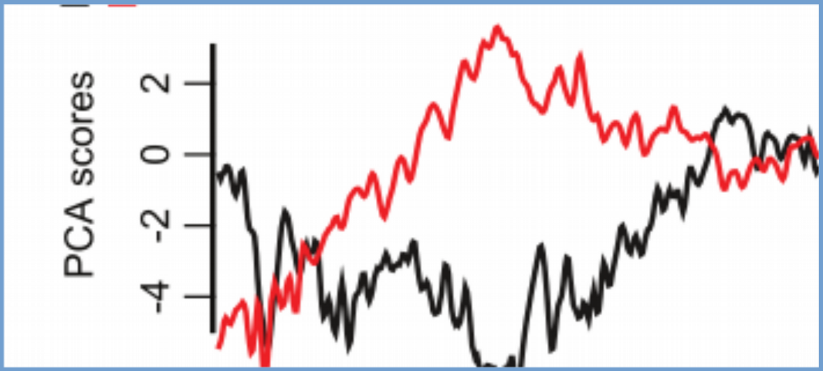
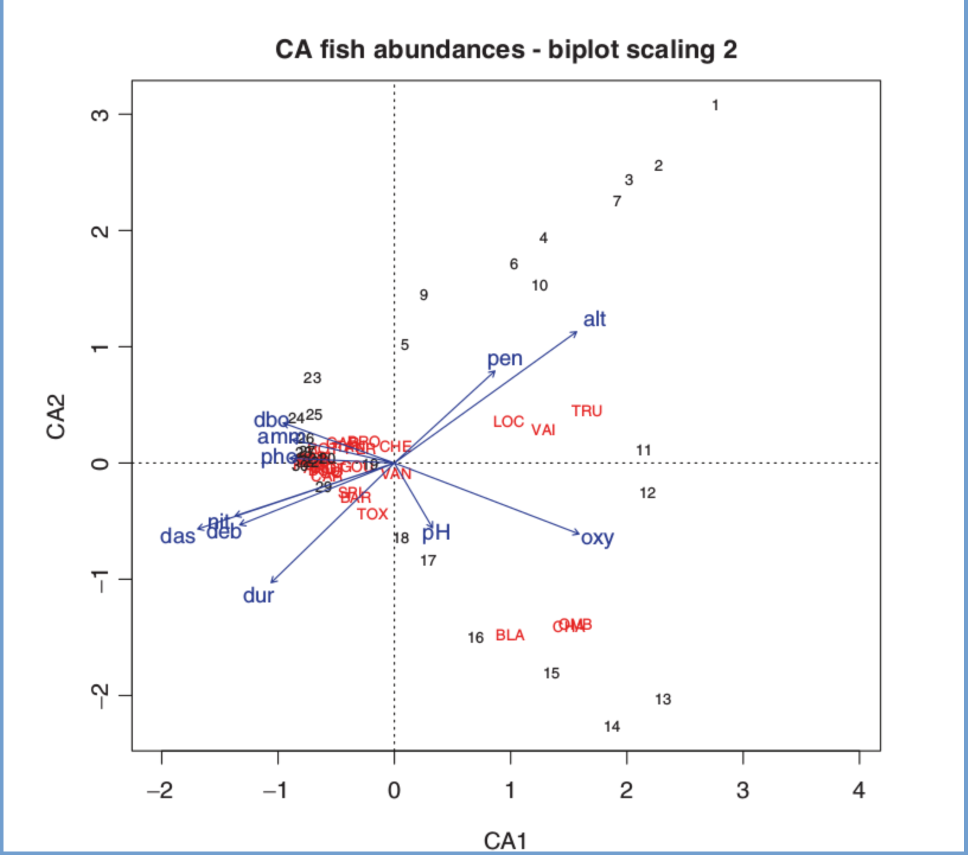
Introduction to ordination: principal components analysis and related methods





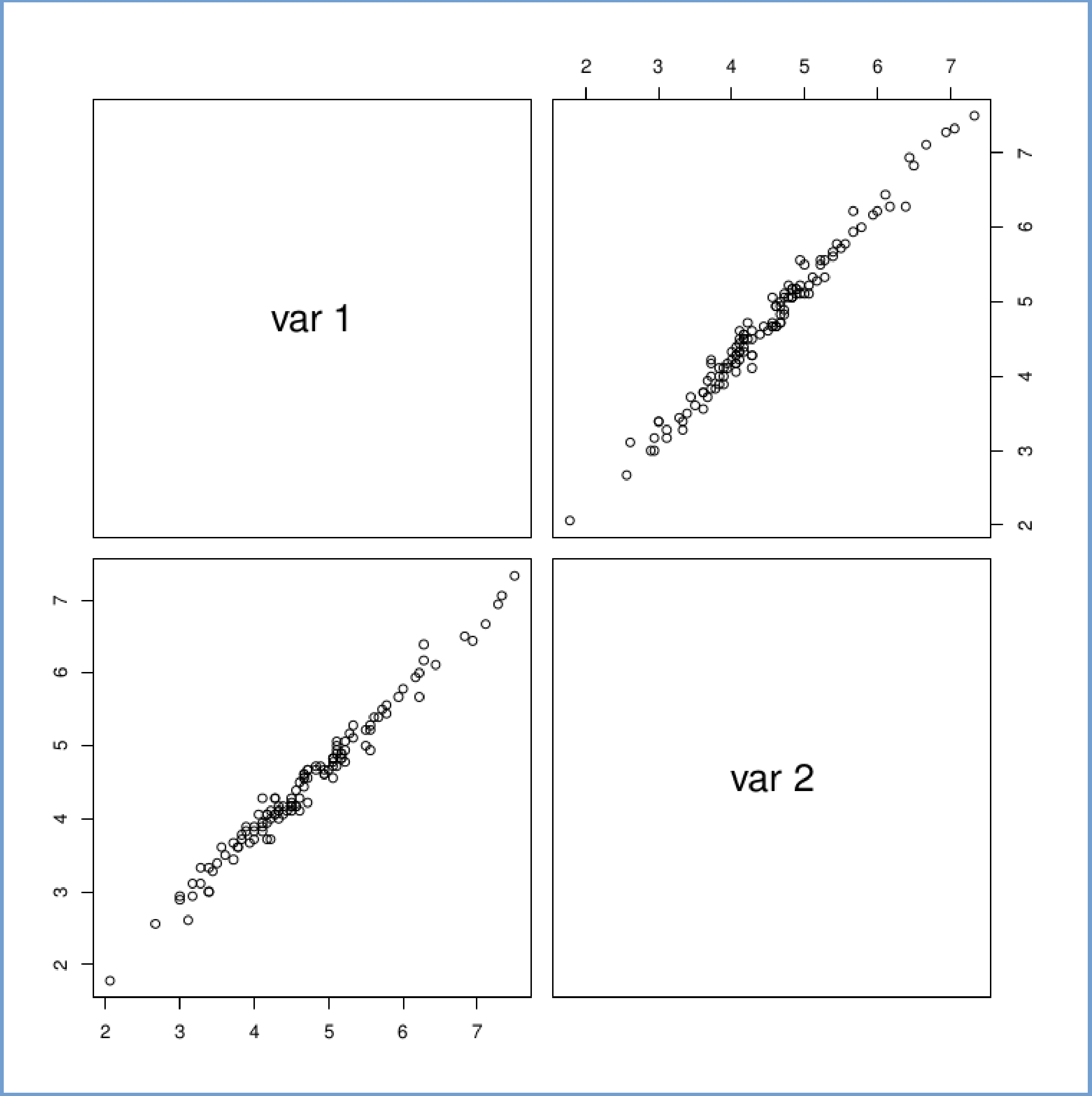
Aims of ordinations

- Simplify complex datasets
  - Find most relevant patterns in a dataset
  - Summarize a dataset in a simple 2-dimensional scatterplot
    - Relation among variables
    - Distance among samples

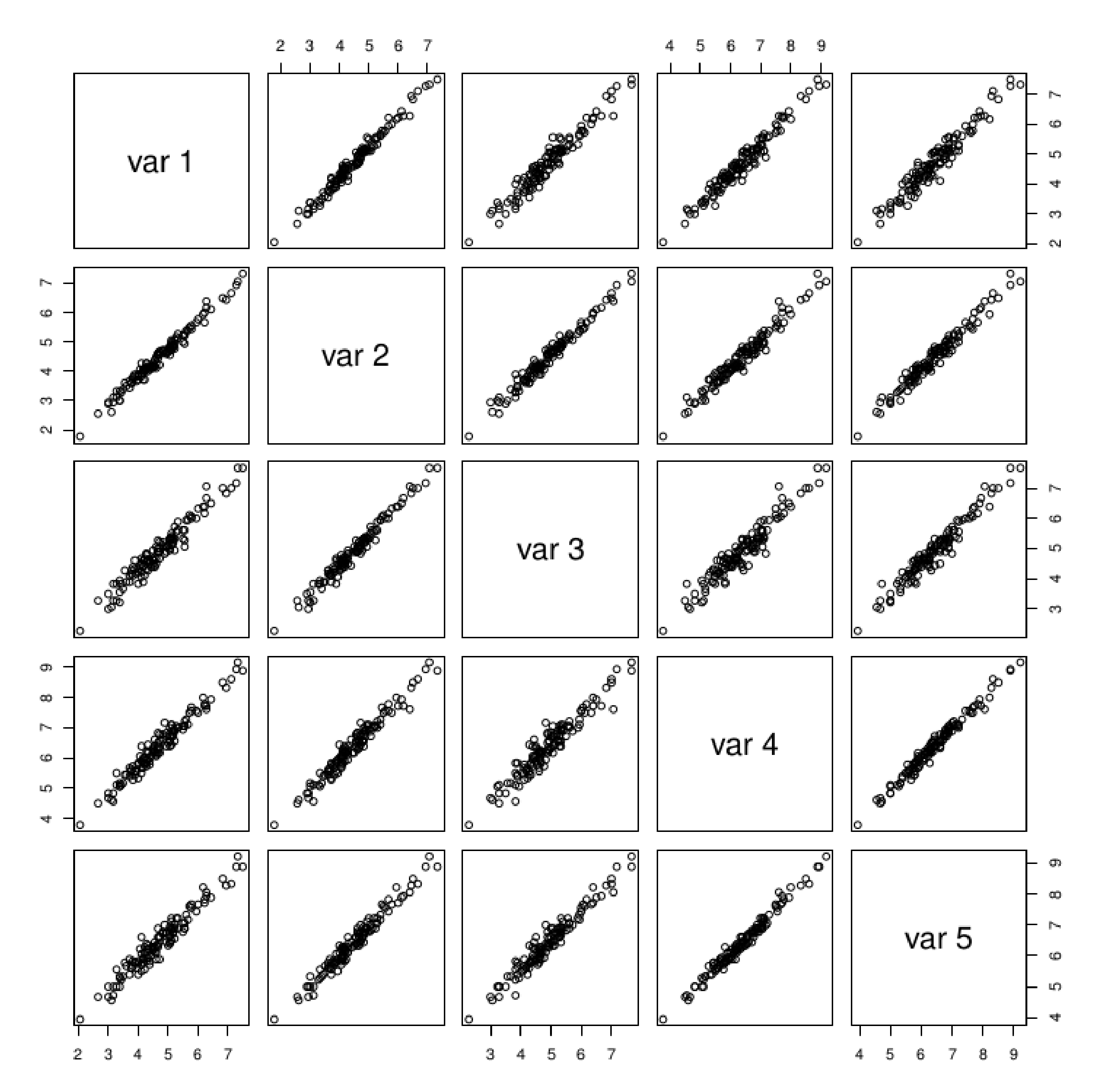
Simplify a dataset

- Dataset with 9 variables
  - Methods to find main patterns:
    - Visually?

Temperatures in Wisconsin

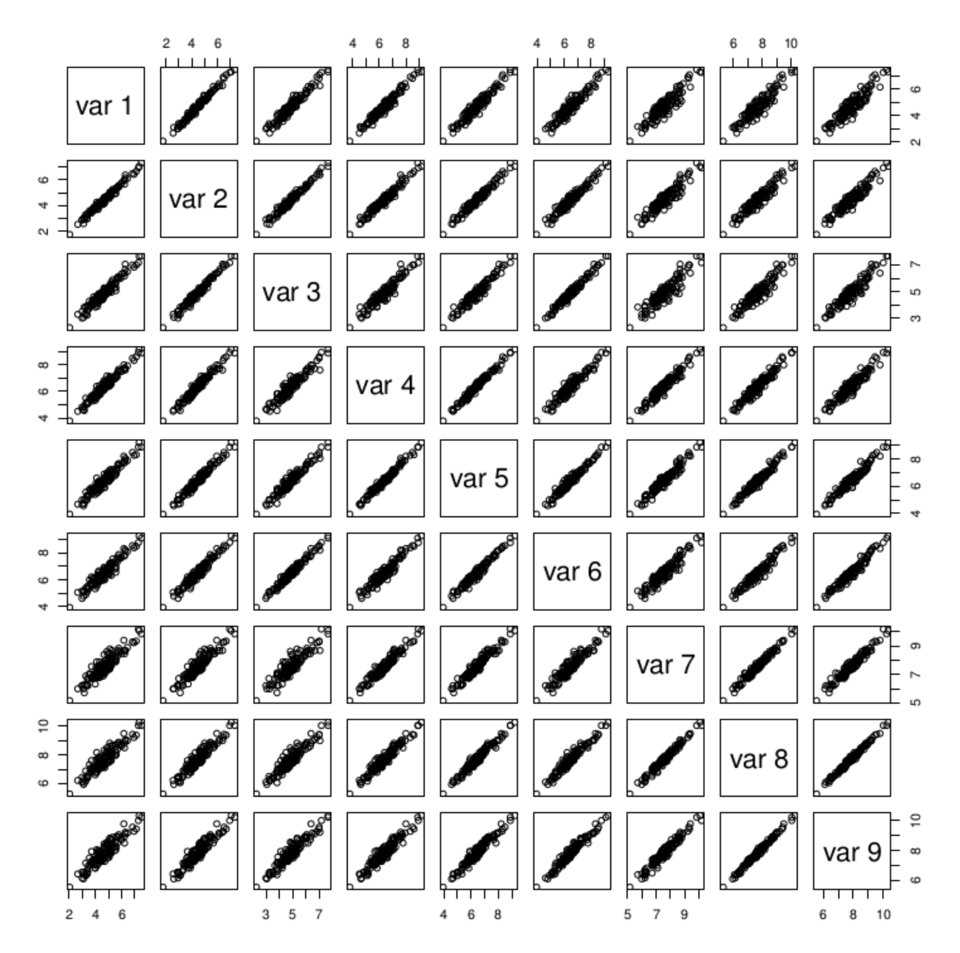


Temperatures in Wisconsin



Simplify a dataset

- Dataset with many variables
  - Methods to find main patterns:
    - Visually?
    - Numerically?

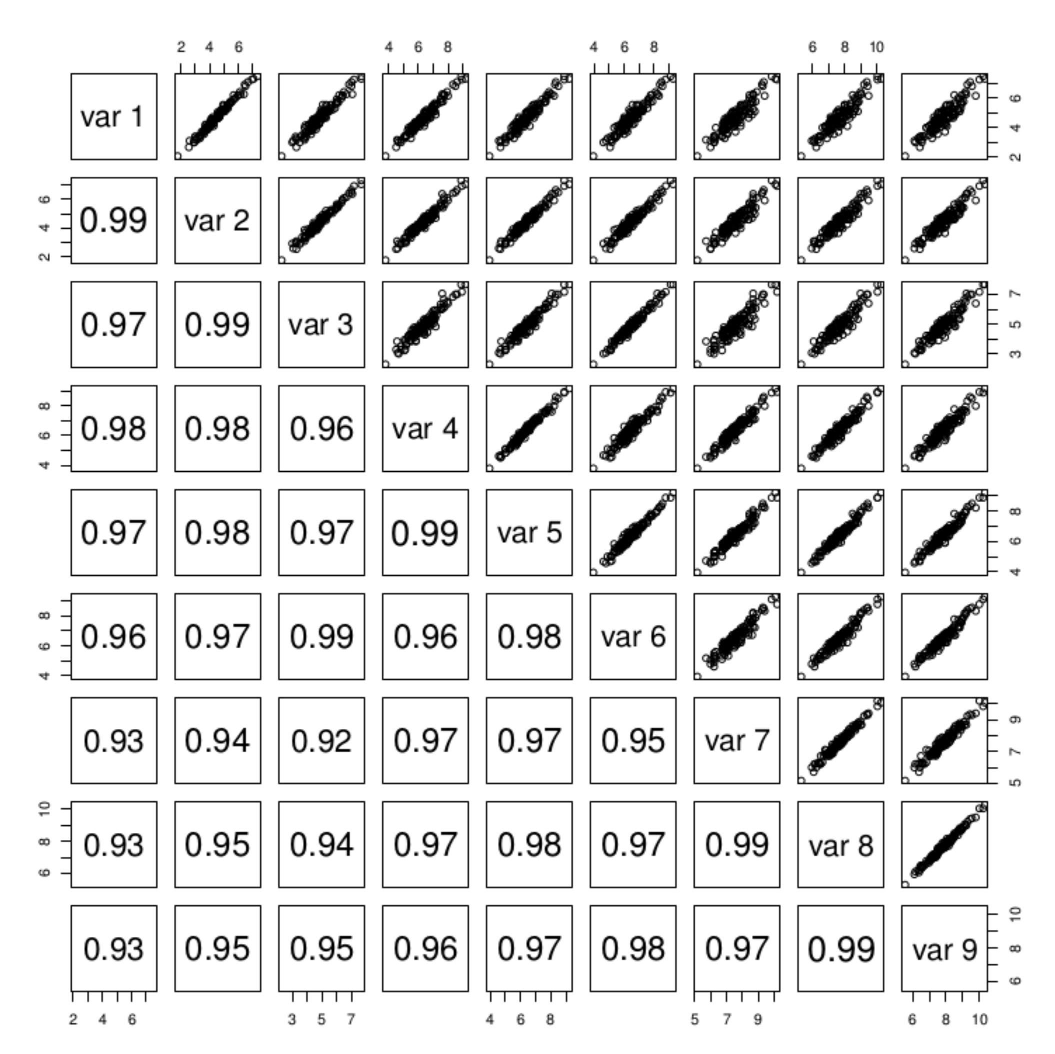


Simplify a dataset

Methods to find main patterns:

Numerically?

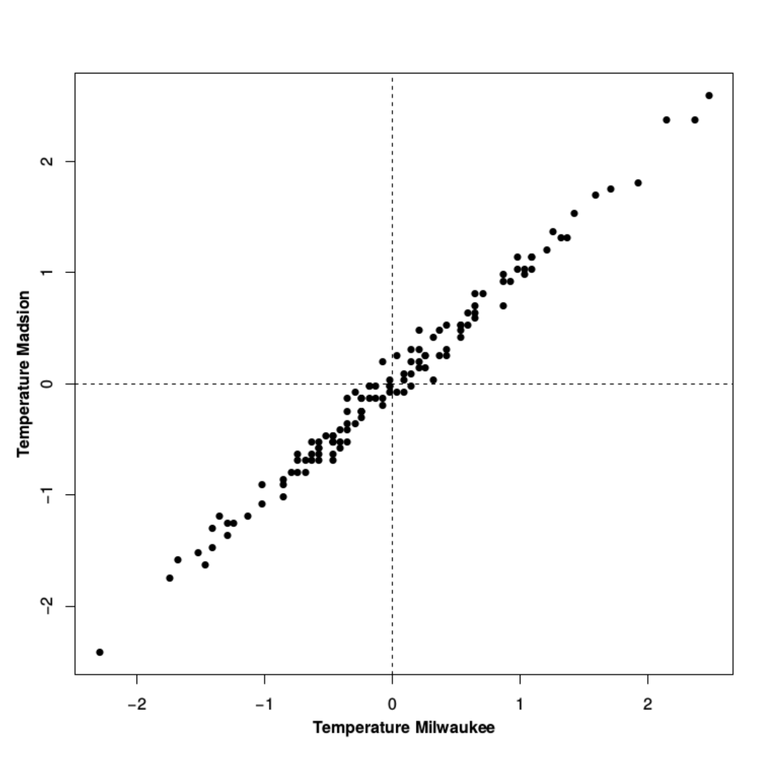
Covariance or correlation matrix



Simplify a dataset

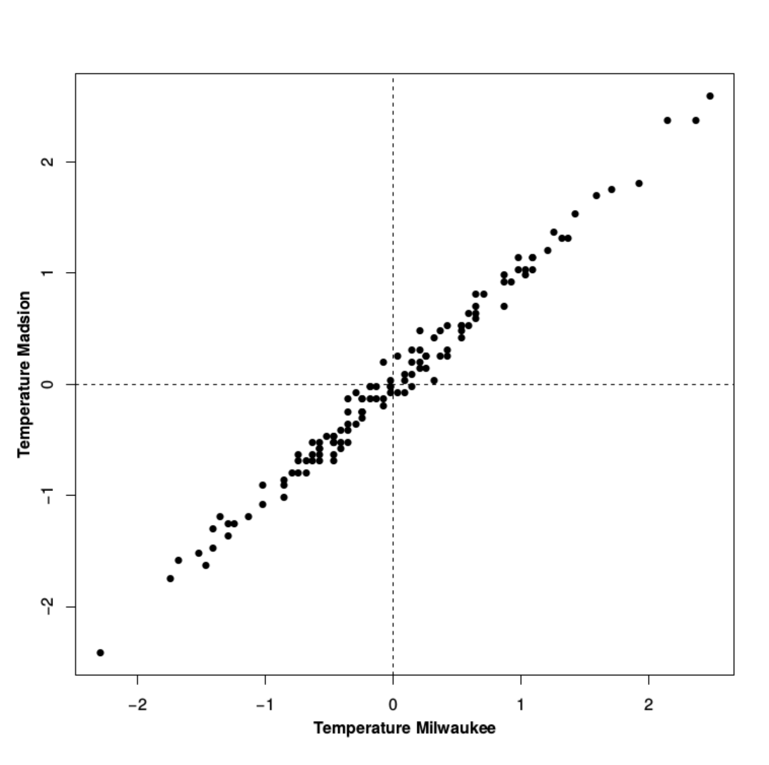
- Dataset with 20 variables
  - Methods to find main patterns:
    - Visually?
      - Show a scatterplot of all variables
    - Numerically?
      - Covariance or correlation matrix

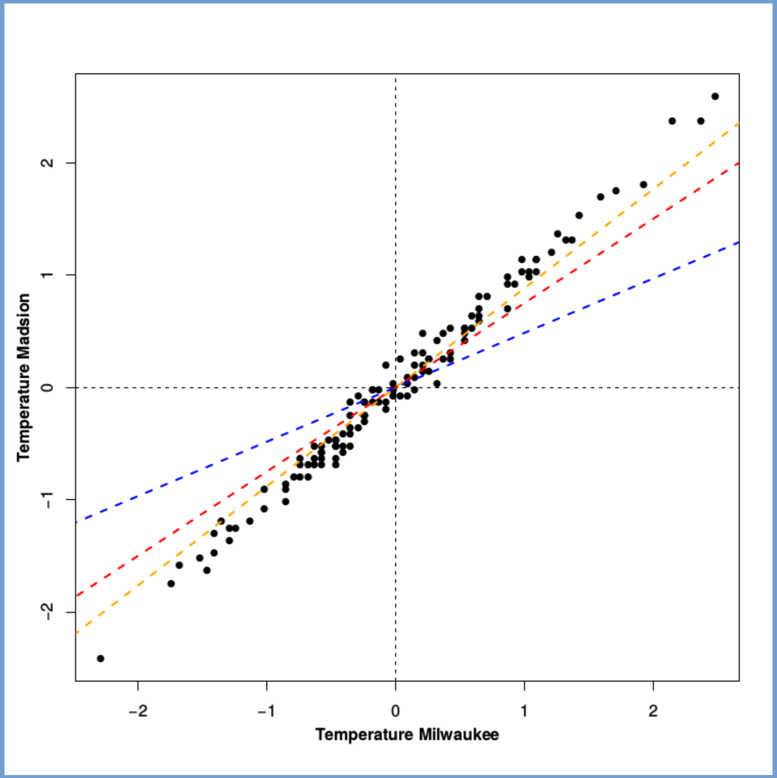
Two dimensional example



- Data represented by a cloud of points in variable space
- Usually our data has structure = correlation between variables = cloud of points is elongated

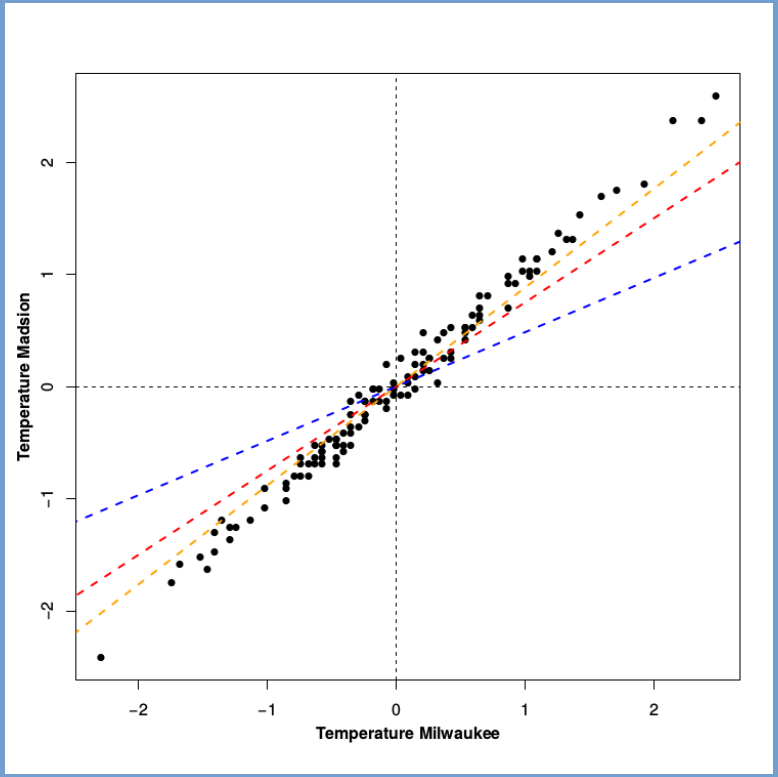
Geometric representation

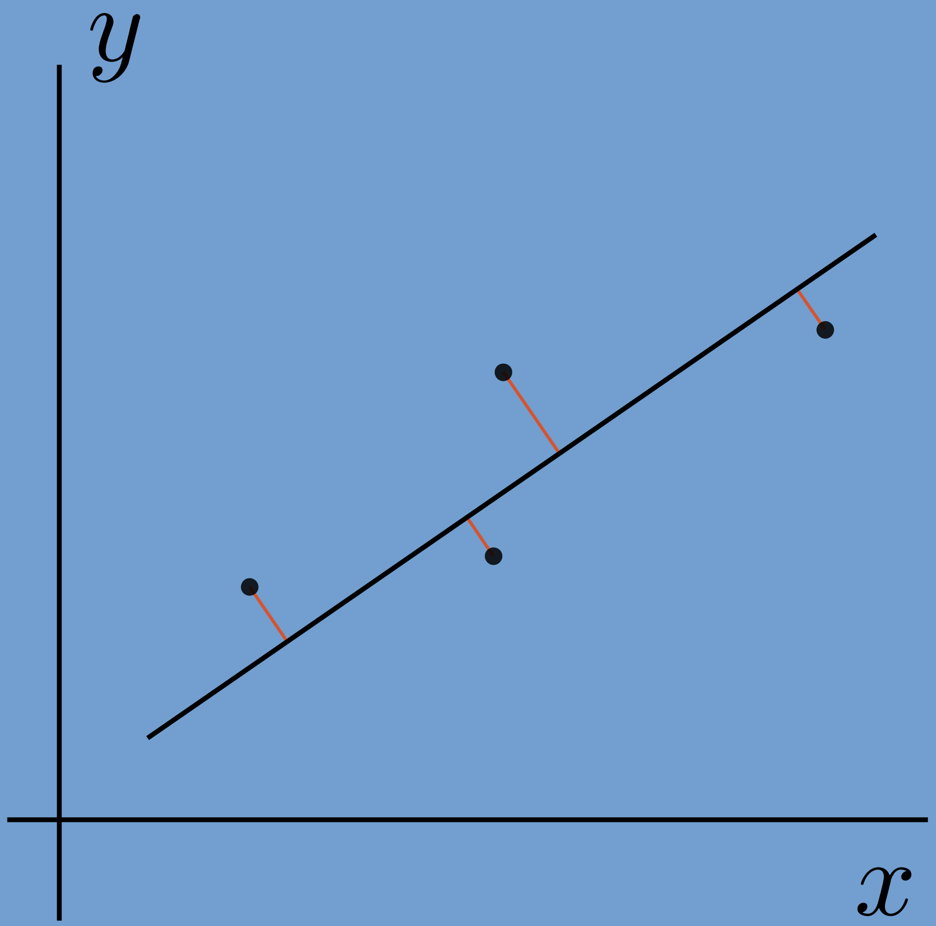




- Draw a line through the direction of maximum elongation that also passes through the center of the swarm:
  - First principal component is the line that explains most variance
  - Minimizes the distance to the observations
  - Does this remind you of another mathematical operation?

PCA as extension of regression





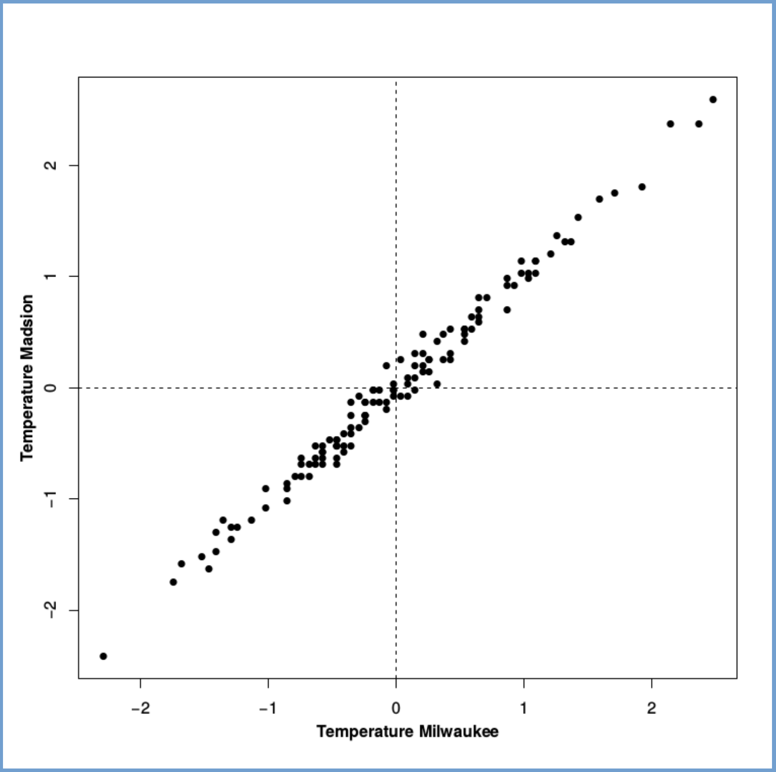
Type II regression: minimizing distance perpendicular to the regression line

Formal solution

- Eigenvalue decomposition of **correlation or covariance matrix**

  - Eigenvalue decomposition (a little magic, no need to know what that is, ***eigen*** in R)
  - **Eigenvalue** indicating how important a component is (as many eigenvalues as variables)
  - **Eigenvector** describing the new coordinate system (as many eigenvectors as variables)

Eigenvalue decomposition 2 dimensions

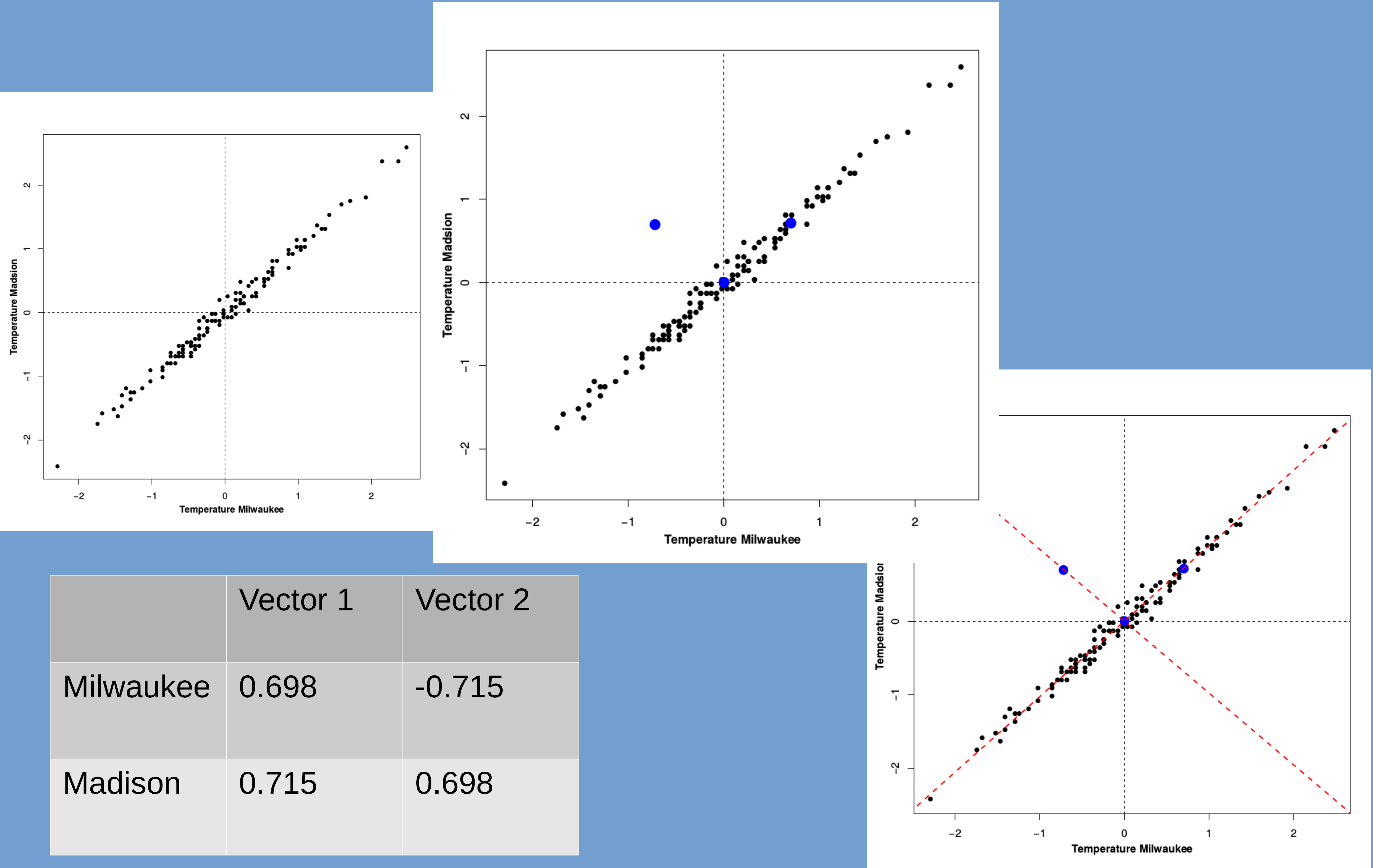


- Eigenvalue (importance):
  - Value 1: 1.985 (99.25%)
  - Value 2: 0.015 (0.75%)

- Eigenvectors (new coordinate system)

	Vector 1	Vector 2
Milwaukee	0.698	-0.715
Madison	0.715	0.698

Two dimensional example

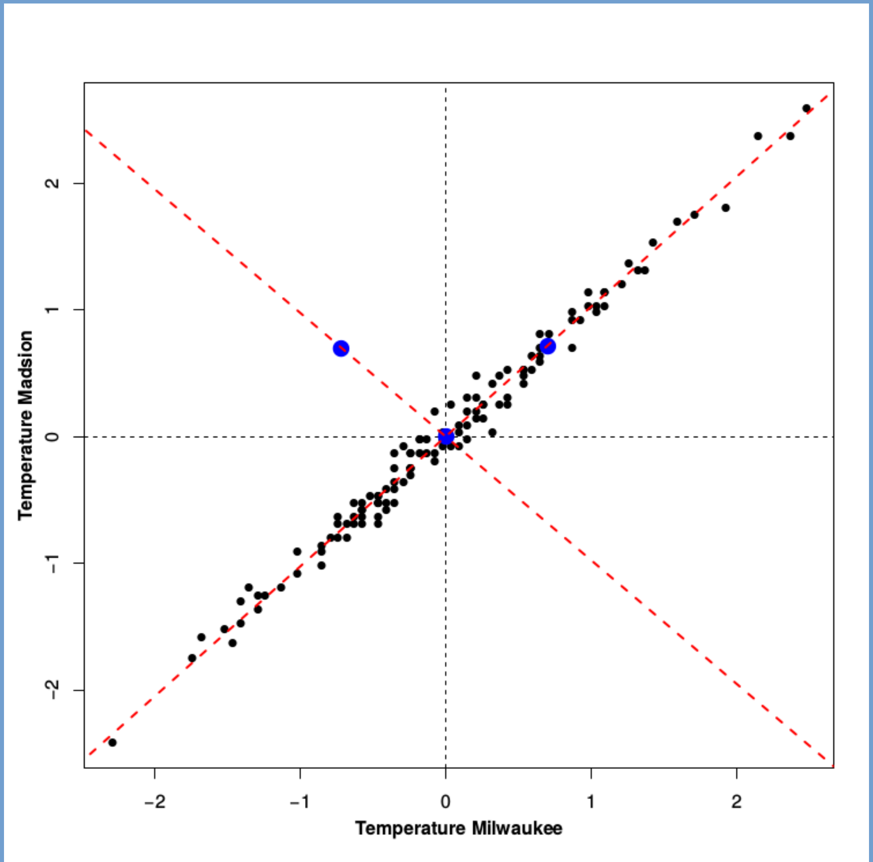


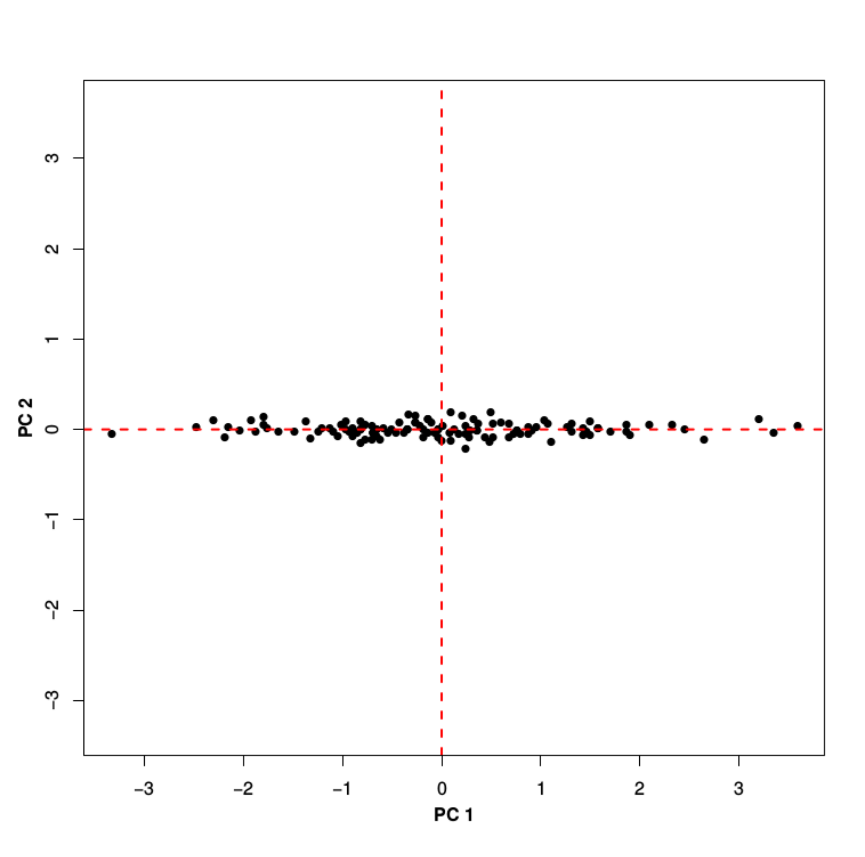
	Vector 1	Vector 2
Milwaukee	0.698	-0.715
Madison	0.715	0.698

PC1 and PC2

- $\mathrm{PC1} = (\mathrm{Vec1}_{\mathrm{Mil}} * \mathrm{T}_{\mathrm{Mil}} + \mathrm{Vec1}_{\mathrm{Ma}} * \mathrm{T}_{\mathrm{Ma}})$
- $\mathrm{PC2} = (\mathrm{Vec2}_{\mathrm{Mil}} * \mathrm{T}_{\mathrm{Mil}} + \mathrm{Vec2}_{\mathrm{Ma}} * \mathrm{T}_{\mathrm{Ma}})$

	Vector 1	Vector 2
Milwaukee	0.698	-0.715
Madison	0.715	0.698





PC1 and PC2

- PC1 = (Vec1$_{Mil}$ * T$_{Mil}$ + Vec1$_{Ma}$ * T$_{Ma}$)
- PC2 = (Vec2$_{Mil}$ * T$_{Mil}$ + Vec2$_{Ma}$ * T$_{ma}$)

	Vector 1	Vector 2
Milwaukee	0.698	-0.715
Madison	0.715	0.698

- PCs are linear combinations of initial variables with eigenvectors as coefficients: **PC Scores** (site score)
- Eigenvectors also called PC **loadings** (species scores)

Variance explained

- Variance PC1 = Eigenvalue 1 = λ_1
- Variance PC2 = Eigenvalue 2 = λ_2

Variance PC1 = 1.539019 Eigenvalue 1 = 1.539018890

Variance PC2 = 0.005447696 Eigenvalue 2 = 0.005447696

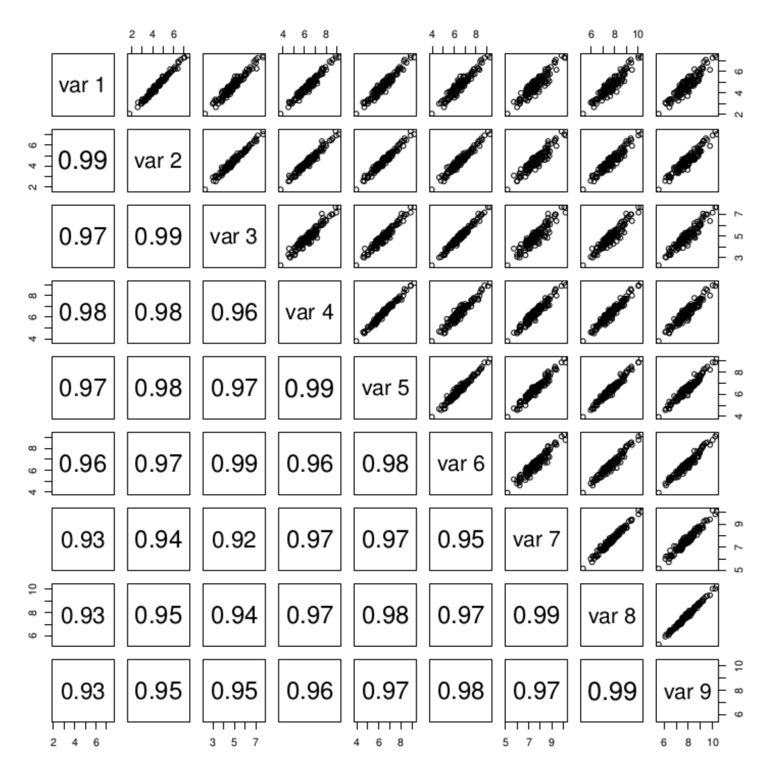
- Total Variance = Var TMil + Var TMad = Var PC1 + Var PC2 = Eig val 1 +Eig val 2
- **Variance explained =** Eig val 1 (Var PC1) / Total Variance

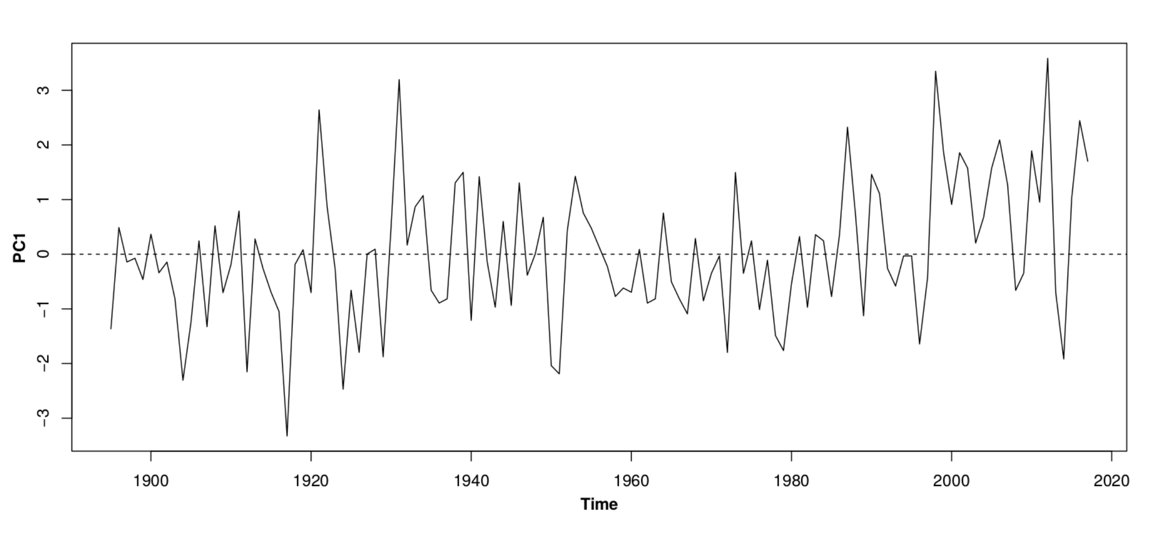
Example: 1.539/1.544 = 0.996

PC1 explains 99.6% of the total variance

Summary

- Simplify dataset:
  - Correlation or covariance matrix
  - A little magic





High dimensional dataset

- Same methodology
- No longer possible to visualize

Visualization of PCA results

- PCA Biplot:
  - Joint representation of variables and observations (samples, sites) on the same diagram

    Usual to represent observations/samples/sites by points and variables by arrows

    Arrows point in direction of maximum rate of change of that variable across the diagram

    Length of arrows indicate relative rate of change in that direction

Example: Climate data Minnesota

- North American Modern Pollen Database (NAMPD)
- Modern climate data at sites with pollen data
- Choose:
  - P Ann
  - P Nov
  - P Dec
  - T ave
- PCA with correlation matrix (standardized variables)

Correlation matrix

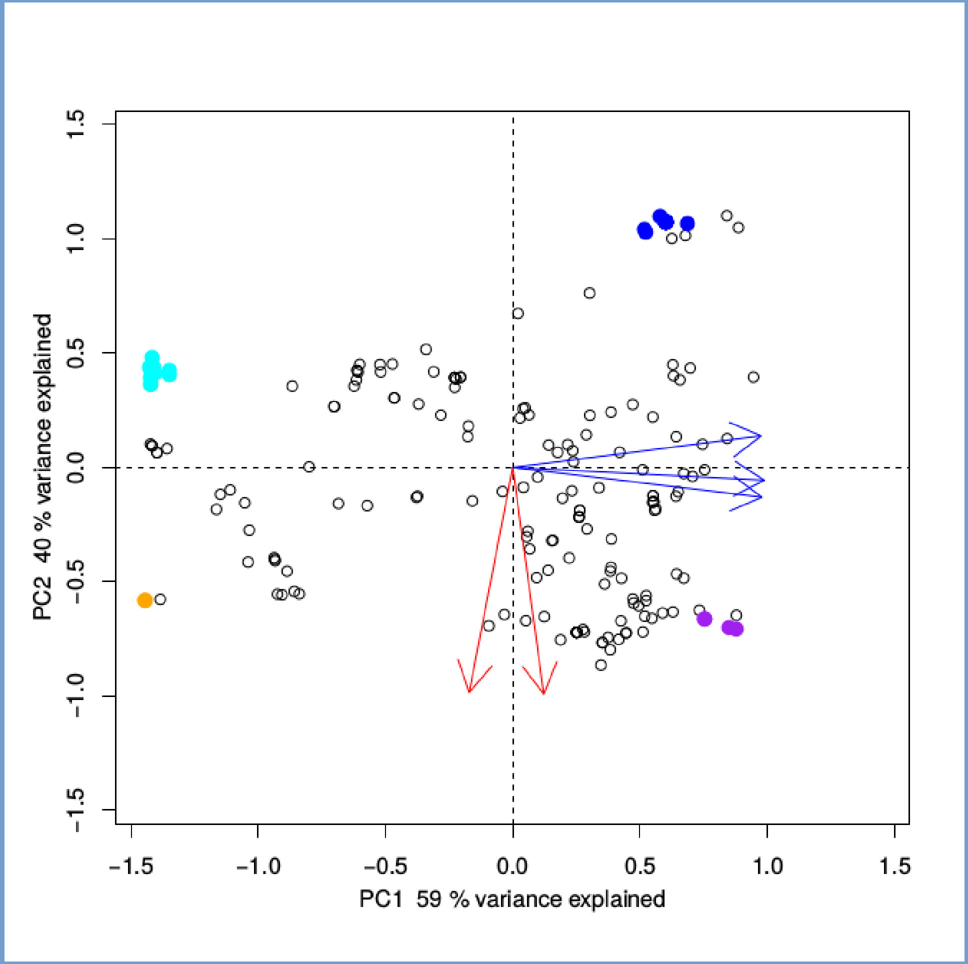
Eigenvector and eigenvalues depend on correlation of variables among each other

	tave	tjul	annp	pnov	pdec
tave	**1**	**0.95**	0.24	0.18	-0.01
tjul	**0.95**	**1**	-0.04	-0.11	-0.3
annp	0.24	-0.04	**1**	**0.96**	**0.92**
pnov	0.18	-0.11	**0.96**	**1**	**0.94**
pdec	-0.01	-0.3	**0.92**	**0.94**	**1**

- Two highly correlated groups of variables
- Two groups are unrelated
- Probably two important PCs

Biplot Example

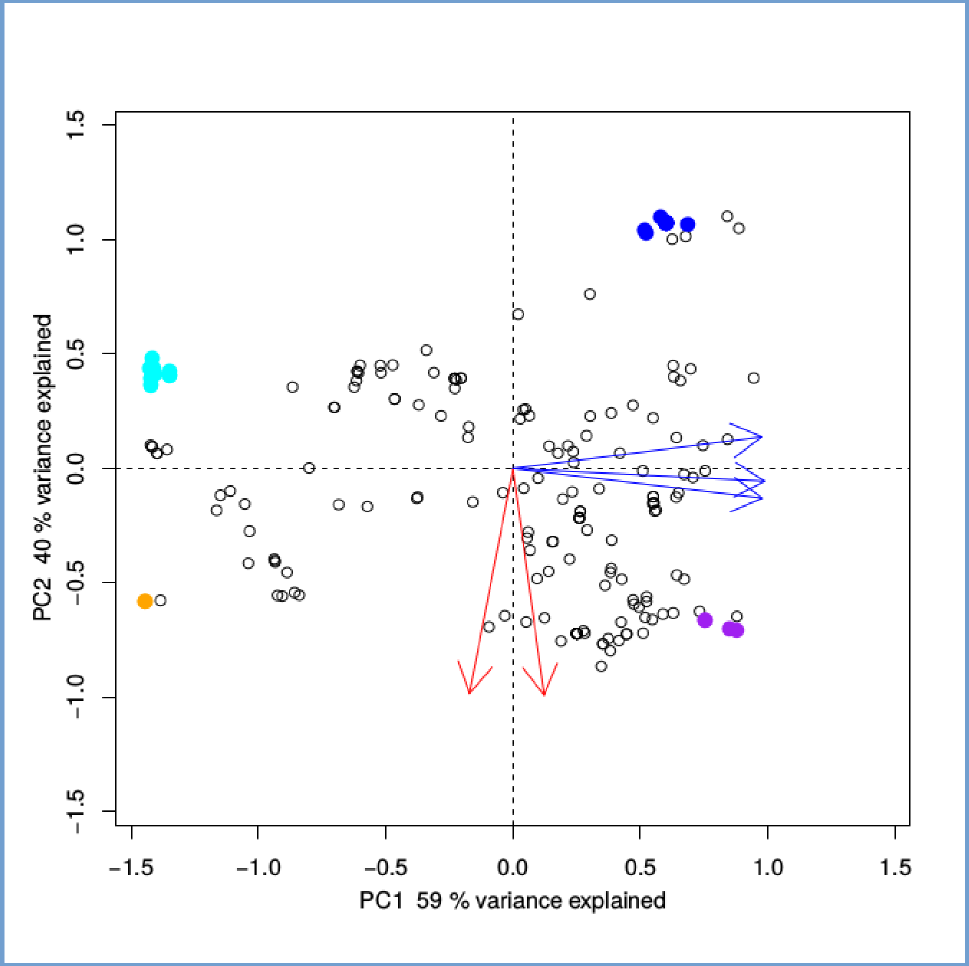
- Dimension reduction 5 to 2



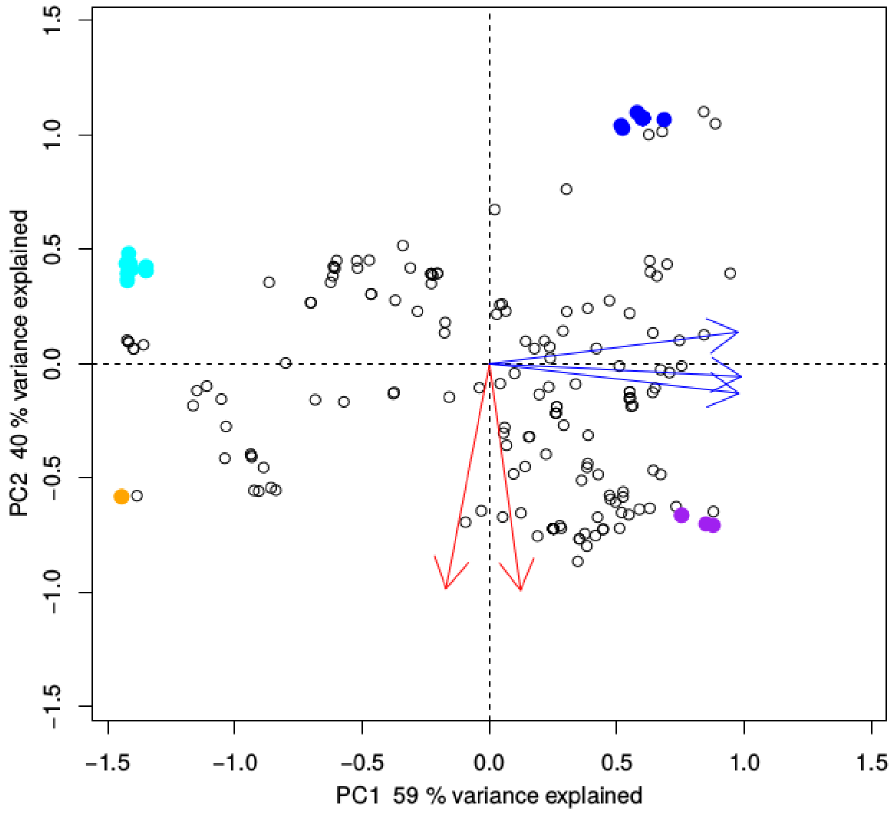
Biplot Example

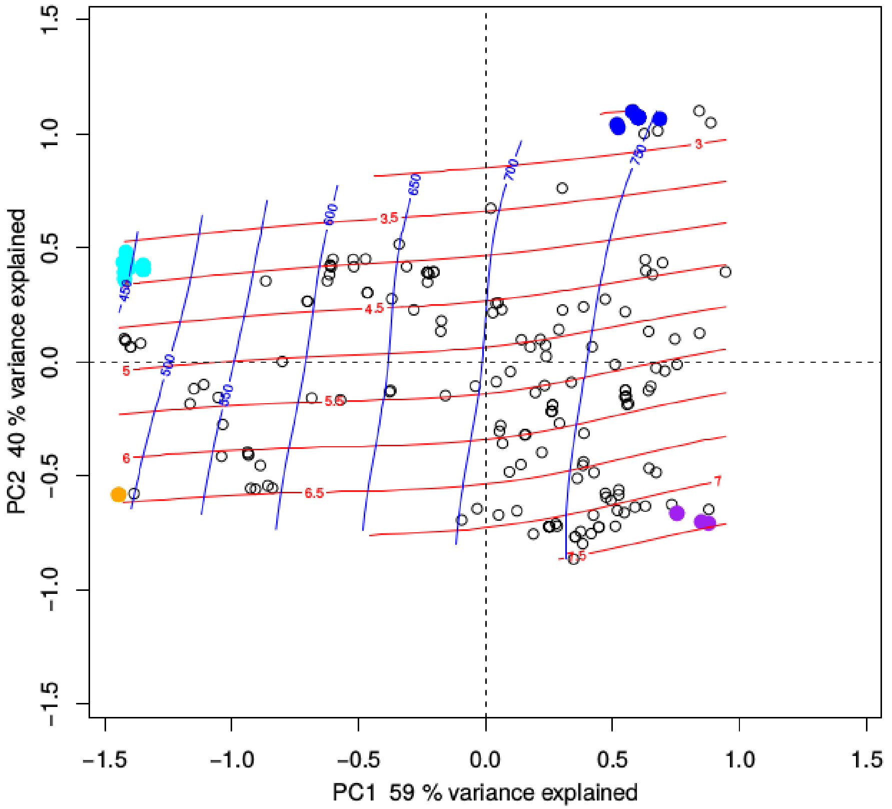
- Angle between arrows:
  - relation between variables
  - Same direction:
    - Positive relation
  - Orthogonal:
    - No relation
  - Opposite direction
    - Negative relation

Describe highlighted points

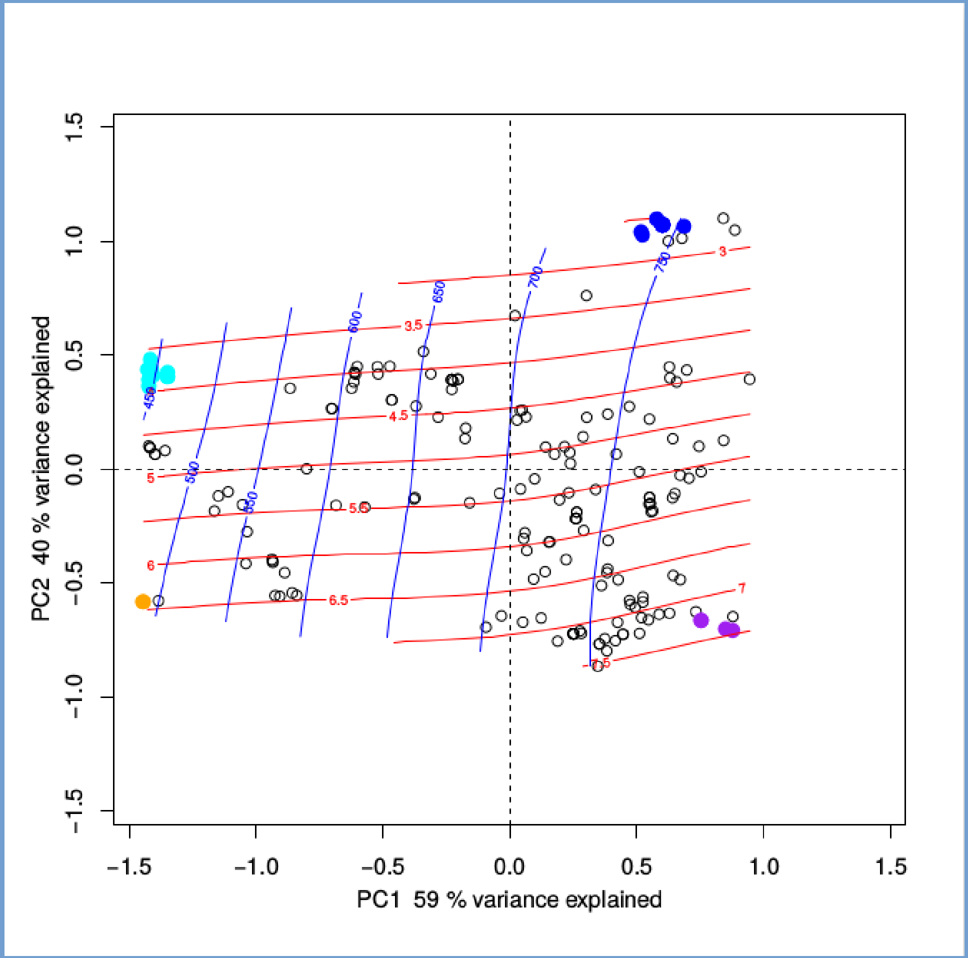


Ordination surface

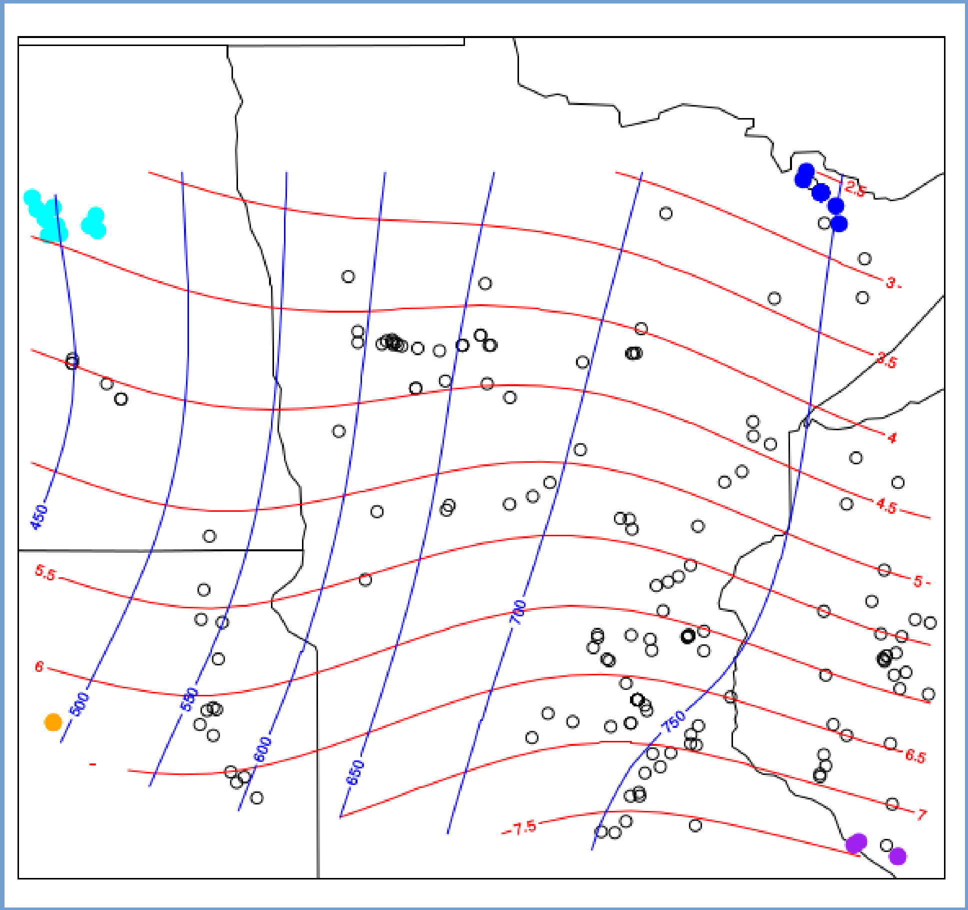




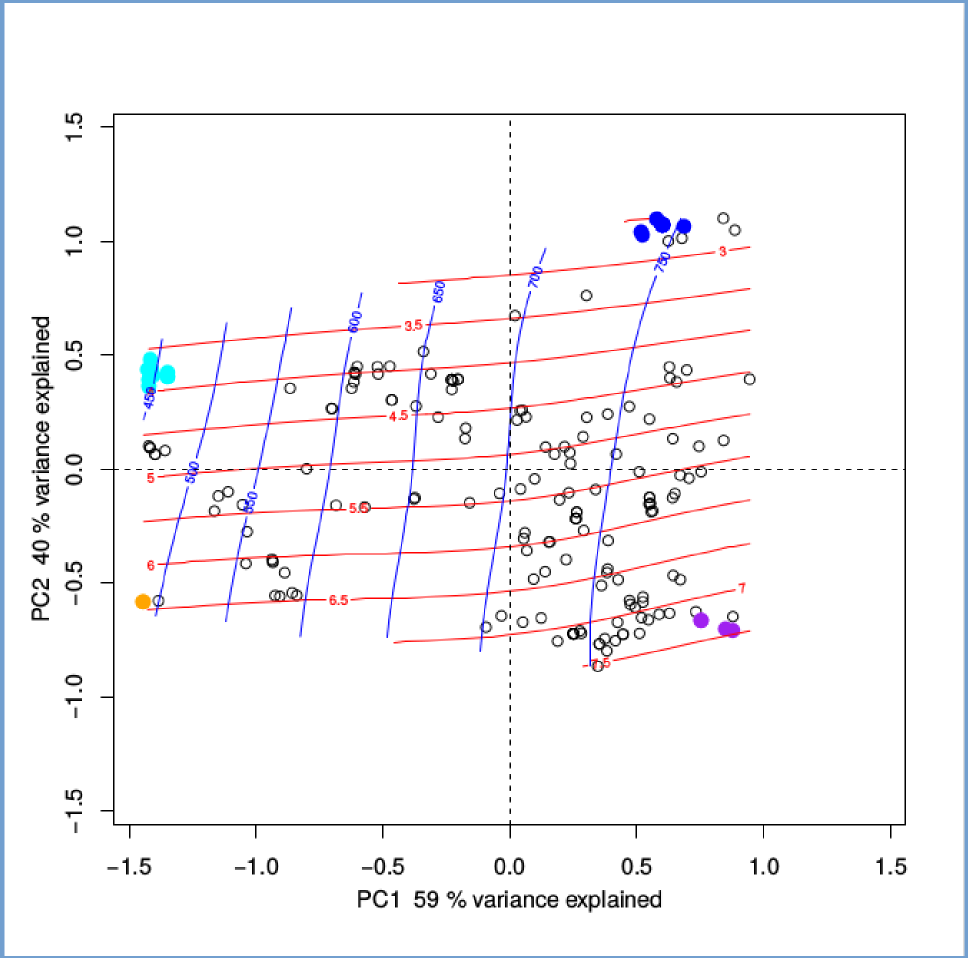
Climate Space



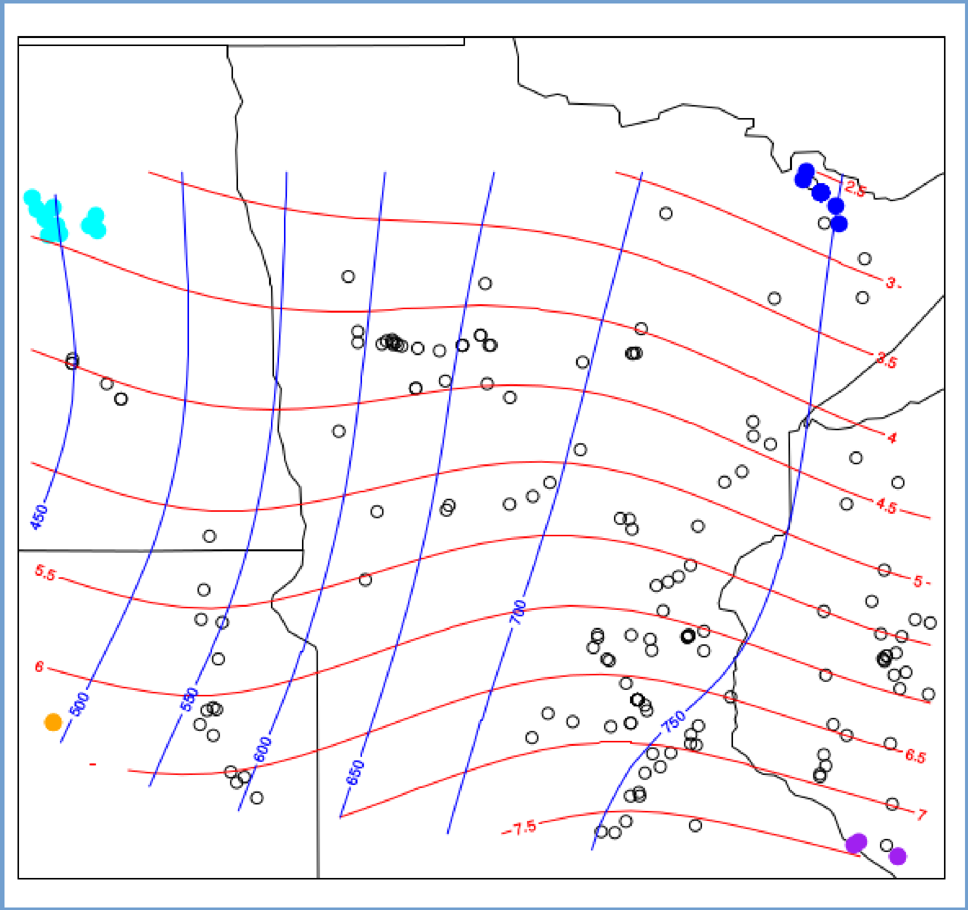
Real Space



Climate Space

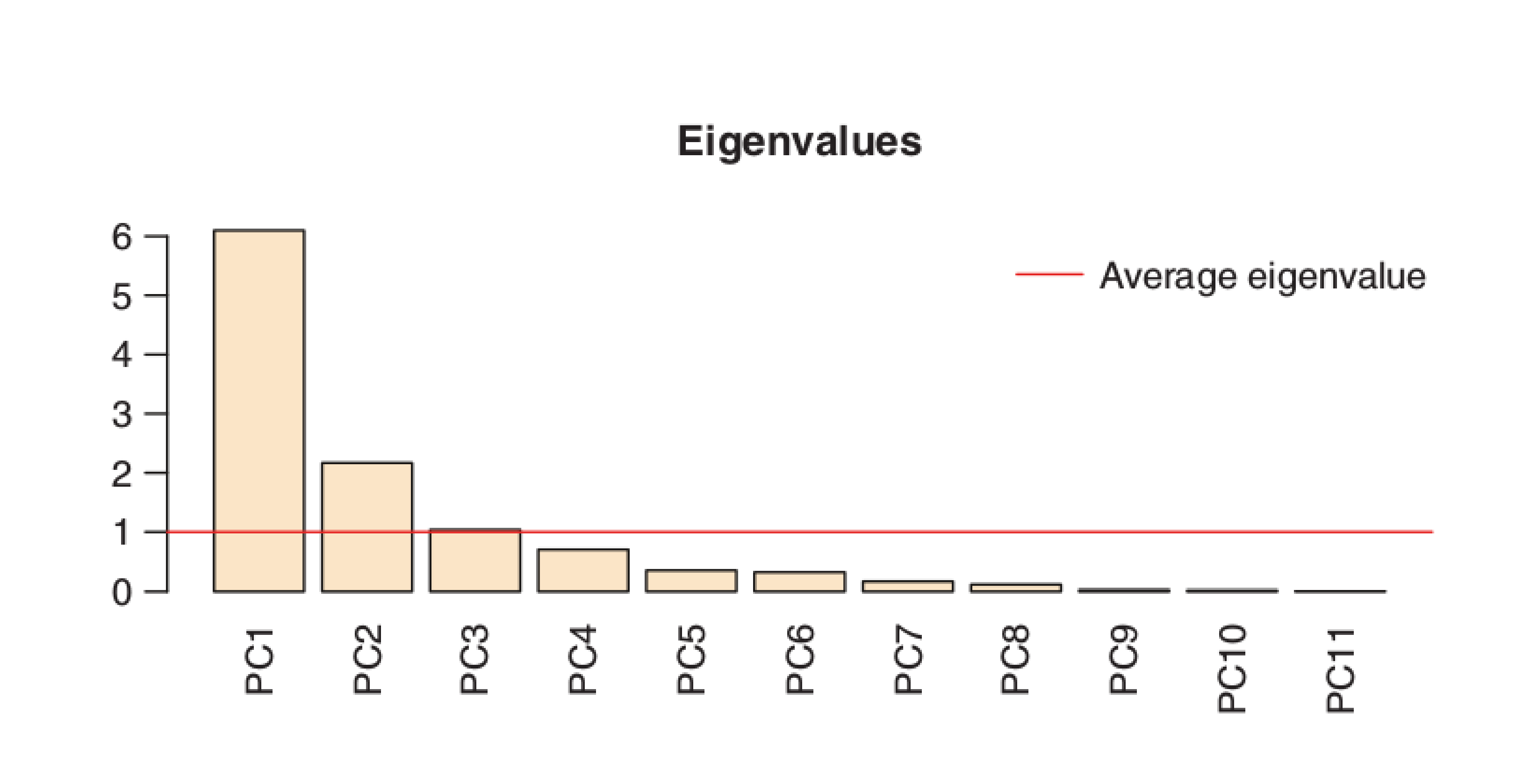


Real Space

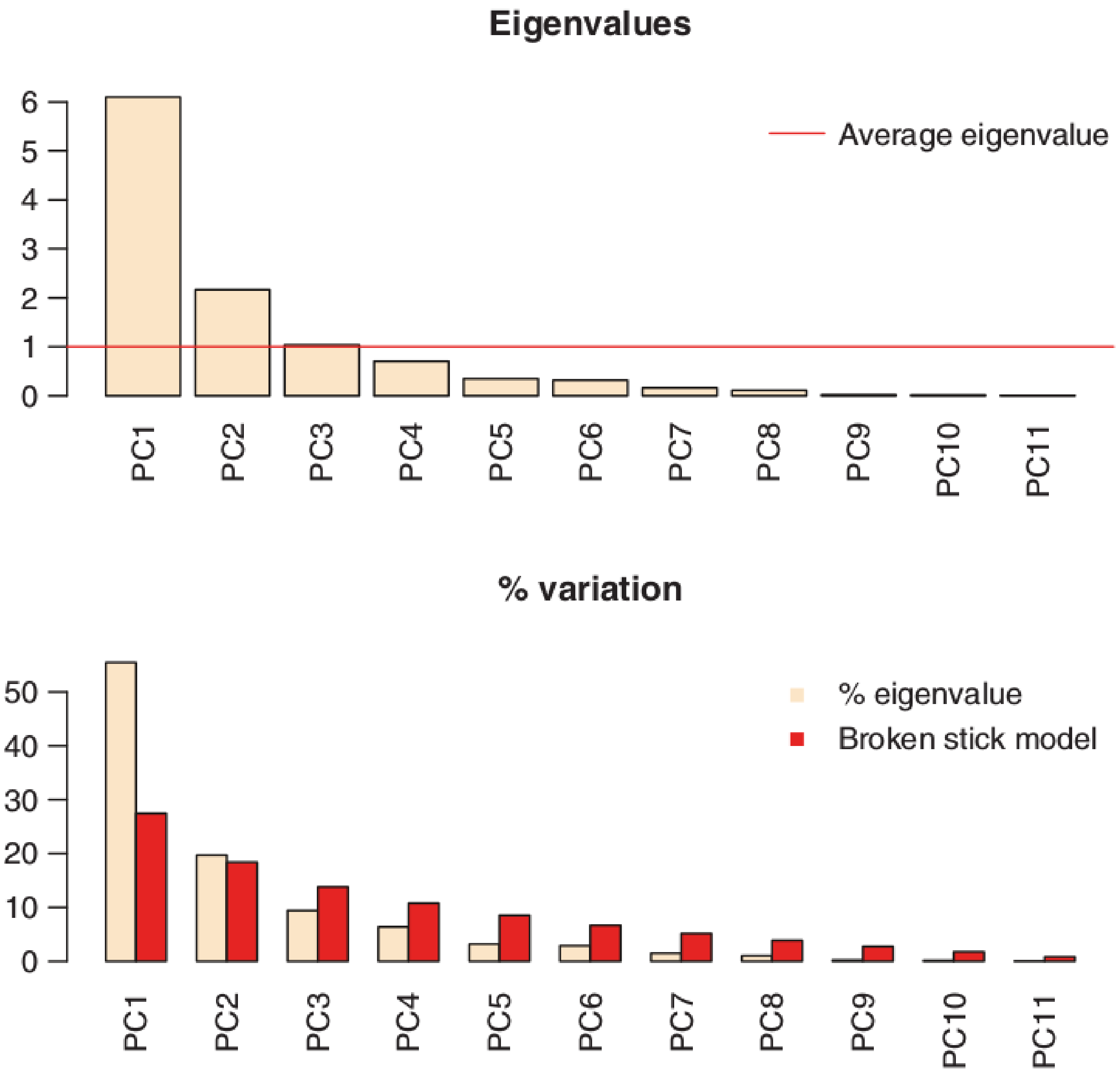


How many axes?

- Screeplot: Histogram of variance explained by each PC axis



How many axes?



Standardized or non-standardized PCA?

- Principal components represent directions of maximum variation through multivariate space
- If original variables are measured on different scales variables with large absolute values will dominate simply because they represent directions of maximum variance
- To avoid this standardize data by centering and dividing by the standard deviation

Standardized or non-standardized PCA?

- A PCA on the non-standardized data is equivalent to a PCA of a covariance matrix between variables
  - Also known as centered variables
- A PCA on the standardized data is equivalent to a PCA of a correlation matrix between variables

Data transformations for PCA

- Usual to center data (subtract mean) – implicit in most software. Variables implicitly weighted by their variances
- For non-species data
  - Standardize data – variables have equal weight
- For species data
  - Log10(x+1) transformation useful for abundance (count) data
  - Square root transformation useful for percentage data

Dealing with closed data

- Closed data: percentage or as a fraction of a fixed total
- Frequent in geochemical and microfossil studies
  - Log-ratio transformation
  - PCA: log transform and center by sample (site) and variable (species)
- Need to consider this for data with small number of variables (species) but usually makes little difference if number of variables is large