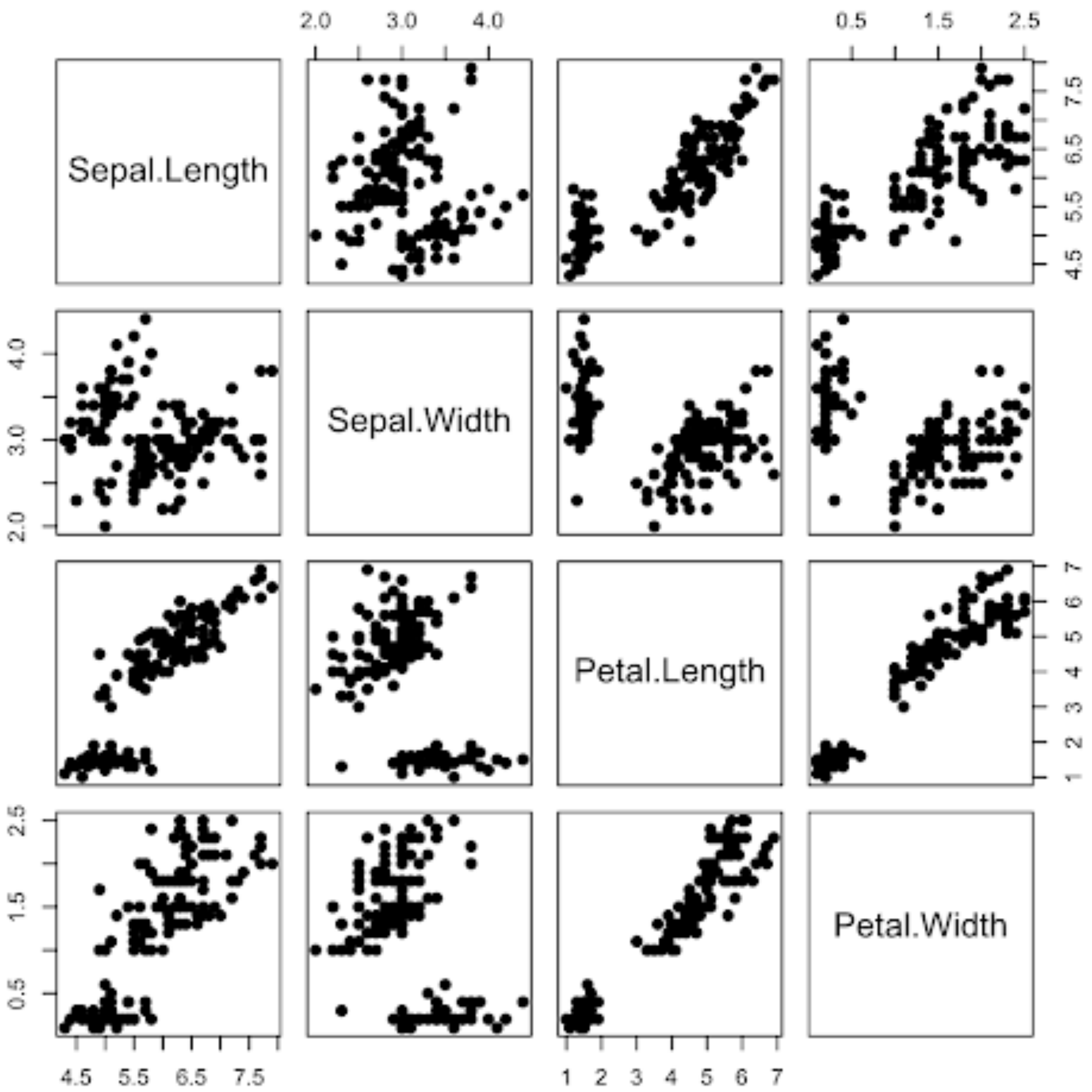
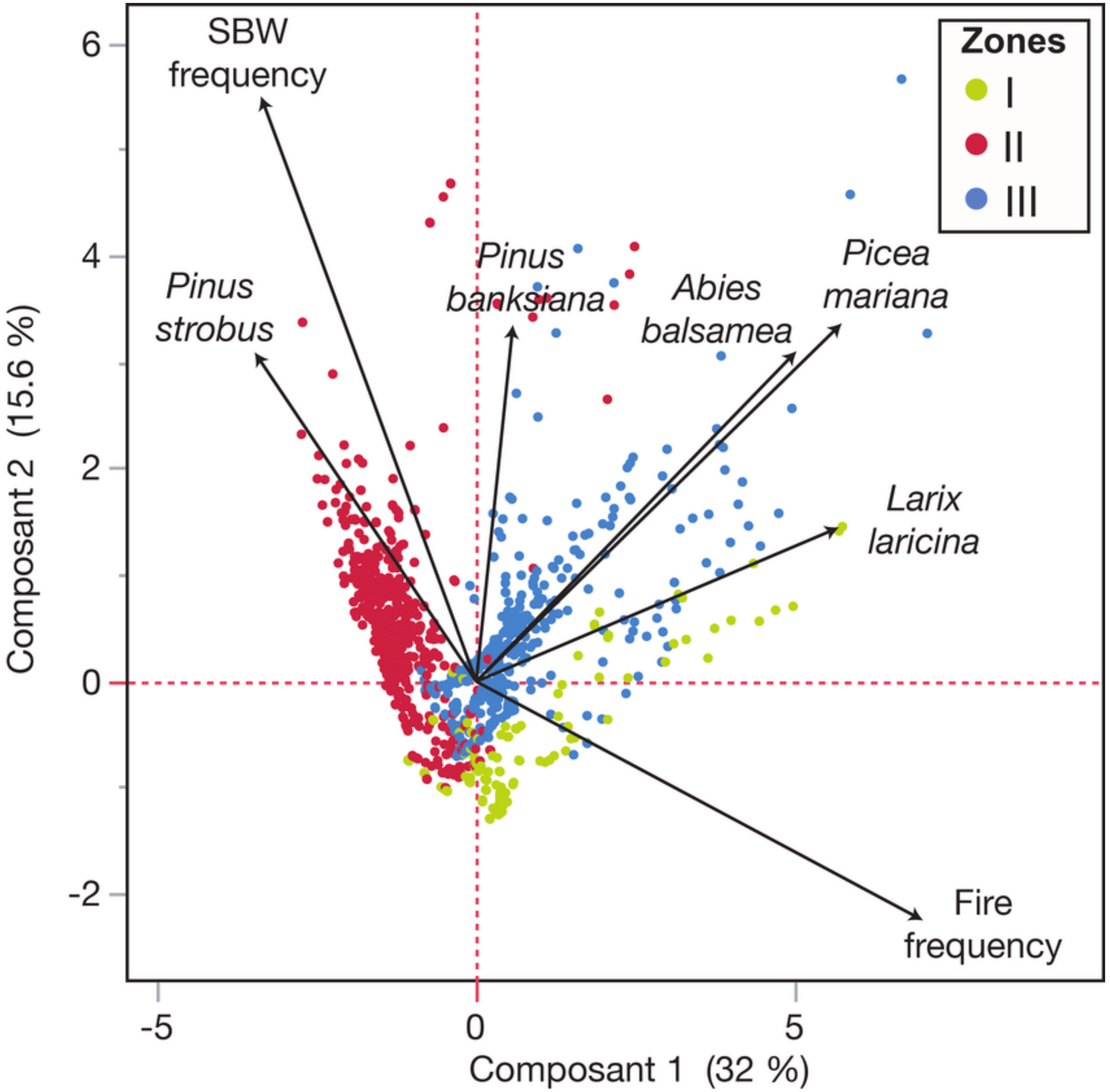
Foundational Statistics

Light intro. to multivariate statistics and Principal Component Analysis (PCA)





What is multivariate statistics?

- **General** - more than one variable recorded from a number of experimental sampling units

- **Specific** - two or more response variables that likely covary

What is multivariate statistics?

- Goals of multivariate statistics
  - testing the effects of a <u>factor</u> on linear combinations of variables (**MANOVA** and **DFA**)
  - Data reduction and simplification (**PCA** and **PCoA**)
  - Organization of objects (**Cluster Analysis** and **MDS**)

What is multivariate statistics?

- i = 1 to n **objects** and j = 1 to p **variables**
- Measure of center of a multivariate distribution = the **centroid**
- Multivariate statistics uses **analysis** of either matrices of **covariances of variables (p-by-p)**, or **dissimilarities of objects (n-by-n)**
- **Matrix and linear algebra** form the mathematical basis of multivariate statistics

Centroid

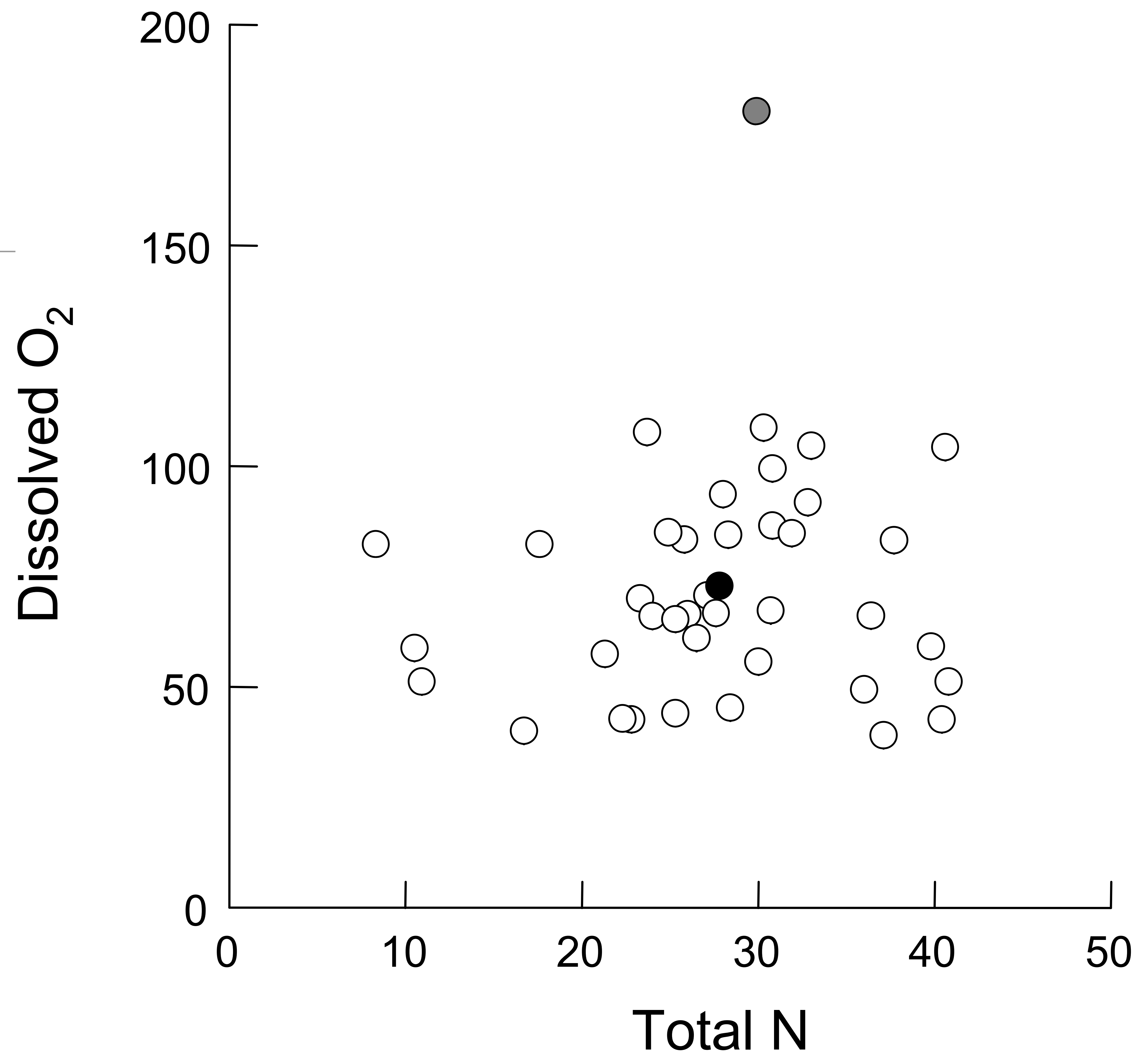


Figure 15.1 Scatterplot of dissolved oxygen against total nitrogen for 39 streams from Lovett *et al.* (2000). The centroid, the point represented by the mean of dissolved oxygen and total nitrogen, is filled. In this example, one object (grey fill) is an outlier for dissolved oxygen and also a multivariate outlier.

Some terminology associated with multivariate approaches:

- Ordination: Arrangement and visualization of observations from a dataset in a space of reduced dimensionality.

- Constrained vs. Unconstrained Ordination

Some terminology associated with multivariate approaches:

- A few points from ordination.okstate.edu, about ordination in ecology:

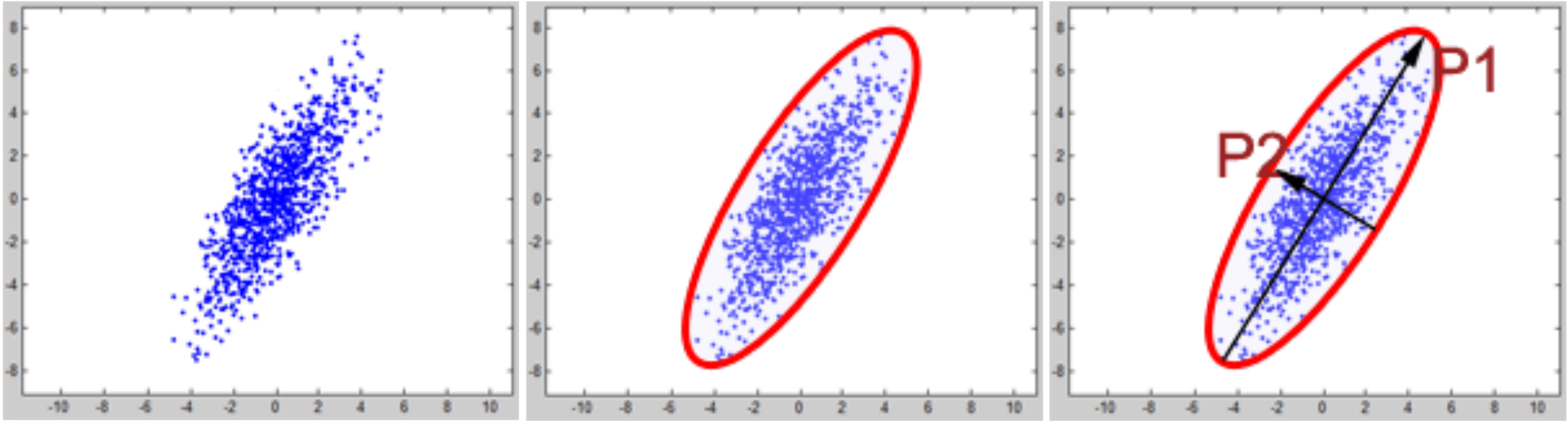
  \

  - A single multivariate analysis can save time
  - Major dimensions often explain a lot of the variation and are linked to ind. vars.
  - Can avoid interpreting “noise”

Two primary aims of Principal Component Analysis (PCA)

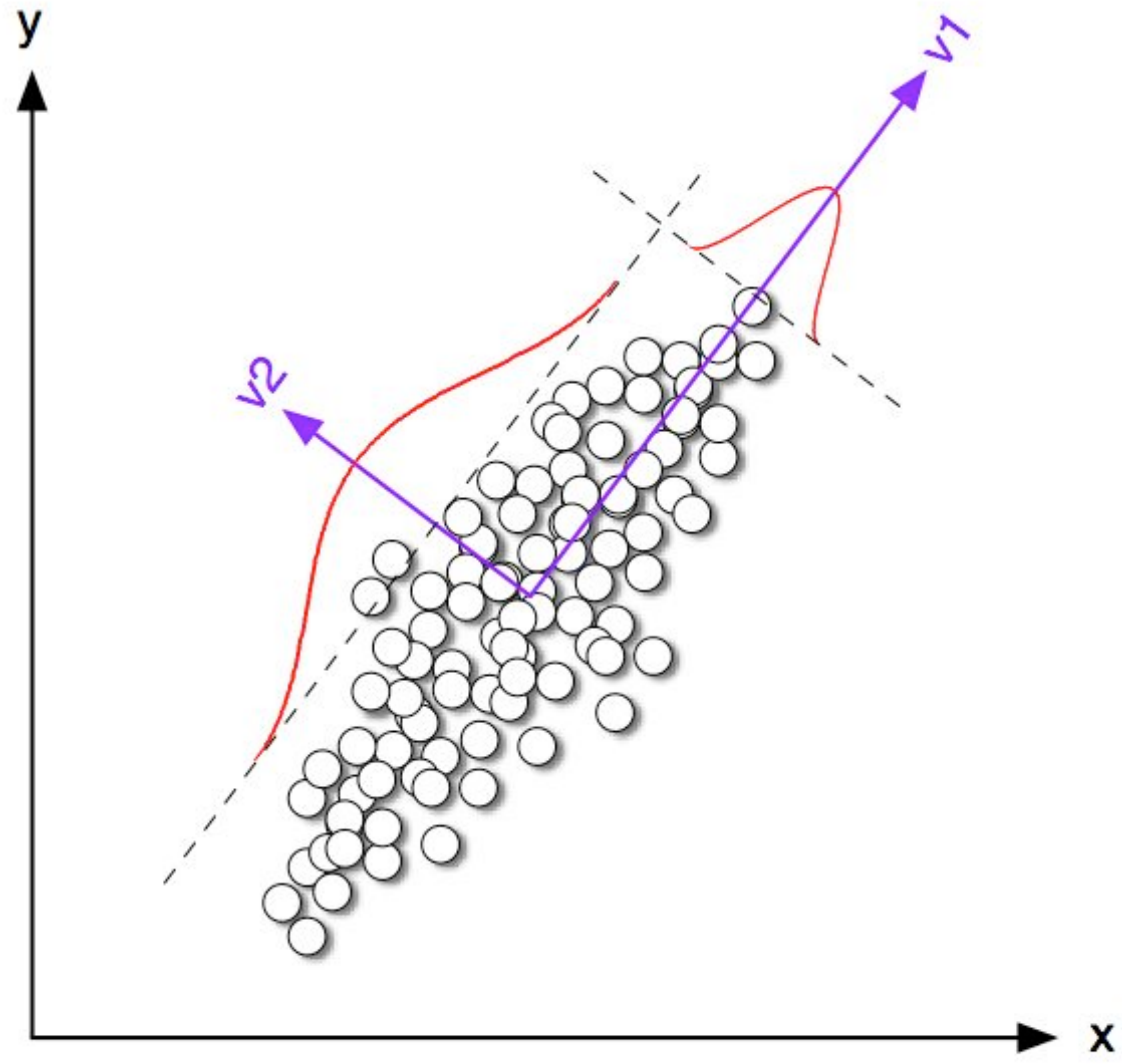
- 1) **Variable reduction** - reduce a lot of variables to a smaller number of new derived variables that adequately summarize the original information (aka **data reduction**).
- 2) **Multidimensional scaling** - Reveal patterns in the data - especially among variables or objects - that could not be found by analyzing each variable separately. A good way to do this is to plot the first few derived variables (this is generally called an **ordination**).
- **General approach** - use eigenanalysis on a large dataset of continuous variables.

Running a PCA analysis



- Start by ignoring any grouping variables
- Perform Eigenanalysis on the entire data set
- After the ordination is complete analyze the objects in the new ordination
- This includes ANOVA using the newly derived PC's and any grouping variables
- LDA is often called '**constrained**' and PCA '**unconstrained**'

Running a PCA analysis



Deriving Principal Components

- $i = 1$ to n objects and $j = 1$ to p variables
- PCA transforms into $k = 1$ to p new uncorrelated variables (z_1, z_2, ... , z_p) or "axes."
- The numeric values for each object on each component are often called "**PC scores.**"
- The first new axis explains the **majority** of the variation, the second uncorrelated axis explains the second most variation, and so on through the rest of the p new variables.
- The fitting of axes can be thought of as fitting lines through the **longest remaining** axis of the multi-dimensional cloud of points, sequentially.

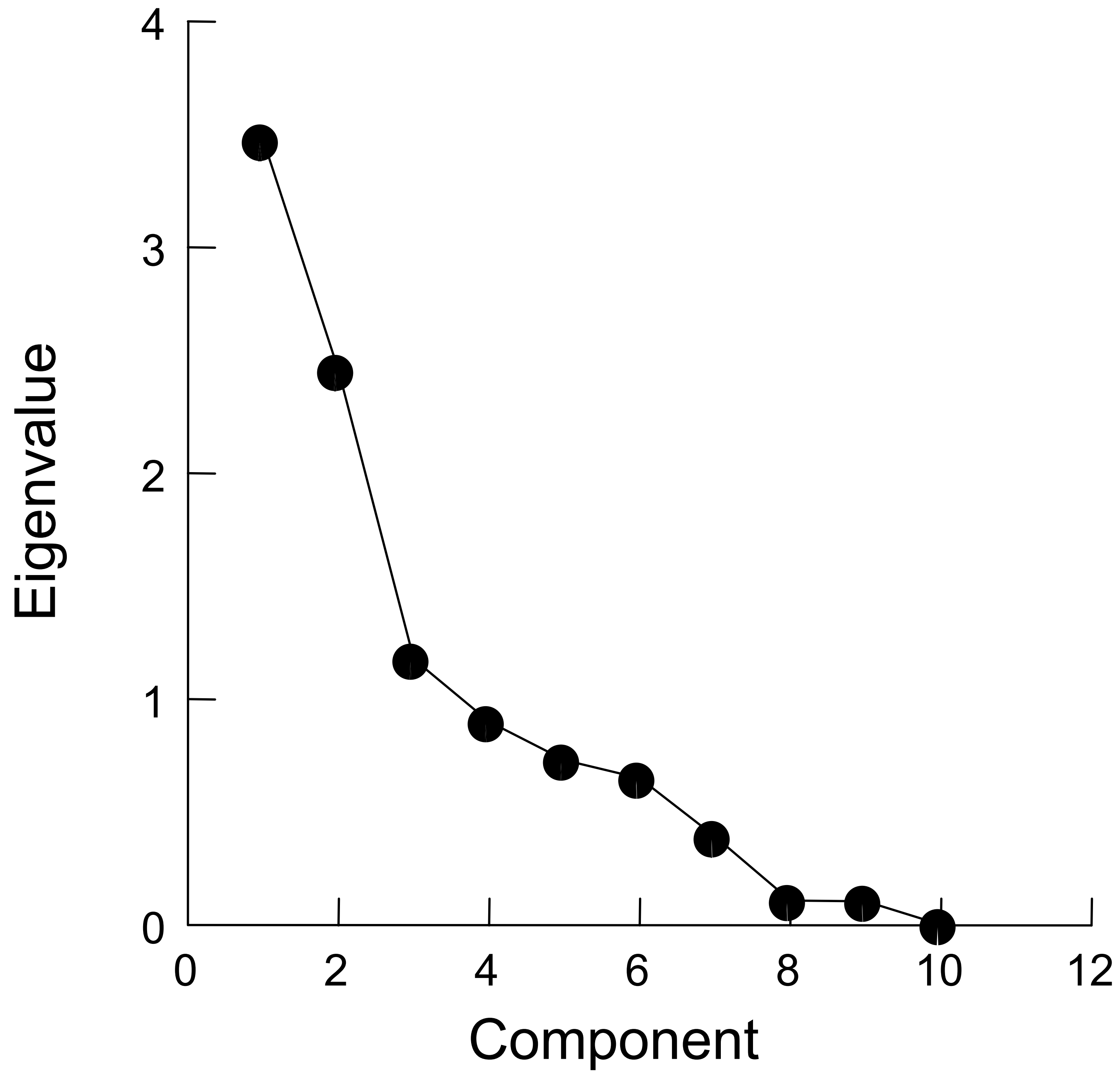
Data standardization

- **Covariances** and **correlations** measure the **linear** relationships among variables and therefore **assumptions** of **normality** and **homogeneity** of **variance** are important in multivariate statistics
- **Transformations** to achieve linearity might be important
- Data on very **different scales** can also be a problem
- **Centering** the data subtracts the mean from all variable values so the mean becomes zero
- **Ranging** divides each variable by its standard deviation so that all variables have a mean of zero and a unit s.d.
- In this case converting to presence or absence (binary) data might be the most appropriate

OR -
choose correlation matrix over var/covar matrix

How many PCs should I retain??

- The full, original variance-covariance pattern is encapsulated in **all** PCs.
- PCA will extract the **same number** of PCs as original variables.
- How many to retain? - really just a question of what to pay attention to.
  - An eigenvalue is the variance explained by a PC
  - **Scree plot** shows “diminishing returns”
  - Consider PCs that capture most of the original variance
- Most of the time, **first few PCs** are enough
- If not, PCA might not be appropriate!



How do I interpret the PCs??

- **High loadings** indicate that a variable is **strongly correlated** with a particular component (can be either **positive** or **negative**).
- The **loadings** and the **coefficients** will show a similar patterns, but different values.
- Ideally, we would like to have a few variables **load strongly on each of a few components**.

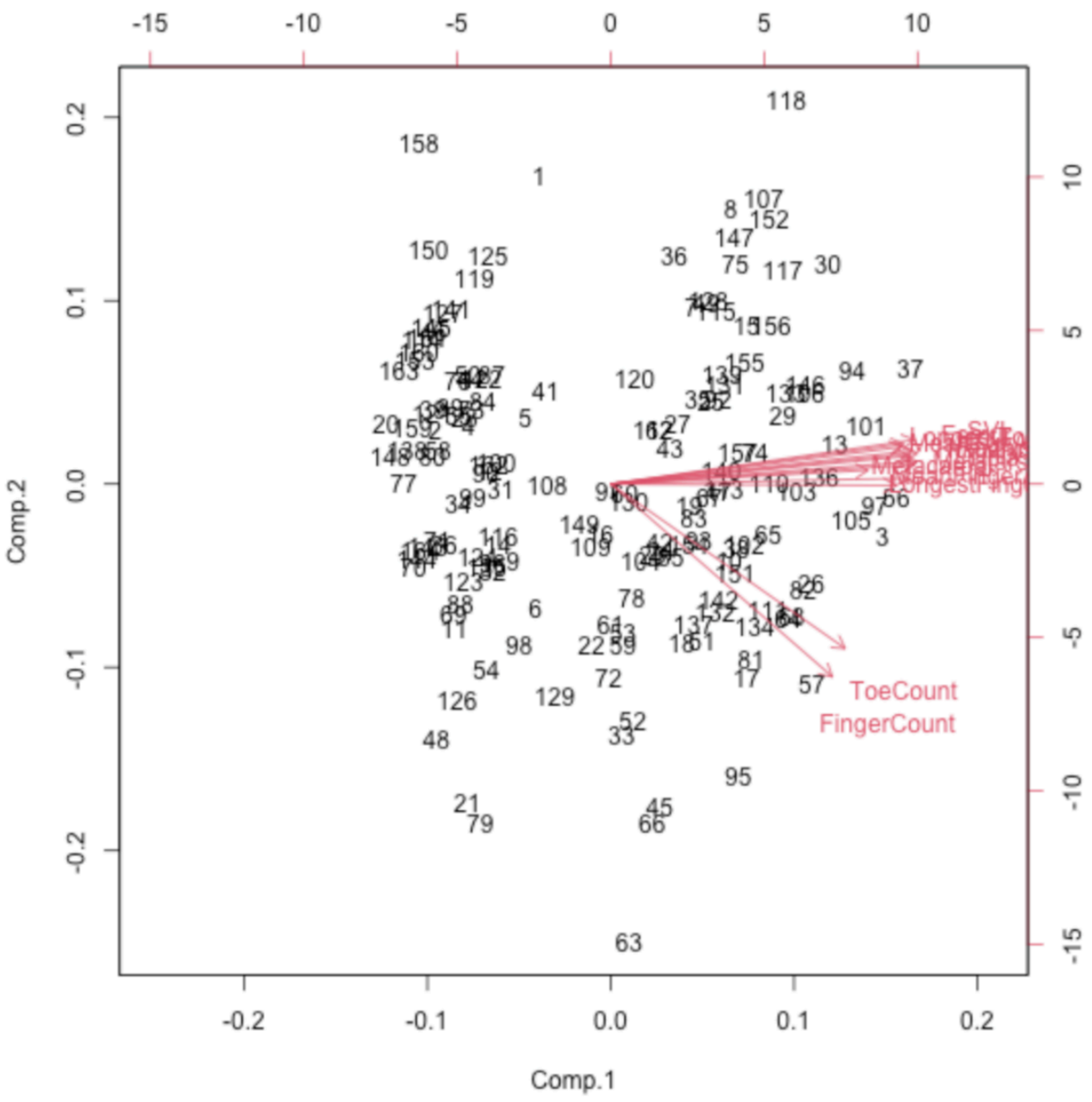
Variable	Component 1	Component 2	Component 3
NO_3	−0.483	−0.816	0.053
Total organic N	0.272	0.471	0.557
Total N	−0.423	−0.802	0.166
NH_4	0.422	0.118	−0.527
Log_{10} dissolved organic C	−0.533	0.231	0.608
SO_4	0.682	−0.354	0.262
Log_{10} Cl	0.662	0.248	−0.019
Ca	0.520	−0.701	0.087
Mg	0.873	−0.024	0.326
Log_{10} H	−0.735	0.443	0.006

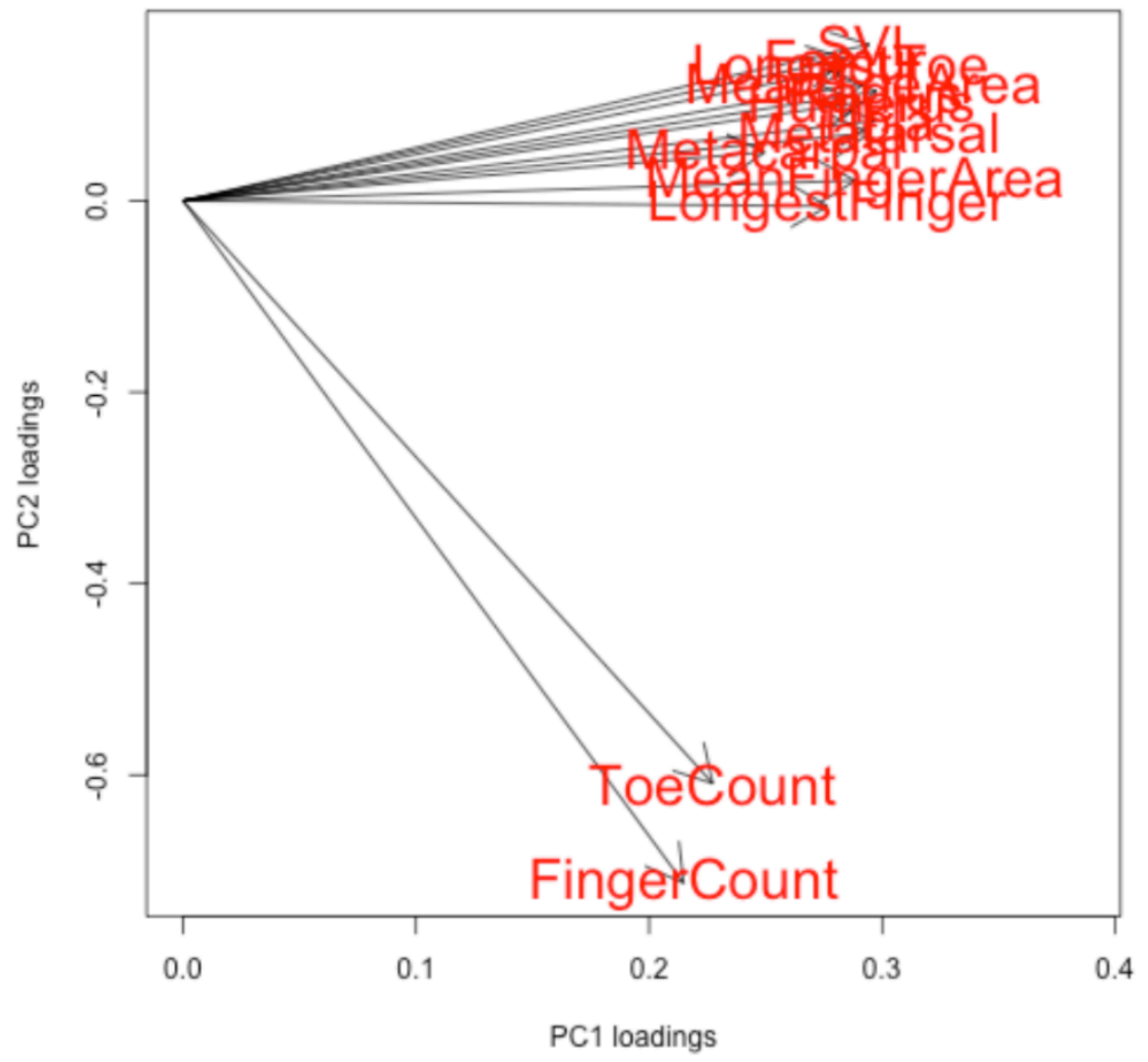
Assumptions of PCA

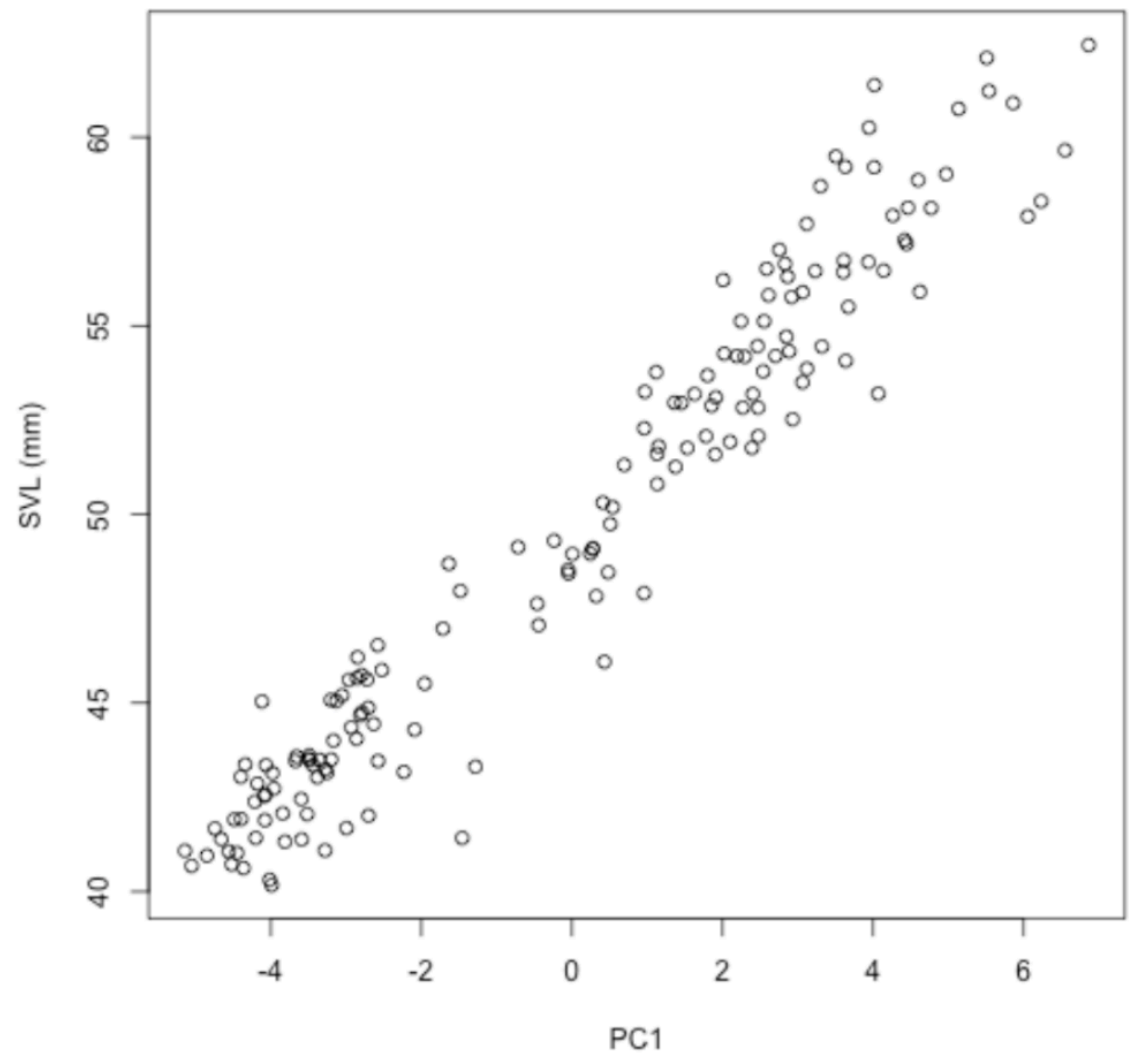
- **Linear relationships** among variables - transformations help!
- **Multivariate outliers** can be a problem, and can be identified via **Mahalanobis** distances (i.e. from centroid).
- **Missing data** are a problem (as with all multivariate analyses), and one of the approaches (removal, imputation, EM) is required.
- **ROBUST PCA** - use Spearman’s rank to form a robust correlation matrix.

PC1 is size:

PCA example (lizard dataset):







PCA example (beer varieties):

```
##
##           Ale    Citrus_IPA    Double_IPA        Helles         NW_IPA
##            21            14            23            20             94
## Oatmeal_Stout   Oktoberfest      Pale_Ale       Red_IPA Vanilla_Stout
##            12             4             8            39             12
##    Winter_Ale
##            10
```

```
beer_vars <- c("Volume", "CO2", "Color", "DO", "pH", "Bitterness_Units", "ABV", "Real_E
beer[,c("Beer_Type", beer_vars)]
```

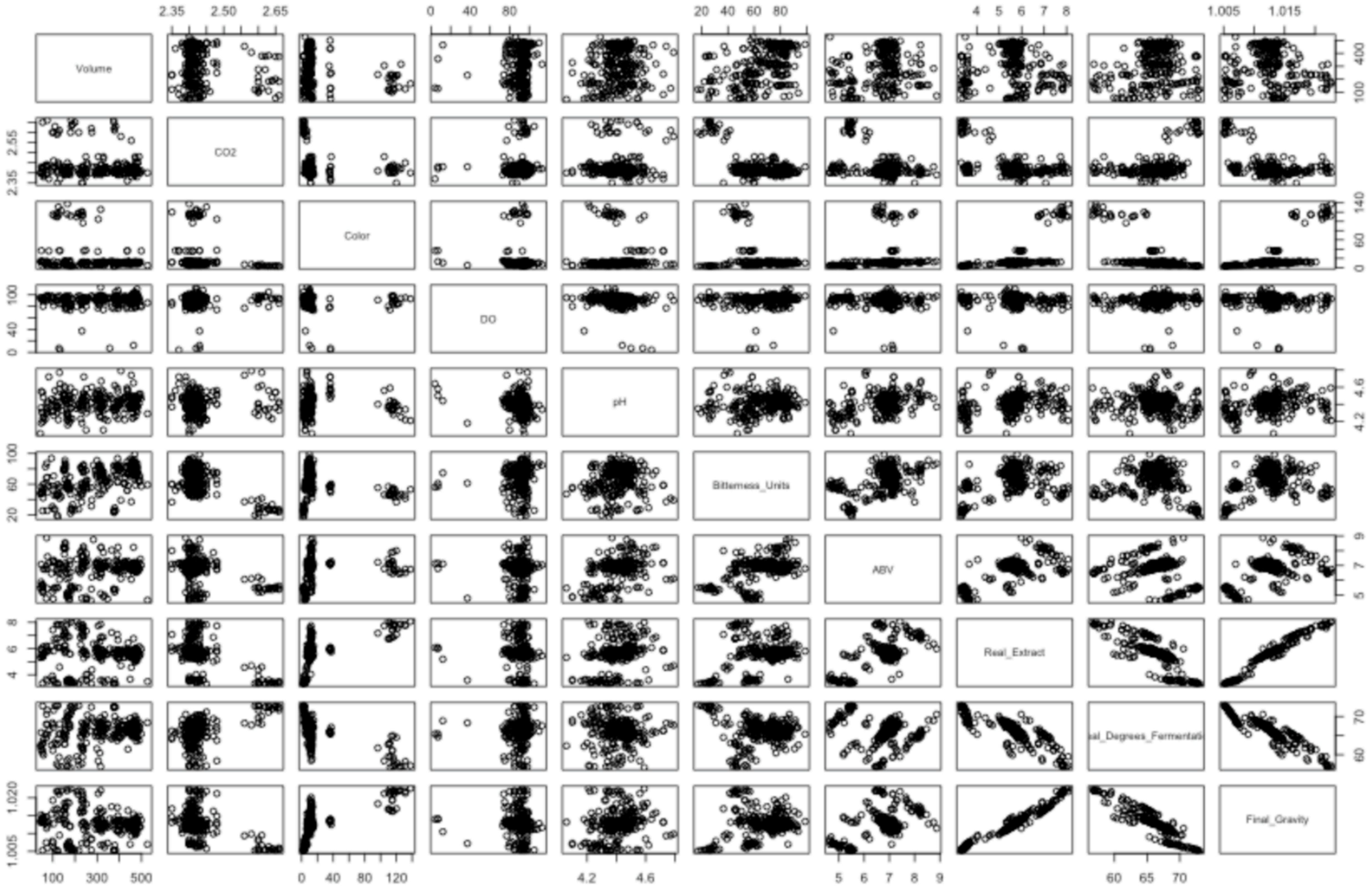
```
##          Beer_Type Volume CO2 Color  DO  pH Bitterness_Units ABV Real_Extract
## 1              Ale    250 2.4   4.7  80 4.1               56 4.9          3.6
## 2              Ale    178 2.4   4.7  86 4.2               53 5.0          3.4
## 3              Ale     70 2.4   4.9  89 4.3               61 4.8          3.6
## 4              Ale    102 2.4   5.0  89 4.3               58 4.7          3.6
## 5              Ale    173 2.4   4.4  82 4.4               59 4.6          3.7
## 6              Ale    254 2.4   4.8  94 4.4               57 4.8          3.5
## 7              Ale    347 2.4    NA  93 4.2               58 4.7          3.6
## 8              Ale    167 2.4   5.2  95 4.2               57 5.2          3.5
## 9              Ale    175 2.4   4.8  98 4.2               54 4.8          3.6
## 10             Ale    163 2.4   5.0  96 4.3               62 4.9          3.5
```

PCA example (beer varieties):

Goals:

1. Describe major axes of variation between beer batches.
2. Look for variation not related to taste/style.
3. Make a pretty plot.

\



PCA example (beer varieties):