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

Table of Contents

[II Summary (less than one page) 2](#_Toc188603143)

[IV Introduction 2](#_Toc188603144)

[Description of data 3](#_Toc188603145)

[Processing of data 3](#_Toc188603146)

[Analyses 3](#_Toc188603147)

[Principal Component Analysis of all data 3](#_Toc188603148)

[PCA on subsettet data I 8](#_Toc188603149)

[PCA on simplefied data 9](#_Toc188603150)

[Results 9](#_Toc188603151)

[Discussion 9](#_Toc188603152)

[Conclusion 9](#_Toc188603153)

[Appendices 9](#_Toc188603154)

[Client decriptionw 9](#_Toc188603155)

[Code 10](#_Toc188603156)

# 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 with a particular emphasis on the diagnostics of PCA. 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?

# 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 1 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 75 plants abundances.

# 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 (%), including 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 percentile yeilding 3 approximately similir sized groups of moisture (low, medium, high).

## Analyses

### Principal Component Analysis of all data

The data is scaled before running the PCA to ensure equal contribution of the variables (plant cover types to the analysis). 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 1 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 %.

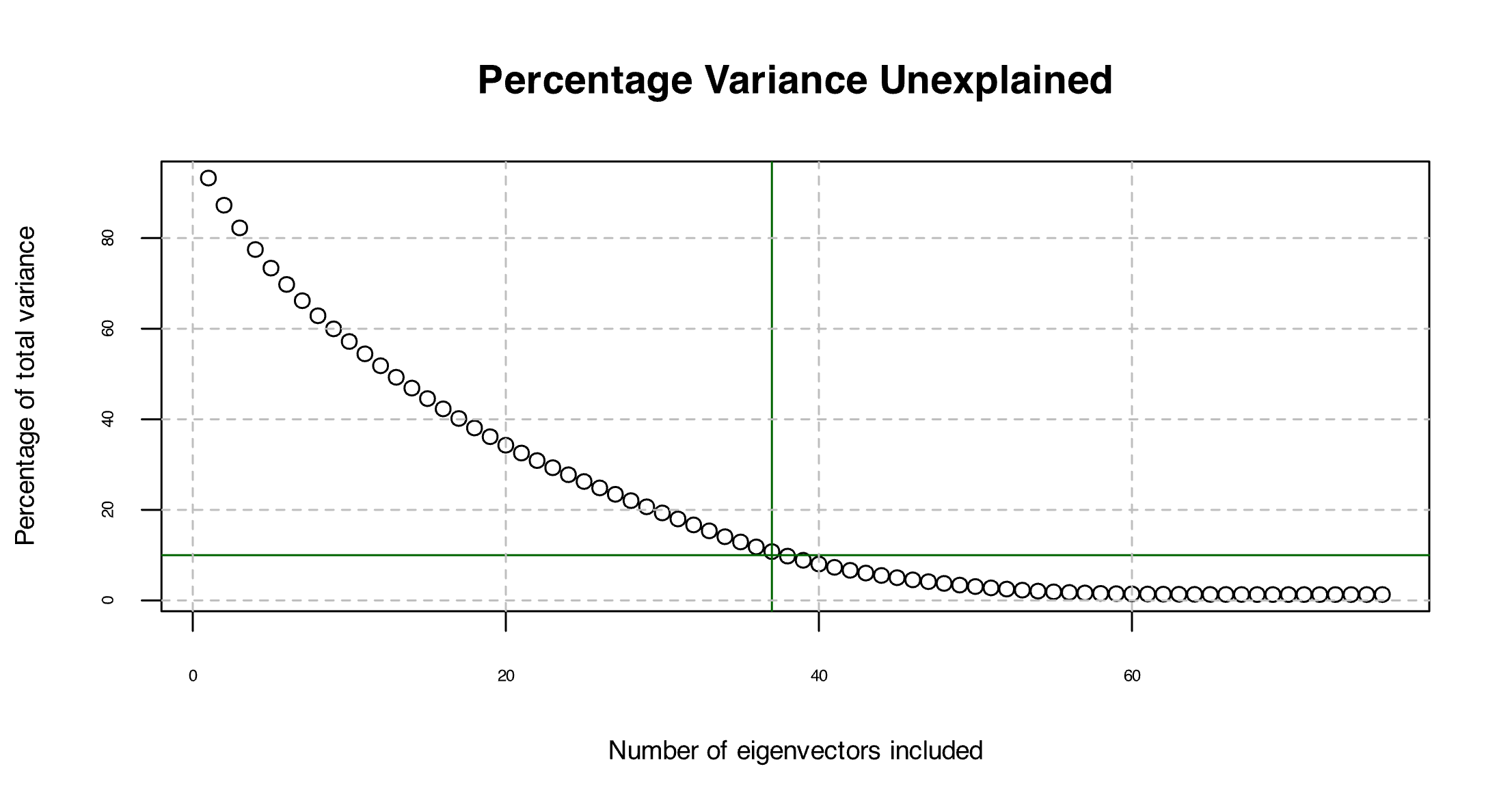


Figure 2 Skree plot. Visualising the percentage of unexplained variance as a function of the eigenvectors included. Note that no 'elbow' is visible.

The scree plot viasulasis the high number of eigenvectors need to explain 90 % of the variance in the data (37) ([Figure 2](#table_02)).

#### It is evident from the score plot that there are no very distinct groupings or structure of the plots (and their species composition) that relates to the measured soil moisture. The elipses, indicators of the groupings of soil moisture, have different orientations, but overlap a great deal. It is also evident from figure 3 that there are 6 suspected outliers (plots MP041, MP051, MP086, MP024, MP034 and MP088). The rest of the plots are quite similar in terms of the species composition described by PC1 and PC2, but could be very different with regards to the remainding PCs (see Diagnostics, Figure 4).

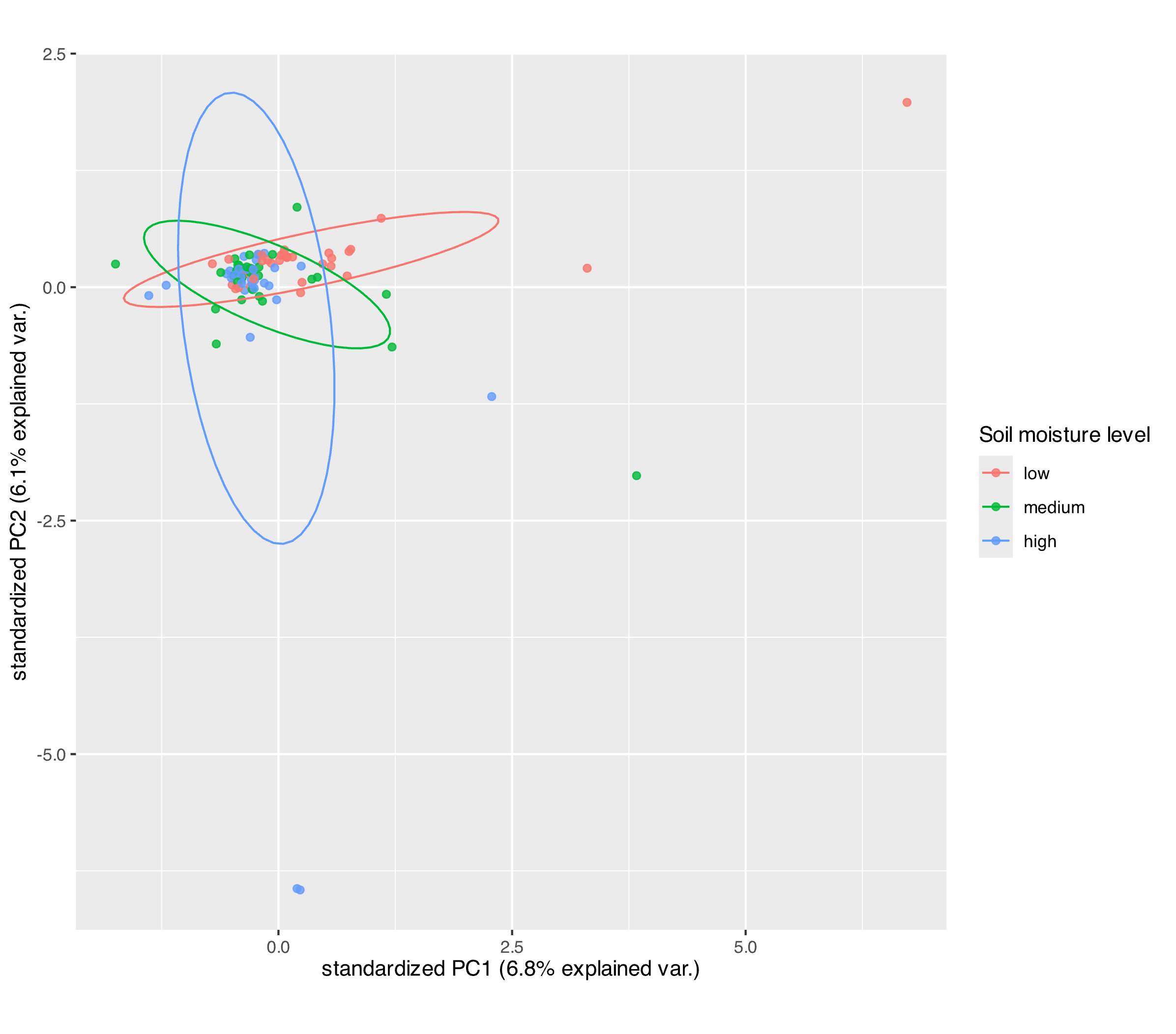


Figure 3 Scoreplot of the PCA with elipses indicating the grouping of soil moisture (biplot without loadings).

#### Diagnostics

Important measures in the diagnostics of the PCA plot is to look at the leverage (score distance) and residuals (orthogonal distance) of the principal components. This can be done both by looking at the the distance on their own ([Figure 3](#figure_003)) and leverage and residual in relation to each other (Figure Y).

**Score distance or leverage** describes how fra from the center of the PCA space the unit is. In other word the higher the score distance the more extreme the observation is, in terms of the pricipal components and has big influence on the area of the PCA space. Very high score distances could be indication of outliers.

**Orthogonal distance or residual** is a measure of the perpindicular distance of observation to the PCA space. In other word is it a measure of how well the PCA represent the observation. When plotting the orthogonal distance and the score distance the key for plot plots is that the values are as close to 0 as possible.   
[Figure 4](#figure_003) shows the orthogonal distance and the score distance of the PCA at hand. Notably there are no distances that are 0. For the orthogonal distance it is expected to have relatively higher values as we already know that relatively little variation was discribed by each PC, meaning there would be a relatively high residuals. But approxemately 10 % are outside the 95 % confidence interval of this PCA. Quite a larger number of observations are outside the confidence interval for the score distance.

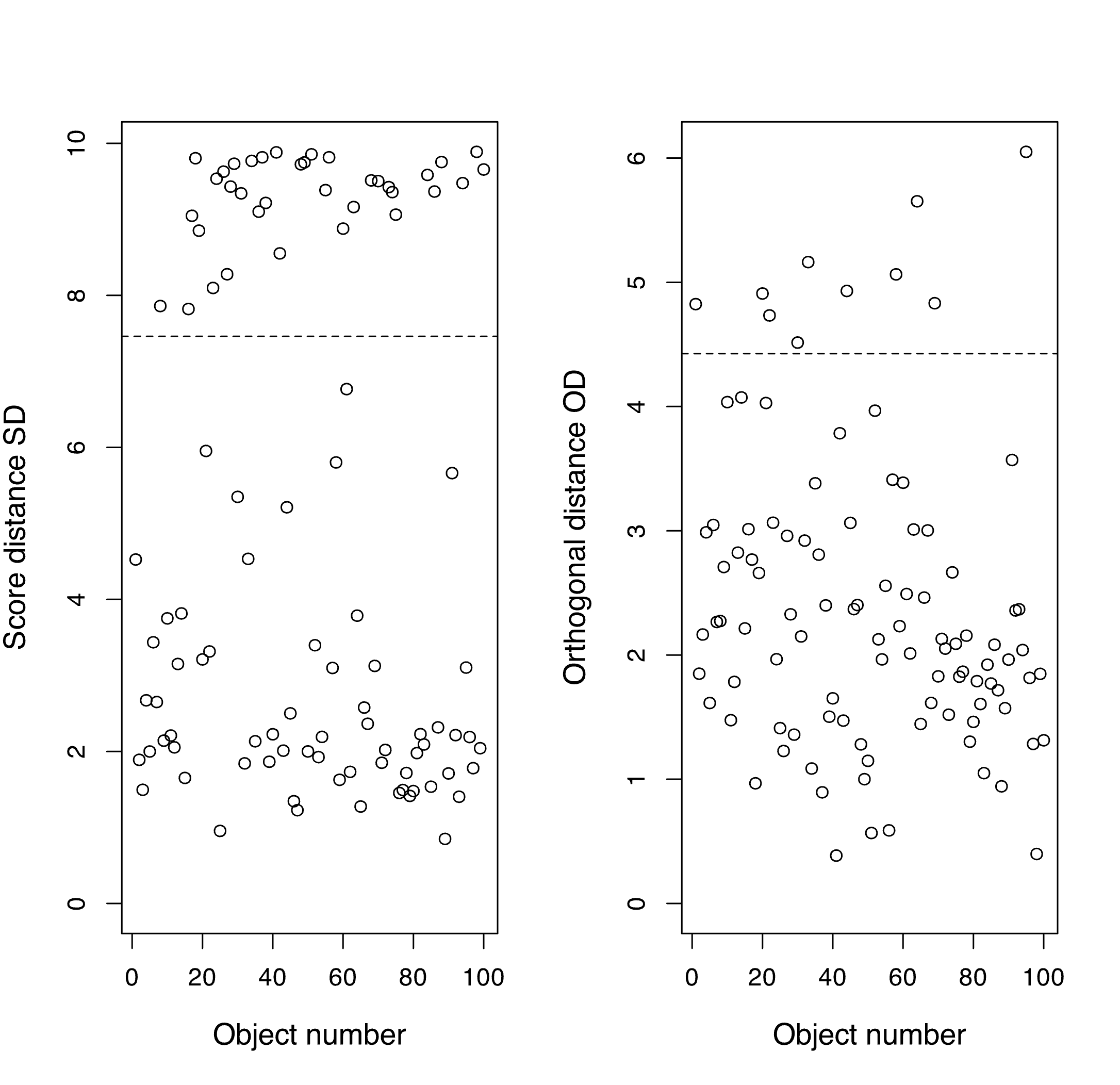


Figure 4 Score distance and orthogonal distance with 37 principal components. Score distance (left) = leverage, Orthogonal distance (right) = residual.

Figure 4

(D)

(C)

(B)

(A)

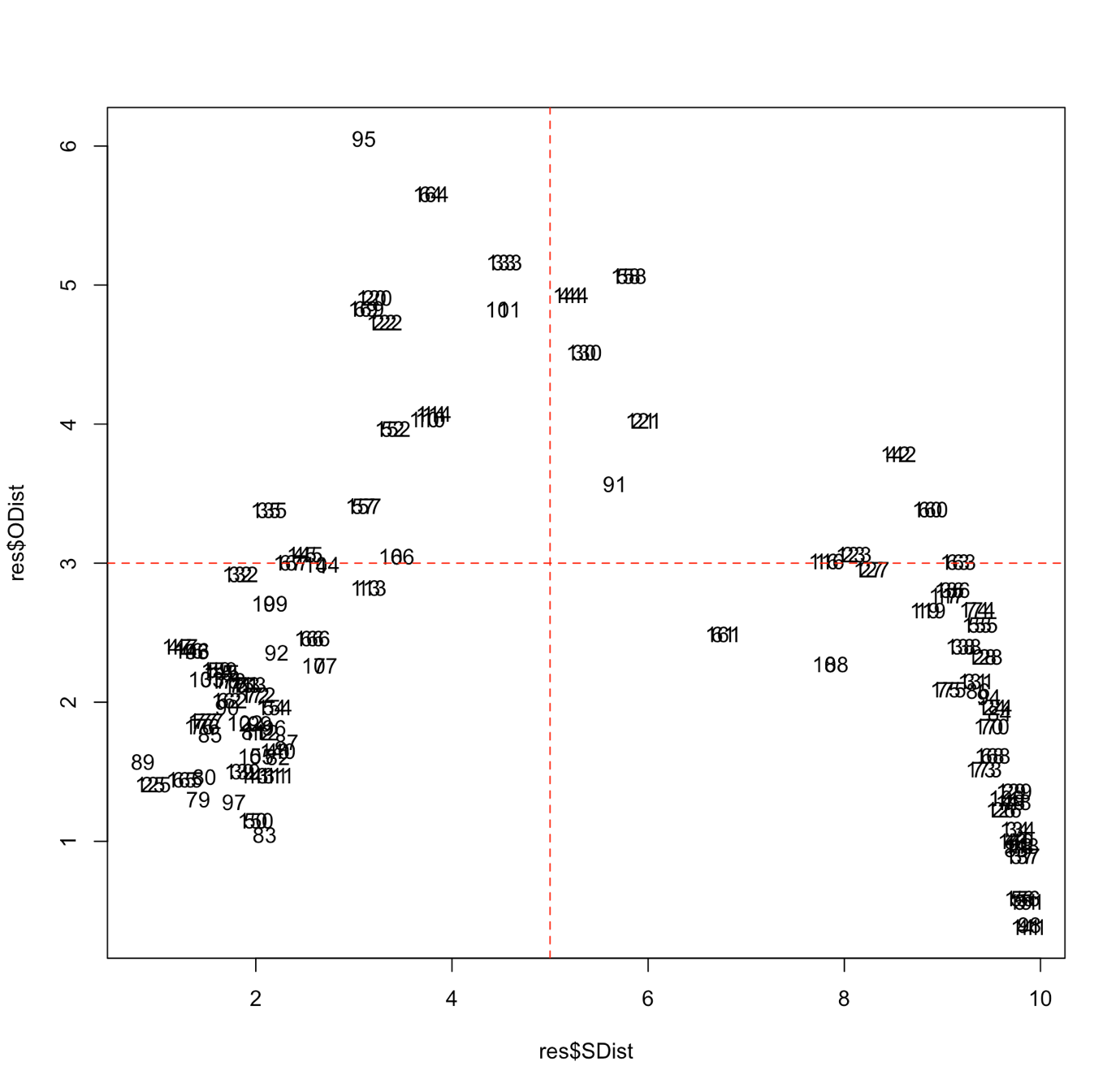


Figure 5 Leverage (score distance) plottet against residuals (orthogonal distance). Red dotted lines represent relative high/low residual or leverage.

[Figure 5](#figure_004) shows each observation in terms of their score distance and orthogonal distance. Plot is split (relatively) in to 4 squares of respectively high/low score distance and high/low residual. Ideally in this type of plotting most observations would be in the **lower left sqaure (A)** (low residual *and* low leverage). Observations in the **lower right square (B)** has hight score distance but low orthogonal distance, meaning that they are far form the center of the PCA, but represented well byt the PCA. Observations in the sqaure could be natural extremes of the data. Observations in the **upper left square (C)** have low score distance/leverage, but high residual. That means that they are close to the center of the score space, but not very well represented by the PCA, that could indicate e.g. noise of the data or features not captured very well by the PCs. Observations in the **upper left corner (D)** are problematic as they have both high leverage/score distance *and* high residuals/othogonal distance. This means that they

### PCA on subsettet data I

To do further exploration of the data and the plot observations very closely lumped in figure X, this section will explore the PCA with out the possible outliers identided in figure 3. These are the plots MP024, MP034, MP051, MP86 and MP088. These plots standout in the way that some have almost no vegetation and others are plot that contain species that have been seen in no other plots. This means that removing those 4 plots also eleminates four species (*Luzula parviflora*, *Listera cordata*, *Pyrola grandiflora* and *Equisetum silvaticum*).

With this further data processing the final dataframe for the PCA is a matrix of 94 rows (plots) and 73 colomns (plant covers).

|  |  |  |
| --- | --- | --- |
|  | Variance | Cumulative variance |
| PC1 | 7.309359 | 7.309359 |
| PC2 | 6.054444 | 13.363804 |
| PC3 | 5.389616 | 18.753419 |
| PC4 | 4.493791 | 23.247211 |
| PC5 | 4.136437 | 27.383648 |

Table 2 Summary of PCA on selected data. Variance ecplained and accumulated variance explained for the first 5 PCs.

Notes that the variance explained of PC1 has increased a bit, but that the cummulated explained variance of 5 PCs is very similar. One less PC is needed to explain 90 % of the variance; 36.

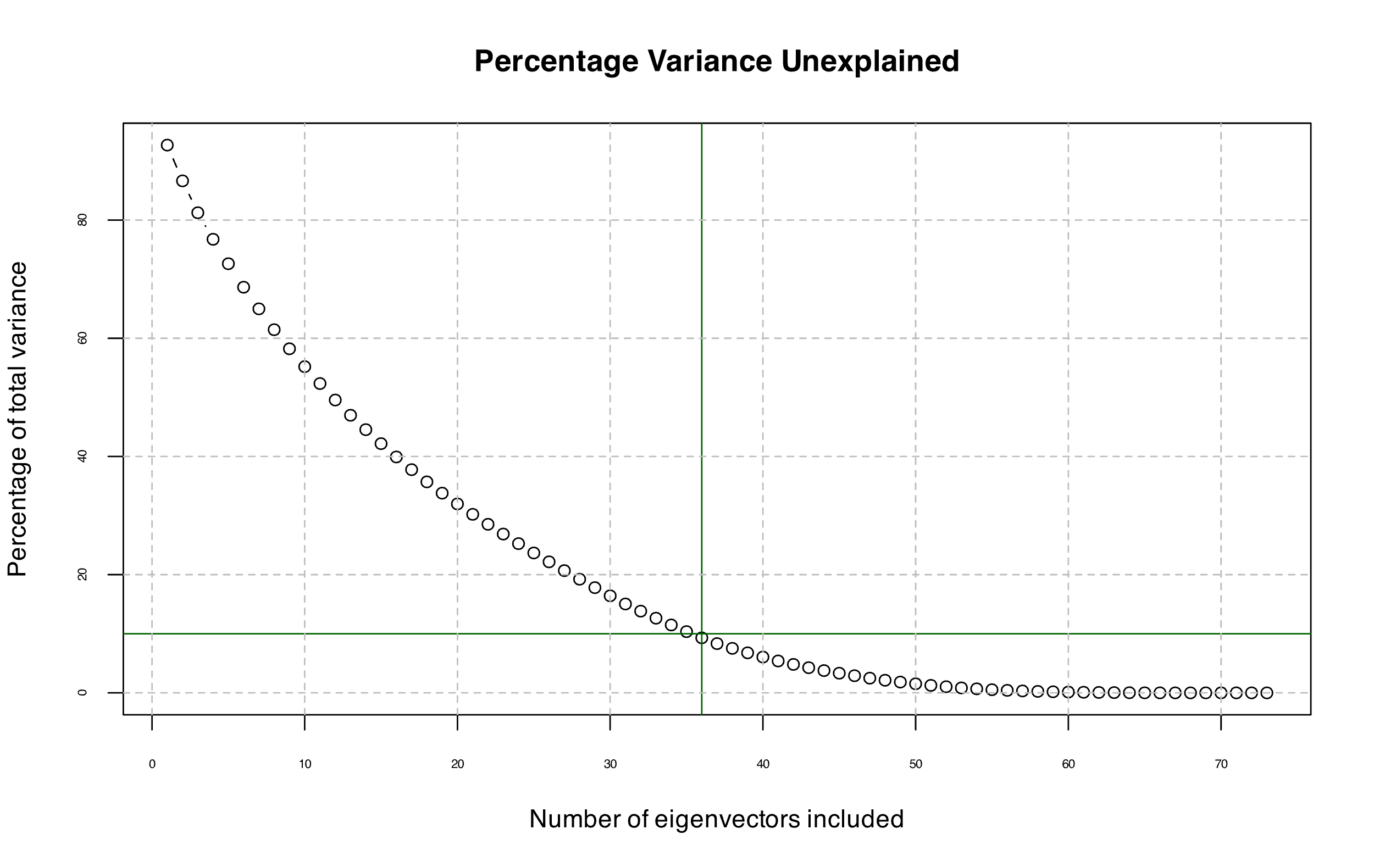


Figure 7 Skree plot of the PCA of the curated data.

Figure 7, depicting the skree plot of curated/subset data, still shows the pattern of many PCs making continues small contributions the eplaining the variance of the data.

#### Diagnostics

### PCA on simplefied data

A secund approach to looking at potential data structures or groupings is to simplify the data and look for structure based only of the functional groups of the plants. This colapses the data in to a dataframe of 100 plots and x variables (% cover of graminoids, evergreen shrubs, decidious shrubs and forbs).

Simplifying the dataset: functional groups, vegetation height, soil moisture

#### Diagnostics

# Results

For PCA on the full data a high number of variables (species) are needed to describe the variation of the dataset. This means that visuliasing two principal components (Figure X) captures relatively little of the variation of the data. In addition the diagnostics of the PCA reveal high levels of score and orthogonal distance respectively indicating (including some points that have both high score level and high orthogonal distance).

# Discussion

While the general sampling scheme of this data has potential 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 and target specific areas with distinct ecotypes to ensure a baseline classification.

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.

# Conclusion

PCA on the full data does not reveal any distinct patterns or structures. This is evident from both lact of structure and groupings within the plots on their own as well as any structures or groupings related to soil moisture. Furthermore, the poor dianogstics of the PCA analysis with the full data indicates that PCA does not adequately represent the data.

• References

# Appendices

## Client decriptionw

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