Statistics for the Sciences

Canonical Correlation Analysis

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Canonical Correlation Analysis
- Example
- Lab

Canonical Correlation Analysis

- Canonical correlation analysis (CCA) is a way of measuring the **linear relationship** between **two multidimensional variables**.
- CCA finds two bases in which the correlation matrix between the variables is diagonal and the correlations on the diagonal are maximized.
- Canonical correlations are invariant with respect to affine transformations of the variables
- One appliction in Environmental Science is to investigate the relationship between environmental factors and biological data.

Canonical Correlation Analysis

- Suppose we are given two vectors of random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \text{ and } \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix}.$$

- The idea is to find an index describing a (possible) link between $\boldsymbol{X}$ and $\boldsymbol{Y}$.
- Consider the linear combinations

$$a^T \boldsymbol{X} = a_1 X_1 + \cdots + a_q X_q$$

and

$$b^T \boldsymbol{X} = b_1 Y_1 + \cdots + b_p Y_p$$

Canonical Correlation Analysis

- Canonical correlation analysis searches for vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ such that the relation of the two indices $a^T \boldsymbol{X}$ and $b^T \boldsymbol{Y}$ is quantified in some interpretable way.
- Suppose the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is (generally assumed to be multivariate normal)

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \left(\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{bmatrix} \right) \quad \text{with} \quad \boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}_{YX}^T,$$

where $Var(\boldsymbol{X}) = \boldsymbol{\Sigma}_{XX}$, $Var(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{YY}$, and $Cov(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}_{YX}^T$.

Canonical Correlation Analysis

- It can be shown that the correlation between $a^T \boldsymbol{X}$ and $b^T \boldsymbol{Y}$ is

$$\rho(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}^T \Sigma_{XY} \boldsymbol{b}}{\left[\boldsymbol{a}^T \Sigma_{XX} \boldsymbol{a}\right]^{1/2} \left[\boldsymbol{b}^T \Sigma_{YY} \boldsymbol{b}\right]^{1/2}}$$

  - Invariance of scale: For any $c \in \mathbb{R}$, $\rho(c\boldsymbol{a}, \boldsymbol{b})$
- So we can re-scale $\boldsymbol{a}$ and $\boldsymbol{b}$. The problem now becomes find $\boldsymbol{a}$ and $\boldsymbol{b}$ to

$$\max(\boldsymbol{a}^T \Sigma_{XY} \boldsymbol{b})$$

under the constraints $\boldsymbol{a}^T \Sigma_{XX} \boldsymbol{a} = 1$ and $\boldsymbol{b}^T \Sigma_{YY} \boldsymbol{b} = 1$.

Canonical Correlation Analysis

- We skip all mathematical derivations, the solution to the CCA problem is a sequence of vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ with
  - $\boldsymbol{a}_i = \boldsymbol{\Sigma}_{XX}^{-1/2}\gamma_i$ and $\boldsymbol{b}_i = \boldsymbol{\Sigma}_{YY}^{-1/2}\delta_i$ maximize the correlation between canonical variables ${\boldsymbol{a}_i}^T\boldsymbol{X}$ and ${\boldsymbol{b}_i}^T\boldsymbol{Y}$, $i = 1, \ldots, k$
    - ⋆ where γ_i is the eigen vector associated withe the ith largest eigen value λ_i of $\boldsymbol{K}\boldsymbol{K}^T$ and δ_i is the eigen vector associated withe the ith largest eigen value λ_i of $\boldsymbol{K}^T\boldsymbol{K}$, with
    - ⋆ $\boldsymbol{K} = \boldsymbol{\Sigma}_{XX}^{-1/2}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2}$
    - ⋆ k is the number of nonzero eigenvalues of $\boldsymbol{K}\boldsymbol{K}^T$ and $\boldsymbol{K}^T\boldsymbol{K}$.

- ${\boldsymbol{a}_i}^T\boldsymbol{X}$ and ${\boldsymbol{b}_i}^T\boldsymbol{Y}$ are called **canonical correlation variables** or **canonical variables**
  - $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are called **canonical correlation vectors** or **canonical directions**

Canonical Correlation Analysis

- Square roots of the nonzero eigenvalues $\lambda_i, i = 1, 2, \ldots, k$ of $\boldsymbol{K}\boldsymbol{K}^T$ and $\boldsymbol{K}^T\boldsymbol{K}$, are called the **canonical correlation coefficients**. They are the correlations between the pairs of canonical variables.

- **Canonical loadings**: $\boldsymbol{\Sigma}_{XX} a_i$ and $\boldsymbol{\Sigma}_{YY} b_i, i = 1, \ldots, k$. They are the correlations between the original variables and their respective canonical variables.
  - Canonical loadings measure how much each original variable contributes to the respective canonical variate.

Example

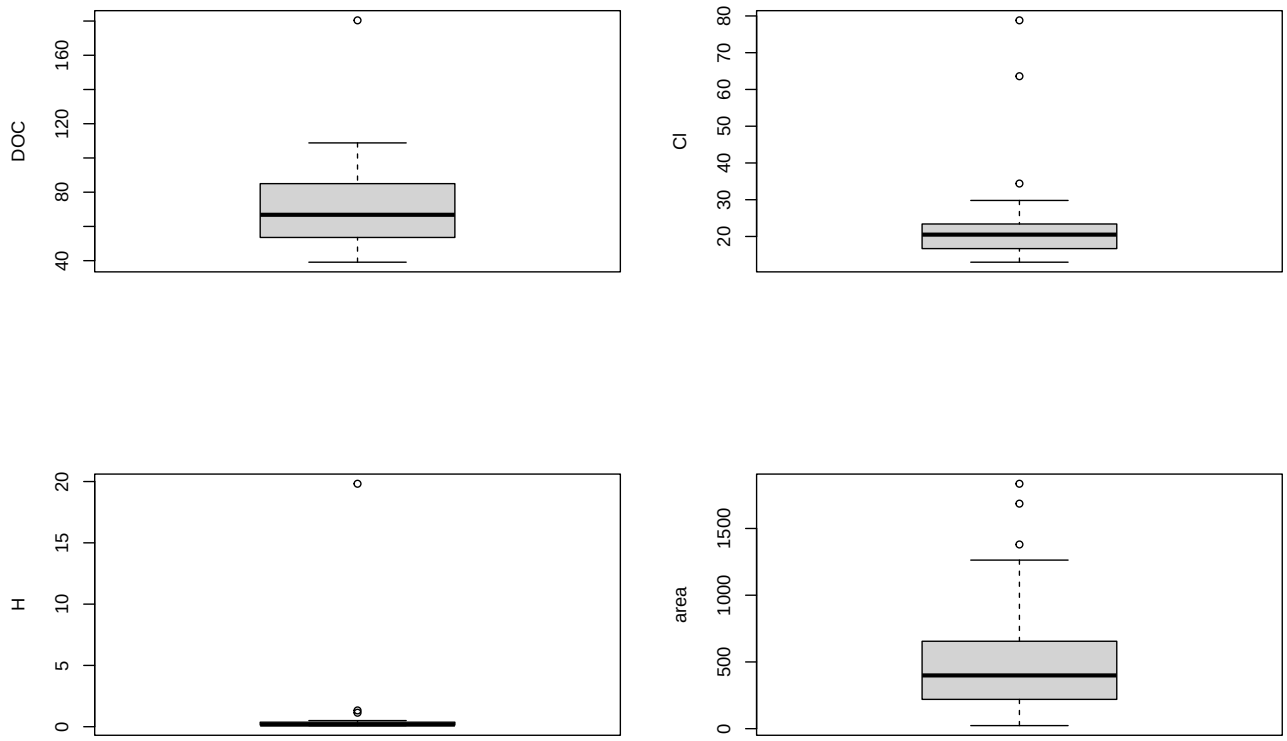
Lovett *et al.* (2000) studied the chemistry of forested watersheds in the Catskill Mountains in New York State. They chose 39 sites (observations) on first and second order streams and measured the concentrations of ten chemical variables (NO_3^-, total organic N, total N, NH_4^-, dissolved organic C, SO_4^{2-}, Cl^-, Ca^{2+}, Mg^{2+}, H^+), averaged over three years, and four watershed variables (maximum elevation, sample elevation, length of stream, watershed area). We will assume that the 39 sites represent a random sample of possible sites in the central Catskills and will focus on point estimation for location and spread of the populations for two variables, SO_4^{2-} and Cl^-, and interval estimation for the population mean of these two variables.

Example

- `lovett.csv`: The variables in the study of 39 stream sites in New York state by Lovett et al. (2000) fell into two groups measured at different spatial scales – watershed variables (elevation, stream length and area) and chemical variables for a site averaged across sampling dates.
- We were interested in testing for correlations between the set of ten chemical variables and the set of four watershed variables (maximum elevation, site elevation, stream length and watershed area) for the 39 stream sites
  - Let's omit the acidified Winnisook site with its extreme concentration of H.
  - For the chemical variables, Let's omit total N `TN` as it was highly correlated with `NO3`
  - Three of the chemical variables (dissolved organic C `DOC`, `Cl`, `H`) and catchment `area` were transformed to log10 to correct skewness.

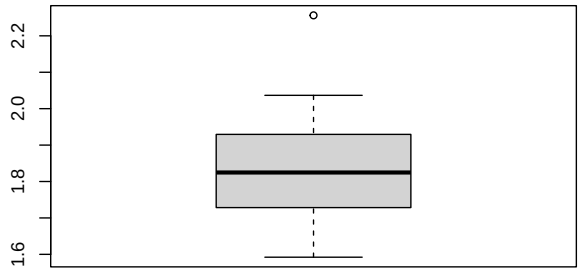
Example

- Boxplots of the variables `Doc`, `Cl`, `H` and `area`

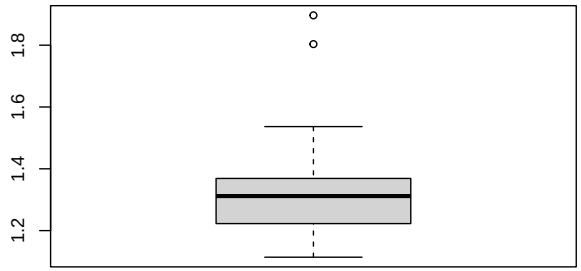


Example

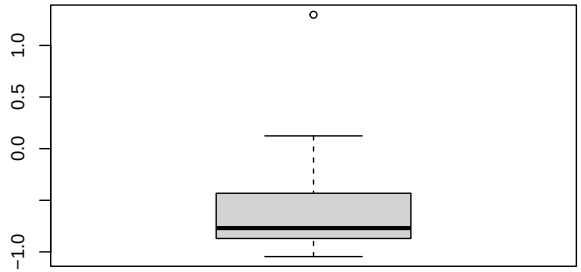
- Boxplots of the variables `Doc`, `Cl`, `H` and `area` after log10 transformation



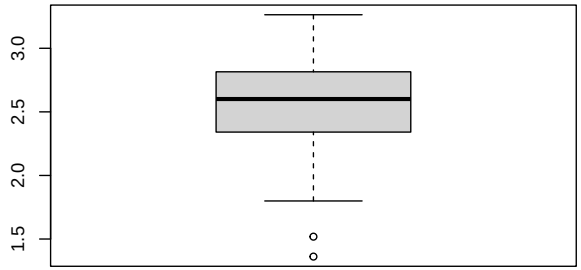
log10 of DOC



log10 of Cl



log10 of H



log10 of area

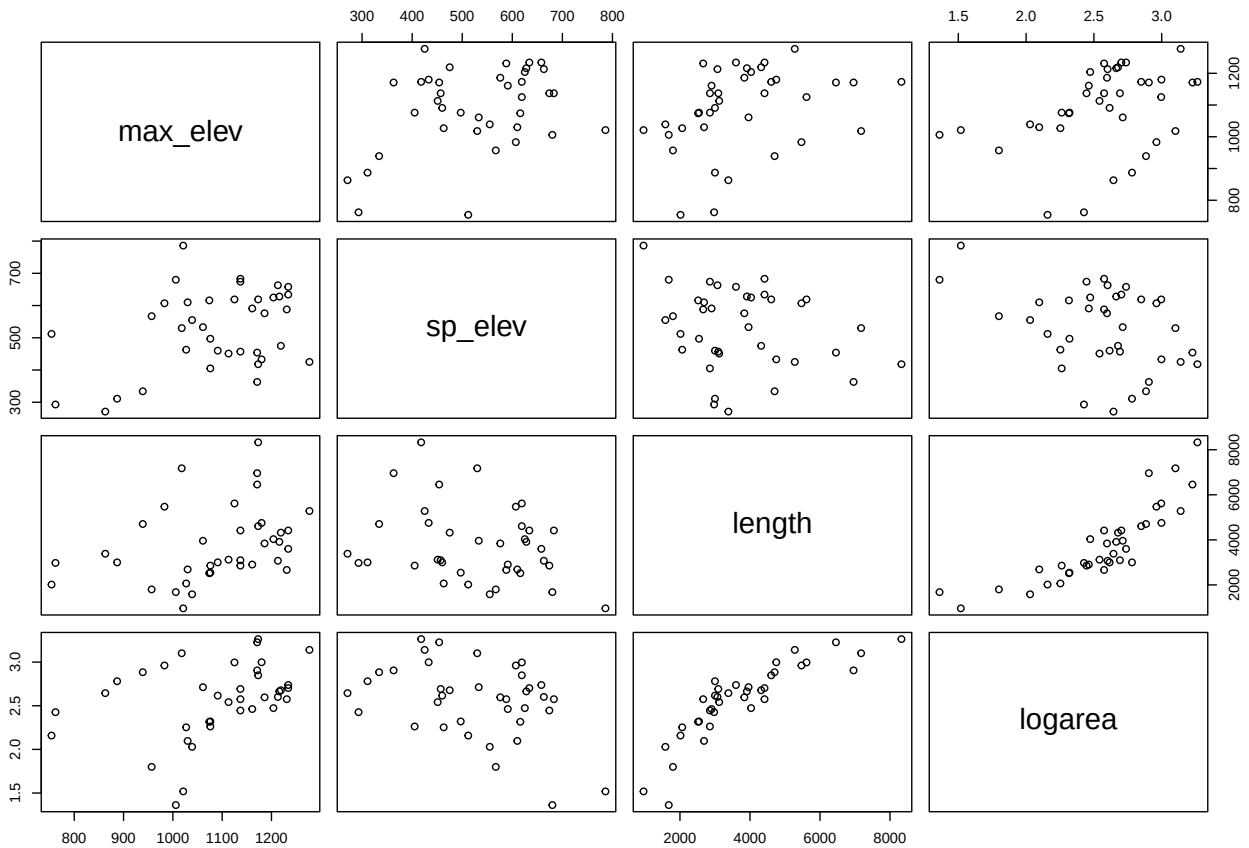
Example

- We use the following data

```
## 'data.frame':    38 obs. of  15 variables:
##  $ ID      : int  2 3 4 5 6 7 8 9 10 11 ...
##  $ name    : chr  "Santa Cruz" "Colgatet" "Halsey" "Batavia Kill" ...
##  $ max_elev: int  1006 1216 1204 1213 1074 1113 1027 1234 1234 1137 ...
##  $ sp_elev : int  680 628 625 663 616 451 463 634 658 674 ...
##  $ length  : int  1680 3912 4032 3072 2520 3120 2064 4416 3600 2856 ...
##  $ NO3     : num  24.2 25.4 29.7 22.1 13.1 27.5 28.1 31.2 22.6 35.9 ...
##  $ TON     : num  5.6 4.9 4.4 6.1 5.7 3 4.7 5.4 3.1 4.9 ...
##  $ NH4     : num  0.8 1.4 0.8 1.4 0.6 1.1 1.4 2.5 3.1 1.4 ...
##  $ SO4     : num  50.6 55.4 56.5 57.5 58.3 63 66.5 64.5 63.4 58.4 ...
##  $ Ca      : num  54.7 58.4 65.9 59.5 54.6 68.5 84.6 73.1 71.1 91.2 ...
##  $ Mg      : num  14.4 17 19.6 19.5 21.9 22.4 26.2 25.4 21.8 22.2 ...
##  $ logDOC  : num  2.26 2.04 2.02 1.93 1.92 ...
##  $ logCl   : num  1.19 1.21 1.23 1.23 1.26 ...
##  $ logH    : num  -0.319 -0.62 -0.328 -0.638 -0.432 ...
##  $ logarea : num  1.36 2.66 2.47 2.6 2.32 ...
```

Example

- Scatter plot matrix of watershed variables



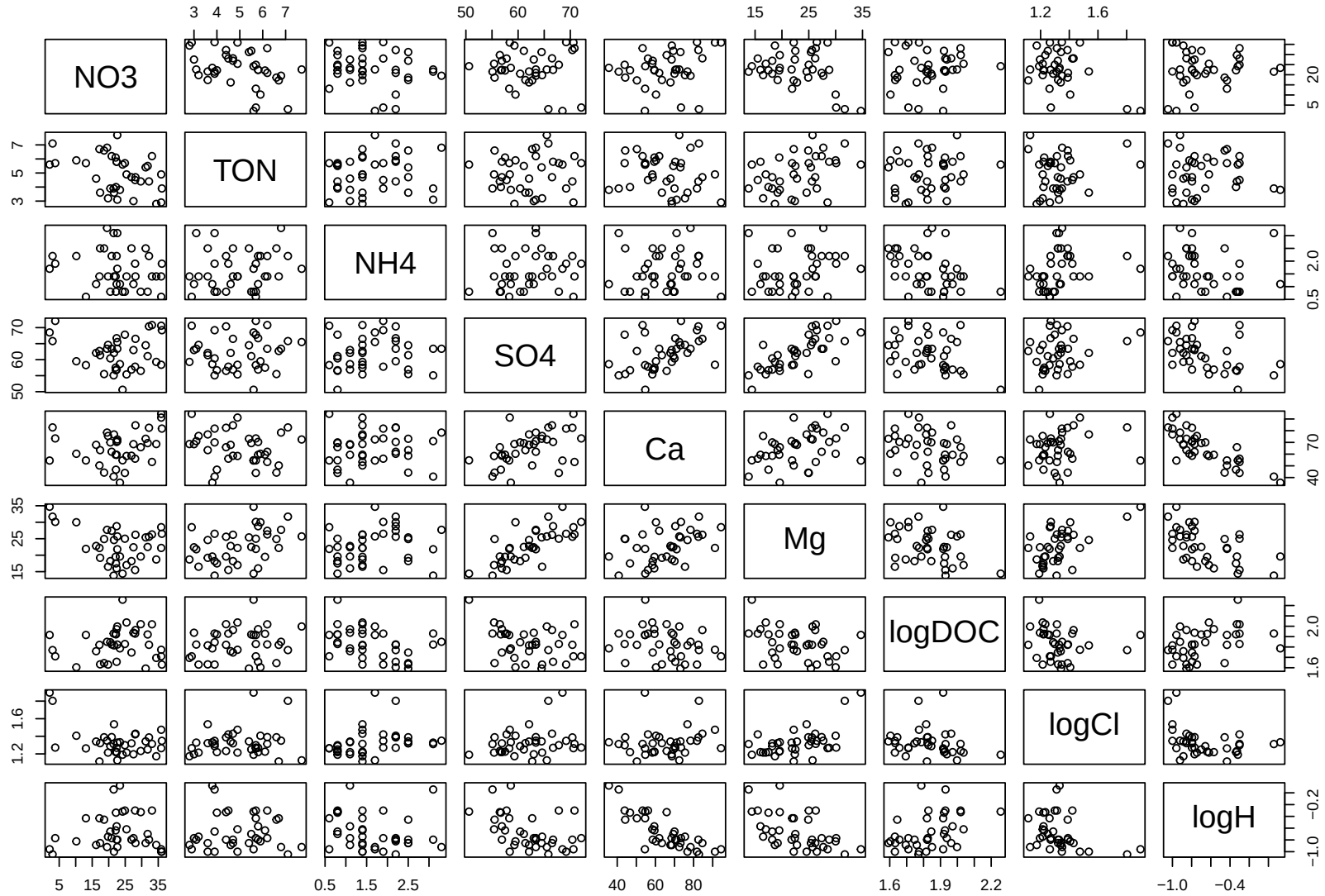
Example

- Correlation matrix of watershed variables

```
##           max_elev sp_elev length logarea
## max_elev     1.000   0.371  0.342   0.364
## sp_elev      0.371   1.000 -0.259  -0.393
## length       0.342  -0.259  1.000   0.851
## logarea      0.364  -0.393  0.851   1.000
```

Example

- Scatter plot matrix of chemical variables



Example

- Correlation matrix of chemical variables

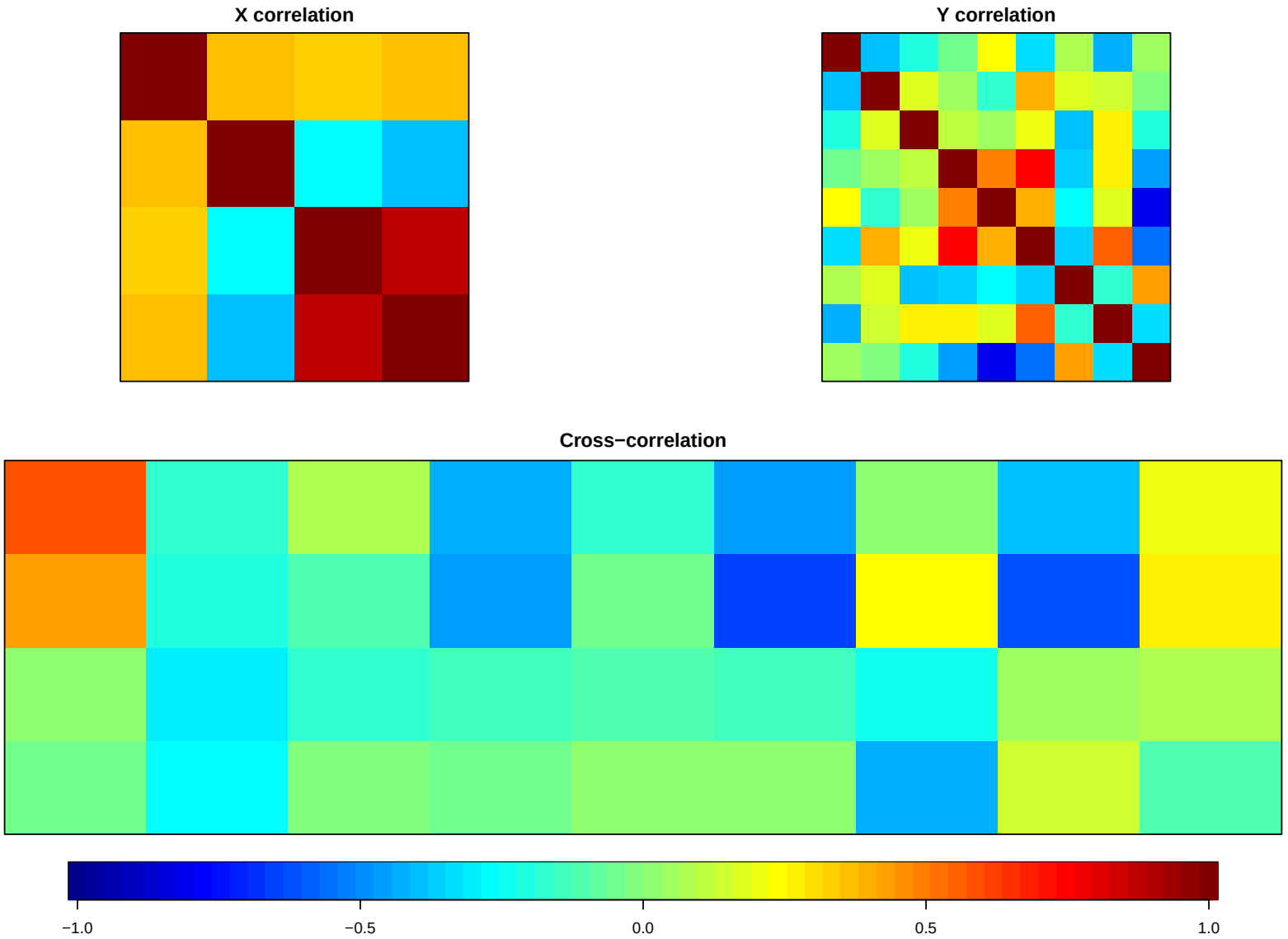
```
##             NO3    TON    NH4    SO4     Ca     Mg logDOC  logCl   logH
## NO3       1.000 -0.396 -0.193 -0.041  0.242 -0.335  0.091 -0.420  0.061
## TON      -0.396  1.000  0.158  0.054 -0.161  0.401  0.185  0.132 -0.010
## NH4      -0.193  0.158  1.000  0.097  0.056  0.208 -0.382  0.280 -0.206
## SO4      -0.041  0.054  0.097  1.000  0.482  0.741 -0.347  0.256 -0.455
## Ca        0.242 -0.161  0.056  0.482  1.000  0.378 -0.266  0.165 -0.812
## Mg       -0.335  0.401  0.208  0.741  0.378  1.000 -0.352  0.552 -0.545
## logDOC    0.091  0.185 -0.382 -0.347 -0.266 -0.352  1.000 -0.161  0.412
## logCl    -0.420  0.132  0.280  0.256  0.165  0.552 -0.161  1.000 -0.334
## logH      0.061 -0.010 -0.206 -0.455 -0.812 -0.545  0.412 -0.334  1.000
```

Example

- Correlation matrix between watershed variables and chemical variables

```
##                NO3    TON    NH4    SO4     Ca     Mg logDOC  logCl
## max_elev     0.590 -0.165  0.077 -0.434 -0.163 -0.455  0.011 -0.386
## sp_elev      0.416 -0.192 -0.117 -0.439 -0.054 -0.646  0.232 -0.605
## length       0.012 -0.311 -0.177 -0.129 -0.123 -0.129 -0.245  0.057
## logarea     -0.033 -0.262 -0.001 -0.048  0.011  0.023 -0.429  0.149
```

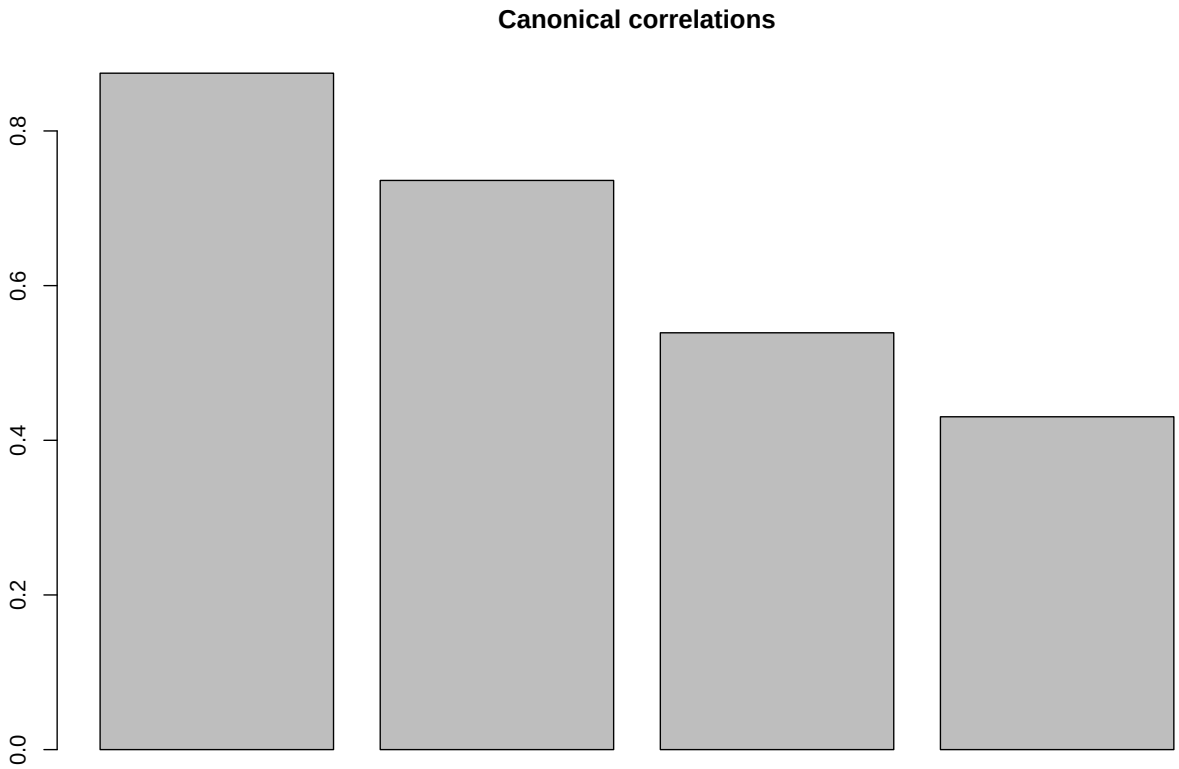
Example


X correlation
Y correlation
Cross−correlation
−1.0
−0.5
0.0
0.5
1.0

Example

- Perform Canonical Correlation Analysis

```
## [1] 0.8746947 0.7360048 0.5390103 0.4304096
```



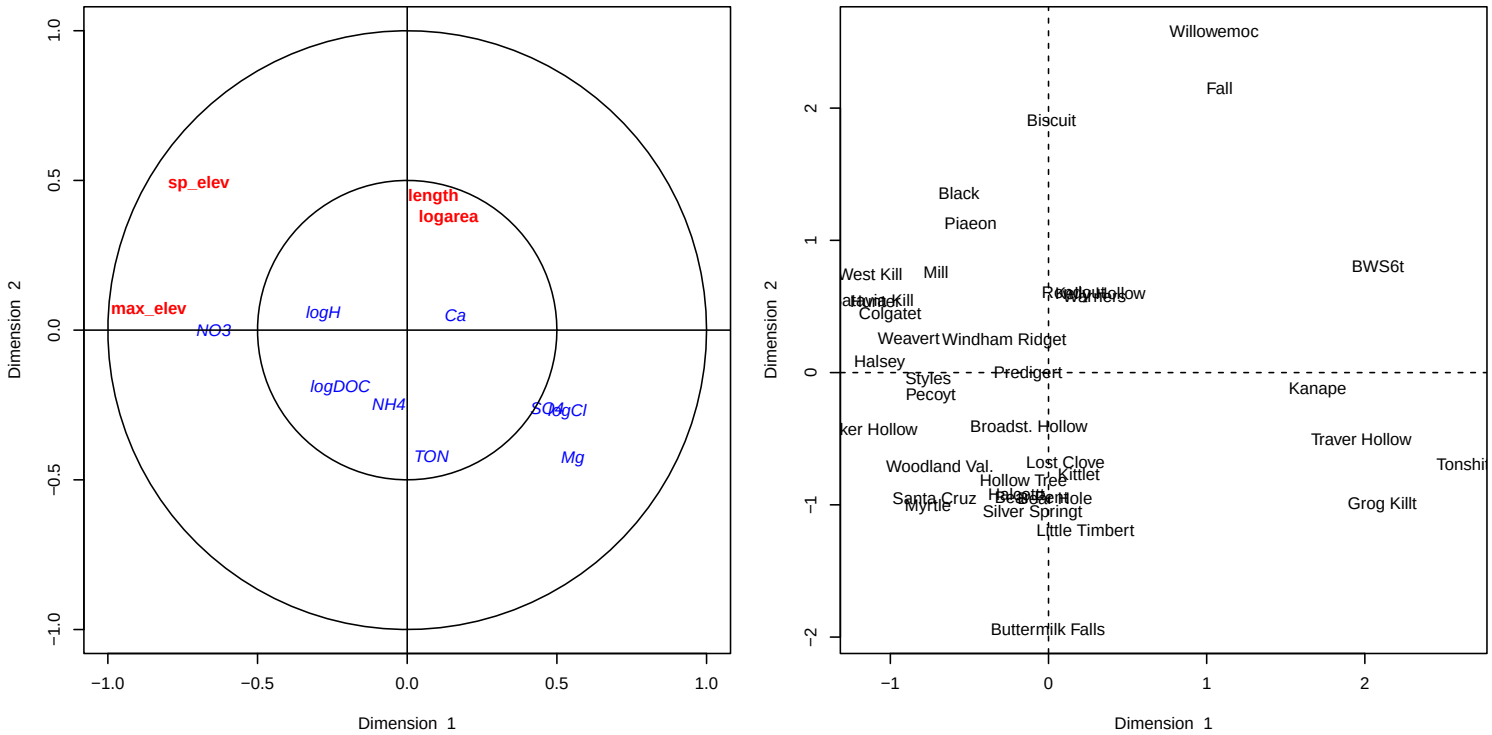
Example

- Estimated linear combinations of the original covariates

```
##             [,1]   [,2]   [,3]   [,4]
## NO3    -0.091 -0.067  0.061  0.047
## TON    -0.157 -0.161 -0.112 -0.203
## NH4    -0.443 -0.575  0.029 -0.734
## SO4     0.066 -0.026 -0.110  0.012
## Ca      0.018  0.044 -0.046 -0.052
## Mg      0.018 -0.142  0.095  0.031
## logDOC -0.605 -4.104 -4.830 -0.553
## logCl   1.735 -0.924  1.660  3.553
## logH    0.683  0.897 -1.980  0.797
```

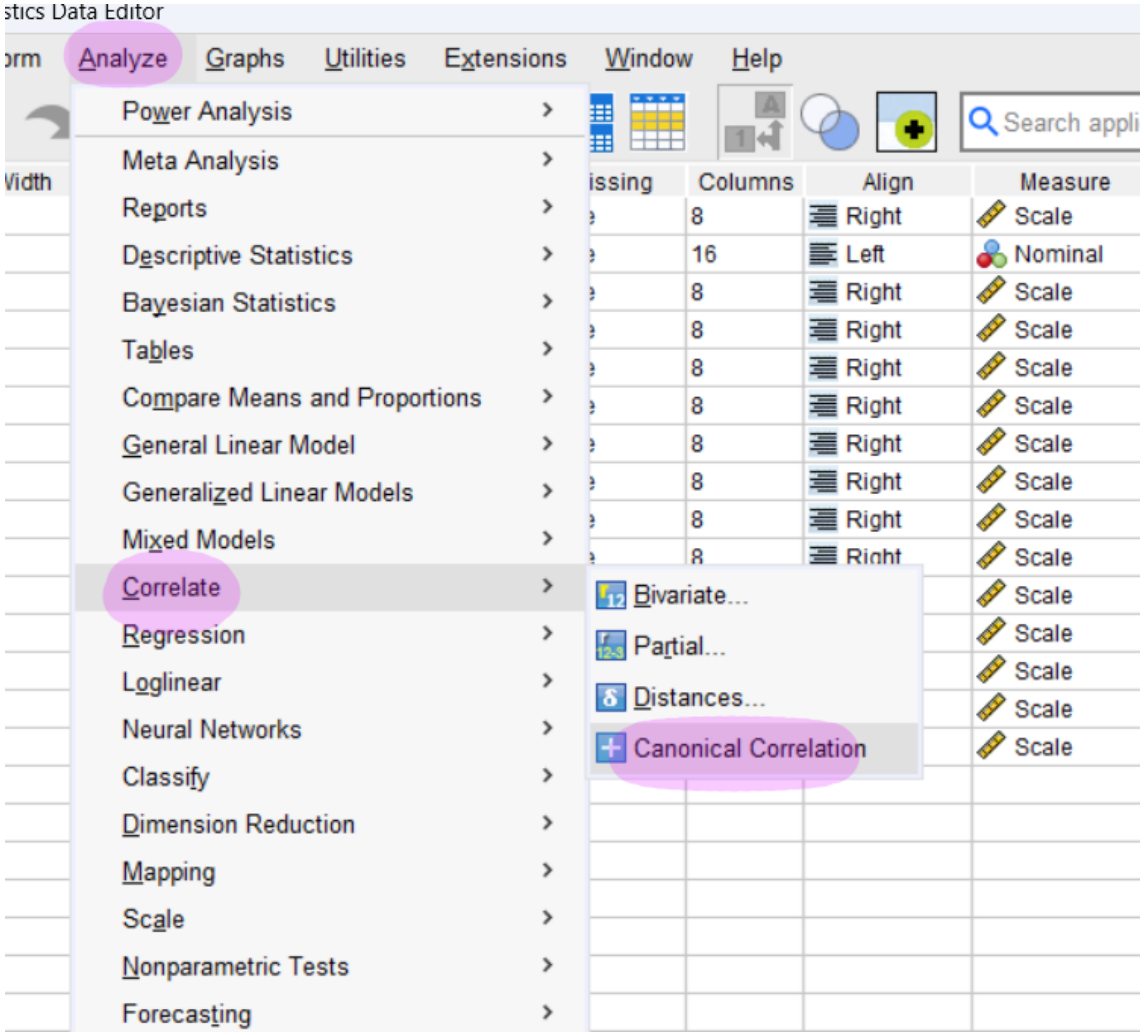
Example

- Display maximized correlations between transformed variables of the two sets of variables
- How to read the left-hand side of the graph
  - Each point represents a variable
  - Points that are close to each other are highly correlated.



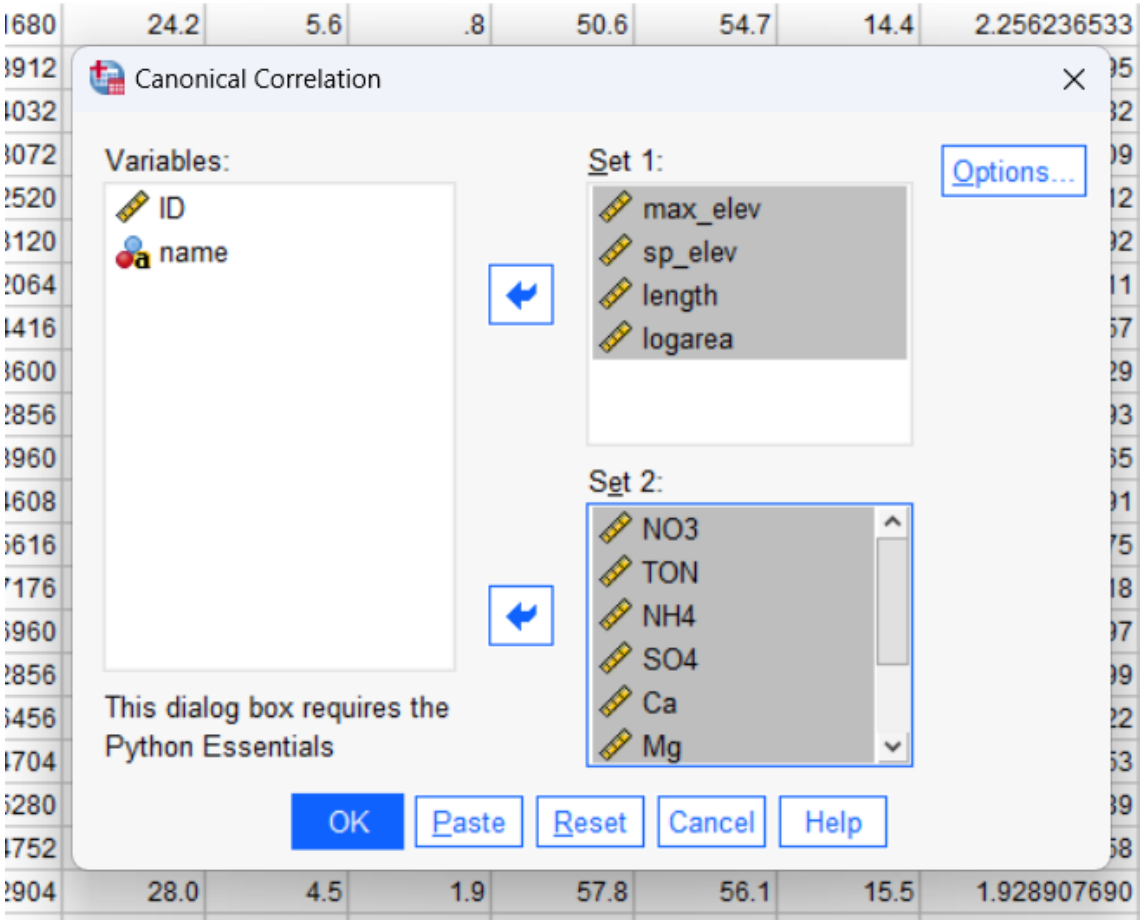
Lab

- After importing data `lovett2.csv`, click on `Analyze` → `Correlate` → `Canonical Correlation`

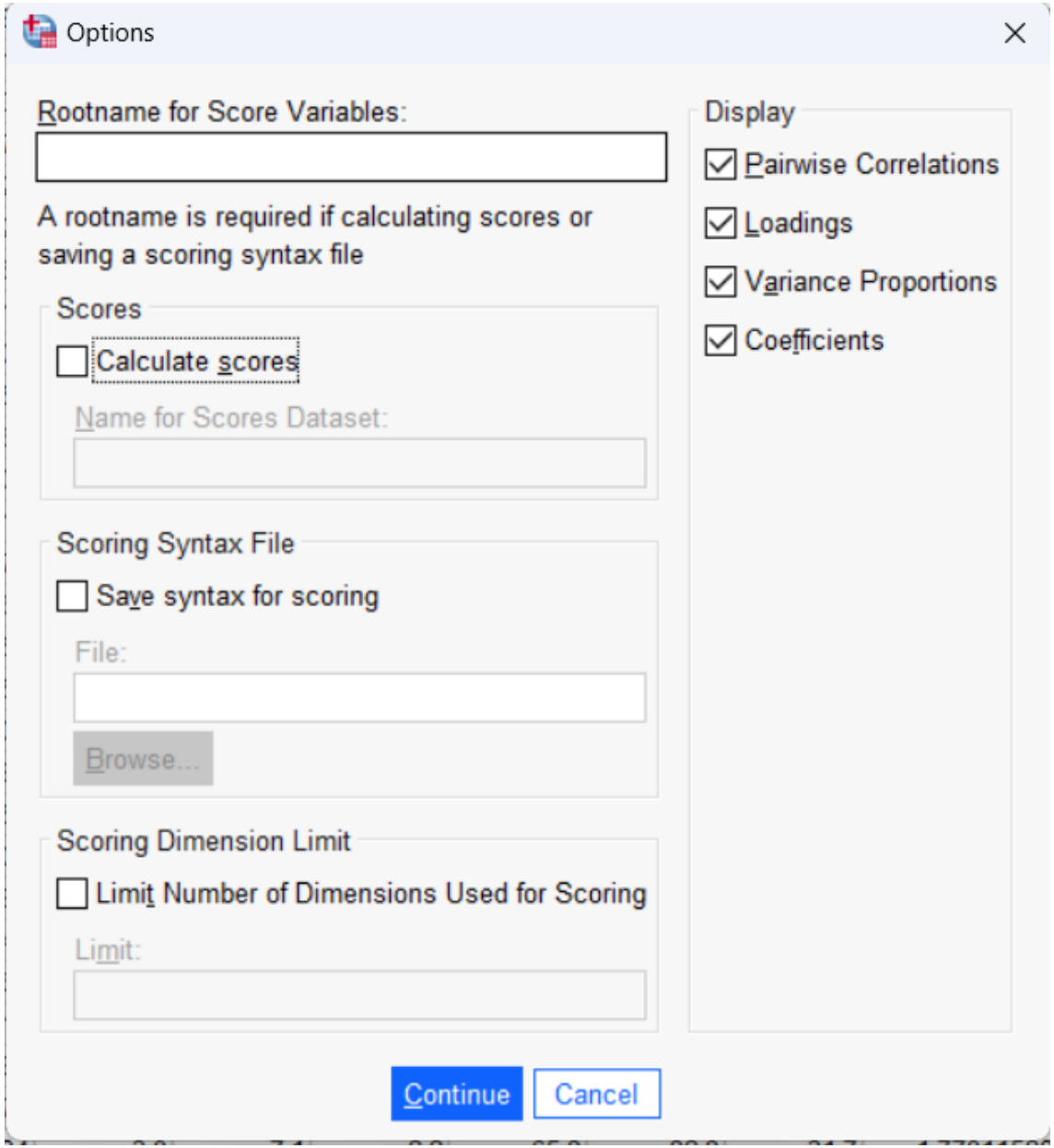


Lab

- Add the numerical variable to two sets:



Lab


Options
Rootname for Score Variables:
A rootname is required if calculating scores or saving a scoring syntax file
Scores
Calculate scores
Name for Scores Dataset:
Scoring Syntax File
Save syntax for scoring
File:
Browse...
Scoring Dimension Limit
Limit Number of Dimensions Used for Scoring
Limit:
Display
Pairwise Correlations
Loadings
Variance Proportions
Coefficients
Continue
Cancel

Lab

Canonical Correlations

[DataSet2]

Canonical Correlations Settings

	Values
Set 1 Variables	max_elev sp_elev length logarea
Set 2 Variables	NO3 TON NH4 SO4 Ca Mg logDOC logCl logH
Centered Dataset	None
Scoring Syntax	None
Correlations Used for Scoring	4

Canonical Correlations

	Correlation	Eigenvalue	Wilks Statistic	F	Num D.F.	Denom D.F.	Sig.
1	.875	3.257	.062	2.911	36.000	95.424	<.001
2	.736	1.182	.265	1.840	24.000	76.009	.024
3	.539	.410	.578	1.216	14.000	54.000	.291
4	.430	.227	.815	1.061	6.000	28.000	.409

H0 for Wilks test is that the correlations in the current and following rows are zero

lab

Set 1 Canonical Loadings

Variable	1	2	3	4
max_elev	-.863	.070	.454	.210
sp_elev	-.696	.489	-.467	-.243
length	.088	.447	.514	.727
logarea	.139	.374	.861	.315

Set 2 Canonical Loadings

Variable	1	2	3	4
NO3	-.739	.003	.075	.151
TON	.093	-.573	-.171	-.263
NH4	-.069	-.336	.367	-.521
SO4	.536	-.355	.021	-.189
Ca	.184	.067	.114	-.608
Mg	.632	-.582	.293	-.214
logDOC	-.255	-.260	-.773	.237
logCl	.612	-.371	.346	.117
logH	-.320	.075	-.407	.631

lab

- In CCA, the proportion of variance explained refers to how much of the variability in each original set of variables is captured by the canonical variates.

Proportion of Variance Explained

Canonical Variable	Set 1 by Self	Set 1 by Set 2	Set 2 by Self	Set 2 by Set 1
1	.314	.240	.202	.155
2	.146	.079	.125	.068
3	.358	.104	.128	.037
4	.183	.034	.143	.026

License