Principal Component Analysis

* 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)

Data 2

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

variable selection

x_1	x_3

$(d = 2)$

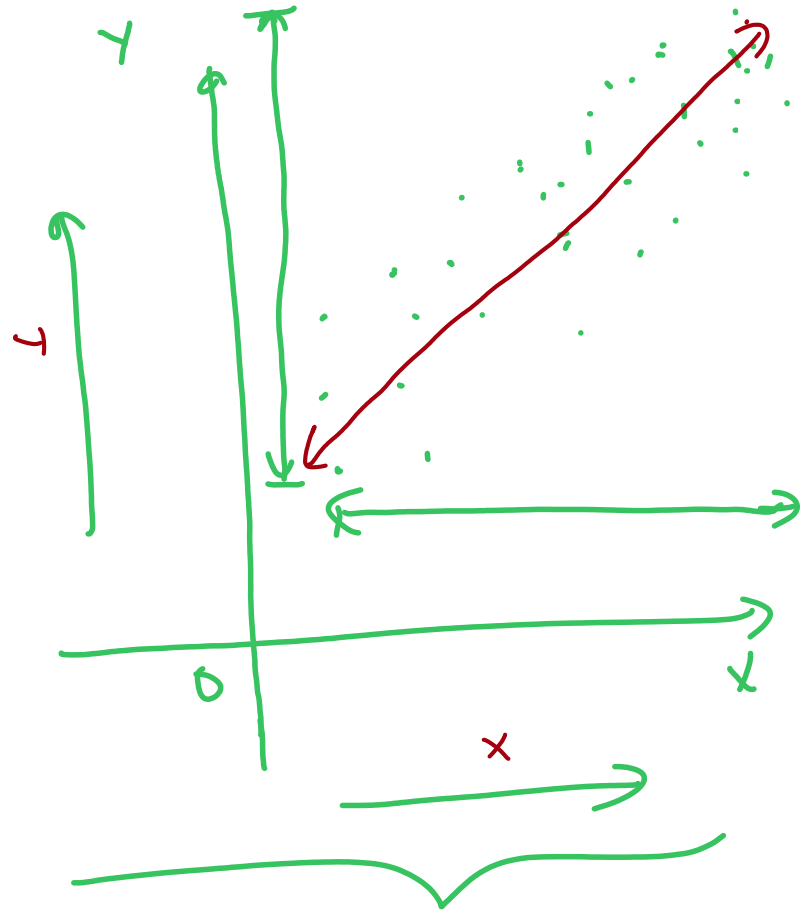
Example:

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

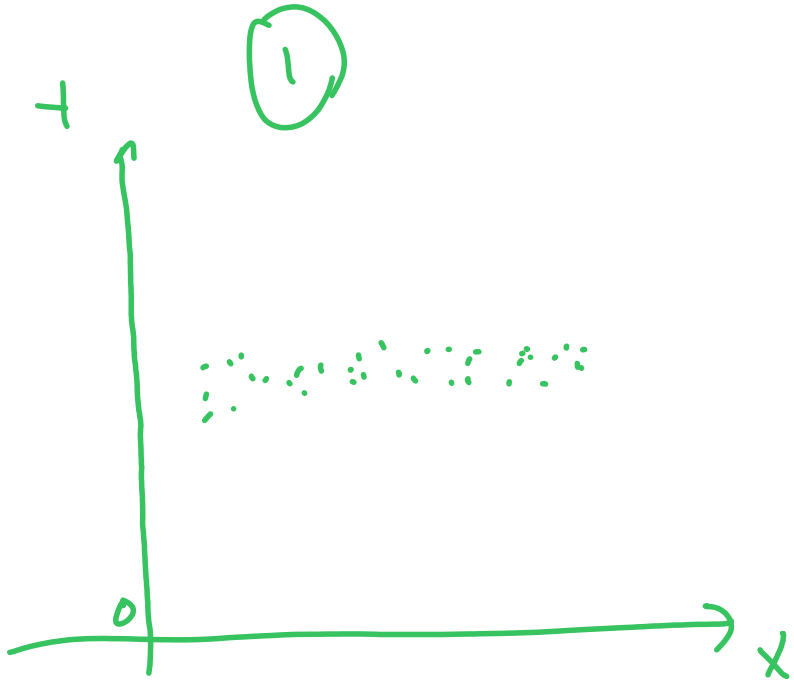
PCA in a view or coordinate rotation

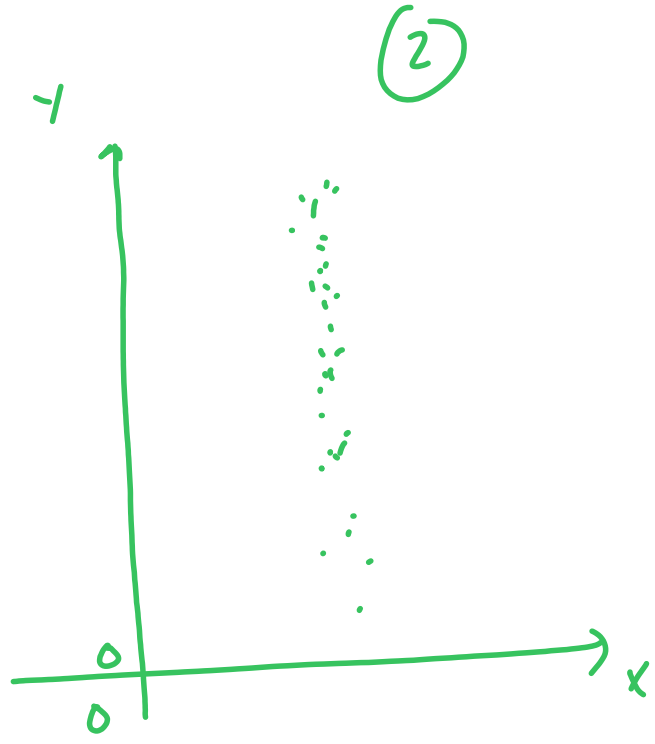
Variance of the Projection

x	y
1	1
2	2
3	5
4	7



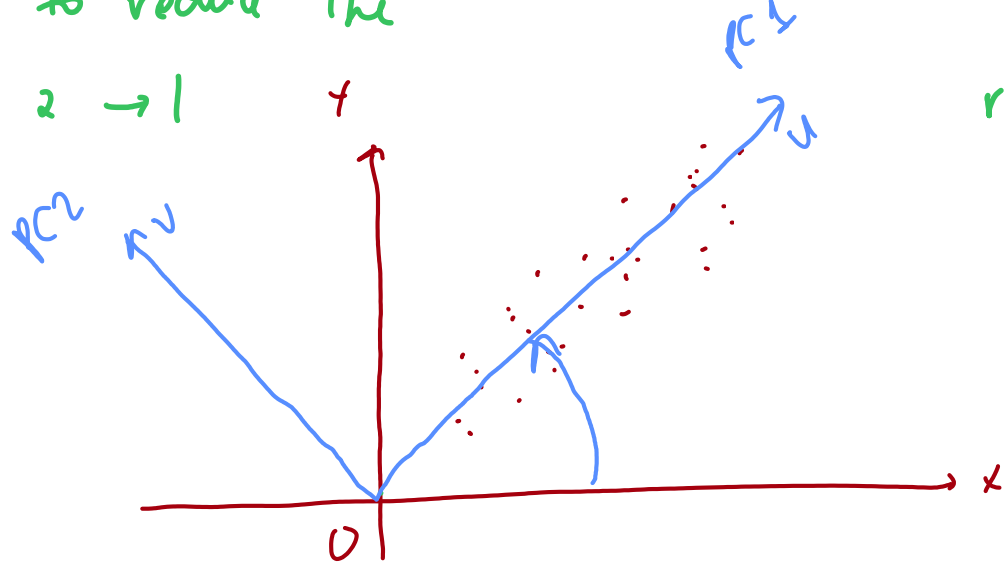
- V(x) = 1.67
- V(y) = 7.58
- Total variance: V(x) + V(y) = 9.25





remove y to reduce the dimension $2 \to 1$

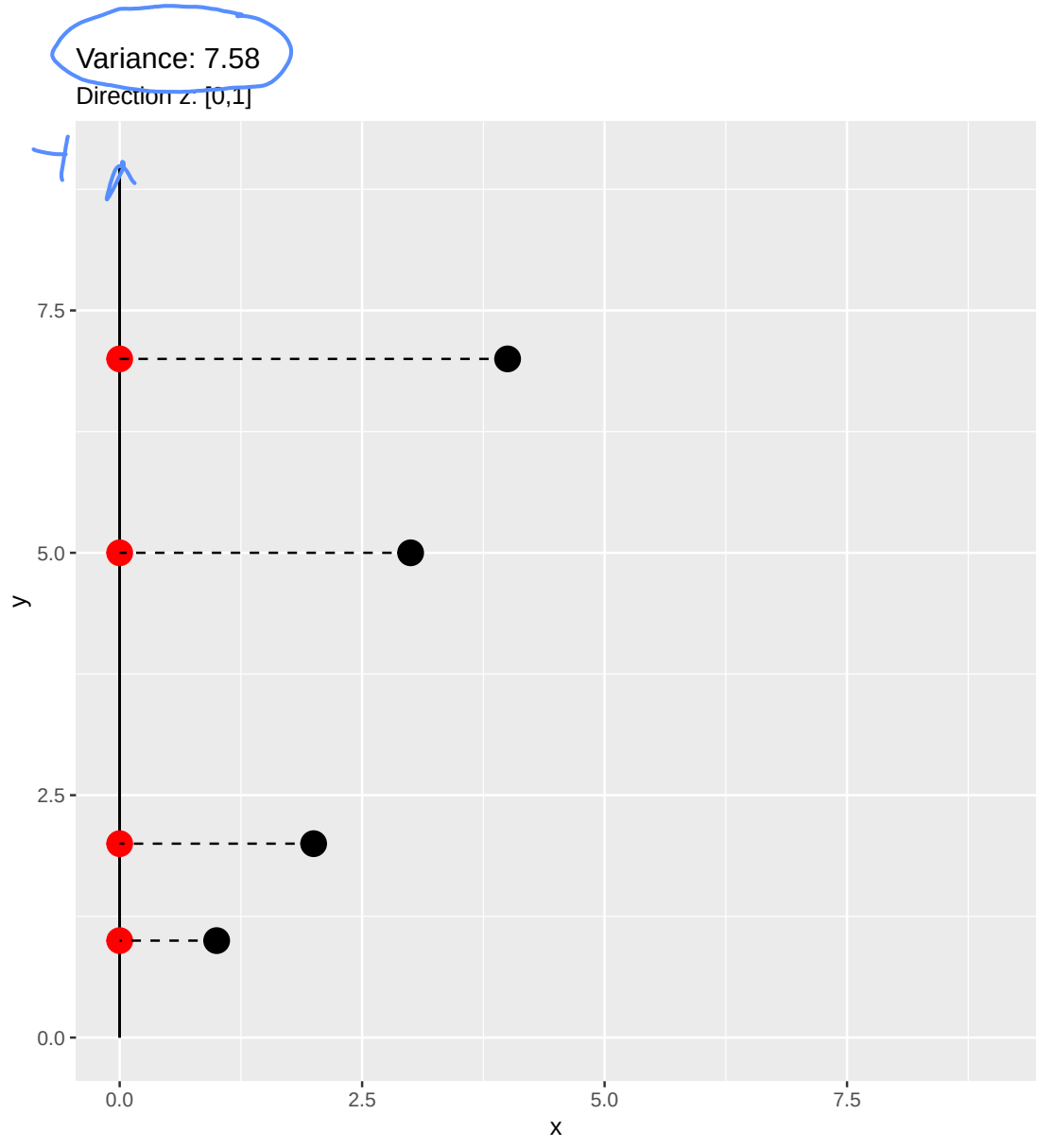
remove x to reduce d : $2 \to 1$

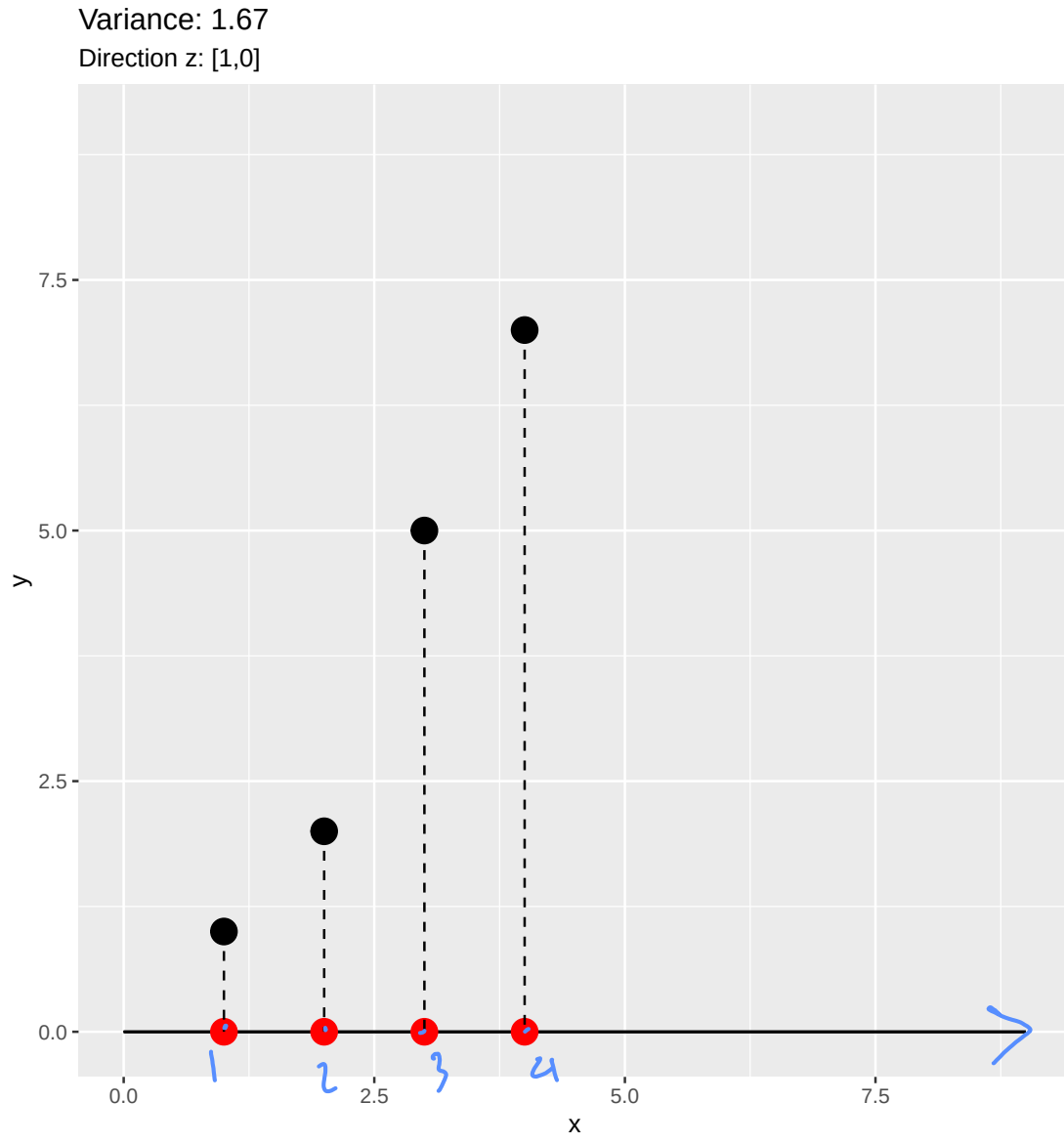


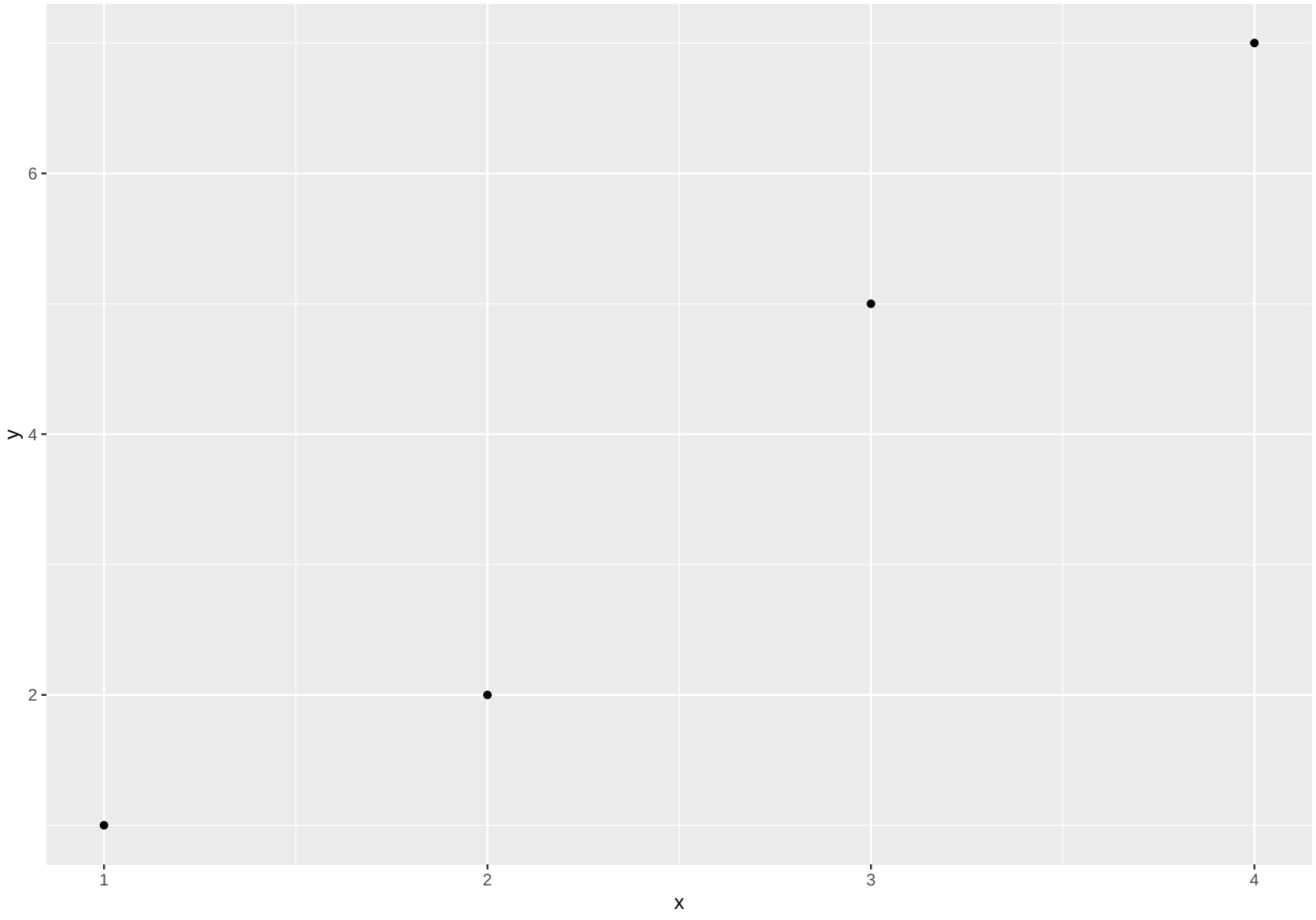
XOY → UOV

↓

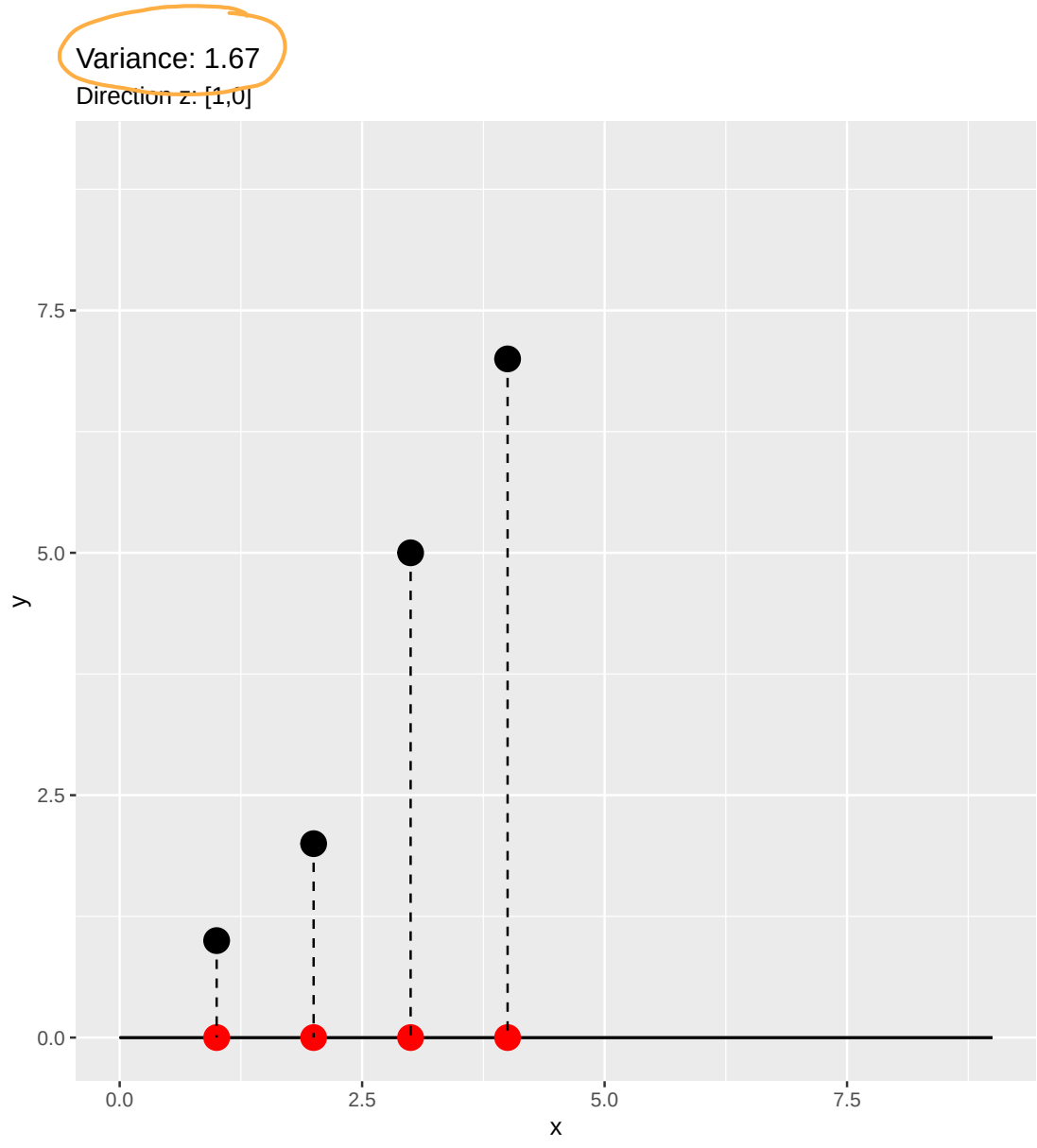
then remove V to reduce d: $2 \to 1$

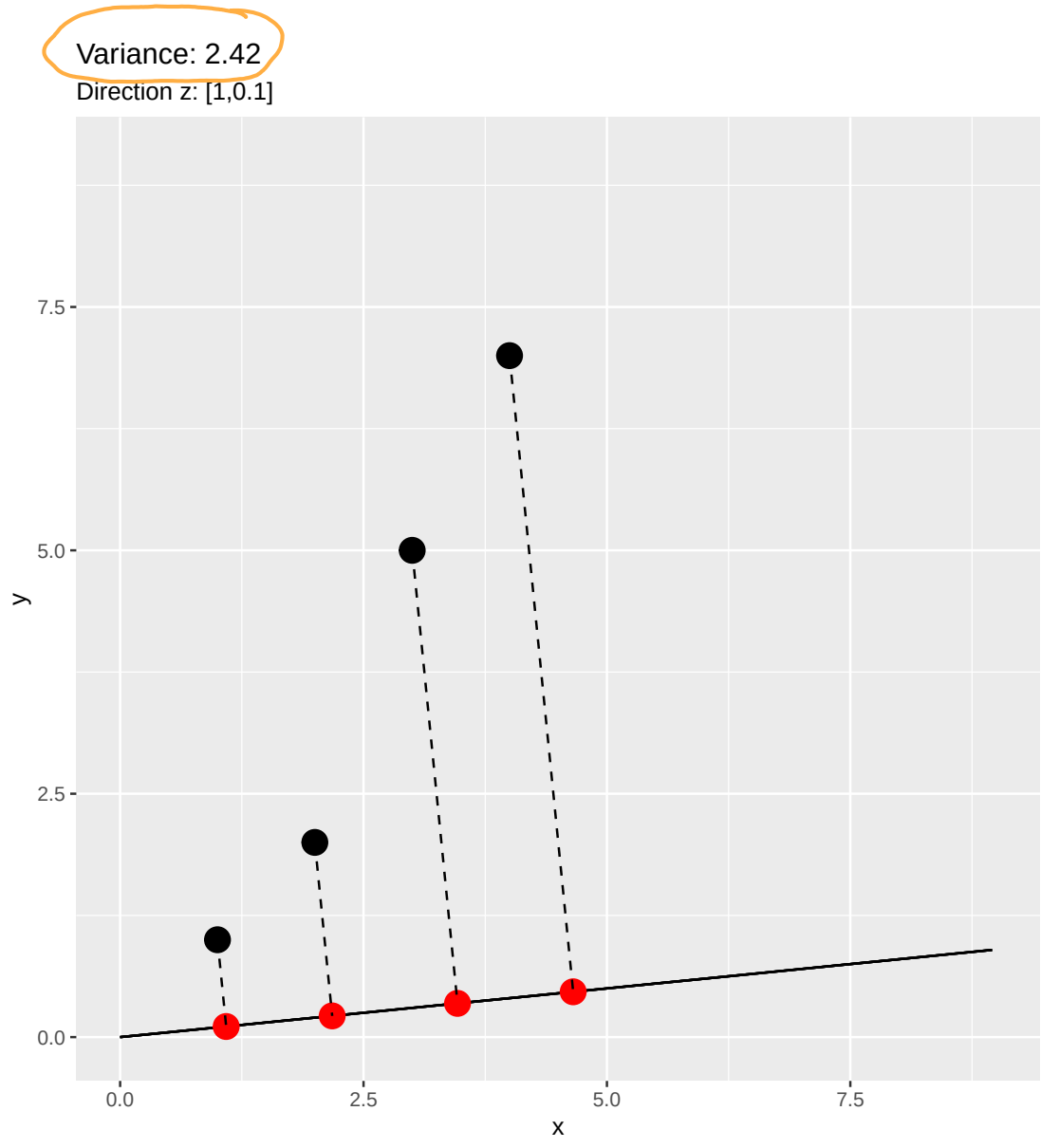

Variance: 7.58
Direction z: [0,1]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

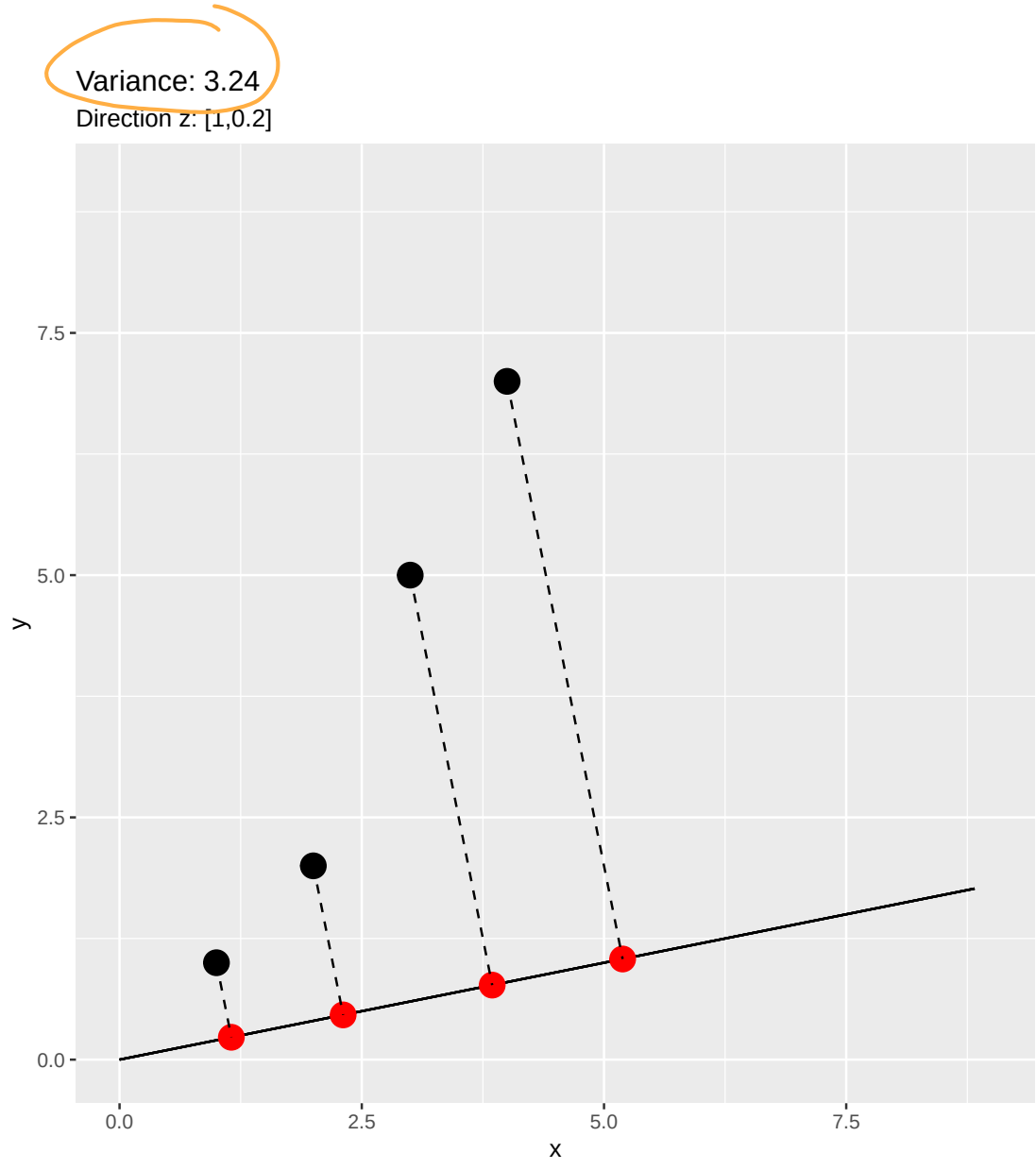

Variance: 1.67
Direction z: [1,0]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x
1
2
3
4


6
4
2
y
1
2
3
4
x

```
[1] 9.25
```

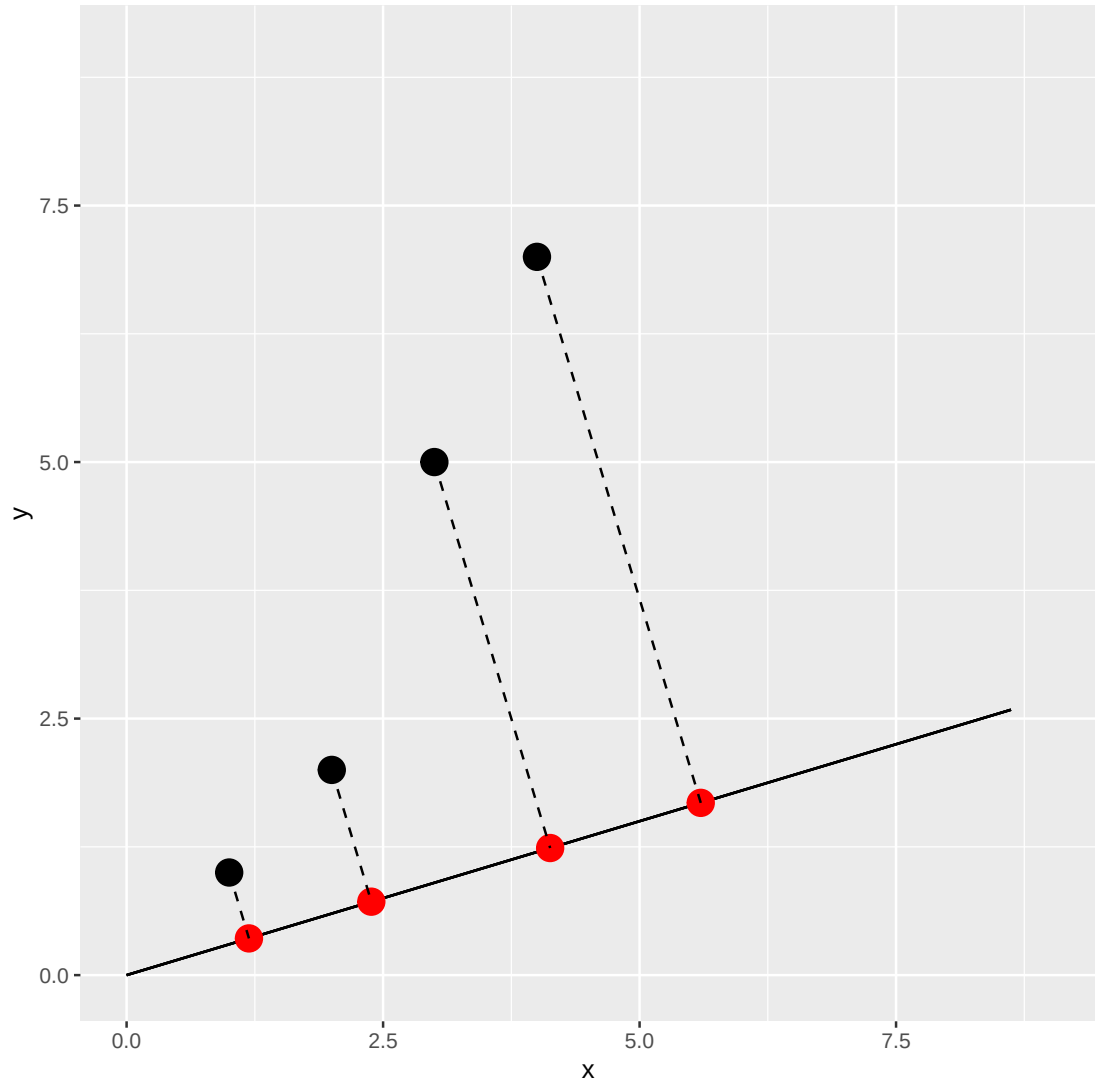

Variance: 1.67
Direction z: [1,0]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

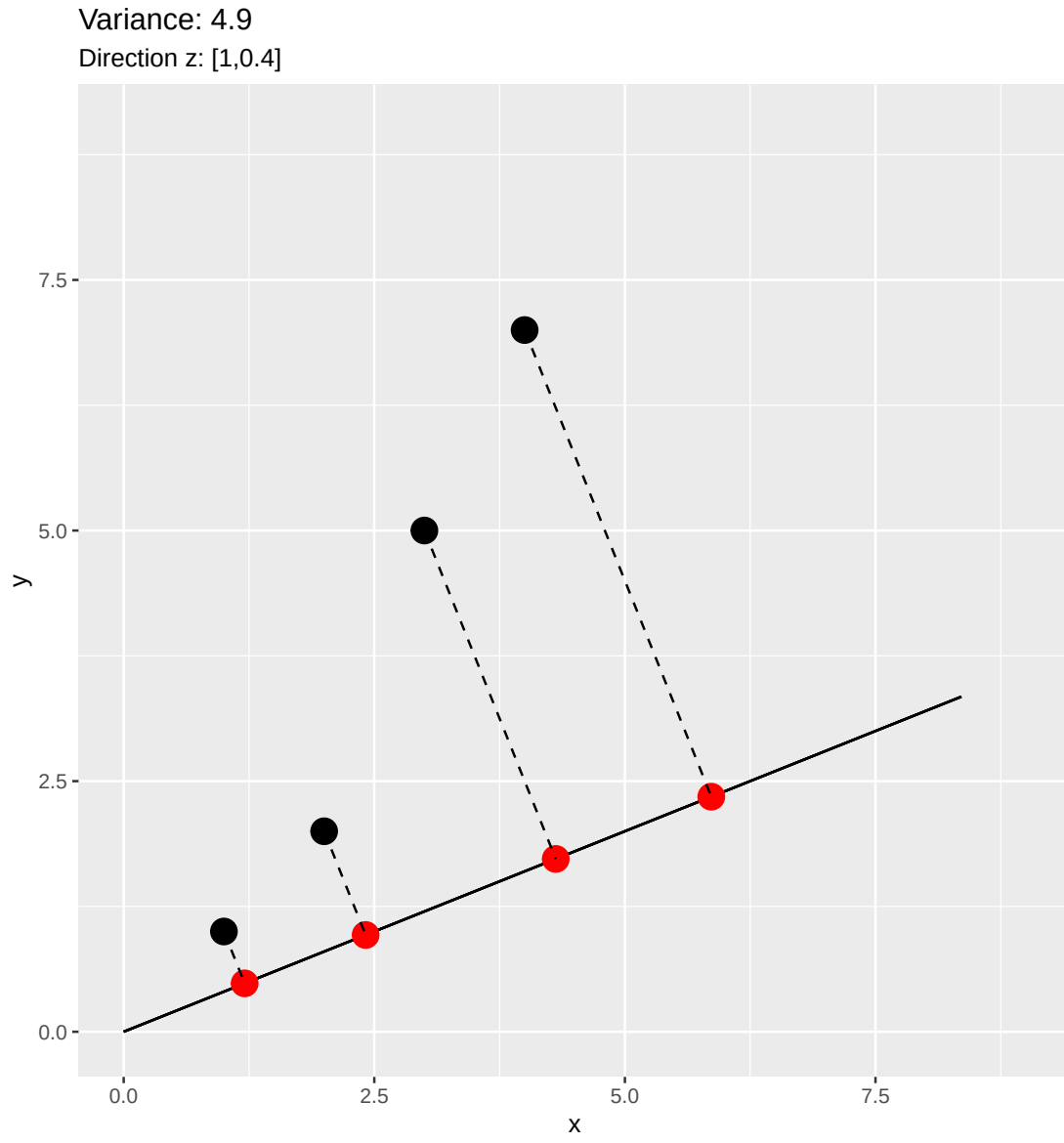

Variance: 2.42
Direction z: [1,0.1]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

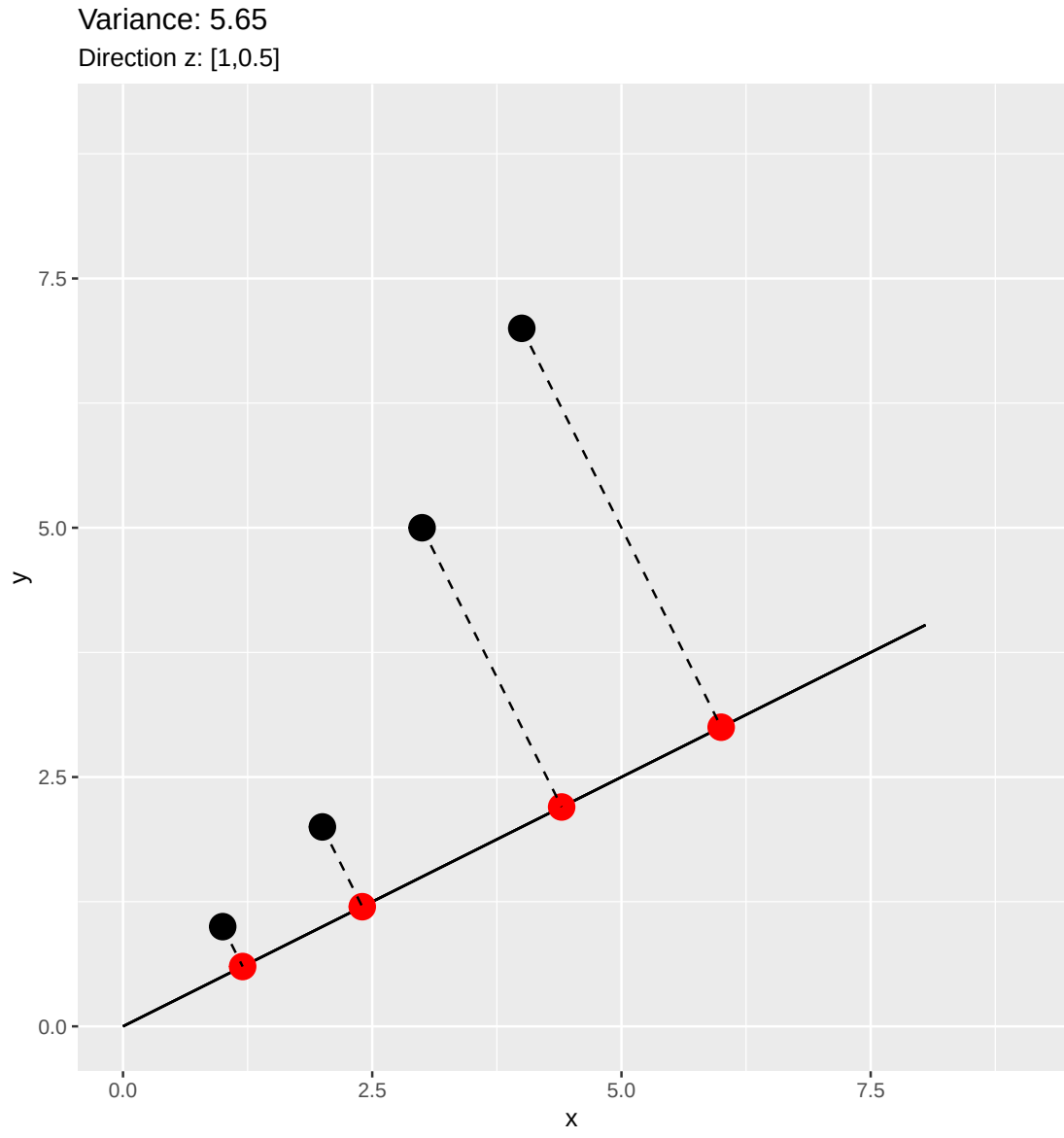

Variance: 3.24
Direction z: [1,0.2]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

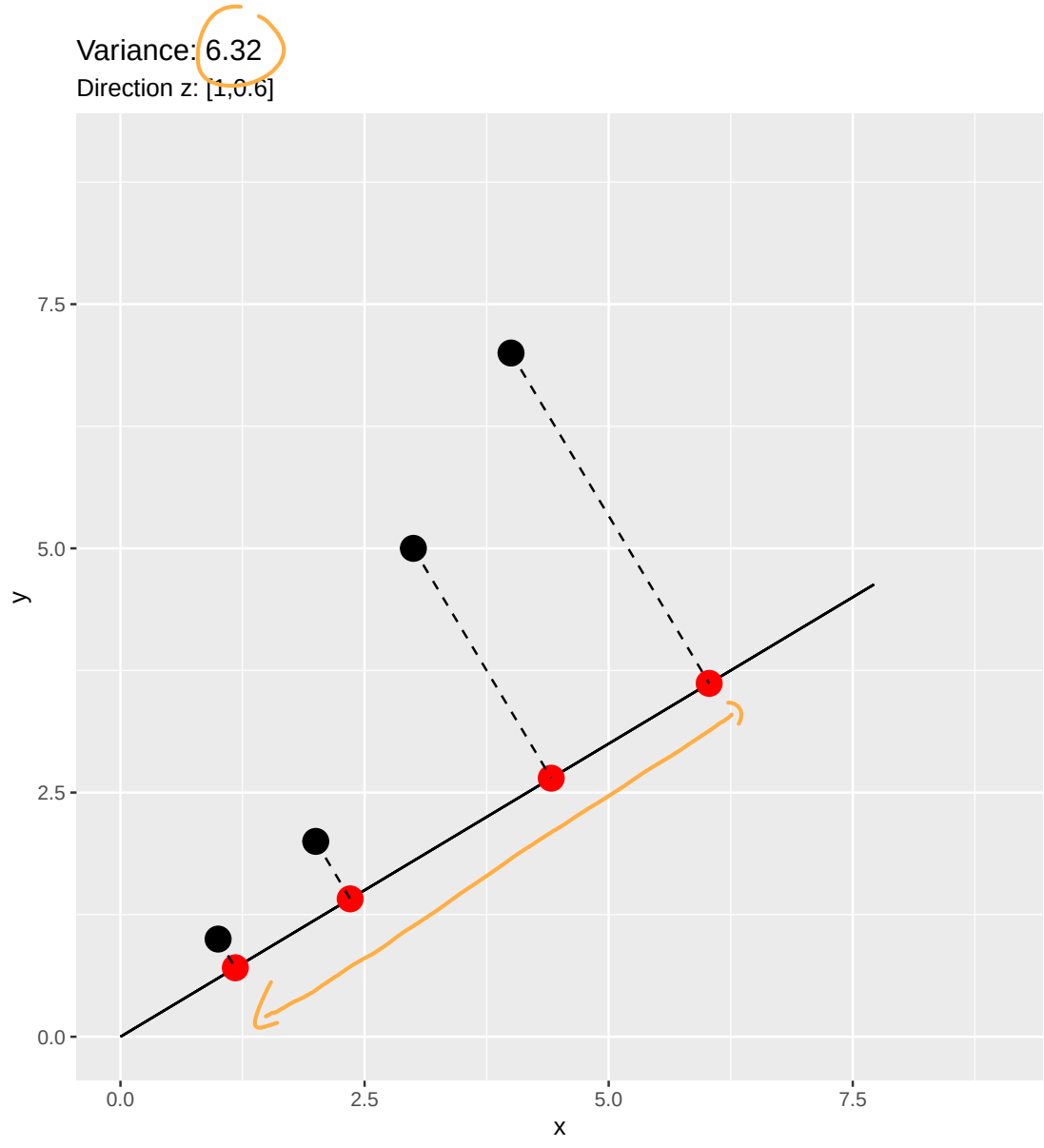
Variance: 4.08

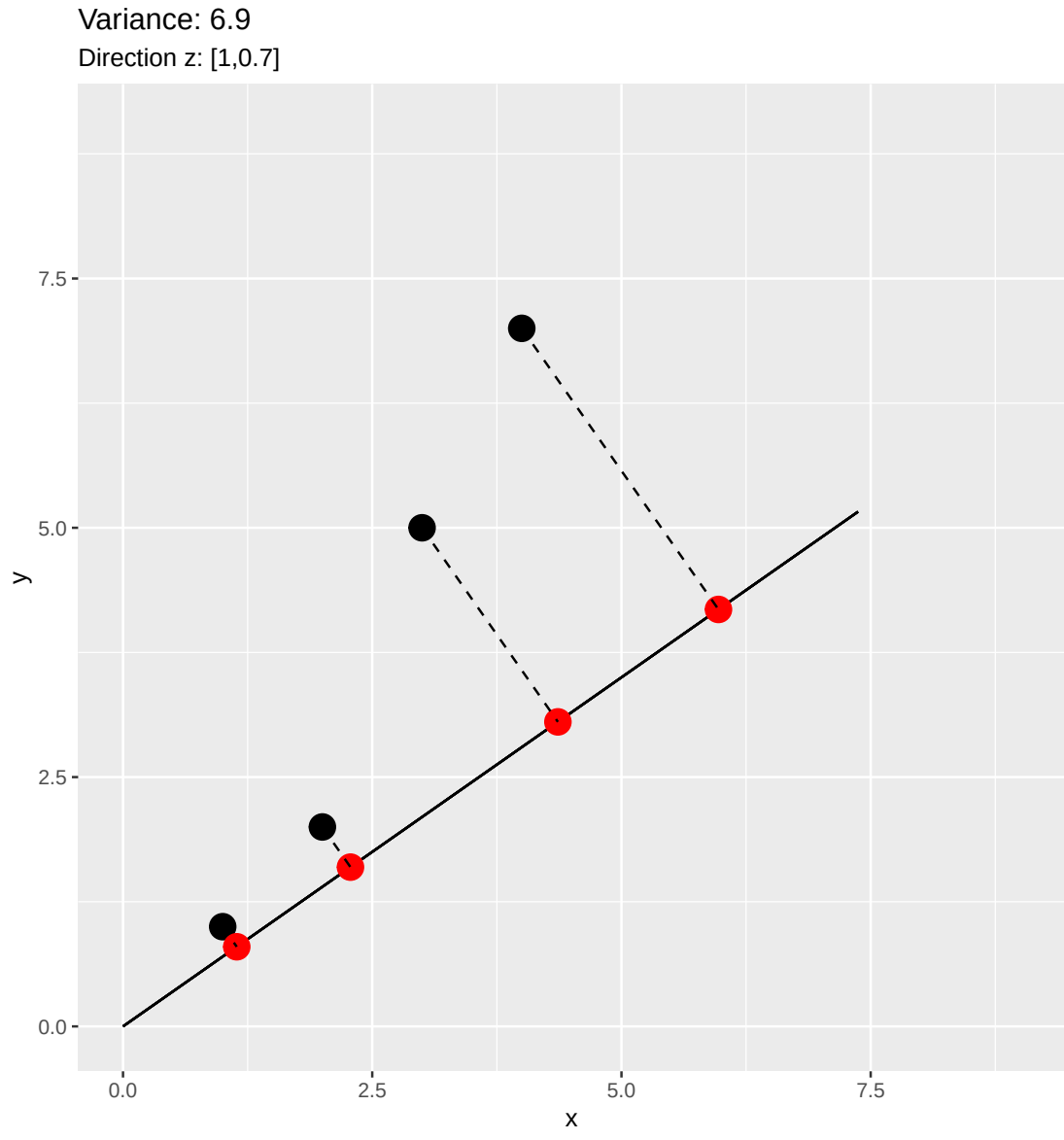
Direction z: [1,0.3]

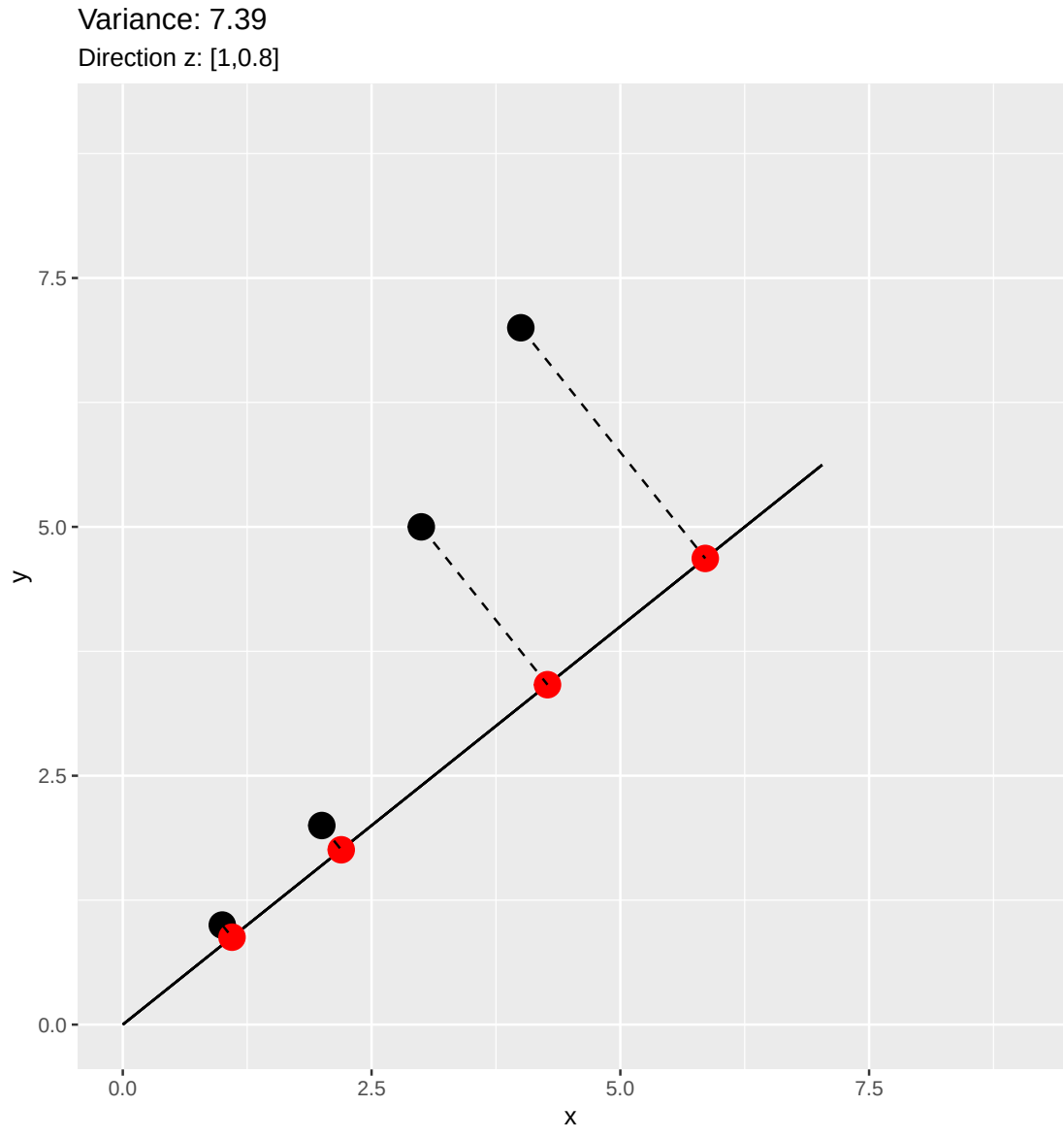


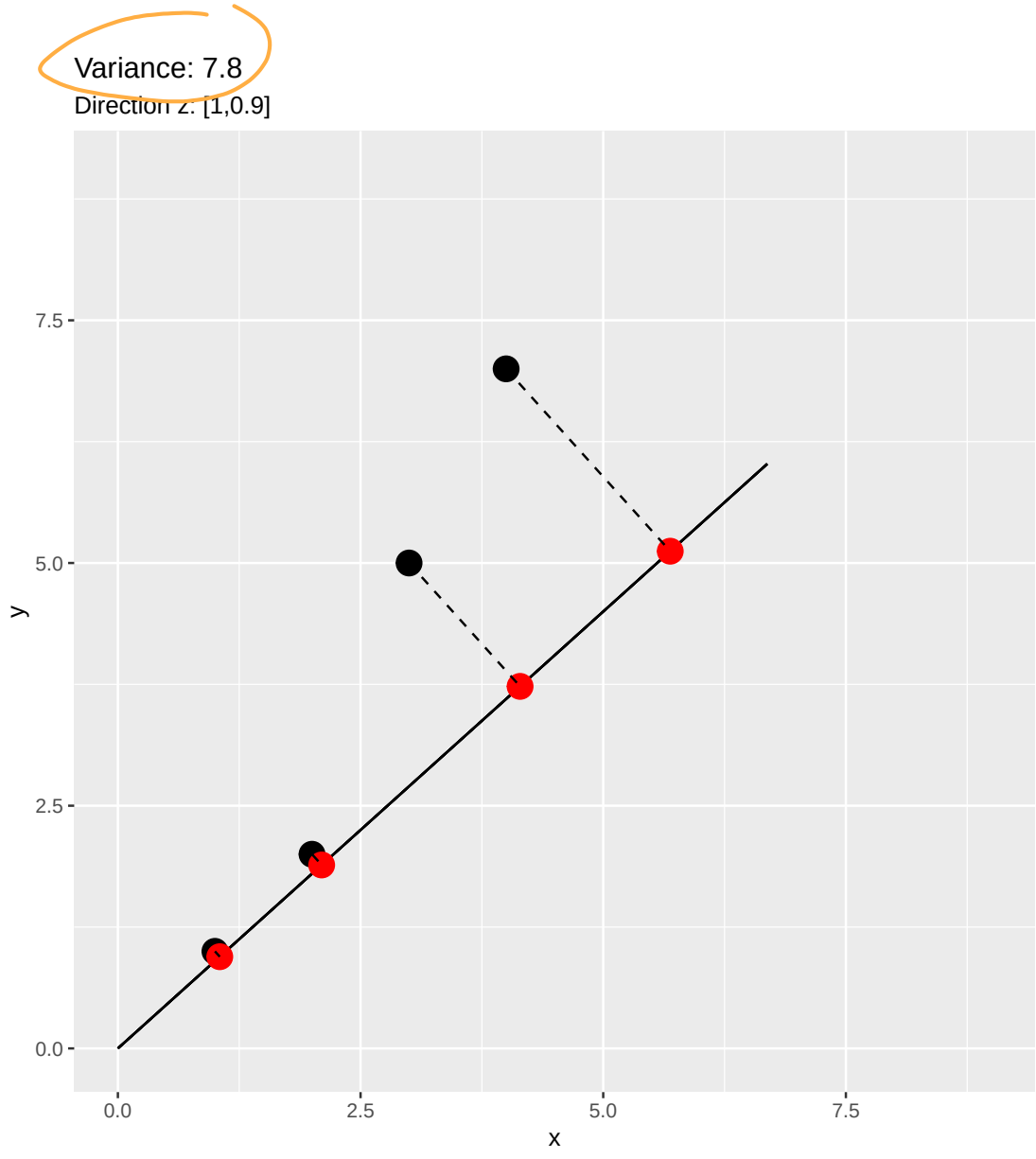

Variance: 4.9
Direction z: [1,0.4]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

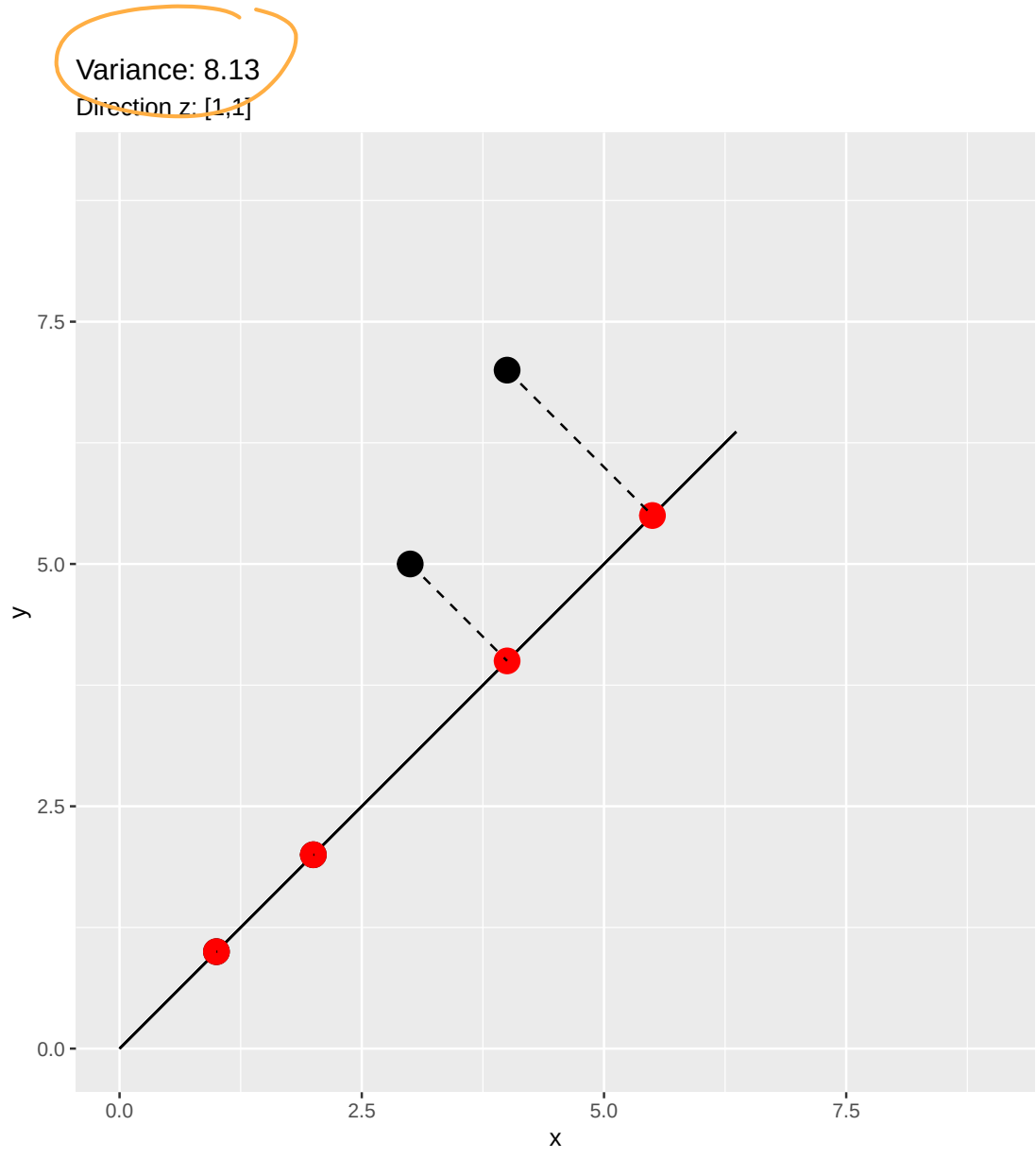

Variance: 5.65
Direction z: [1,0.5]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

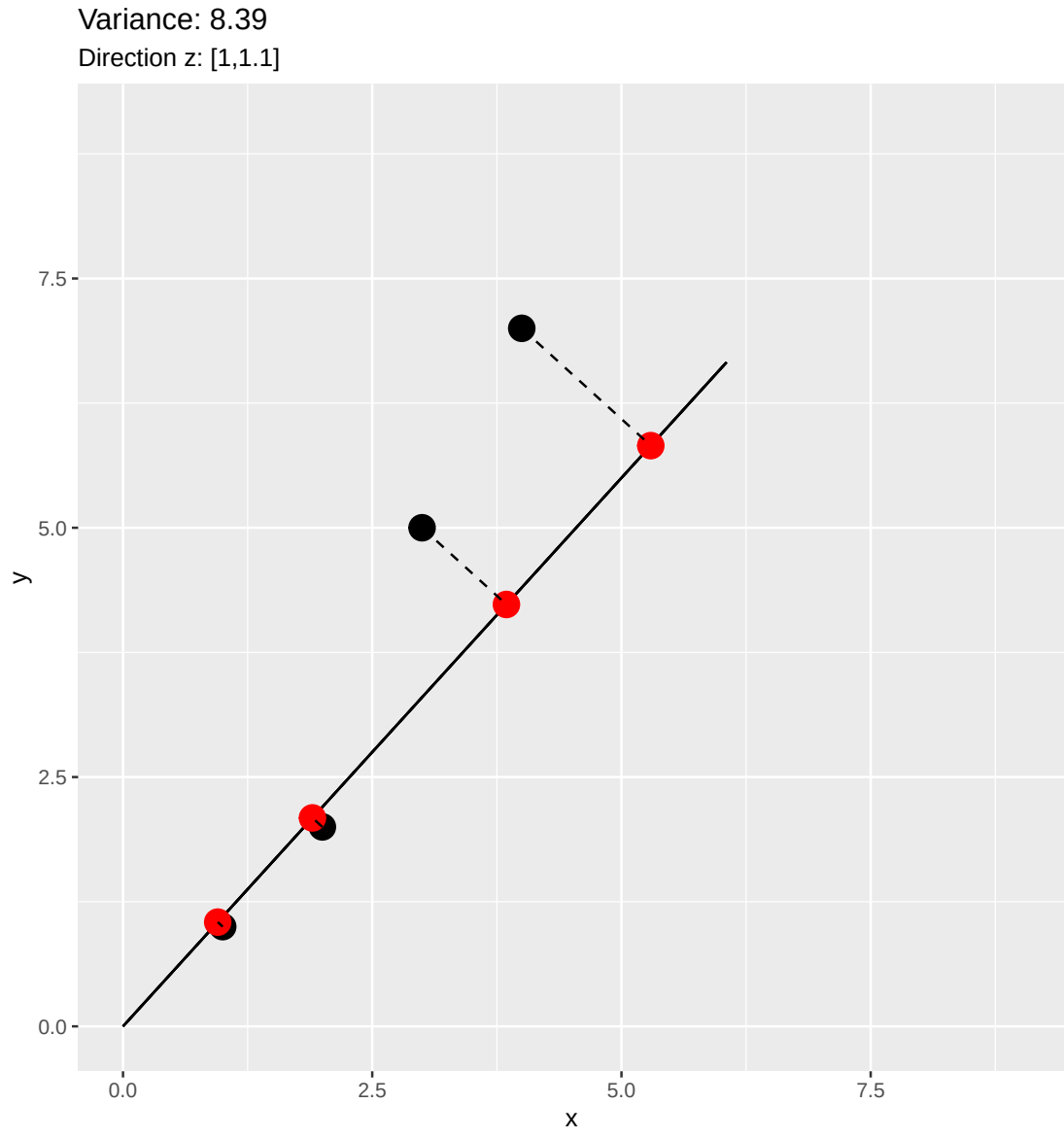

Variance: 6.32
Direction z: [1,0.6]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

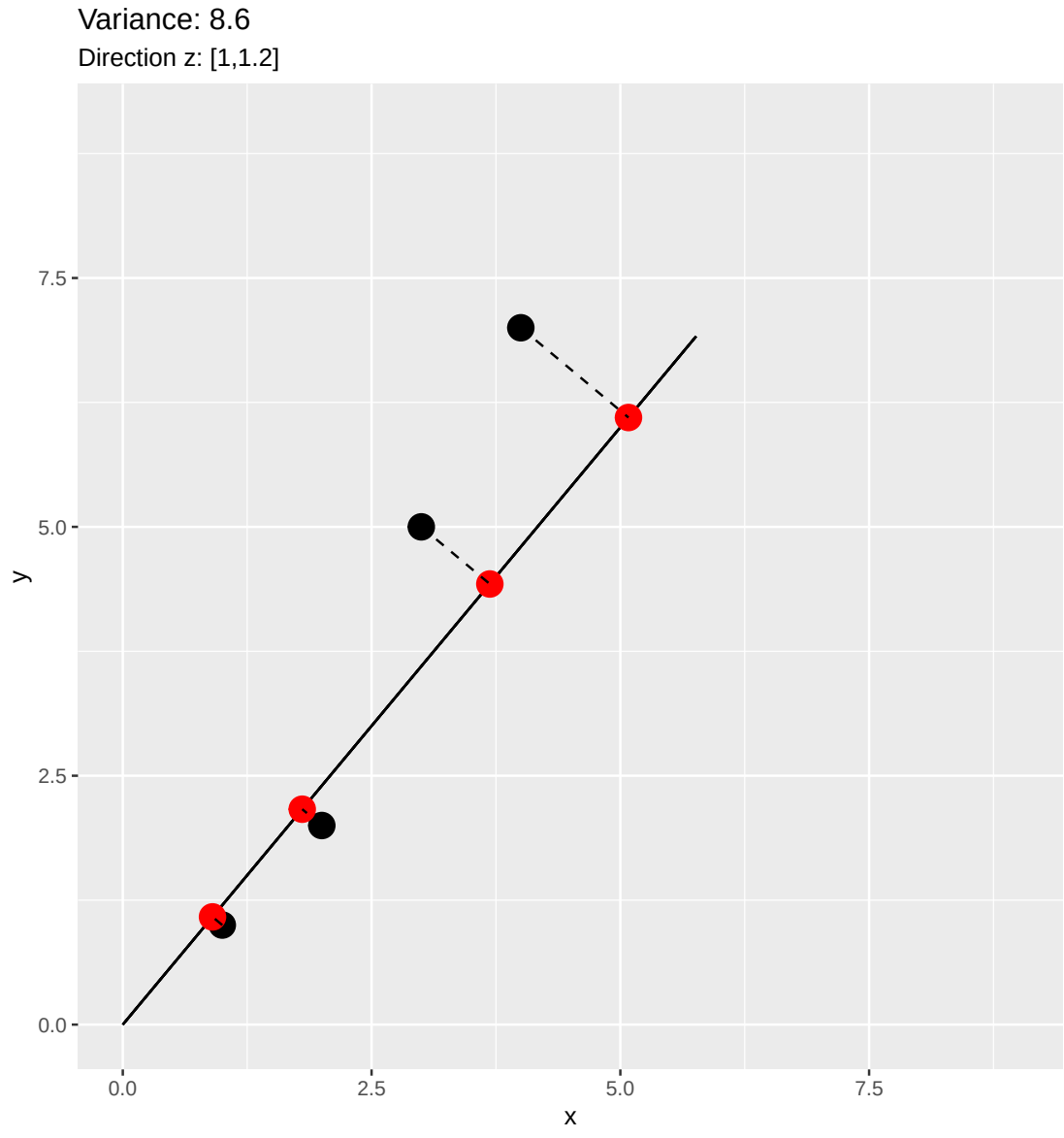

Variance: 6.9
Direction z: [1,0.7]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

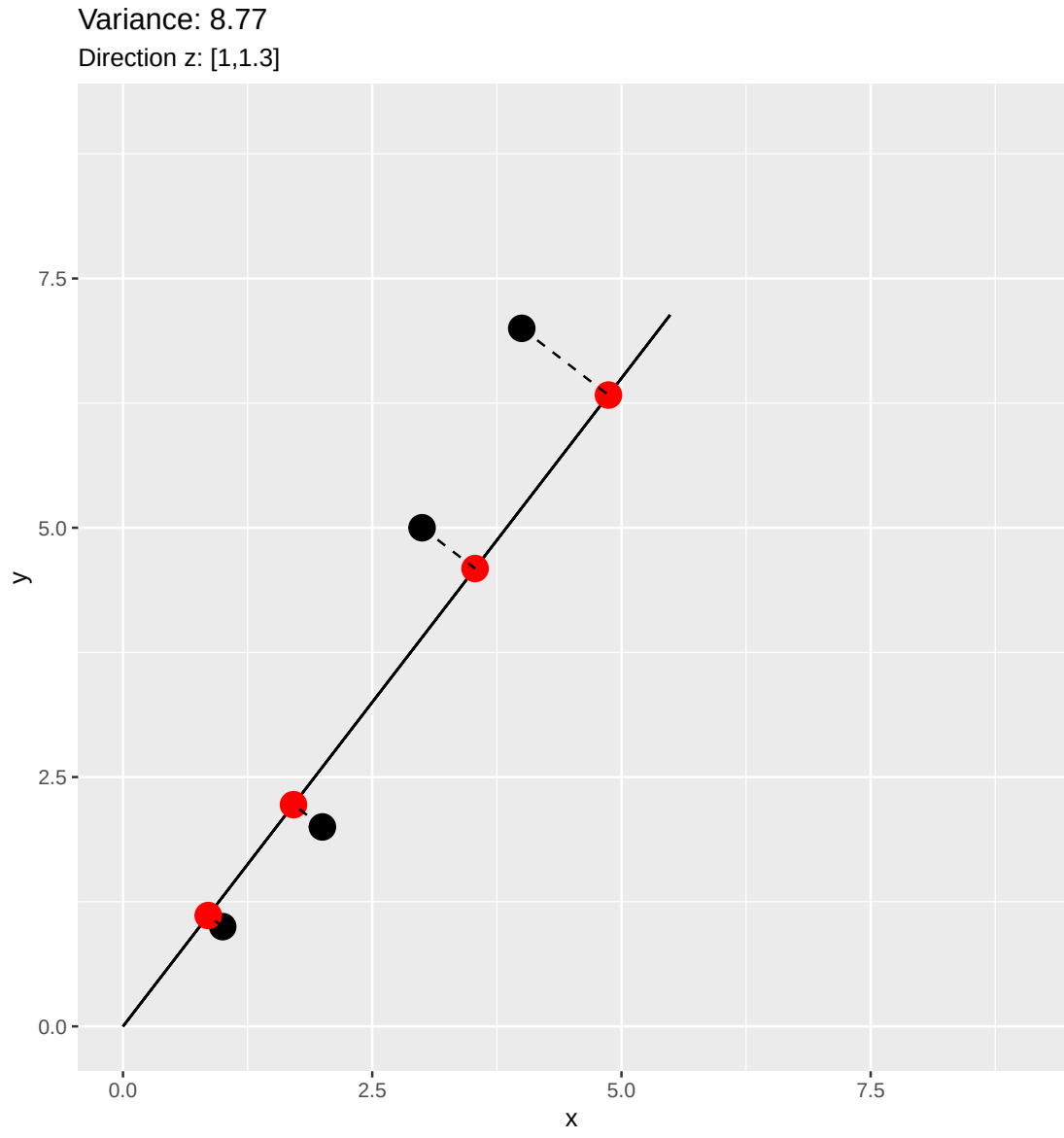

Variance: 7.39
Direction z: [1,0.8]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

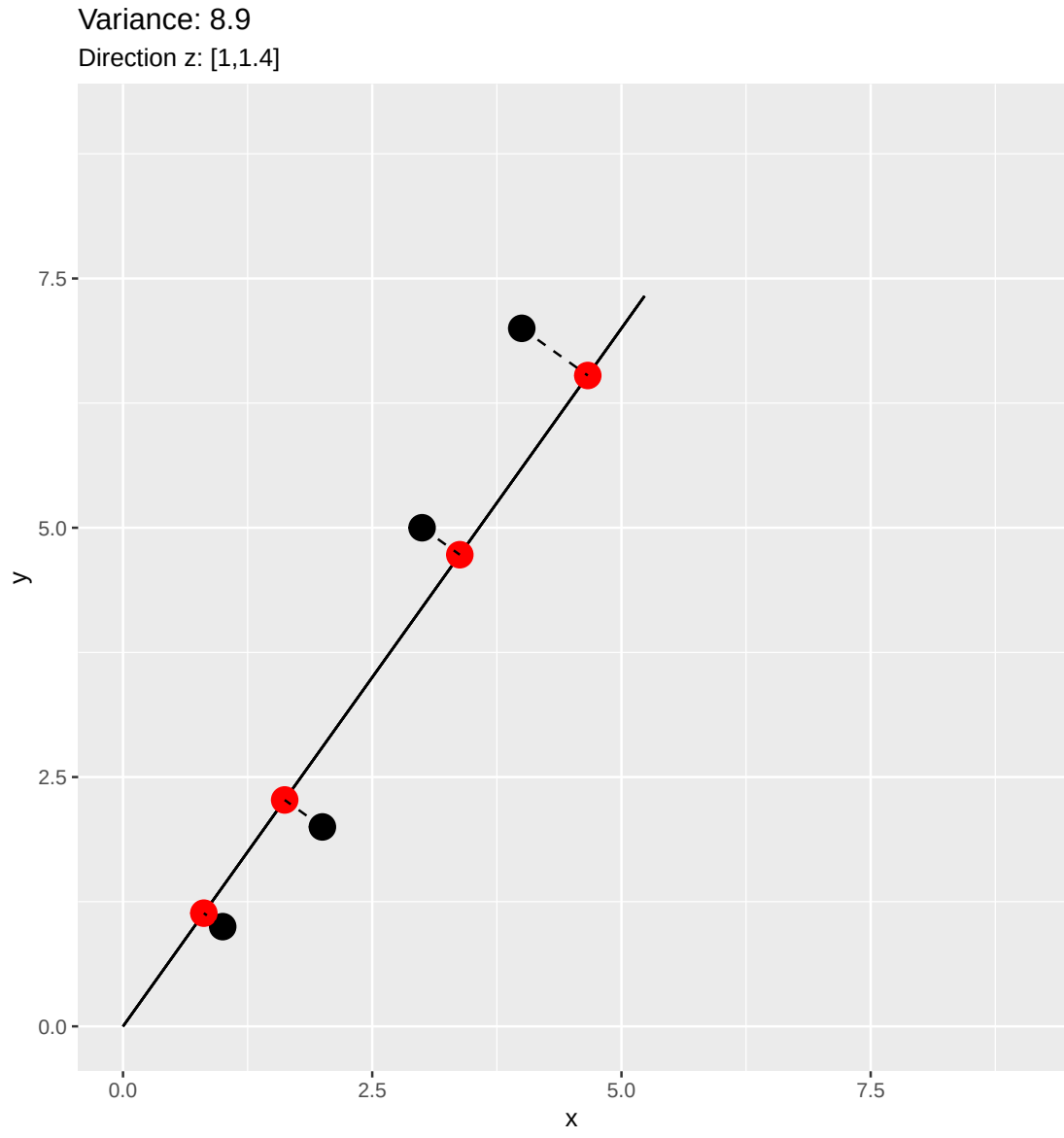

Variance: 7.8
Direction 2: [1,0.9]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

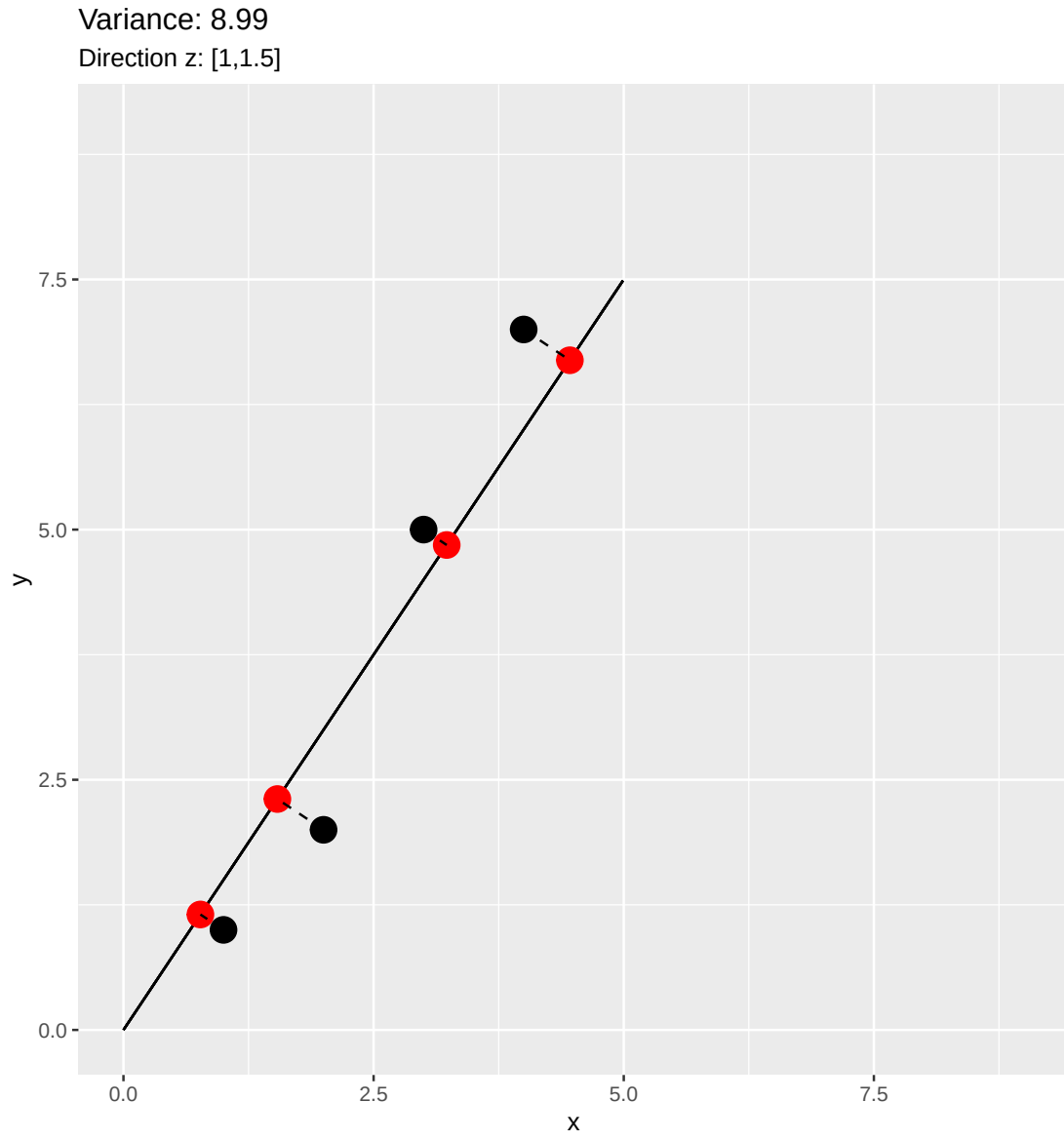

Variance: 8.13
Direction z: [1,1]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

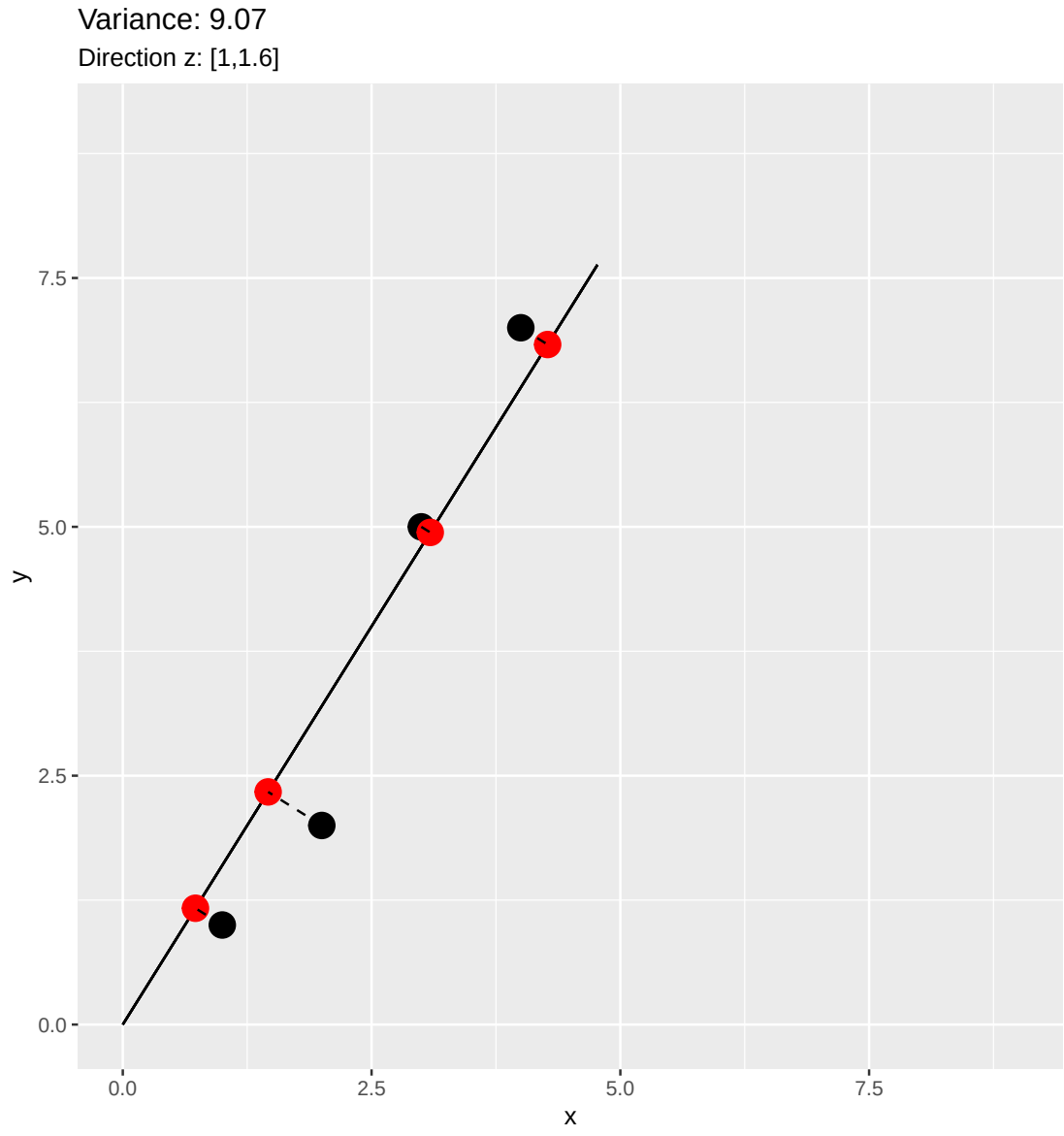

Variance: 8.39
Direction z: [1,1.1]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

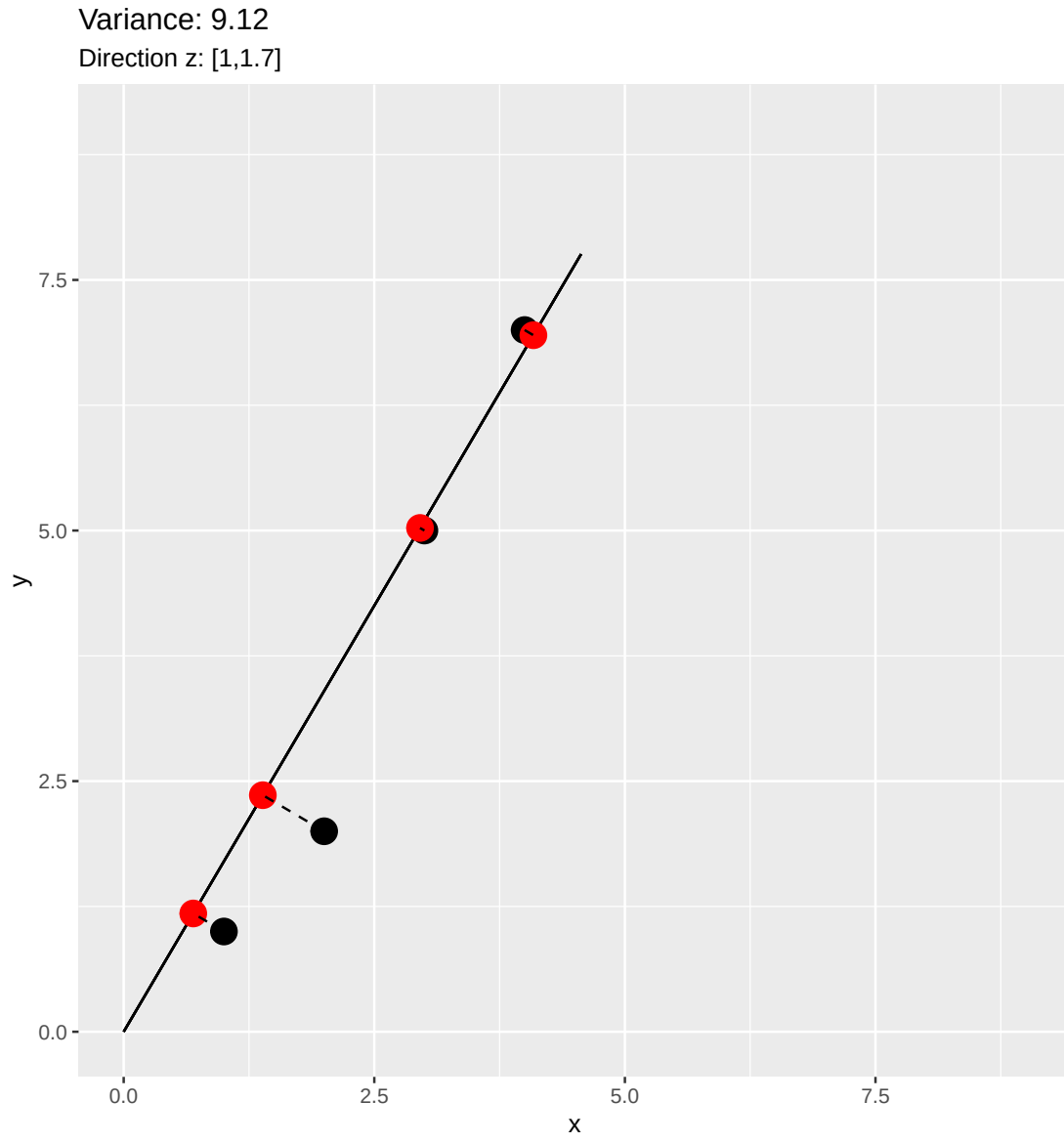

Variance: 8.6
Direction z: [1,1.2]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

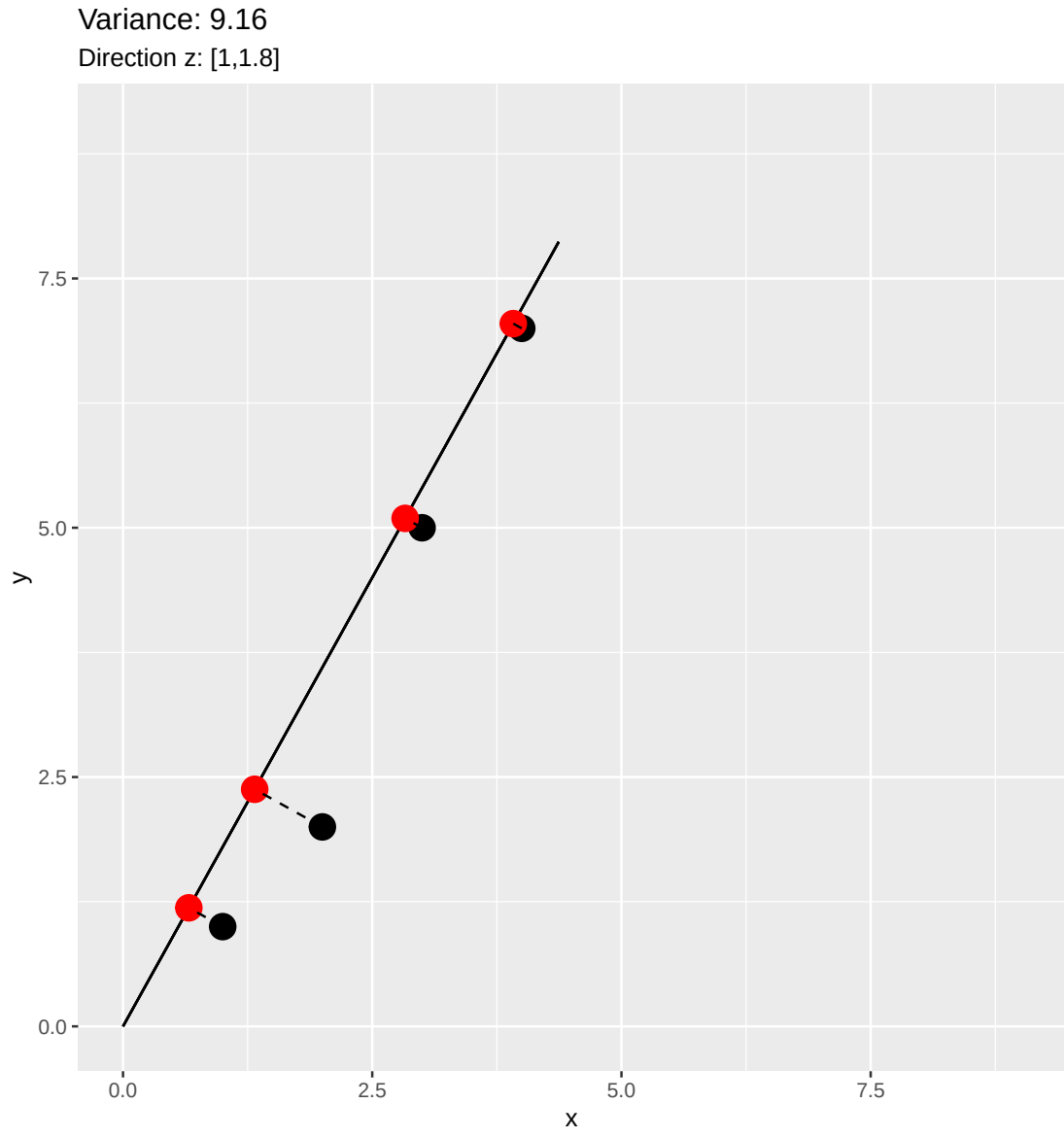

Variance: 8.77
Direction z: [1,1.3]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

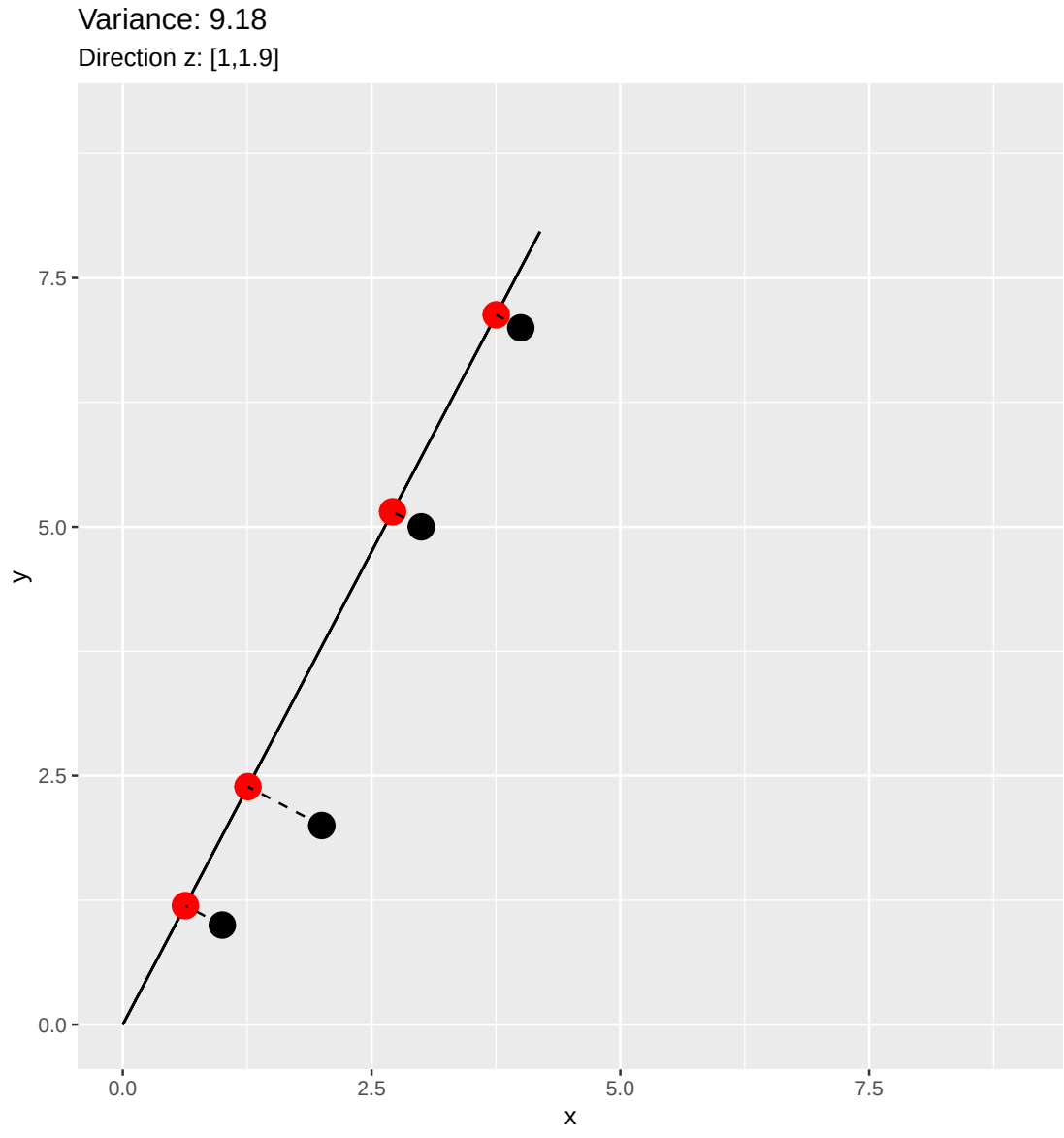

Variance: 8.9
Direction z: [1,1.4]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

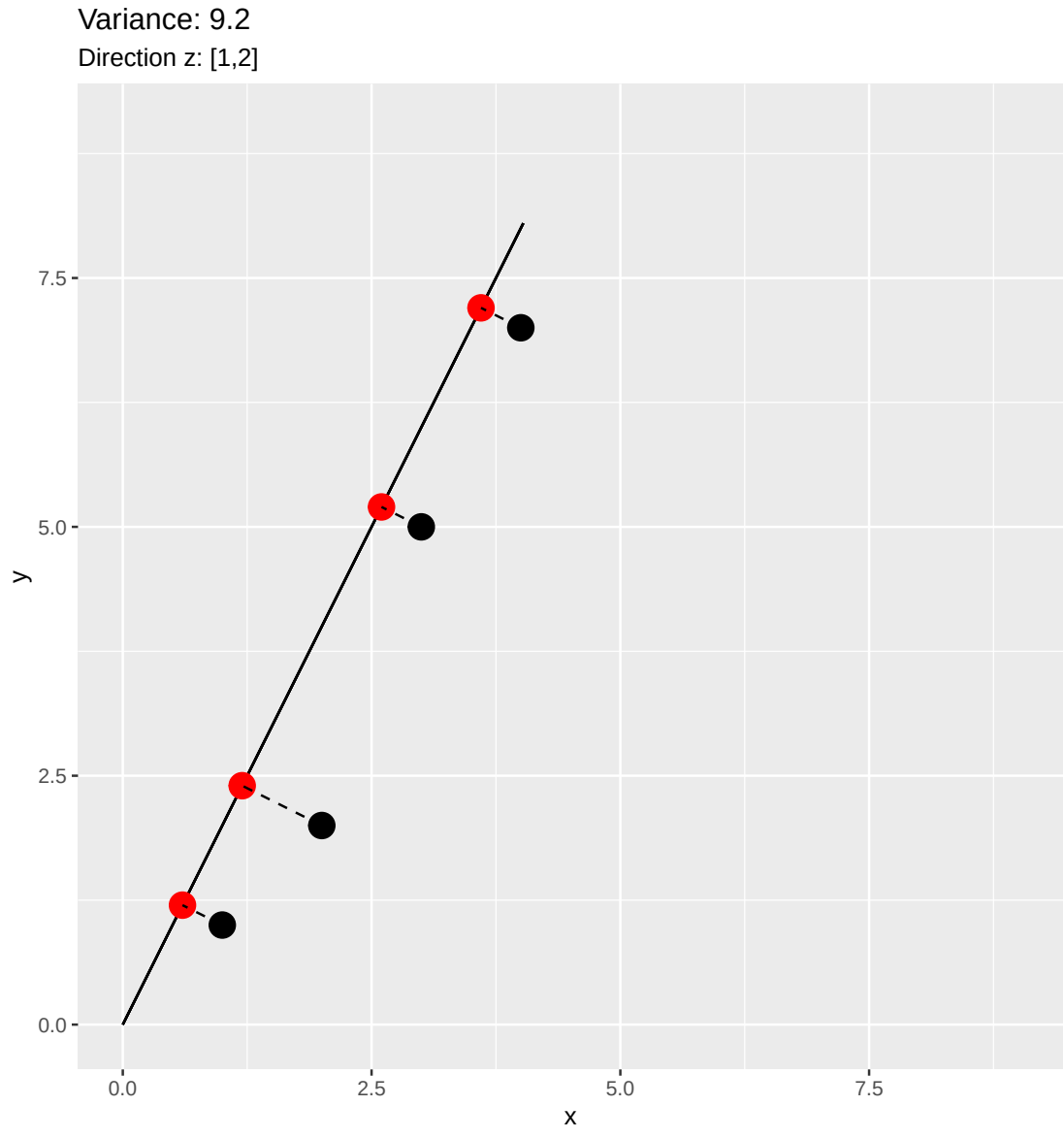

Variance: 8.99
Direction z: [1,1.5]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

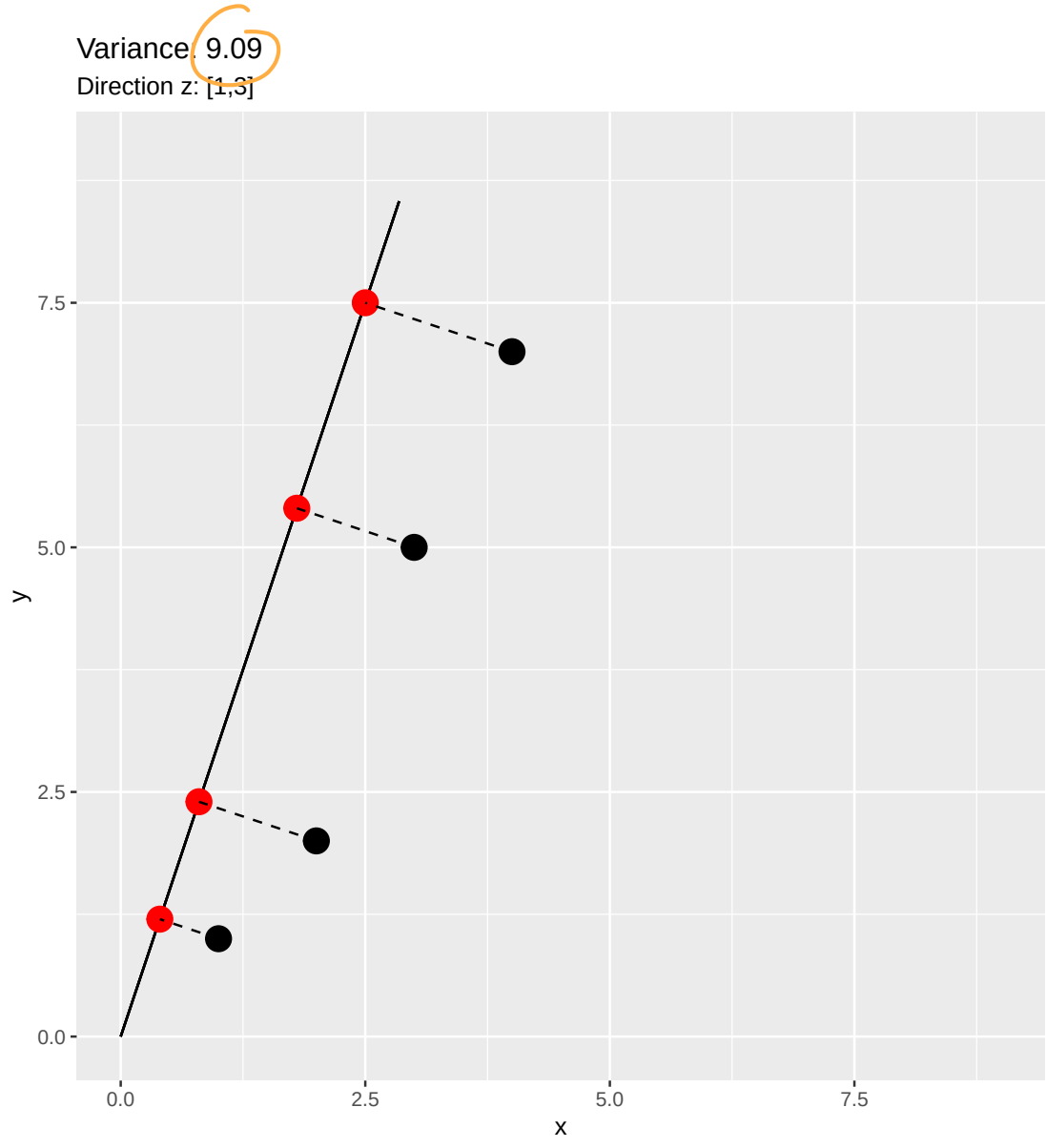

Variance: 9.07
Direction z: [1,1.6]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

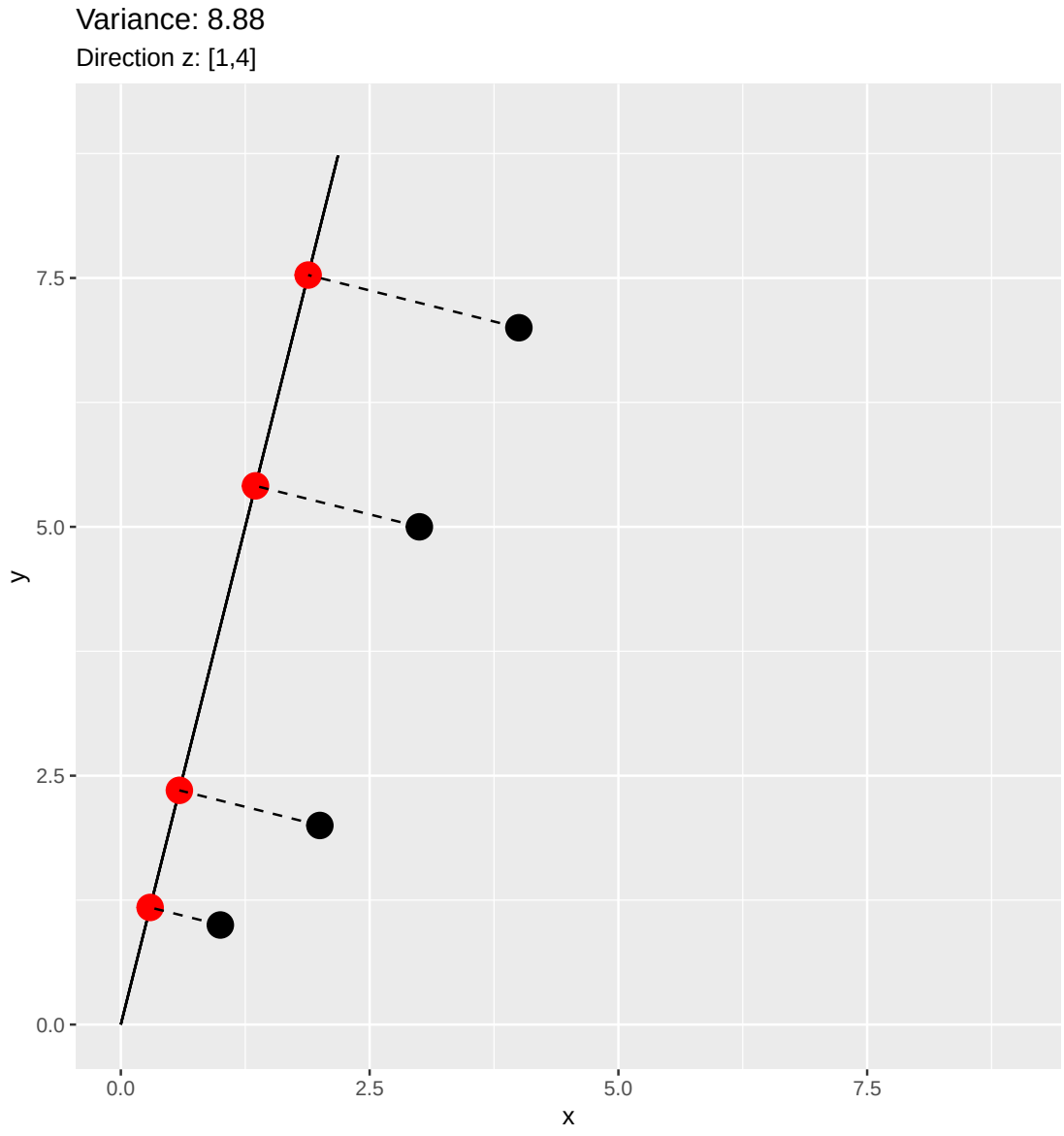

Variance: 9.12
Direction z: [1,1.7]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

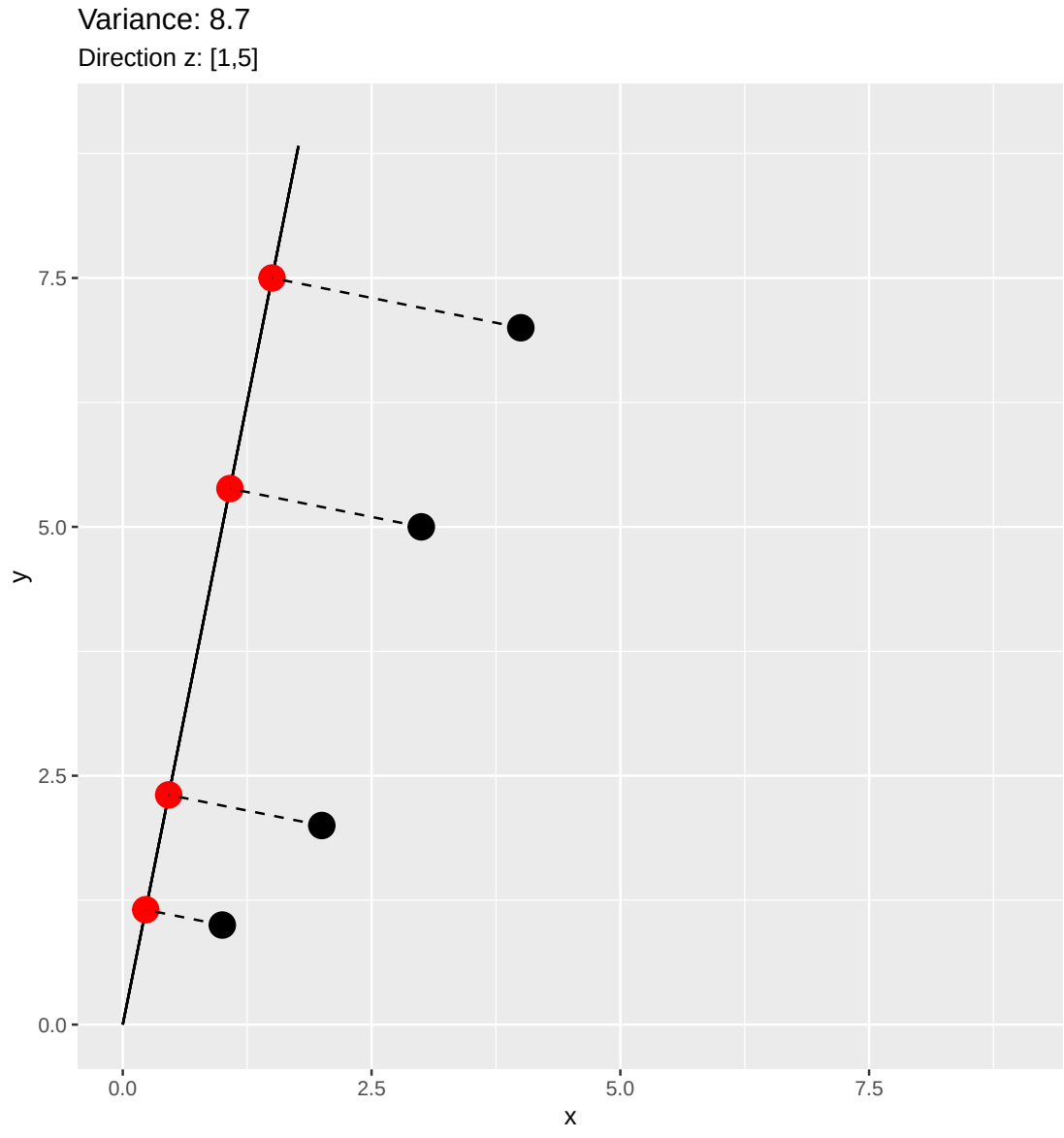

Variance: 9.16
Direction z: [1,1.8]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

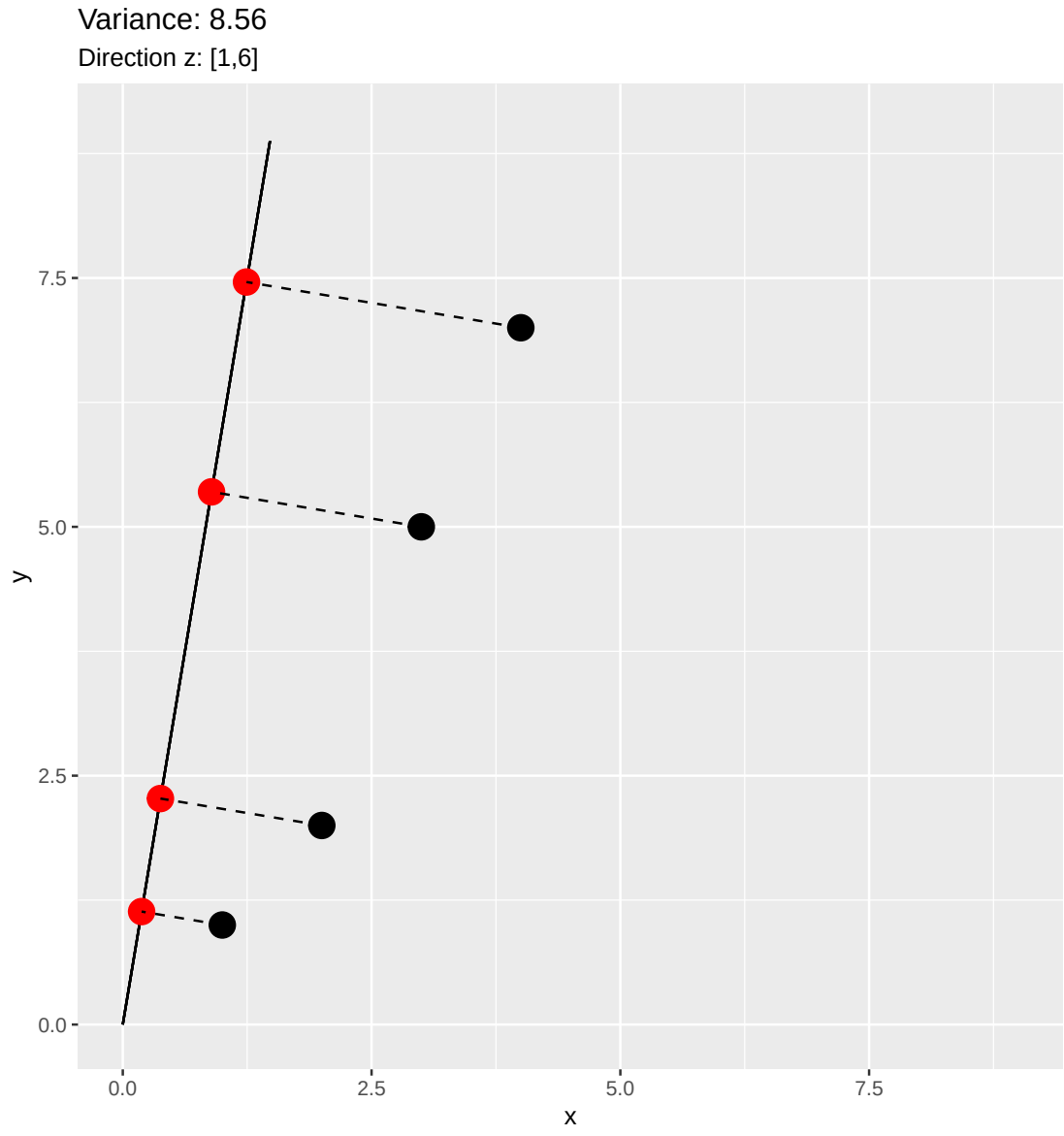

Variance: 9.18
Direction z: [1,1.9]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

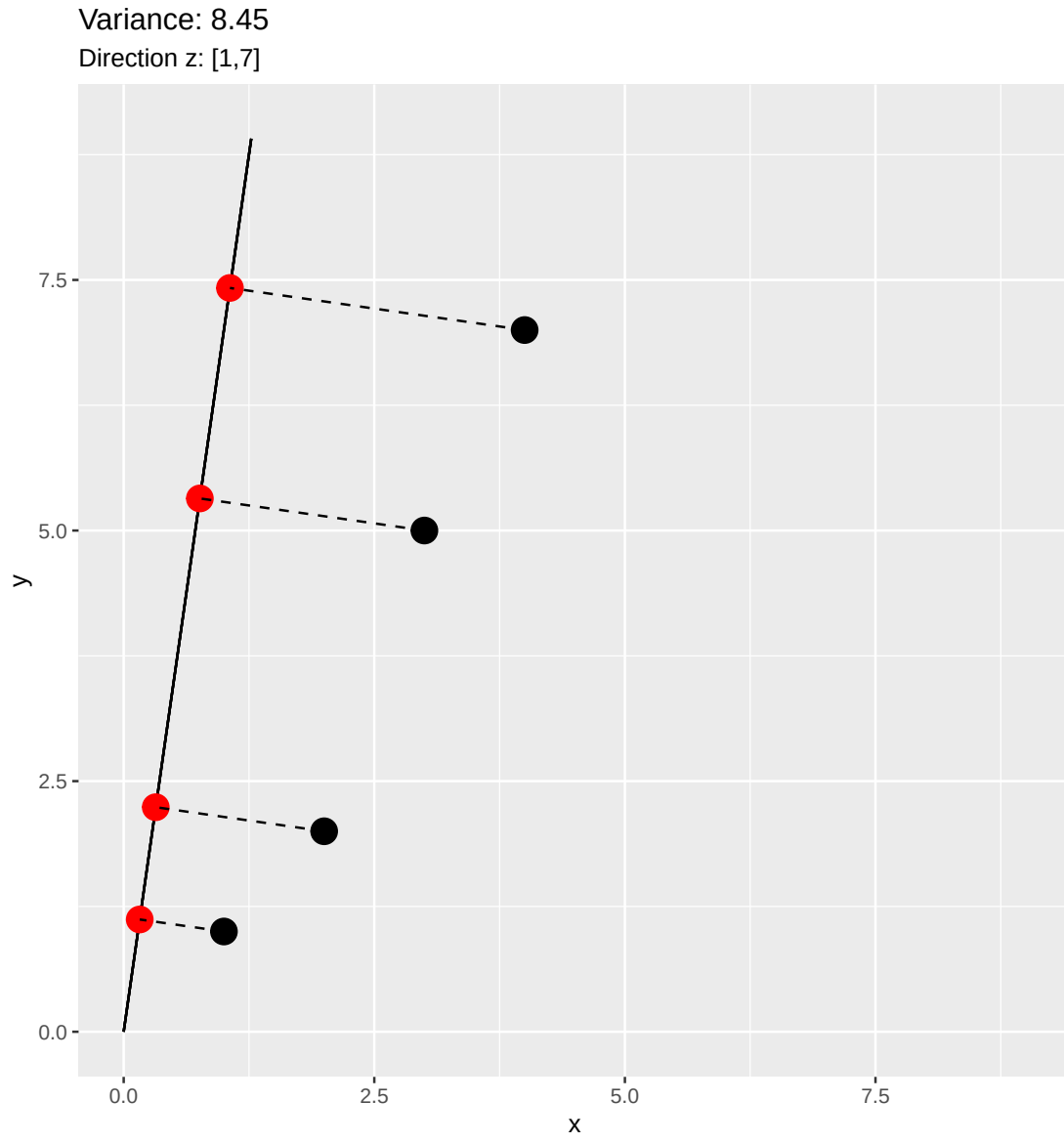

Variance: 9.2
Direction z: [1,2]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

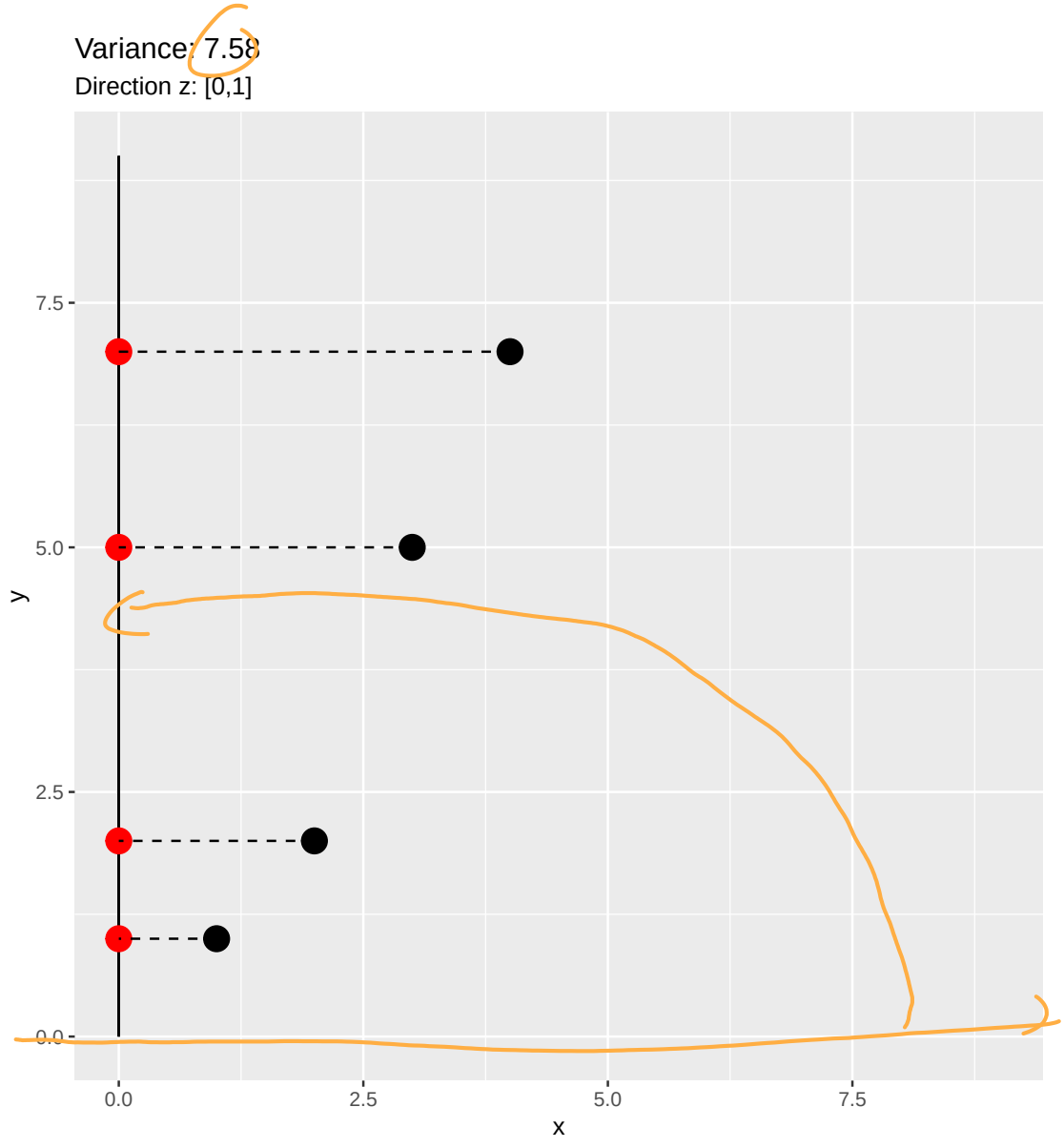

Variance: 9.09
Direction z: [1,3]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

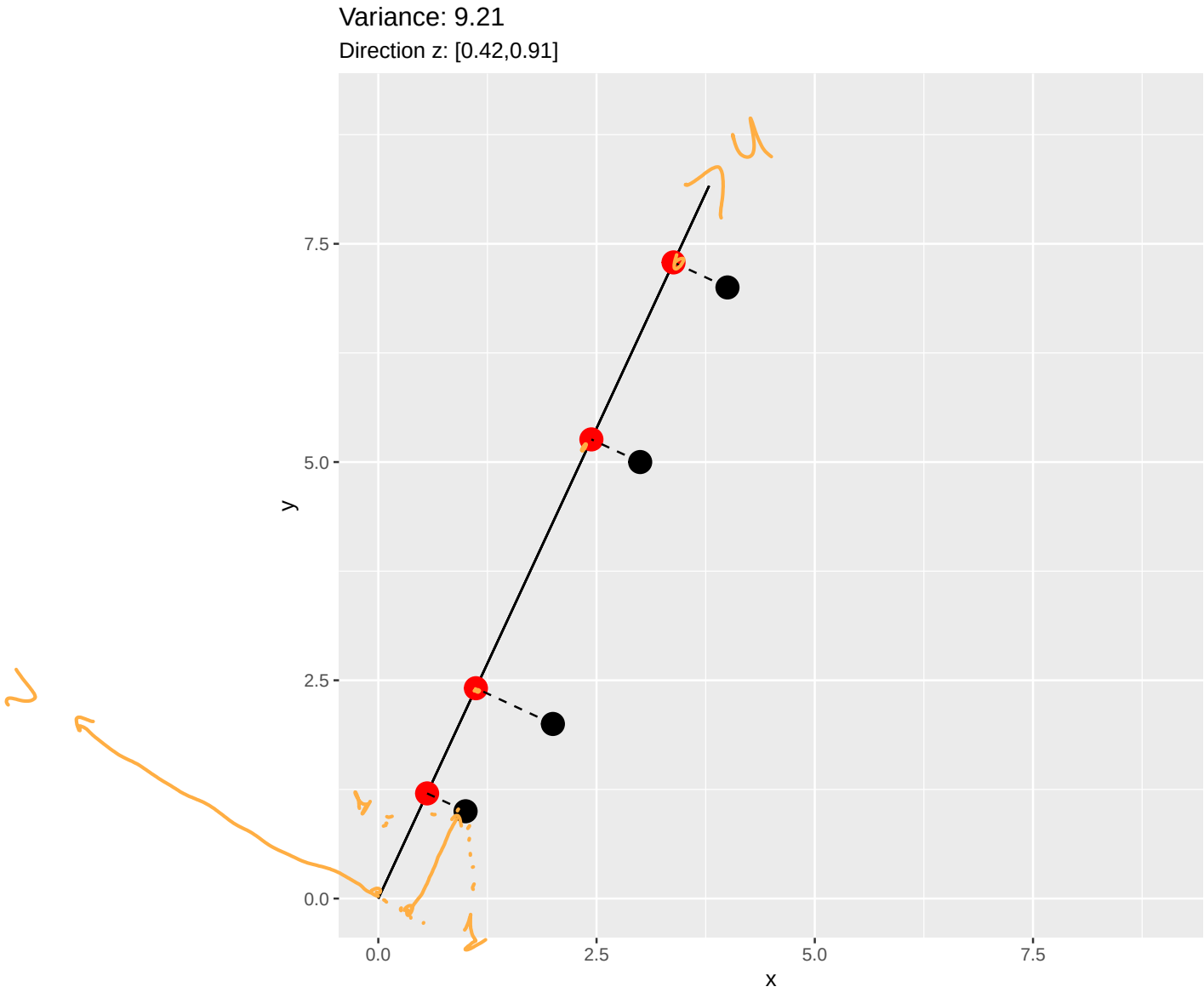

Variance: 8.88
Direction z: [1,4]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x

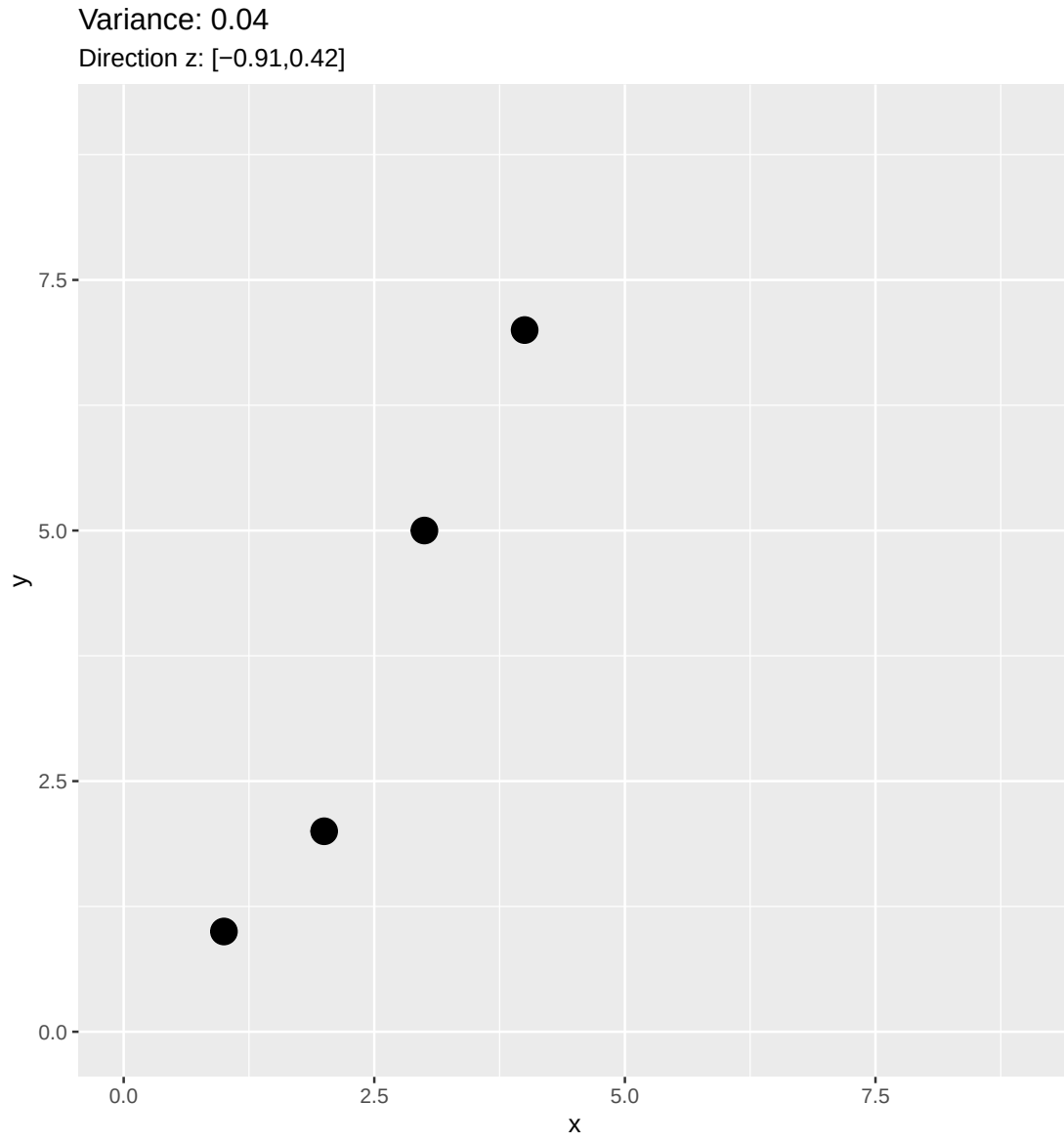

Variance: 8.7
Direction z: [1,5]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x


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


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


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: 9.21
Direction z: [0.42,0.91]
7.5
5.0
2.5
0.0
y
0.0
2.5
5.0
7.5
x


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

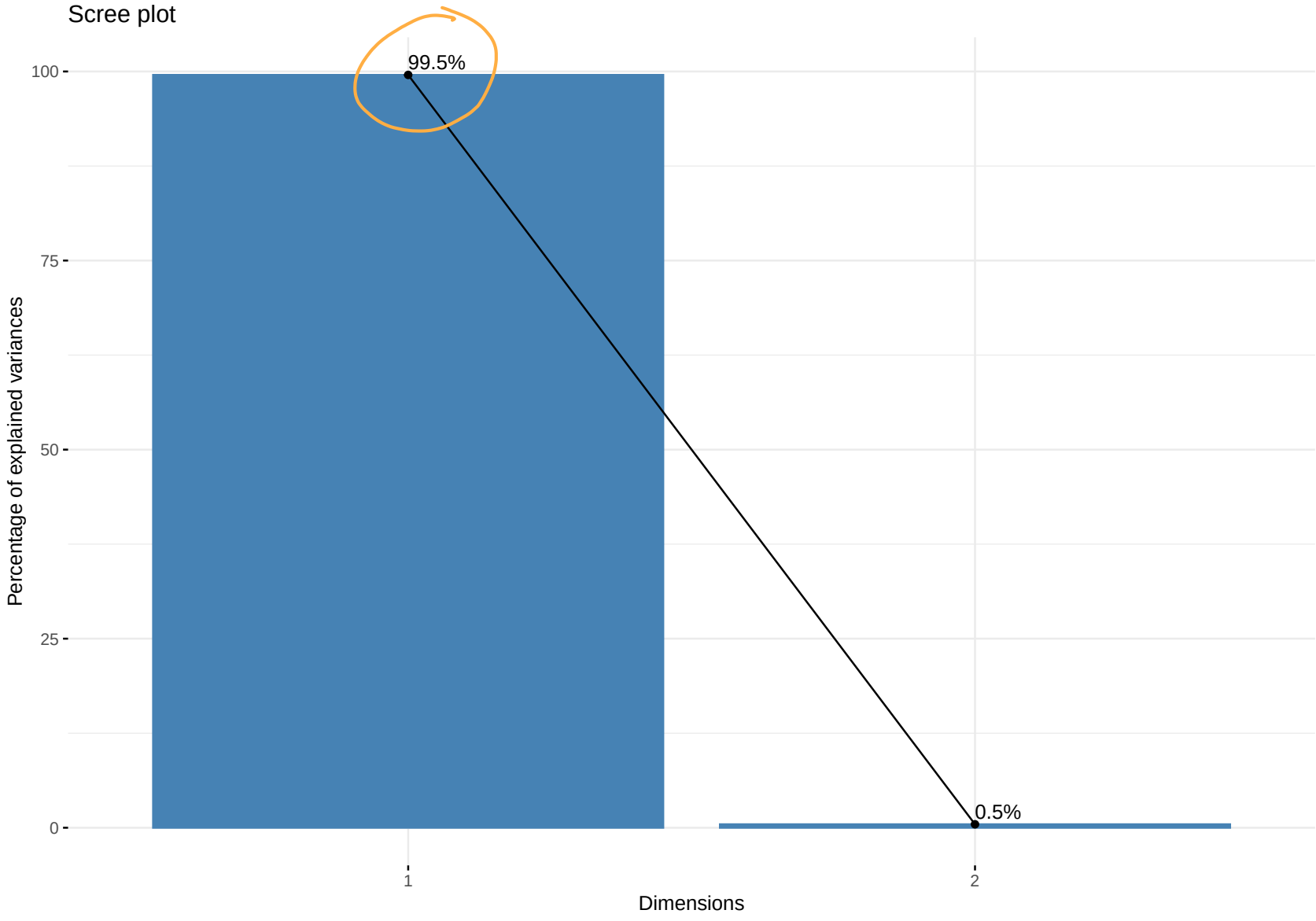
- $\Phi =$

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

PC Scores

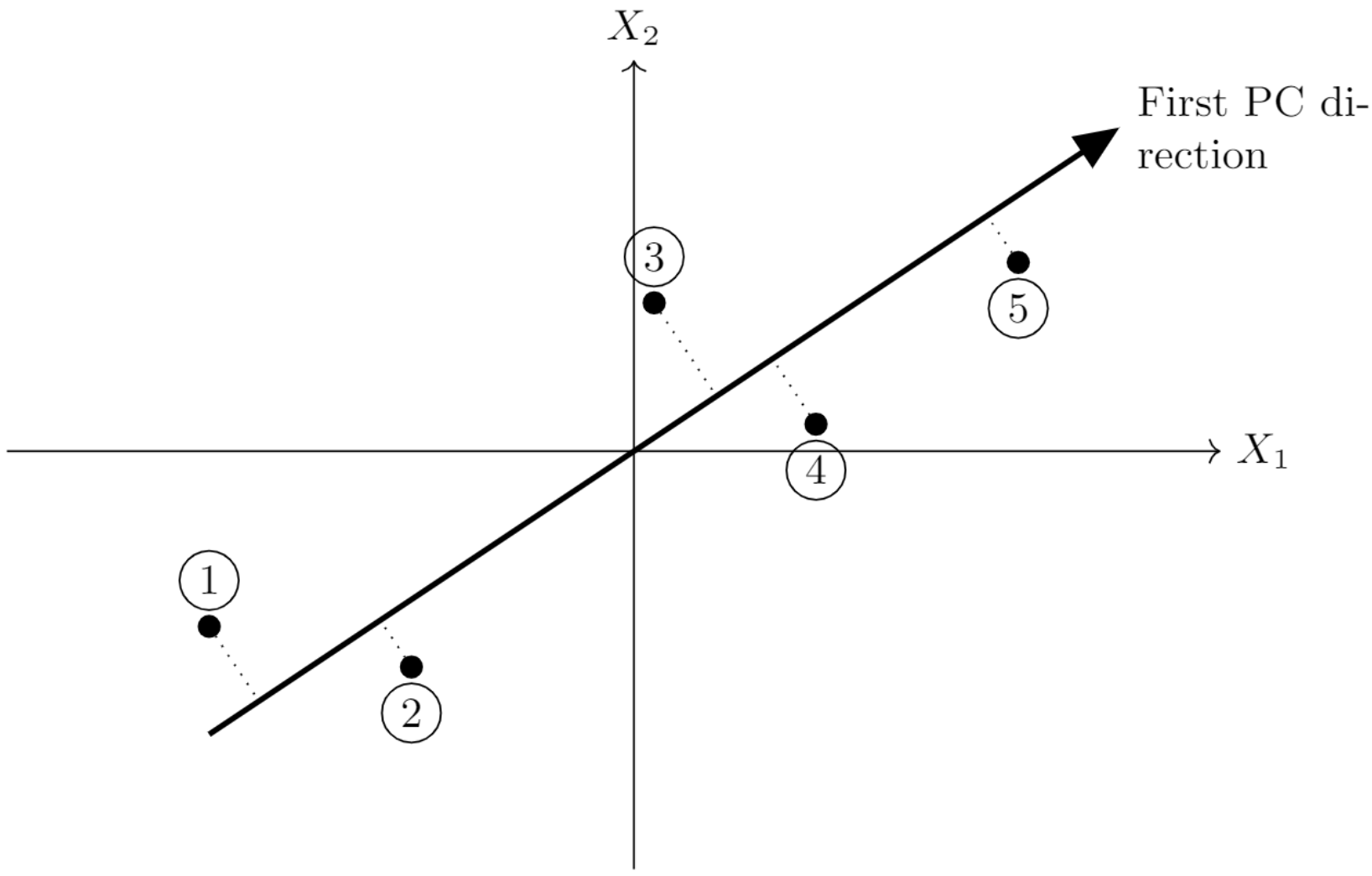
- $Z = X \cdot \Phi =$

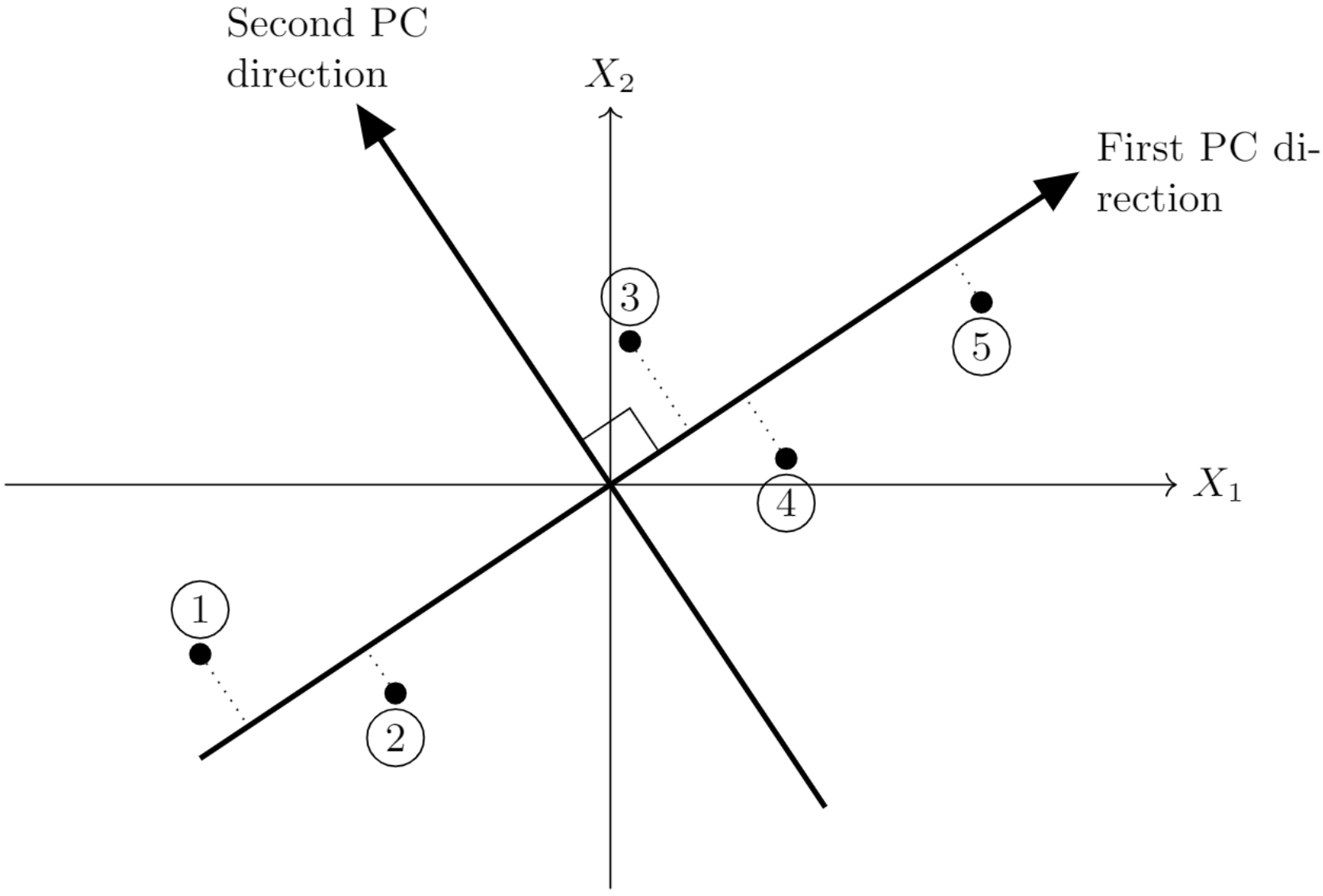
```
        PC1   PC2
[1,] 1.33 -0.49
[2,] 2.66 -0.97
[3,] 5.80 -0.62
[4,] 8.03 -0.68
```


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
First PC direction
3
5
X_1
4
1
2


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

Formula

- Write down matrix form of the example

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

- ϕ is PC loading
- z is PC scores

In general

Original data matrix (fat matrix!)

$$\begin{array}{cccccc} \underline{X_1} & \underline{X_2} & \cdots & \cdots & \underline{X_p} \\ \\ \left(\begin{array}{ccccc} x_{11} & x_{12} & \cdots & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & \cdots & x_{np} \end{array} \right) \end{array}$$

$$\mathbf{X}$$

$\xrightarrow{\text{compressed}}$

New data matrix (thin matrix!)

$$\begin{array}{ccc} \underline{Z_1} & \cdots & \underline{Z_M} \\ \\ \left(\begin{array}{ccc} z_{11} & \cdots & z_{1M} \\ z_{21} & \cdots & z_{2M} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nM} \end{array} \right) \end{array}$$

$$\mathbf{X}\boldsymbol{\phi}_1 \quad \cdots \quad \mathbf{X}\boldsymbol{\phi}_M$$

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

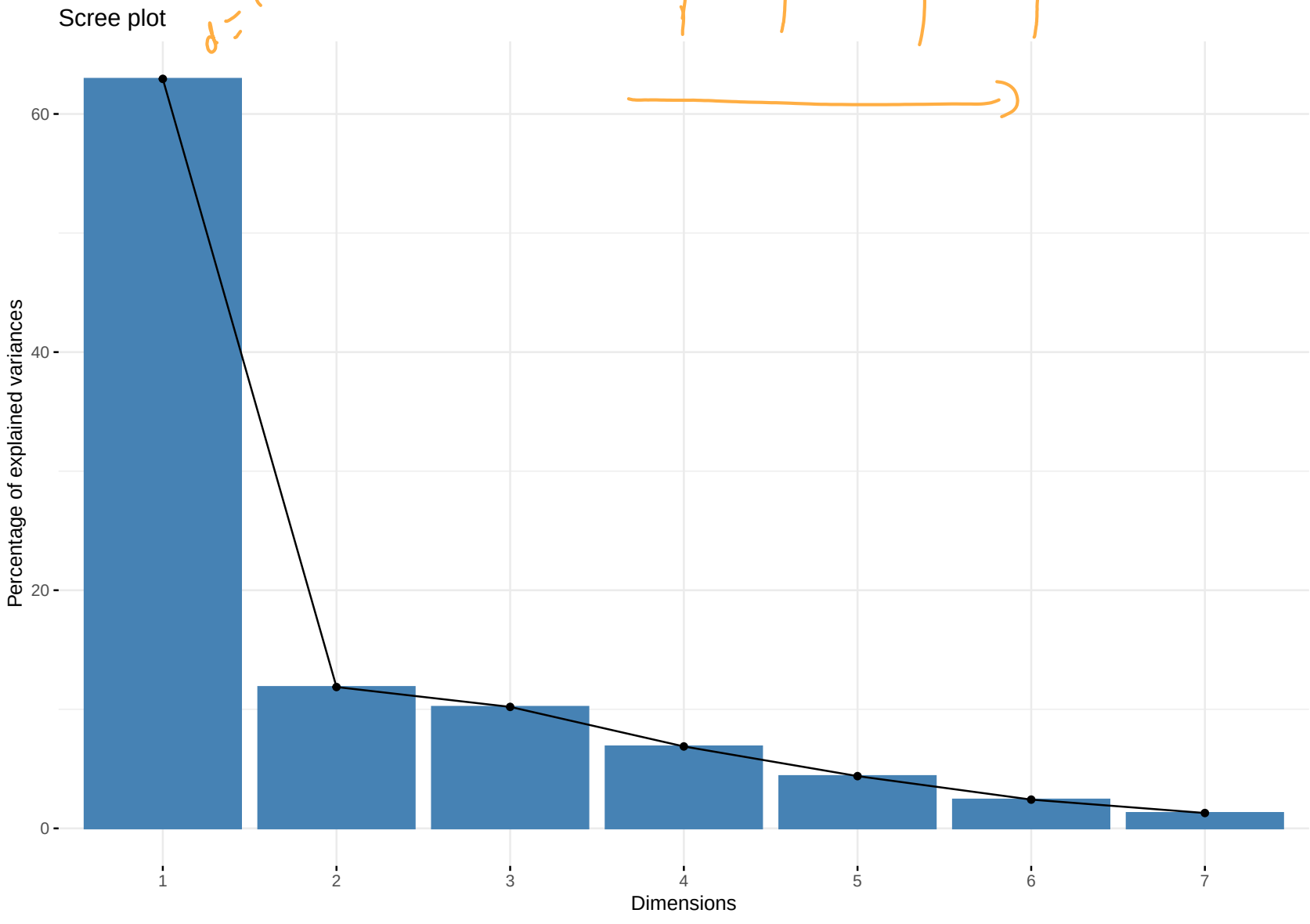
	First PC	Second PC
Murder	0.5359	−0.4182
Assault	0.5832	−0.1880
UrbanPop	0.2782	0.8728
Rape	0.5434	0.1673

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