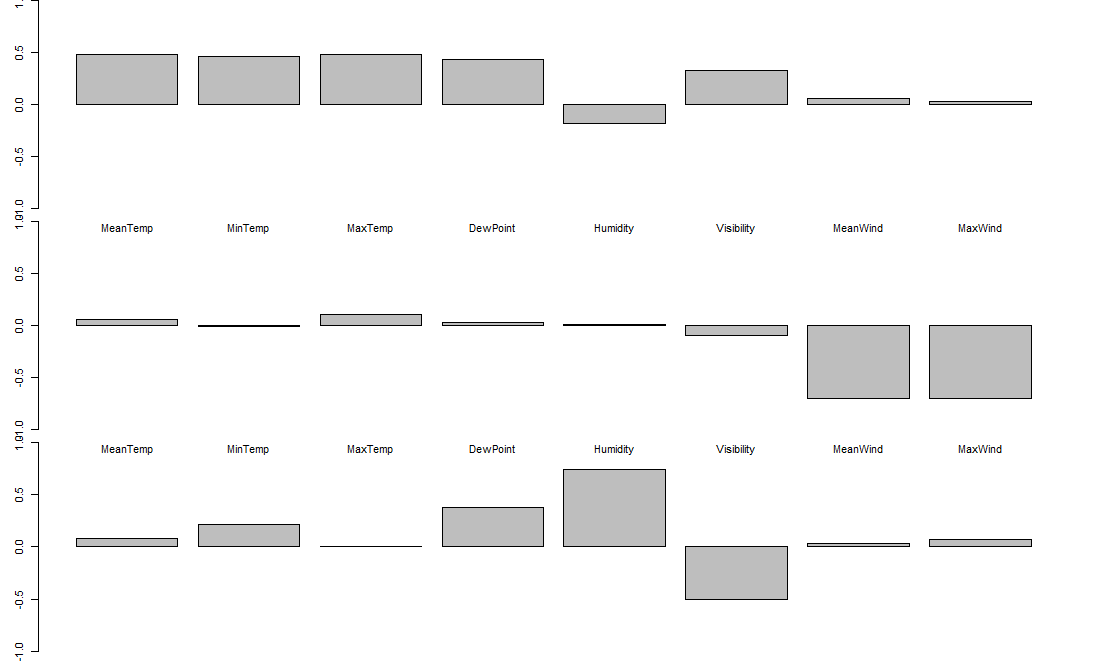
Ex1)

1. We proceed by scaling the data since from the boxplot we can observe that the variances of visibility and meanwind are significantly smaller than the one of humidity so we scale in order to have more or less the same variance for all the variables.

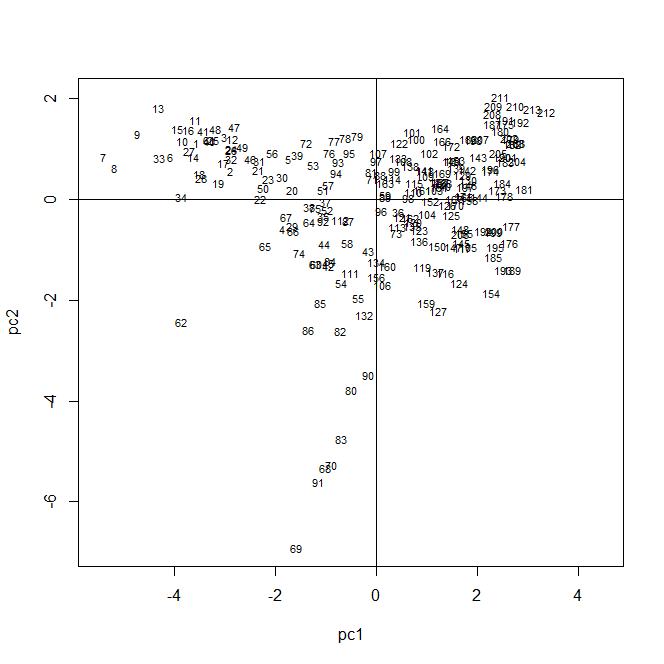
Plot loadings



Interpretation:

* PC1 is a weighted sum of all the variables except for humidity, meanwind and maxwind which have loading almost 0 (this can be seen more clearly plotting only the loadings above a certain threshold)
* PC2: is the opposite of the mean of the variables related to wind namely meanwind and maxwind
* PC3 shows a contrast between principally dewpoint and humidity against the visibility

1. Scatterplot



In the quadrant in the topright: Pc1 and Pc2 are positive so there are those data which have positive weighted sum of the variables and negative mean of the wind variables

In the quadrant on the bottomright: pc1 >0 and pc2<0 so again data which have positive weighted sum of the variables but now a positive mean of the wind variables

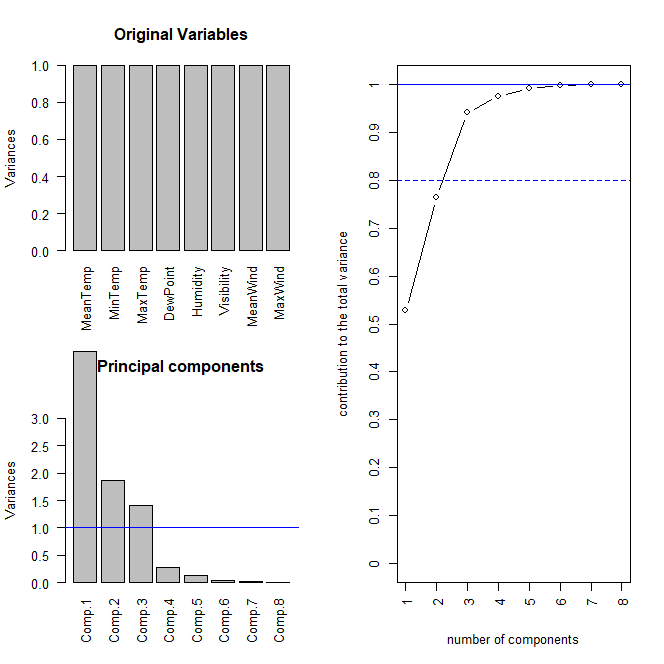
In the topleft: pc1<0 and pc2>0 so those data have a negative mean of the variables and negative mean of the wind variables.

In the bottomleft: pc1<0 and pc2<0 so data with negative weighted mean of the variables and positive mean of the wind variables.

Plotting also the number of the days we see that more or less data in the first 100 days are characterized by a pc1 score <0 and data next the first 100 by pc1 score >0 so the weighted mean goes from negative to positive going on with the days.

Increases going on with the days

1. Scree-plot



To perform dimensionality reduction we would keep 3 PC to reach a good amount of explained variance indeed we see that with this choice we reach 0.9407453 of explained variability and adding the 4 for instance we would not add so much of explained variability.

Indeed the variance explained by each of the 3 Pc is:

PC1: 0.5291589777

PC2: 0.2350543981

PC3: 0.1765319087

1. Projection

Ex2)

1. The 2 populations candle and sunshine are independent so we perform a test for the difference of the means in this frame.

The hypothesis to be satisfied are:

# - independent population

# - gaussian distribution

# - same covariance structure

Check the gaussian distribution in each population via a mcshapiro test, the pvalues are quite high so we assume gaussianity and then we compare the covariances matrix of both the population and we deduce that we can assume same cov structure

Test H0: mu1-mu2 == c(0,0) vs H1: mu1-mu2 != c(0,0)

1. Pvalue = 3.09412e-08

* At level 95% we reject H0 so we state that the means are different

1. Bonf intervals for the components of the mean

LM1 0.5302967 1.7092 2.8881033

LM2 -2.9142181 -1.6934 -0.4725819

We see that either the interval for the measurements with the first meter and either the one for those with the second meter do not contain 0 which means that the difference is significantly different from 0. In particular for the case LM1 the mean of the measurements of the population candle is higher than the one of sunshine instead for the measurements for LM2 it is the opposite.

1. I built 2 new populations: the first is given by the diff of LM1-LM2 for the candle bulbs and the second population is given by LM1-LM2 for the sunshine bulbs.

Performing the test to check

# H0 : mu1 -mu2 <= 0

# H1 : mu1 -mu2 > 0

We obtain that at level 95% we reject H0 -> we can state that the diff in brightness of the candle bulbs is higher than the one of the sunshine bulbs

* Come traduce il fatto che ora è un test unilaterale?

Ex3)

1. Formulate the model defining

Y = volume

Z1 = time

Z2 = time^2

Dummy\_by = 1 if yeast = by and 0 otherwise

Model

# y = beta\_0 + beta\_1 \*z1 + beta\_2 \*z2 + beta\_3 \* z1:dummy\_no + beta\_4 \* z2:dummy\_no + Eps

# with Eps ~ N(0, sigma^2)

Parameters:

(Intercept) z1 z2 z1:dummy\_by z2:dummy\_by

0.9839522351 0.0670422710 -0.0005003801 -0.0529347109 0.0122343277

So dividing the 2 groups

(Intercept) z1 z2

by 0.9839522 0.01410756 0.0117339476

sd 0.9839522 0.06704227 -0.0005003801

the first column is the intercept, the second the parameter related to the variable z1 and the third parameter related to the variable z2.

Estimates of sigma^2 = 0.002207288

Verify the assumprion:

* Residuals centered in 0 and homoschedastic -> from the plot of the residuals this is ok
* Normality of the residuals : Shapiro test with pvalue 0.9238 -> ok

1. – performing a linear hypothesis test we check if we can put at 0 borh the interaction coeff so beta3 and beta4

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

The pvalue is almost 0 so we reject -> there is significant dependence on the resing from the type of yeast

* Since the coeff of z2 (not accounting) for the interaction is almost 0 the degree of polynomial for by is higher than the one for sd if the regressor z2:dummy\_by is significant -> from the output of the fit er read the pvalue of the test

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

Which is the numerical 0 so it is important -> reject H0 and there is evidence to state that the the degree is higher for the by.

1. Z2 does not seem to be significant -> remove it

# y = beta\_0 + beta\_1 \*z1 + beta\_2 \* z1:dummy\_no + beta\_3 \* z2:dummy\_no + Eps

# with Eps ~ N(0, sigma^2)

Parameters:

(Intercept) z1 z1:dummy\_by dummy\_by:z2

0.99239281 0.06208154 -0.05098783 0.01196397

So diving by groups

by 0.9923928 0.01109371 0.01196397

sd 0.9923928 0.06208154 0.00000000

the first column is the intercept, the second the parameter related to the variable z1 and the third parameter related to the variable z2.

Sigma^2 = 0.002187187

1. The pointwise estimates for the volume considering time = 2 are:

* Using yeast by = 1.062436
* Using yeast sd = 1.116556

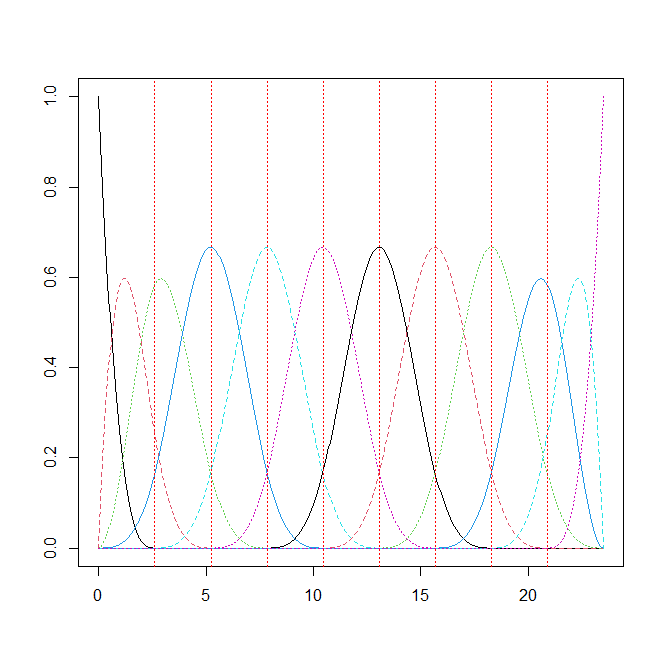
so I would suggest to use the yeast sd in order to obtain a larger volume, in this case the

PI = 1.021119; 1.211993

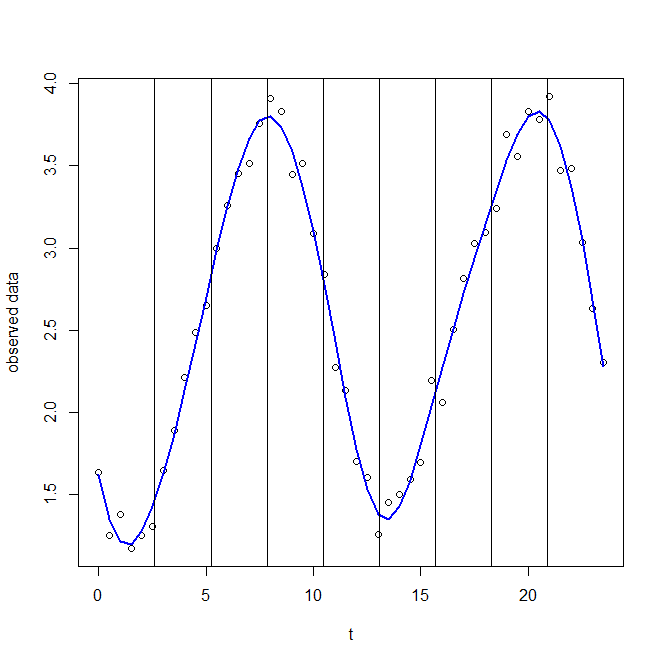
Ex4)

1. Perfoming CV we obtain that nbasis\_opt = 12

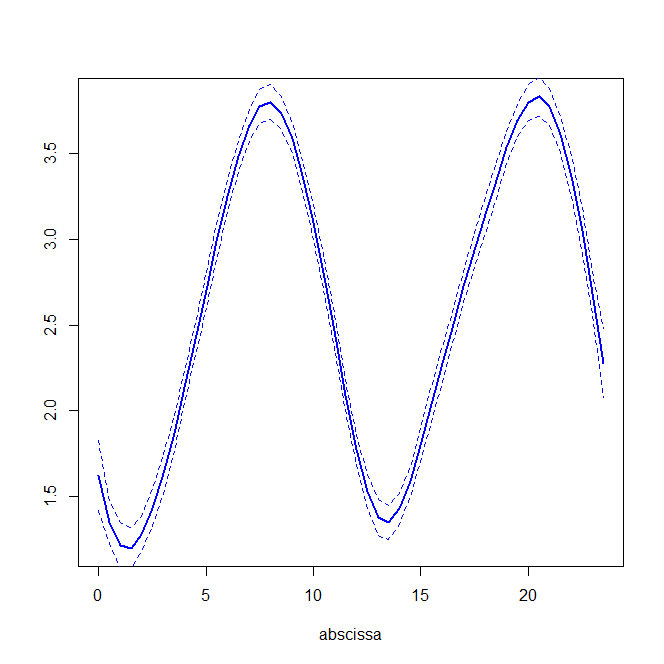
Basis:



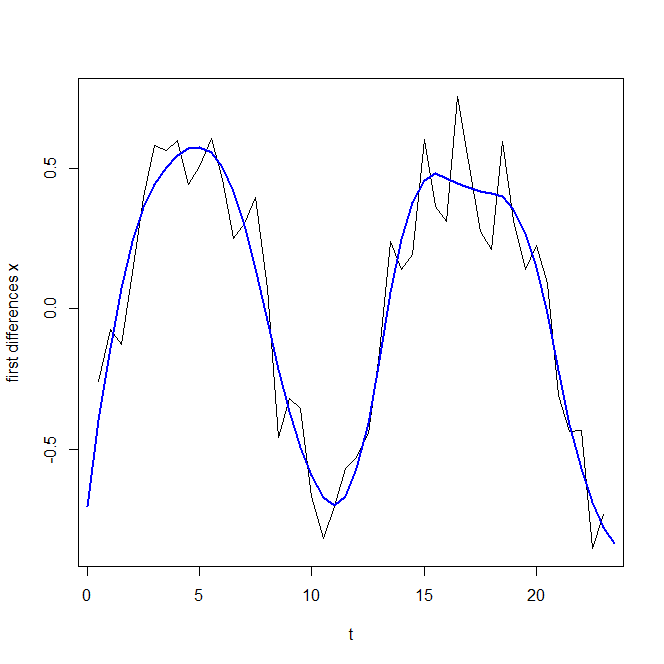
Smoothed data:



1. CI:



1. First derivative:



The black line is the first derivative estimated by the data and the blue line is from the smoothed data, we can appreciate that the smoothed one is capturing the behaviour of the one of the data smoothing the noise

1. The data seems to be periodic so maybe it could be a good idea to use the Fourier basis in order to better capture this characteristic.