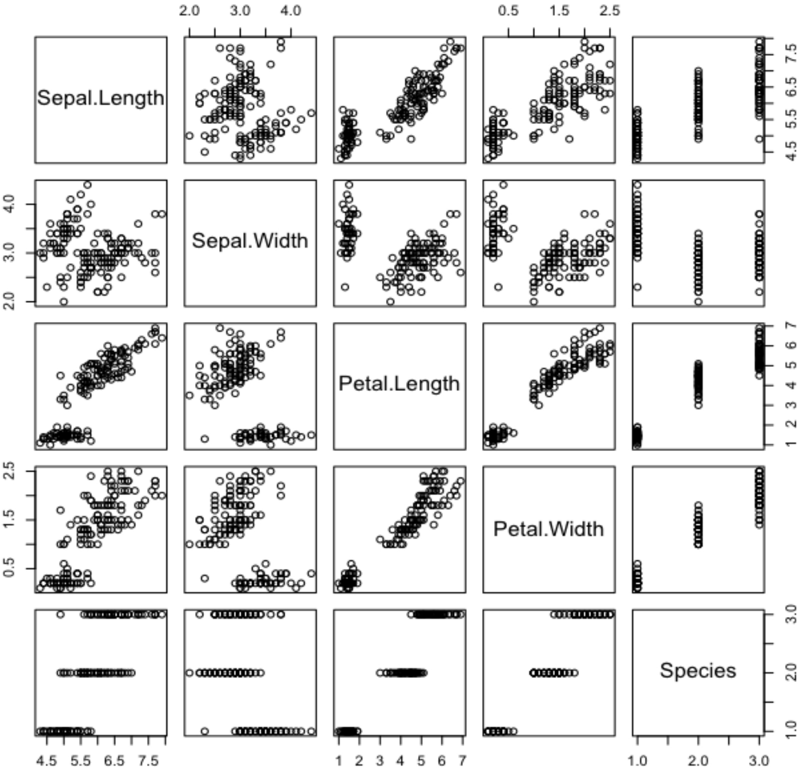
TheTrophicLink

{ 2014.03.28 }

Some notes on Principal Component Analysis

Principal Component Analysis (PCA) is an ordination method that reduces the dimensionality of multivariate data by creating few new key explanatory variables called principal components (PCs).
Each PC accounts for as much variance in the data as possible, provided that all the PAs are uncorrelated: therefore all PCs are independent and orthogonal.
It is possible to order the PCs according to the amount of total variation they explain, as well as to determine the relative contribution of each of the original variables to each PA.
A practical example follows using the software R on the "iris" dataset:

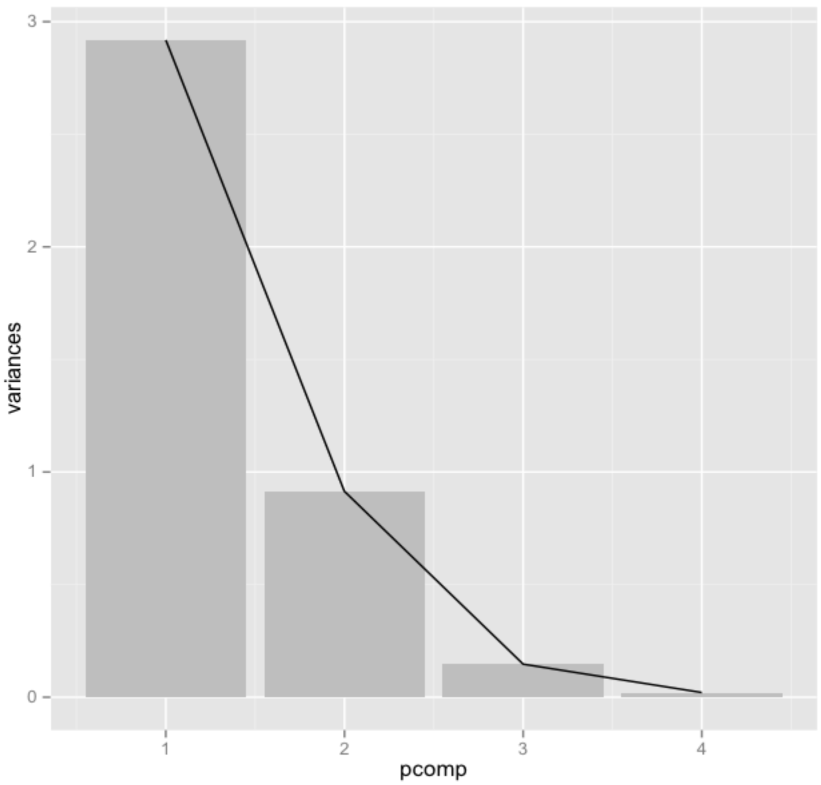
```
rm(list=ls()) # it's always good practice to clear R's memory

data(iris)
#?iris # gives info about the dataset
str(iris)
head(iris)
plot(iris)
```



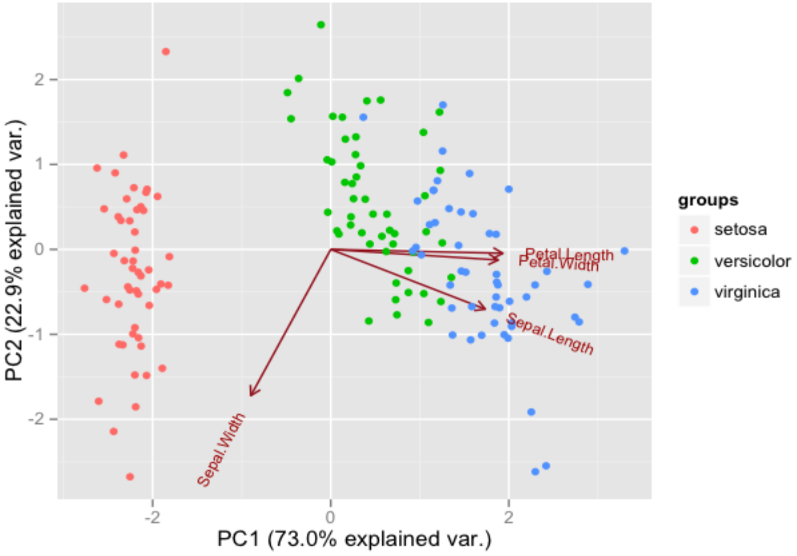
```
                                                                        # by plotting the variables against each other it becomes
that some are strongly correlated: in other words, there is an
overlap in the power of some variables at explaining/accounting for
the data variability. A PCA will help disentangling these correlations.

# functions prcomp() performs PCA:
fit<-prcomp(iris[-5], scale=TRUE)
# scale=T standardizes the variables to the same relative scale,
avoiding some variables to become dominant just because of their
large measurement units.

summary(fit)
# the summary indicates that four PCs where created: the number
of possible PCs always equals the number of original variables.

# PC1 and PC2 explain respectively ~73% and ~23% of the data's
total variability, summing up to a more-than-respectable 96% of
the total variability. There is no fixed rule about this, but
this already tells us that all the other PCs can be ignored as
they explain only crumbs of the total data variability.

plot(fit,type="lines")
# a "scree plot" allows a graphical assessment of the relative
contribution of the PCs in explaining the variability of the data.

fit[2] # the "Rotation" matrix contains the "loadings" of each
of the original variables on the newly created PCs.
The concept of eigenvalue would require to be introduced for
understanding how the loadings are estimated, and in general
for a quantitative understanding of how the principal
components are calculated: the interested reader might look
it up in Ref. 2 and 3.

biplot(fit)
# the arrows provide a graphical rendition of the
loadings of each of the original variables on the used PCs.

# Package GGPLOT2 and its derivative GGBIPLOT have a somehow esoteric
```

syntax, but they produce much fancier plots:

```
library(ggplot2) # this might need to be installed

# Variances of principal components
variances <- data.frame(variances=fit$sdev**2, pcomp=1:length(fit$sdev))
# **2 means ^2

#Plot of variances
varPlot <- ggplot(variances, aes(pcomp, variances))
+ geom_bar(stat="identity", fill="gray") + geom_line()
varPlot
```



```
# run these two lines only if ggbiplot is not installed ye
#library(devtools)
#install_github("ggbiplot", "vqv")
# load ggbiplot
library(ggbiplot)

Species<-iris$Species
iris_pca <- ggbiplot(fit,obs.scale = 1,
         var.scale=1,groups=Species,ellipse=F,circle=F,varname.size=3)
# iris_pca <- iris_pca + theme_bw() # try this for a less
ink-demanding background
iris_pca
```



We can see that petal width and length are highly correlated and their variability across the three Iris species is accounted mainly by PC1, which also explains a large part of variability in sepal length. This would be good to know if we wanted to study the correlation between floral elements and some other variable, say water availability, or the size or relative abundance of a certain pollinator: instead of measuring the correlation between this variable and all the measured response variables separately, we can simply use PC1 and PC2, which are syntetic and uncorrelated to each other. Other uses of PCA include, in pollination ecology, the identification of floral syndromes, namely common sets of floral traits evolved to suit a certain group of pollinators: in our case study, the two PC derived from the size of petals and sepals succeed in identifying the three Iris species and PC1 is enough to discern Iris setosa from the other two species very clearly, possibly suggesting some kind of floral specialization that might be related to pollination strategies (e.g. avoiding shared pollinators with the other two species).

UPDATE (04/11/2013): Also see "An introduction to applied multivariate analysis with R" by Everitt and Hothorn, Springer 2011: http://www.springer.com/statistics/statistical+theory+and+methods/book/978-1-4419-9649-7. Thanks to Torsten Hothorn for his help!

REFERENCES

1. Gotelli and Ellison, "A Primer of Ecological Statistics"
2. Manly, "Multivariate Statistical Methods"
3. Pielou, "The Interpretation of Ecological Data"
4. http://www.instantr.com/2012/12/18/performing-a-principal-component-analysis-in-r/
5. http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf
6. http://www.vince.vu/software/
7. http://www.statmethods.net/advstats/factor.html

[reposted from www.marcoplebani.com]

Like 0

Edit this entry.

Posted by Marco Plebani on Friday, March 28, 2014, at 8:52 am. Filed under Figure / graph.

Follow any responses to this post with its comments RSS feed. Comments are closed, but you can trackback from your blog.