Case of Study: Water Quality

Eudald Romo y Laura Santulario Verd?

05/21/2018

Prior Analysis

Let's first examine the main characteristics of our data, to base the following points of our study on them.

As stated in the dataset information slides, the data represents the contamination of sea ecosystems at different points. We have 57 different data points and each of them provides information for 7 different variables.

All variables are real non-negative numbers and some of them clearly represent similar kinds of contamination. For instance, **Colif_total**, **Colif_fecal**, and **Estrep_fecal** represent all organic contamination and are highly correlated, as seen below.

Lastly, it is important to note that there's a really reduced number of samples. The size of the data will be further analized when deciding upon a discriminant for the discriminant analysis, but a first observation is that no cluster analysis is going to be very robust or meaningful if any group has less than 10 variables, this means that it is highly unlikely that we have to analyse our groups hypothesis for more than around 6 groups (we'll do it anyway for up until 8 groups for the sake of illustrating what we qualitatively observed by using the standard indicators for clustering quality.).

```
##              Colif_total Colif_fecal Estrep_fecal
## Colif_total        1.000       0.988        0.841
## Colif_fecal        0.988       1.000        0.861
## Estrep_fecal       0.841       0.861        1.000
```

With this information, we can already choose a distance matrix.

Distance matrix

Our dataset clearly does not intend to evaluate profiles, it just characterizes different samples of a certain set of variables of interest to (presumably) check similatities or dissimilarities at certain points. Furthermore, no specific value of any of the variables has more weight than others, in particular, it does not make any difference if there is a really low level of contamination of any kind in a certain point or if there's none (handling the case where a variable has 0 value only makes sense in situations where we are studying variables like human behaviour, where theres a clear difference between showing a slight interest in a topic or showing none).

This leaves out with the choice of using either Euclidean or Mahalanobis distances. The high correlations shown above would advocate in favour of the Mahalanobis one, but it was explained in one of the practical sessions that in the original study that used this dataset, it was intended that certain aspects of the ecosystem, like organic contamination, had a higher weight than others. Because of that, we decided to keep those weights and use the Euclidean distance. As the variables have different units (and in particular, quite different means/standard deviations) it is important to standardize the data before computing the distances. Below is a subset of the obtained distance matrix.

```
##          1        2        3        4        5
## 1 0.000000 5.183625 5.618755 6.067342 3.350898
## 2 5.183625 0.000000 2.471269 5.817088 3.721250
## 3 5.618755 2.471269 0.000000 3.787447 3.037862
## 4 6.067342 5.817088 3.787447 0.000000 3.665658
```

```
## 5 3.350898 3.721250 3.037862 3.665658 0.000000
```

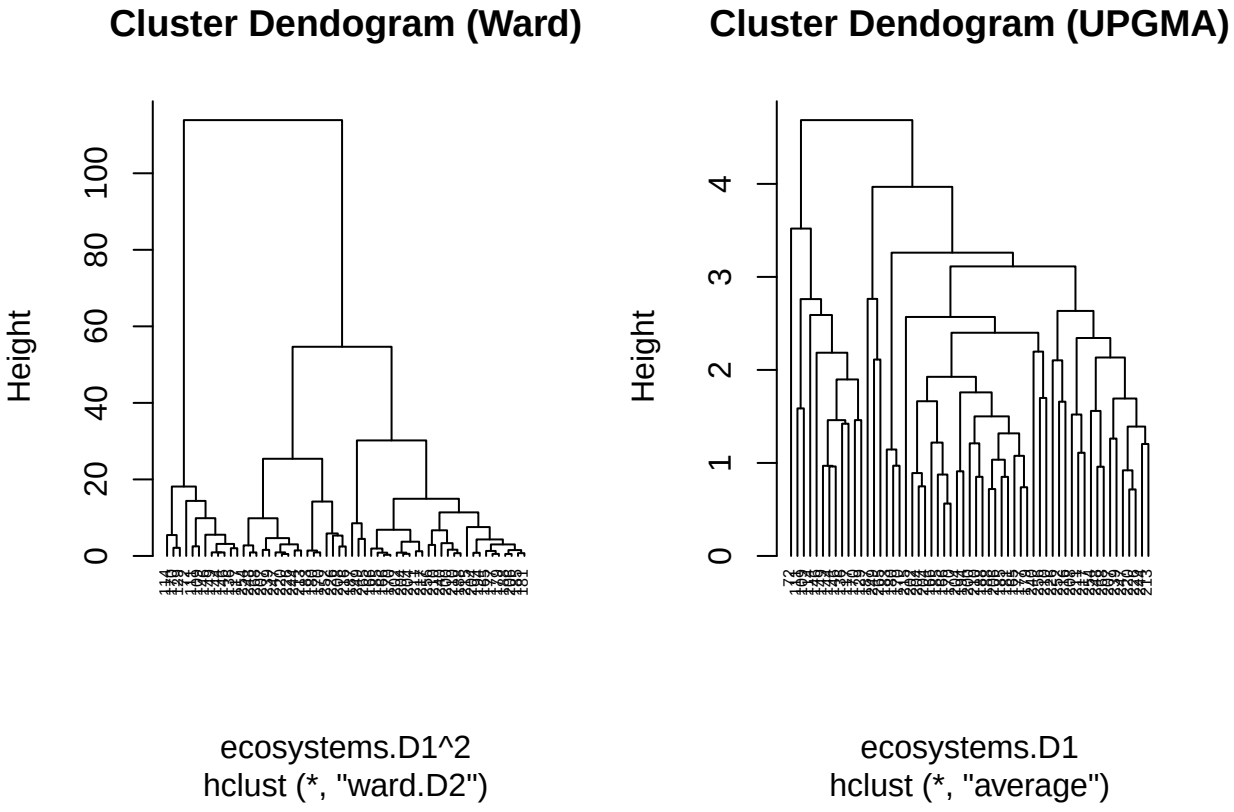
Group Structure

We're going to both do a first qualitative analysis of the group structure through hierarchical clustering methods and then corroborate and refine our first results through a more objective analysis using clustering quality indicators as $\Delta TESS$, *pseudoF*, and *silhouette* for 1 to 8 groups (as we already explained before, more than 5 groups would result in some groups having less than 10 samples, thus makin any further inference or discriminant analysis really poor.).

Hierarchical Clustering

At class we have considered the *single*, *complete*, *average (UPGMA)* and *ward.D2* methods. In this report, we chose to use the *ward.D2* as it tries to minimize the variance. This is coherent with most analysis that use Euclidean distance, like PCA. We expect this method to be more congruent with the distance chosen than others like *UPGMA*, which don't even need a fully defined distance, just a dissimilarity matrix, which doesn't even need to fulfill the triangular inequallity. We know that the hierarchichal clustering can be subjective, as the election of the clustering method can be biased or even have some arbitrariety. So, for the sake of completeness, we are providing two different hierarchical analysis, *ward.D2* and *UPGMA*.

As can be seen below, a much clearer analysis can be done in the *ward.D2* clustering, and it seems that the most possible choices for number of groups are 2 or 3.

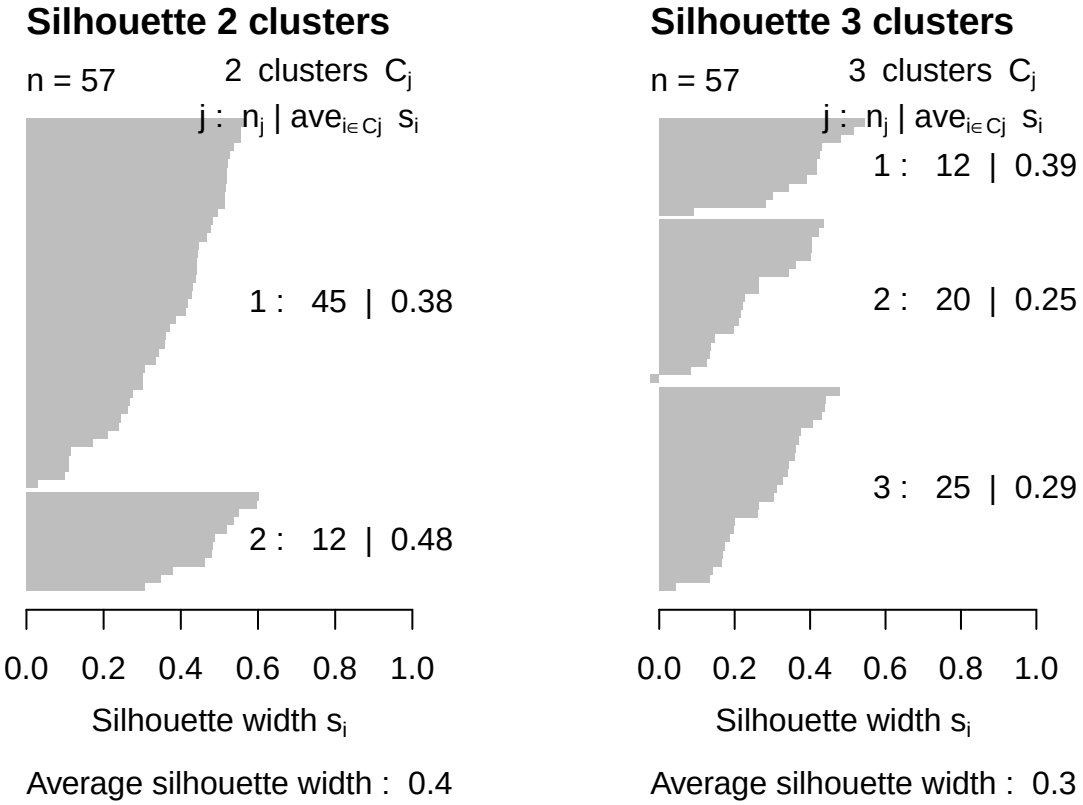


Non - Hierarchical Clustering

For the non-herarchical clustering we will use *kmeans* to obtain intra and intercluster data. As said before, we will handle the cases from 1 to 8 groups and we will compute the indicators mentioned before for the situations in which it makes sense. sult among different values of k.

As can be seen below, the hierarchical analysis is ratified by the non-hierarchical one. For all indicators,

2 and 3 groups have a higher score (in particular, the difference is specially big for $\Delta TESS$). Specifically, all the indicators for 2 groups are higher than for 3. Furthermore, the plot of the individual silhouettes for individual elements of each group clearly shows that the difference between 2 and 3 groups is likely to be a splitting of the biggest group into two smaller ones. This splitting reduces the average silhouette of the smallest group and the silhouettes of the two new groups are also smaller than the silhouette of the original one. Thus, all seems to indicate that increasing the number of clusters to 3 results in a poorer quality cluster structure. Because of that we will use 2 groups for the remaining of this study.




```
##   K TOTSS  WITHINSS  PseudoF   AvgSilh  DeltaTESS
## 1 1   392 392.00000       NA        NA         NA
## 2 2   392 235.39776 36.58966 0.3974478 0.39949551
## 3 3   392 168.38508 35.85592 0.2955183 0.28467849
## 4 4   392 142.55117 30.91472 0.2752161 0.15342161
## 5 5   392 122.64389 28.55119 0.2709973 0.13965007
## 6 6   392 109.48461 26.32020 0.2623816 0.10729669
## 7 7   392  96.46200 25.53147 0.2516608 0.11894469
## 8 8   392  87.14405 24.48809 0.2426378 0.09659709
```

Groups characterization

Representatives characterization

Below are shown the values of the representatives of each of the two groups. The analysis has been done with standardized values, so the representatives are shown accordingly. It can be seen that the second group representative has relatively low values in all the contamination indicators. The group 1 representative has overall positive values (which mean higher than the mean values in the original data), in particular for the organic contamination variables mentioned at the beginning and in **DQO_M** (which also determines organic contamination). Hence, group 2 represents low contamination and group 1 represents high organic

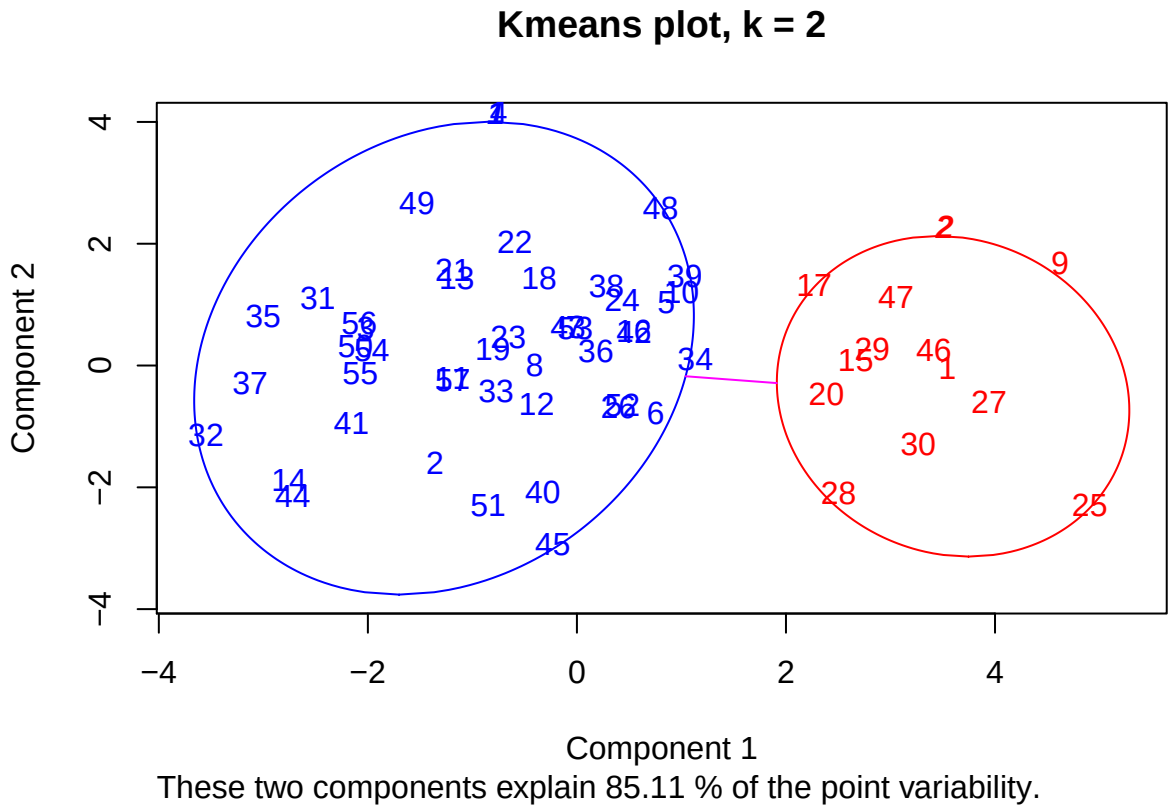
contamination. This will be expanded further when analizing the groups in a 2 dimentional space, but we don't feel confident enough yet to infer anything about the inorganic contamination.

```
result.km.2 <- kmeans(countries2, centers=2, nstart=1000)
result.km.2$centers
```

```
##   Colif_total Colif_fecal Estrep_fecal Cont_mineral Conductivitat
## 1   0.3907419   0.3954644    0.4227471    0.1087381    0.08126691
## 2  -1.4652823  -1.4829914   -1.5853018   -0.4077678   -0.30475090
##   Solids_susp      DQO_M
## 1    0.298371  0.3706778
## 2   -1.118891 -1.3900417
```

```
ecosystems$cluster.km.2 <- result.km.2$cluster
```

Representation in a low dimensional space We will use the *clusplot* command (which internally uses PCA) to draw the data, as it provides nice plots. We will also compute the regular PCA of the data so that we can interpret the principal components.



Above is the PCA plots for 2 groups, together with a cluster encircling purely for representation purposes. With only 2 dimensions the explained variability is higher than 80%, so we deem 2 dimentions as sufficient.

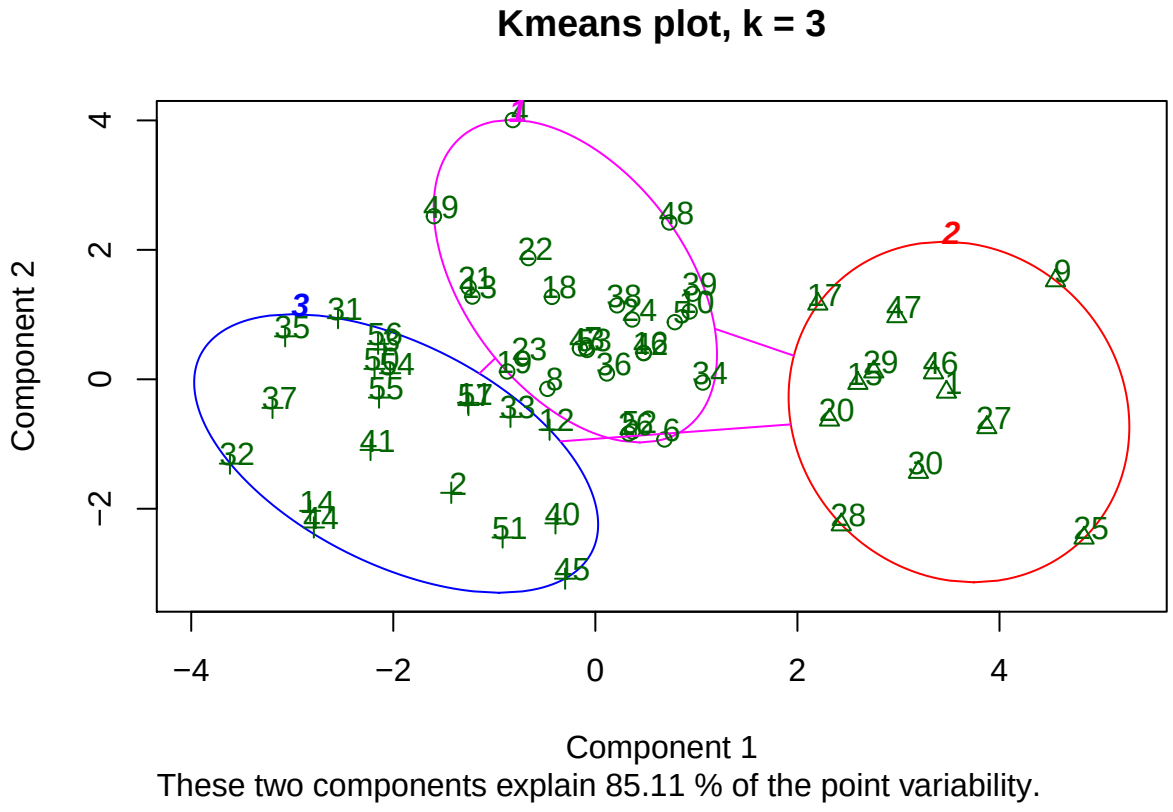
For interpretation of the principal components, we look at the loadings of the PCA. It seems that the first component (which accounts for around 60% of the explained variability) strongly depends on most of the organic contamination variables. Some inorganic contamination variables (like mineral contamination and solids in suspension) also affect it but to a lesser degree, and conductivity doesn't affect it in any appreciable way. For this reason we only deem to conclude that PC1 grows as organic contamination decreases. On the other hand the PC2 seems to be more related to the inorganic contamination as it grows importantly when the mineral contamination decreases or when the conductivity increases. As can be seen in the plot above, both groups are centered in the y axis, so we don't feel confident enough to infer anything about the inorganic contamination of each of the two groups.

```
##
## Loadings:
##               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## Colif_total   -0.451  0.251                -0.556         0.642
## Colif_fecal   -0.459  0.239                -0.393        -0.752
## Estrep_fecal  -0.429  0.137 -0.531 -0.306  0.563 -0.323
## Cont_mineral  -0.206 -0.547 -0.423  0.691
## Conductivitat         0.656  0.196  0.643  0.338
## Solids_susp   -0.384 -0.320  0.650         0.107 -0.559
## DQO_M         -0.459 -0.170  0.261 -0.110  0.304  0.755  0.133
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.143  0.143  0.143  0.143  0.143  0.143  0.143
## Cumulative Var  0.143  0.286  0.429  0.571  0.714  0.857  1.000
```

```
##    Comp.1
## 0.5902417
```

It is tempting to add a new group, given that it adds some inorganic separation in the contaminated group, as can be seen in the figure below. We think it is important to keep our analysis impartial and follow the indicators that we computed before viewing the 2D representation of the data, otherwise we could add bias to our study. Also, the two contaminated groups in the 3-cluster plot seem to be hard to distinguish, so this seems to agree with our Silhouettes analysis: adding a new group will only decrease the quality of our group structure. Thus we reject the 3-clusters hypothesis.

```
result.km.3 <- kmeans(countries2, centers=3, nstart=1000)
clusplot(countries2, result.km.3$cluster, main = "Kmeans plot, k = 3", color = TRUE, labels=2)
```



Group distinction

We'll use an inference analysis to check whether the two obtained groups can be really considered different groups.

We first check whether the groups follow normal multivariates using the Anderson-Darling test. As can be seen below, we can assume they follow it with a p-value of 0.05.

```
g1<-as.matrix(ecosystems[which(ecosystems$cluster.km.2==1),-8]) # Multivariate normality
AD.test(g1, qqplot = FALSE)
```

```
##             Anderson-Darling test for Multivariate Normality
##
##   data : g1
##
##   AD              : 0.4142843
##   p-value         : 0.6444356
##
##   Result  : Data are multivariate normal (sig.level = 0.05)
```

```
g2<-as.matrix(ecosystems[which(ecosystems$cluster.km.2==2),-8]) # Multivariate normality
AD.test(g2, qqplot = FALSE)
```

```
##             Anderson-Darling test for Multivariate Normality
##
##   data : g2
##
##   AD              : 0.7579276
##   p-value         : 0.819618
##
##   Result  : Data are multivariate normal (sig.level = 0.05)
```

As they follow multivariate normals, we can test for the homogeneity of their Variance/Covariance matrices and their mean vectors.

We first ran a chi-squared test using the boxM command and we got a p-value of around 0.7, so we cannot reject the null hypothesis of the variance-covariance matrix being homogenious.

On the other hand, with a negligible p-value, we can reject the null hypothesis of both groups having the same mean, as can be seen below. With this, we can conclude that both groups are really different and our analysis up to this point holds.

```
HotellingsT2(g1, g2, test = "chi")
```

```
##
##  Hotelling's two sample T2-test
##
## data:  g1 and g2
## T.2 = 175.3652, df = 7, p-value < 2.2e-16
## alternative hypothesis: true location difference is not equal to c(0,0,0,0,0,0,0)
```

Prediction

Lastly, we need want to predict the cluster that corresponds to a new sample. In order to do that we decide to use either a linear or a quadratic discriminant. The previous inference advocates in favour of the linear one, as the covariance matrices seem to be the same (or, at least, we cannot reject the possibility that they are the same).

Furthermore, we have $57 * 7 = 399$ distinct pieces of data (seven for each of the samples). If we were to use a linear discriminant, we would need to determine $2 * 7 + 7 * 8/2 = 42$ estimators (7 for each element of the mean vector of a multinormal distribution of 7 variables and $7 * 8/2$ for the variance/covariance matrix). This leaves us with a proportion of around 10 pieces of data per estimator, a low but assumible amount. If we were to use a quadratic one, we would need to determine $2 * 7 + 2 * 7 * 8/2 = 70$ estimators, having only 5.7 pieces of data per estimator, clearly insufficient. Thus, we decide to use a linear one (we also double checked the quadratic one out of curiosity and it gave a lower prediction rate when using leave-one-out crossvalidation).

Below we show the quality estimation of our linear discriminant. According to this analysis, the discriminant should have a low error rate of under 5%.

```
fit.l.cv <- lda(cluster.km.2 ~ ., data = ecosystems, na.action = "na.omit", CV = T) # Linear and crossv
sum(diag(prop.table(table(ecosystems$cluster.km.2, fit.l.cv$class))))
```

```
## [1] 0.9649123
```

Finally, we predict the cluster that corresponds to the new point (251,241,109,42,25,972,715). As can be seen below, the new point belongs to the cluster number 1 (which corresponds to organically contaminated samples) with almost total certainty. Looking at the values of the new point, this result is congruent with our analysis (as this point has higher-than-the-mean levels for most of the organic contamination variables).

```
element <- data.frame(Colif_total = 251, Colif_fecal = 241, Estrep_fecal = 109, Cont_mineral = 42, Condu
fit.l.ncv <- lda(cluster.km.2 ~ ., data = ecosystems, na.action = "na.omit", CV = F) # Should use the n
predict(fit.l.ncv, element)
```

```
## $class
## [1] 1
## Levels: 1 2
##
## $posterior
##           1            2
## 1 0.9999937 6.268562e-06
##
## $x
##         LD1
## 1 -1.231824
```