Universidad del Valle
Programa Académico de Estadística
Estadística Aplicada III
Cali, febrero de 2019

Laboratorio 3:Análisis Factorial Múltiple (AFM)

Kevin Steven García[a], Alejandro Vargas[b], Alejandro Soto[c], Natalia Buitron[d]

1. Introducción

En el presente informe veremos la aplicación del AFM a la base de datos data(orange) de la librería missMDA, la cual corresponde a la descripción sensorial de 12 jugos de naranja por 8 atributos, esta base presenta cerca del 20 % de datos faltantes, por lo cuál se hará primero un proceso de imputación para posteriormente realizar el método AFM. Se analizará e interpretará el porcentaje de Inercia explicado, la nube de individuos,la nube de variables, la nube de los grupos, los coeficientes Lg y Rv de Escoufier y se realizará el gráfico de representación Superpuesta y de los ejes parciales, todo esto luego de un debido análisis descriptivo de las variables del estudio.

La base de datos es la siguiente:

	Color.intensity	Odor.intensity	Attack.intensity	Sweet	Acid	Bitter	Pulp	Typicity
1	4.791667	5.291667	NA	NA	NA	2.833333	NA	5.208333
2	4.583333	6.041667	4.416667	5.458333	4.125000	3.541667	4.625000	4.458333
3	4.708333	5.333333	NA	NA	4.291667	3.166667	6.250000	5.166667
4	6.583333	6.000000	7.416667	4.166667	6.750000	NA	1.416667	3.416667
5	NA	6.166667	5.333333	4.083333	NA	4.375000	3.416667	4.416667
6	6.333333	5.000000	5.375000	5.000000	5.500000	3.625000	4.208333	4.875000
7	4.291667	4.916667	5.291667	5.541667	5.250000	NA	1.291667	4.333333
8	NA	4.541667	4.833333	NA	4.958333	2.916667	1.541667	3.958333
9	4.416667	NA	5.166667	4.625000	5.041667	3.666667	1.541667	3.958333
10	4.541667	4.291667	NA	5.791667	4.375000	NA	NA	5.000000
11	4.083333	5.125000	3.916667	NA	NA	NA	7.333333	5.250000
12	6.500000	5.875000	6.125000	4.875000	5.291667	4.166667	1.500000	3.500000

2. Análisis descriptivo

- Definición de variables:
  - Intensidad del color: Cuantitativa continua. Escala de intervalos.
  - Intensidad del olor: Cuantitativa continua. Escala de intervalos.
  - Intensidad del ataque(sensación inicial del jugo en la boca): Cuantitativa continua. Escala de intervalos.
  - Dulce: Cuantitativa continua. Escala de intervalos.
  - Ácido: Cuantitativa continua. Escala de intervalos.
  - Amargo: Cuantitativa continua. Escala de intervalos.
  - Pulpa: Cuantitativa continua. Escala de intervalos.
  - Tipicidad: Cuantitativa continua. Escala de intervalos.

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1525953. E-mail: jose.alejandro.vargas@correounivalle.edu.co
[c]Código: 1532457. E-mail: asotomurillo@gmail.com
[d]Código: 1526135. E-mail: osorio.natalia@correounivalle.edu.co

- Resumen Estadístico:

	Color.intensity	Odor.intensity	Attack.intensity	Sweet	Acid	Bitter	Pulp	Typicity
Mínimo	4.083	4.292	3.917	4.083	4.125	2.833	1.292	3.417
Cuartil 1	4.448	4.958	4.833	4.510	4.375	3.104	1.510	3.958
Mediana	4.646	5.292	5.292	4.938	5.042	3.583	2.479	4.438
Media	5.083	5.326	5.319	4.943	5.065	3.536	3.312	4.462
Cuartil 3	5.948	5.938	5.375	5.479	5.292	3.792	4.521	5.042
Máximo	6.583	6.167	7.417	5.792	6.750	4.375	7.333	5.250
Des. Estándar	0.9809135	0.630481	1.005842	0.6318818	0.7971605	0.5529671	2.22528	0.6521764
NA's	2	1	3	4	3	4	2	0

En el resumen estadístico, se puede ver como las variables en rasgos generales son muy parecidas, las variables intensidad del color, intensidad del olor, intensidad del ataque, dulce y ácido poseen una media bastante parecida, al igual que el rango en el que se están moviendo estas variables, vemos como el rango más amplio lo tiene la variable pulpa con un rango de 6,041 lo cual esta sustentado también en que posee la desviación estándar (2.22528) más grande de todas las variables. Las variables dulce y amargo son las que mayor cantidad de datos faltantes tienen, con 4 datos faltantes cada una y la que menos datos faltantes tiene es la variable tipicidad, la cual se encuentra completa (0 datos faltantes).

- Matriz de correlaciones:
A partir de la matriz de datos imputada que se mostrará posteriormente, se obtuvo la siguiente matriz de correlaciones.

	Color.intensity	Odor.intensity	Attack.intensity	Sweet	Acid	Bitter	Pulp	Typicity
Color.intensity	1.0000000	0.61111536	0.7441370	-0.7630744	0.7183662	0.8327382	-0.29700410	-0.4570463
Odor.intensity	0.6111154	1.00000000	0.4143692	-0.5650579	0.2786459	0.7312423	0.04552839	-0.3173757
Attack.intensity	0.7441370	0.41436921	1.0000000	-0.7829896	0.9584735	0.8725642	-0.73475086	-0.8447980
Sweet	-0.7630744	-0.56505792	-0.7829896	1.0000000	-0.7904037	-0.9269297	0.50992903	0.6527152
Acid	0.7183662	0.27864586	0.9584735	-0.7904037	1.0000000	0.8149900	-0.74605657	-0.7572032
Bitter	0.8327382	0.73124233	0.8725642	-0.9269297	0.8149900	1.0000000	-0.50555139	-0.7165638
Pulp	-0.2970041	0.04552839	-0.7347509	0.5099290	-0.7460566	-0.5055514	1.00000000	0.8739956
Typicity	-0.4570463	-0.31737571	-0.8447980	0.6527152	-0.7572032	-0.7165638	0.87399562	1.0000000

Podemos ver que la variable Bitter(Amargo) esta altamente correlacionada con la mayoría de las otras variables, positivamente con las variables intensidad del color(0.8327), intensidad del olor(0.7312), intensidad del ataque(0.8725) y ácido(0.8149), y negativamente con las variables dulce(-0.9269), pulpa(-0.5) y tipicidad(-0.7165). También se observa que la variable intensidad del ataque esta altamente correlacionada con la variable ácido, con una correlación casi perfecta(0.9584), entre otras correlaciones importantes que se observan en la matriz.

3. Análisis Factorial Múltiple

Para realizar el AFM, debemos dividir la tabla en subtablas dependiendo de los tipos de variables y su definición. En este caso, dado que todas las variables son cuantitativas continuas, nuestro criterio de división se baso en la definición de las variables. Decidimos dividir la tabla en dos subtablas, la primera consta de las tres primeras variables(columnas), Intensidad del color, intensidad del olor e intensidad del ataque, esta primera tabla fue denominada percepción previa, ya que las tres variables se miden antes de degustar el jugo. La segunda subtabla consta de las cinco últimas variables, dulce, acido, amargo, pulpa y tipicidad, esta segunda tabla fue denominada percepción posterior, ya que todas estas variables involucradas son medibles solamente después de catar o degustar el jugo.

3.1. Imputación AFM

Dado que la base de datos tiene valores faltantes en una cantidad considerable(20 % de valores faltantes), se realizó una imputación o estimación de estos valores por el método EM.

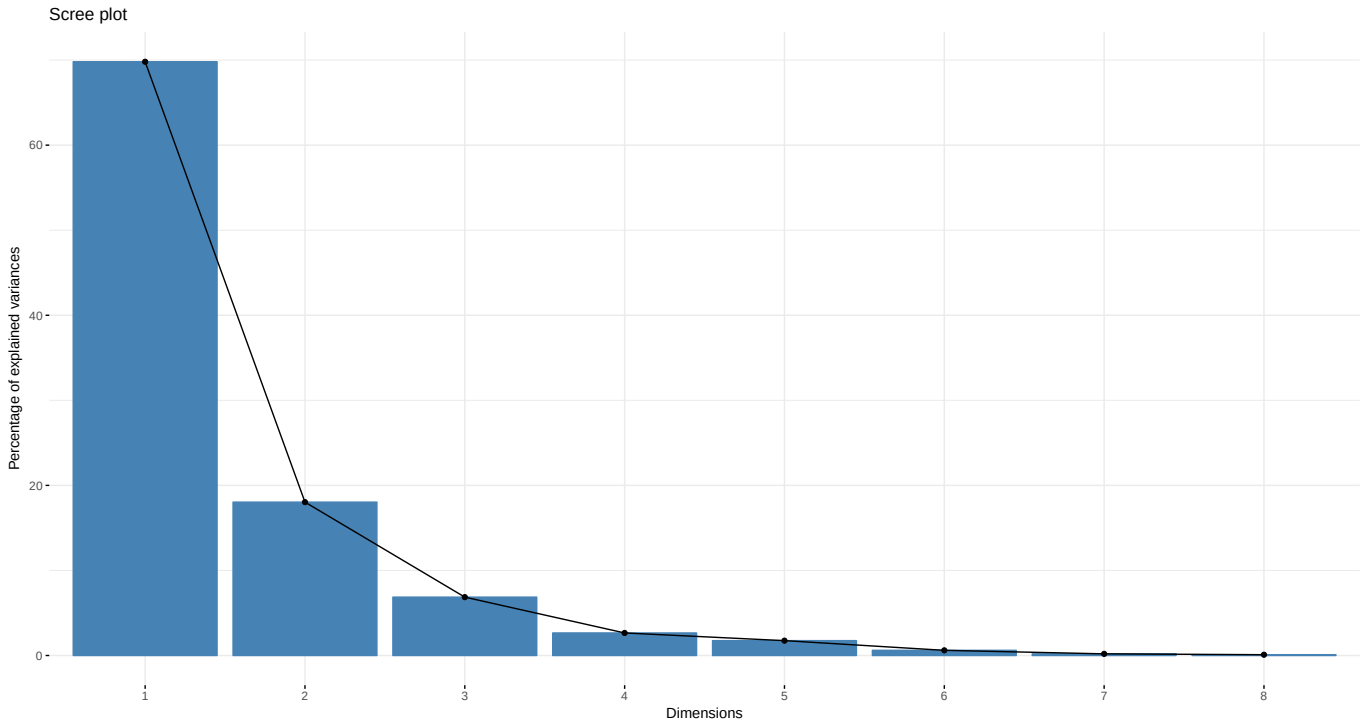
La matriz imputada es:

	Color.intensity	Odor.intensity	Attack.intensity	Sweet	Acid	Bitter	Pulp	Typicity
1	4.791667	5.291667	**4.124561**	**5.501018**	**4.154464**	2.833333	**5.574109**	5.208333
2	4.583333	6.041667	4.416667	5.458333	4.125000	3.541667	4.625000	4.458333
3	4.708333	5.333333	**4.168132**	**5.449296**	4.291667	3.166667	6.250000	5.166667
4	6.583333	6.000000	7.416667	4.166667	6.750000	**4.748854**	1.416667	3.416667
5	**6.169035**	6.166667	5.333333	4.083333	**5.332292**	4.375000	3.416667	4.416667
6	6.333333	5.000000	5.375000	5.000000	5.500000	3.625000	4.208333	4.875000
7	4.291667	4.916667	5.291667	5.541667	5.250000	**3.213539**	1.291667	4.333333
8	**4.536606**	4.541667	4.833333	**5.430180**	4.958333	2.916667	1.541667	3.958333
9	4.416667	**4.957997**	5.166667	4.625000	5.041667	3.666667	1.541667	3.958333
10	4.541667	4.291667	**4.278448**	5.791667	4.375000	**2.724521**	**3.593447**	5.000000
11	4.083333	5.125000	3.916667	**5.712512**	**3.833623**	**2.790614**	7.333333	5.250000
12	6.500000	5.875000	6.125000	4.875000	5.291667	4.166667	1.500000	3.500000

3.2. Porcentaje de inercia explicado

El porcentaje de varianza explicado por cada dimensión se puede ver en la siguiente tabla:

	Valor propio	Porcentaje de Inercia	Porcentaje de Inercia acumulado
Dim.1	1.840695280	69.6179879	69.61799
Dim.2	0.501038381	18.9500589	88.56805
Dim.3	0.169785432	6.4215518	94.98960
Dim.4	0.065450856	2.4754542	97.46505
Dim.5	0.044459815	1.6815401	99.14659
Dim.6	0.016131006	0.6101000	99.75669
Dim.7	0.004671617	0.1766879	99.93338
Dim.8	0.001761408	0.0666192	100.00000



Tanto en la tabla como en la gráfica observamos que las dos primeras componentes son las que más porcentaje de inercia explican, entre las dos explican un porcentaje de inercia acumulado del 88.57 %. Teniendo en cuenta lo anterior, seleccionamos los dos primeros ejes para realizar el AFM.

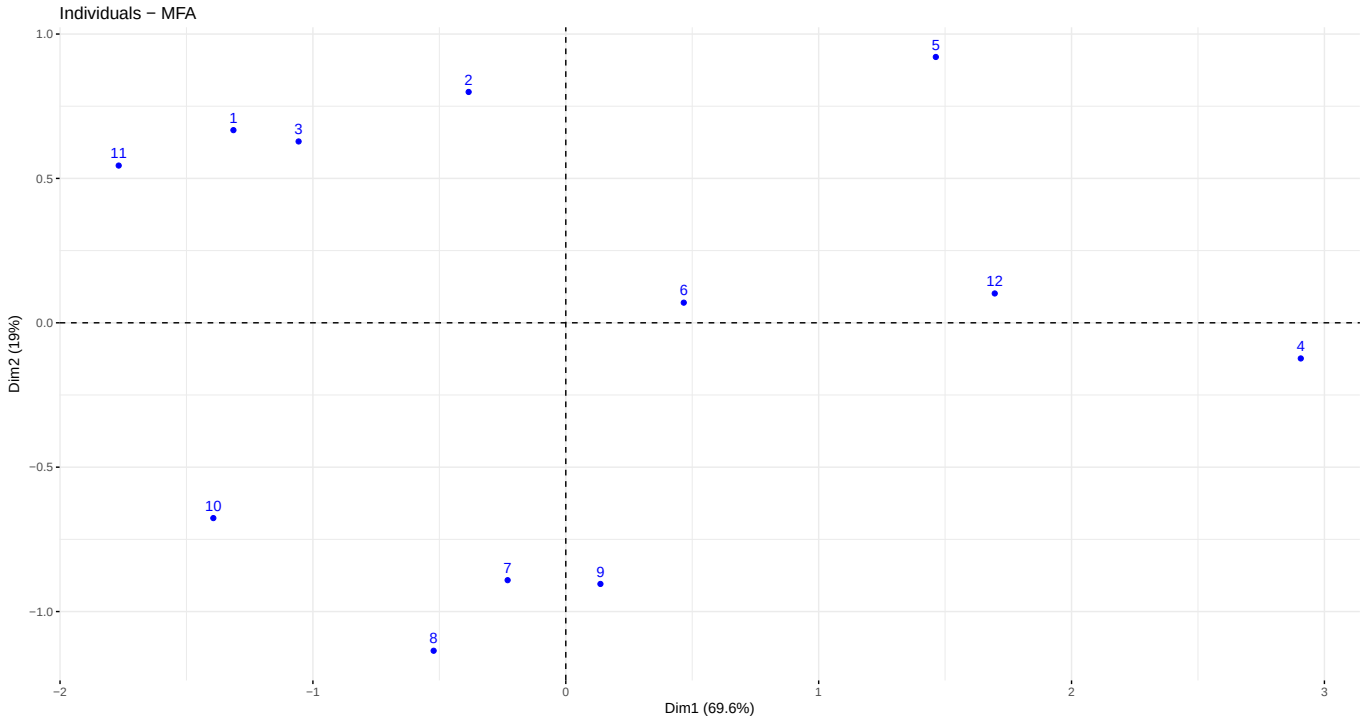
3.2. Resultados para los individuos

Los resultados para los individuos en las dos primeras componentes son:

3.2.1. Coordenadas

	Dim.1	Dim.2
1	-1.3143364	0.66717540
2	-0.3841853	0.79937921
3	-1.0565914	0.62799661
4	2.9064639	-0.12345710
5	1.4632792	0.92058312
6	0.4665257	0.06981062
7	-0.2300072	-0.89133817
8	-0.5225866	-1.13568796
9	0.1370254	-0.90435088
10	-1.3937676	-0.67628794
11	-1.7680431	0.54456430
12	1.6962234	0.10161280

3.2.2. Nube de individuos



Respecto a la gráfica anterior, se ve que se forman grupos de jugos de naranja que son muy parecidos entre ellos en cuanto a las 8 variables medidas, ya que las distancias son muy pequeñas. Por ejemplo, los jugos 1,2,3 y 11 se podría decir que tienen cualidades muy parecidas en las variables medidas, al igual que los jugos 7,8 y 9 que también se encuentran muy cerca entre ellos.

3.2.3. Contribuciones

	Dim.1	Dim.2
1	7.82077792	7.40334197
2	0.66821834	10.62804681
3	5.05418322	6.55937355
4	38.24430069	0.25350113
5	9.69373756	14.09528211
6	0.98534439	0.08105705
7	0.23950776	13.21395346
8	1.23638399	21.45190194
9	0.08500387	13.60259270
10	8.79463025	7.60695852
11	14.15215481	4.93226139
12	13.02575720	0.17172937

Las contribuciones nos dicen que tanto contribuyen las variables o los individuos a la construcción de cada eje. En este caso, los individuos (jugos) que más aportan a la construcción del eje 1 son 4,11 y 12, mientras que los individuos 2,5,7,8 y 9 son los que más aportan a la construcción del eje 2.

3.2.3. Cosenos cuadrados

	Dim.1	Dim.2
1	0.78366123	0.201927428
2	0.10365480	0.448759438
3	0.72290537	0.255377089
4	0.97704556	0.001762857
5	0.65512731	0.259296895
6	0.17896249	0.004007318
7	0.04884725	0.733571370
8	0.16466255	0.777671827
9	0.01490332	0.649165033
10	0.73853508	0.173881625
11	0.88421345	0.083882184
12	0.88317768	0.003169413

Los cosenos cuadrados nos muestran que tan bien representados están las variables o individuos en cada dimensión. En este caso, podemos ver que la dimensión 1 ó el eje 1 representa muy bien a los individuos(jugos) 1,3,4,5,10,11 y 12, mientras que la dimensión 2 ó eje 2 representa mejor los jugos 7 ,8 y 9.

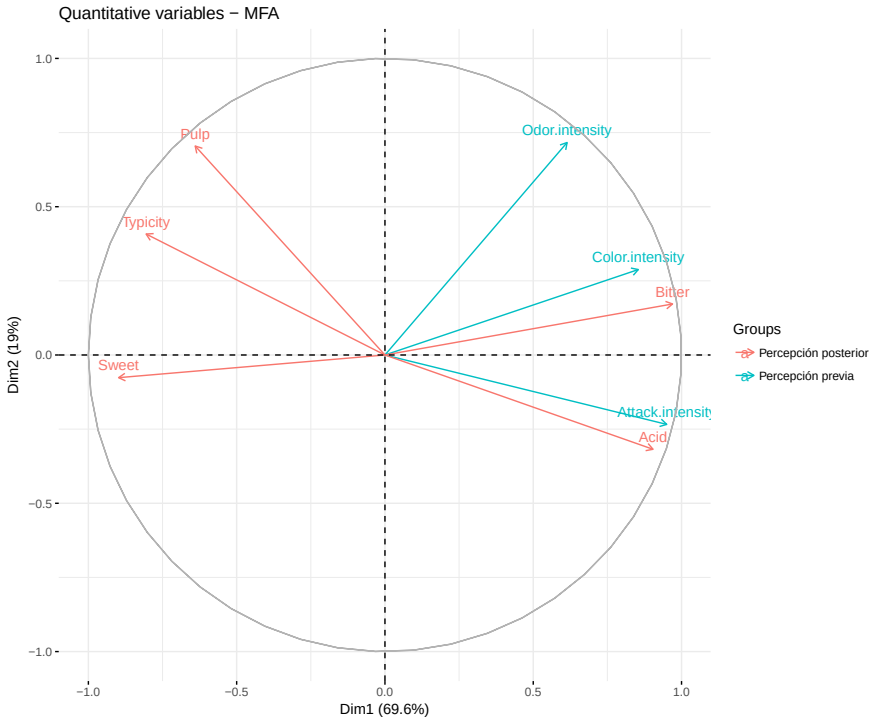
3.3. Resultados para las variables

Los resultados para las variables en las dos primeras componentes son:

3.3.1. Coordenadas

	Dim.1	Dim.2
Color.intensity	0.8542630	0.28809072
Odor.intensity	0.6141905	0.71649921
Attack.intensity	0.9502405	-0.23400493
Sweet	-0.8979722	-0.07642941
Acid	0.9039337	-0.31789113
Bitter	0.9697080	0.17204423
Pulp	-0.6398293	0.70466313
Typicity	-0.8053443	0.40789439

3.3.2. Nube de variables



En esta gráfica podemos observar que la variable intensidad del color correspondiente al grupo de percepción previa, esta correlacionada positivamente con la variable amargo y negativamente con la variable dulce, ambas del grupo percepción posterior. Esto nos indica que mientras mas intenso sea el color del jugo, este será mas amargo y menos dulce. También observamos que la intensidad del ataque(primera sensación del jugo en la boca) coorespondiente a la percepción previa, esta correlacionada positivamente con el ácido y negativamente con la tipicidad, ambas correspondientes a la percepción posterior. Con esto, podemos concluir que mientras mayor sea la intensidad del ataque, el jugo será más ácido y tendrá menor tipicidad.

3.3.3. Contribuciones

	Dim.1	Dim.2
Color.intensity	18.104829	7.5645086
Odor.intensity	9.358741	46.7900601
Attack.intensity	22.401565	4.9908234
Sweet	11.162149	0.2970666
Acid	11.310847	5.1391316
Bitter	13.016792	1.5052657
Pulp	5.666961	25.2520138
Typicity	8.978116	8.4611302

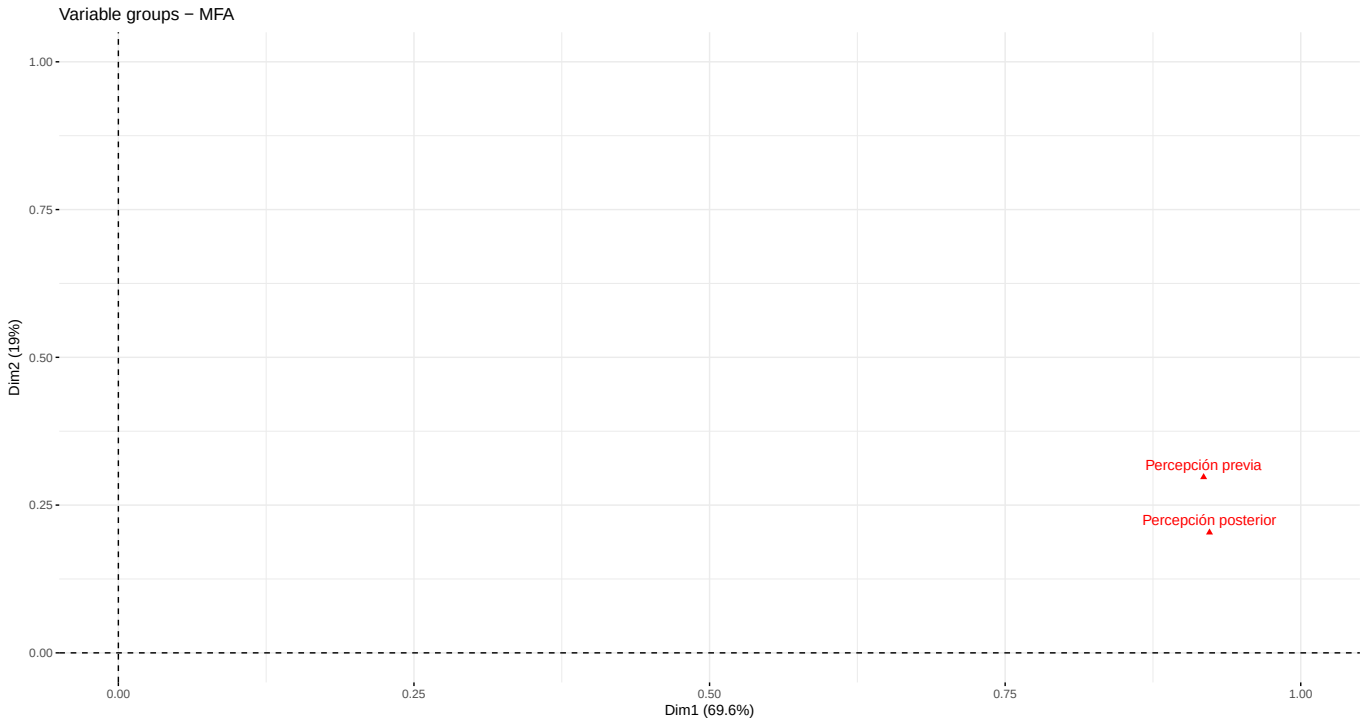
Las variables ó atributos que más aportan a la construcción del eje 1 son la intensidad del color y del ataque, mientras que la pulpa y la intensidad del olor son los atributos que mas aportan a la construcción del eje 2.

3.3.4. Cosenos cuadrados

	Dim.1	Dim.2
Color.intensity	0.7297652	0.082996266
Odor.intensity	0.3772299	0.513371117
Attack.intensity	0.9029570	0.054758309
Sweet	0.8063541	0.005841454
Acid	0.8170961	0.101054769
Bitter	0.9403336	0.029599218
Pulp	0.4093815	0.496550127
Typicity	0.6485794	0.166377831

Podemos ver que la dimensión ó eje 1 representa muy bien los atributos de intensidad del ataque, intensidad del color, amargo, dulzor y acidez, mientras que la dimensión ó eje 2 representa mejor la intensidad del olor y la pulpa.

3.4. Nube de los grupos



En esta gráfica podemos observar que los dos grupos están relativamente cerca, esto nos indica que la percepción previa de los jugos de naranja coincide con la percepción posterior, es decir, existe una relación fuerte entre el aspecto del jugo en cuanto al olor, al color y a la primera sensación en la boca, y el sabor (acidez,dulce, amargo) que este tendrá cuando se deguste.

3.5. Coeficientes Lg y Rv de Escoufier

- Coeficiente Lg: Es un indicador del grado de similitud o deformación con respecto a un foco (homotecia) entre los conjuntos de indicadores, y cuando se calcula para un solo conjunto de ellos. Esto se conoce como indicador de la dimensionalidad de la nube, que es igual al número de direcciones ortogonales de inercia no cero, es decir, el número de valores propios no cero. Esta cantidad es 0 cuando todas las variables de un grupo son ortogonales a todas las variables del otro grupo. Es mas alto en cuanto cada una de las variables de un grupo este más relacionada con el conjunto de variables del otro grupo.
  Se define por:

$$Lg = \frac{Traza(S'T)}{\alpha_1^2 x \lambda_1^2}$$

donde α_1 y λ_1 es el primer valor propio de S y de T respectivamente.

Los coeficientes Lg se pueden observar en la siguiente tabla:

	Percepción previa	Percepción posterior	MFA
Percepción previa	1.0839432	0.7761085	1.0105158
Percepción posterior	0.7761085	1.0384112	0.9857795
MFA	1.0105158	0.9857795	1.0845333

El valor del coeficiente $Lg_{(P.Previa)} = 1.0839$ para la percepción previa indica claramente que es de dimensionalidad uno, es decir, que puede sintetizarse en un solo factor; $Lg_{(P.Posterior)} = 1.0384$ indica que la percepción posterior también tiene una dimensión o factor que lo caracteriza. El coeficiente Lg cruzado $Lg_{(P.Prev,P.Post)} = 0.7761$ indica que estos dos grupos comparten un factor; y finalmente, el coeficiente $Lg_{(MFA)} = 1.0845$ indica que éste se puede sintetizar como mínimo en un factor.

- Coeficiente Rv de Escoufier: Es una generalización multivariada del coeficiente de correlación de Pearson al cuadrado. Este coeficiente mide el vínculo entre dos grupos o dos matrices de variables. Este coeficiente, al igual que el de correlación de Pearson, se encuentra entre 0(todas las variables del primer grupo o matriz, son ortogonales a todas las variables del segundo grupo o matriz) y 1(los dos grupos o matrices son homotéticos, es decir, las dos matrices son muy parecidas).

El coeficiente de RV se define como (Robert y Escoufier, 1976; Schlich, 1996):

$$RV(W_i, Wj) = \frac{T(W_i, W_j)}{[T(W_i, W_i) \cdot T(W_j, W_j)]^{\frac{1}{2}}}$$

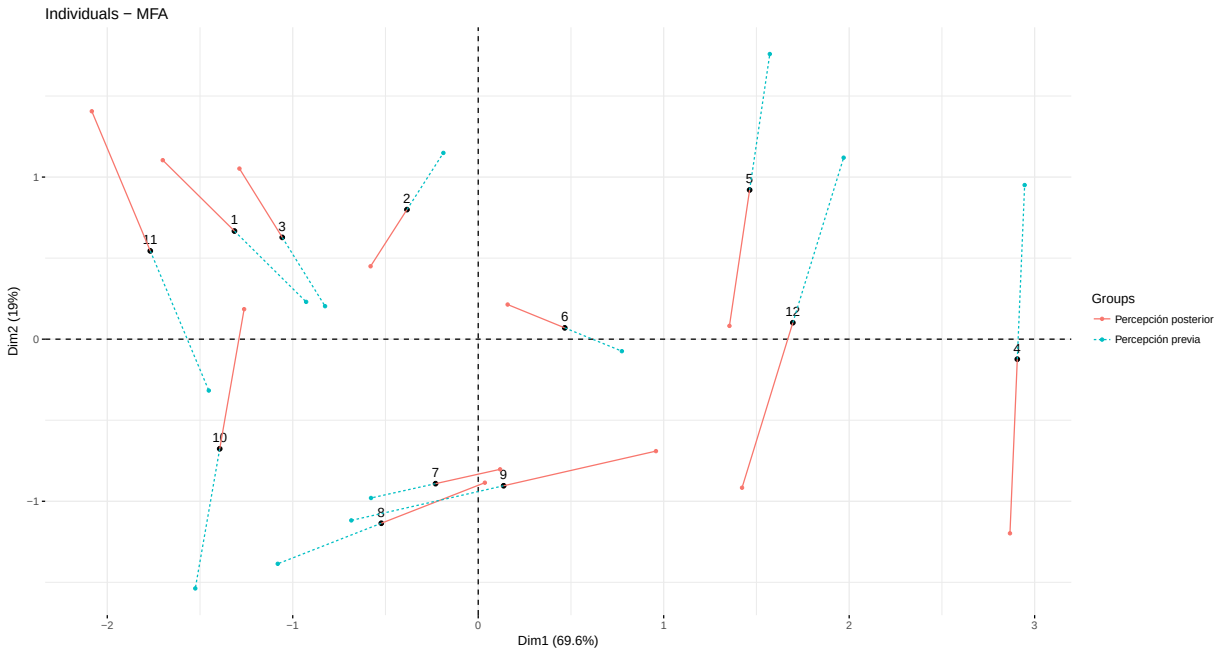
Donde $T(W_i, W_j) = \sum_{l,m} w^i_{l,m} w^j_{l,m}$ es un coeficiente de covarianza generalizado entre las matrices W_i y W_j, $T(W_i, W_i) = \sum_{l,m} {w^i_{l,m}}^2$ es una varianza generalizada de la matriz W_i y $w^2_{l,m}$ es el (l,m) elemento de la matriz W_i.

Los coeficientes Rv se pueden observar en la siguiente tabla:

	Percepción previa	Percepción posterior	MFA
Percepción previa	1.0000000	0.7315340	0.9320054
Percepción posterior	0.7315340	1.0000000	0.9289101
MFA	0.9320054	0.9289101	1.0000000

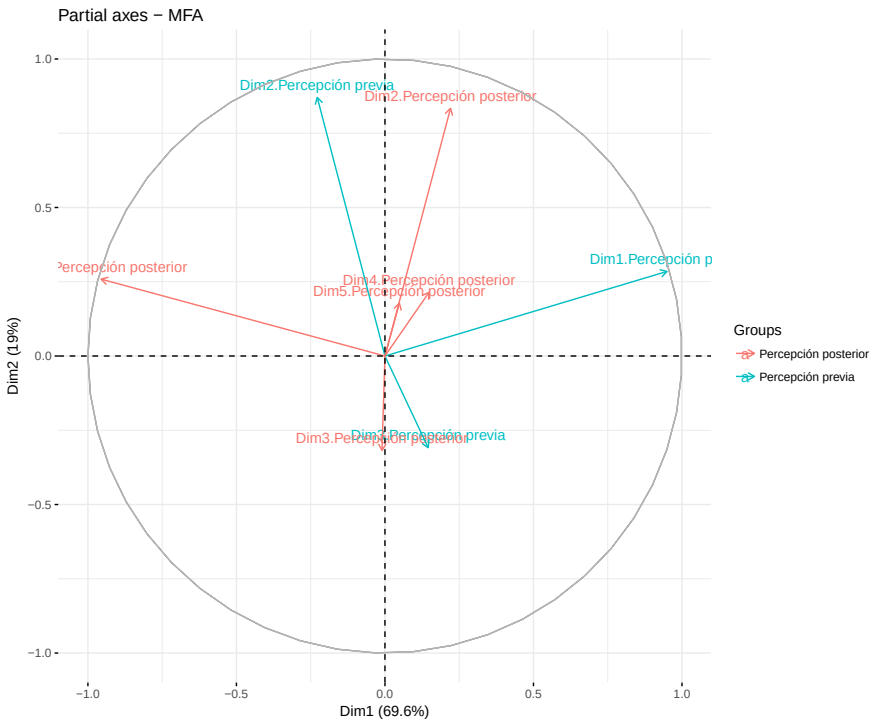
Los valores de los coeficientes $Rv_{(MFA,P.Previa)} = 0.932$ y $Rv_{(MFA,P.Posterior)} = 0.9289$ nos indican que ambos grupos (percepción previa y posterior) tienen una estructura cercana a la de toda la degustación ó en otras palabras, tienen un grado considerable de asociación con el AFM. Es decir, que su representación sobre los planos generados por el AFM es adecuada. Además, entre la percepción previa y posterior el coeficiente Rv es de 0.7315340 lo que significa que existe un vinculo considerable entre estos dos grupos (algunas de las variables del primer grupo están asociadas con las del segundo grupo).

3.6. Representación superpuesta



La representación superpuesta nos muestra gráficamente en un mismo plano los diferentes individuos en cada uno de los grupos en el AFM(tablas individuales ó grupos, y tabla yustapuesta), esta gráfica nos sirve para ver que tan distantes son los individuos de un grupo al otro. En este caso observamos la percepción previa en azul y la percepción posterior en rojo, donde el punto medio entre estos dos es el individuo en la nube de puntos del ACP global(el cual se realiza a la tabla yustapuesta). En otras palabras, esto nos dice que tan similares son los individuos en cada grupo de acuerdo con la descripción global; podemos apreciar como los individuos 2 y 6 en ambos grupos son muy similares, mientras que individuos como el 12 y el 4 se comportan diferentes de un grupo a otro.

3.7. Ejes parciales



Esta gráfica muestra la proyección de los componentes principales normalizados de cada uno de los 2 grupos (percepción previa y percepción posterior) en términos de los dos primeros componentes del AFM. El primer componente del AFM, vemos que esta muy relacionado como era de esperarse, con el primer componente de cada grupo (percepción previa positivamente y percepción posterior negativamente), y el segundo componente del AFM, esta relacionado con el segundo componente de cada grupo, ambos se relacionan positivamente. Se observa que las otras componentes de cada grupo no tienen mucha relación con las dos primeras componentes del AFM.

4. Construcción indice

Las coordenadas de las variables para las dos primeras dimensiones son:

	Dim.1	Dim.2
Color.intensity	0.8581049	0.2236407
Odor.intensity	0.6316557	0.7028833
Attack.intensity	0.9522195	-0.2260020
Sweet	-0.8881581	-0.1346408
Acid	0.9028145	-0.3139184
Bitter	0.9640321	0.1981328
Pulp	-0.6320766	0.7018089
Typicity	-0.8054955	0.3978624

Se observa que el primer grupo de variables que corresponde a Color, Odor y Attack tienen el mismo signo en sus coordenadas de la primera componente, lo cuál no sucede con el otro grupo conformado por las 5 variables restantes. Se puede conformar un indice para el primer grupo (percepción previa) pero no es posible hacerlo para el segundo (percepción posterior) y por consiguiente no se puede realizar un indice global. Los indices de los jugos para el primer grupo (percepción previa) se muestran en la siguiente tabla:

Jugo	Indice
1	11.03415
2	11.823
3	11.16971
4	16.35666
5	14.28656
6	13.58883
7	11.71434
8	11.12588
9	11.61305
10	10.57015
11	10.35774
12	14.9813

Se logra apreciar que los jugos que sobresalen del resto en cuanto a la percepción previa, es decir, los que mayor puntaje tuvieron en esta percepción fueron en orden decreciente el 4,12,5 y 6.

Referencias

Brigitte Escofier, J. P. (2008), *Analyses factorielles simples et multiples : Objectifs, méthodes et interprétation*, sciences sup, 4ème édition edn, Dunod.

Josse, J. & Husson, F. (2016), 'missMDA: A package for handling missing values in multivariate data analysis', *Journal of Statistical Software* **70**(1), 1–31.

Kassambara, A. & Mundt, F. (2017), *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.5.
*https://CRAN.R-project.org/package=factoextra

Lê, S., Josse, J. & Husson, F. (2008), 'FactoMineR: A package for multivariate analysis', *Journal of Statistical Software* **25**(1), 1–18.

Salamanca, J. A. C. (2017), 'Análisis factorial múltiple para clasificación de universidades latinoamericanas', *Comunicaciones en Estadística* .

Vaissie, P., Monge, A. & Husson, F. (2017), *Factoshiny: Perform Factorial Analysis from 'FactoMineR' with a Shiny Application*. R package version 1.0.6.
*https://CRAN.R-project.org/package=Factoshiny

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
*http://ggplot2.org