

Evaluation de la stabilité de la carte des produits obtenue par l'ACP en fonction des données considérées

Evaluation of the stability of the PCA products' maps in function of the data taken into consideration

Thierry WORCH & Pieter PUNTER

*OP&P Product Research BV
Burgmeester Reigerstraat 89
NL-3581 KP Utrecht
thierry@opp.nl*

Abstract

Although in theory, the multivariate analysis of sensory data should be done on the products' profiles (products averaged over the consumers), some experts run it on the "complete" dataset (*consumer x product x attributes*). In this peculiar case, the analysis is noised with an "undesirable" consumers' effect.

Through a real example, these two different configurations are compared, and the importance of the consumers' effect, is evaluated. Some suggestions for reducing this effect are also proposed.

Keywords: *PCA, Two sets Procrustean Analysis, Stability of the products' maps*

Résumé

Bien qu'en théorie, l'analyse multidimensionnelle de données sensorielles se fait sur les profils de produits (moyenne calculée par produit pour tous les consommateurs), certains praticiens la réalisent sur le tableau dit « complet » (croisant individus, produits et attributs). Dans ce cas bien précis, l'analyse est bruitée par un effet juge « indésirable ».

A travers un exemple réel, les différentes configurations énoncées ci-dessus sont comparées, et l'importance de l'effet juge existant dans le second cas est évaluée. Certaines pistes permettant de limiter cet effet sont également proposées.

Mots-clés : *ACP, Analyse Procrustéenne de deux configurations, Stabilité de l'espace produit*

1. Introduction

In the world of research and science, theory and practice often differ. Indeed, it is not always possible to apply in practice all the theoretical recommendations, and conversely, the theorists should sometimes close their eyes on the way practical studies are run.

The sensory world is not an exception to this rule. Indeed, in multivariate analysis of sensory data, the theorists often explain, that the Principal Component Analysis should be run on the products' profiles (*product x attributes* data, averaged over the consumers) and not on the complete dataset (*consumers x product x attribute* data). Their argument is that in the second case, the results of the PCA are noised with the consumers' variability. This consumers' effect, which is considered as a random error, is removed while working on the products' profiles, whereas it is maintained in the analysis when the total dataset is used. Nevertheless, some experts, who have an interest in consumers' variability, run PCA's on the *consumers x product x attributes* table.

The aim of this study is to compare the results obtained by PCA on the complete profiles (*consumers x products x attributes*) and on the products' profiles (*products x attributes*).

2. Materials and methods

2.1 Data

The data set used to illustrate this study concerns the test of 8 different fruit yoghurts (coded 1104, 1263, 1346, 1428, 1587, 1692, 1815 and 1971). They were all tasted by the same 130 consumers in two one-hour sessions (4 products are tasted in each session). The products were presented in a sequential monadic order, which takes care of the rank and carry over effects. The 130 consumers gave marks to both 28 intensity and ideal attributes, and 6 acceptance attributes.

Here, only the 28 intensity attributes are considered. They can be separated into the 5 following groups:

- *appearance* (gloss, color, amount fruit, recognfruit, thickness)
- *odor* (odor, fruity_ouour, sweet_o, sour_o)
- *taste* (taste, fruity, fresh fruit, sweetness, sourness, bitterness, astringent, creamy, mildness, fresh taste, off taste)
- *mouth feel* (thick_mf, smooth_mf, structure_mf, dairy_mf, firmfruit_mf, amountfruit_mf)
- *after taste* (intensity_at, length_at)

2.2 Methods

2.2.1 “Standard” multivariate analyses

In order to compare the different products’ configurations through the position of the mean products (averaged over the consumers), first three analyses are done:

- A first PCA is run on the products’ profiles. This analysis is called PRODIND, and return the table of scores \mathbf{X}^P .
- A second PCA is run on the individual judgements. The products’ profiles are projected as illustrative on the product space. This PCA is called INDPROD and return the table of scores \mathbf{X}^{SP} .
- The third products’ space calculated is the Procrustean consensus space for the 130 consumers: the best GPA-consensus space between the 130 consumers is calculated. It returns the table of scores \mathbf{X}^C .

The two first configurations are then compared together through two sets Procrustean analysis (see 2.2.3) and are then compared to the GPA consensus’ space (see 3.3).

2.2.2 “Additional” multivariate analyses

If a consumer effect is observed, different transformations can be applied on the dataset before calculation in order to minimize it. Here, three different transformations by consumer are considered; the total dataset is either centred, or standardized, or scaled.

2.2.2.1 Data centred by consumer (X^{TC})

It is known, that the consumers often don’t use the same scale. To correct it, we can centre the data by consumer; for each consumer, the mean of the attributes by product are computed. Then, for each consumer and each attribute, the mean is subtracted from the data.

Hence, for a product p , a descriptor k and a consumer n , we have:

$$y_{pk}^{(n)} = x_{pk}^{(n)} - \sum_{p=1}^P \frac{x_{pk}^{(n)}}{P}$$

The new products' profiles are computed, and the PCA on the transformed total dataset is performed. The new products' profiles are projected as illustrative, and the corresponding scores are extracted and added in the table \mathbf{X}^{TC} . The PCA is run on the correlation matrix.

2.2.2.2 Data standardized by consumer (\mathbf{X}^{TS_t})

The data can also be standardized; the data centred by consumer are divided by the standard deviation calculated for each attribute, and for each consumer, over the P products. In other words, z-scores are calculated for each consumer.

For a product p, a descriptor k and a consumer n, we have:

$$y_{pk}^{(n)} = \frac{x_{pk}^{(n)} - \sum_{p=1}^P \frac{x_{pk}^{(n)}}{P}}{\sigma_k^{(n)}}$$

We can assume that the standardization done here is sufficient. A new standardization on the whole data set is not considered as relevant. So the PCA done on the Y matrix ($Y = \cup_{n=1}^N y_{pk}^{(n)}$) is run on the covariance matrix. Again, the corresponding average table is projected as illustrative, and its scores are extracted in \mathbf{X}^{TS_t} .

2.2.2.3 Data scaled by consumer (\mathbf{X}^{TS_c})

Instead of standardizing the data by attribute and by consumer, we can compute a scale coefficient for each consumer. In this case, we give to each consumer the same sum of variance, without taking care of the variability around the attributes. Indeed, the same coefficient is applied to all the attributes of a consumer. In other words, the scale coefficient given for a consumer n is:

$$\alpha^{(n)} = \frac{\sqrt{\frac{\sum_{n=1}^N \sum_{k=1}^K \text{Var}(x_k^{(n)})}{N}}}{\sqrt{\sum_{k=1}^K \text{Var}(x_k^{(n)})}}$$

Hence, for a product p, a descriptor k and a consumer n, we have:

$$y_{pk}^{(n)} = \alpha^{(n)} \left(x_{pk}^{(n)} - \sum_{p=1}^P \frac{x_{pk}^{(n)}}{P} \right)$$

Here again, the PCA run on Y ($Y = \cup_{n=1}^N y_{pk}^{(n)}$) is done on the covariance matrix. The new products' profiles are computed and projected as illustrative on the PCA products' map, and the corresponding scores are extracted in \mathbf{X}^{TS_c} .

2.2.3 Comparison between the “standard” configurations \mathbf{X}^{P} and \mathbf{X}^{SP}

In order to compare the different configurations, the first 5 factor scores related to the 8 products of each analysis are computed (Table 1). The remaining factors are considered as noise.

\mathbf{X}^p					\mathbf{X}^{sp}				
	Product	Dim 1	Dim i	Dim 5		Product	Dim 1	Dim i	Dim 5
	1					1			
			
	p		Y_{ip}			p		Y'_{ip}	
			
	8					8			

Table 1: Factor scores of the averaged products given by \mathbf{X}^p and \mathbf{X}^{sp}

Y_{ip} : scores of the product p on the dimension i of \mathbf{X}^p (PCA on the *products x attributes*).

Y'_{ip} : scores of the product p on the dimension i of \mathbf{X}^{sp} (PCA on the *subject x products x attributes*)

A two set Procrustean Analysis is run on the scores. The coefficient of similarity is used to measure the link between both configurations. The higher this coefficient of similarity, the more similar the configurations, and the less important the consumers' effect, compared to the products' variability. As the five dimensions extracted from each analysis are not necessarily relevant, the comparison is also done between the three-dimensional configurations (only the three first factors are then used).

2.2.4 Comparison of the “additional” configurations (\mathbf{X}^{TC} , \mathbf{X}^{Tst} , \mathbf{X}^{TSc}) with the “standard”

Here again, the factor scores computed for each additional analysis are submitted to a two-set Procrustean Analysis in order to compare the configuration \mathbf{X}^{TC} (centered), \mathbf{X}^{Tst} (standardized) and \mathbf{X}^{TSc} (scaled) with the \mathbf{X}^p and the \mathbf{X}^{sp} configurations. Again, the criterion used is the coefficient of similarity. This procedure has two aims:

- to see if a part of the consumers' effect can be removed
- if yes, to compare the different transformations together in order to see which one has the best results.

2.2.5 Stability of the results when a strong consumers' effect does exist

In a discussion in the sensory egroups in October 2005 (sensory@yahoogroups.com) concerning the “PCA on raw data?” topic, Harry Lawless stated the problem that, when PCA's are done on the complete dataset, “disagreements of judges [can] feed the correlations”. He stated:

“Consider a one-product PCA done where there are only judges. I score the product high on scale X and Y; you score it low on X and Y. This disagreement then creates a correlation. Most of us would call that error variance. Why base your PCA on error?”

To estimate this effect, a modification of some existing consumers is done, in such a way that pairs of “new consumers” hardly disagree without changing the global products' profiles.

In order to measure the stability of the results when a structure in the consumers through their disagreements does exist, many simulations have been done. The factors taken in consideration in these simulations are:

- the size of the modified panel (going from 10 to 30)
- the degree of disagreement between two “new consumers” (going from -0.5 to -0.95)
- the correlation between the pairs of modified consumers with the other (going from 0 to 1)
- the correlation between the pairs of modified consumers with the rest of the panels (going from 0 to 0.3)

The new configuration calculated on the total dataset is compared to \mathbf{X}^p and to \mathbf{X}^{sp} .

3. Results

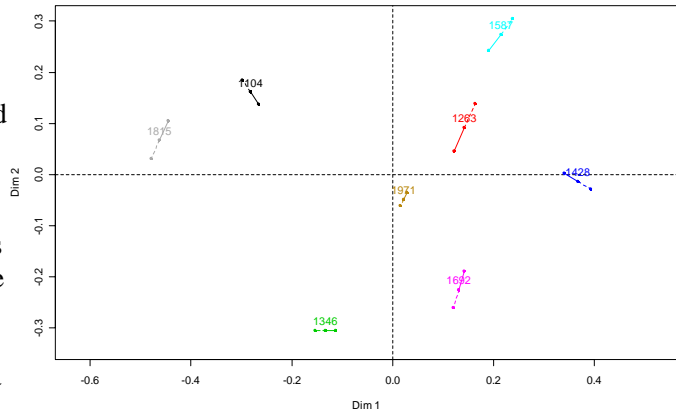
3.1 Comparison of the configurations X^P and X^{SP}

The X^P and X^{SP} configurations are given in Appendix 1. The two sets Procrustean Analysis run on the X^P and X^{SP} configurations show the following results (*Graph 1*):

The value of the coefficient of similarity between both configurations, when three (resp. five) dimensions are kept, is 0.87 (resp. 0.93). These values show that the configurations X^P and X^{SP} are not identical, but very close.

More precisely, the high values of the coefficient show that the main variability of this dataset is related to the products, and not to the consumers' variability.

But since both configurations are not identical, a consumers' effect does exist.



Graph 1: Two-set Procrustean Analysis consensus' space calculated between X^P and X^{SP} (5 dimensions are for interest in each case)

3.2 Comparison of the “standard” and the “additional” configurations

After applying the three different transformations (centering, standardizing, scaling), the PCA's are done, and the factor scores associated to the average products are computed. Like previously, these scores are submitted to a two-set Procrustean Analysis in order to compare these new configurations with the X^P and X^{SP} ones. The *Table 2* summarizes the results:

	Three dimensions		Five dimensions	
	PRODIND X^P	INDPROD X^{SP}	PRODIND X^P	INDPROD X^{SP}
PRODIND X^P	1.00	0.87	1.00	0.93
INDPROD X^{SP}	0.87	1.00	0.93	1.00
Centered X^{TC}	0.96	0.94	0.96	0.95
Standardized X^{TSt}	0.95	0.92	0.98	0.94
Scaled X^{TSc}	0.96	0.92	0.95	0.93

Table 2: Coefficient of similarity between the different configurations in function of the number of dimensions taken in consideration

Whatever the number of dimensions kept, the transformation of the data gives a higher coefficient of similarity with the X^P configuration than the X^{SP} configuration (all coefficients after transformation are superior to 0.87). In other words, the transformations do minimize the consumers' effect present in the data.

It also appears that the configurations after transformation are always closer to the X^P than to the X^{SP} configuration.

Moreover, the high values of the coefficients of similarity show that the most part of the variability present in the dataset is due to the products' variability.

Finally, no conclusion can be drawn on the most successful type of transformation: the three seem to have the same impact on the reduction of the consumers' effect.

3.3 Comparison of the PCA configurations with the GPA consensus' space

In order to compare the GPA consensus configuration (given in *Appendix 2*) with the configurations given by the PCA on the different transformed data, two-set Procrustean analyses have been run between the \mathbf{X}^C and \mathbf{X}^P , \mathbf{X}^{sp} , \mathbf{X}^{TC} , \mathbf{X}^{Tst} and \mathbf{X}^{Tsc} . The results are given *Table 3*, in function of the number of dimensions of interest.

	Three dimensions	Five dimensions
	GPA \mathbf{X}^C	GPA \mathbf{X}^C
PRODIND \mathbf{X}^P	0.92	0.89
INDPROD \mathbf{X}^{sp}	0.75	0.79
Centered \mathbf{X}^{TC}	0.87	0.82
Standardized \mathbf{X}^{Tst}	0.88	0.86
Scaled \mathbf{X}^{Tsc}	0.88	0.85

Table 3: Coefficient of similarity between the consensus configuration and the different PCA configurations in function of the number of dimensions

We can see an evolution of the coefficient of similarity in the opposite direction of the consumer's effect. Hence, the GPA removes a part of the systematic error due to the consumers' effect.

Indeed, the value of the coefficient of similarity increases while the consumers' effect is reduced. It is growing from 0.75 (comparison between \mathbf{X}^C and \mathbf{X}^{sp} ; consumers' effect is maximal) to 0.92 (comparison between \mathbf{X}^C and \mathbf{X}^P ; the consumers' effect is removed) via a coefficient of 0.87 (comparison between \mathbf{X}^C and the \mathbf{X} transformed; the consumers' effect is reduced) when looking at the first three dimensions. When the first five dimensions are taken, another small evolution is observed within the transformations. It appears here that centering the data removes less consumers' effect than scaling or standardizing.

It also shows that going from three to five dimensions reduces the coefficient of similarity. This is maybe due to the fact that the factors added in the consensus space bring some additional information not present in the PCA's.

3.4 Stability of the results when a strong consumers' effect does exist

Whatever the structure given to the consumers (strong or weak link), it appears that the presence of disagreements between consumers doesn't affect the PCA results. The configurations of the products space run on the total modified-data is always close to \mathbf{X}^P ; indeed, the coefficient of similarity vary between 0.85 and 0.90.

These results are maybe due to the fact that the size of the modified panel (maximum 30 consumers) is always low compared to the rest of the panel (minimum $130-30 = 100$ consumers). But it has to be said that the simulations taken in consideration are extreme in the fact that in practice, we will hardly find consumers, which are that much in disagreement.

4. Conclusions

In this case study, the high similarity between the \mathbf{X}^P and \mathbf{X}^{SP} configurations suggests that the part of variability related to the consumers is always small compared to the one related to the products, even if we create a structure between consumers. Nevertheless, a consumers' effect does exist, and some transformations can reduce it.

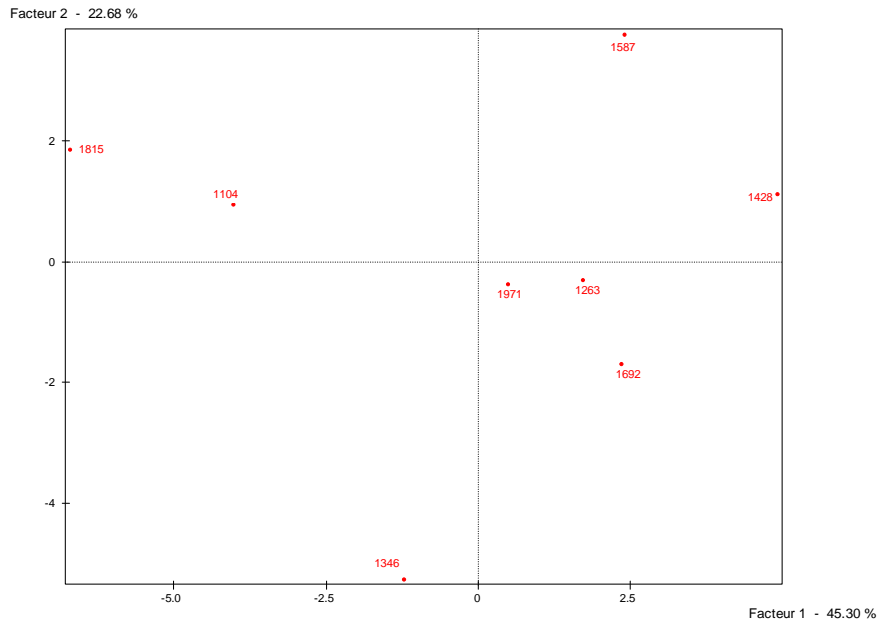
One of the main differences not mentioned in this study, between the two configurations \mathbf{X}^P and \mathbf{X}^{SP} lies in the number of relevant dimensions of the products' spaces. For \mathbf{X}^P , only 2 or 3 dimensions are usually extracted for the interpretation of the results, whereas for \mathbf{X}^{SP} , as each dimension explains less variability, more (5 to 8 usually) are extracted.

Hence, the projection of the individual judgments on the PCA products' space obtained with the products' profiles compresses the consumers' variability. As a more elaborate segmentation of the attributes is sometimes used (in practical situations, the overall liking is regressed on the individual factor scores in order to define the main driver's of liking), this compression is not necessarily beneficial.

In practice, the use of the GPA instead of the PCA on the total dataset can be a good alternative, as it removes a part of the systematic error in the consumers' effect. But in this case, compared to the PCA, the number and the interpretation of the relevant factors could be more problematic...

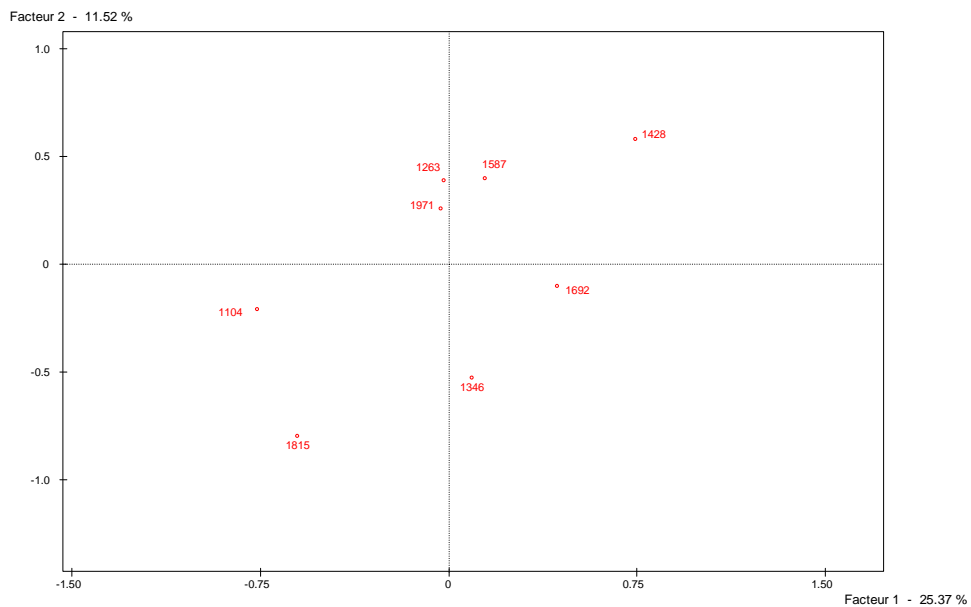
Appendix 1: X^P and X^{SP} configurations

PRODIND:



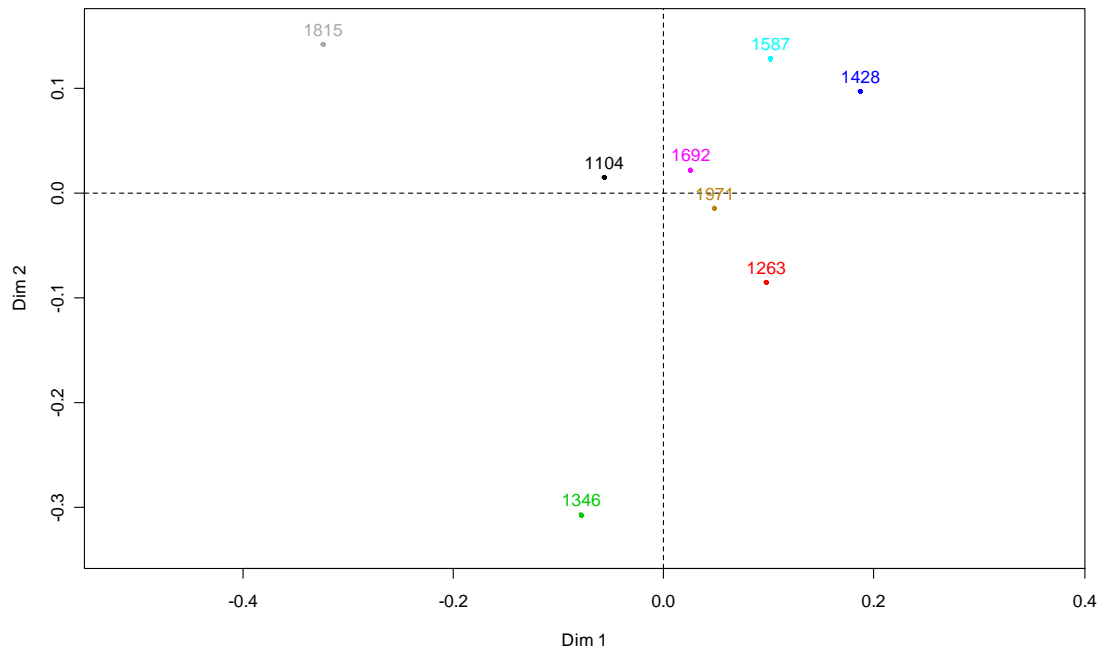
Graph 1.1: *PRODIND* products space (X^P configurations)

INDPROD:

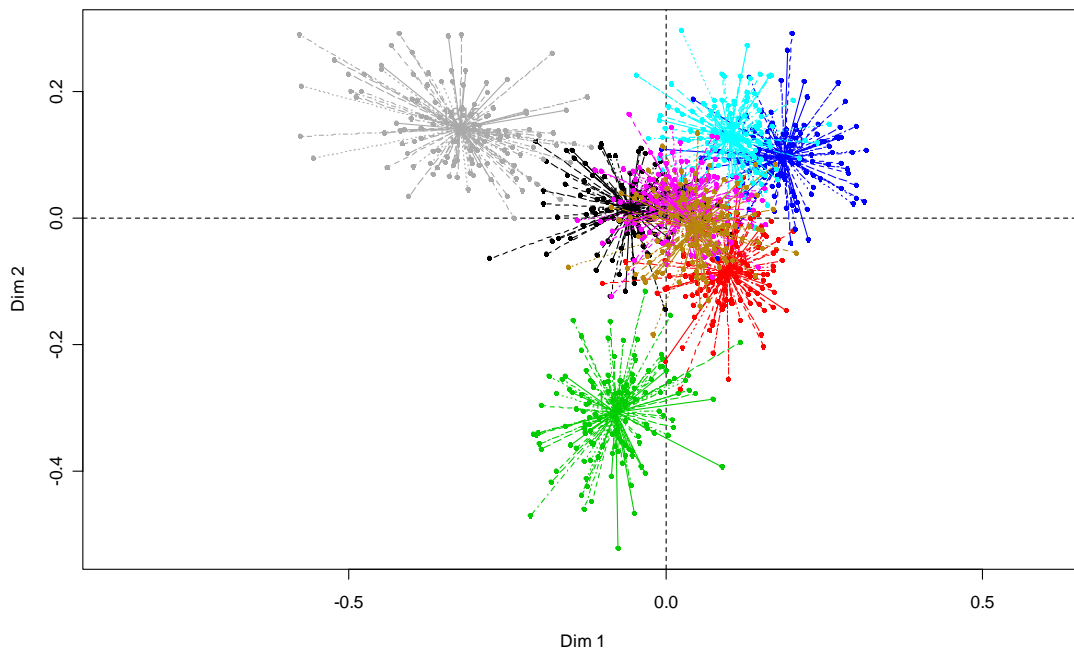


Graph 1.2: *INDPROD* products space (X^{SP} configurations)

Appendix 2: GPA (X^C configuration)



Graph 2.1: GPA consensus' space for the 130 consumers (X^C configuration)



Graph 2.2: GPA consensus' space for the 130 consumers (X^C configuration) with the individual judgements represented (please note that the scales are changed)