Section 1

Review session

Generating some data

Generate a matrix of 60 random normals, in 3 columns:

```
set.seed(457299)
z = matrix(rnorm(60, mean = 10, sd = 3), ncol = 3)
head(z)
```

```
##          [,1]   [,2]   [,3]
## [1,] 14.866 13.482  9.502
## [2,]  7.761 11.416  8.067
## [3,]  9.193 12.834  3.525
## [4,]  7.901  7.578  6.525
## [5,] 10.640 10.838 11.557
## [6,] 12.127 12.059  7.012
```

Making some correlated variables

x and y are related, but z has nothing to do with them:

```
w = data.frame(x = z[, 1], y = z[, 1] + 0.9 * z[, 2], z = z
cor(w)
```

```
##         x      y      z
## x 1.0000 0.8268 0.2908
## y 0.8268 1.0000 0.2361
## z 0.2908 0.2361 1.0000
```

```
rm(z)
```

Principal components

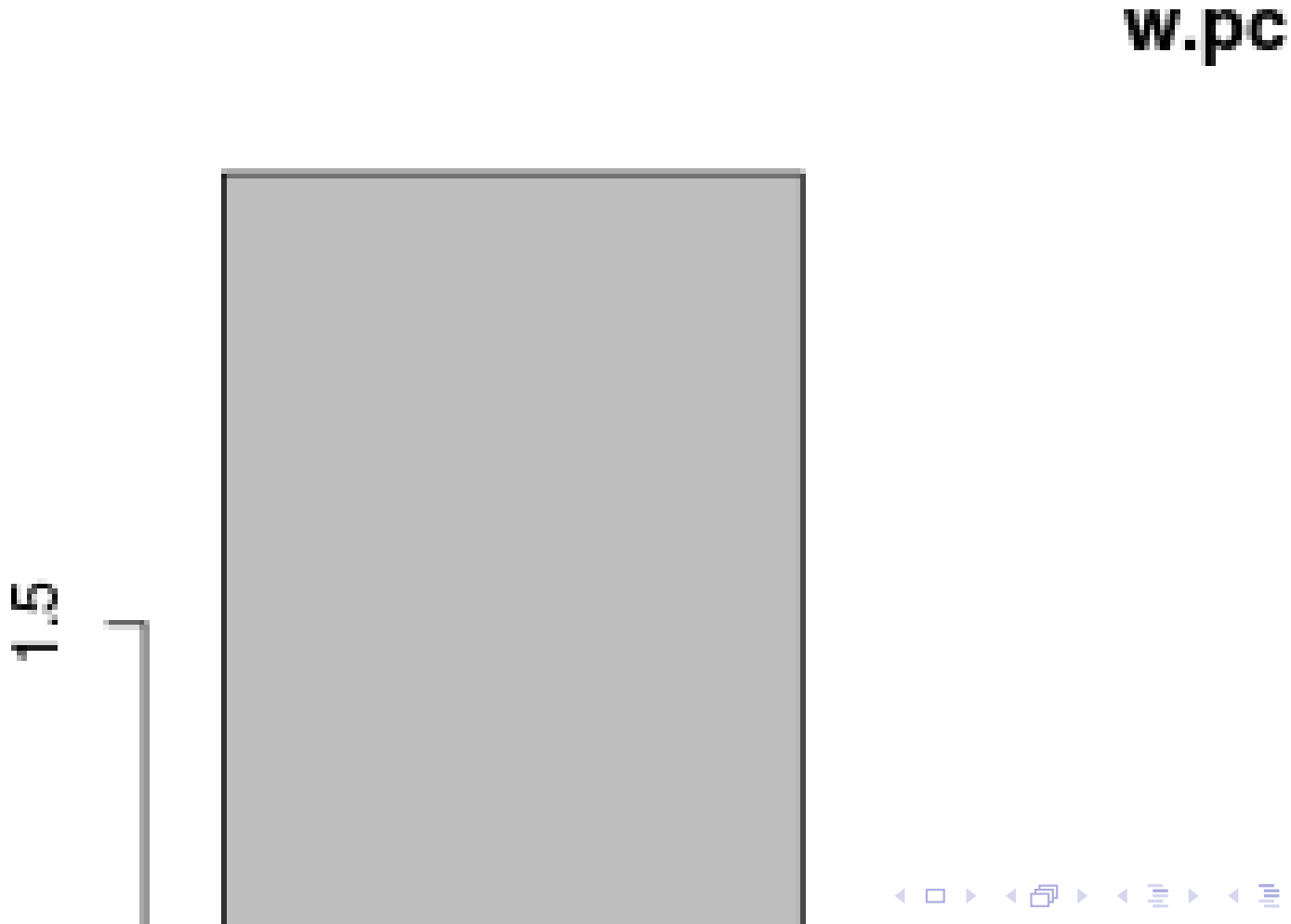
```
w.pc = princomp(w, cor = T)
summary(w.pc)
```

```
## Importance of components:
##                           Comp.1 Comp.2  Comp.3
## Standard deviation     1.4036 0.9267 0.41372
## Proportion of Variance 0.6567 0.2863 0.05705
## Cumulative Proportion  0.6567 0.9429 1.00000
```

2 components explain 94% of variability.

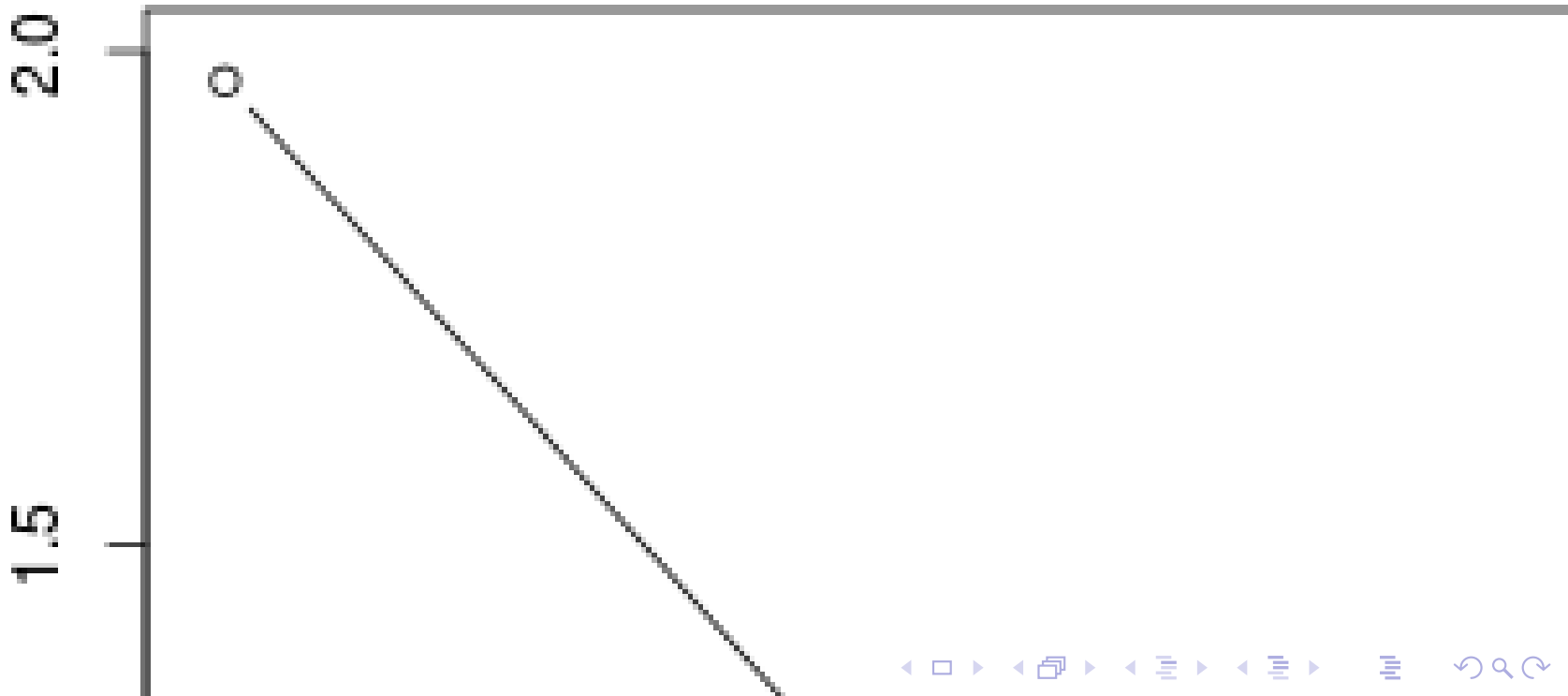
Scree plot (version 1)

```
plot(w.pc)
```



Scree plot (version 2)

```
plot(w.pc$sdev^2, type = "b")
abline(h = 1, lty = "dashed")
```



Decision to make:

- elbow at 2, suggests 1 component.
- 2nd eigenvalue close to 1, suggests 2 components.
- 1 component explains 66% of variability
- 2 components explain 94% of variability.

I go with 2 components.

Loadings

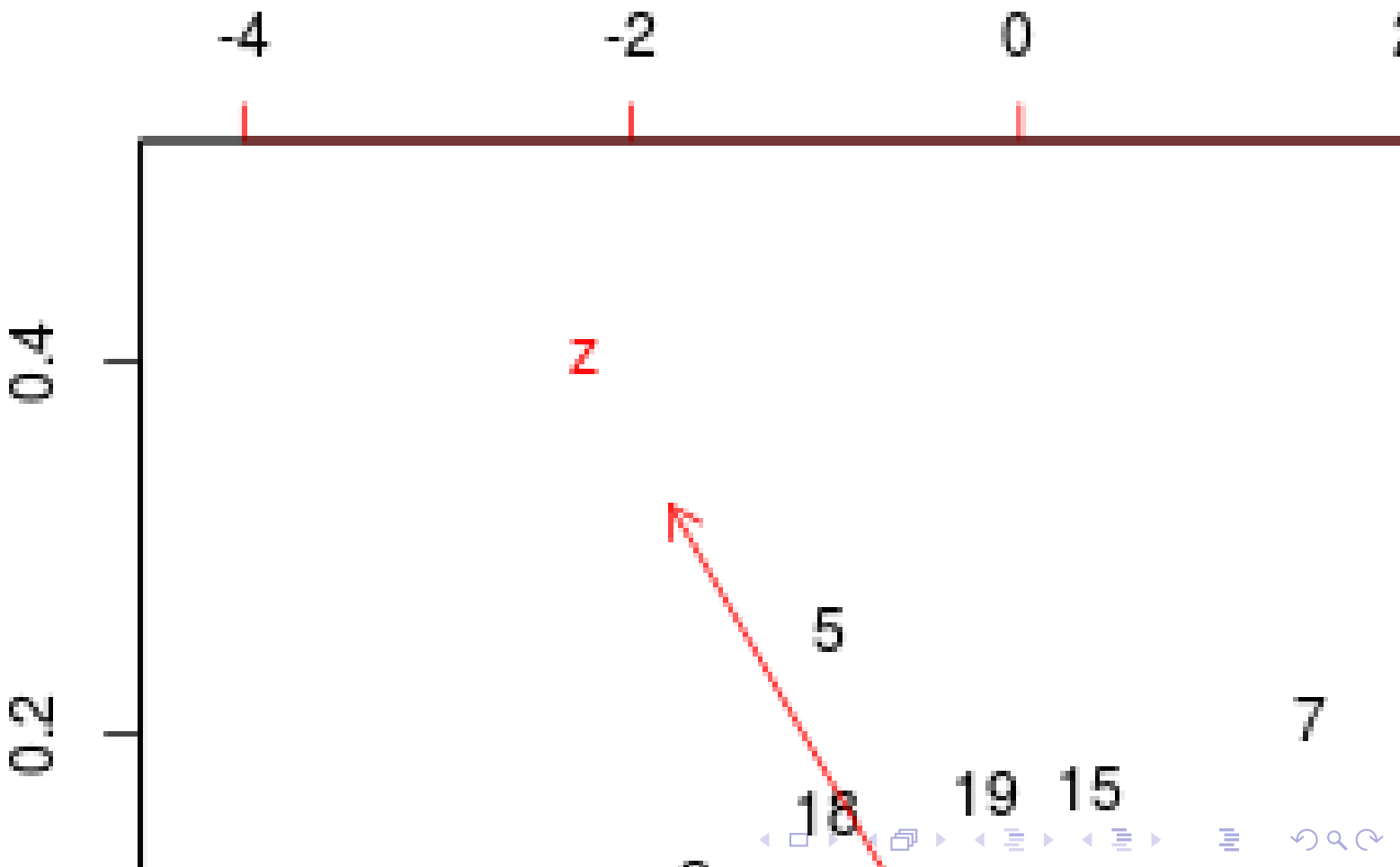
```
w.pc$loadings
```

```
##
## Loadings:
##   Comp.1 Comp.2 Comp.3
## x -0.665 -0.216  0.714
## y -0.655 -0.291 -0.698
## z -0.359  0.932
##
##                Comp.1 Comp.2 Comp.3
## SS loadings     1.000  1.000  1.000
## Proportion Var  0.333  0.333  0.333
## Cumulative Var  0.333  0.667  1.000
```

Component 1 mostly x and y (negatively), component 2 mostly z. (z had nothing to do with x and y, which were related.)

Biplot

```
biplot(w.pc)
```



Individuals

Individual 1 should be high on x and y, 12 (or 20) low on both.
Individual 3 should be low on z, 11 high (and also low on x and y)

```
summary(w)
```

```
##        x                y               z        
##  Min.   : 4.84   Min.   :13.6   Min.   : 2.74  
##  1st Qu.: 7.76   1st Qu.:17.9   1st Qu.: 7.32  
##  Median :10.09   Median :19.8   Median : 9.34  
##  Mean   :10.02   Mean   :20.0   Mean   : 8.47  
##  3rd Qu.:12.41   3rd Qu.:22.6   3rd Qu.:10.19  
##  Max.   :14.87   Max.   :27.0   Max.   :11.75
```

```
pickout = c(1, 12, 20, 3, 11)
w[pickout, ]
```

```
##          x     y      z
## 1   14.866 27.00  9.502
## 12   6.379 13.64  7.420
## 20   4.839 14.86  2.736
```

Summary

Without using a biplot

Look at loadings first to determine which variables have to do with which components:

```
w.pc$loadings
```

```
##
## Loadings:
##   Comp.1 Comp.2 Comp.3
## x -0.665 -0.216  0.714
## y -0.655 -0.291 -0.698
## z -0.359  0.932
##
##                Comp.1 Comp.2 Comp.3
## SS loadings     1.000  1.000  1.000
## Proportion Var  0.333  0.333  0.333
## Cumulative Var  0.333  0.667  1.000
```

- Component 1 mostly x and y (negative)
- Component 2 z (positive)

Plotting component scores

```
labels = as.character(1:20)
plot(w.pc$scores, type = "n")
text(w.pc$scores, labels)
```

Adding a group variable

```
cbind(w, group)
```

```
##          x     y      z group
## 1   14.866 27.00  9.502     c
## 2    7.761 18.04  8.067     b
## 3    9.193 20.74  3.525     a
## 4    7.901 14.72  6.525     a
## 5   10.640 20.39 11.557     d
## 6   12.127 22.98  7.012     a
## 7    6.765 17.41  9.662     c
## 8   12.374 21.57 10.609     d
## 9   10.012 18.60  3.847     a
## 10 13.288 19.11 10.249     d
## 11  5.034 14.07 11.750     d
## 12  6.379 13.64  7.420     b
## 13 13.806 26.59  9.019     b
## 14 12.515 19.18  9.174     b
## 15  7.762 20.34 10.023     c
```

Manova: are the groups different on any of the variables?

```
gf = factor(group)
attach(w)
response = cbind(x, y, z)
detach(w)
w.man = manova(response ~ gf)
summary(w.man)
```

```
##           Df Pillai approx F num Df den Df Pr(>F)
## gf         3  0.947     2.46      9     48  0.021 *
## Residuals 16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Yes, something distinguishes groups.

Which variable(s) distinguish groups?

Discriminant analysis.

```
library(MASS)
w.lda = lda(group ~ x + y + z, data = w)
w.lda$scaling
```

```
##         LD1      LD2     LD3
## x -0.09778  0.25938  0.4814
## y  0.04896 -0.37284 -0.1843
## z  0.95083  0.05488 -0.0723
```

LD1 best distinguishes groups, and is almost entirely z.

Discriminant predictions

Or, how separate are the groups?

```
w.lda2 = lda(group ~ x + y + z, data = w, CV = T)
table(group, pred = w.lda2$group)
```

```
## Error: all arguments must have the same length
```