Sweetness Intensities for Chocolate Milk Products

Authors: Anurag Miglani, Thomas Zamojski
Course: Penalized Regression
Submitted to: Professor D.Causeur

18 nov. 2015

1 Introduction

The agroindustry is interested in testing its products to control their qualities. In this report, we investigate five chocolate milk products that have been tasted and registered according to their sweetness intensities. The experiment consists of having ten judges tasting six products without them knowing which one they are. The six products presented to them are one of each chocolate milk plus one product we label E being repeated once. It was presumably asked of them to record in realtime their impression on the level of sweetness felt in the first 100 seconds by tracing a curve on scaled paper. The experiment is itself repeated three times. We do not have any information about how this has been done, but supposedly the order of the products through these sittings have been shuffled. This would ensure a better estimate of modelling errors.

As for most human experiences, the intensities of sweetness felt during tasting depends on the person and even on other factors like fatigue, the sentiments of the day, and so on. We therefore expect the data to be quite noisy, both in amplitude and phase. This makes the investigation inherently difficult, but some information can nonetheless be extracted from the experiments.

We propose to investigate two main questions that arise naturally in this situation.

Q1: **Can we find intrinsic features of the products that can distinguish them?** These features will have to be somewhat universally agreed upon by most individuals to be called intrinsic to the product. This will amount for us to fit a single classifying model to all judges.

Q2: **Can a person learn through time to distinguish the products?** If so, after an initial phase of tasting many products, an individual could then have a definite preference for one product over the others.

Ideally, one would like to answer positively to both questions. However, more realistically, we expect some *set* of products to be similar to each other, giving a coarser classification into groups of products rather than individual products.

Our findings on the data were that the product labeled C was by far the less sweet. Of the remaining four products, the product labeled D could be sometimes distinguished from the others as well, but

not always. In general, the products were mostly indistinguishable, except for C which was clearly different.

2 A Look at the Data

There are ten judges labeled J1 through J10, six products labeled A,B,C,D,E and E2, where presumably E2 is the repetition of E in a tasting session, and there were 3 sessions, labeled 1, 2 and 3. For each triple of judge, product and session, a curve of sweetness intensity is recorded at a resolution of 0,1 second from 0 to 100 seconds. This gives us 180 curves with a quite good resolution.

We propose to visualise the data by fitting smooth splines with 8 degrees of freedom. Smoothing splines fit into the subject of the course as it is a form of penalized regression, where the penalty is given by the integral of the square of the second derivative.

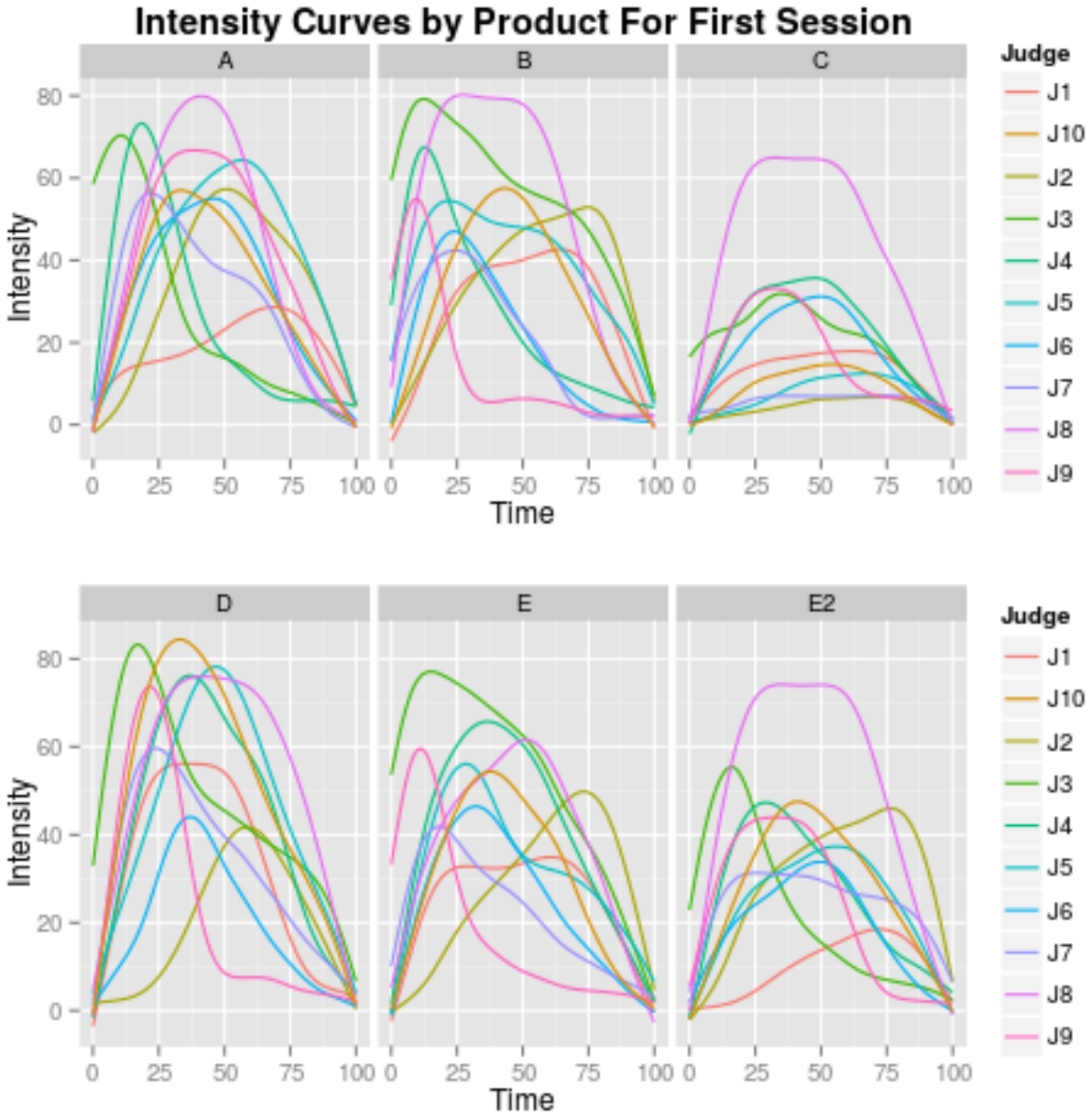

Figure 1: Judges exhibit high variance in between themselves for same products.

A quick glance at these plots enable to make simple conclusions that we announced in the introduction already. The graphs by product show that indeed different judges have a quite different view

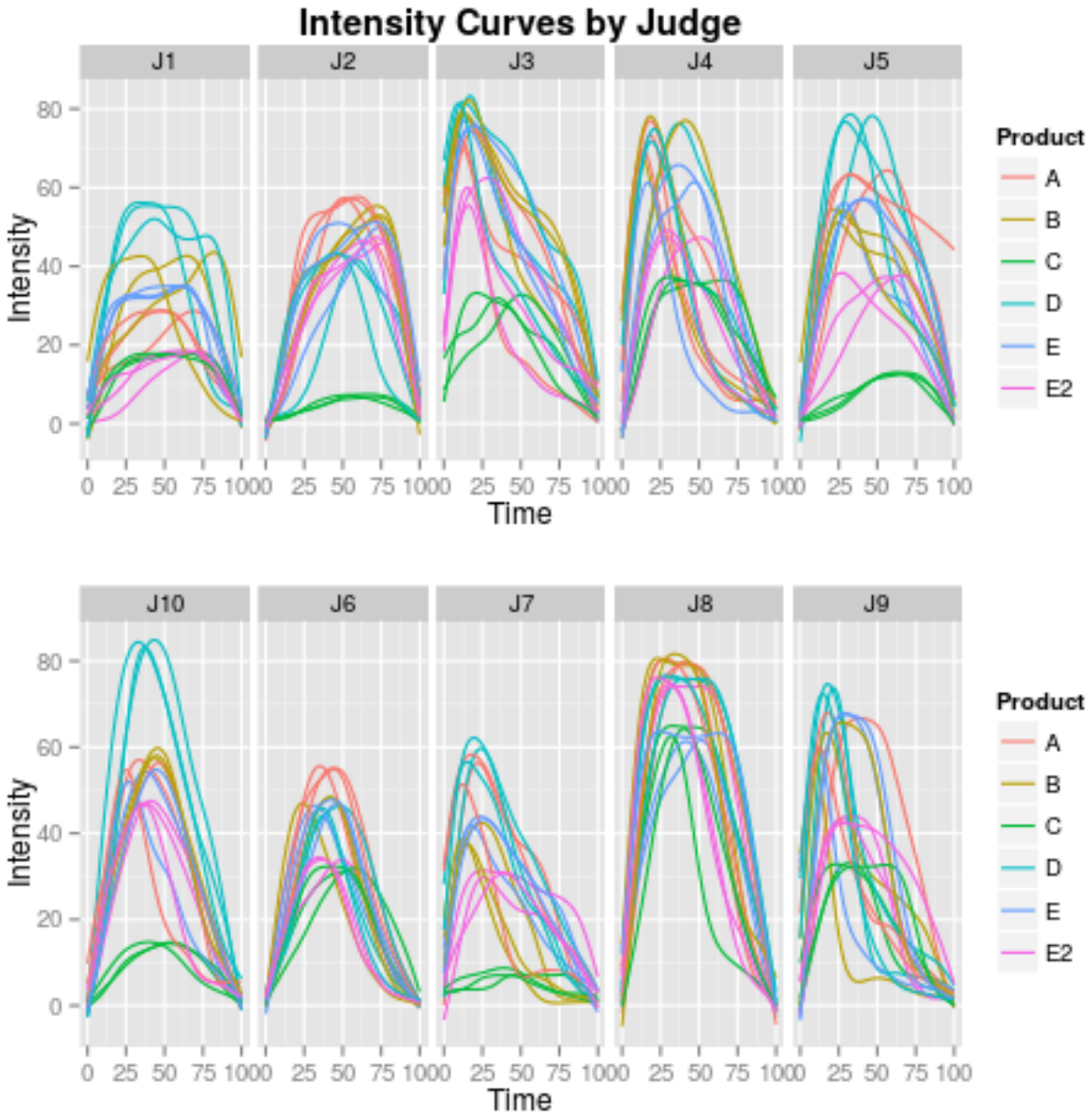

Figure 2: Most products don't seem to be distinguished by judges, except C.

of the sweetness of products. The graphs by judge show the view of each judge on the products. We distinguish in it that for most judges, products A,B,E'E2 seem indistinguishable, and product C has lower intensities. We see however that product D is sometimes distinguished by some judges as being sweeter.

2.1 Registration

While the curves plotted above were done without registering the curves, that is without any phase-change, we looked for ways to do so. The shape of the curves is mostly the same and consist of only one "hill". Registration of such curves could be done via aligning the peaks, i.e. the maximum, through out the sessions, but within each judge and product. As the peaks last for some time, we could register either both the start and the end of the peaks, or the middle of the peaks. In either case, we could average the times over the different sessions, and make a piecewise linear time-change to maintain the 0 to 100s time interval.

Registration could be an important factor here, as it could reduce the noise in the phase, but we

have not found evidence for it. Other methods for registering could be investigated in further analysis, such as registering also the 0.75 of the max levels as well (two more points).

2.2 Maxima

Registering the maxima brought us to the following observation: the maxima for each judge and product are all equal through the sessions (shown in the next figure). This is most unlikely for the raw data, especially considering that it is noisy. Therefore, we are forced to conclude that the data was preprocessed, at least at the level of amplitude. This is reinforced by the fact that all maxima are integers divided by the number of sessions, suggesting further that maxima were registered using a mean over the sessions. This is going to cause us considerable problems later on.

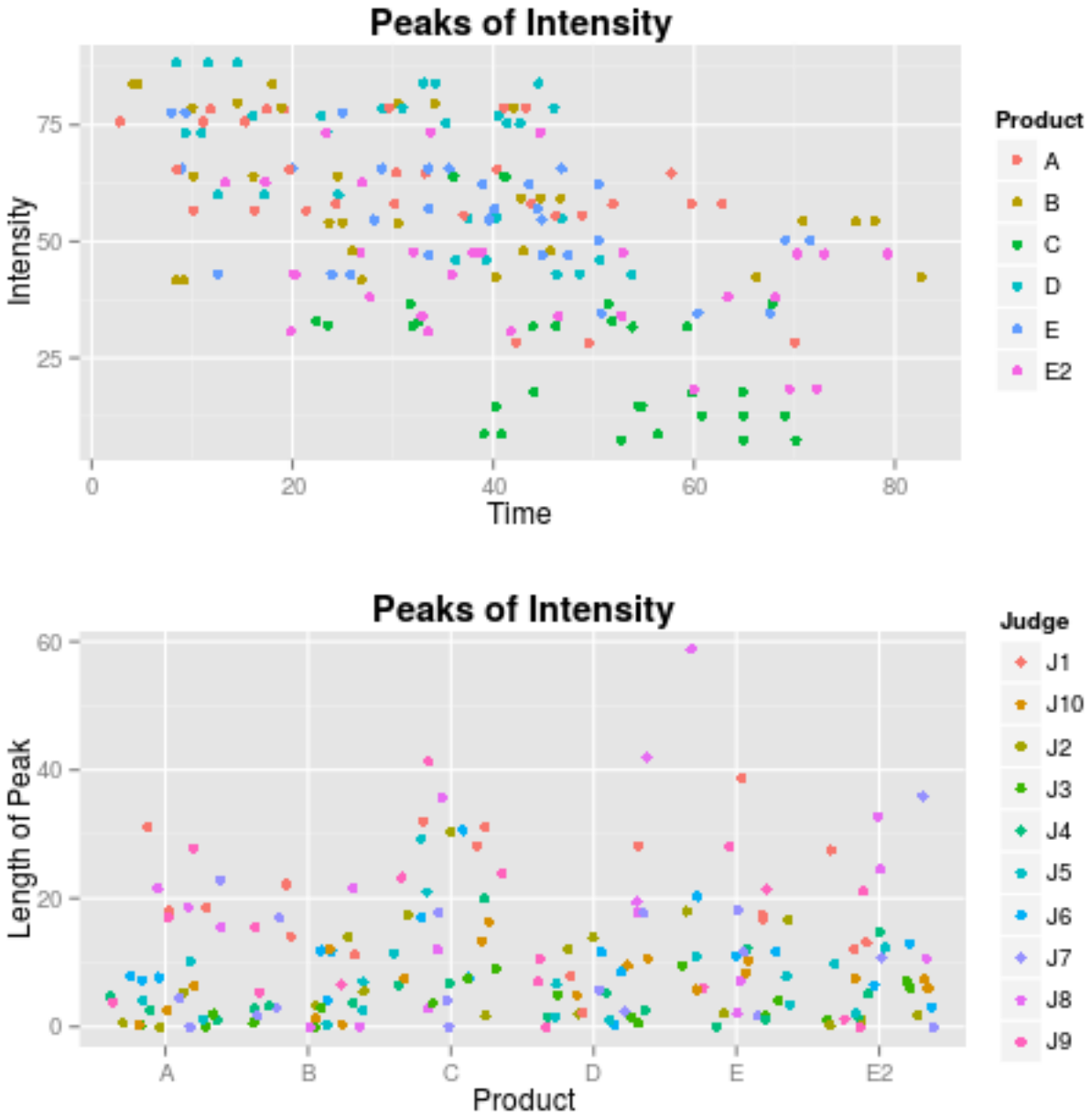

Figure 3: Above: plot of the maximum intensities against the time they occur. The amplitude of the maxima do not depend on the sessions, showing that they must have been prepocessed.

3 Prediction Errors and Cross-Validation

The following stratified cross-validation scheme is used to estimate the prediction error: for each judge, for each product, we reserve at random one session for independent testing, the two others going for training the model. This gives us 3^{60} possibilities for the splitting, quite enough to avoid any overfitting of the cross-validation procedure.

However, as noted above, the maximum of curves seems to have been registered already before the data arrived to us. We face therefore the difficulty of having an underestimated prediction error. In fact, if a classification model uses greatly this maximum value, it will be largely overfitting. For example, in the case of fitting a model to one judge, the maximum sweetness intensity is a perfect classifier, leading to no error whatsoever. This of course cannot be.

Therefore, to avoid overfitting, we avoid registering the curves according to their peaks. Indeed, this would lead to peaks happening at the same time and same height within each judge and product. This could be picked up by the linear models below, and would lead to an underestimated prediction error.

4 Modeling using Regularized Logistic Regression

The intensity curves exhibit high auto-correlation, and it is therefore not appropriate to use multivariate analysis on the time mesh to model them. We prefer to use regularized linear models, as they take into account that if the intensity at a given time is given some importance, then the intensities over the time close to it should also be given importance. Therefore, the coefficients of the linear model should also be treated as a continuous curve.

A linear decision boundary can therefore be obtained in the following way: if $X(t)$ denotes a random curve of sweetness intensities, then the linear terms in the logistic regression will be of the form:

$$\beta_0 + \int_0^{100} X(t)\beta(t)\,dt, \quad \text{for some coefficient function } \beta(t) \text{ and constant } \beta_0.$$

As usual in logistic regression, the log-odds are assumed to be such a linear form of the intensity curves.

We pick a big enough dictionary of basis functions ϕ_k consisting of Fourier basis functions. Specifically, we took the 31 first ones, giving us frequencies up to 0.15Hz. Note that the resolution of our data being 0.1s, this will amount to a certain amount of compression. However, as it is often the case, we find that not much information is contained in higher frequencies, because of the high auto-correlation. Each intensity curve $X_i(t)$ is expanded in the basis functions using ridge regression with about 8 degrees of freedom. Here, the penalty is as in smoothing splines, the integral of the square of the second derivatives. Similarly, each coefficient function $\beta(t)$ is also expanded in terms of the basis functions. However, to avoid overfitting, the coefficients are expanded using only the first 5 or 11 basis elements, that is up to frequency 0.02Hz or 0.05Hz. The problem of fitting the coefficients is then brought back to multivariate analysis. We have made use of cv.glmnet with ridge regression (alpha=0) and the multinomial family.

Explicitely, if t_i denotes the i^{th} time point, we have:

$$\begin{aligned}
\Phi_{ij} &= \phi_j(t_i) \\
R_{ij} &= \int_0^{100} \Phi_i''(t)\Phi_j''(t)\,dt \\
Y_{ij} &= X_j(t_i) \\
J_{ij} &= \int_0^{100} \Phi_i(t)\Phi_j(t)\,dt
\end{aligned}$$

Then the curves $X_j(t_i) = \Phi A$, where the matrix of coefficients A solves the equation:

$$(\Phi^T\Phi + \lambda R)A = \Phi^T Y.$$

Also, the input data to the logistic regression model is now $A^T J$, that we truncate as mentioned to avoid overfitting the β's. The output coefficients are the coefficients of β's in the Fourier basis expansion.

To seek features that would distinguish the products independently of the individual tasting it, we fit the model to all intensity curves altogether, independently of the judge or session. For simplicity, we decided to exclude product E2 from the data, as relabeling it as E would bring even more noise. The following curve is a graph of the coefficient of the log-odds, picking E as the reference class. Given the oscillation, it seems to fit the data more than anything, an indicator that the model did not find any strong feature that could distinguish the classes. However, as usual the curve for product C is different.

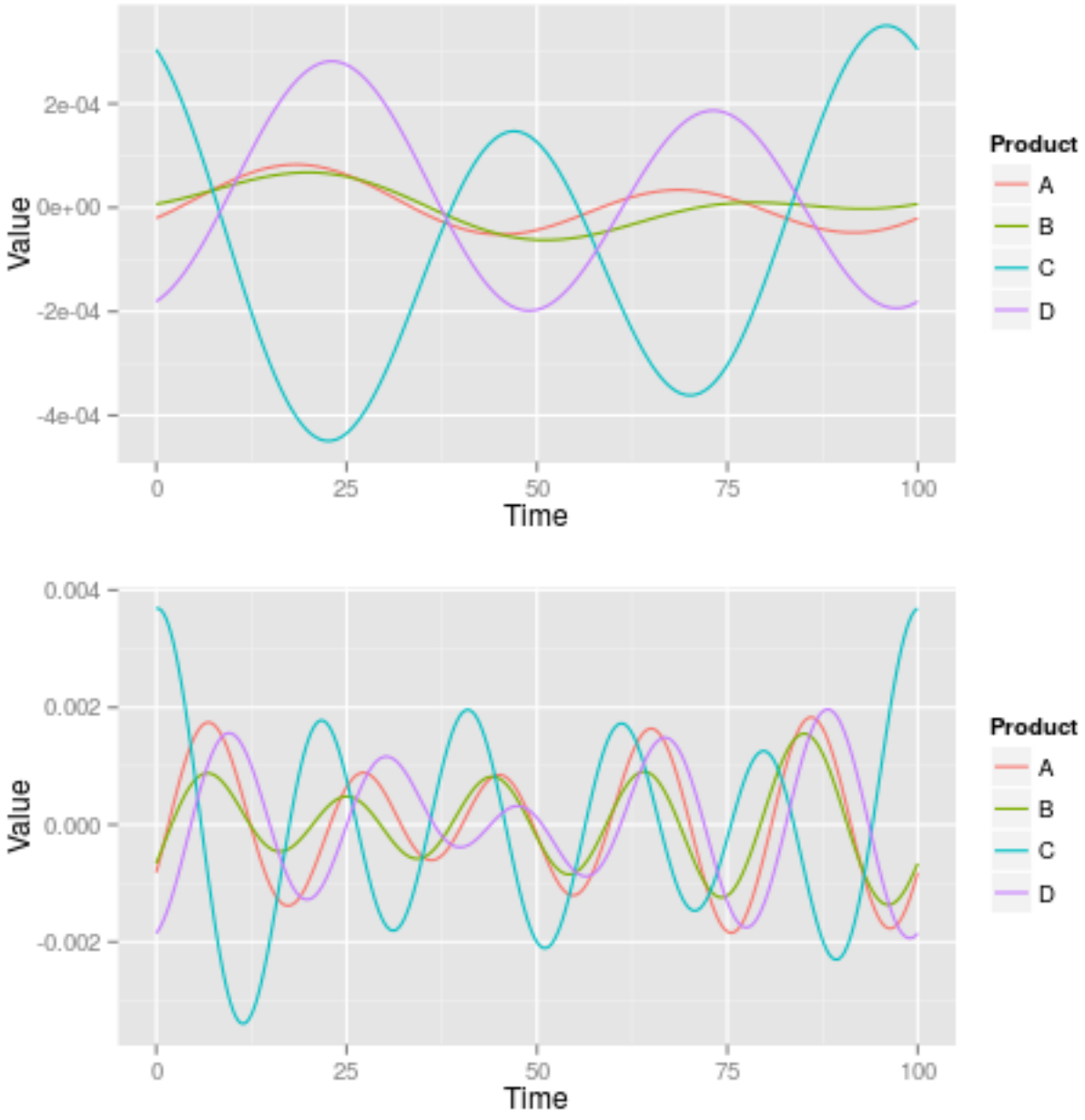

Figure 4: The coefficients of the log-odds with product E as reference,for k=5 and k=11 basis elements. The oscillation indicates that it did not find strong classifying features in the data.

Using cross-validation as described above, we can get an estimate of the prediction error. Using 50 rounds of the cross-validation process and using the misclassification error, we have found the following total prediction error and errors by product class:

We note that random guessing would give us a 0.8 error across all products. This the improvement of the model really is in product C alone, and perhaps product D.

This suggests to try to put products A,B and E together into a single group that we label class ABE. This however did not improve much the estimates, meaning that there remains confusion between

Total	A	B	C	D	E
0.650	0.900	0.855	0.040	0.470	0.985

Table 1: Prediction errors total and by product class.

product D and ABE. It indicates that overall, even product D seem to be only distinguishable from C alone.

Total	ABE	C	D
0.509	0.638	0.070	0.560

Table 2: Prediction errors total and by product class, putting A,B and E together in one class.

5 Pooling over Individuals

We would like now to address the second question: can individuals classify the products better on their own? We would like to allow judges to each have their own model. This however poses two problems: one is that there are only 3 sessions, so in the training set only 2 observations per label. The second problem comes from the registered max mentioned earlier. Here, there is real danger of overfitting, as the max would give a perfect classifier for each individual. In fact, in our early attempts, this was indeed happening, leading to the false conclusion that we could improve our prediction error by making a vote over each judge for classification.

As of now, we have left this idea for further investigation, but we believe that having the raw data in which there is some within class variance over the maximum to be primordial. The low number of observations could also be address by having more data.

6 Conclusion

We briefly summarise our findings here. Product C was very well distinguished from the others. This is due to the fact that it is far less sweet than the other products, which was shown for example by the low coefficients in a regularised logit regression. Product D was distinguished by some Judges, as shown by the graphs and the cross validated prediction error table for the model. The other products cannot be classified using sweetness intensity curves, at least, not according to the data at hand.

We believe that the estimates presented could be improved, but not so drastically. The prediction error is an indicator and is underestimated in our method due to what seems to be pre-registration of the max intensities.

Further lines of investigation would include curve registration (but not according to peaks), selection of time intervals, weighting regularisation with an estimate of variance of an additive error using AR models (which are good even if we have few observations) and using boosting.

For references, data and scripts, please feel free to ask any of us.