# Dimensionality reduction

*Data Scientist: Tung Dang*

*6/12/2019*

## Data source

This data set was analyzed in Zhao 2011 (Nature Communications 2:467)

```
line <- read.csv("RiceDiversityLine.csv")
pheno <- read.csv("RiceDiversityPheno.csv")
geno <- read.csv("RiceDiversityGeno.csv")
line.pheno <- merge(line, pheno, by.x = "NSFTV.ID", by.y = "NSFTVID")
alldata <- merge(line.pheno, geno, by.x = "NSFTV.ID", by.y = "NSFTVID")
```

```
mydata <- data.frame(
  # Flowering time
  flower.Aber = alldata$Flowering.time.at.Aberdeen,
  flower.Ark = alldata$Flowering.time.at.Arkansas,
  flower.Fari = alldata$Flowering.time.at.Faridpur,
  # Morphology
  culm = alldata$Culm.habit,
  leaf.length = alldata$Flag.leaf.length,
  leaf.width  = alldata$Flag.leaf.width,
  # Yeild components
  plant.height = alldata$Plant.height,
  panicle.length = alldata$Panicle.length,
  pri.panicle.branch = alldata$Primary.panicle.branch.number,
  seed.panicle = alldata$Seed.number.per.panicle,
  flor.panicle = alldata$Florets.per.panicle,
  panicle.fertility = alldata$Panicle.fertility,
  # Seed morphology
  seed.length = alldata$Seed.length,
  seed.width = alldata$Seed.width,
  seed.volum = alldata$Seed.volume,
  seed.surface = alldata$Seed.surface.area,
  brown.length = alldata$Brown.rice.seed.length,
  brown.width = alldata$Brown.rice.seed.width,
  brown.surface = alldata$Brown.rice.surface.area,
  brown.volume = alldata$Brown.rice.volume,
  # Stress tolerance
  straighhead = alldata$Straighthead.suseptability,
  blast = alldata$Blast.resistance,
  # Quality
  amylose = alldata$Amylose.content,
  alkali.spreading = alldata$Alkali.spreading.value,
  protein = alldata$Protein.content
)
missing <- apply(is.na(mydata), 1, sum) > 0
mydata <- mydata[!missing, ]
subpop <- alldata$Sub.population[!missing]
```

# PCA analysis

## 1. Computation PCA function

```
res <- prcomp(mydata, scale = T)
summary(res)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6    PC7
## Standard deviation     2.7544 2.1393 1.7467 1.33991 1.20363 0.95575 0.9178
## Proportion of Variance 0.3035 0.1831 0.1220 0.07181 0.05795 0.03654 0.0337
## Cumulative Proportion  0.3035 0.4865 0.6086 0.68038 0.73833 0.77487 0.8086
##                            PC8     PC9    PC10    PC11    PC12    PC13
## Standard deviation     0.88292 0.85743 0.77320 0.71279 0.67196 0.62265
## Proportion of Variance 0.03118 0.02941 0.02391 0.02032 0.01806 0.01551
## Cumulative Proportion  0.83975 0.86915 0.89307 0.91339 0.93145 0.94696
##                           PC14   PC15    PC16    PC17   PC18    PC19
## Standard deviation     0.59088 0.5830 0.44918 0.44626 0.3905 0.22120
## Proportion of Variance 0.01397 0.0136 0.00807 0.00797 0.0061 0.00196
## Cumulative Proportion  0.96092 0.9745 0.98259 0.99056 0.9967 0.99861
##                           PC20    PC21    PC22    PC23    PC24    PC25
## Standard deviation     0.13123 0.09688 0.07385 0.03862 0.02488 0.02136
## Proportion of Variance 0.00069 0.00038 0.00022 0.00006 0.00002 0.00002
## Cumulative Proportion  0.99930 0.99968 0.99990 0.99996 0.99998 1.00000
```

## 2. Results

```
library("factoextra")
```

```
## Loading required package: ggplot2
```

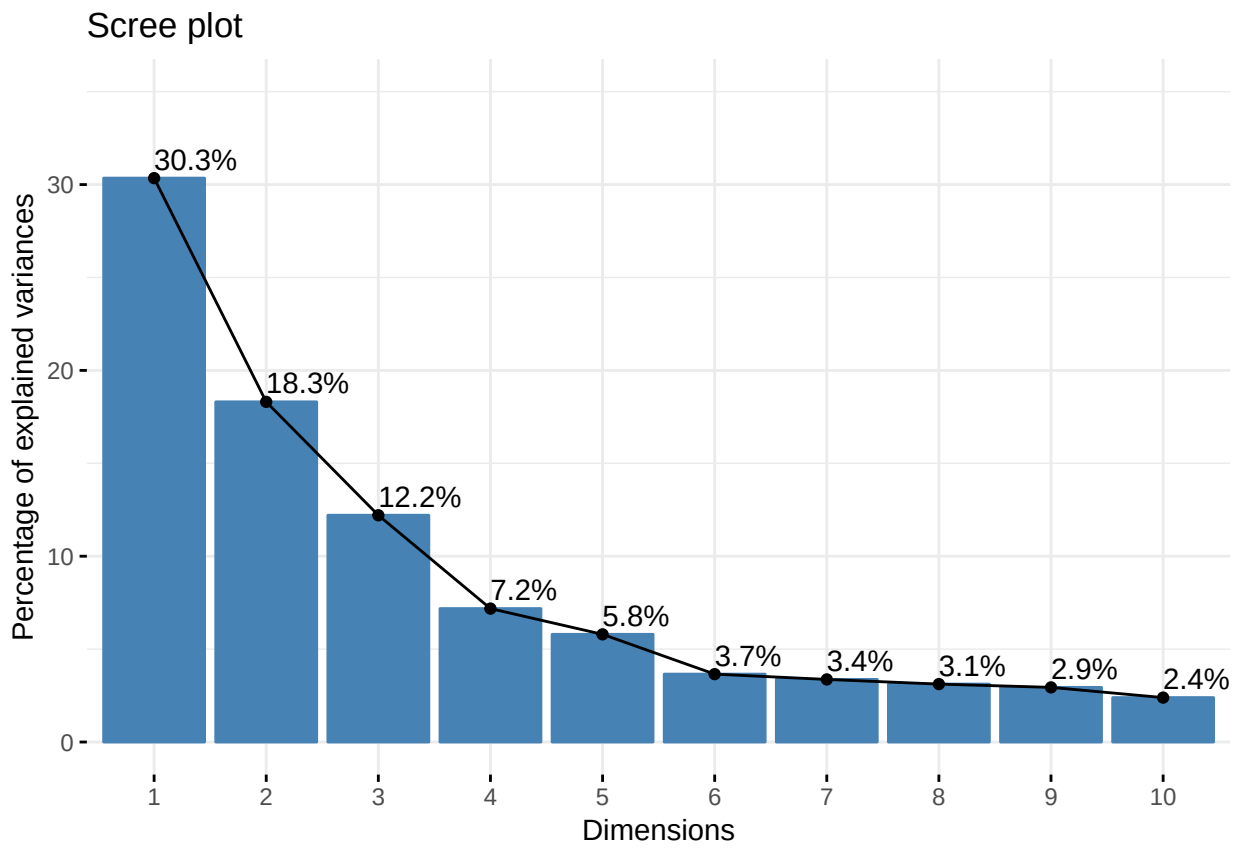
```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```
library("corrplot")
```

```
## corrplot 0.84 loaded
```

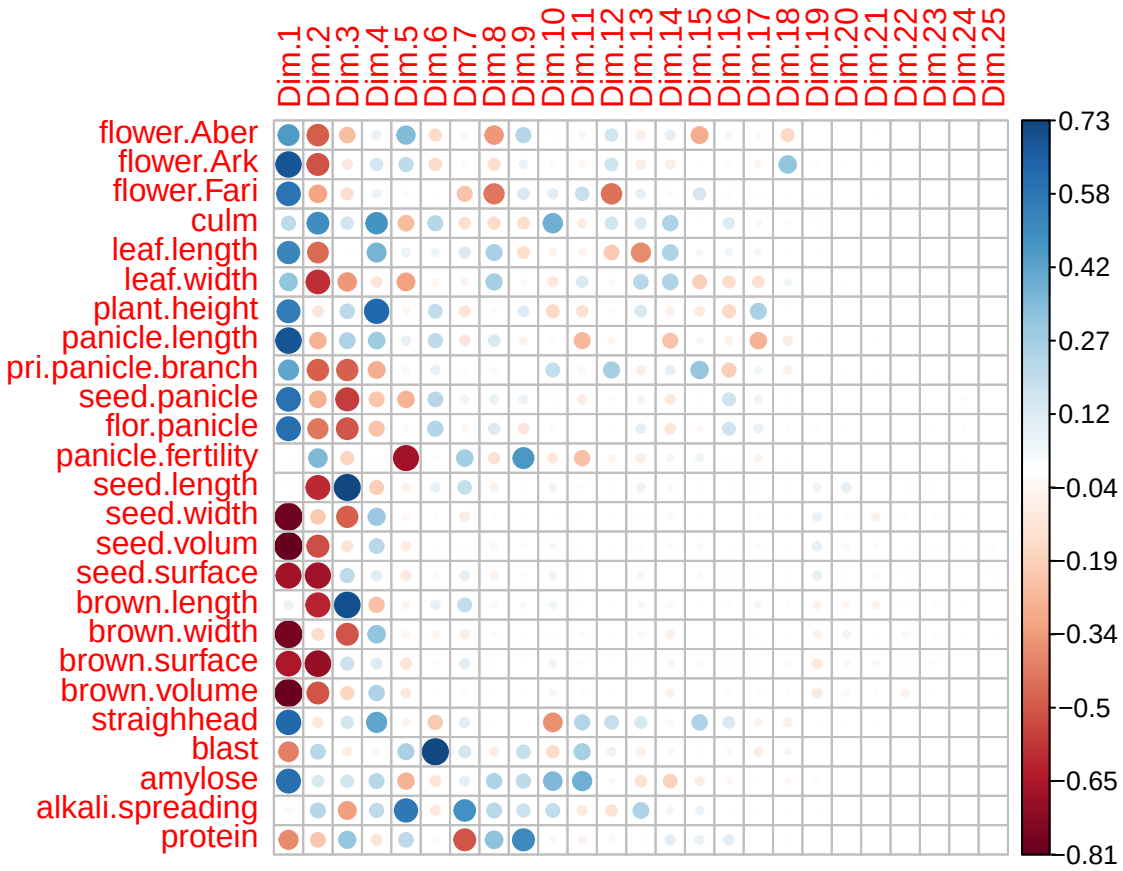
```
fviz_eig(res, addlabels = TRUE, ylim = c(0, 35))
```



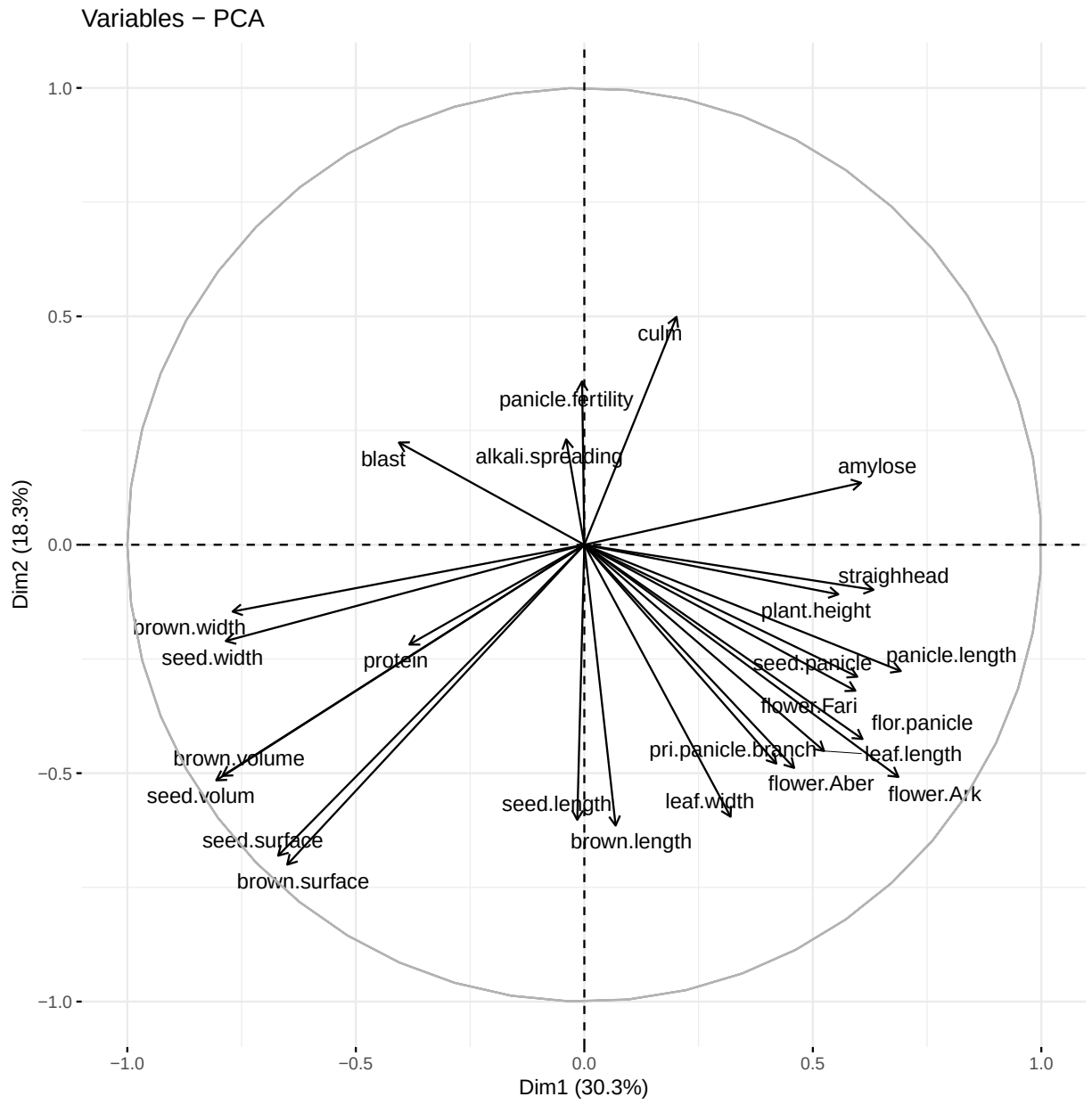

The results show that if we based on the first rule is 80%, the first six principal components with a cumulative contribution of 85.4% are selected. Next, based on the second rule, the first four principal components whose contribution rate exceeds $1/20 = 5\%$ are selected. Moreover, the eigenvalues decreases rapidly until the fourth principal component, and then decreases gradually. Combining the above rules, the first four or six components are considered to be appropriate number of principal components.

### 2.1. Correlations between variables and dimensions

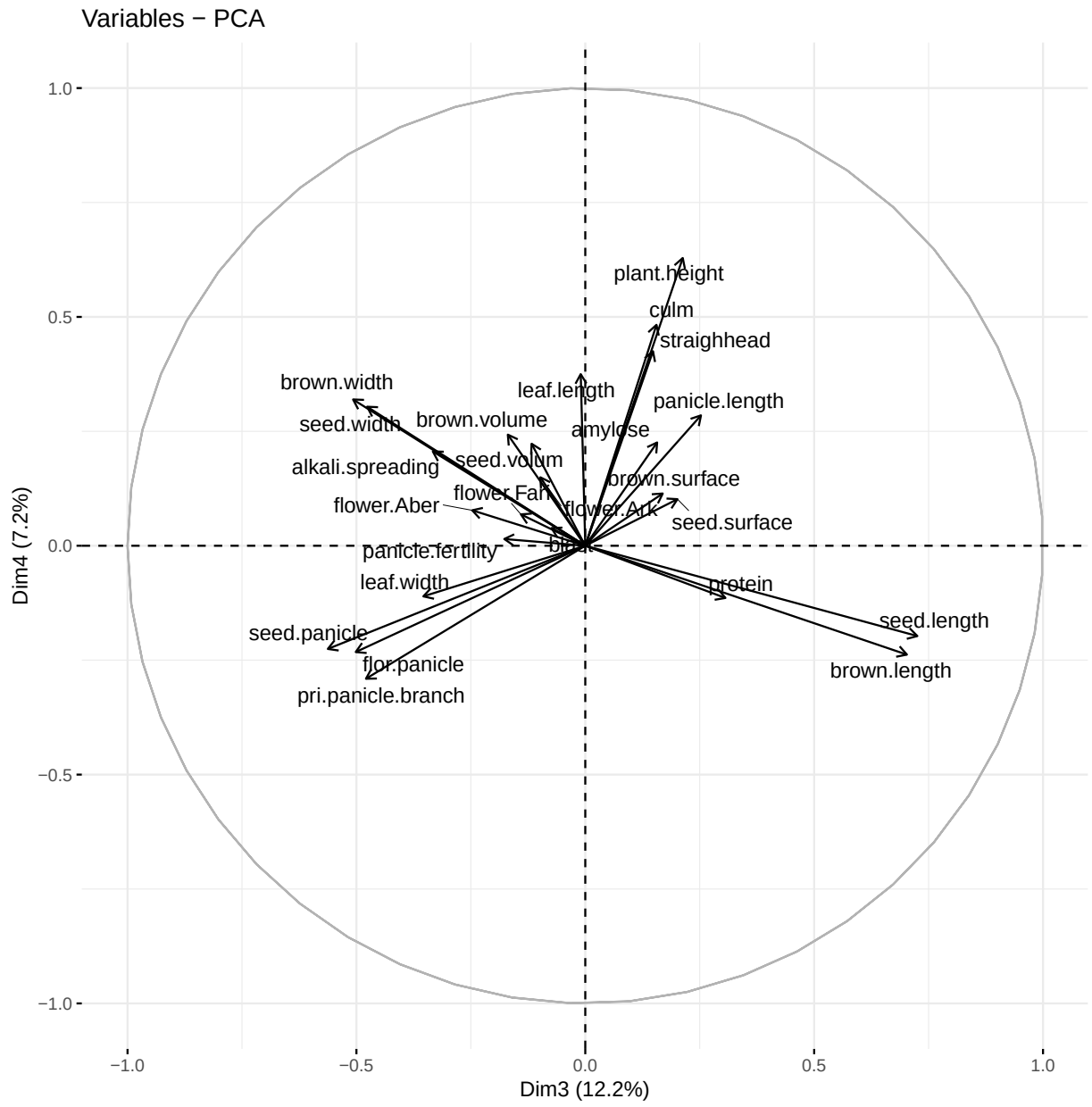
```
res.var <- get_pca_var(res)
corr <- res.var$cor
corrplot(res.var$cor, is.corr=FALSE)
```



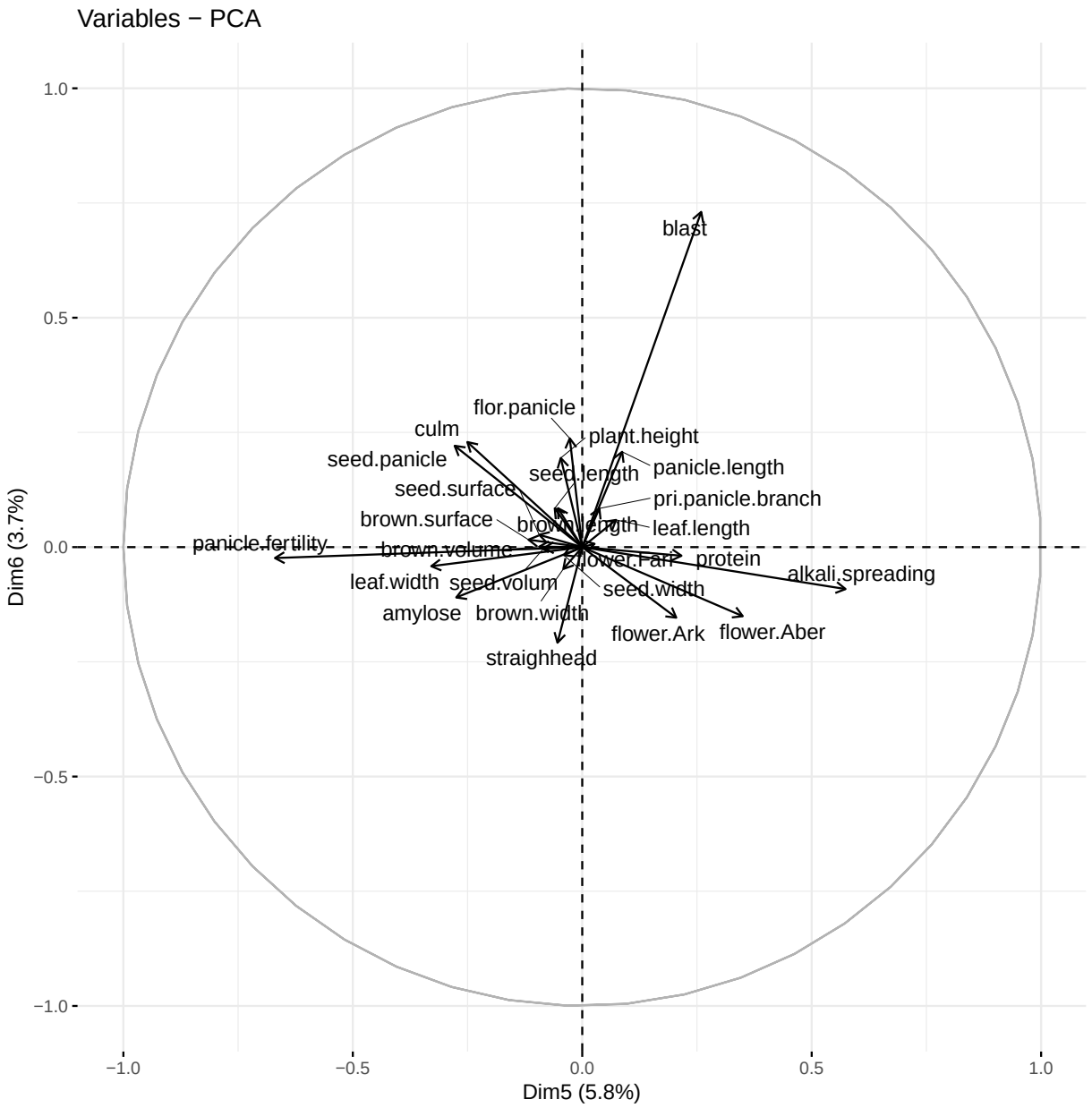

```
fviz_pca_var(res, axes = c(1,2), col.var = "black", repel = TRUE)
```




```
fviz_pca_var(res, axes = c(3,4), col.var = "black", repel = TRUE)
```




```
fviz_pca_var(res, axes = c(5,6), col.var = "black", repel = TRUE)
```




- The results show the correlation coefficient between the value of original variables and the principal component score. Most of the original variables had the strong correlation with first principal component score, meanwhile there are weak correlation with five and six principal component score.
- With first and second principal component scores, grain morphology group has the strong negative correlations with first principal component but there are strong positive correlations with the second principal component. In contrast, groups of plant morphology, flowering time have the strong positive correlations with both first and second principal component scores.
- There are the large numbers of original variables in three groups that have the negative correlations with the four principal component score. The grain morphology group, excepted Brown rice seed length and Seed length has the negative correlation with both third and four principal component scores.

### 2.2. Coordinates of variabels

```
corrd <- res.var$coord
corrd[,1:4]
```

```
##                          Dim.1       Dim.2       Dim.3       Dim.4
## flower.Aber          0.459085624 -0.48843310 -0.24703437  0.07743825
## flower.Ark           0.687234221 -0.50821202 -0.09821413  0.14960240
## flower.Fari          0.593182964 -0.31955060 -0.14063310  0.06859592
## culm                 0.201823586  0.49889361  0.15551705  0.48268014
## leaf.length          0.524636530 -0.45144444 -0.00998311  0.37502845
## leaf.width           0.320062080 -0.59531631 -0.35378530 -0.11058266
## plant.height         0.555861685 -0.10851459  0.21297018  0.62871486
## panicle.length       0.692590074 -0.27648221  0.25278048  0.28486699
## pri.panicle.branch   0.420059051 -0.47903416 -0.47918851 -0.29060071
## seed.panicle         0.597359431 -0.28878762 -0.56245690 -0.22547480
## flor.panicle         0.608757369 -0.42518203 -0.50149439 -0.23223792
## panicle.fertility   -0.004859004  0.35771094 -0.17694923  0.01470204
## seed.length         -0.015444437 -0.60211455  0.72554648 -0.19729256
## seed.width          -0.784661279 -0.21072264 -0.47665310  0.30376853
## seed.volum          -0.805501164 -0.51565874 -0.11734539  0.22288555
## seed.surface        -0.670085749 -0.68060318  0.20138092  0.10178065
## brown.length         0.068479767 -0.61466149  0.70267760 -0.23786349
## brown.width         -0.769702943 -0.14623188 -0.50730227  0.31961818
## brown.surface       -0.650093501 -0.70018406  0.16921042  0.11415164
## brown.volume        -0.792050710 -0.50591525 -0.16924518  0.24280468
## straighhead          0.632756307 -0.09855987  0.14790871  0.42505336
## blast               -0.405758528  0.22373230 -0.07390106  0.03859682
## amylose              0.605614854  0.13557565  0.15701535  0.22568001
## alkali.spreading    -0.039678410  0.23074714 -0.33312881  0.20613916
## protein             -0.382852895 -0.21839988  0.30624976 -0.11409225
```

### 2.3. Quality of representation

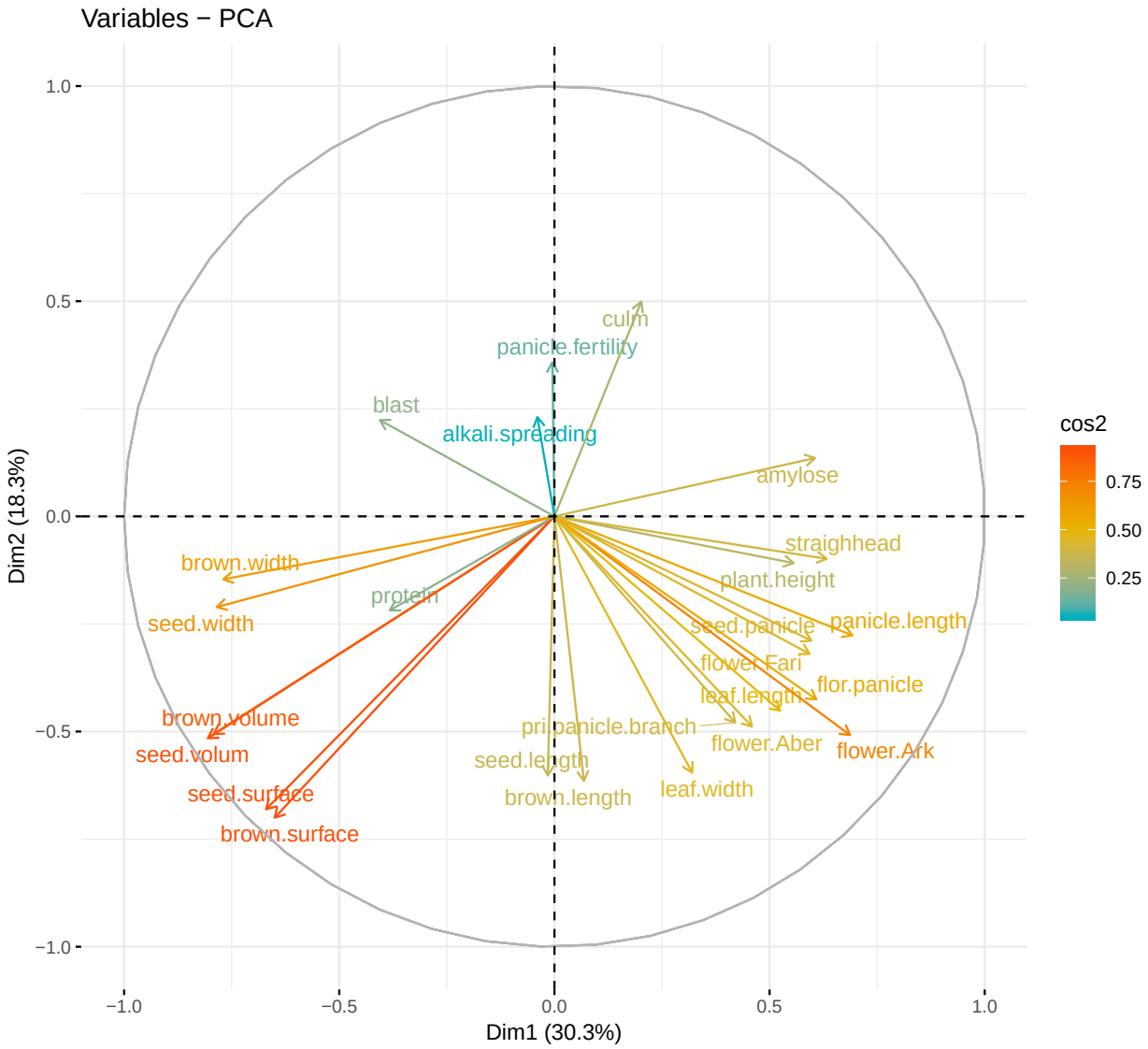
- Indicate the contribution of a component to the squared distance of the observation to the origin
- Components with a large value of cos2 contribute a relatively large portion to the total distance and therefore these components are importance for that observation
- The closer a variable is to the circle of correlation, the better its representation on the factor map (the more important it is to interpret components)

```
cos2 <- res.var$cos2
cos2[,1:4]
```

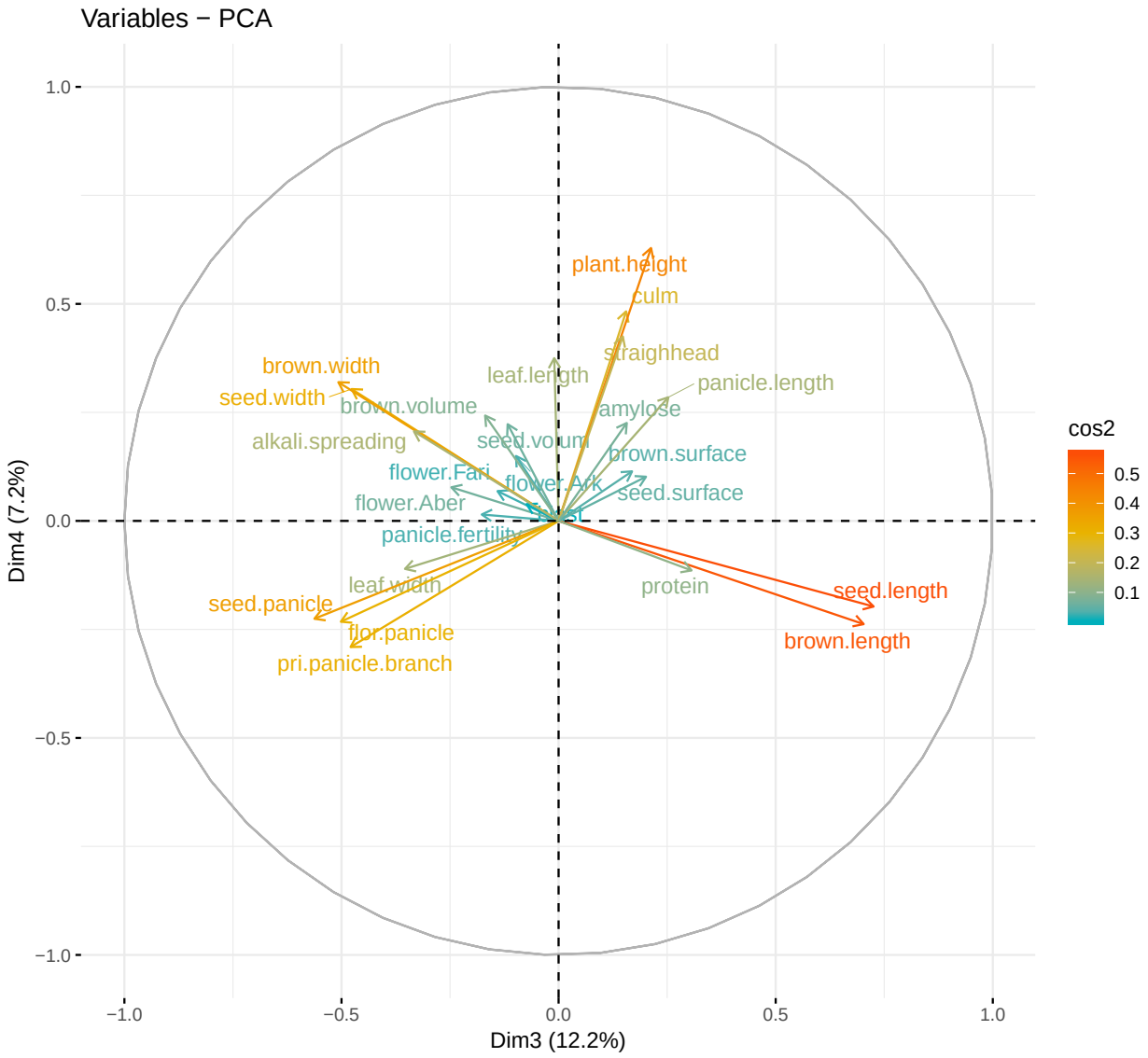
```
##                          Dim.1       Dim.2        Dim.3        Dim.4
## flower.Aber          2.107596e-01 0.238566892 6.102598e-02 0.0059966822
## flower.Ark           4.722909e-01 0.258279461 9.646016e-03 0.0223808769
## flower.Fari          3.518660e-01 0.102112587 1.977767e-02 0.0047054007
## culm                 4.073276e-02 0.248894835 2.418555e-02 0.2329801128
## leaf.length          2.752435e-01 0.203802082 9.966248e-05 0.1406463348
## leaf.width           1.024397e-01 0.354401505 1.251640e-01 0.0122285256
## plant.height         3.089822e-01 0.011775417 4.535630e-02 0.3952823766
## panicle.length       4.796810e-01 0.076442412 6.389797e-02 0.0811492028
## pri.panicle.branch   1.764496e-01 0.229473731 2.296216e-01 0.0844487732
## seed.panicle         3.568383e-01 0.083398291 3.163578e-01 0.0508388870
## flor.panicle         3.705855e-01 0.180779757 2.514966e-01 0.0539344512
```

```
## panicle.fertility  2.360992e-05 0.127957114 3.131103e-02 0.0002161498
## seed.length        2.385306e-04 0.362541930 5.264177e-01 0.0389243559
## seed.width         6.156933e-01 0.044404031 2.271982e-01 0.0922753172
## seed.volum         6.488321e-01 0.265903940 1.376994e-02 0.0496779696
## seed.surface       4.490149e-01 0.463220686 4.055427e-02 0.0103593003
## brown.length       4.689479e-03 0.377808749 4.937558e-01 0.0565790411
## brown.width        5.924426e-01 0.021383762 2.573556e-01 0.1021557803
## brown.surface      4.226216e-01 0.490257715 2.863217e-02 0.0130305959
## brown.volume       6.273443e-01 0.255950238 2.864393e-02 0.0589541125
## straighhead        4.003805e-01 0.009714048 2.187699e-02 0.1806703572
## blast              1.646400e-01 0.050056142 5.461367e-03 0.0014897142
## amylose            3.667694e-01 0.018380757 2.465382e-02 0.0509314682
## alkali.spreading   1.574376e-03 0.053244241 1.109748e-01 0.0424933525
## protein            1.465763e-01 0.047698507 9.378891e-02 0.0130170418
```

```
fviz_pca_var(res, axes = c(1,2), col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE)
```




```
fviz_pca_var(res, axes = c(3,4), col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE)
```




### 2.4. Contributions of the variables

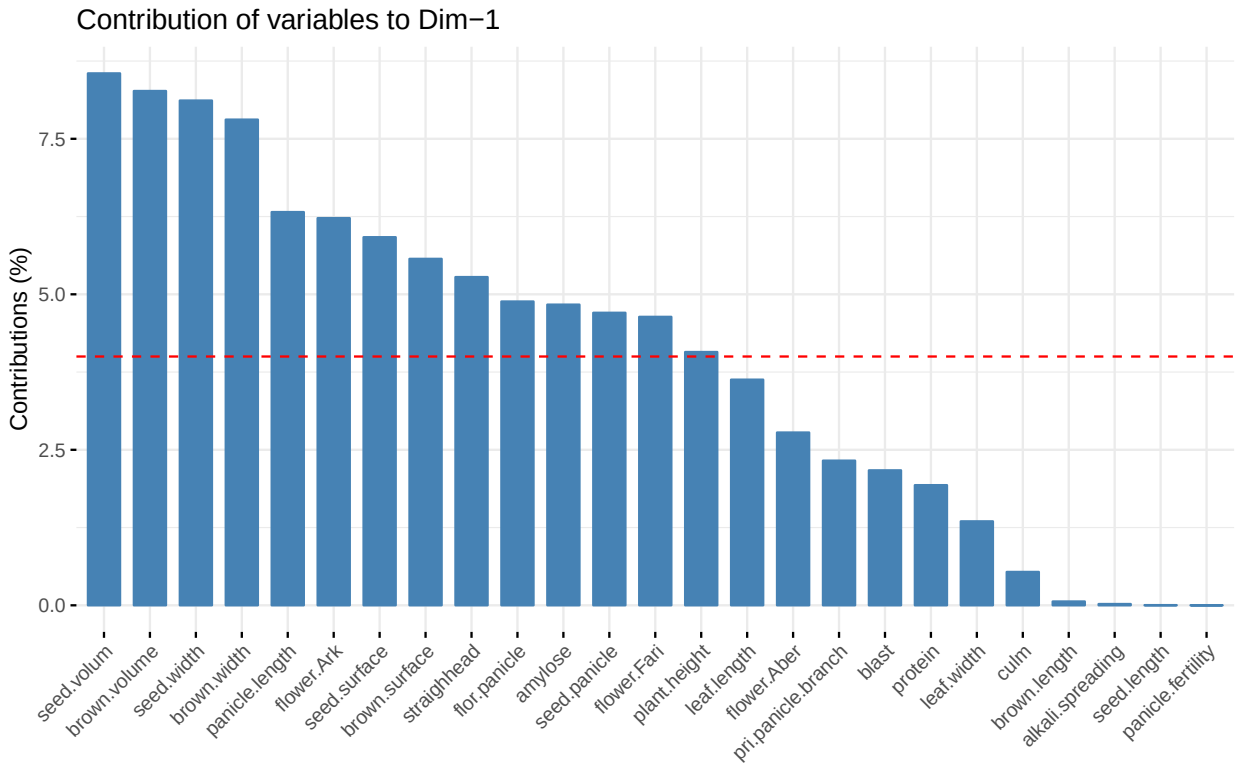
- Contributions of the variables in accounting for the variability in a given principal component are expressed in percentage
- Variables that are correlated with PC1 and PC2 are the most important in explaining the variability in the data set
- Variables that do not correlated with any PC or correlated with the last dimension are variables with low contribution and migh be removed to simplify the overall analysis

```
contrib <- res.var$contrib
contrib[,1:4]
```

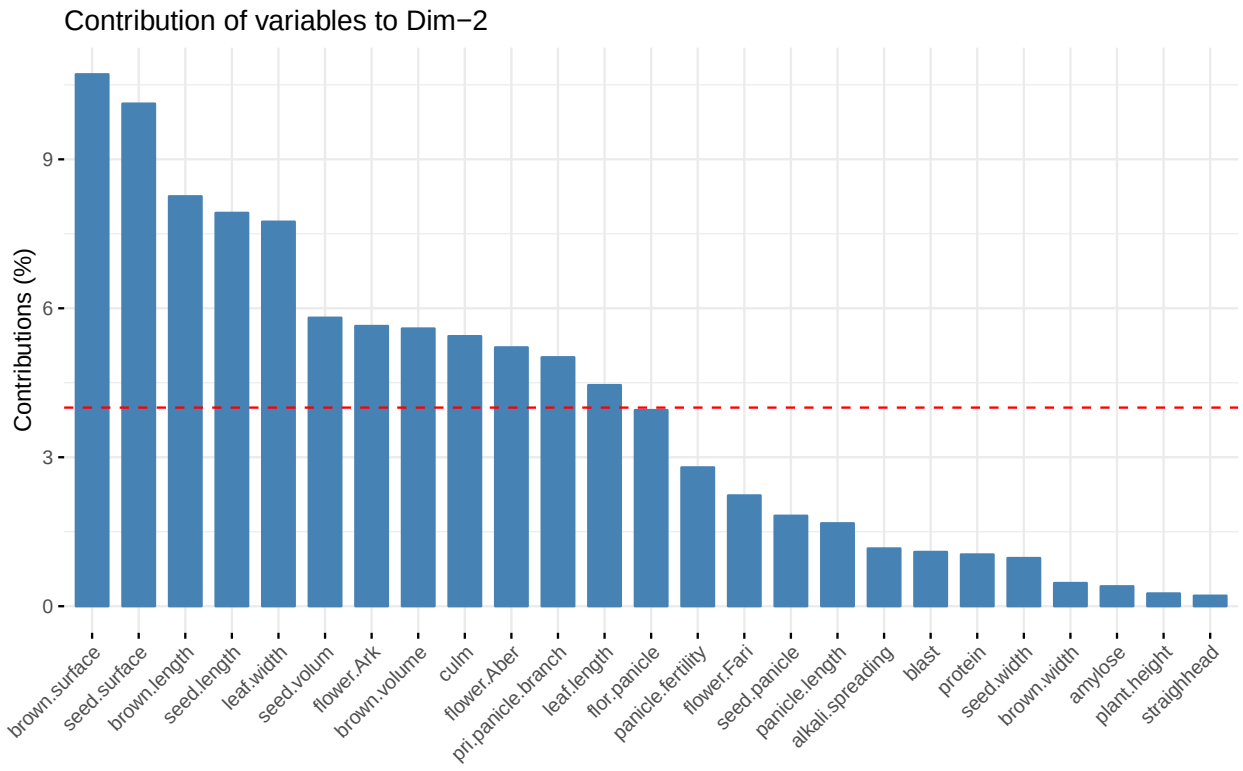
```
##                        Dim.1      Dim.2        Dim.3       Dim.4
## flower.Aber        2.7780105439  5.2129260  2.000180516  0.33400886
## flower.Ark         6.2252394069  5.6436654  0.316156703  1.24659120
```

```
## flower.Fari          4.6379262981  2.2312625  0.648230563  0.26208585
## culm                 0.5368962143  5.4386020  0.792702899 12.97674622
## leaf.length          3.6279689083  4.4532800  0.003266526  7.83385230
## leaf.width           1.3502523773  7.7440286  4.102362135  0.68111596
## plant.height         4.0726771371  0.2573047  1.486592709 22.01681088
## panicle.length       6.3226483687  1.6703434  2.094312581  4.51992489
## pri.panicle.branch   2.3257723219  5.0142313  7.526051827  4.70370747
## seed.panicle         4.7034653934  1.8223364 10.368905513  2.83167231
## flor.panicle         4.8846670420  3.9502191  8.243024342  3.00409197
## panicle.fertility    0.0003112011  2.7959914  1.026246638  0.01203932
## seed.length          0.0031440590  7.9219050 17.253805175  2.16804551
## seed.width           8.1154189783  0.9702726  7.446621023  5.13963771
## seed.volum           8.5522196829  5.8102680  0.451321978  2.76701044
## seed.surface         5.9184402410 10.1218369  1.329202198  0.57700208
## brown.length         0.0618117524  8.2555003 16.183283216  3.15139283
## brown.width          7.8089527836  0.4672567  8.435057236  5.68996907
## brown.surface        5.5705509657 10.7126231  0.938444592  0.72579043
## brown.volume         8.2689902163  5.5927696  0.938830179  3.28368180
## straighhead          5.2773933846  0.2122617  0.717037563 10.06314808
## blast                2.1701103393  1.0937769  0.179001123  0.08297551
## amylose              4.8343661555  0.4016380  0.808050790  2.83682899
## alkali.spreading     0.0207517645  1.1634401  3.637297405  2.36683485
## protein              1.9320144638  1.0422602  3.074014571  0.72503548
```

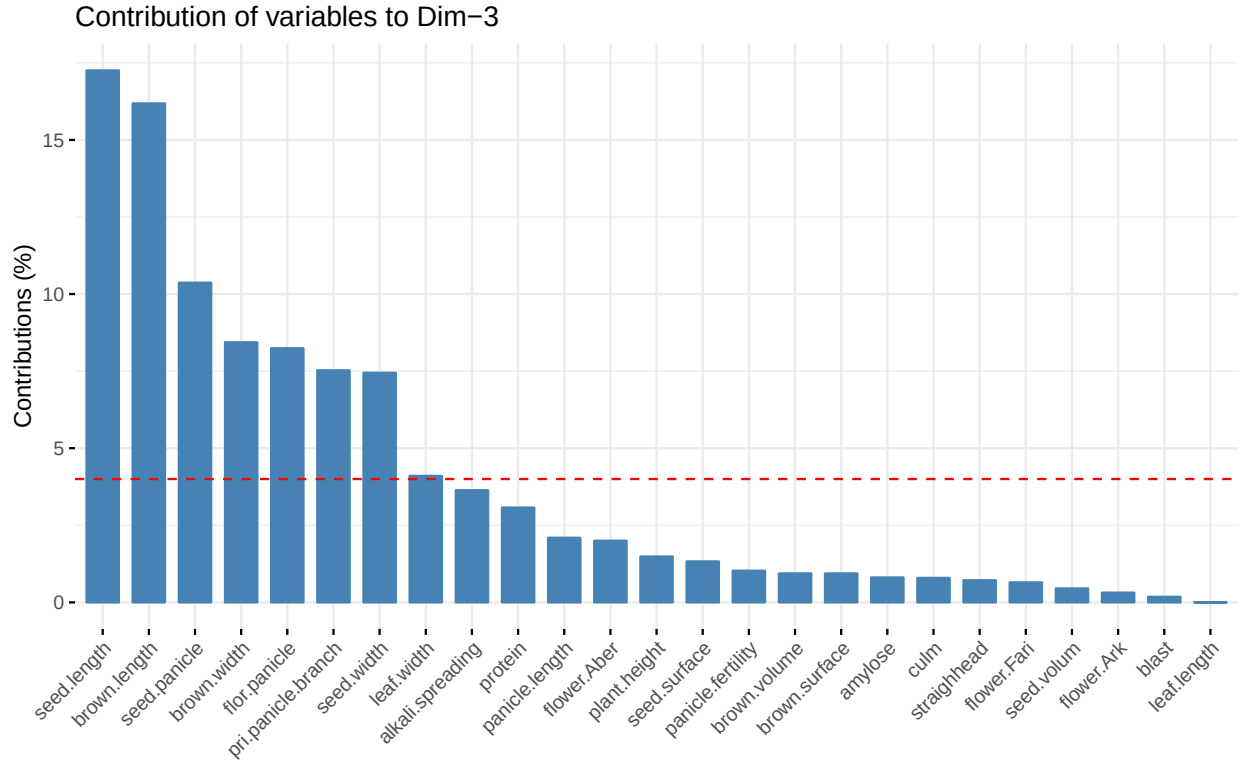
```
fviz_contrib(res, choice = "var", axes = 1)
```



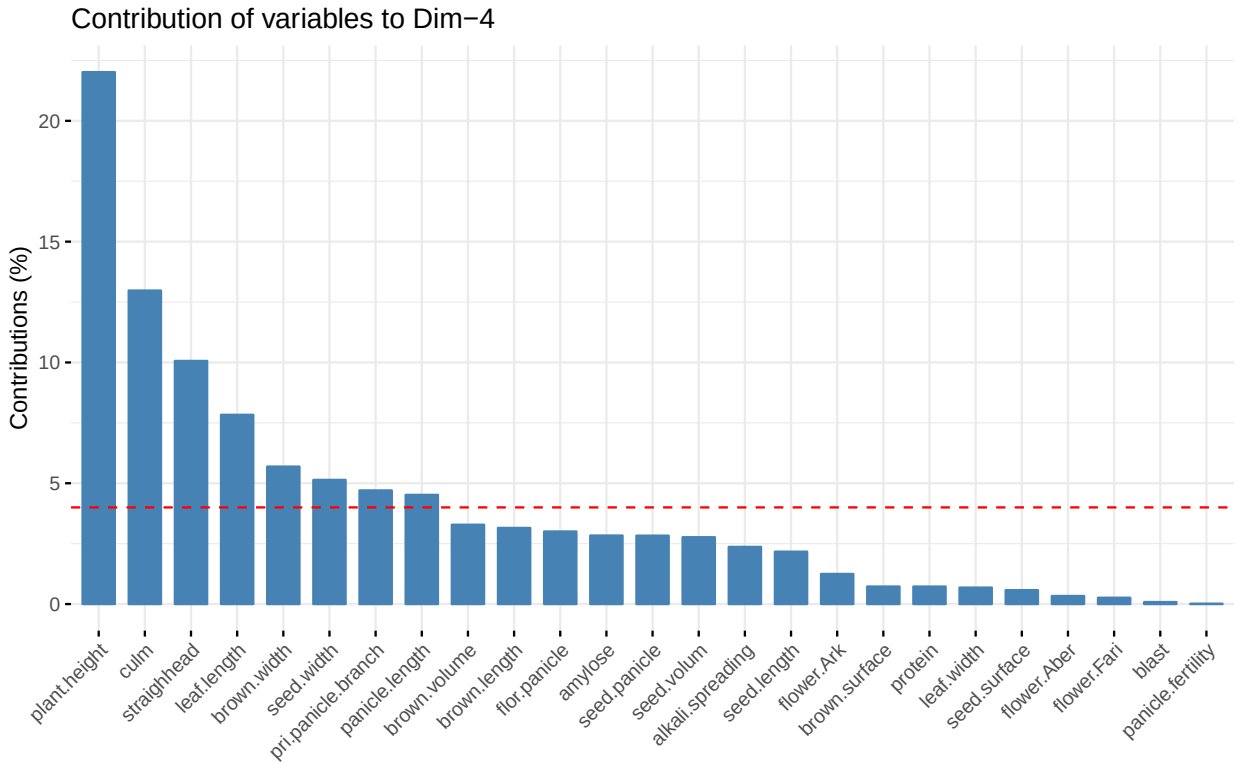

```
fviz_contrib(res, choice = "var", axes = 2)
```




```
fviz_contrib(res, choice = "var", axes = 3)
```

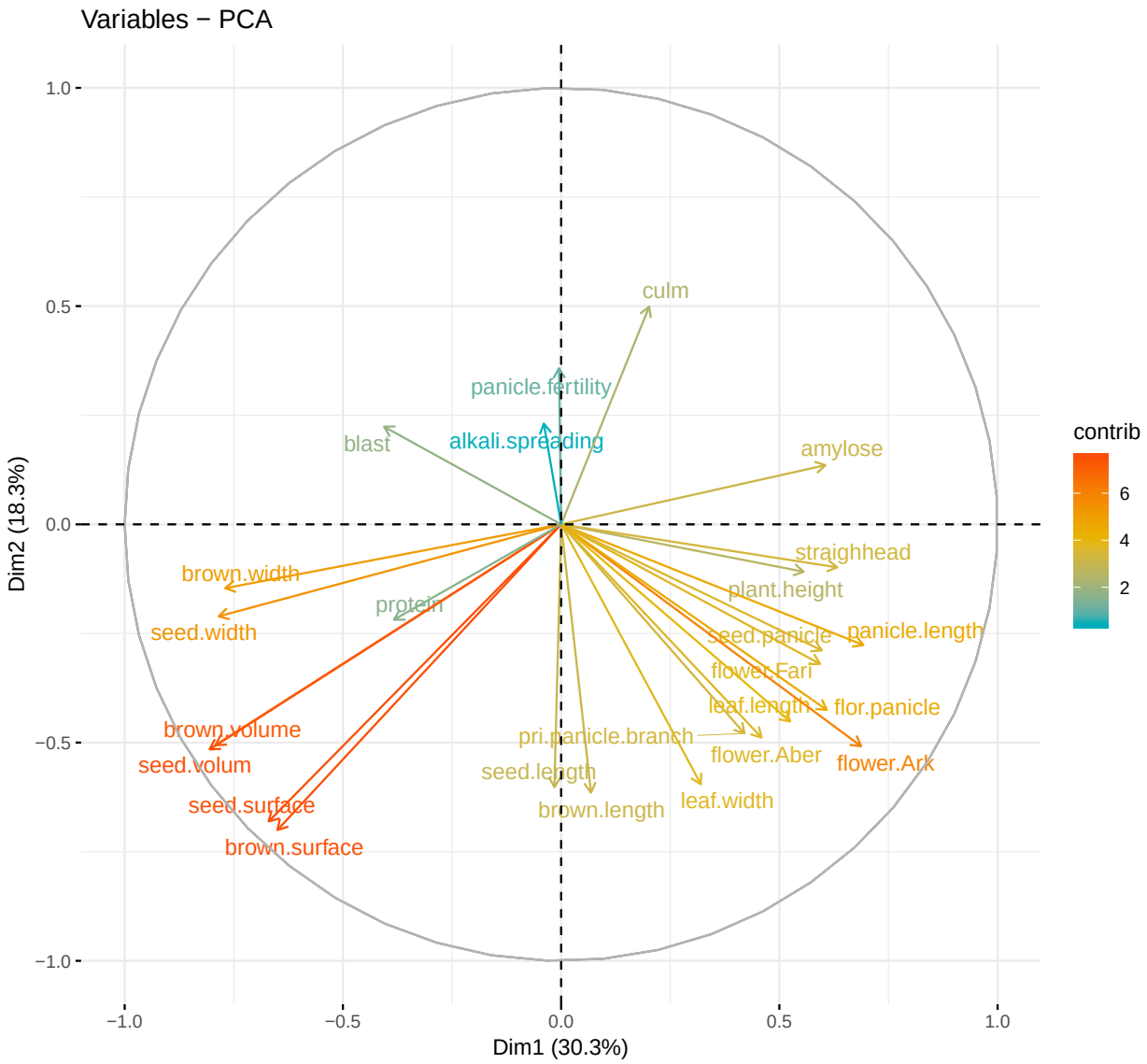


```
fviz_contrib(res, choice = "var", axes = 4)
```

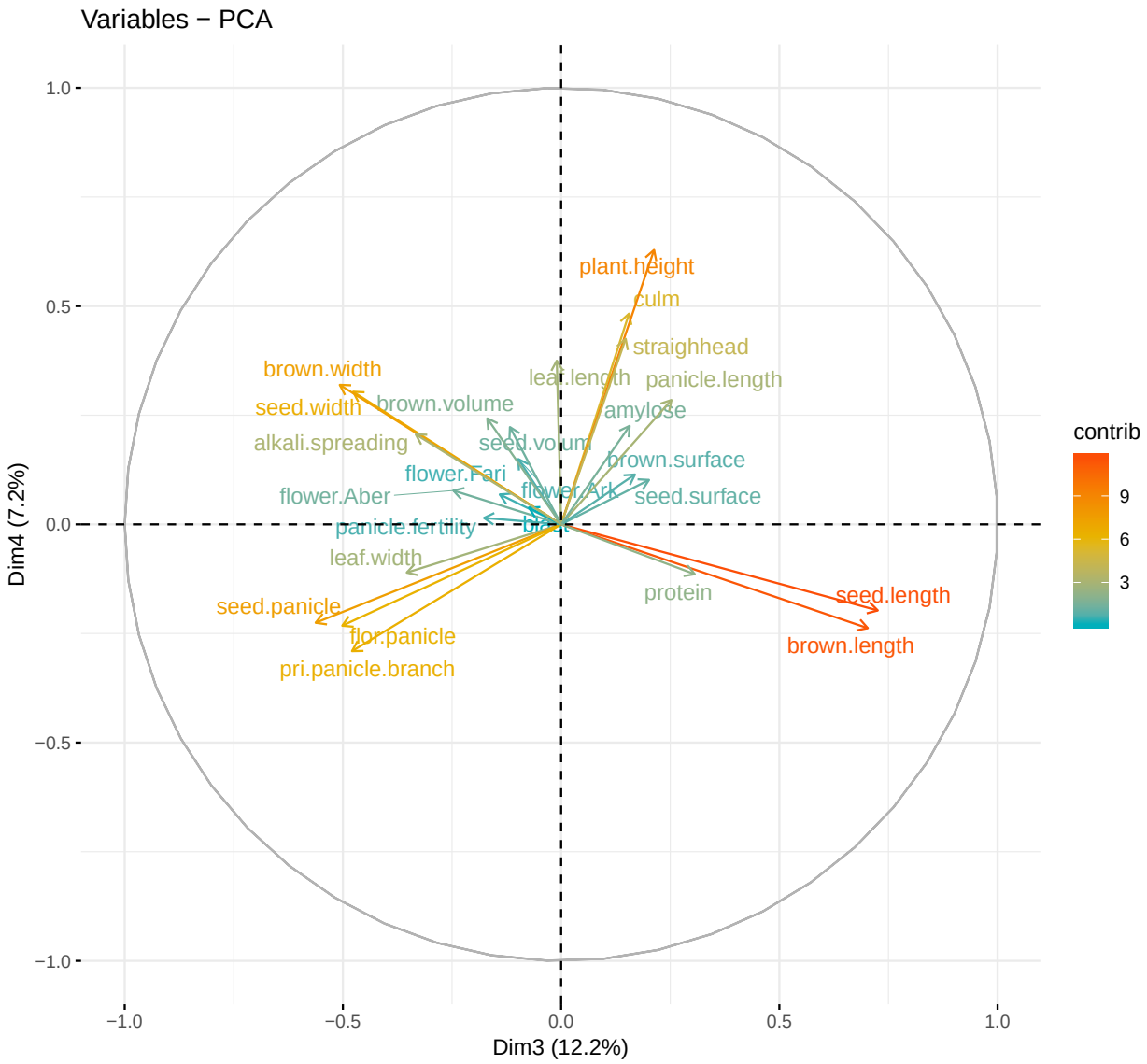



The red dashed line on the graph above indices the expected average contribution. If contribution of the variables were uniform, the expected value would be $1/length(variables) = 1/20$.

```
fviz_pca_var(res, axes = c(1,2), col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE)
```

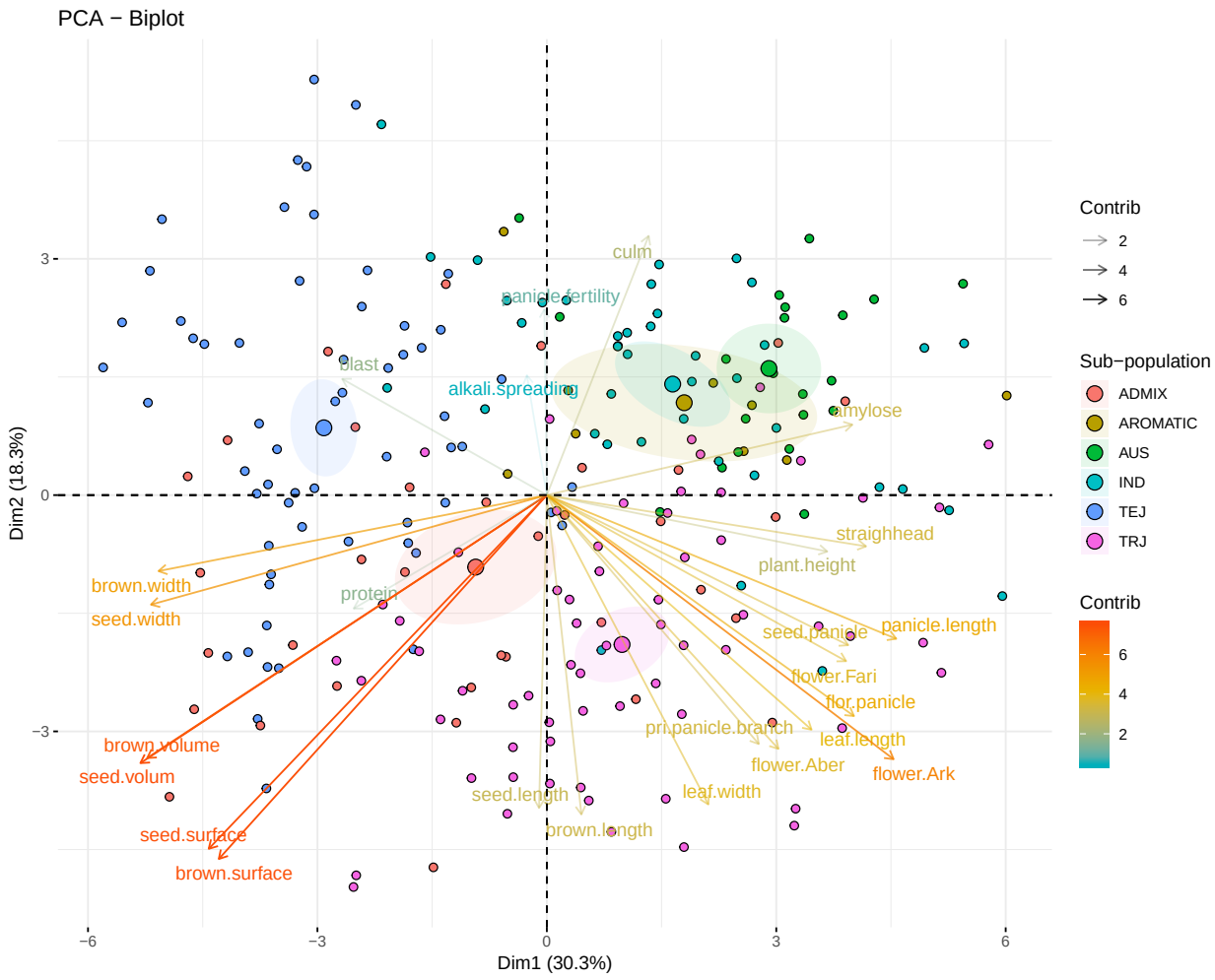



```
fviz_pca_var(res, axes = c(3,4), col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE)
```

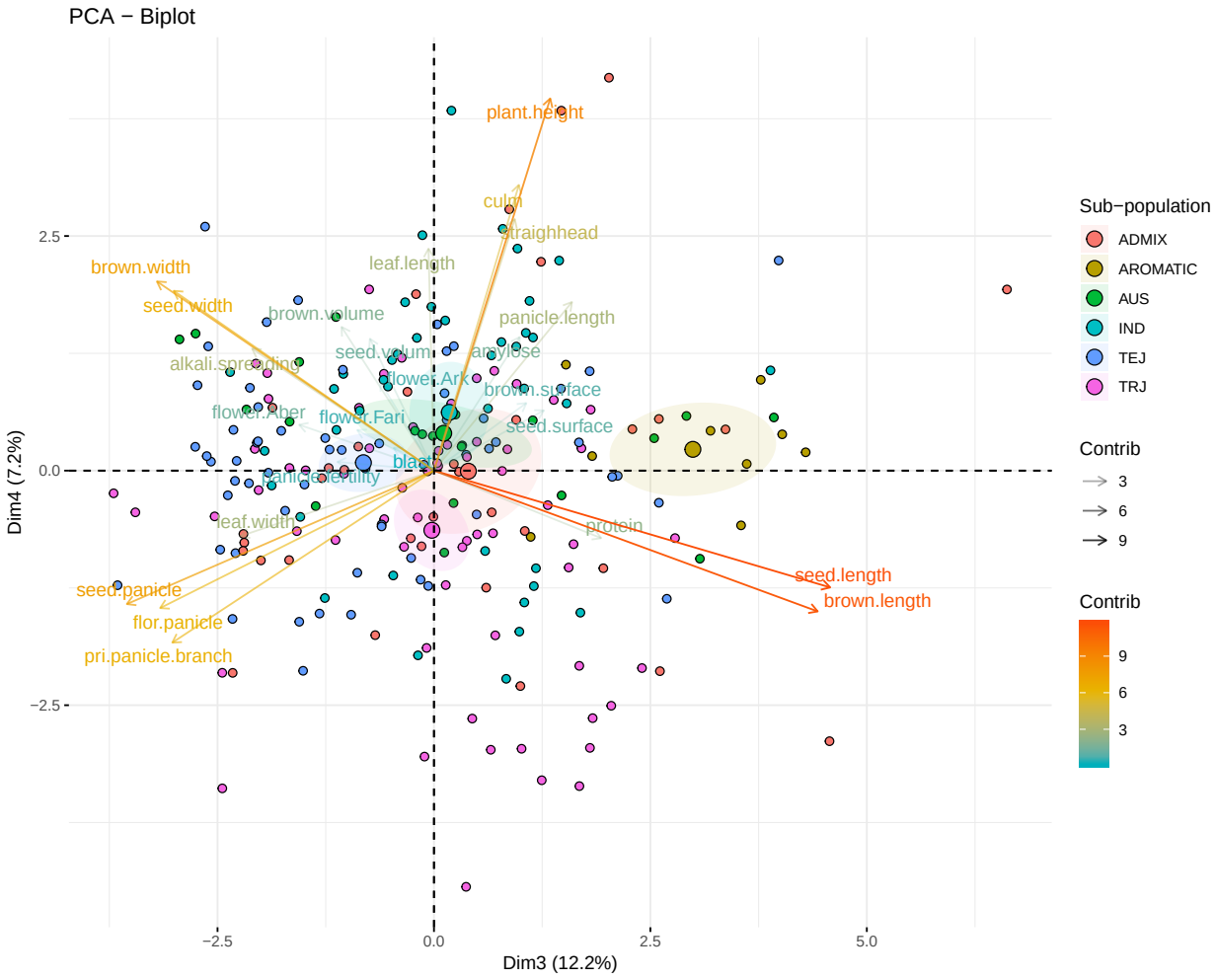



### 2.5. Biplot

```
fviz_pca_biplot(res, axes = c(1,2),
                # Individuals
                geom.ind = "point",
                fill.ind = subpop,
                pointshape = 21, pointsize = 2,
                addEllipses = TRUE, ellipse.type = "confidence",
                # Varibles
                alpha.var = "contrib", col.var = "contrib",
                gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE,
                legend.title = list(fill = "Sub-population", color = "Contrib",
                                    alpha = "Contrib"))
```




```
fviz_pca_biplot(res, axes = c(3,4),
                # Individuals
                geom.ind = "point",
                fill.ind = subpop,
                pointshape = 21, pointsize = 2,
                addEllipses = TRUE, ellipse.type = "confidence",
                # Varibles
                col.var = "contrib",alpha.var = "contrib",
                gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE,
                legend.title = list(fill = "Sub-population", color = "Contrib",
                                    alpha = "Contrib"))
```


PCA − Biplot
plant.height
culm
straighhead
leaf.length
brown.width
seed.width
brown.volume
panicle.length
seed.volum
amylose
alkali.spreading
flower.Ark
brown.surface
flower.Aher
flower.Fari
seed.surface
blast
panicle.fertility
leaf.width
protein
seed.length
brown.length
seed.panicle
flor.panicle
pri.panicle.branch
Dim4 (7.2%)
Dim3 (12.2%)
2.5
0.0
−2.5
−2.5
0.0
2.5
5.0
Sub−population
ADMIX
AROMATIC
AUS
IND
TEJ
TRJ
Contrib
3
6
9
Contrib
9
6
3