Ex 1:

a) Verify assumptions:
   - Normality in each subgroups (6 subgroups given by the 2 groupings) , we perform a multivariate Shapiro which allow us to assume gaussianity since we don't reject the normality assumption
     P = 0.2369077 0.4612391 0.9702243 0.6525198 0.5279147 0.6612603
   - Homogeneity in the 6 subgroups:
     Perorm a barlett test :
     # H0: sigma.1 = .. = sigma.g
     # H1: there exist i,j s.t. sigma.i!=sigma.j

     P value = 98% -> accept homogeneity

     ➔ Hypothesis verified

   Build the anova two ways model

b) Df Sum Sq Mean Sq F value Pr(>F)
   group1 1 32.79 32.79 62.176 7.12e-13 ***
   group2 2 0.16 0.08 0.149 0.862
   group1:group2 2 0.21 0.11 0.201 0.818

   by this output we can see:
   - ### H0: tau.1 = tau.2 = 0 vs H1: (H0)^c
     ### i.e.,
     ### H0: The effect Fact1 doesn't significantly influence the alchol
     ### H1: The effect Fact1 significantly influences the alchol
     Reject H0 ->
     group1 (color of the wine) is significant so there seems to be a difference in the alcohol between wines with different colors
   - H0: beta.1 = beta.2 = 0 vs H1: (H0)^c
     ### i.e.,
     ### H0: The effect Fact2 doesn't significantly influence the alchol
     ### H1: The effect Fact2 significantly influences the alchol
     Accept H0 ->
     group 2 (region) does not ignificantly influence the alchol
   - the interaction between the groupings do not seem to be significant

   We can go on removing the interaction and so building an additive model:
   - group1 still significant
   - group 2 not significant
     ➔ remove group2

   model : x.jk = mu + tau.i + eps.jk; eps.jk~N(0,sigma^2)

   Estimate mu, tau.i:

   mu = 7.981784

   tau1 = 0.4675426

   tau2= -0.4675426

estimate of thesigma^2: 0.5156073

c) BF for the means in the 2 groups identified by colors (grouping 1)
red 8.201944 ; 8.696709
white 7.266859 ; 7.76162
BF for the variance = 0.3748061 0.7442128

Commento?

Ex2)

a) I want to use LDA or QDA to build the classifier
Check the assumptions fir LDA:

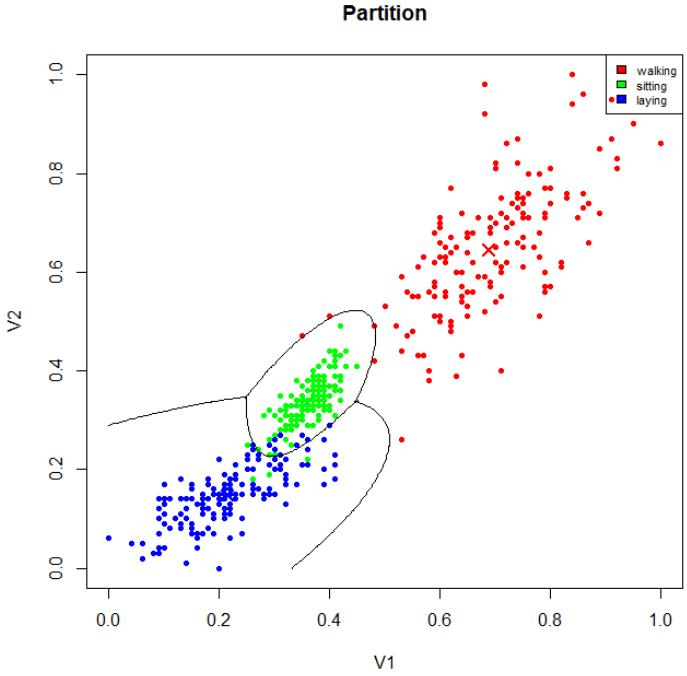
- Gaussianity in each group:
Pvalue in the 3 groups = 0.424 0.416 0.256
➔ Accept normality
- Homogeneity:
The homogeneity does not seem to be met

➔ We use QDA which only need the assumption of normality

Prior probabilities of groups:
walking sitting laying
0.125 0.500 0.375

Group means:

	accel	gyro
walking	0.6877333	0.6447333
sitting	0.3588000	0.3376667
laying	0.2147333	0.1474000



b) APER

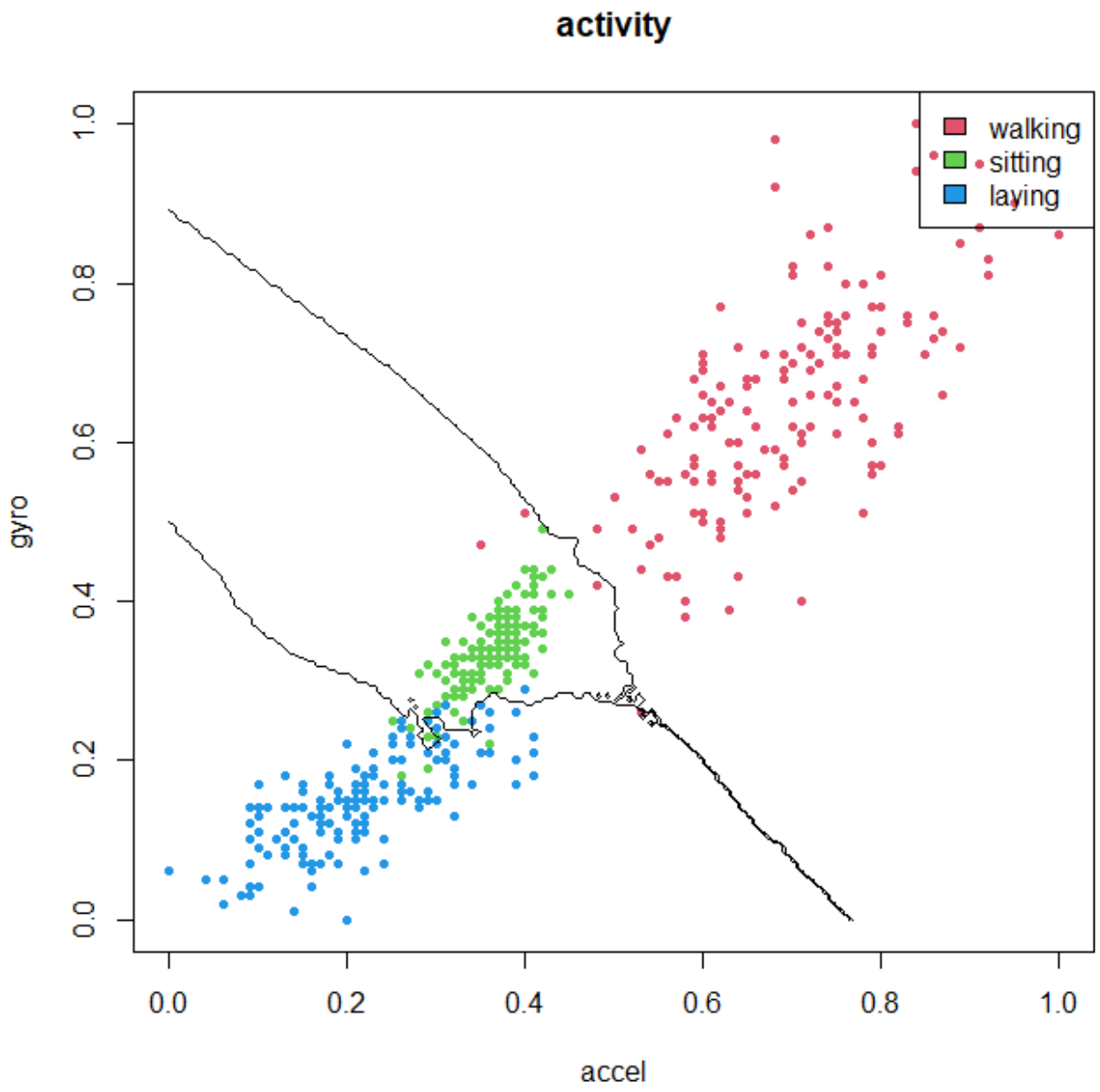
	class.assigned		
class.true	walking	sitting	laying
walking	147	3	0
sitting	0	145	5
laying	0	10	140

APER = 3/150 * 0.125 + 5/150*0.5 + 10/150* 0.375 = 0.04416667

Small -> good

c) The predicted activity is Sitting with a posterior probability of 0.5372111

d) KNN



Come calcolo l'error rate? Ok

Ex3:

a)

Construct a linear model with z1= mean_temp and z2=mean_wind numerical regressors and the categorical regressor Holiday/not holiday (dummy variable = 1 if Holiday)

y = B0[g] + B1[g]*z1 + B2[g]*z2 + eps

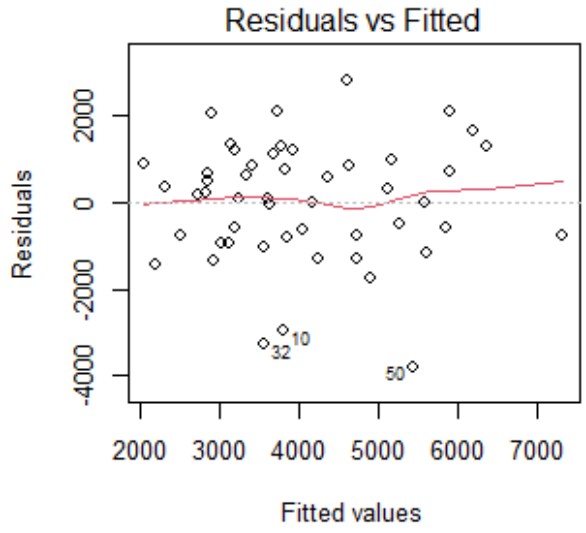
Estimated parameters:

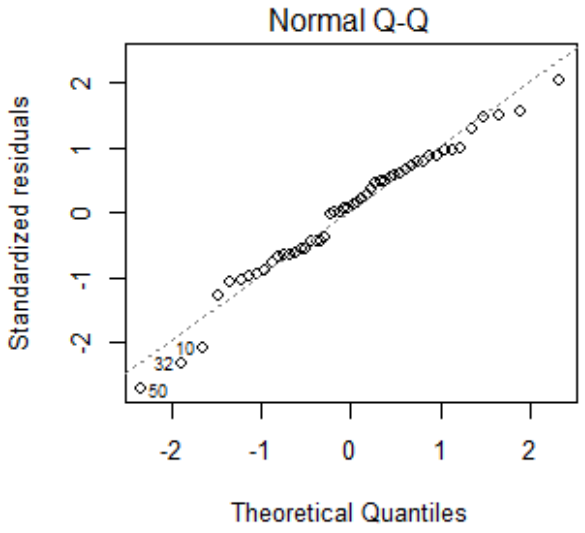
	B0	B1	B2
Holiday	4084.388	118.60701	-224.9999
No_holiday	3286.108	87.42783	-441.7284

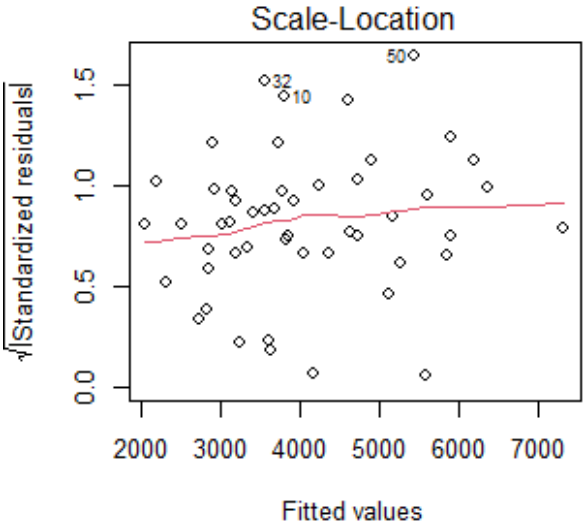
Sigma^2 =

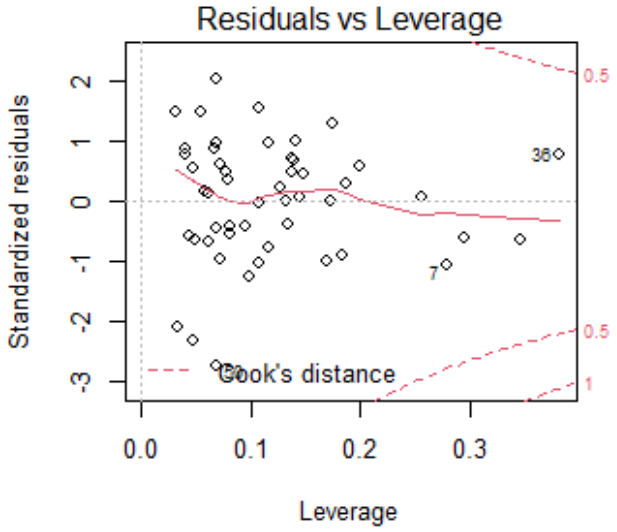
b)

verify assumption:









The residuals seem to have mean zero and seem to be homogeneus ->
We can say that maybe the data 10,32,50 are outliers and we should look at them to understand if we can remove them
The qqplot is quite good indeed also the Shapiro test allow us to assume gaussianity

➔ Assumptions verified

Test about the weather: I perform a linear hypothesis test to check if we can put at 0 all the coefficients related to z1 and z2 -> the pvalue= 0.0002863 -> at 5% we reject H0 (ie all the coeff are 0) -> there is statistical evidence of a dependence of the mean number of bikes rented on weather information

Test about the holiday: perform again a linear hypothesis test to check if we can put at 0 all the coeff related to the dummy variable : pvalue = 0.009922 -> at 5% we reject -> there is statistical evidence of a dependence of the mean number of bikes rented on holiday information.

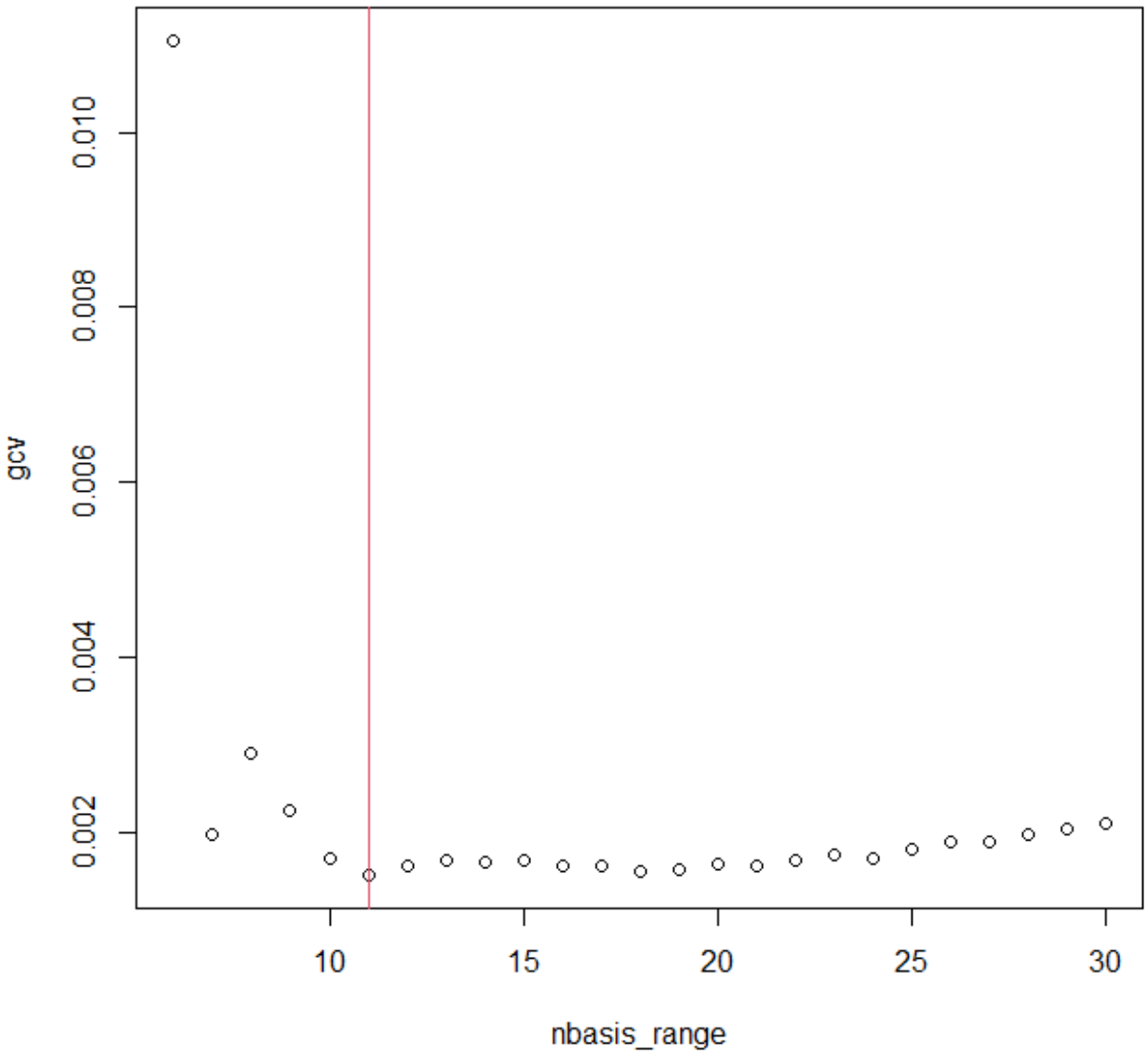
c) From the output of the model we see that a lor of regressors are not significant and this may impact the goodness of the model indeed we see that the R^2 adj is quite low-> reduce the model
The final model is:
fit <- lm(y ~ z1 + dummy , data = dataset)

all the regressors are now significant.
Estimates:

	B0	B1
Holiday	3880.735	97.49175
No_holiday	2456.943	97.49175

d) PI = [1121.848 ,7029.589]
Point estimate = 4075.718

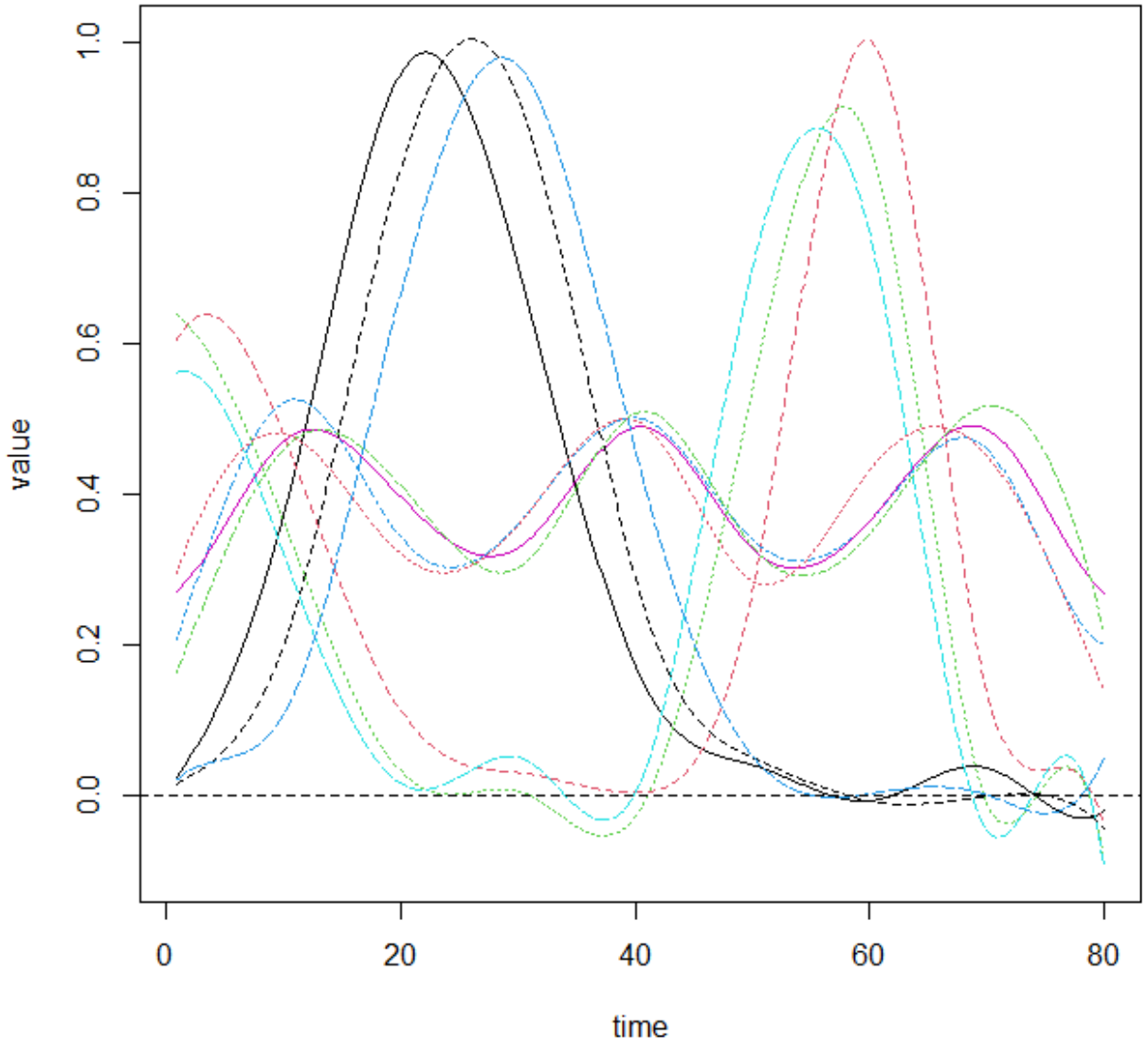
Ex4)



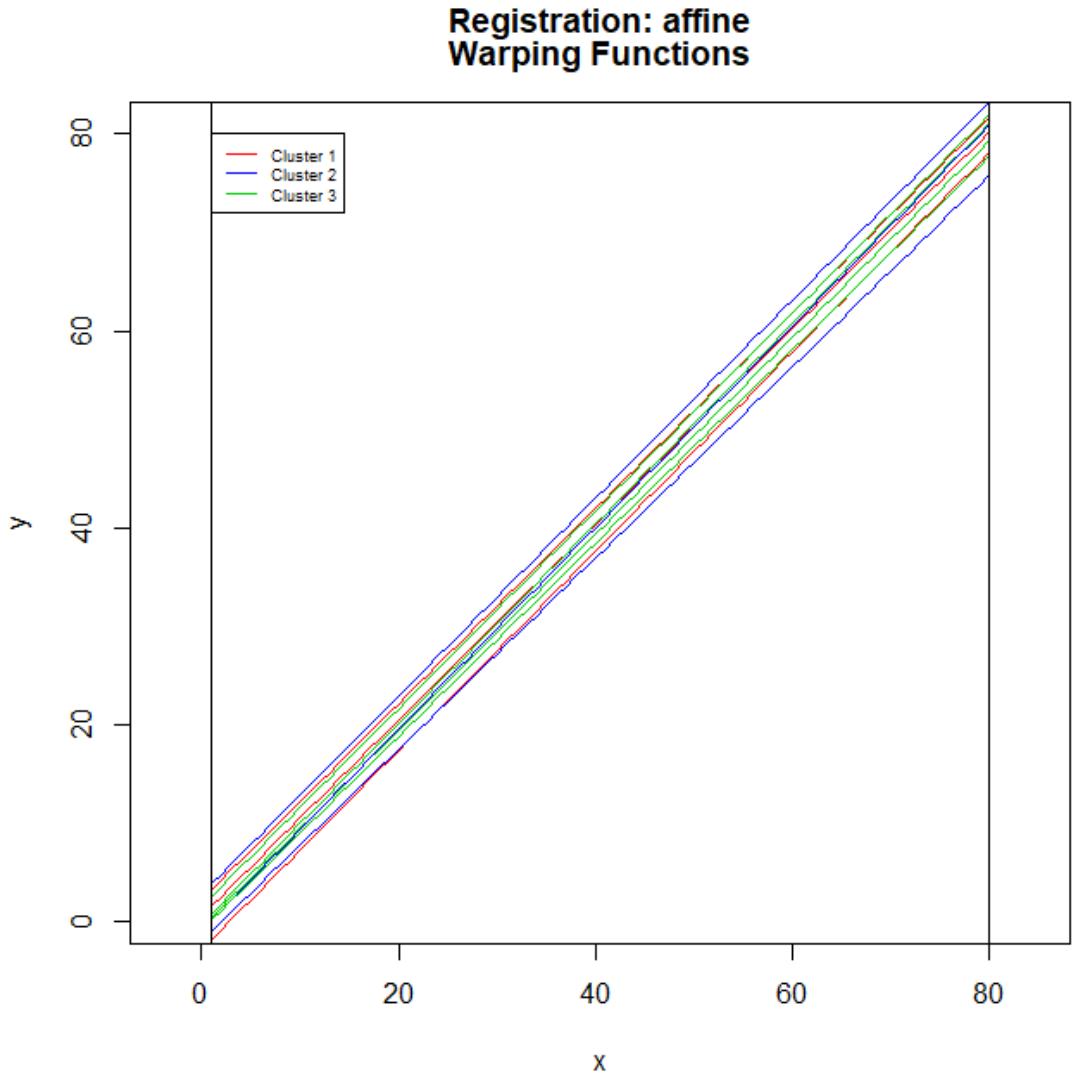
As we can see from the plot the optimale number of basis is 11.

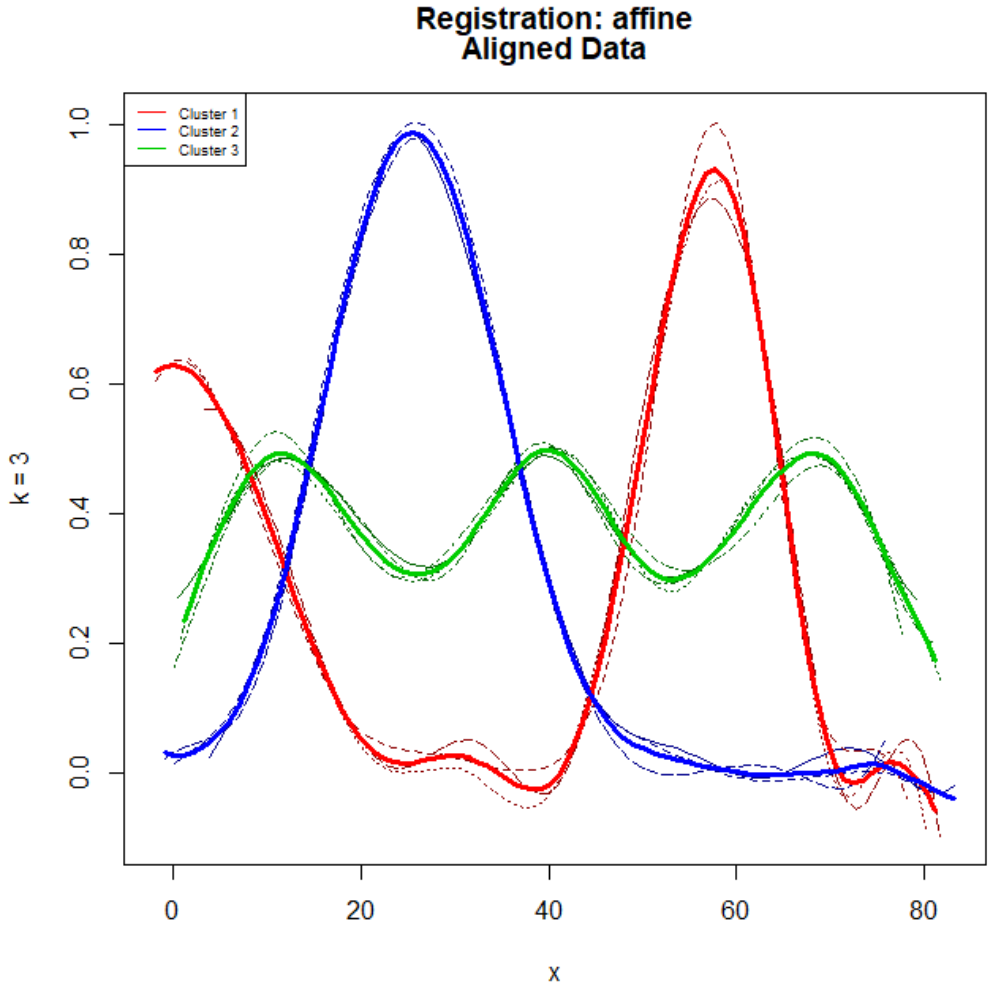
Estimated coeff = 0.02401024 0.10070402 0.33763038 -> check -> sì

b) smoothed data



d) K-mean allignment





The result of the k-mean alignment is good for the functions but we don’t see a clustering in the warping functions

Devo farlo su quelle smooth? Come?