Ex1)

1. We are in the frame of paired data since we have the same products in 2 different stores.

Formulate the test as:

# Ho: mu1 = mu2 vs H1: mu1 != mu2

# with the difference it becomes

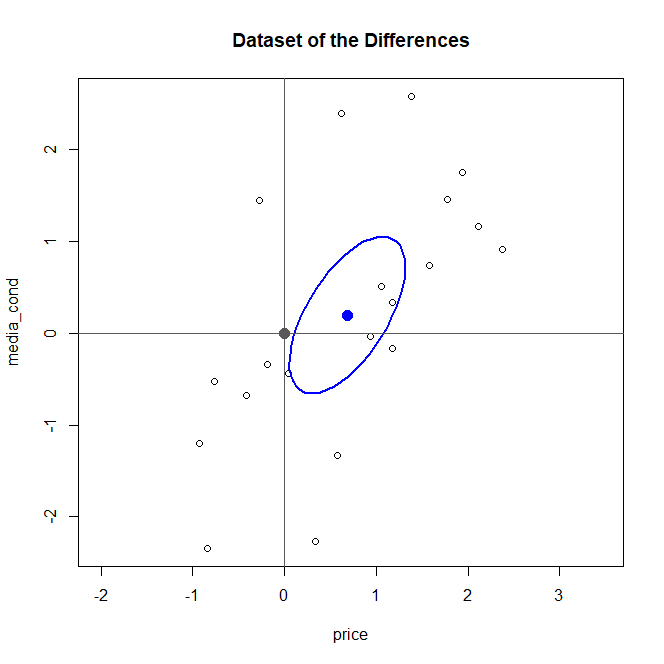
# H0: mean\_diff = (0,0) vs H1: mean\_diff != (0,0)

Where mean\_diff is in R^2 and contains the mean of the difference between the prices in the 2 stores and between the conditions.

At level 95% we reject H0 so we have evidence to state that the mean prices and conditions in the 2 stores differ.

The pvalue of the test is 0.01625523.

1. To perform the test the population given by the difference must be a bivariate gaussian so we test the normality via a mcshapiro test which returns a pvalue of 0.744 and so we can assume gaussianity and the hyp is met.
2. Plot of the CR for the difference of price and condition at level 95%



1. Bonf CI at global level 95%:

inf center sup

mean.diff\_price 0.0480422 0.6830 1.317958

mean.diff\_cond -0.6597456 0.2005 1.060746

var.diff\_price 0.5322847 1.059612 2.841861

var.diff\_cond 0.9770104 1.944921 5.216247

we observe that the interval for the mean of the difference of the prices does not contain the 0 which validate the result of the test performed at point a). Indeed we know that we reject H0 if at least in one direction we have evidence that the difference is different from 0. Through the bonf intervals we show that the mean of the difference of the prices is significantly different from 0.

Ex2)

1. We want to build a LDA or a QDA classifier based on the verification of the assumptions.

For both classifier we need the data in each group to be a bivariate gaussian so we perform two mcshapiro test which return as pvalue 0.1344 0.9852 -> we can accept the hypothesis of gaussianity.

Then we check if the 2 groups have same covariance structure but this assumption (needed for LDA) does not seem to be met so we reject this hypothesis and we choose to build a QDA classifier which does not rely on this hypothesis.

Building the QDA classifier we obtain these means and covariances:

* Means

price average.length

Germany 59.82982 5.891951

US 29.74888 4.377556

* Covariances

Germany:

price average.length

price 108.936681 -7.621866

average.length -7.621866 8.505943

US

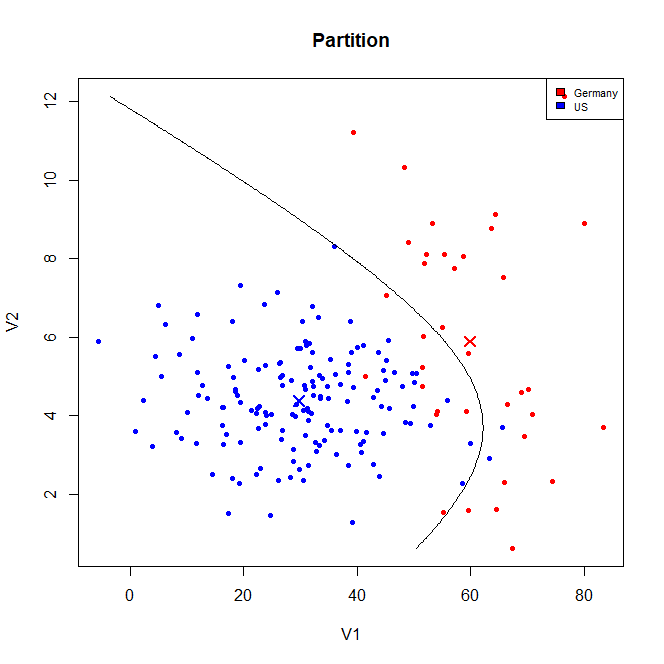
price average.length

price 173.53448 -1.524040

average.length -1.52404 1.543192

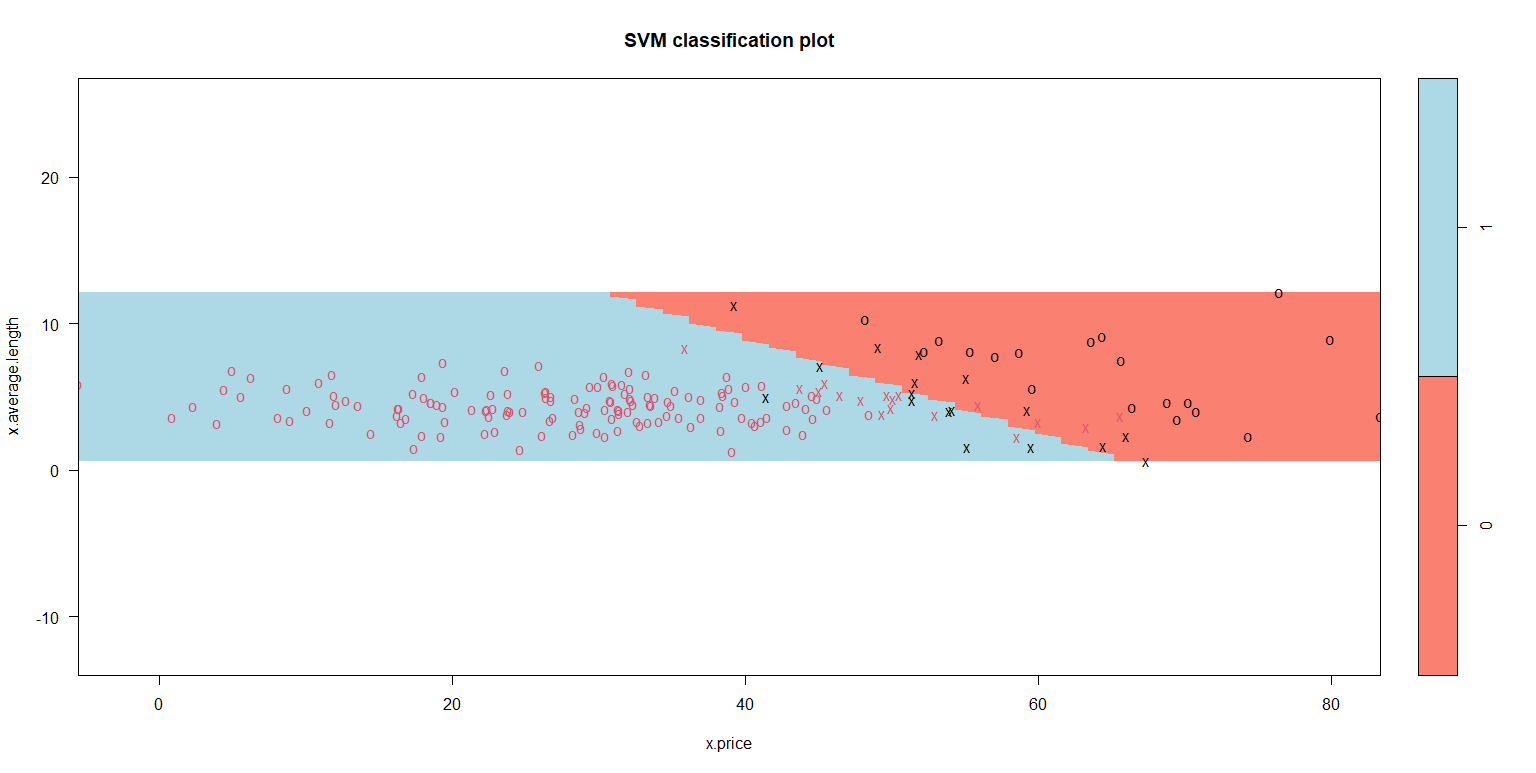
* Calcolate prima!

Plot of the classification region:



1. APER = 4/152\*0.9 + 9/36\*0.1= 0.04868421
2. 0.9 since we use the prior
3. I would classify it as US with a posterior probability of 0.9231916
4. The chosen cost is 0.01.

Classification region:



Where 1 corresponds to US and 0 to Germany.

The observation would be classified as US.

Ex3)

1. Estimates for beta:

Beta0 beta1 beta2 beta3

9.18676786 0.09930218 0.07258300 -0.00888967

Estimate for sigma = 0.935

1. Assumption: Eps ~ N(0, sigma^2) -> normality needed in order to make inference on the model

In order to check the assumption we do:

* Plot the residuals against the fitted value to see if we can assume that they are centered in 0 and homogeneous -> in this case tha assumption seems to be met.
* Perform a Shapiro test for the residuals to check the gaussianity assumption -> the pvalue is 0.6654 so we can assume it

1. We perform a linear hypothesis test:

H0: (beta1, beta2) == (0, 0) vs H1: (beta1, beta2) != (0, 0)

But the pvalue is 2.112e-06 so at 5% we reject tìH0 and so we state that loudness and energy cannot be removed both from the model

1. We remove only energy since the test for the significance of beta2 have a pvalue of 0.07 so at level 5% we can accept H0.

In this way we obtain a model where all the regressors are significant

Danceability = beta0 + beta1\*loudness + beta2\*tempo + eps

Parameters:

* Beta

(Intercept) loudness tempo

9.182772292 0.170266165 -0.008962479

Oppure tieni la scrittura con i parametri beta numerati come prima e dì che beta2=0 perché lo abbiamo rimosso

* Sigma = 0.9376

1. Fit the lmm model including the random inetercpet depending on genre:

Dance\_ability\_it = beta\_0 + beta\_1\*loudness\_it + beta\_2\*tempo\_it + b\_i + eps\_it

eps\_ij ~ N(0, sigma2\_eps)

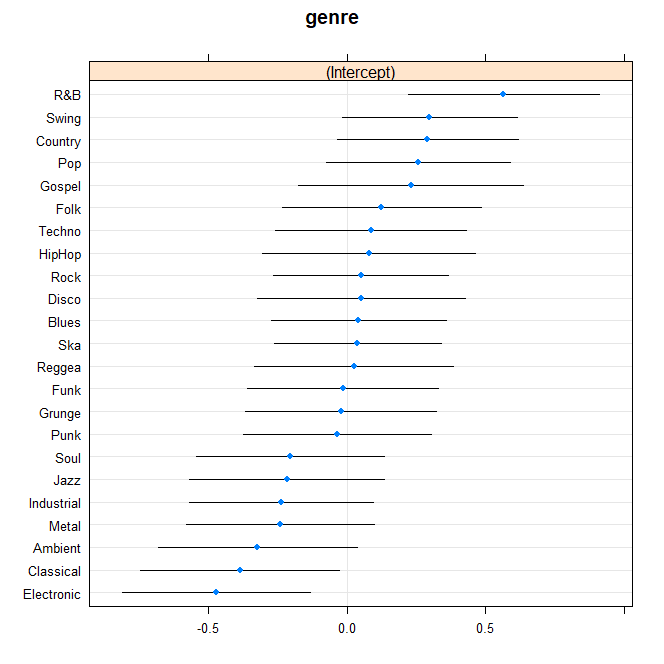
b\_i ~ N(0, sigma2\_b)

Where I is the index which accounts for the genre.

PVRE = 0.1034443 -> mi viene diverso da paolo ma codice uguale

Which can be considered quite high and mean that around the 10% of the variability of danceability is explained by the grouping given by genre

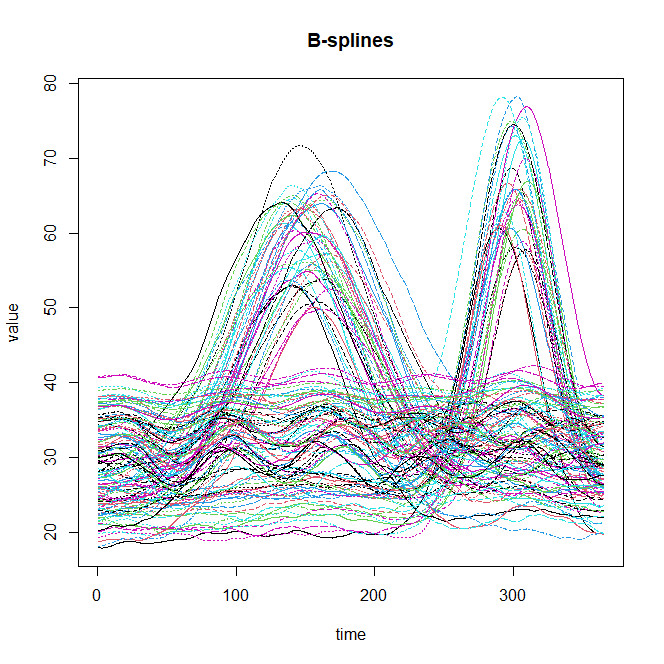
1. Dot-plot:



Net to the effect of the fixed covariate the genre associated to thehigher danceability is R&B because it has the highest value of the random intercept

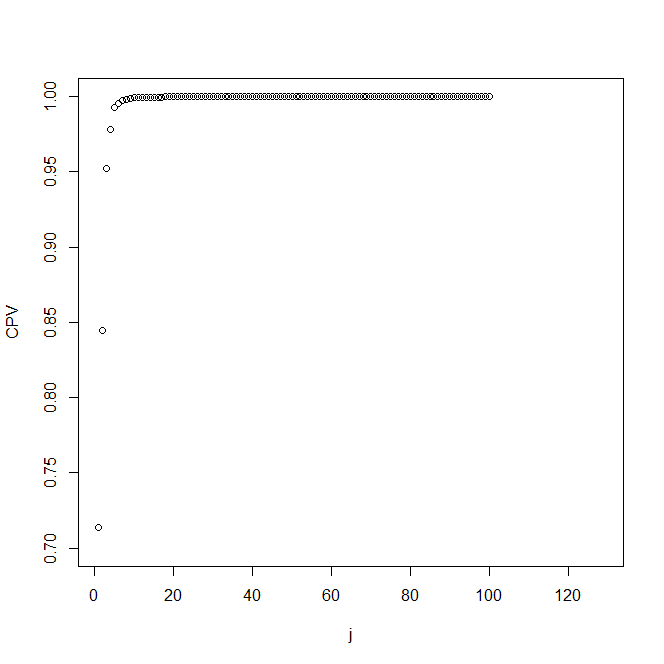
Ex4)

1. CAPISCI



First 3 coeff of the first song: 18.03452 , 17.94459 , 18.05387

1. Scree-plot:



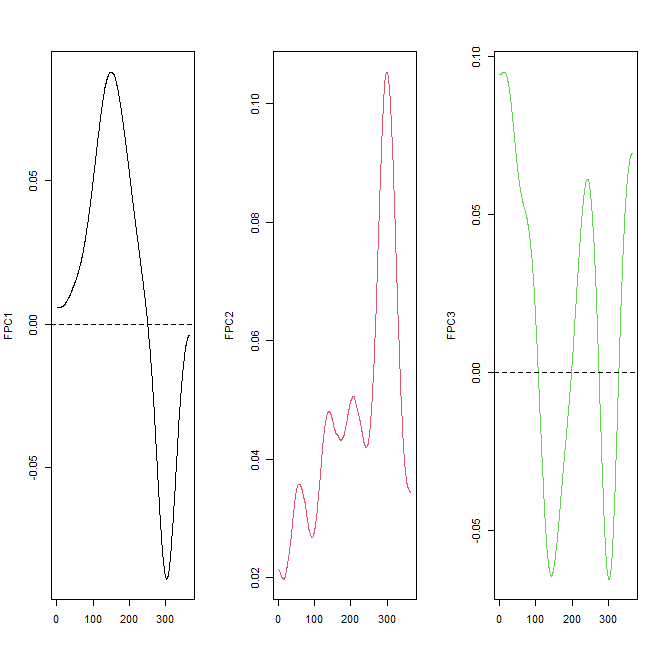
The variance explained by the first 5 FPC:

FPC1 FPC2 FPC3 FPC4 FPC5

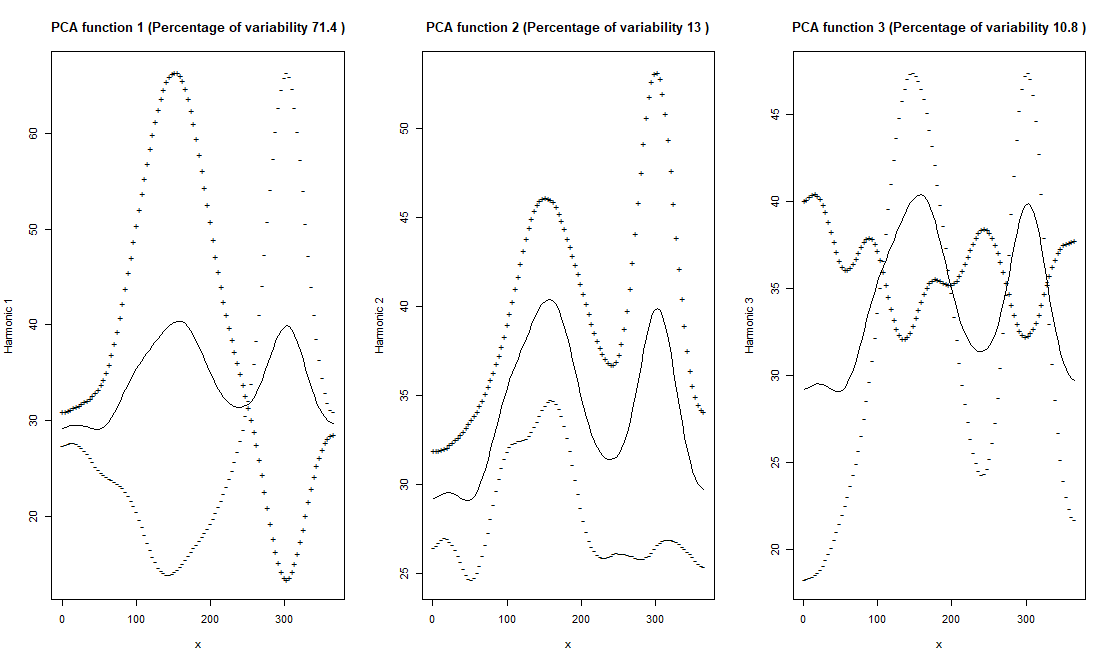
0.71390733 0.13041587 0.10791398 0.02623626 0.01428483

1. We can keep 3 FPC in order to reach a cumulative proportion of explained variability around 95% and moreover we see that from the fourth FPC on the explained variability added is quite low so it does not seem worth it to add them.

Plot:



1. Perturbation of the mean:



Interpretation:

+ = HIGH PC\_i , - = LOW PC\_i

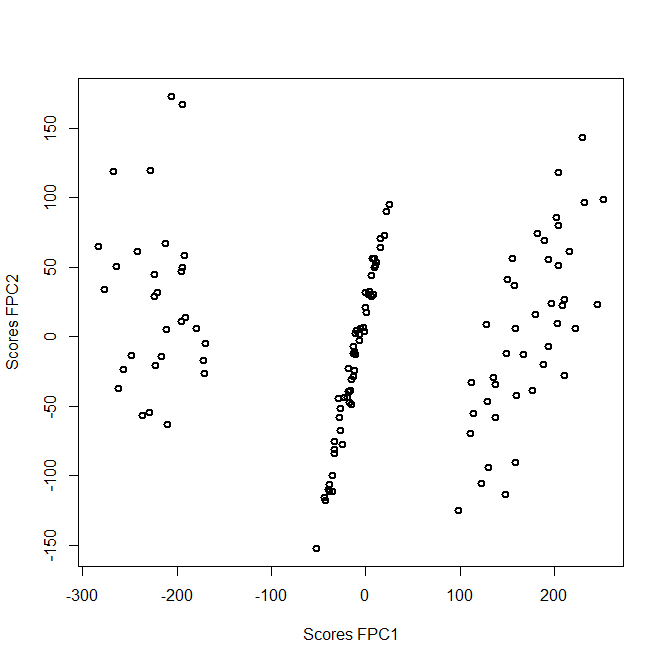
FPC1 shows a contrast between the first part of the year up to the 250-th day more or less and the second part of the year. the functions with positive scores are characterized by more listening in the first part and less in in the second.

FPC2 shows a difference in amplitudine, functions with positive scores are associated with a higher level of listening

FPC3 is again a contrast and we can individuate 4 parts of the year and this is a contrast between the first and the third part against the second and the fourth . in the first and in the third we have higher level for function with positive scores and viceversa for the second and the fourth part.

Paolo: FPC3 -> song with high PC1 have less variation of the listening during the year, songs with low PC3 have instead high variation

1. Scatterplot:



From this we can identify 3 clusters based mainly on the value of scores FPC1: one with negative scores, one with slmost 0 scores and one with positive scores. So recalling the interpretation of the FPC1 the cluster of the left is characterized by more listenings in the final part of the year and less in the first one, the cluster with positive scores the opposite and the cluster with scores around 0 is characterized by listenings in the mean during all the year.

Negative PC1 scores -> low level of PC1

Positive PC1 scores -> high level of PC1