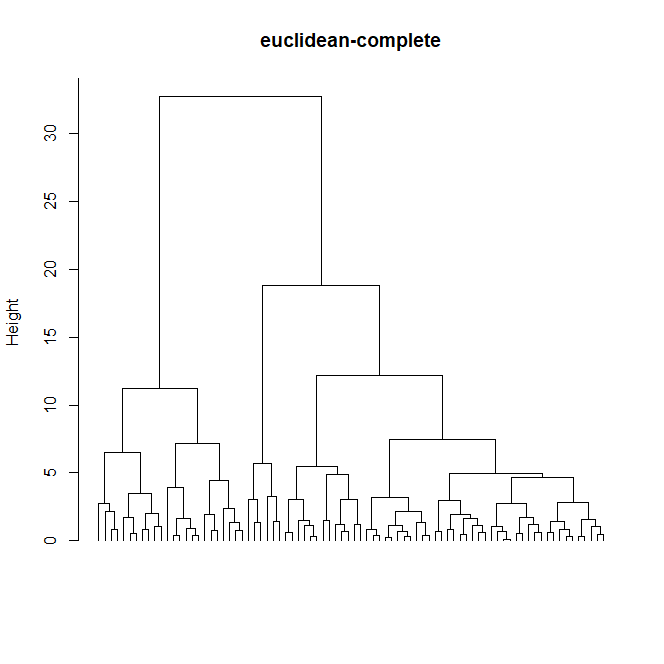
Ex1)

1. Dendogram



From the dendogram we can choose K=2 or k=3 but choosing k=3 we would obtain a group (the one in the middle) a lot smaller than the others and moreover the choice k=2 seems to be the more robust so we choose k=2 clusters. This choice is also coherent with the scatterplot of the data which seems to suggest 2 groups.

1. MANOVA:

Verify normality inside each of the 2 clusters -> the pvalues are 0.5240 0.2448

* Accept normality assumption

Verify homogeneity:

Plotting the 2 covariances structure we are not sure that this assumption is met, anyway we proceed keeping in mind that the homogeneity assumption may be not satisfied.

Model: X.ij = mu + tau.i + eps.ij; eps.ij~N\_2(0,Sigma^2), X.ij, mu, tau.i in R^2

Fitting the model we obtain pvalue = 0 which indicates that the grouping is statistically effecting the mean features analized.

Parameters:

* Mu :

height width

29.29782 11.05665

- tau1

height width

5.269634 1.604267

-tau2

height width

-12.734949 -3.876977

* Sigma^2 = 15.14519

1. Bonf intervals for the difference of the means

$group1\_group2

inf12 sup12

height 16.368443 19.640724

width 4.555844 6.406644

We can see that neither the interval for height neither the one for width contain the 0 which means that there is statistical difference between the mean in the 2 clusters , moreover the intervals are both positive so the mean of both the features in the cluster1 is larger than the one in cluster2.

Ex2)

1. Non fatto
2. Benj -> vedi codice
3. Bonf -> vedi codice

Ex3)

1. # Model:

price = beta0 + beta1\*dummy\_w + beta2\*length +

+ beta3\*power + beta4\*draught + beta5\*crew + beta6\*year + eps

with eps~ N(0, sigma^2)

so the dummy is modifying only the intercept : dummy\_w = 1 if the material = wood.

Estimates of the parameters:

(Intercept) length power draught crew year

wood -10636.38 316.6172 0.1041251 42.97902 652.5758 4.787693

fiberglass -11073.77 316.6172 0.1041251 42.97902 652.5758 4.787693

so only the intercept is different between the 2 groups.

The name of the columns indicates the regressor to which that coeff is associated.

Estimates of sigma^2 = 21007.11

Verify assumptions: plotting the residuals they seem to be centered in zero and homoschedastic -> 0

And performing the Shapiro test we can assume that they are normal since we have a pvalue of almost 21%.

1. Perfotming a linearHypothesis test to check if we can put at 0 at the same time the coeff of the 3 variables related to the dimension , so:

H0: (beta2, beta3, beta4) == (0,0,0) vs H1: (beta2, beta3, beta4) != (0,0,0)

We obtain as pavlue the numerical 0 -> reject H0 -> it is necessary to include into the models these 3 variables

1. Performing the same test for these others parameters:

H0: (beta1, beta5) == (0,0) vs H1: (beta1, beta5)!= (0,0)

Again we obtain as pavlue the numerical 0 -> reject H0 -> it is necessary to include into the models the variables related to the accessories

1. From the output of the model the variable draught does not seem to be significant so we can remove it, then also year is not much significant we can try to remove it. Removing also year we obtain that all the regressors are now significant and the R^2adj is practically the same as the model which keep year but discard draught so we prefer the simpler model without year.

Final model:

price = beta0 + beta1\*dummy\_w + beta2\*length +

+ beta3\*power + beta4\*crew + eps

with eps~ N(0, sigma^2)

Parameters estimates:

(Intercept) length power crew

wood -1019.914 321.6404 0.09561734 648.5048

fiberglass -1450.105 321.6404 0.09561734 648.5048

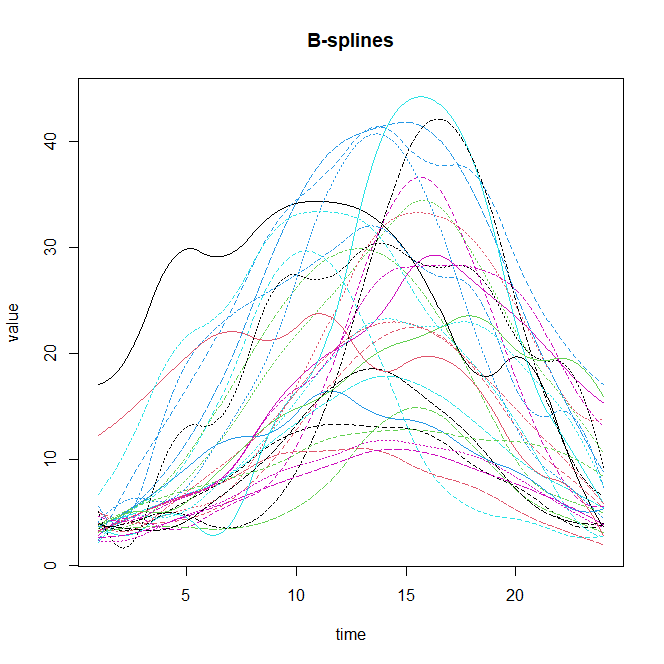
Estimates for sigma^2 = 21726.95

1. Pointwise estimate = 2517.115

Prediction Interval = [2204.685, 2829.544]

Ex4)

1. Smoothed data



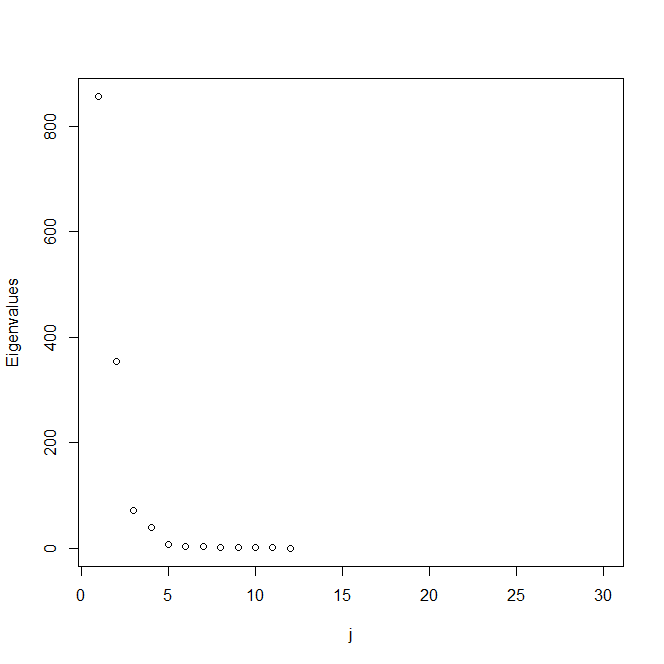
Coeff of the first 3 basis for Day1:

17.045819, 17.634977, 31.079588

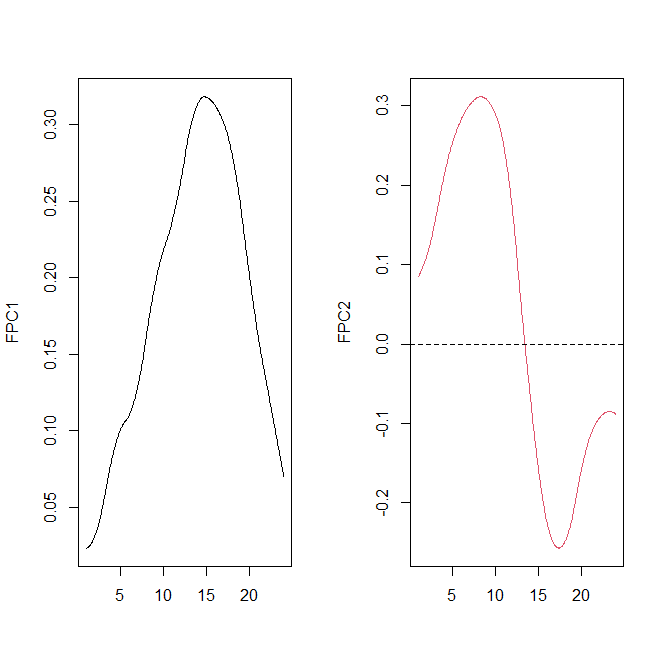
1. Variance explained along the first 3 PC

0.639975910 , 0.265055813 , 0.053557439

Scree-plot:



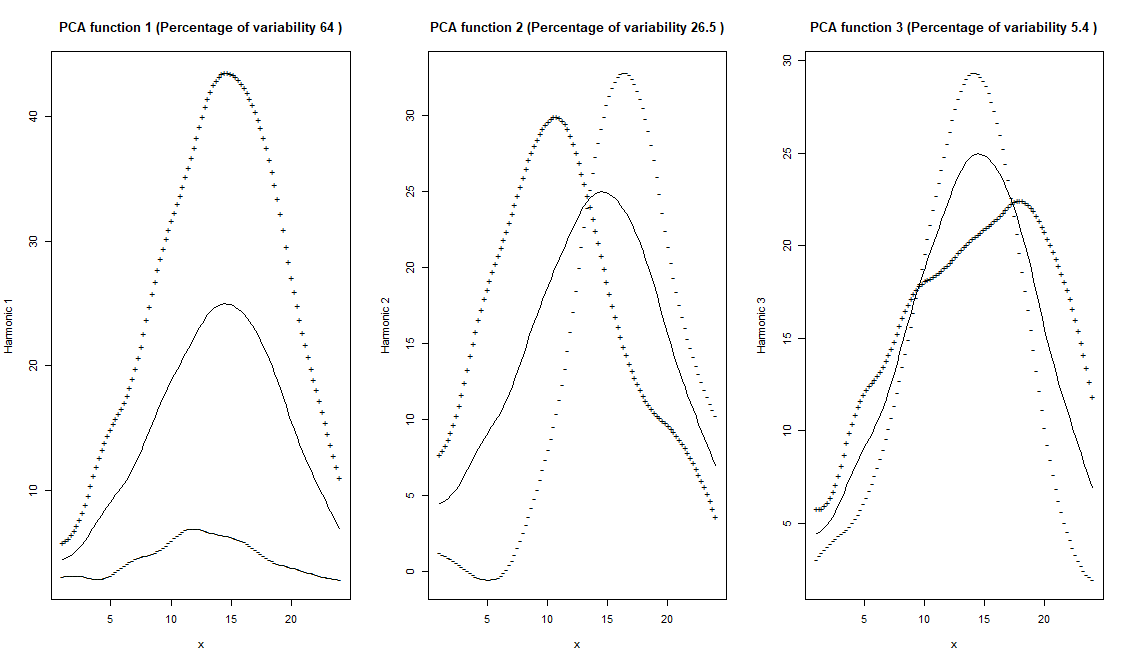
FPC: sono gli eigenfuncions -> sì



Me ne fa fare solo due -> ? aggiustato

1. Looking at the cumulative proportion of explained variability we see that with 2 FPC we explain around the 90% of the variability and with the first 3 we explain around 95% so we can keep 2 or 3 PC, we try to keep 3 and look at the interpretation.

Plot of the first 3 FPC



Interpretation:

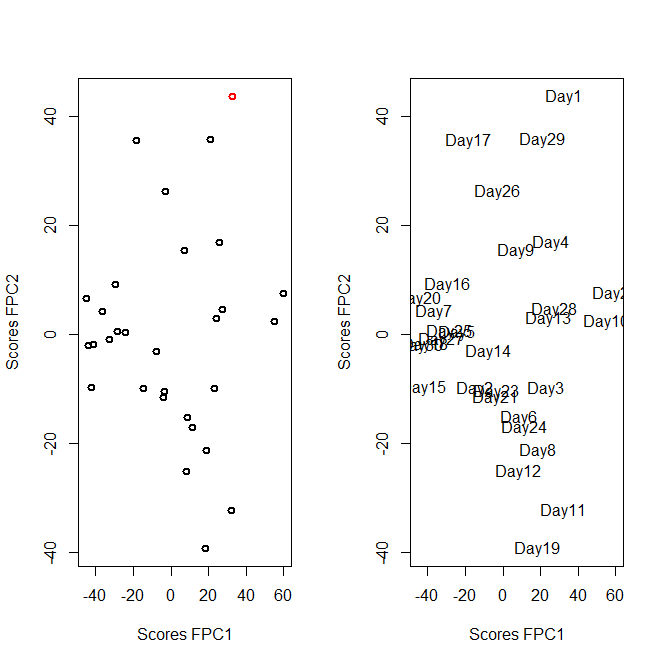
For the first PC we see a large difference in amplitude between those functions which have positive scores and the ones with negative ones , the one with positive scores are characterized by a larger intensity of the wind. -> days with high PC1 are characterized by stronger wind

In the second PC we can observe a delay of those functions with positive scores wrt those with negative ones. -> days with low Pc2 have an andament of the wind which is in delay

Finally in the third PC we observe a contrast between the central part of the day between 9 and 16 and the rest of day, in particular functions with negative scores are characterized by less wind in the tail part of the day but greater intensity in the middle part. -> days with low Pc3 have less wind in the extreme parts of the day but reach higher values of wind in the central part of the day

d)

Day 1



Day 1 is characterized by positive and large values of both scores so the wind in Day1 has higher amplitudine wrt the mean, it is early in the day and

ERR: we observe a contrast between the middle part of the day and the rest

* Avevo messo anche questa ma non ci va perché quella si riferisce a pc3