Advanced Research Methods Cookbook

Cristina Rivera

10/1/2022

Table of contents

Preface

This is my cookbook for my Advanced Research Methods Class with Dr. Abdi. The purpose of this book is to have a reference for all the Advanced Research Methods I have learned in this class and give an overview of each method and an interpretation of the findings using each method. This is for my future self, or anybody who find this.

1 Introduction

Hello, my name is Cristina, and I am a student in the Masters of Applied Cognition and Neuroscience program here at the University of Texas at Dallas. This book is my Advanced Research methods Cookbook created in R studio using Quarto. Each "Recipe" will include a method I learned in class. As I understand more and more, I will add to this book! So far, I have one method, PCA, and I will add more as I learn more. The purpose of the Cookbook is to serve as a reference for myself later in the future, although just making the book has been so helpful in solidifying those connections, so I may not even need it! Each chapter has three sections:

1. Method Explanation
2. About the Data
3. Interpretation of the Data

Each of the data analysis graphs were rendered using R studio. The Interpretation of the Data will be based off graphics and given results. Thats it. Enjoy.

Part I

PCA

2 PCA Basics

PCA stands for principal component analysis.

2.1 What does PCA do?

1. PCA reduces dimensions
2. Helps us visualize the variables since we can't see beyond 2 dimensions and its hard to visualize things in more then 3 dimensions
3. Reduces noise in the data

2.2 What is PCA's goal and why?

1. Reduce Noise
   Reducing Noise in your dataset can be useful when you want to focus on what is important. Its easier to visualize the data when you have less dimensions. Not all data carry the same weight.
2. Compressing the dataset
   Compressing the dataset can make it easier and faster for machine learning applications or other types of analysis to go through the data
3. Helps use Visualize our dataset
   Each component will have variables that will often correlate with each other. Running a PCA can help us see those variables that have high levels of correlation with other variables.

2.3 Where is PCA used?

- In machine learning applications
- PCA technique is particularly useful in processing data where **multi-colinearity** exists between the **features/variables**.
- When there are a lot of dimensions in your data
- Can be used for denoising and data compression

2.4 PCA Terms

Below are some terms that you will want to know before you jump in to the analysis.

Observations These will be composed of our variables and their values. Observations can be separated by time, by categories (for example individual people can have their own observations) or other factors.

Variables Part of an observation, could be something like a rating for a magazine. Needs to be continous.

Standardization and Mean centering A way of making sure that each variable is on the same scale as all the other variables. if one variable is on a larger scale then the larger variable value could over take its contribution to the components. For example think of housing prices and number of rooms. If they were standardized then housing prices would over be over represented while number of room would be under represented.

Components These are the new variables which are linear combinations of the initial variables.

Component 1 is always the component with the most variance with each subsequent component having less variance then the previous but more variance then the components after. Each component is orthogonal to each other.

Covariance How two variables relate to each other . Can be positive, negative, or zero.

Correlation Similar to covariance except it is restricted to a certain range.

Eigenvectors contains eigenvalues from highest to lowest of all of variables.

Eigenvalues we can think of eigen values as weights. It is important to keep eigenvectors with eigenvalues that contribute significantly to the components.

2.5 How are the Components created?

The components are created using the eigenvectors and eigenvalues calculated from the covariance heat matrix. One thing that happens when we are creating components is that we don't use all of our eigenvectors which results in dimension reduction with some loss of accuracy. This loss in accuracy is always something to keep in mind as you will want to find the right balance of keeping accuracy loss low while not using the dimensions that don't contribute much.

3 The Data

Before we analyze our datas findings, lets talk some about the data we will be working with. Our Data can be described as some kind of strawberry yogurt tasting with 8 Panelist per yogurt. The panelist each taste 7 yogurts and rates the yogurts based on 24 different sensory factors which we will call our variables. Since we have 8 panelist and 7 yogurt we will have 56 **observations** in total. Each observation has a total of 24 different sensory attributes that the panelist rate the yogurts so we don't know what the data looked like before it (but could extract the original values if need). In this example our sensory attributes are our **variables** Some examples of our variables are: Overall Aroma, Overall Flavor, Cheesiness, Fruitiness , Sweetness etc. The Data is mean normalized.

Our data:

- 8 panelist
- 7 different strawberry yogurts
- 56 observations
- 24 variables per observation

4 Interpreting the Data

In this section I will be interpreting the Data

4.1 Covariance Heat map

First, lets take a look at our covariance heat map of our strawberry yogurts and there sensory inputs. Our covariance heat map is great for looking at our variables at a glance to see which ones may have a larger/smaller effect on our components.

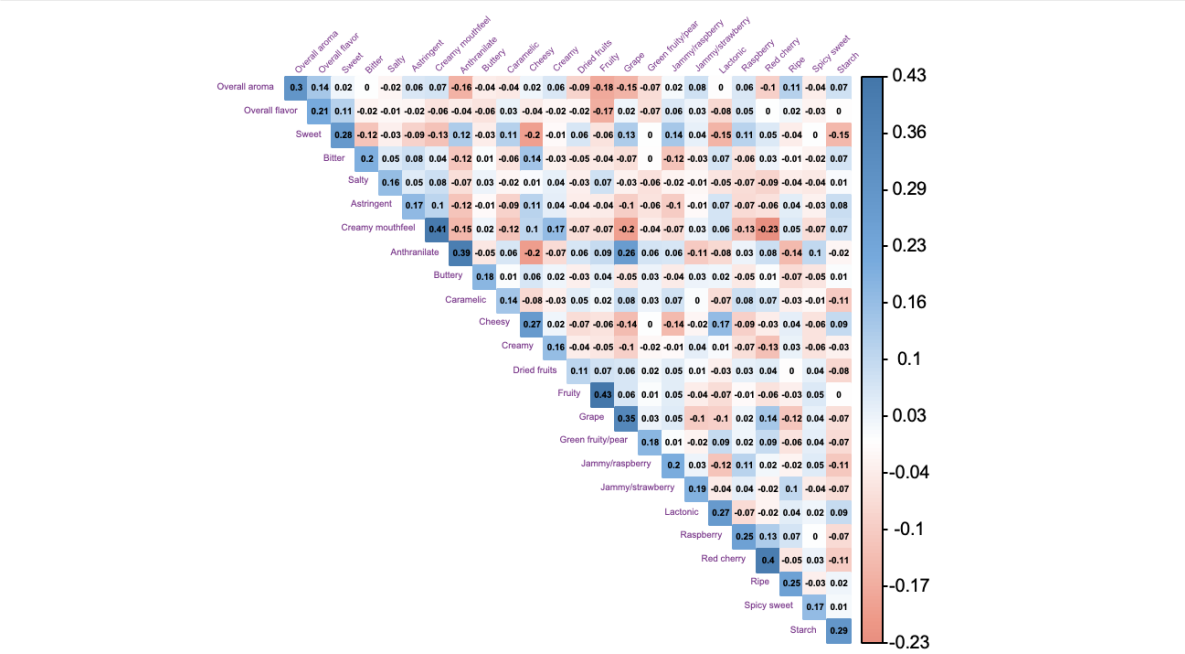
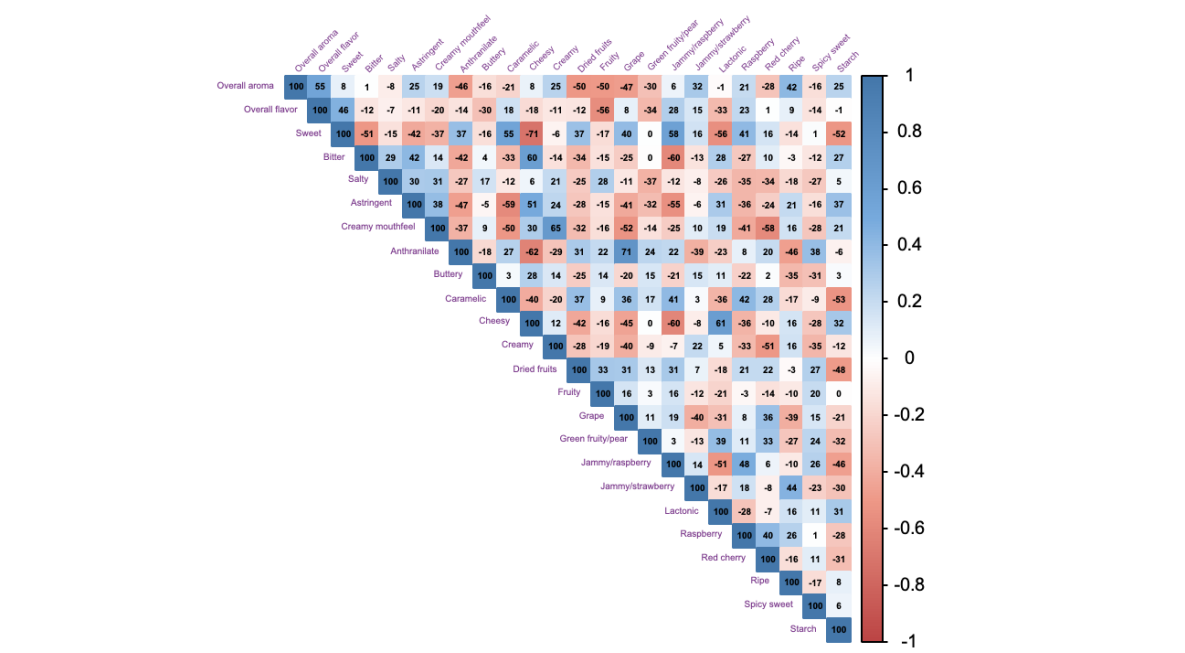


Figure 4.1: Covariance Heat Map

First, we will look at some of the variances on the covariance heat map, which include all our variables on the left-hand side. Note that all the variable's variances are positive since they are not negative. One thing to note is that the larger the variance, the more the variable will co-vary, indicating it will be an essential variable. So our top variance is Fruity with .43. This one may be one to look out for (or maybe not).

4.2 Correlation Heat map

Next, lets take a look at some of our correlations.



For a PCA, it is essential to look at our correlations; after all, if two variables are highly correlated with each other, then it is time to take a look at whether it would be advantageous to weigh them equally. In this data, the most notable correlations are:

4.2.1 Top Correlations:

1. **Sweet and Cheesy** -71
2. **Anthranalite and Grape** 71
    - Note: Anthranalite is actually an artifical grape flavoring that is used in products like grape kool-aid so it makes sense that they have one of the strongest correlations!
3. **Anthranalite and Cheesy** -62
4. **Cheesy and Lactonic** 61
    - Note: Lactonic in this context is a "milky" type of flavor. So, again, it makes sense that they are highly correlated with each other
5. **Bitter and Cheesy** 60
6. **Bitter and Jammy/Raspberry** -60
7. **Astringent and Caramelic** -59

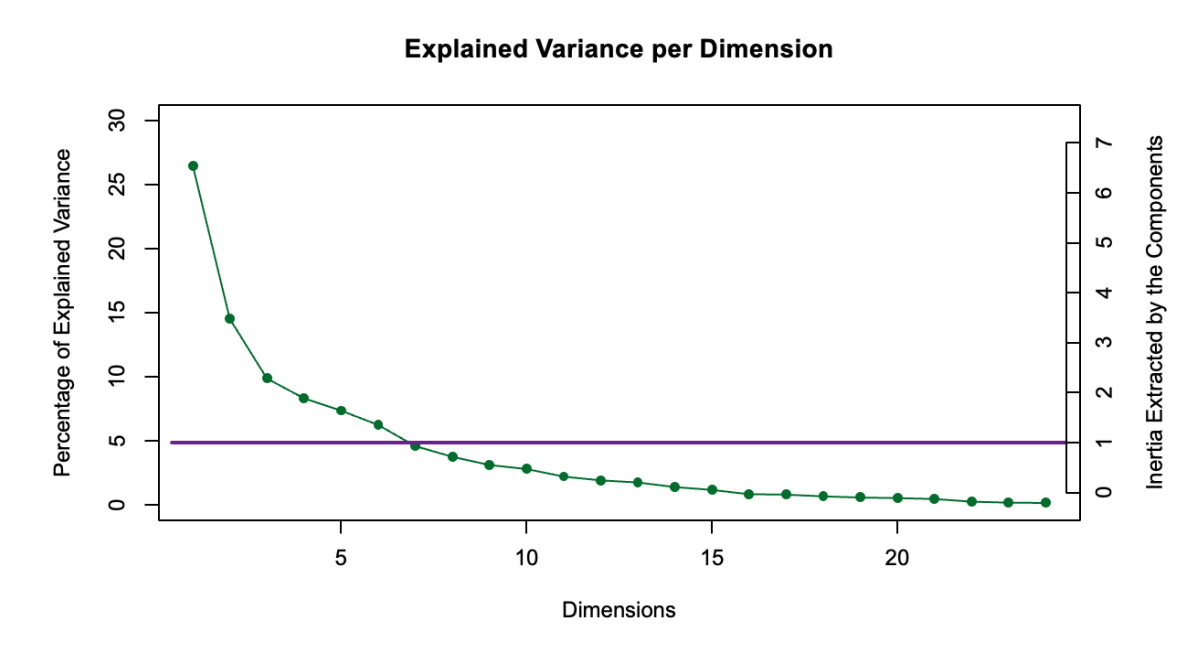
Another thing to note when looking at this graph is how many variables have high correlations with other variables. For example, take a look at the "sweet" column, and you can notice that it has about six variables that have a strong correlation with it (Pearson correlation>.5). This could mean many things. If there are many positive correlations, those correlations mean that the variables represent similar things. The negative correlations mean that they represent the same thing also, but in a different direction. For example, our Sweet and Cheesy variables have a correlation of -71, representing a strong negative correlation with each other. This correlation means sweet yogurts are not typically Cheesy and vice versa.

4.3 Explained Variance Scree plot

Calculating our eigenvalues and dividing each by the cumulative eigenvalues will give us our percentage of explained variance. We will want to have the maximum explained variance while reducing our dimensions. To do this, we will look at a **Scree Plot** that will plot our dimensions, aka eigenvalues, aka components, to their corresponding explained variance percentage.

Explained Variance Percentage This is our Eigenvalue / Cumulutive Sum of Eigenvalues which will represent exactly how much variance this eigenvalue or component brings.

Scree Plot A line plot of the cumulative eigenvalues of the Explained Variance.



Once we have calculated our components, aka dimensions, we will want to know how many components we should keep in our analysis. There are two different tests that I will be talking about that we can use here (although there are probably other ways to do this analysis!).

1. **Elbow test**

   The "elbow" is where the percentage of Explained Variance only makes marginal gains. Typically, it will be the "bend" of the elbow when you look at the graph. In this plot, we can see that the elbow starts around dimension 3. There are two ways to interpret this. The first is to keep all the components at the bend. The second option is to keep the components before the bend. If we keep all the components at the bend, then we would keep three components. If we keep before the bend, then we would keep two components.

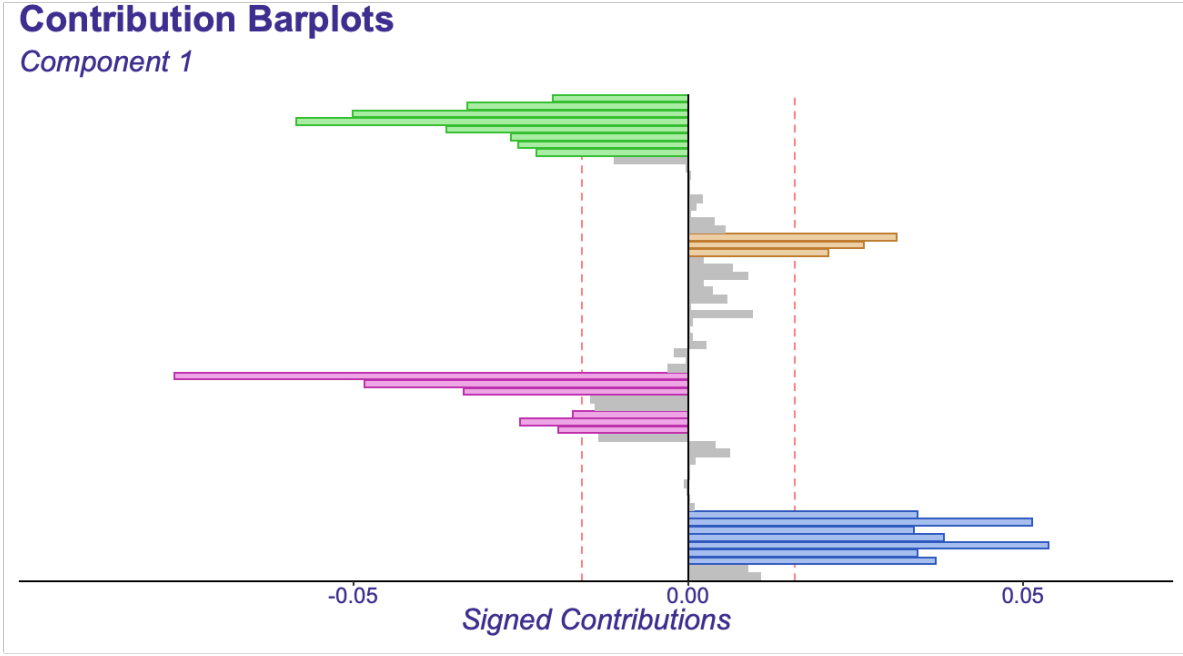
2. **Kaisar Criterion**

   Instead of using the elbow test, which often produces too few components and is more subjective, use the Kaisar Criterion, which is what the purple line above represents. The Kaisar Line is created by taking the average of all the explained variances and then plotting that line onto the graph. So, everything above the line, keep, and below the line, do not keep.

4.4 Contributions

Next, we will take a look at our contributions. For this, we have two different factors: our observations and our variables. Specific observations can make more significant contributions than other observations (same thing with the variables!), which simultaneously means they contribute a lot to the component and that they are more likely to be better represented by our components.

4.4.1 Observations



In this bar plot, each different color represents a different yogurt. For example the lime green bars represent the contribution from each observation of the yogurt Alpura Frutal. Anything over the line represents contribution that is larger then the average of the contribution and is a notable contribution. In this graph all the yo-plait Greek yogurt (in yellow) seems to make the smallest contribution to component 1.

4.4.2 Variables

When looking at contribution it is also important to look at how much each of the variables contributed to our components.

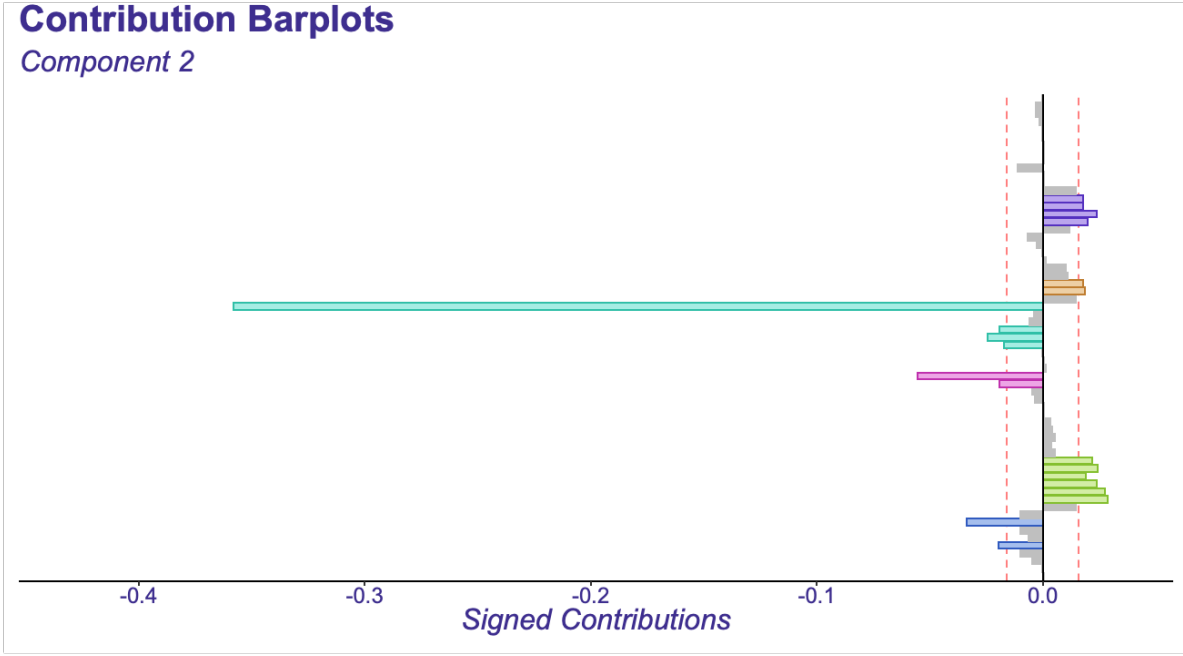
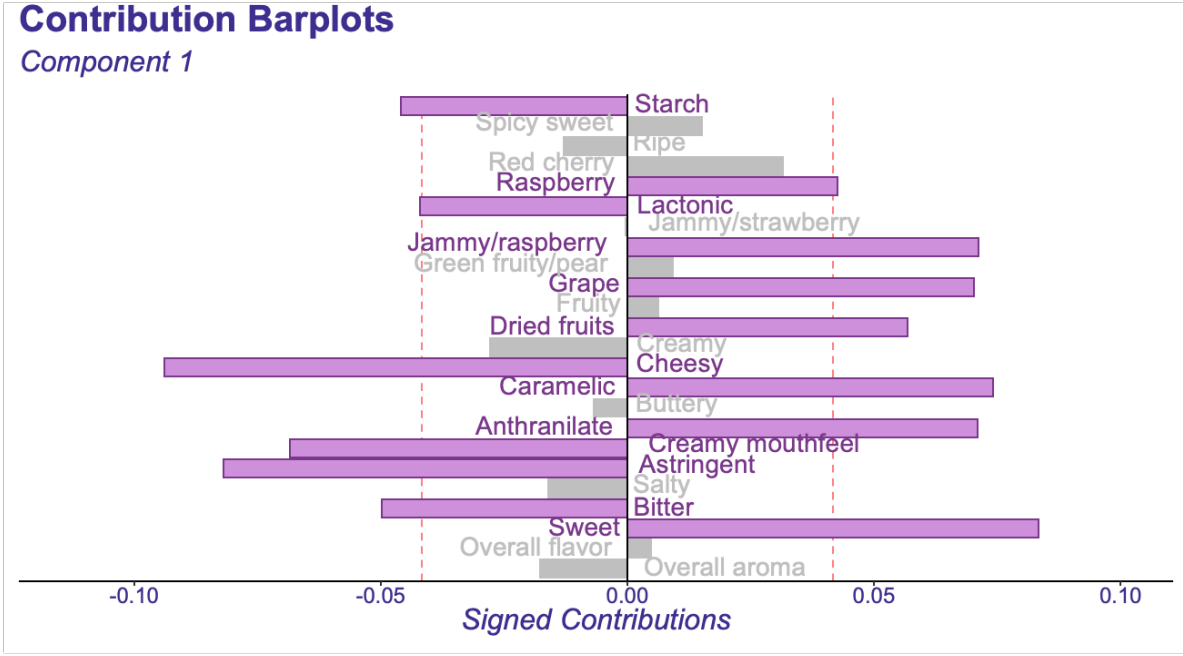
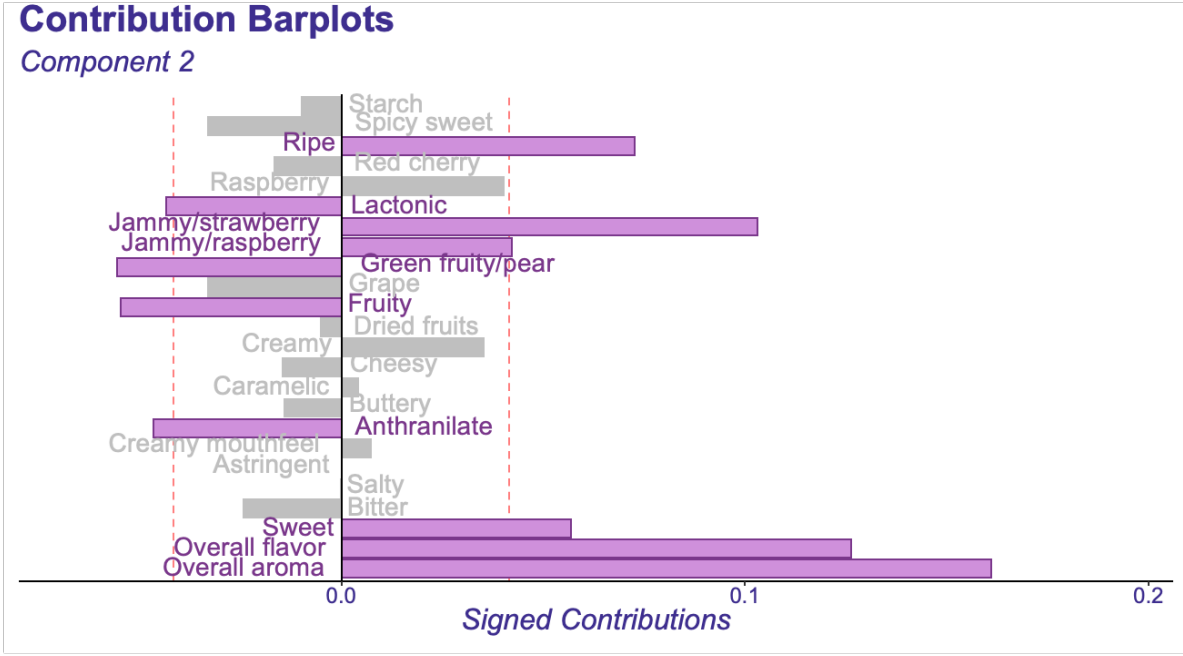


Figure 4.2: This bar plot for component 2 has one very notable observation that seems to have made a overwhelmingly large contribution to component 2.



For component 1 our largest contributor to this component was Sweet and Cheesy which also

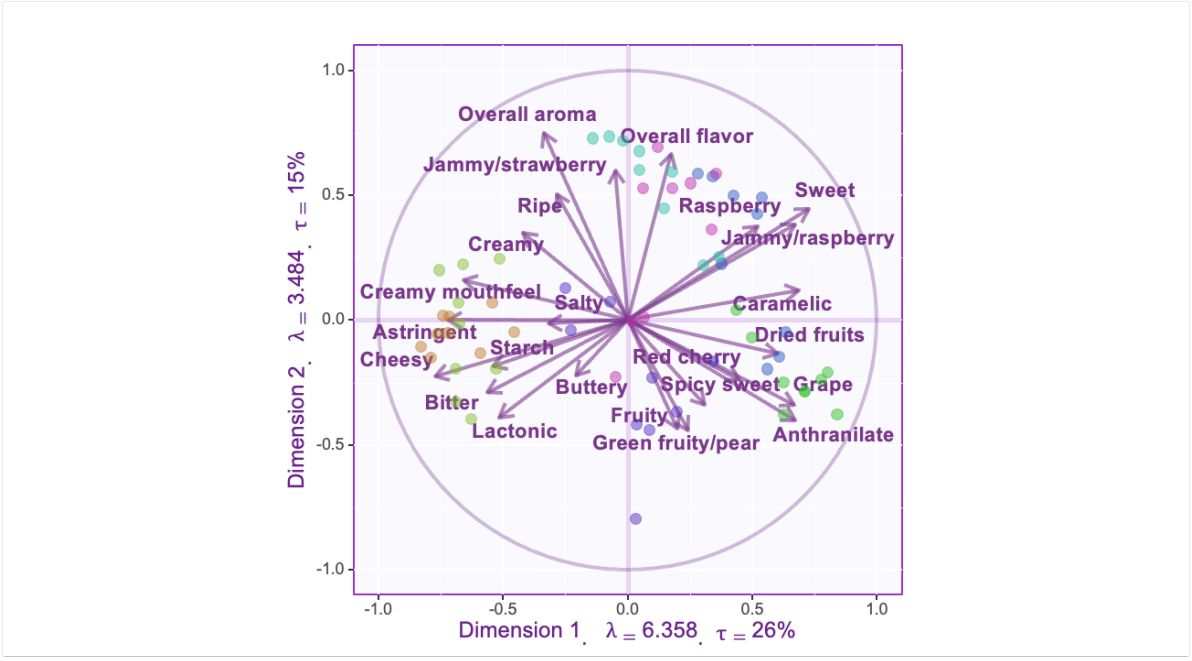
happen to be 1 of 2 of our strongest correlation. There are also other variables that have similar levels of contribution. Most of these variables are correlated with each other but it will be easier to see it on the correlation circle then here.



In this graph, Overall flavor and Overall aroma seemed to be the biggest contributors then Jammy/Strawberry and Ripe.

4.5 Correlation Circle

Lets take a look at our correlation circle and make some obervations about it.



Our correlation circle here has a few things plotted . We have our observations that is plotted using the colorful circles, and then we have ours variables plotted as well each with their corresponding correlation score to our components. For example, We can notice the lime green clusters on the right side are best represented by component one and a little by component 2. Then we can also notice within those green clusters, the variables grape and anthranalite pop up, which means that these this yogurt may be one that is more of a grape/anthranalite flavor.

Another element we have on our graph is our correlation circle itself, which will help us determine how well our observations and variables are represented by our data.

First thing I notice when I look at this correlation circle is that none of our variables touch the circle. This means that their variance is not explained perfectly by our components. We do have some components that are a bit close to the circle such as sweet and cheesy. Since we have none that are on the circle itself we have to take some precautions when making a statement about the variables correlation with each. Here are some general guidelines for the angles between two variables:

Degree of Angle	Correlation
around 180 degrees	Negative correlation
around 90 degrees	No correlation or very weak correlation
less then 90 degrees	Positive correlation

Since none of our variables touch the circle however Its important to point out that the angles may not be entirely accurate. For example since sweet and buttery are at 180 degrees you may assume that they are negatively correlated but since buttery is not represented well by either of our components (which we can tell because it is closer to the center of our graph) we cannot trust the angles. For variables like sweet and cheesy however, you can make this connection and they do seem to be negatively correlated with each , with even our correlation heat map making this connection.

One thing to note, since our components don't represent salty and buttery well this could mean they are probably better represented by a different component. Perhaps if we take a look at component three or beyond they are better represented in those.

4.6 Confidence and Tolerance Interval

4.6.1 Tolerance Intervals

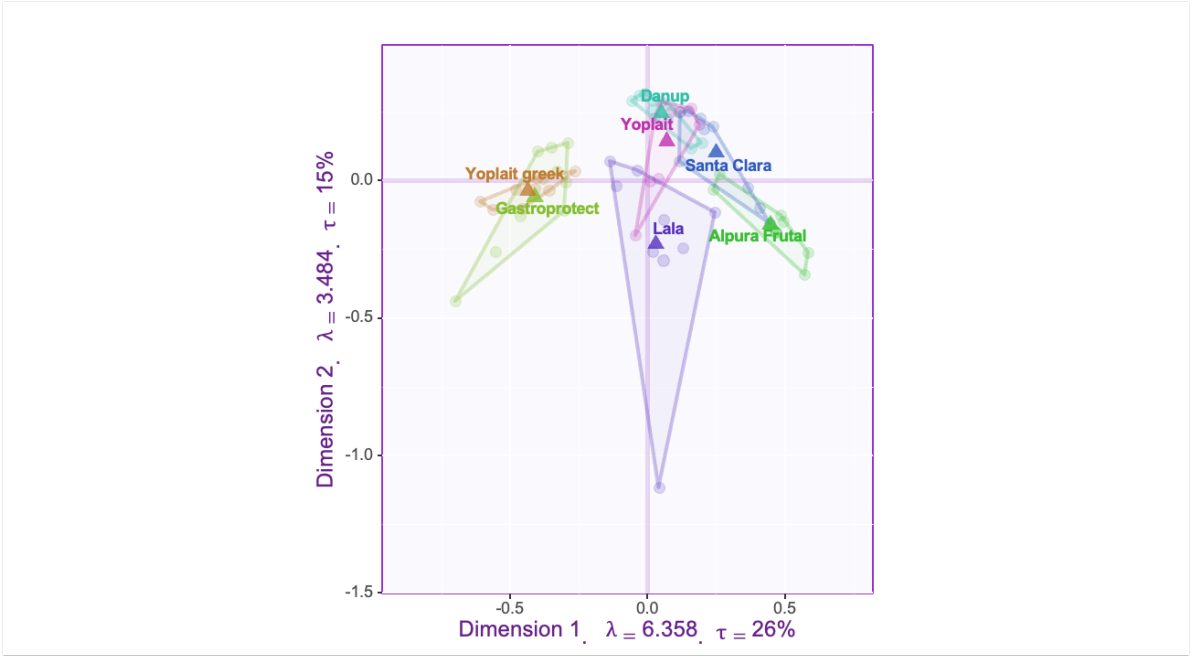
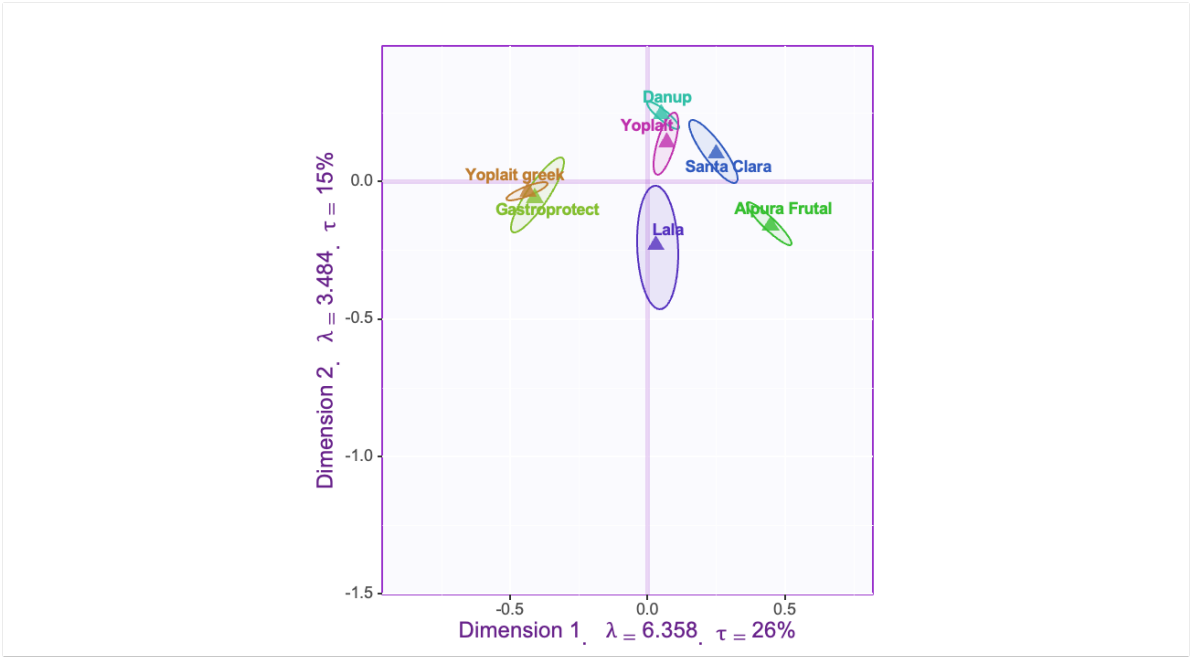


Figure 4.3: Tolerance Intervals with mean factor Scores

Our Tolerance Intervals Gives is our maximum surfuce area of the points along with our mean factor scores which are illustrated by the triangles in the middle of the areas. This graph is great for illustrating the spread of the points and also showing us how our points compare to the average of the points. For example the Lala yogurt has a outliar observation that doesn't seem to be anywhere close to the mean factor score for it.

4.6.2 Confidence Intervals



When analyzing components its important to think about the accuracy of the components. These confidence intervals represent the accuracy of our components to generalizing our yogurts. The larger the surface area of the yogurt, the less accurate we can that the component will represent for that yogurt. In this instance, we have a LaLa yogurt which seems to have a low confidence intervals. In contrast, our Danup yogurt will have a high degree of accuracy.

Part II

Correspondance Analysis

5 CA

5.1 What does Correspondence Analysis do?

- Correspondence analysis is a way of analyzing similarities and differences between different variables based on data given between a contingency table.

5.2 What is CA's goal and why?

- Its goal is to analyze relationships between variables based on relativeness.

5.3 Where is Correspondence Analysis used?

- These types of analysis are often used in brand mapping to show the relationship between the brands and other competitors and the image and demographics these brands cultivate. For example, let us look at three different soda brands, Pepsi, Coke, and Diet Coke, and among those three brands, we look at other variables, such as age range, innovativeness, and income levels. We can plot our rows: Pepsi, Coke, and Diet Coke, to see how similar they are to one another. Not only does it give us that information, but it also tells us the demographics, such as age range and inventiveness, that each brand fits into based on the closeness of that brand to each variable.
- Some questions that correspondence analysis may answer for a brand keeping in mind this is in relation to other brands.

    1. What does our brand represent?
    2. Who is attracted most to our brand?
    3. What gaps in the market can we fill?

6 About the Data

Before we go into interpreting the data it is important to define what kind of data we are using in our correspondence analysis. In this correspondence analysis we are going to be using different brands of perfumes for our rows, and attributes of the perfumes for the columns.

6.1 Our data

- 8 brands of perfume.
- 24 Different rated attributes for each perfume ex: Joyful, Classic, Cheap, Masculine. These ratings are an average of different observations.

6.2 The goal:

- To interpret the brands relationships to each other, and also to interpret the brands proximity to the different attributes.

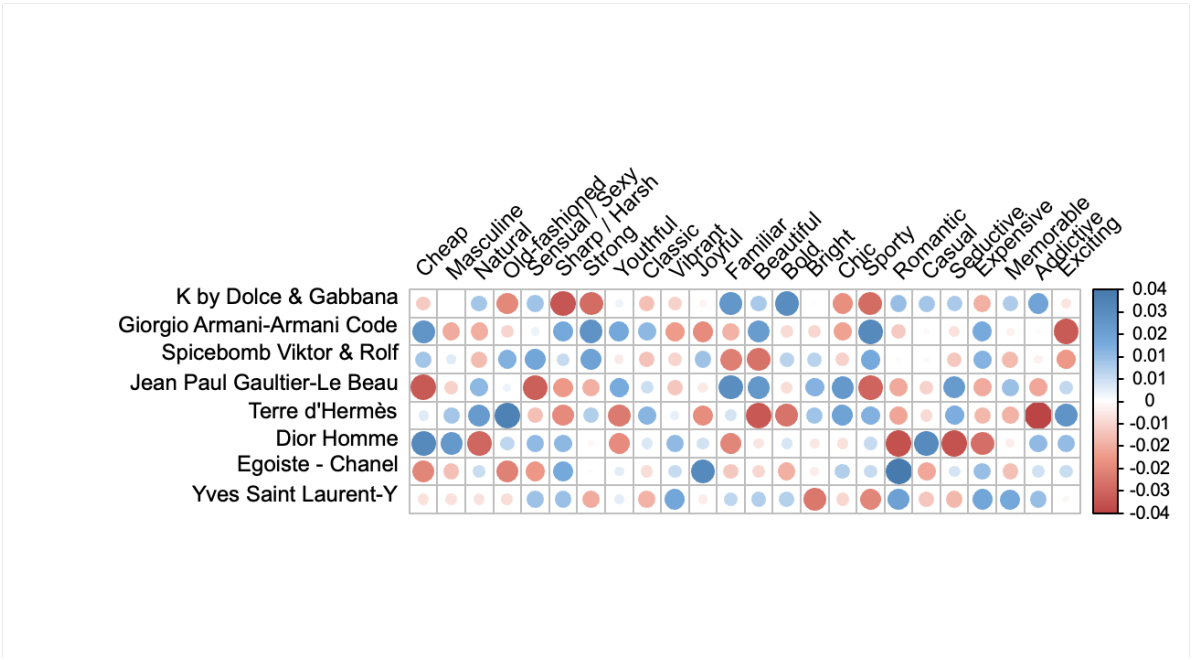
7 Interpretation of the data for Both sets

In our interpretation, we will be looking at symmetric and asymmetric results. For both the results , they have similarities in the data that we are using for it so we will take a look at those similarities first in isolation.

7.1 Chi Squared Residuals

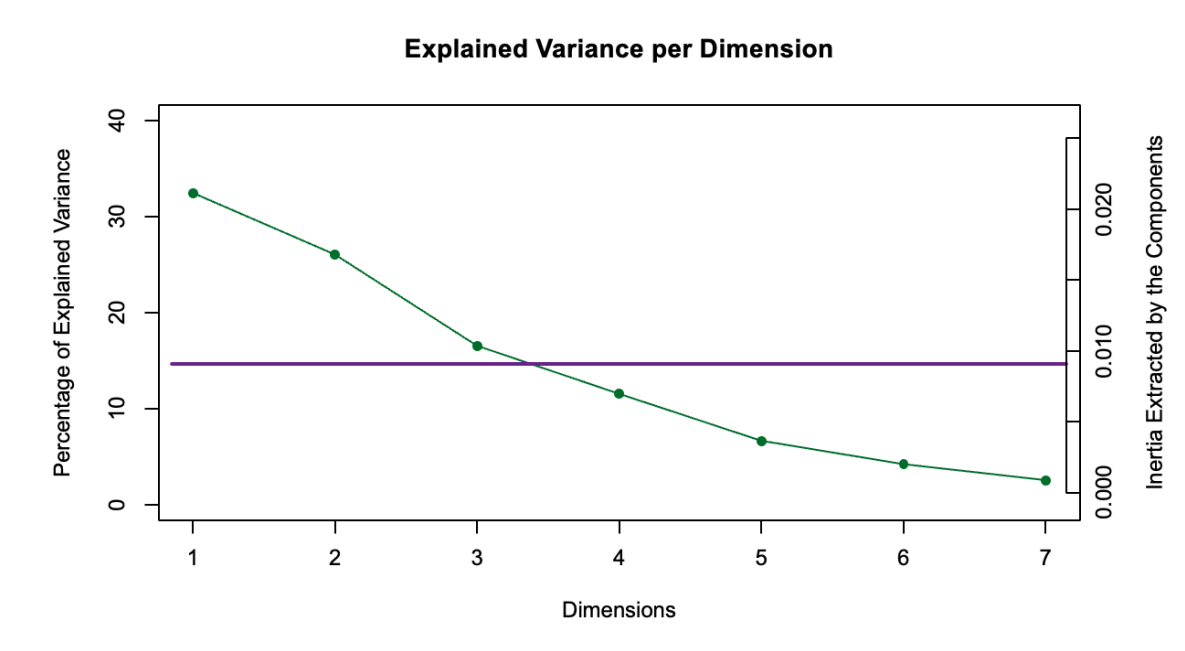
Now that we have taken a look at the data we can look at our the first thing:

Chi Squared Residuals which is going to tell us which cells contribute most to our Chi square test. Cells with the highest absolute pearson residuals will tell us which one of them is contributing most.



7.2 The Eigenvalues Scree plot

Next, we will want to take a look at the eigenvalues scree plot. In this plot, we want to look at the dimensions above the kaiser line. There are three variables about the kaiser line, so we will want to keep those three variables. One thing to note here is that there is no steep decline, and the elbow test would not work in this case.

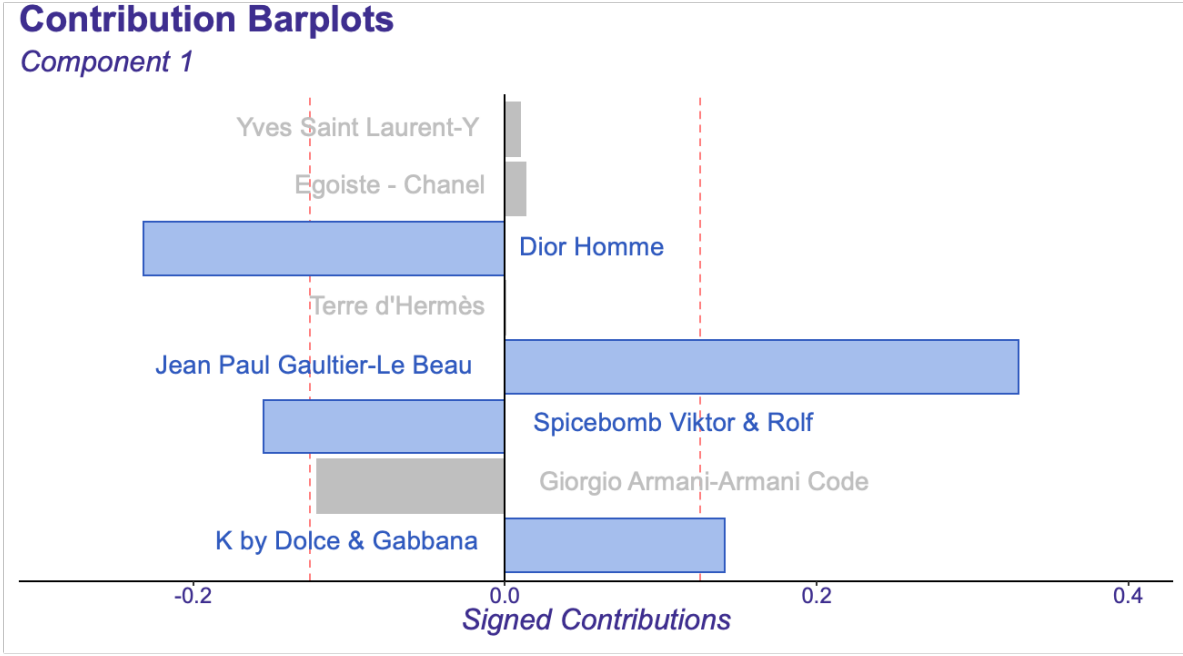


7.3 Contribution Bar plots

7.3.1 Contribution bar plot Rows:

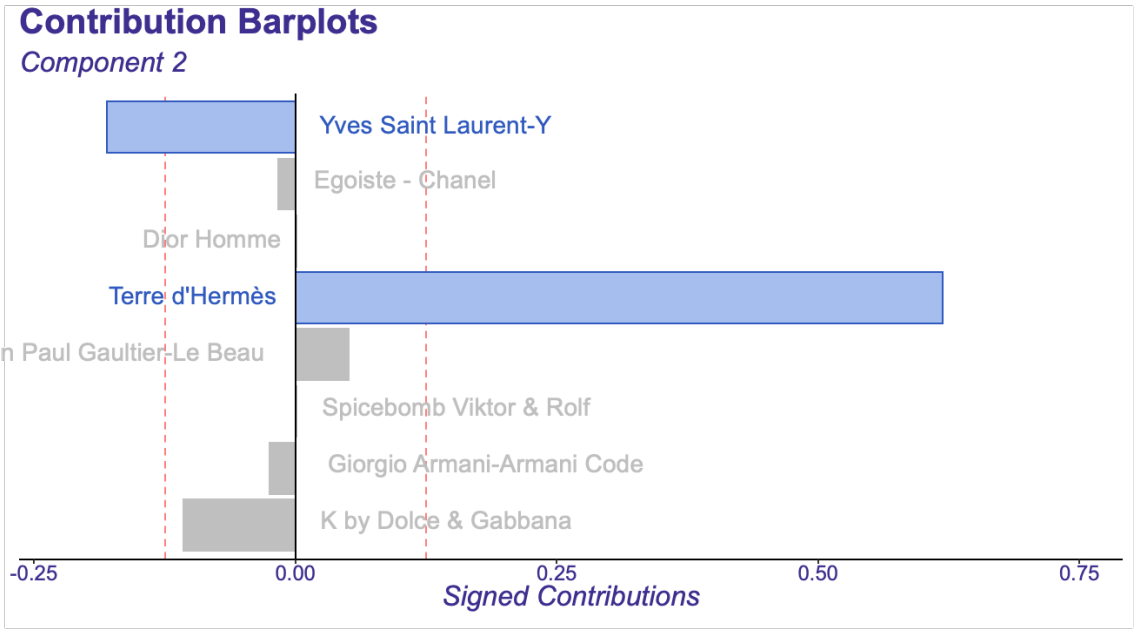
7.3.1.1 Component 1:

Here we will want to see which of the rows will contribute most to our components. Looks like 4/8 of the perfume brands contribute to component one with the largest contribution being from Spicebomb Viktor and Rolf



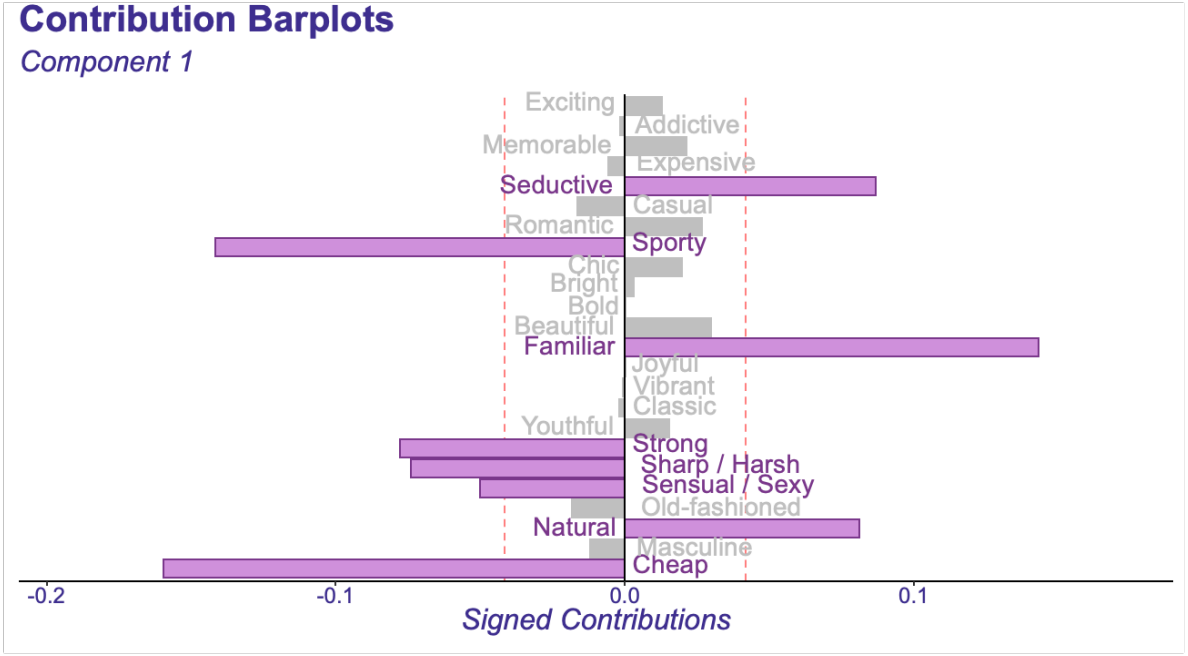
7.3.1.2 Component 2:

Below for component 2 we have two perfumes that contribute the most with Terre d'Hermes beings the largest contributor by about three times as much as the next significant contributer Yves saint Laurent-Y by far coming in at .61.



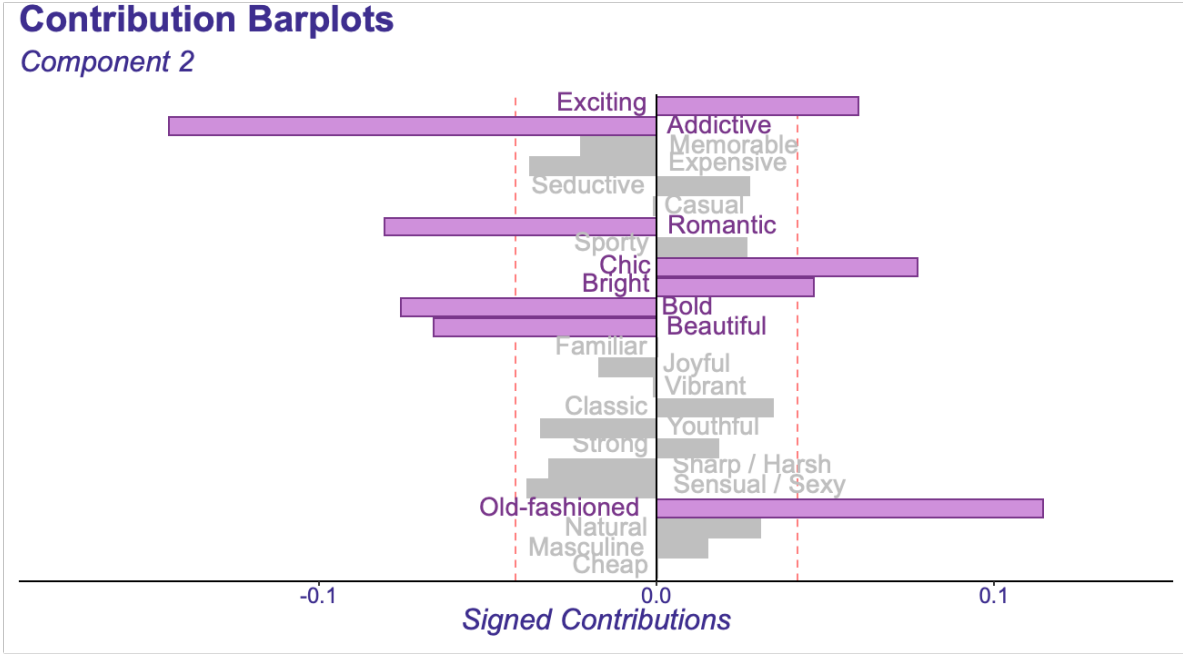
7.3.2 Contribution bar plot Columns:

7.3.2.1 Component 1



Cheap, sporty and Familiar seem to be our top 3 contributors here.

7.3.2.2 Component 2:



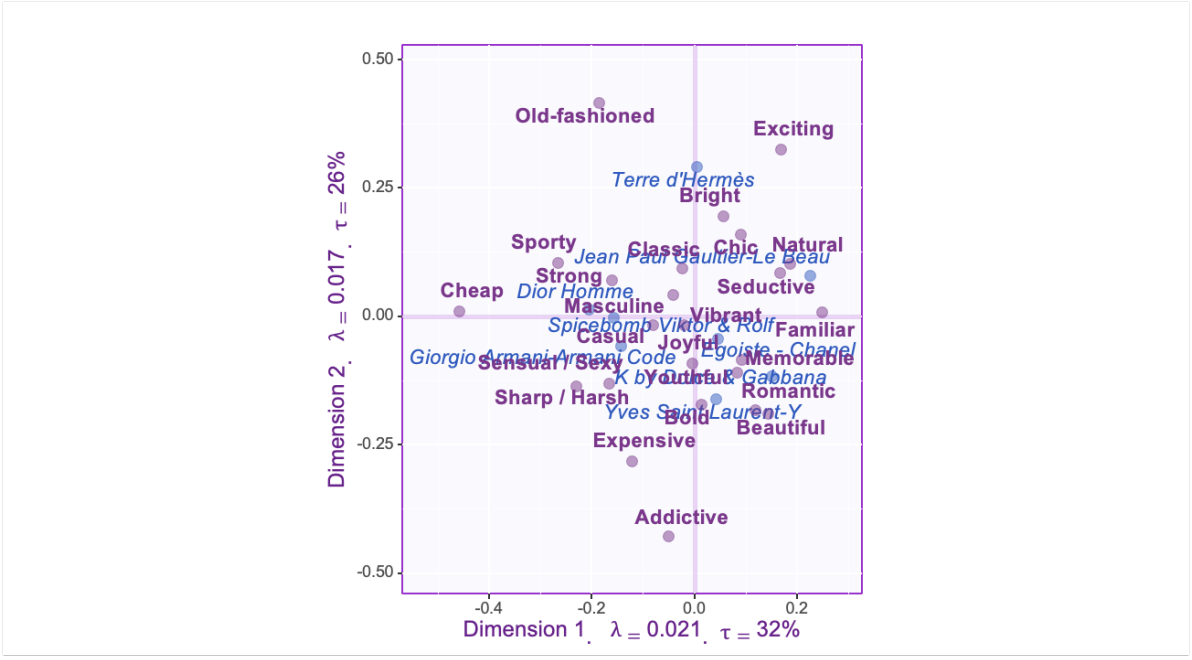
Next we will take a side by side look at our asymmetric vs symmetric results side by side.

8 Symmetric vs Asymmetric results

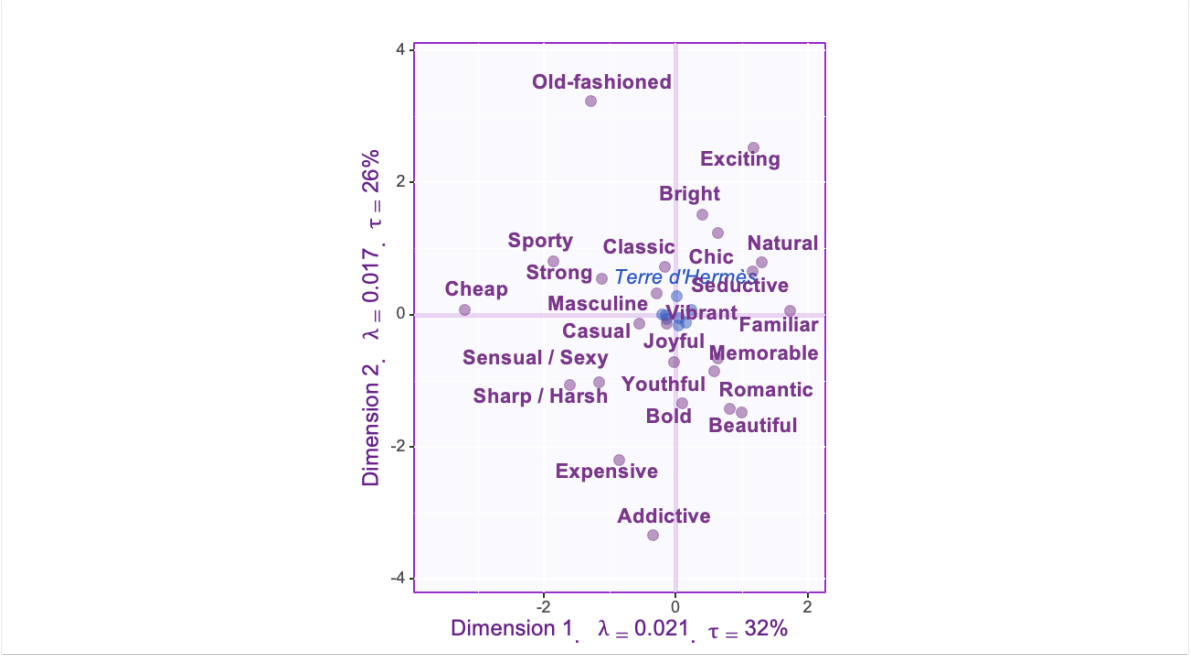
For our we need to take a look at both the symmetric and Asymmetric results. In these cases the only thing that has changed it s the column factor scores, which makes a big difference on some maps.

8.1 First i want to take a look at the factor scores:

Symmetric



Asymmetric



For the Asymmetric graph it does not look like we can trust the row data to be accurate when it comes to its proximity to any attributes, and its also a mess to look at when it comes to looking at the rows as well. The symmetric map in this case appears to be a better representation even if some of the distances are a bit exaggerated.

8.2 Final observations

So now that we have all our data lets interpret it.

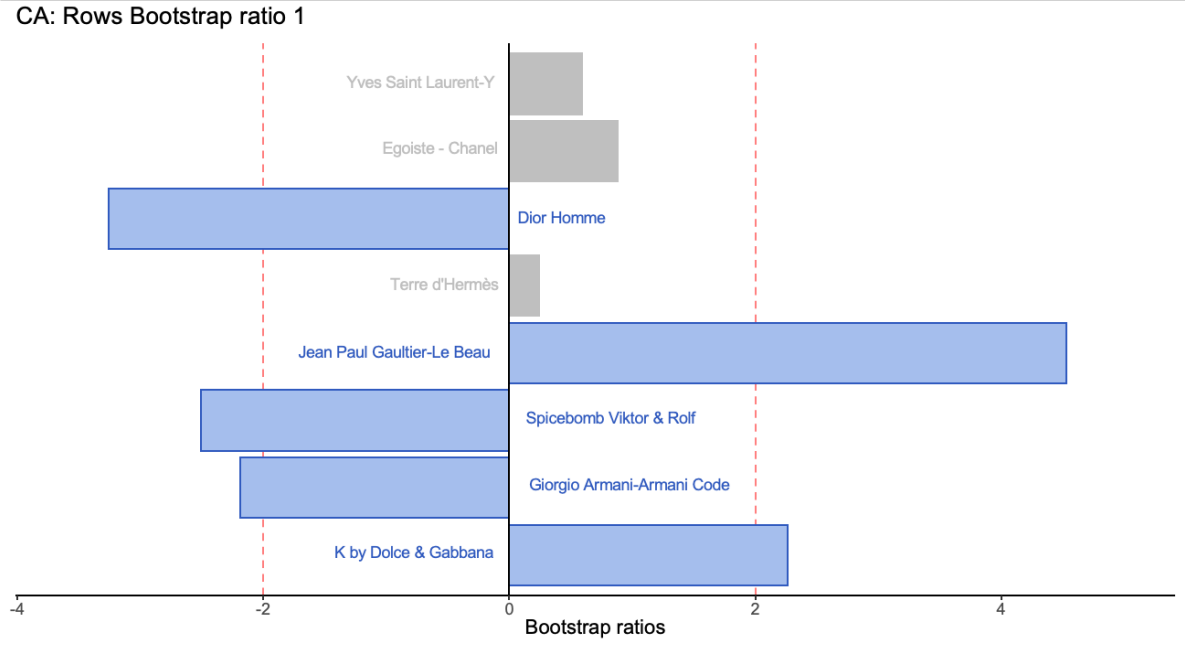
Dimension 1

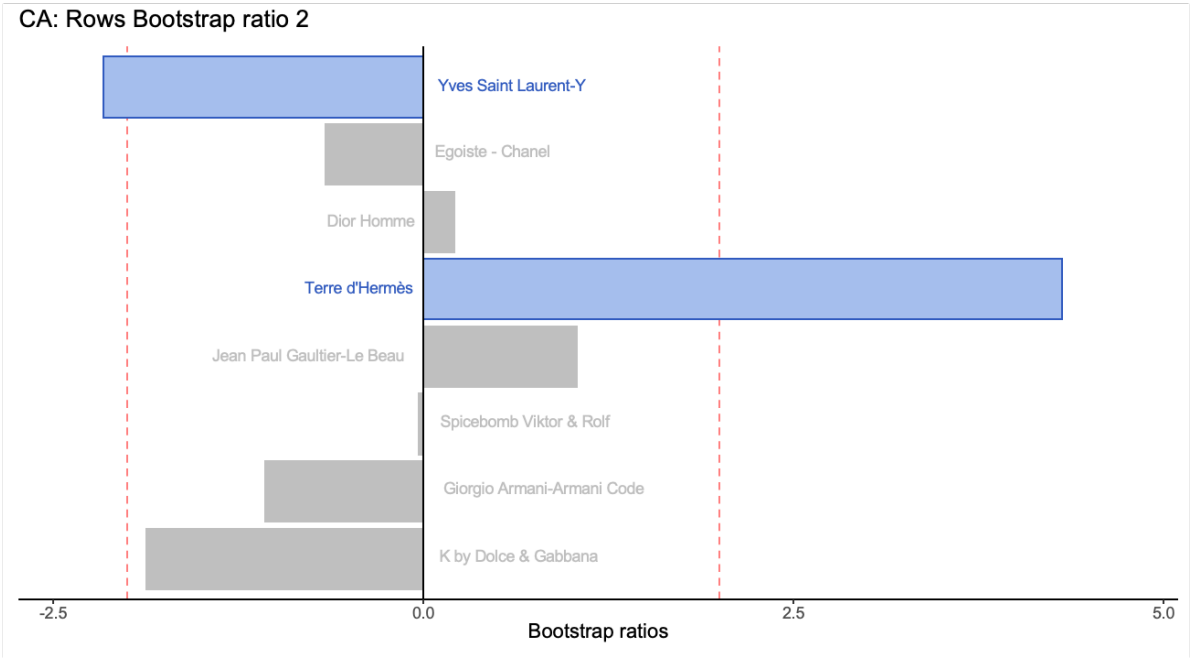
- Represents perfumes that are on one end, cheap,sharp,sporty,strong, not familiar
  - Spicebomb
  - Dior Homme
- Represents perfumes on the other hand that are natural, seductive, and familiar
  - Jean Paul
- Lastly, represents perfumes described as beautiful
  - K by dolce and cabana

- Dimension 2
  - represents perfumes that are not addictive and old fashioned
    - Terre D Hermes
  - Represents perfumes that are bold and addictive
    - Yves Saint laurent

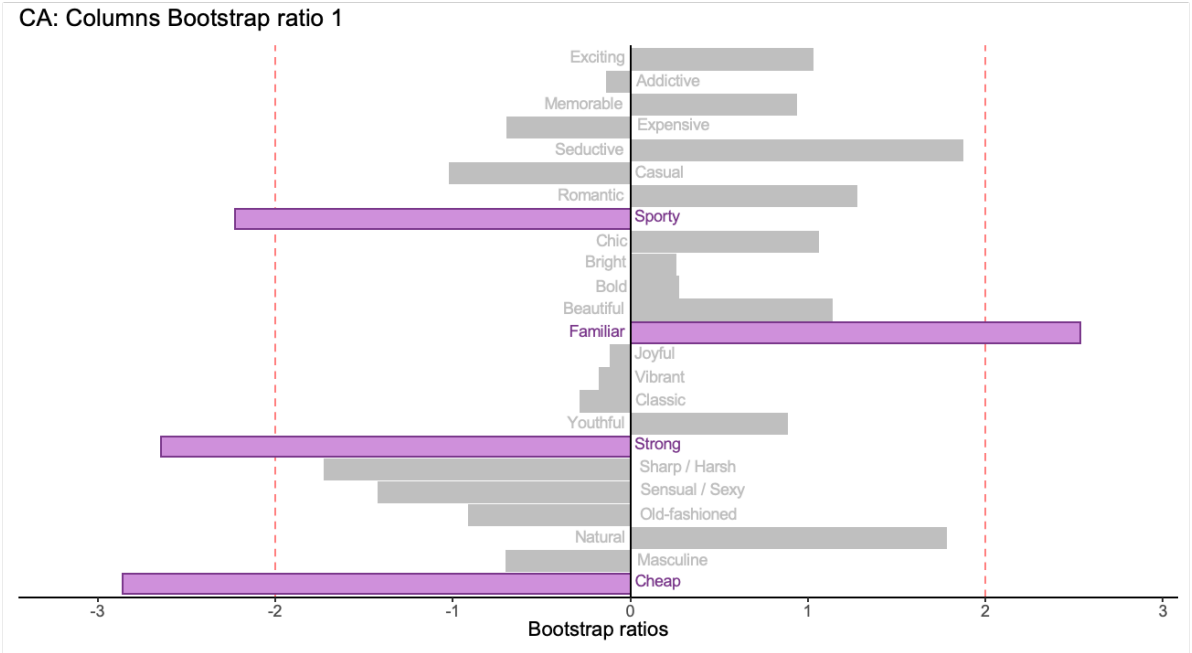
9 Inference

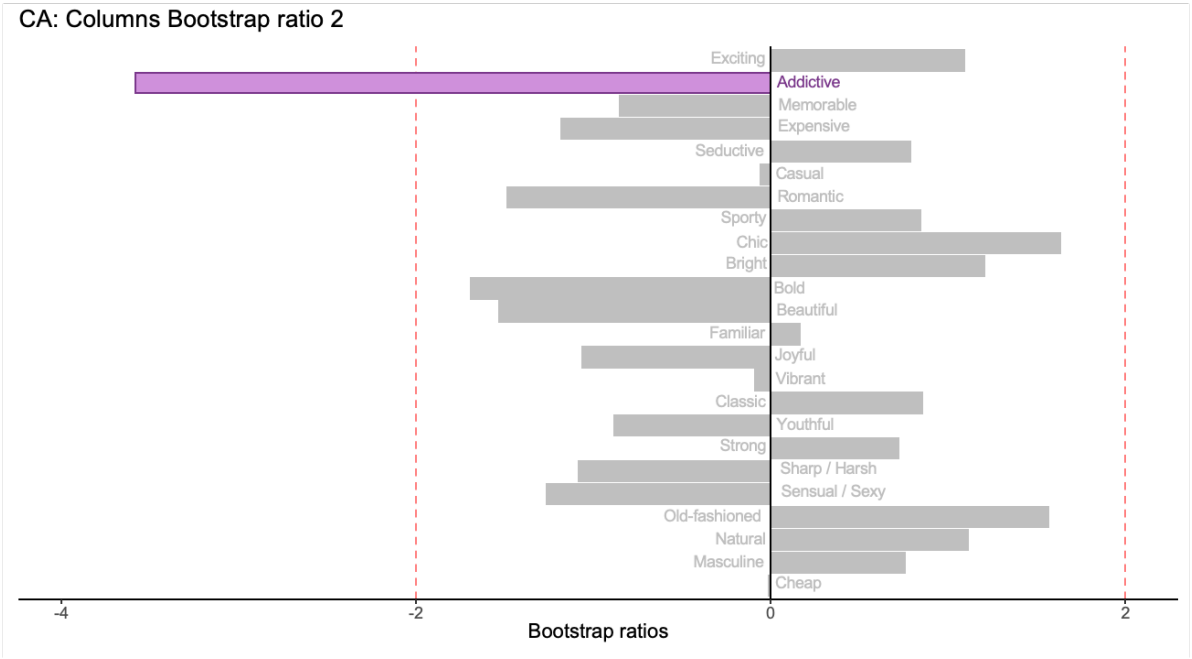
Now lets take a look at our inference





Nothing surprising here, these are the same perfumes with the same ratio we get for our contribution bar plots.





For the first dimension, 4/8 had statistical significance.

For our bootstrap ratios it appears only 1 of 8 had statistical significance four the second dimension

References