Report for

02935 Introduction to applied statistics and R for PhD students, Winter 2025

Ida Bomholt Dyrholm Jacobsen, sxxxx

# II Summary (less than one page)

Summary

# III Table of content

Table of Contents

[• Introduction. 2](#_Toc188266271)

[II Summary (less than one page) 2](#_Toc188266272)

[III Table of content 2](#_Toc188266273)

[IV Introduction 2](#_Toc188266274)

[• V Description of data. 2](#_Toc188266275)

[• VI Statistical analyses, results and discussion 3](#_Toc188266276)

[• VII Conclusion. 3](#_Toc188266277)

[• VIII Appendices. 3](#_Toc188266278)

# IV Introduction

This report serves as an introduction to the use of Principal Component Analysis (PCA) in the context of ecological studies and vegetation composition analysis. PCA is a widely used method in vegetation ecology as it allows researchers to explore structures and patterns within datasets. The goal of PCA is not to determine the effect of changes in a certain response variable to a predictive varaible, but rather to explore structures. For instance, PCA can help detect groupings of vegetation types or ecotones in the dataset being analyzed. The classification of vegetation into distinct types or ecotones is a common practice in ecology because generalizations and groupings facilitate upscaling and the application of ecological insights across broader contexts.

However, determining clearly distinguishable vegetation types is challenging due to the gradient-like nature of ecological factors influencing plant species’ distributions, occurrences, and ranges and their abiotic drivers, not to mention biotic interactions that may influence these patterns (such as plant-plant competition and symbiotic requirements of some plants). These factors create transitional zones or intermediate types, making it difficult to delineate distinct groups. This is where PCA proves particularly valuable: it provides a means to visualize and understand relative differences between plots (or other units of analysis) in terms of species composition.

PCA additionally has applications in time series analysis, enabling comparisons over time. By doing the PCA scores/positions of plots (units surveyed) on repeat survey data, researchers can assess changes in species composition or abiotic variables. This capability makes PCA a powerful tool for understanding temporal dynamics in ecological datasets.

To keep the focus on the application of PCA in an ecological context and the specific case answering the following research question will be the aim:

Can PCA revial structures (groupings) in the species composition of the vegetation data at hand?

* What species does most contribute to scribe the variance of the data?
* How might this relate to measured soil moisture?

# • V Description of data

The data process in the following report was collected in [Kangerluasunnguaq](https://da.wikipedia.org/wiki/Kangerluarsunnguaq) (Kobbefjord, Nuuk) in Southeast Greenland in 2024. The data was collected in 100 plots placed in an area of interest of aproxemately 12 km2 by randomised stratified sampling. Stratification was based on elevation (5 bins) and NDVI (normalised difference vegetation index, aka ‘greeness’) (4 bins).



Figure While surveying the plot was marked by a metal cross. Soil moisture and soil temperature was measured at each end of the cross (blue triangles).

Within each circular plot of 1 m2 the abundance and maximum height of all vascular plant species was recorded. Abundance was assessed by means of the well known Braun-Blanquet scale (8 step version). Abundance was assessed for both bryophytes and lichens collectively. For data processing purposses the 8

In all plots soil temperature, soil moisture, and general vegetation height, was measured. Soil temperature was measured with a generic thermometer (°C, 4 measurements, mean calculated). Soil moisture was measured with a ThetaProbe (% water content, 4 measurements, mean calculated). General vegetation height was measured with a generic ruler (cm, 4 measurements, mean calculated)

The final processed data contains data from 100 plots and 72 species of vascular plants.

# • Processing of data

Data was processed in R. The raw data consists of a dataframe with the dimensions of 100 rows (plots) and 80 colomns (75 plants species abundance (%), abundance of lichens (%), abundance of bryophytes (%), soil moisture (%), soil moisture group (3 level factor). The soil moisture groups were based on splitting the data by the 33.33 % and 66.66 % quantiale yeilding 3 approximately similir sized groups of moisture (low, medium, high).

## Analyses

### Principal Component Analysis

The data is scale before running the PCA. The results of the PCA with the PCA function from the ChemometricsWithR package in R revieal that for this full data the PCA model is a mean centered matrix of 100 (plots) by 73 (plant species ground cover). 37 principal components are needed to cover 90 % of the variance of the data.

|  |  |  |
| --- | --- | --- |
|  | Variance | Cumulative variance |
| PC1 | 6.990096 | 6.990096 |
| PC2 | 6.189466 | 13.179562 |
| PC3 | 5.185294 | 18.364856 |
| PC4 | 4.912269 | 23.277125 |
| PC5 | 4.147000 | 27.424125 |

Table Variance and cumulated variance of the first 5 PCs (principal components).

The head of the table ([Table 1](#table_01)) from the list of principal components shows that PC1 explains 6.99 % of the variance. The accumulated explained variance up to the 5th principal componant is only approximately 27.4 %.

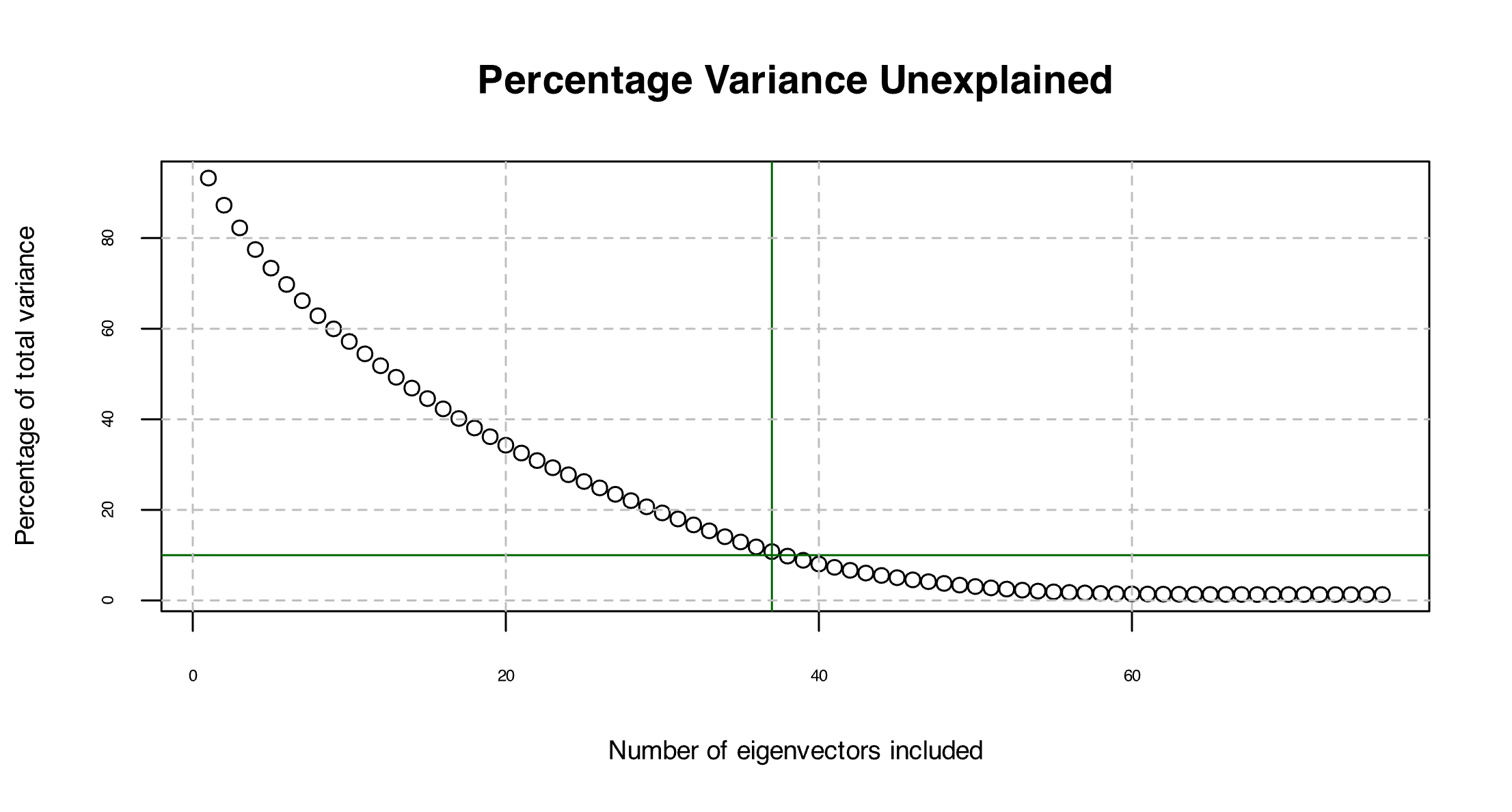


Figure Skree plot. Visualising the percentage of unexplained variance as a function of the eigenvectors included. Note that no 'elbow' is visible.

The scree plot viasulasis the high number of eigenvectors need to explain 90 % of the variance in the data (37).

### Diagnostics

### PCA on subsettet data I

## Diagnostics

### PCA on subsettet data II

## Diagnostics

## (Cluster analysis)

## Results

It is the result of the PCA that a high number of variables (species) are needed to describe the variation of the dataset. This means that visuliasing two principal components captures relatively little of the variation of the data.

## Discussion

While the general sampling scheme of this data could be use to collect data for classification this data has not been collected for that specific purpose, nor in a way to couple this directly to soil moisture. For this purpose more plots would have to be inventoried in stratefied sampling that includes the soil moisture.

Furthermore, the soil moisture data at hand was collected at one specific time that might be indfluenced by that days weather. A more robust picture of the differences in soil moisture between the plots will come from loggers installed at the plots to be collected this coming summer.

# • VII Conclusion

• References

# • Appendices

## Client decription

The client, i.e. the receiver, of this report is my future self. I have basic understanding of statistics, statistical methods, and want to further my expertise in this areas both to explore the data I collect and have available as well as to document known phenomena of this same data. I have advance knowledge in biology and ecology. I do not have extensive of intuitive understanding of statistics and this report is aimed at document the learning outcomes of the data processing with the purpose of statistical reporting.

It is my interest to gain an applied and hand on approach to statistics, answer the reserach question at hand, explore the data I have collected

- What does the client already know? (basic/advanced science on the subject, statistical methods, project circumstances)

- What does the client not know? (basic/advanced science on the subject, statistical methods, project circumstances)

- What is the interest of the Client? (research question, p-values, effect parameters, issues with data handling)

- What is NOT the interest of the Client? (R code, issues with data handling, intermediate analyses)

- Adapt the contents and structure (not the results though ) to fit the knowledge and interests of the Client.

## Code

Importing data

Plotting data

PCA analysis