principal Component Analysis

Data:

x_1	x_2	...	x_{1000}

dimension = 1000

$\xrightarrow[\text{reduction}]{\text{dimension}}$

u_1	u_2	...	u_{20}

dim = 20

Dimension Reduction

- remove some "less important" variables: (LASSO, Decision tree, Forest, Gradient boosting ...)

  variable selection | Feature selections techniques

- create a set of fewer variables from the original variables: PCA ...

  variable extractions | Features extractions techniques.

Data

X_1	X_2	X_3	X_4	X_5	Y

$\rightarrow$

u_1	u_2	Y

"Truth"

$$Y = \underbrace{(X_1 + 2X_2 - X_3)}_{u_1}^2 + \cos(\underbrace{X_1 + 6X_5}_{u_2}) + \varepsilon$$

$$E(Y \mid X_1, X_2 \ldots X_5) = E(Y \mid u_1, u_2)$$

$$u_1 = X_1 + 2X_2 - X_3$$

$$u_3 = X_1 + 6X_5$$

* Dimension Reduction

Data 1

x_1	x_2	x_3	x_4	x_5

Five variables or the dimension is 5

$d = 5$

variable extraction (PCA)

variable selection

x_1	x_3

$(d = 2)$

Data 2

$u_1 =$ $2x_1 + x_3$	$u_2 =$ $x_4 + 6x_5$

Example:

$$y = (2x_1 + x_3)^3 + \log(x_4 + 6x_5)^2$$

PCA in a view or coordinate rotation

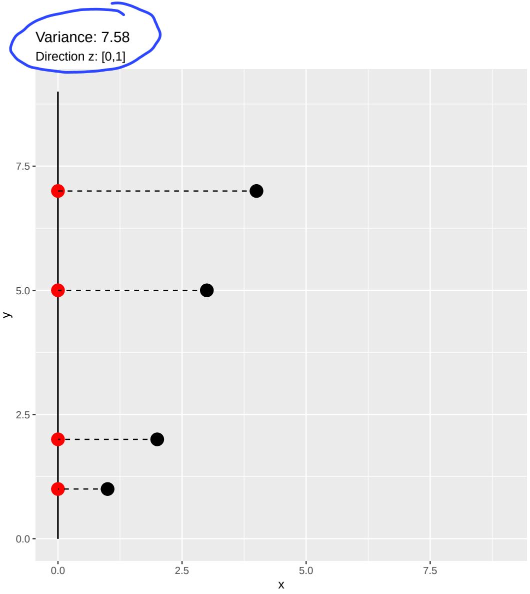
Variance of the Projection

Data :

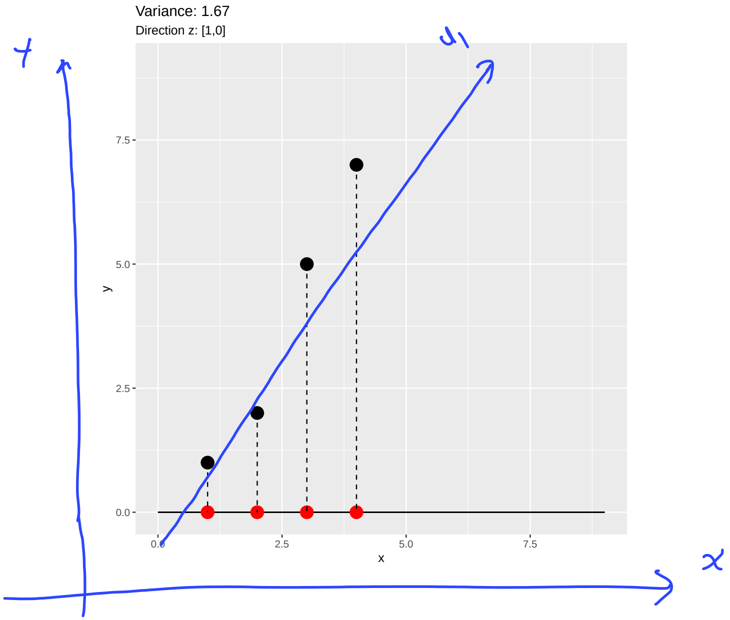
x	y	z
1	1	2024
2	2	2024
3	5	2024
4	7	2024

$\rightarrow V(z) = 0$

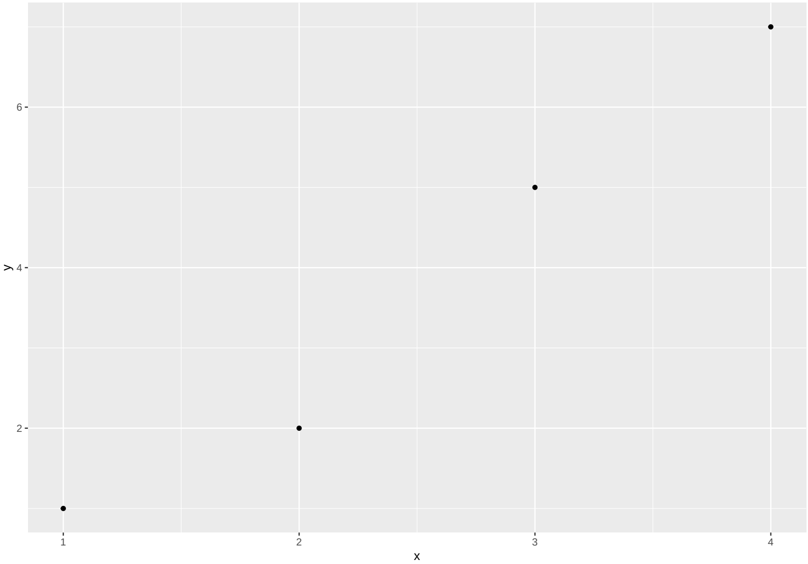
- $V(x) = 1.67$
- $V(y) = 7.58$
- Total variance: $V(x) + V(y) = 9.25$



Variance: 7.58
Direction z: [0,1]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

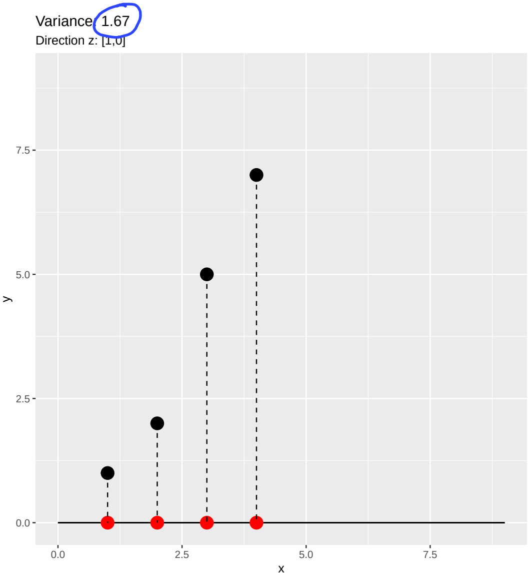


Variance: 1.67
Direction z: [1,0]
y
x
0.0
2.5
5.0
7.5
0.0
2.5
5.0
7.5
y
x
u1



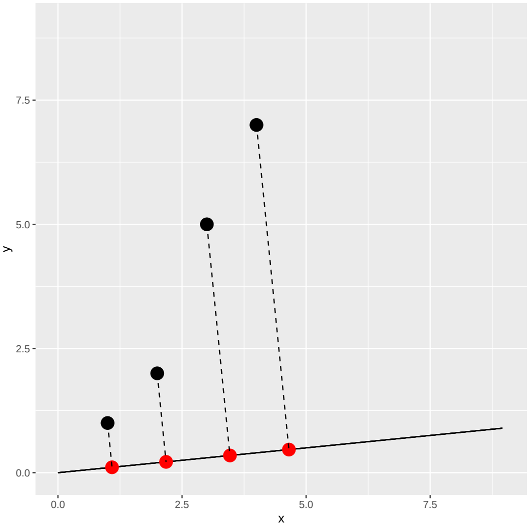
6
4
2
y
1
2
3
4
x

```
[1] 9.25
```


Variance: 1.67
Direction z: [1,0]
y
x
0.0
2.5
5.0
7.5
0.0
2.5
5.0
7.5

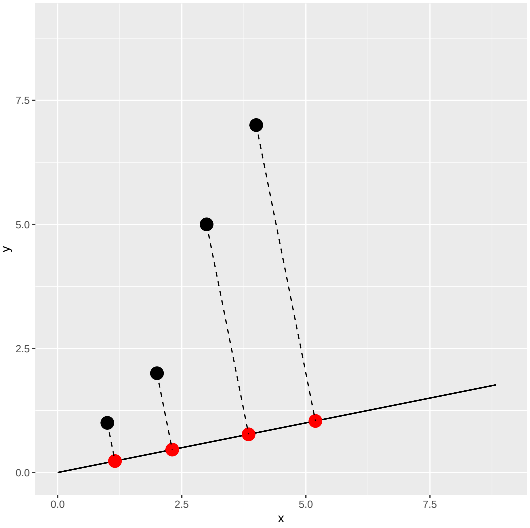
Variance: 2.42

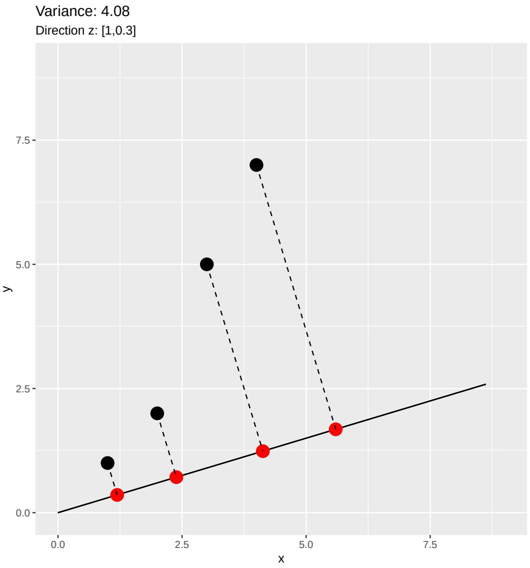
Direction z: [1,0.1]

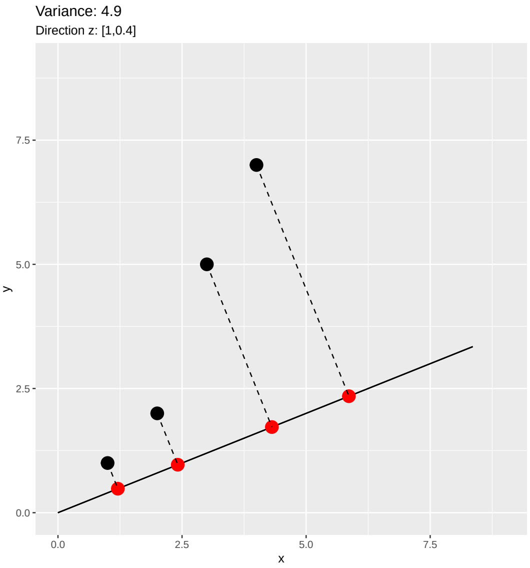


Variance: 3.24

Direction z: [1,0.2]

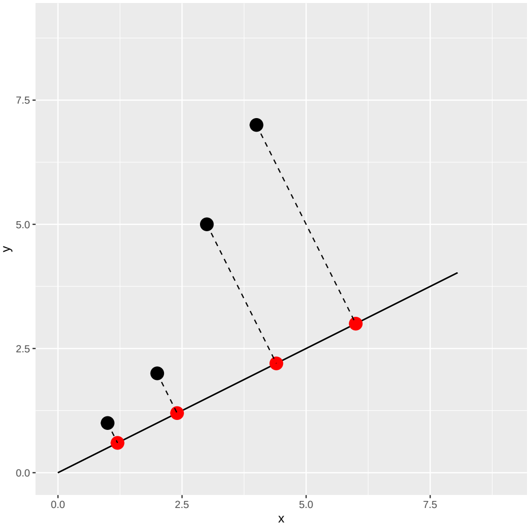


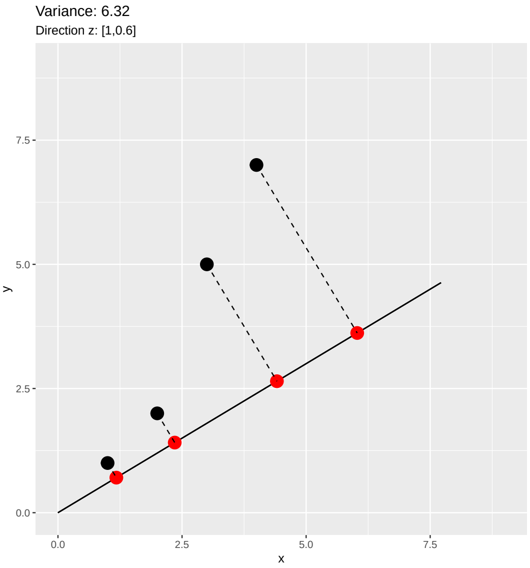

Variance: 4.08
Direction z: [1,0.3]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x



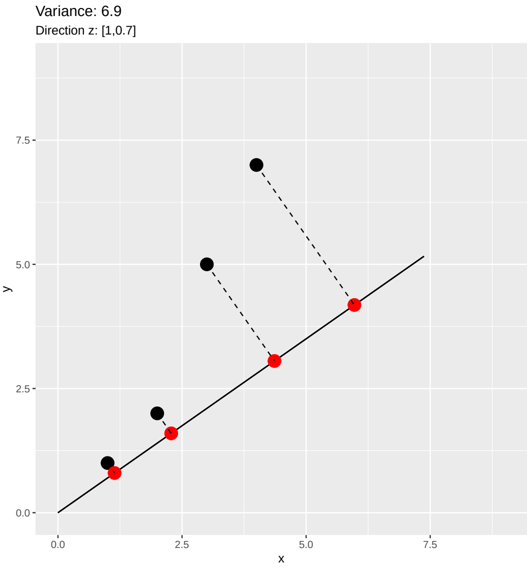
Variance: 4.9
Direction z: [1,0.4]
y
x
0.0
2.5
5.0
7.5
0.0
2.5
5.0
7.5

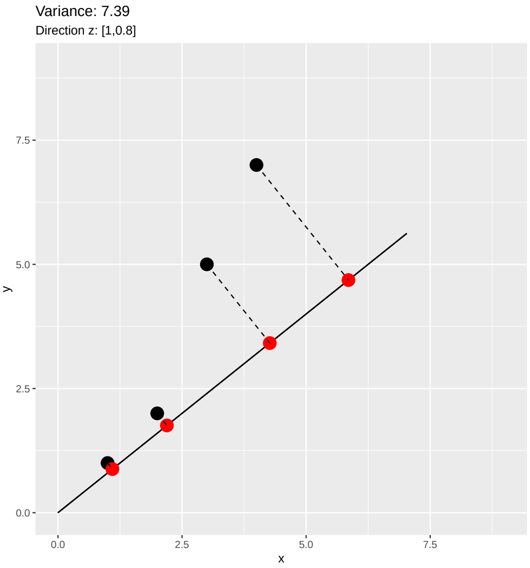
Variance: 5.65
Direction z: [1,0.5]

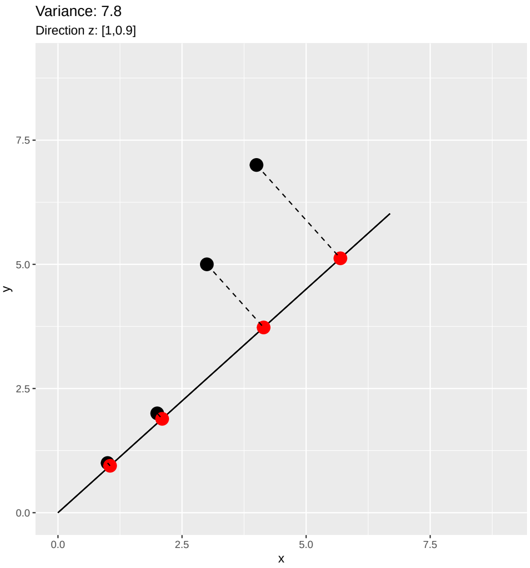


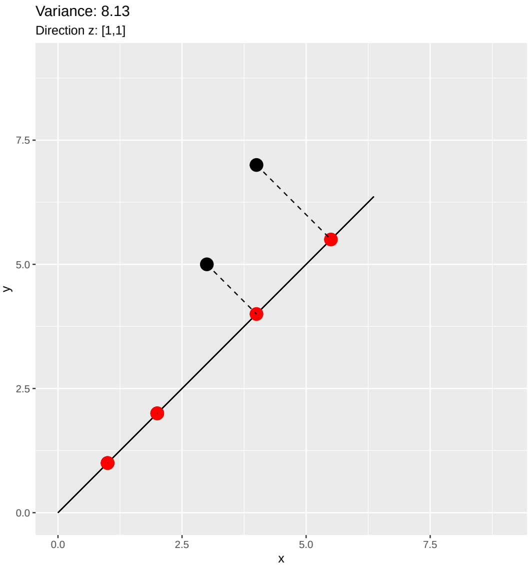


Variance: 6.32
Direction z: [1,0.6]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x


Variance: 6.9
Direction z: [1,0.7]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x


Variance: 7.39
Direction z: [1,0.8]
y
x
0.0
2.5
5.0
7.5
0.0
2.5
5.0
7.5

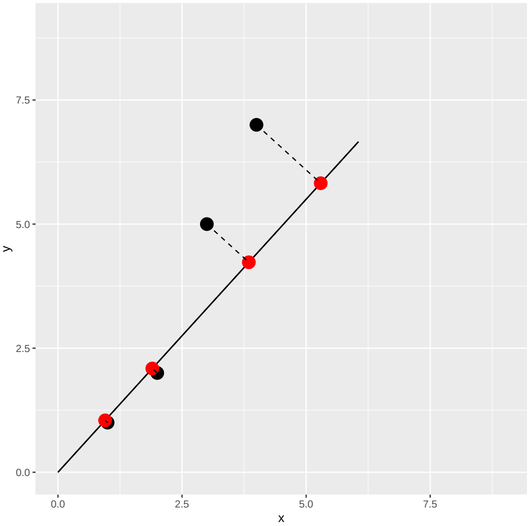

Variance: 7.8
Direction z: [1,0.9]
y
x
0.0
2.5
5.0
7.5
0.0
2.5
5.0
7.5

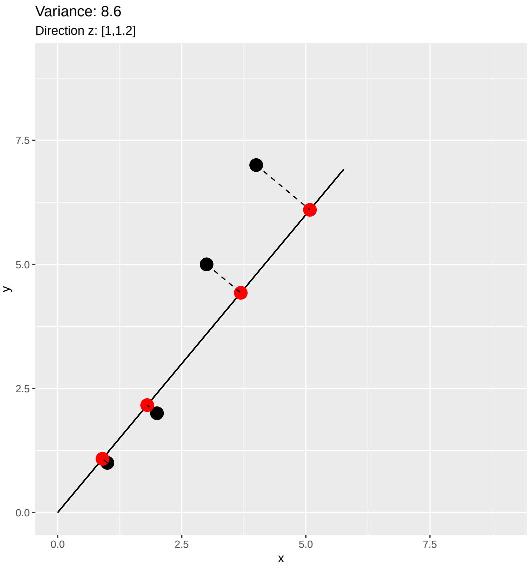


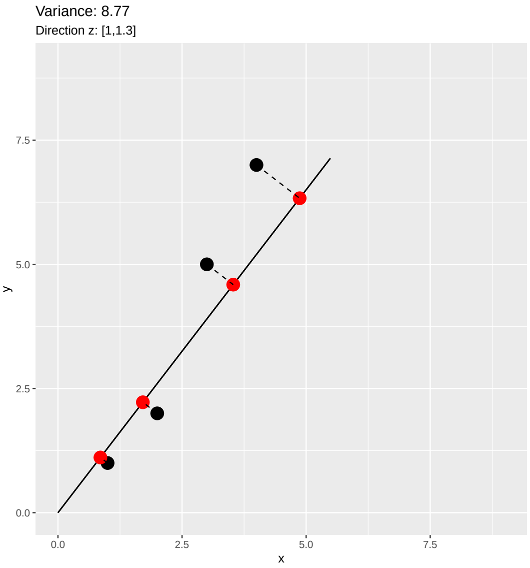
Variance: 8.13
Direction z: [1,1]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

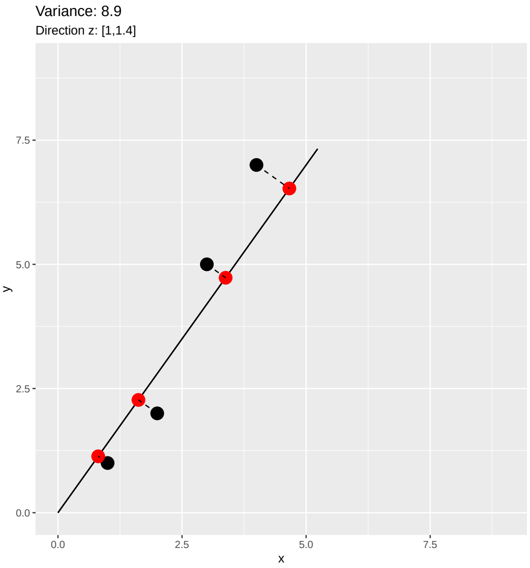
Variance: 8.39

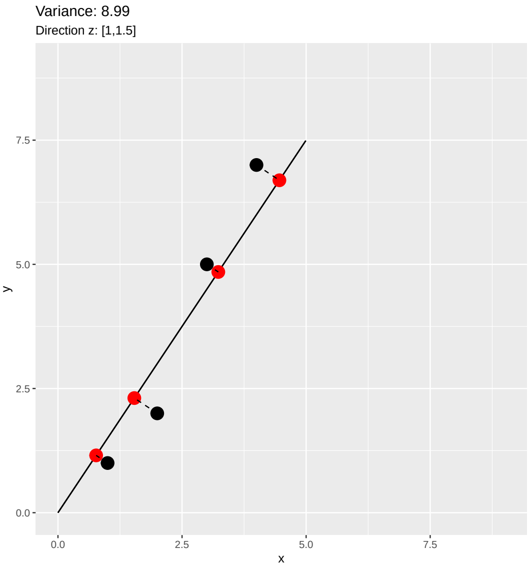
Direction z: [1,1.1]




Variance: 8.6
Direction z: [1,1.2]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

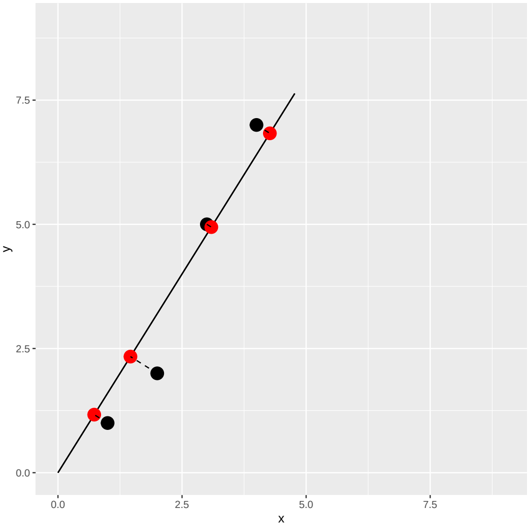

Variance: 8.77
Direction z: [1,1.3]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

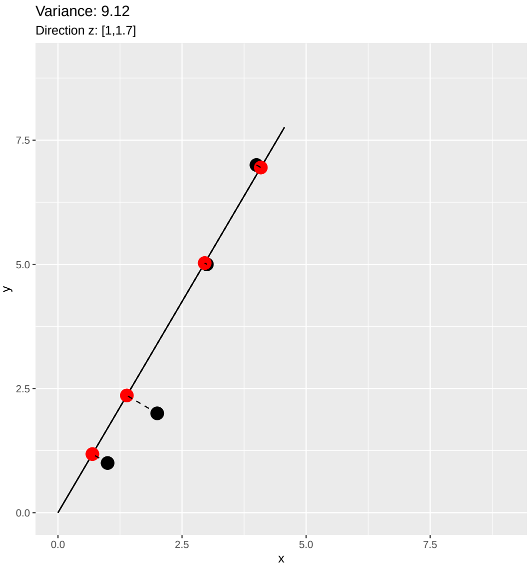

Variance: 8.9
Direction z: [1,1.4]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x


Variance: 8.99
Direction z: [1,1.5]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

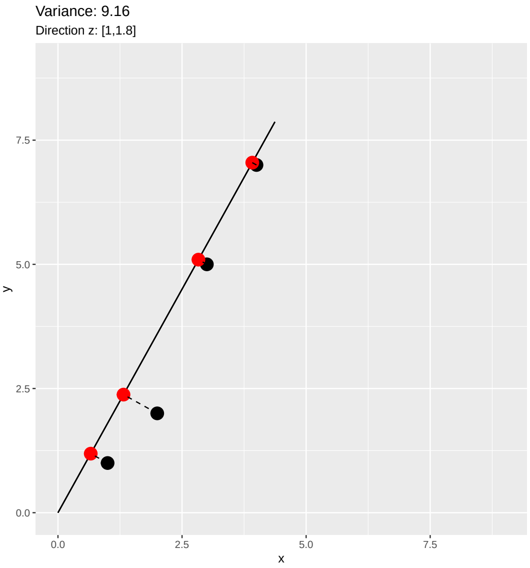
Variance: 9.07

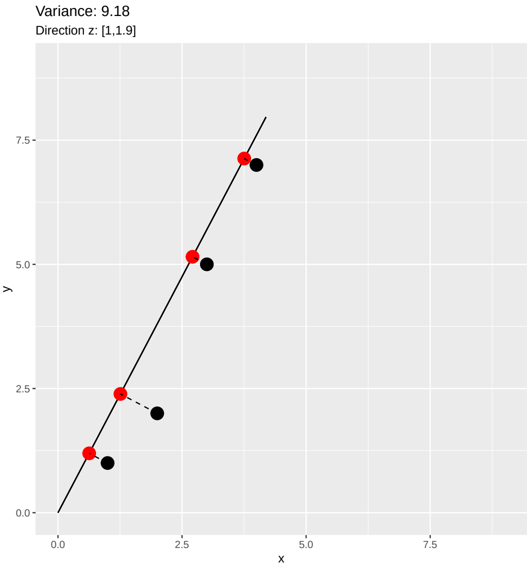
Direction z: [1,1.6]



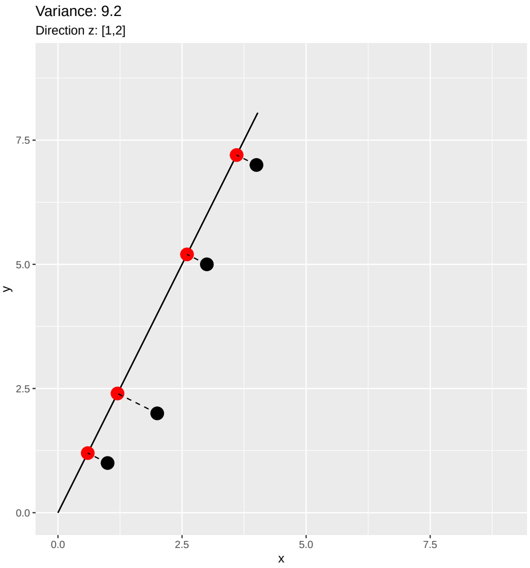


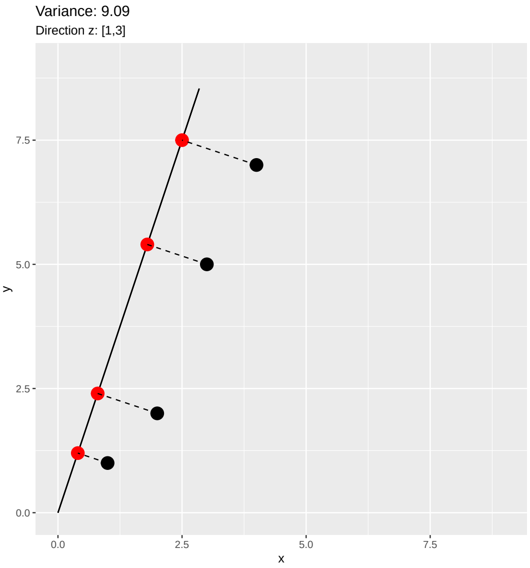
Variance: 9.12
Direction z: [1,1.7]
y
x
0.0
2.5
5.0
7.5
0.0
2.5
5.0
7.5

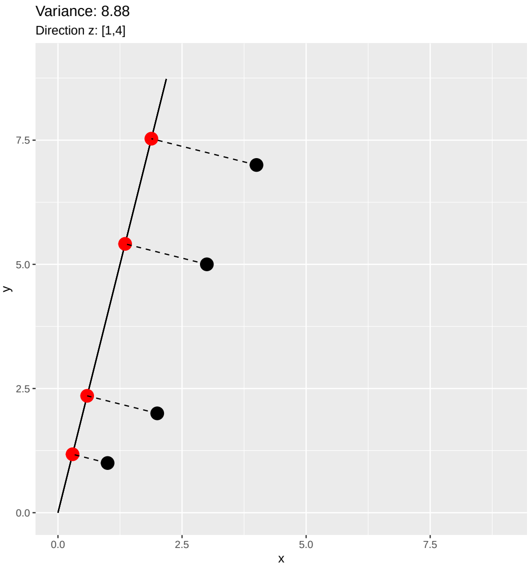

Variance: 9.16
Direction z: [1,1.8]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

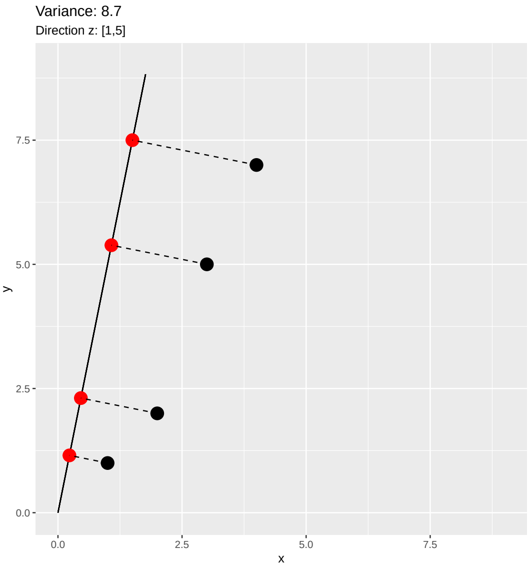


Variance: 9.18
Direction z: [1,1.9]
y
x
0.0
2.5
5.0
7.5
0.0
2.5
5.0
7.5

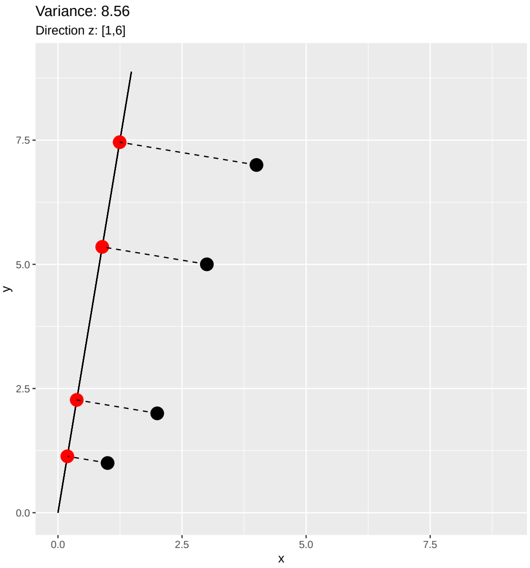

Variance: 9.2
Direction z: [1,2]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

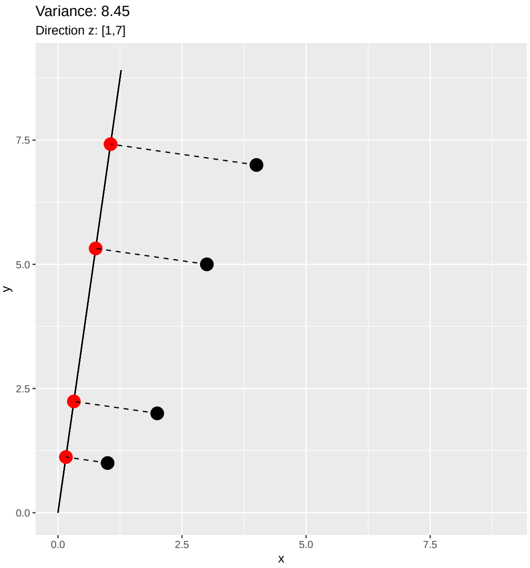

Variance: 9.09
Direction z: [1,3]
y
x
0.0
2.5
5.0
7.5
0.0
2.5
5.0
7.5

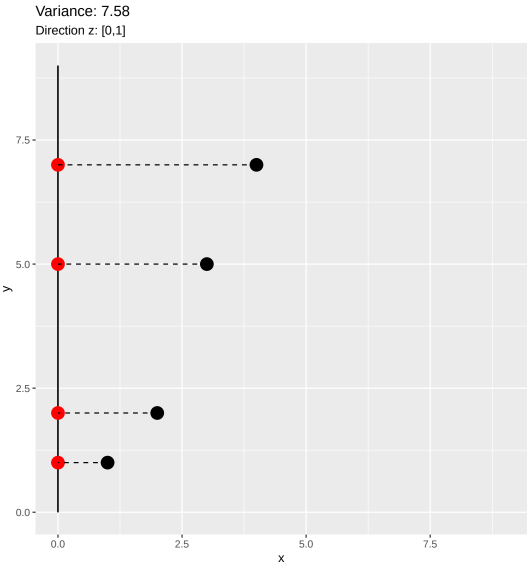

Variance: 8.88
Direction z: [1,4]
y
x
0.0
2.5
5.0
7.5
0.0
2.5
5.0
7.5



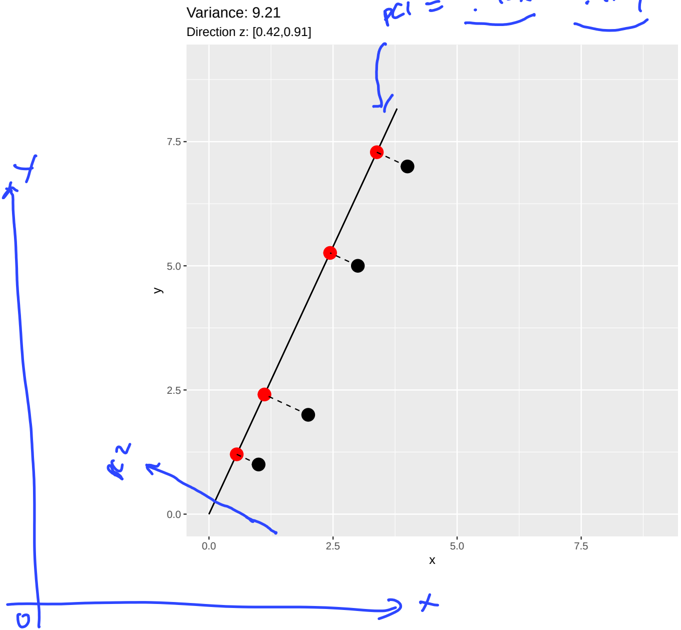
Variance: 8.7
Direction z: [1,5]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x


Variance: 8.56
Direction z: [1,6]
y
x
0.0
2.5
5.0
7.5
0.0
2.5
5.0
7.5


Variance: 8.45
Direction z: [1,7]
y
x
0.0
2.5
5.0
7.5

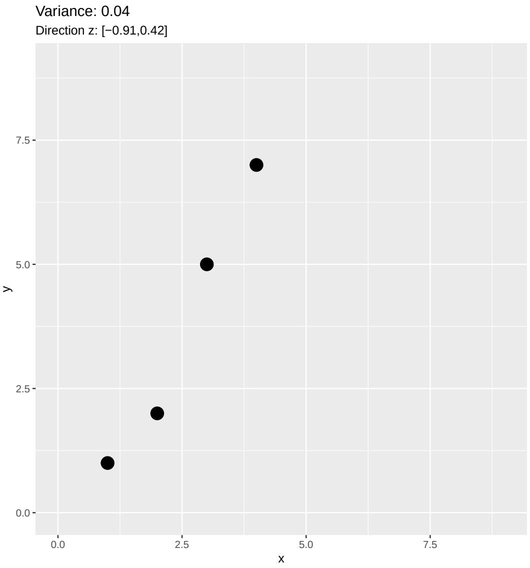

Variance: 7.58
Direction z: [0,1]
y
x
0.0
2.5
5.0
7.5
0.0
2.5
5.0
7.5

$pc1 = \underbrace{.42}x + \underbrace{.91}y$



9.21 out of 9.25 variation/information

what about .05 variation left


Variance: 0.04
Direction z: [−0.91,0.42]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

Rotation Matrix or PC Loading

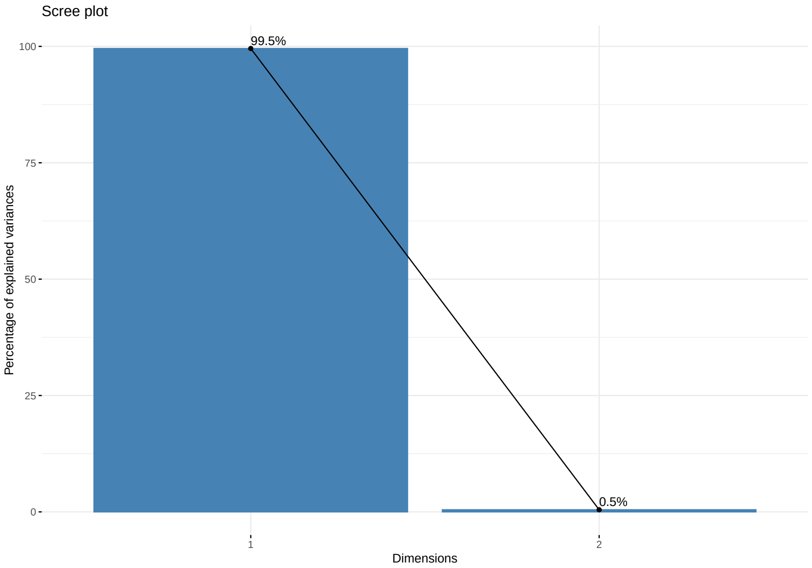
- $\Phi =$

```
   PC1   PC2
x 0.42 -0.91
y 0.91  0.42
```

PC Scores

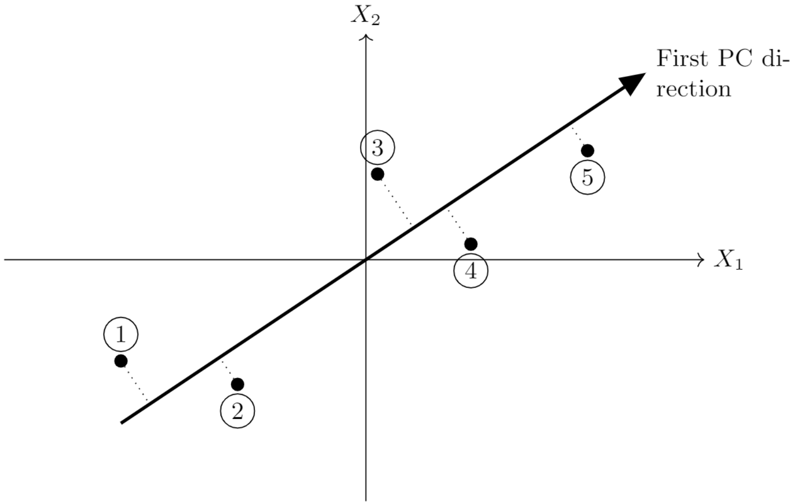
- $Z = X \cdot \Phi =$

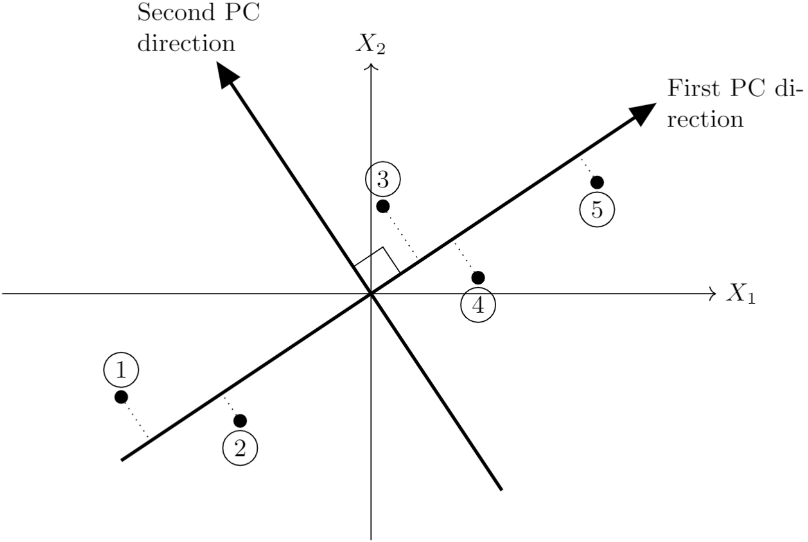
```
     PC1   PC2
[1,] 1.33 -0.49
[2,] 2.66 -0.97
[3,] 5.80 -0.62
[4,] 8.03 -0.68
```


Scree plot
99.5%
0.5%
100
75
50
25
0
Percentage of explained variances
1
2
Dimensions

- What direction maximizes the variance?

- What direction maximizes the variance?
- The first principal component


X_2
X_1
First PC direction
1
2
3
4
5


Second PC
direction
X_2
First PC di-
rection
X_1
1
2
3
4
5

Formula

- Write down matrix form of the example

$X \to X \cdot \phi = Z$

- ϕ is PC loading
- z is PC scores

In general

$$\begin{array}{ccc}
\begin{array}{c}\textbf{Original data matrix}\\ \textbf{(fat matrix!)}\end{array} & & \begin{array}{c}\textbf{New data matrix}\\ \textbf{(thin matrix!)}\end{array} \\
\begin{array}{ccccc}\underline{X_1} & \underline{X_2} & \cdots & \cdots & \underline{X_p}\end{array} & & \begin{array}{ccc}\underline{Z_1} & \cdots & \underline{Z_M}\end{array} \\
\begin{pmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & \cdots & x_{np} \end{pmatrix} & \xrightarrow{\text{compressed}} & \begin{pmatrix} z_{11} & \cdots & z_{1M} \\ z_{21} & \cdots & z_{2M} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nM} \end{pmatrix} \\
\mathbf{X} & & \mathbf{X}\boldsymbol{\phi}_1 \;\cdots\; \mathbf{X}\boldsymbol{\phi}_M
\end{array}$$

Example

	Independent variables	
Observation	X_1	X_2
1	-2	2
2	2	-2

- The data set consists of only these two observations.
- The first principal component loading for X_1, ϕ_{11}, is 0.7071.
- The first principal component loading for X_2, ϕ_{21}, is negative.

Calculate the first principal component score for Observation 1.

PC Loadings

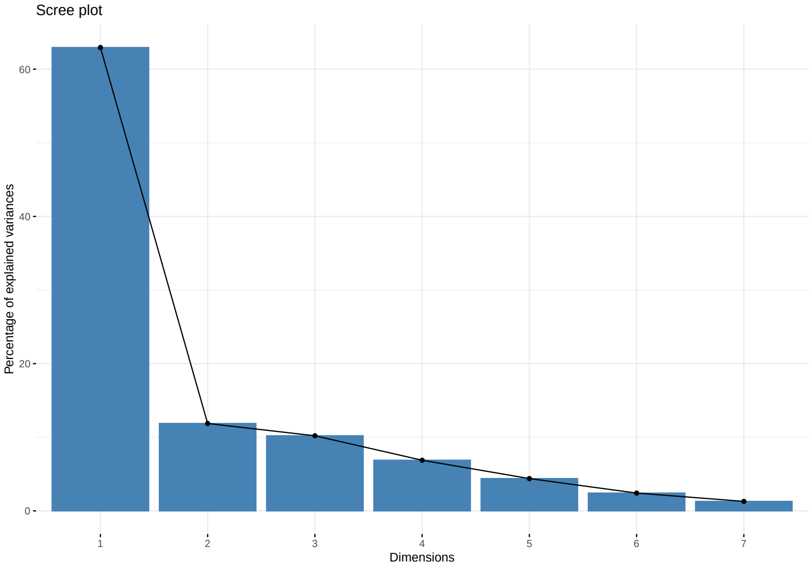
	First PC	Second PC
Murder	0.5359	−0.4182
Assault	0.5832	−0.1880
UrbanPop	0.2782	0.8728
Rape	0.5434	0.1673

How many PC should we use?

- Performance during two sporting events

X100m	Long.jump	Shot.put	High.jump	X400m	X110m.hurdle	Discus
11.04	7.58	14.83	2.07	49.81	14.69	43.75
10.76	7.40	14.26	1.86	49.37	14.05	50.72
11.02	7.23	14.25	1.92	48.93	14.99	40.87
11.34	7.09	15.19	2.10	50.42	15.31	46.26
11.13	7.30	13.48	2.01	48.62	14.17	45.67
10.83	7.31	13.76	2.13	49.91	14.38	44.41

Scree Plot



Scree plot
Percentage of explained variances
60
40
20
0
1
2
3
4
5
6
7
Dimensions