High Dimensional Regression Modeling

Time–Intensity Signal analysis of Chocolate Milk

Gaurav. Remaniche
Shishir. Dubey

Objective

The objective is to fit a classification model for any pair of products out of 6 products, assess the fit and use the model to explain how perception of sweetness differs over time between those two products.

Data

During the tasting tests, sensory evaluation of food is usually based on the award by the judges notes characterizing flavor, flavor or texture of the products. These notes can be regarded as instantaneous ed evaluations of products, thus making no account of Evolution during the tasting of sensations perceived by the judges. However, many phenomena occurring during the tasting of a product are dynamic, like the Evolution of the texture during the mastication or persistence of flavor. It is possible to enrich the static sensory evaluations by the acquisition during the presence of the product in the mouth, a signal called Time-Intensity describing the Evolution over time the intensity of perception.

The variables contained in the data frame are, 10 judges, 6 products, the duration of perception, the signal x (t) describing the evolution of the intensity of perception as a function of time . To allow comparison of 180 signals, the time variation of the beach is the same for all signals , from 0 (beginning of perception) at 1000 (end of the collection), with a 1 unit .

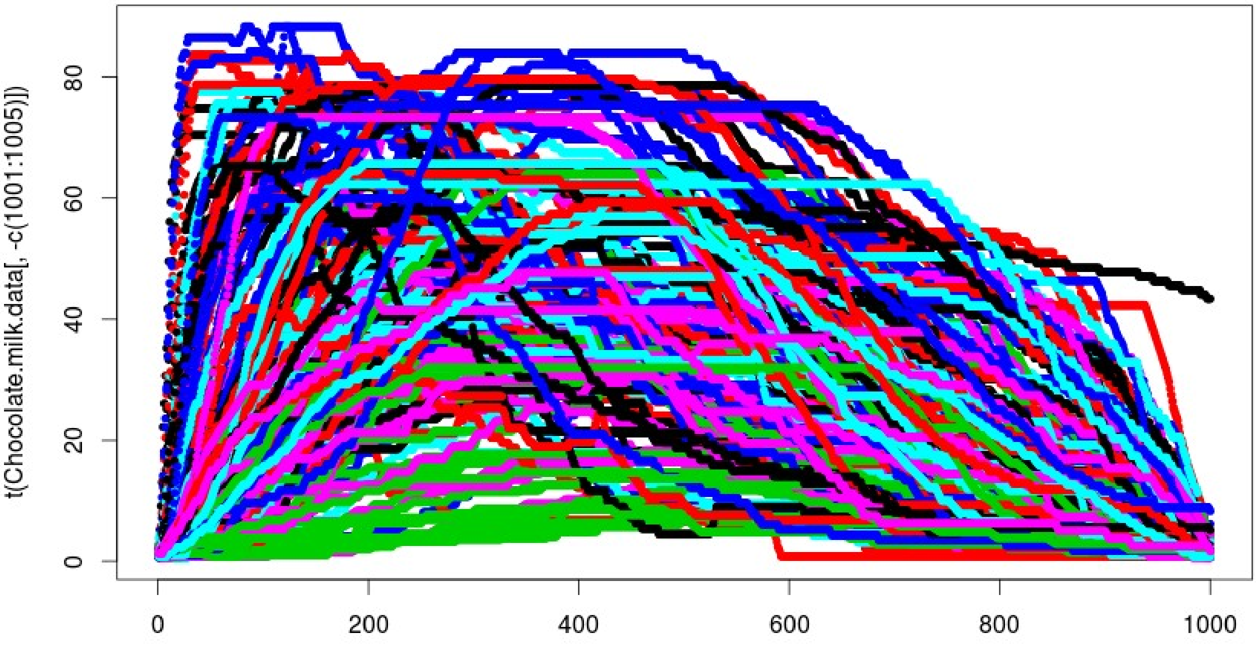

Figure 1

The statistical models used to analyze the variability of IT signals consist , mostly , a reduction in the signal has a finite number of components (area under the curve, duration of perception, slope at zero time , ...). The objective is rather to propose an approach privileging a consideration of TI signal in its raw version . The properties of this approach will illustrate by their application has a set of TI signal collected during the sensory evaluation of the flavor of sweetened chocolate milk drink .

The plot in fig.1 shows the TI signal for all the six products. The TI signal is for perception of sweetness for 180 observations of 6 products. In this dataset PRODUIT is the explanation variable (Y) and a classification model is adapted to analyze the relationship between the explanatory variables and Y.

We have made pair of products C and D for the analysis of perception of sweetness. From the plot in fig.2, we can say that the perception of sweetness of one product is less as compared to another. The variation of sweetness by time according to the different judges for the 2 products. From which the variation of one product is more as compared to other.

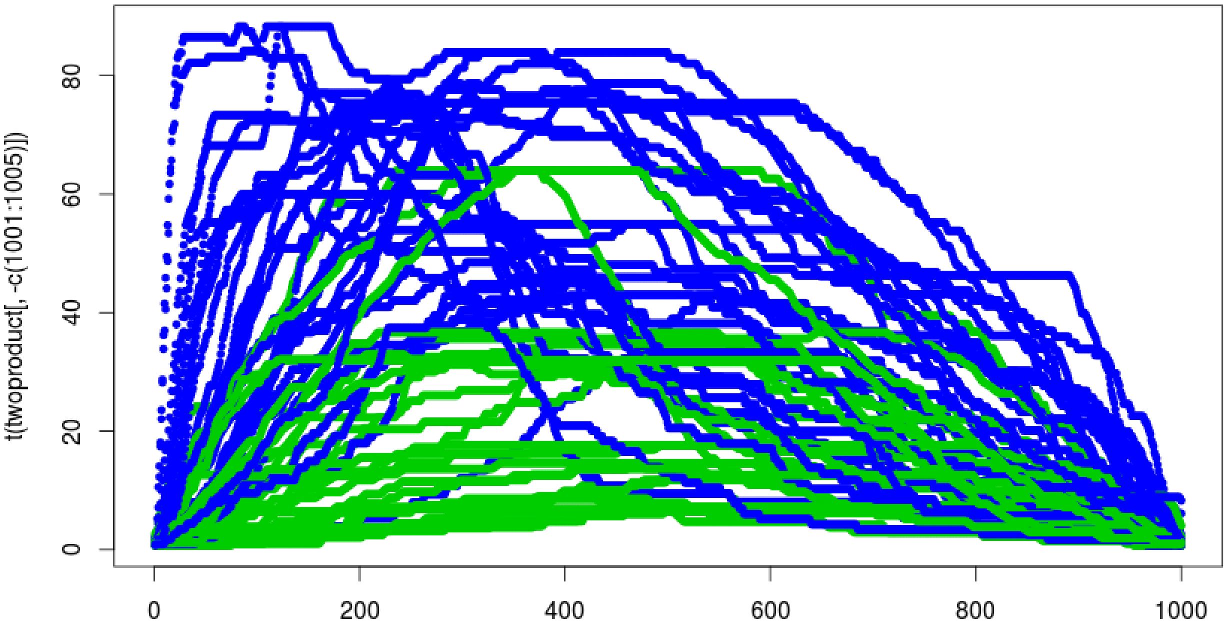

Figure 2

Selection of a regularized regression technique

As we are in the case where p is very high we will need to select the most relevant variables. In consequence we will prefer the use of Lasso over Ridge. If we were to apply Ridge over our data set obtaining 1000 shrined estimates would not facilitate our study. We can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero. It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance. The two best-known techniques for shrinking the regression coefficients towards zero are ridge regression and the lasso.

Method Applied

LASSO

The lasso shrinks the coefficient estimates towards zero. However, in the case of the lasso, the 1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. Hence, much like best subset selection, the lasso performs variable selection. As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression.

Here, we are applying the Lasso regression and cross validation on the two conditions separately in order to identify differences among the two situation.

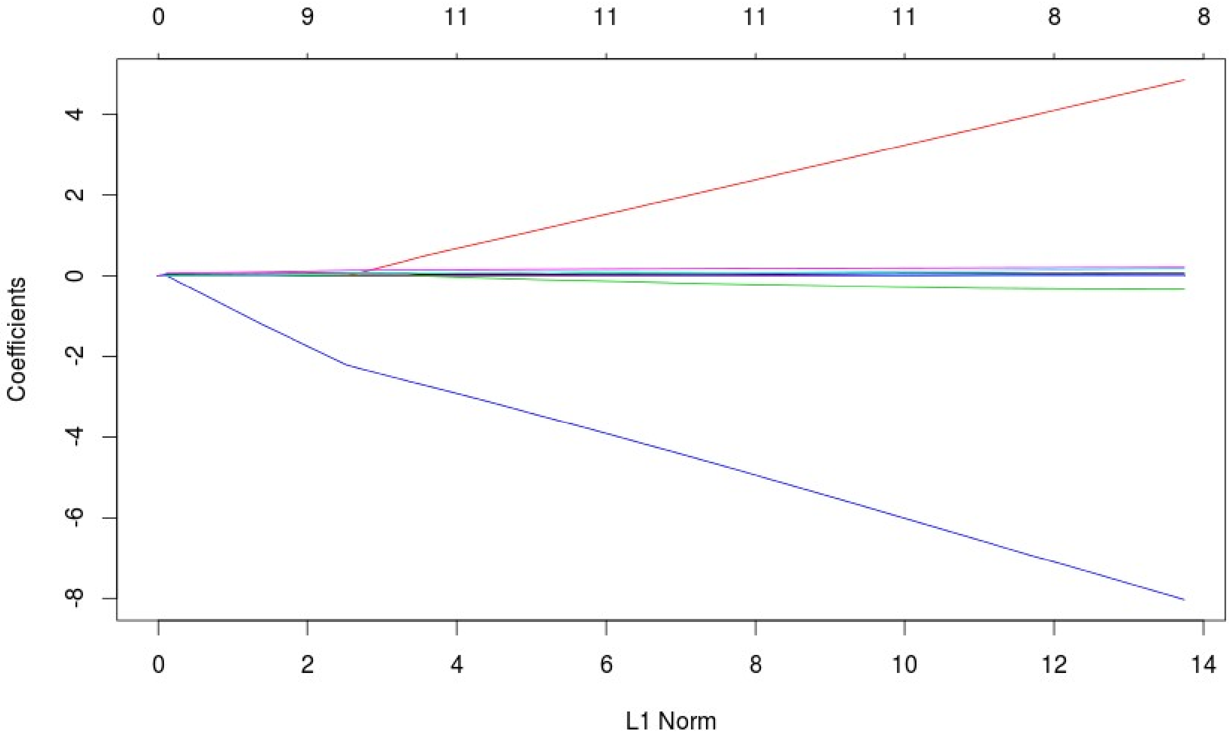

Figure 3 - Lasso for Paired product

The function of glmnet cv.glmnet in R package can choose the λ value that can achieve the best compromise between quality and complexity of adjustment within the meaning of the number of non-zero parameters , the model.

Firstly, we apply Lasso regression on paired product of training dataset. In order to fit a lasso model, we use the glmnet() function, with use the argument alpha = 1. We can see from the coefficient plot that depending on the choice of tuning parameter, some of the coefficients will be exactly equal to zero. We now perform cross-validation and compute the associated test error.

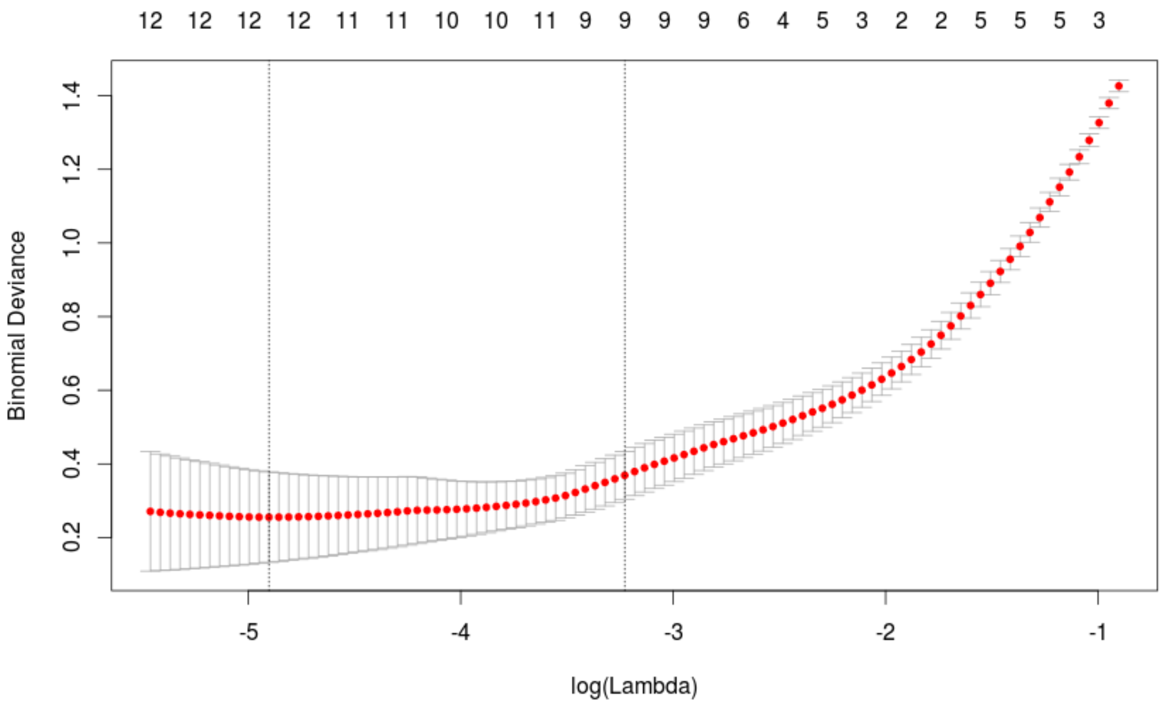

Figure 4 – CV on training dataset

If we observe the plot of the results using the default values for lambda, we can see that the lambda that yields the lowest RMSE is the smallest, meaning that we should start with a lower value for our evaluation. We will choose the lambda 0.004249503 that keeps about 11 variables and minimizes the Cross-validation RMSE .The CV is done keeping 10 folds. The variables of importance are: "T225", "T226", "T227", "T282", "T298", "T593", "T701", "JUGEJ8", "DEBUT", "DUREE". The first seven variables are the TI signals for perception of sweetness. Here, we fit the model with keeping lambda minimum and calculating good fitted rate, which is 0.9777778.

```
> fitted.value=predict(lasso2,x[-train,],type = "class")
> good.fitted.rate<-mean(y[-train]==fitted.value)
> good.fitted.rate
[1] 0.9777778
```

Similarly, we analyze predicted rate for same,

```
> pred.value=predict(lasso2,x[train,],type = "class")
> good.predicted.rate<-mean(y[train]==pred.value)
> good.predicted.rate
[1] 0.9333333
```

Now, we apply cross validation on the overall dataset. If we observe the plot of the results using the default values for lambda, we will choose the lambda 0.003871132 that keeps about 11 variables and minimizes the Cross-validation RMSE .The CV is done keeping 10 folds. The variables of importance are: "T298", "T299", "T300", "T542", "T548", "T707", "JUGEJ2", "JUGEJ3", "JUGEJ8", "DEBUT", "DUREE". The first six variables are the TI signals for perception of sweetness. Here, we fit the model with keeping lambda minimum and calculating good fitted rate, which is 0.983333.

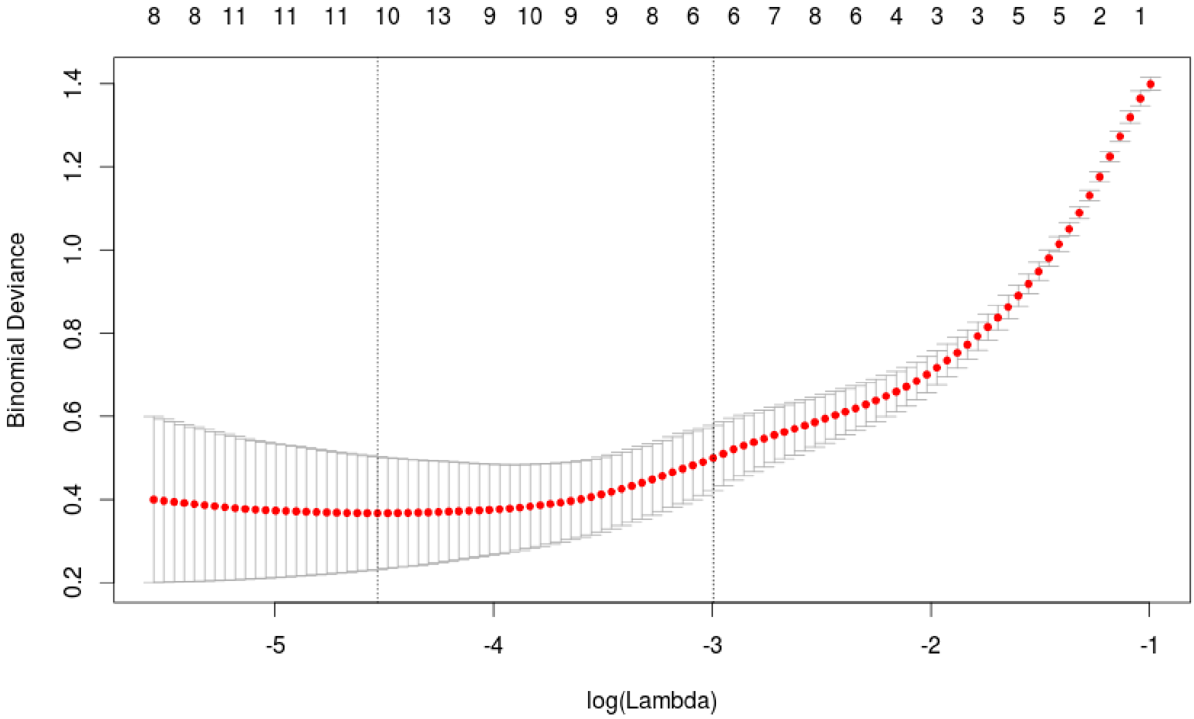

Figure 5 – CV of whole dataset

```
> fitted.value=predict(lasso3,x,type = "class")
> good.fitted.rate<-mean(y==fitted.value)
> good.fitted.rate
[1] 0.983333
```

Similarly, we analyze predicted rate for same,

```
> pred.value=predict(lasso3,x,type = "class")
> good.predicted.rate<-mean(y==pred.value)
> good.predicted.rate
[1] 0.967777
```

Re-sampling

To compare the respective contributions of the length of perception and Time - Intensity signal is adjusted by non- regularized method of estimating the model including the effects Judge, Duration of perception and Intensity of perception, or intensities are reduced values selected in the model. This model is sampled 100 times and mean of good fitted rate(gfr) and mean of good predicted rate(gpr) are calculated.

```
gfr<-c()
gpr<-c()
for(i in 1:100){
  train<-sample(1:length(y),length(y)/4)
  mod.multi <- multinom(PRODUIT~., data=twoproduct[-
train,c("PRODUIT","T298","T299","T300","T530","T548","T707","JUGE","DEBUT
","DUREE")], trace=F)

  fitted.value=predict(mod.multi,twoproduct[-train,],type = "class")
  gfr[i]<-good.fitting.rate<-mean(as.vector(twoproduct$PRODUIT[-
train])==fitted.value)

  pred.value=predict(mod.multi,twoproduct[train,],type = "class")
  gpr[i]<-good.predicting.rate<-
mean(as.vector(twoproduct$PRODUIT[train])==pred.value)
}

plot(gfr,xlab="Resampling (1 to 100) ", ylim=0:1)
points(gpr,type = "l",col="red")
```

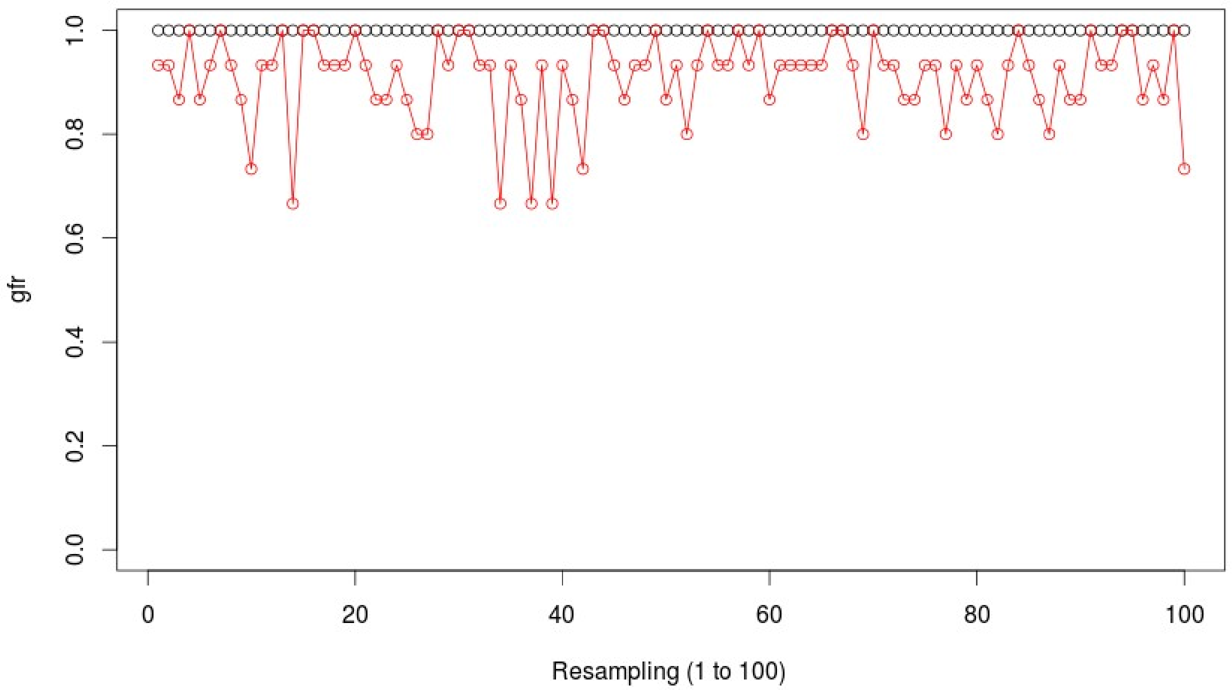

Figure 6 – Re-sampling for Selected Variables

So form above plot we can say that the least minimum predicted value is greater than 0.6. The mean of fitted rate is 1 and mean of predicted rate is 0.903333.

The confidence interval can be calculated as below,

```
> error <- qt(0.975,df=length(gpr)-1)*sd(gpr)/sqrt(length(gpr))
> error
[1] 0.01623049
> left <- mean(gpr)-error
> right <- mean(gpr)+error
> left
[1] 0.8931028
> right
[1] 0.9255638
```

Conclusion

The objective of this study was to highlight perception of sweetness by analyze the shape of TI curves, associated with the extraction of key features such as maximum of perceived intensity, duration of perception and time of beginning of perception. As we are in a high-dimensional setting we turned to applied commonly used method LASSO. The technique was applied using “Leave-One-Out” cross validation as we have a small number of individuals. So we seek that, the variations of trained dataset of paired product and that of whole dataset of paired product is less. The least minimum predicted value is greater than 0.6. The mean of fitted rate is 1 and mean of predicted rate is 0.903333. The perception of sweetness for two products is good as the mean of fitted rate and predicted rate are near 1.

References

- [COURSE] D.Causeur “High dimensional regression modelling”, 2015-2016, ENSAI
Master Big Data

- “An Introduction to Statistical Learning with Applications in R”, G. Casella , S. Fienberg , I. Olkin.

- http://math.agrocampusouest.fr/infoglueDeliverLive/digitalAssets/84891_exam13.pdf