Matériel de cours

Analyse en composantes principales

Laura Symul

12 janvier 2023

UCLouvain - ISBA

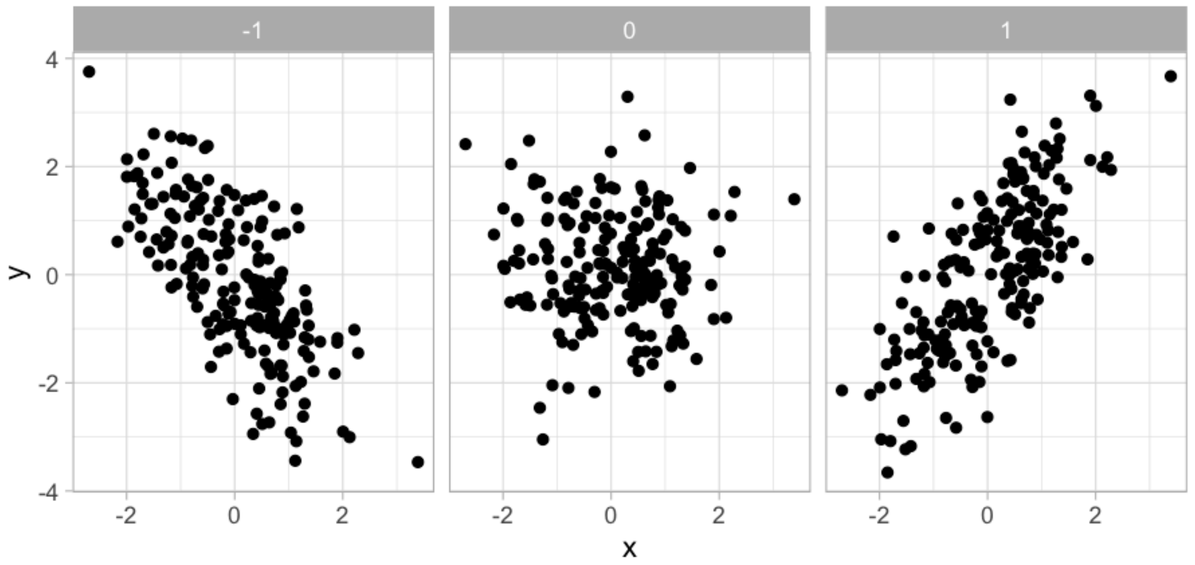
Analyse en composantes principales

Outil d'analyse descriptive / d'exploration de jeux de données **multivariés**

Multivarié = plusieurs variables

2 variables:

	x	y
	<dbl>	<dbl>
1	-2.55	2.48
2	-2.11	1.29
3	-1.93	0.277
4	-1.85	3.19
5	-1.69	1.44
6	-1.67	1.57
7	-1.62	3.20
8	-1.50	-0.700
9	-1.42	2.62
10	-1.35	1.42



p variables:

	a	b	c	d	e	f	g	h	i	
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<db
1	1.36	-1.11	0.491	-0.221	-2.02	0.591	-0.269	-2.32	-1.47	-0.18
2	-0.787	0.134	-0.502	0.330	-0.665	-0.317	0.760	0.683	-0.645	1.00
3	-0.385	0.626	0.288	1.29	0.435	-0.703	-1.23	0.721	-1.35	-0.85
4	0.331	0.873	-0.687	0.334	-0.0729	-1.37	-1.27	0.629	0.309	-1.72
5	-0.555	0.816	0.788	1.48	-1.28	-0.729	0.127	-0.411	0.421	-0.49
6	0.122	-0.968	0.691	-0.648	1.02	-0.122	0.760	0.0993	-1.37	1.07
7	-0.0476	-1.31	1.24	1.35	1.16	-0.635	-0.408	-1.43	0.0479	-1.36
8	-0.776	-2.01	1.98	0.593	-0.940	0.595	-0.577	0.360	-0.914	-0.24
9	0.831	0.505	-0.646	-1.19	-0.185	0.368	-2.84	-0.817	0.494	0.96
10	0.846	0.828	0.966	0.402	0.731	-1.61	0.374	1.32	-1.55	0.04
11	1.02	0.336	-1.43	-1.24	0.864	0.317	-0.903	-0.926	0.865	0.56
12	1.27	-1.06	-0.457	0.947	-0.529	0.542	0.403	-2.06	-0.0444	1.62
13	-0.506	1.57	0.945	-0.540	-0.879	-0.261	1.43	0.173	-1.04	0.90
14	-0.464	-0.370	-0.738	-0.236	0.345	-0.0265	-0.496	-1.63	-0.0829	0.59
15	0.261	1.78	0.346	-0.680	-1.92	-0.432	0.697	-0.928	-0.266	0.49
16	0.630	0.551	-0.900	-0.852	-0.950	-0.293	-0.153	0.804	-0.364	-1.16
17	-0.340	1.19	-0.370	1.70	0.774	-1.01	0.369	1.87	-1.50	0.83
18	-0.423	0.331	-0.0408	0.992	-1.70	0.425	-1.16	-0.263	-0.585	0.54
19	-0.618	-0.0647	-0.612	0.675	-1.20	-1.22	0.418	0.972	-0.173	0.40
20	1.48	-1.01	-1.95	0.0736	-0.241	-0.590	1.30	-1.38	-1.48	0.05
21	-2.51	-0.559	0.243	0.736	-0.302	-0.914	1.80	1.73	0.302	-1.27
22	-0.168	-0.0471	0.475	0.662	-1.75	-0.871	-1.19	-0.00509	1.37	0.15
23	0.0382	0.282	0.137	1.60	0.509	-0.369	0.534	-0.780	0.782	1.25
24	-1.06	-0.0332	-0.489	0.850	-1.02	-0.485	-0.508	0.398	-0.799	-0.61
25	0.385	-0.178	0.900	-0.206	-0.158	-0.700	0.523	-0.809	-0.657	-1.80

Mise en bouche

Des chercheurs des Pays-Bas ont mené une étude pour obtenir des données sur la perception des saveurs des aliments les plus consommé aux Pays-Bas.

Les testeur.se.s ont préalablement été formé à reconnaître 6 saveurs primaires (**sucré, salé, amer, acide, umami, sensation de gras**) sélectionnées par les chercheurs.

CC Photo by Mae Mu https://unsplash.com/photos/H5Hj8QV2Tx4

Combien de saveurs gustatives primaires percevons-nous?

La base de donnée contient les valeurs moyennes et écart-type pour plus de 600 produits.

Quelques exemples:

Valeurs moyennes pour chaque saveur.

Product_description_EN	Food_group_EN	Number of panelist	m_sweet	m_sour	m_bitter	m_umami	m_salt	m_fat
Coffee prepared	(non) alcoholic beverages	11	2	9	63	1	3	4
Fruit juice concentrated	(non) alcoholic beverages	8	62	46	0	0	0	6
Pineapple	Fruit	11	34	34	1	0	2	5
Banana	Fruit	12	29	2	1	1	1	24
Flan filled with rice pudding	Pastry, Cakes and Biscuits	7	35	1	0	2	13	44
Endive raw	Vegetables	9	4	1	10	1	1	2
Onion juice	Vegetables	9	6	10	74	2	3	2
Cheese La Vache qui rit	Cheese	12	10	18	3	11	25	73
Herring pickled (sweet)sour	Fish	11	4	73	3	18	36	39
Tuna in oil tinned	Fish	8	3	19	2	31	37	36
…								

Que pouvons-nous observer de ces exemples? Il y a-t-il des tendances ou corrélations entre les saveurs?

Quels outils d’exploration ou d’analyses statistiques peut-on utiliser pour appréhender ces données?

Objectifs et applications d'une PCA

PCA: principal component analysis

Une **analyse en composantes principales (ACP)** nous permet de répondre à ces questions:

Dans un jeu de données multivarié:

- Existe-il des corrélations entre les variables?
- Si oui, peut-on **résumer** (compresser) l'information contenue dans ces données dans un **espace de dimension moindre** que celui des données originales?
- Quelles sont les variables corrélées?
- Comment les échantillons sont-ils disposés dans cet espace de dimension moindre?
  (Quels sont les échantillons similaires?)

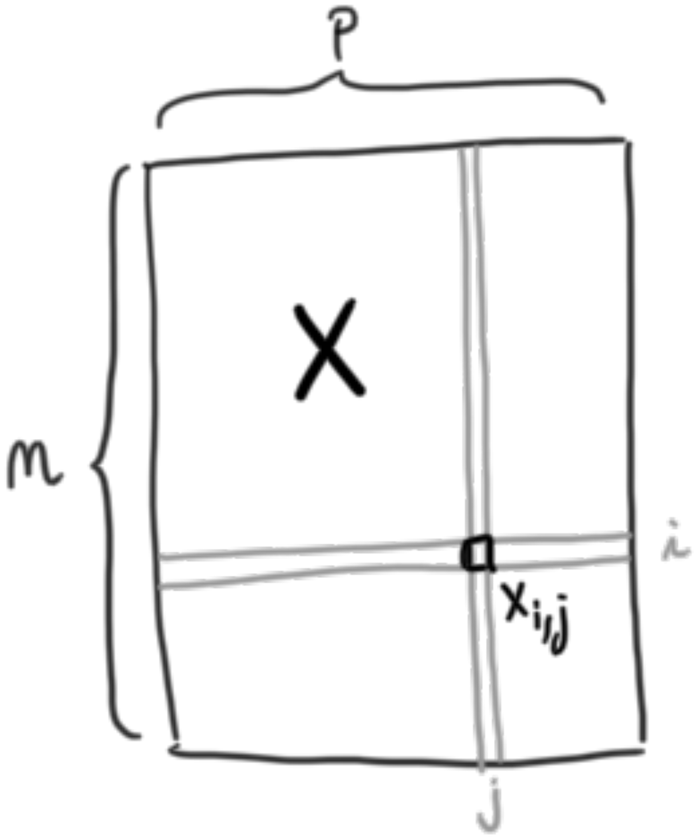


Table des matières

Aujourd'hui

Introduction générale à l'analyse en composantes principales (PCA)
- Objectifs et applications de l'analyse en composantes principales
- Exemple de résultats d'une PCA
- Comment identifier les composantes principales (principe général)
- Cercles des corrélations
- Projection dans l'espace des composantes principales (+ biplot)

Composante principales (calcul, propriétés, et preuves)
- Calcul des composantes principales via la décomposition en valeurs singulières (SVD)
- Choix du nombre de composantes; exercises de simulations de données
- Distribution des variables (scaling / standardizing)

Applications et limitations
- Expression génétique dans différent tissus / tumeurs / types de cellule
- Temporal trajectories / geospatial variations
- Cas où le nombre de variables > nombre d'échantillons
- Cas de relations non-linéaires entre les variables

Problèmes et méthodes connexes (e.g., LDA: Linear Discriminant Analysis)

Objectifs d'apprentissage (à la fin du chapitre)

- Comprendre les objectifs d'une PCA et reconnaître les situations dans lesquelles une PCA serait utile
  - Savoir quand et comment transformer les données avant une PCA
- Savoir effectuer une PCA avec du code R et en visualiser les résultats
  - Cercles des corrélation, projections des échantillons, "biplot"
- Pouvoir interpréter les résultats:
  - justifier le choix du nombre de composantes sélectionnées
  - interpreter un cercle de corrélation
- Savoir démontrer que la SVD permet d'identifier les composantes principales
- Comprendre comment la sélection des échantillons / variables, et leur transformations potentielles influencent les résultats d'une PCA

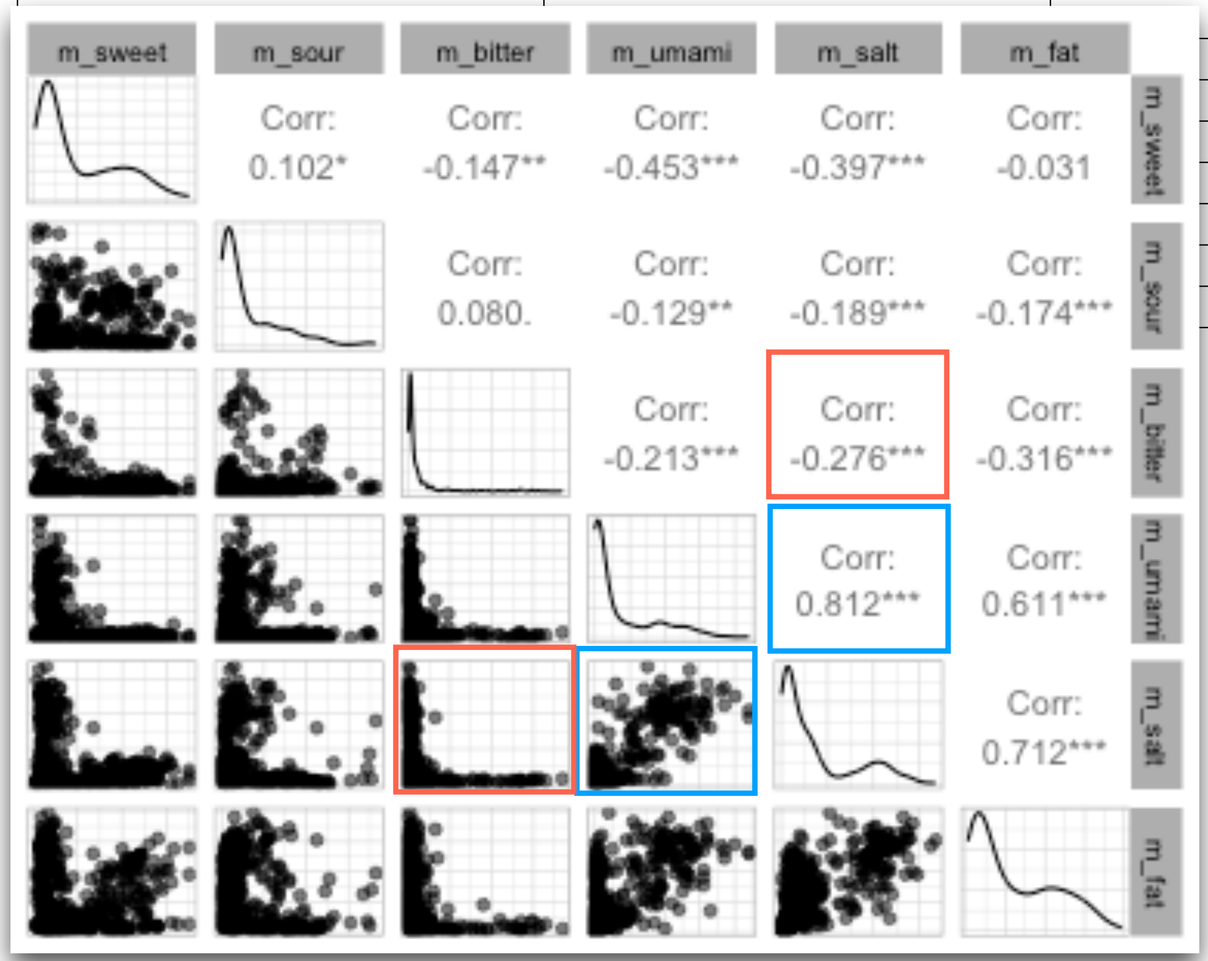
Rappel: Trouver des fautes/erreurs dans le matériel de cours contribue à la note de participation.

Matériel de cours

- Disponible en ligne
- Code R et données utilisées en classe disponible en ligne

Corrélation entre paires de variable

Product_description_EN	Food_group_EN	Number of panelist	m_sweet	m_sour	m_bitter	m_umami	m_salt	m_fat
Coffee prepared	(non) alcoholic beverages	11	2	9	63	1	3	4
Fruit juice concentrated	(non) alcoholic beverages	8	62	46	0	0	0	6
Pineapple	Fruit	11	34	34	1	0	2	5
			29	2	1	1	1	24
			35	1	0	2	13	44
			4	1	10	1	1	2
			6	10	74	2	3	2
			10	18	3	11	25	73
			4	73	3	18	36	39
			3	19	2	31	37	36

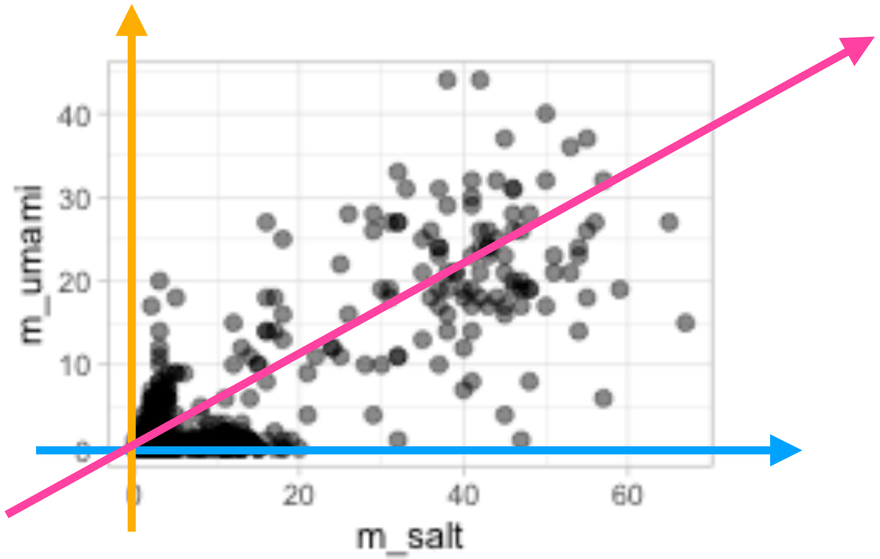


Les saveurs “salé” et “amer” s’excluent l’une l’autre.

Les saveurs “salé" et “umami" sont très corrélées.

Puisque ces saveurs sont très (anti-)corrélées, est-il nécéssaire de garder ces deux variables?
Pourrait-on les résumer par un seul axe de variation?

“Résumer” deux variables corrélées



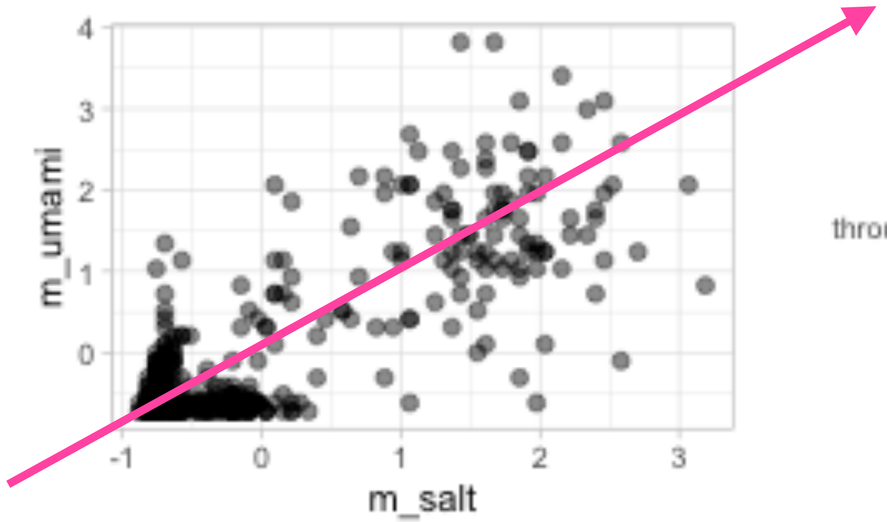
Puisque ces saveurs sont corrélées, nous pourrions ne **garder qu’une de ces deux variables**, puisque connaître la valeur d’une nous donne une bonne idée de la valeur de l’autre.

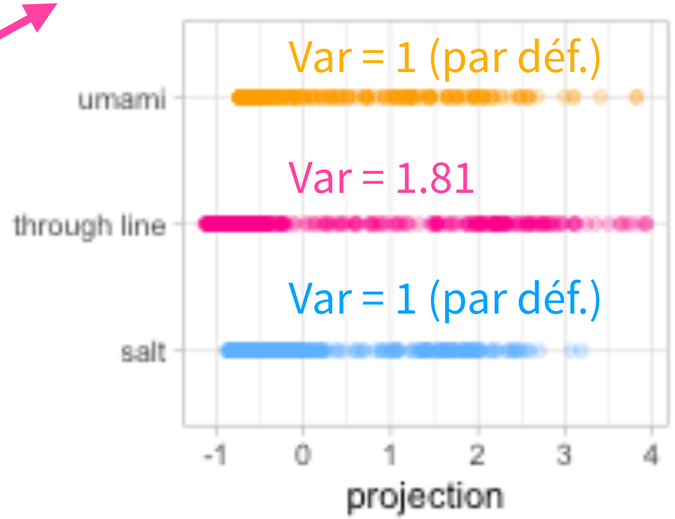
Ou **créer un nouvel axe** qui passerait par leur nuage de point.

Quelle solution vous parait la meilleure?

Le nouvel axe maximise la variance de la projection des points.

Standardisation des variables



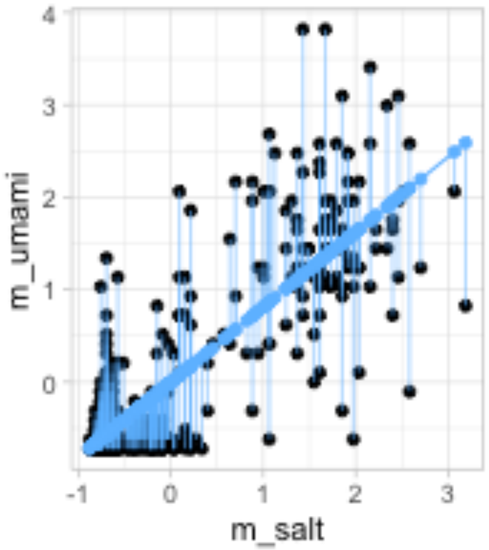


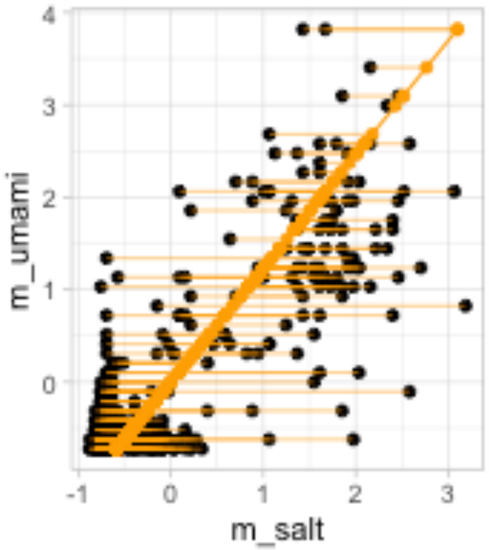
Note: si les variables ne sont pas standardisée (on soustrait la moyenne et divise par l’écart-type), alors la variable avec le plus de variance aura (artificiellement) un poids plus important que l’autre

Direction qui maximise la variance

La direction qui maximise la variance “simultanée” est **différente de celle des modèles linéaires** qui maximisent la variance d’une variables expliquée par l’autre variable.

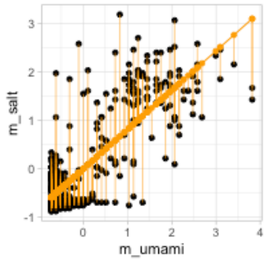
Modèles linéaires: on minimise la **distance verticale** à la droite de régression



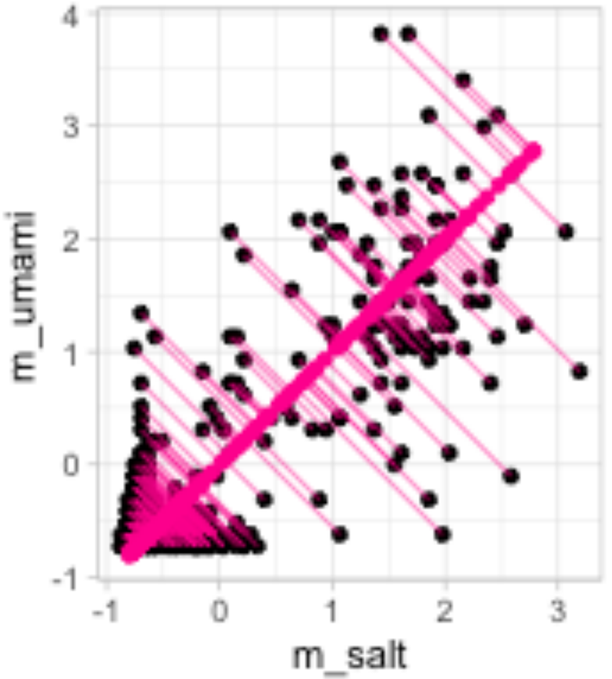


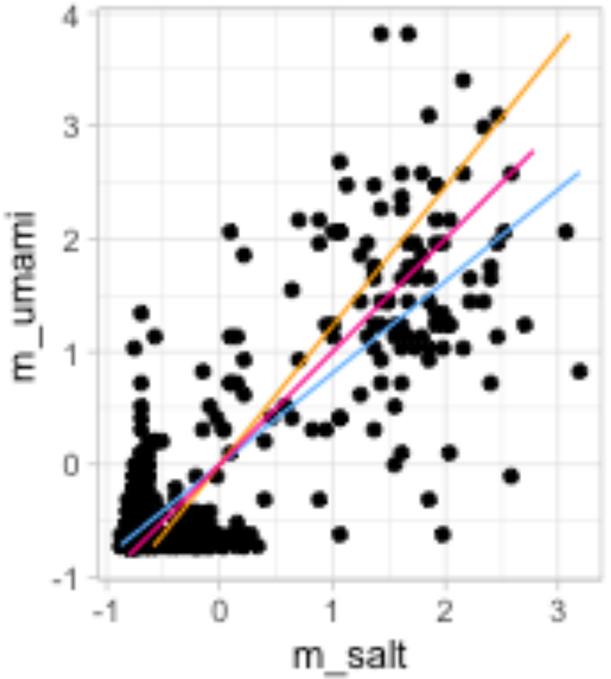
$$y = ax + b$$

$$\vec{v} = \begin{pmatrix} 1 \\ a \end{pmatrix}$$



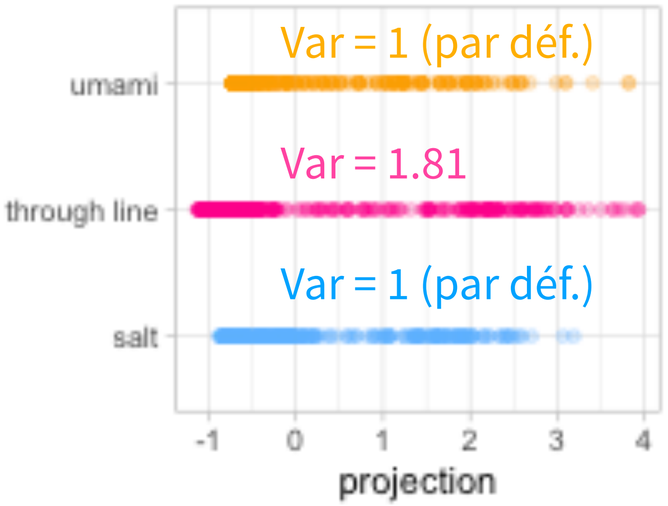
Composantes principales: on minimise la **distance orthogonale** à la droite de composante principale

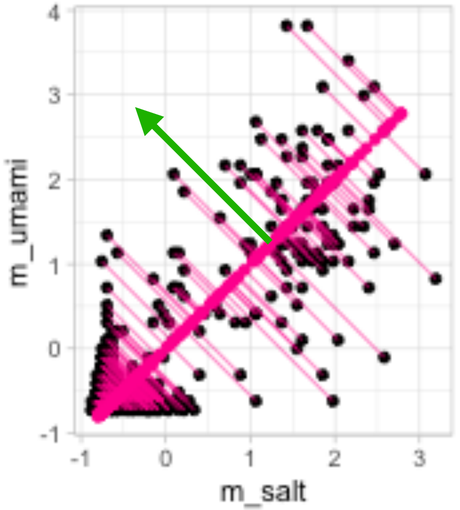




Maximiser la variance totale = minimiser la distance orthogonale

Deuxième C.P. et variance expliquée par les 2 C.P.





La première composante principale cherchait à capturer le plus de variance possible.

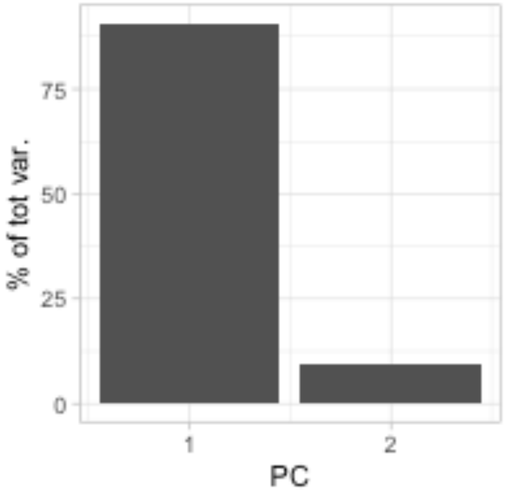
La deuxième composante principale est **orthogonale** à la première et explique le reste de la variance.

Variance totale = variance CP1 + variance CP2

Variance totale = 1 + 1 = 2
Variance CP1 = 1.81
Variance CP2 = 2 - 1.81 = 0.19

Pourcentage de variance expliquée:
CP1 = 1.81/2 = 90 %
CP2 = 0.19/2 = 10%

“Scree plot”

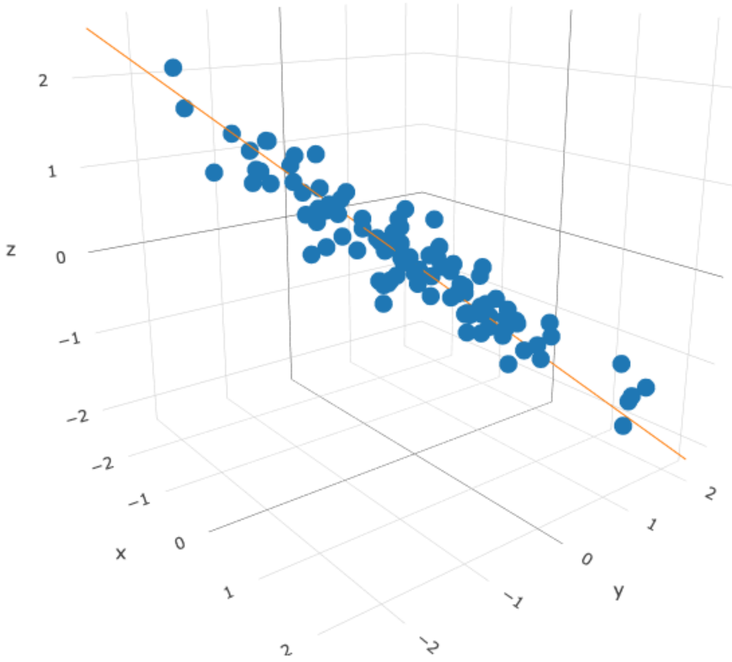


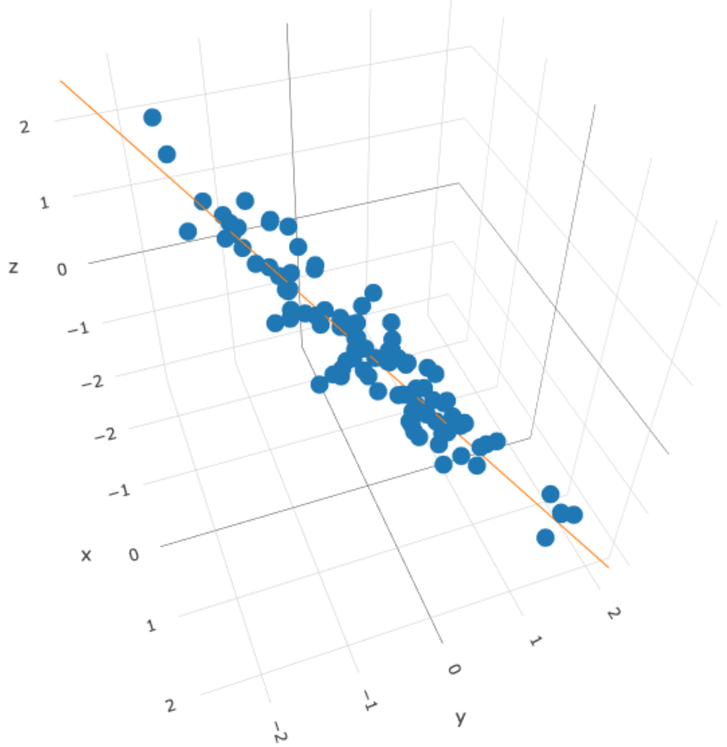
En 3 dimensions

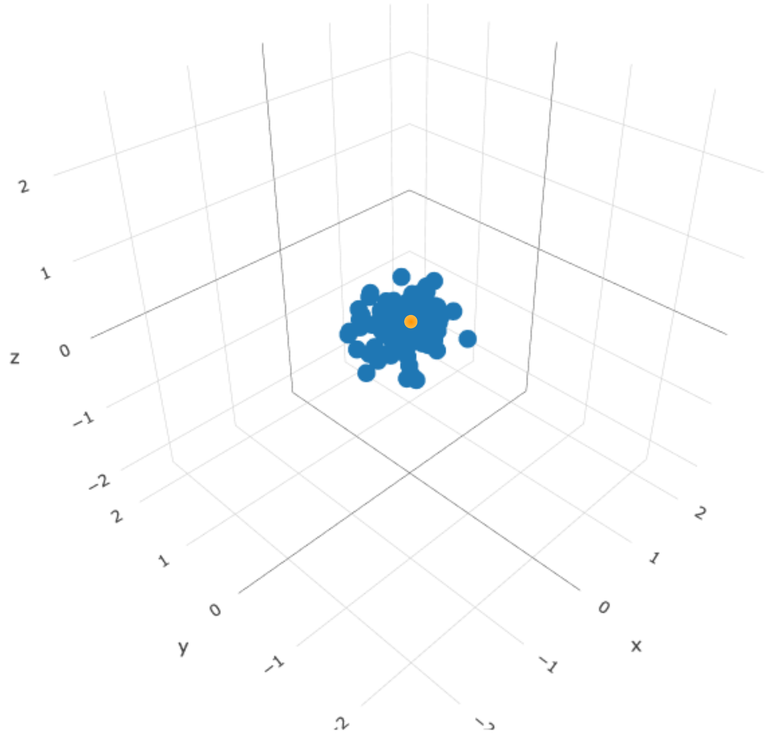
En 2 dimensions, nous avons cherché la direction qui maximise la variance totale.

Cette direction peut être décrite par un vecteur 2D.

En 3 dimensions, cette direction sera décrite par un vecteur 3D.







Une fois la première C.P. trouvée, on cherchera la **2eme C.P.** en trouvant la direction maximisant la variance des projections des points dans le plan orthogonal à la première C.P.

En dimension *p*

Si X est une matrice de dimension $n \times p$ centrée (la moyenne des colonnes est 0),
alors la matrice des covariances empiriques est $\Sigma = X^t X/(n-1)$

Pour trouver la première C.P., on cherche le vecteur $\vec{v}$ tel que

$$\arg \max_{v \in R^p} \frac{v^t \Sigma v}{v^t v}$$

On calcule ensuite les projections de X dans l'espace orthogonal à $\vec{v}$

$$X' = X - Xvv^t$$

On réitère pour X'

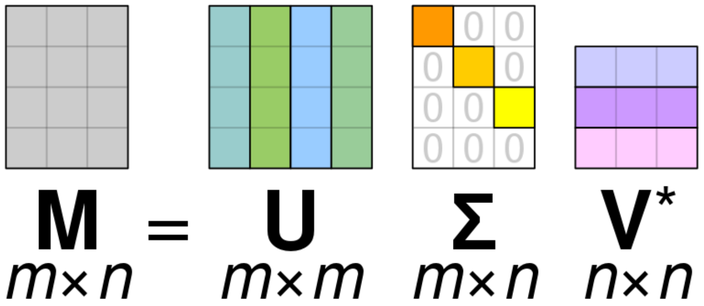
x $(p-1)$

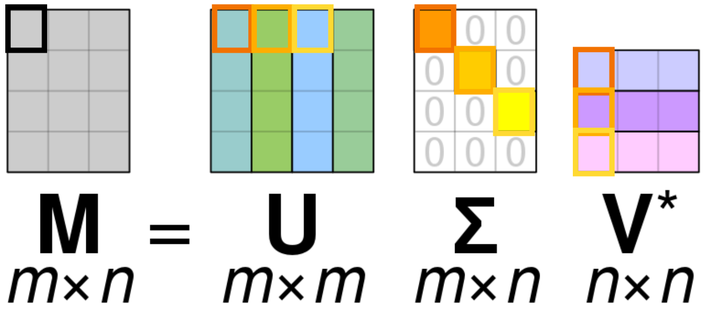
Nous verrons en détail au prochain cours que, au lieu d'utiliser cette procédure d'optimisation itérative, nous pouvons simplement obtenir les composantes principales en utilisant la **décomposition en valeurs singulières (SVD)**.

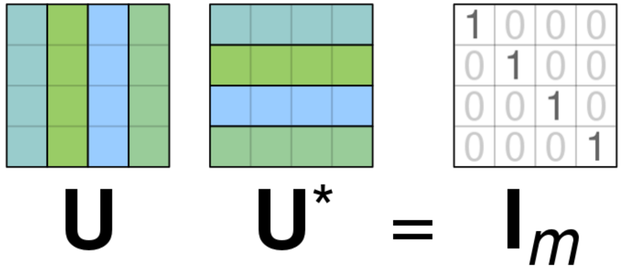
Décomposition en valeurs singulières

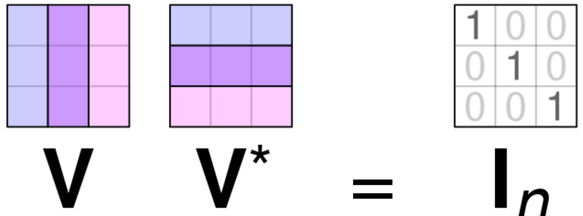
Toute matrice peut être décomposée en un produit de trois matrices:

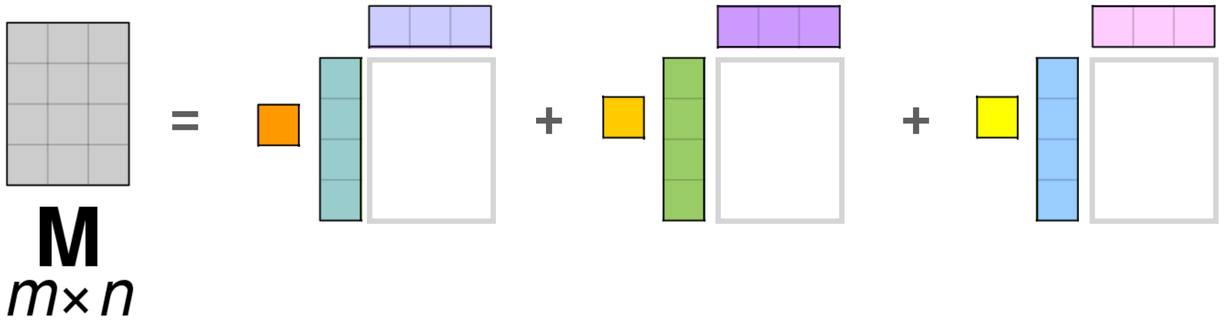
$$M = U\Sigma V^t$$







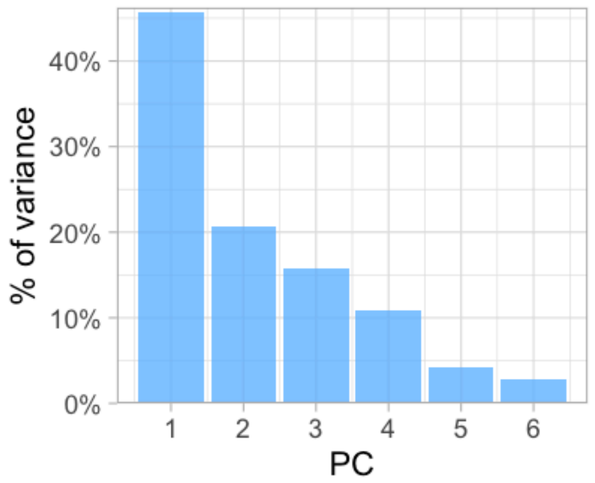


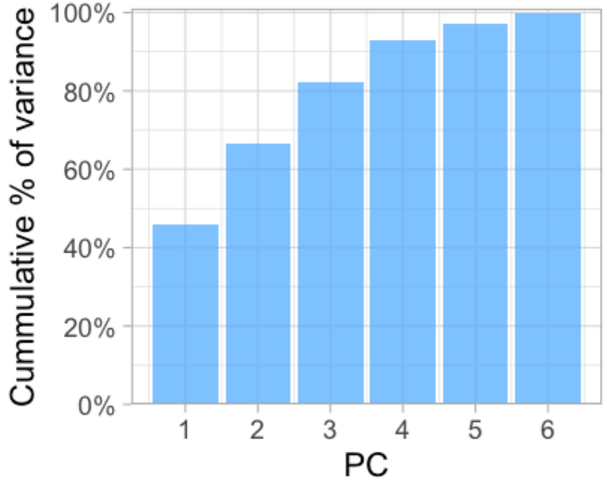


Résultats d'une PCA: exemple

Product_description_EN	Food_group_EN	Number of panelist	m_sweet	m_sour	m_bitter	m_umami	m_salt	m_fat
Coffee prepared	(non) alcoholic beverages	11	2	9	63	1	3	4
Fruit juice concentrated	(non) alcoholic beverages	8	62	46	0	0	0	6
Pineapple	Fruit	11	34	34	1	0	2	5
Banana	Fruit	12	29	2	1	1	1	24
Flan filled with rice pudding	Pastry, Cakes and Biscuits	7	35	1	0	2	13	44
Endive raw	Vegetables	9	4	1	10	1	1	2
Onion juice	Vegetables	9	6	10	74	2	3	2
Cheese La Vache qui rit	Cheese	12	10	18	3	11	25	73
Herring pickled (sweet)sour	Fish	11	4	73	3	18	36	39
Tuna in oil tinned	Fish	8	3	19	2	31	37	36
...								

```
pca_tastes ← prcomp(t, scale = TRUE) # this returns a `prcomp` object
```





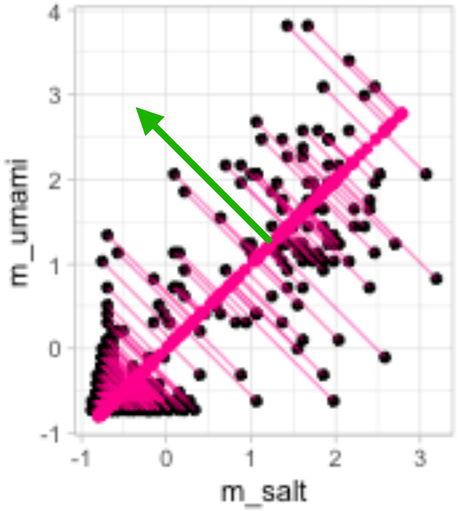
La **première composante** explique à elle-seule **plus de 45%** de la variance totale.

Les **3 premières composantes** expliquent **plus de 80%** de la variance totale.

Quels sont liens entre les variables originales (salé, sucré, etc.) et ces nouvelles composantes?

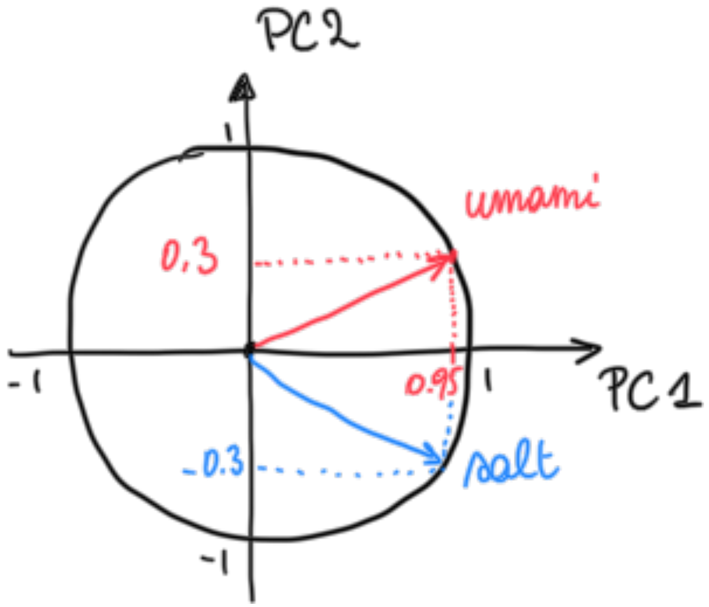
Résultats d'une PCA: cercle des corrélations

Rappel en 2D:



Il y a une + forte corrélation entre la première composante principale et les deux variables originale qu'entre la deuxième composante principale et les deux variables originales.

Nous pouvons ensuite les représenter visuellement:



Nous pouvons calculer les corrélations entre les coordonnées des échantillons dans l'espace original et les coordonnées des échantillons dans l'espace des composantes principales.

```
cor(t[, c("salt","umami")], pca_t_salt_umami$x)
```

```
              PC1        PC2
salt  0.9519234 -0.3063363
umami 0.9519234  0.3063363
```

Cercle des corrélations en pD

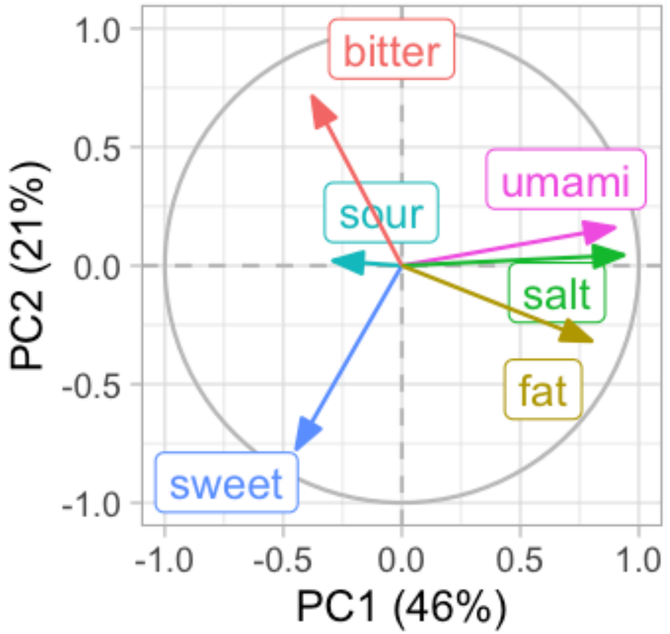
```
> cor(t, pca_tastes$x)
              PC1         PC2           PC3         PC4         PC5         PC6
sweet  -0.4448449 -0.77345800  0.0002141907 -0.39029816  0.22179440 -0.04848000
sour   -0.2900437  0.02078561 -0.9560593562  0.01046957 -0.03436360 -0.01013062
bitter -0.3775682  0.71391585  0.0388450188 -0.58725613  0.03231013 -0.01854063
umami   0.8976746  0.16054381 -0.1484182933 -0.01788606  0.32933176  0.19390399
salt    0.9357987  0.04458061 -0.0846034478 -0.07056367  0.04567312 -0.32874049
fat     0.8003419 -0.31776761 -0.0626406982 -0.38761622 -0.29671965  0.12752987
```

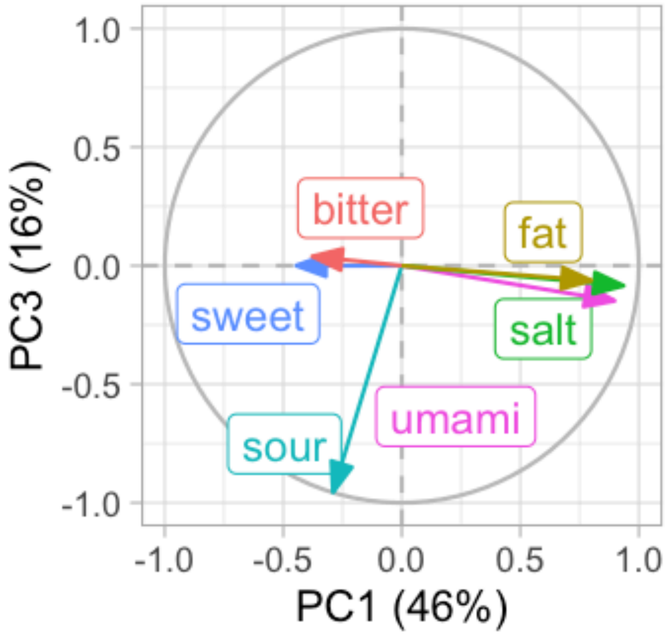
Corrélations entre les 6 variables originales et la première composante principale.

Les flèches sont les projections de vecteurs en 6D dans les plans 2D

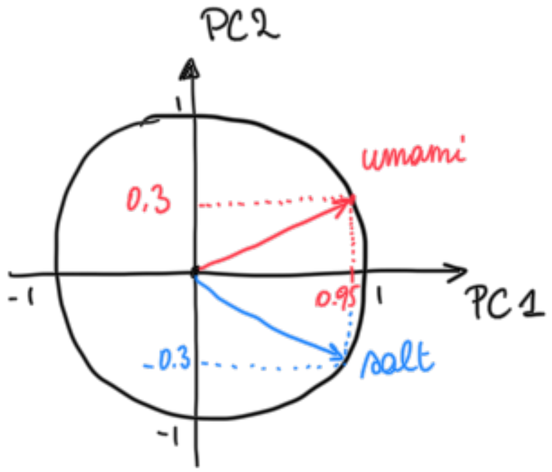
Toutes les flèches sont de longueur 1 en 6D (touchent les contours de la “sphère” 6D)

Puisque les 3eres CPs expliquent + de 80% de variance totale, nous pouvons représenter les corrélations avec CP1-CP3



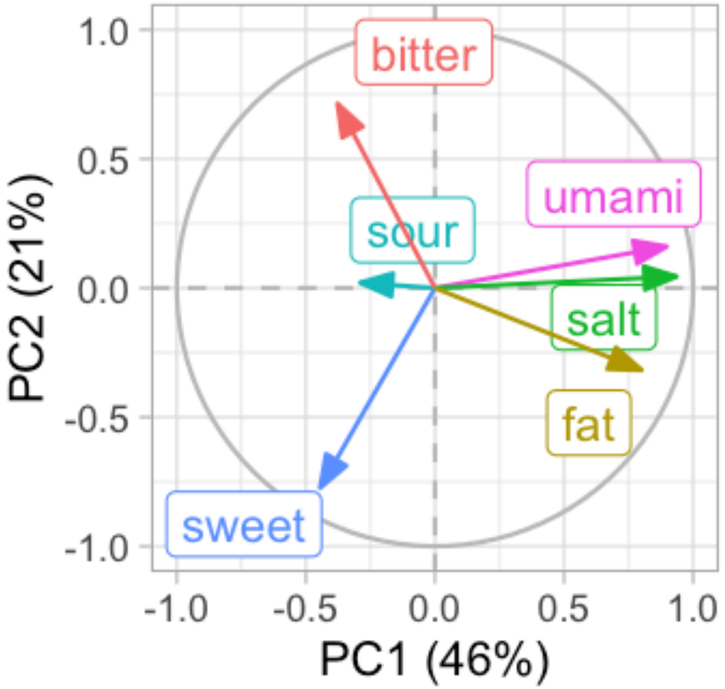


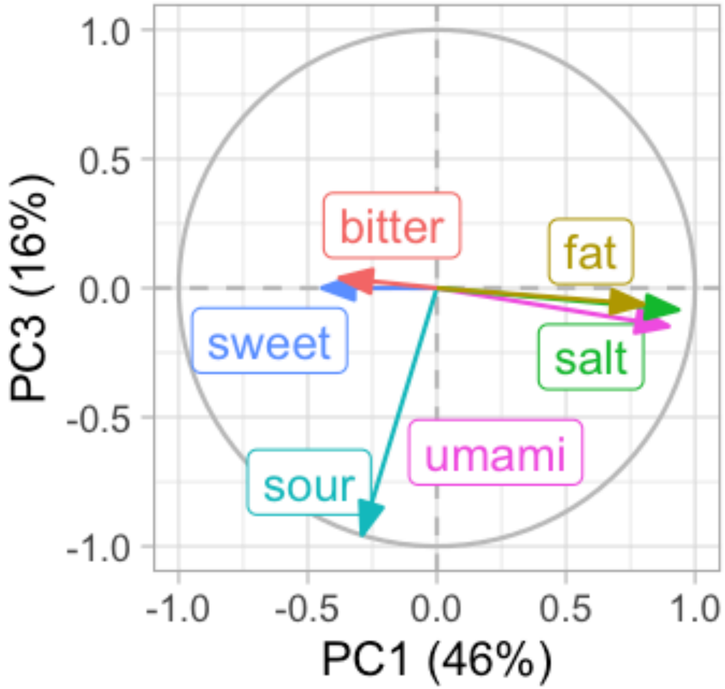
Cercle des corrélations: observations



En 2D, les deux “flèches” avaient une longueur unitaire (toute la variation selon les variables originales étaient expliquée par les 2 CPs).

En pD, la plupart des flèches ont une longueur <1 en 2D.
Plus la flèche est longue (touche le cercle), mieux cette variable est représentée dans le plan des CPs considérées.





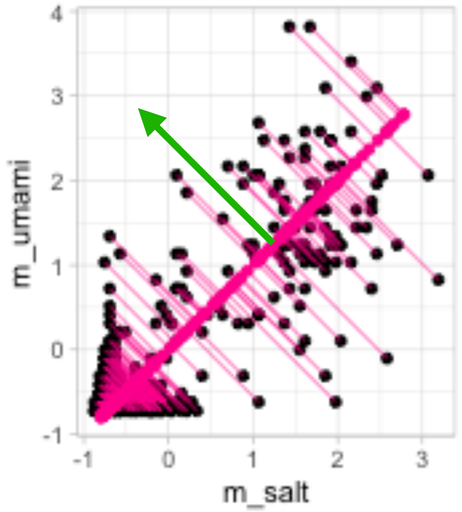
Les saveurs *salée* et *umami* sont fortement corrélées avec la 1ere CP.

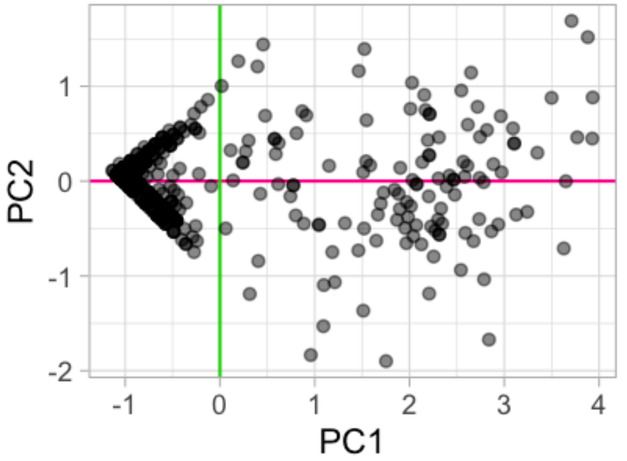
Les saveurs *sucrée* et *amère* sont majoritairement (anti-)corrélées avec la 2eme CP.

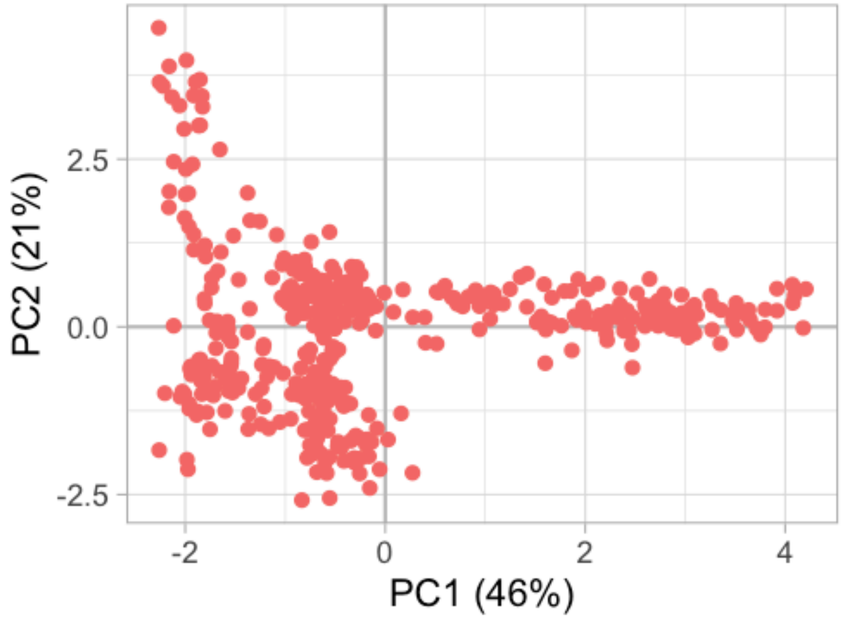
La saveur *acide* est fortement corrélée avec la 3eme CP et n’est bien représentée que dans le plan CP1 - CP3.

Représentation des échantillons dans l'espace des CPs

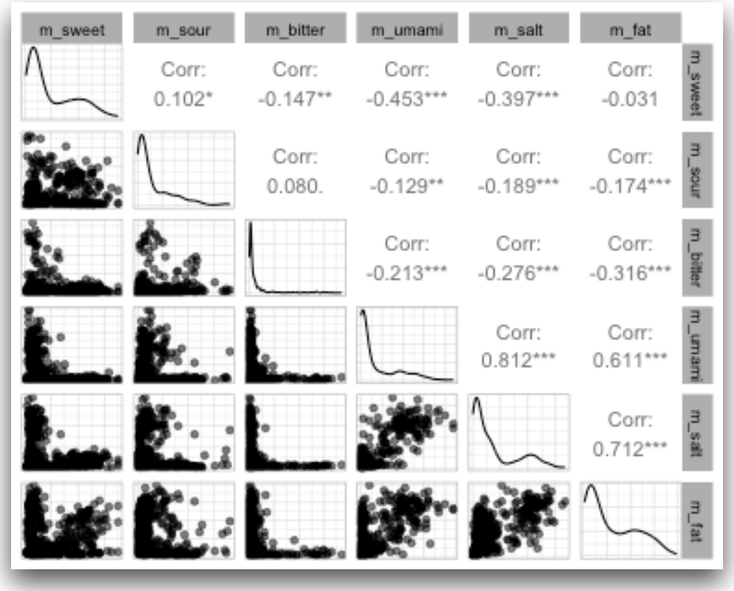
En 2D:



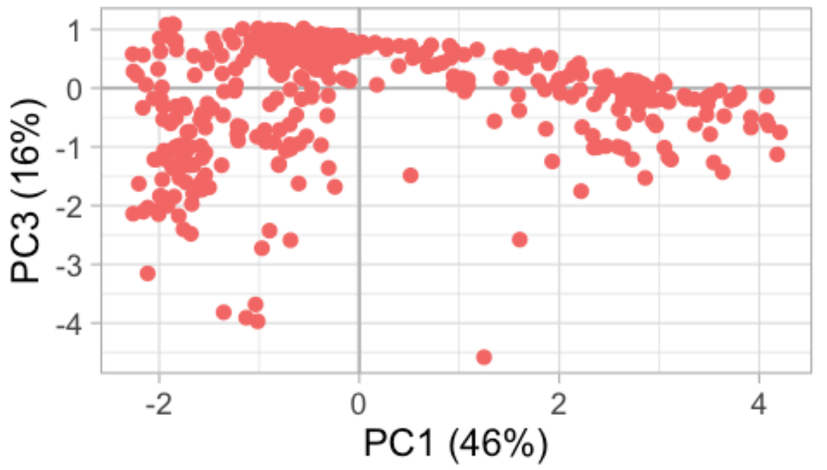




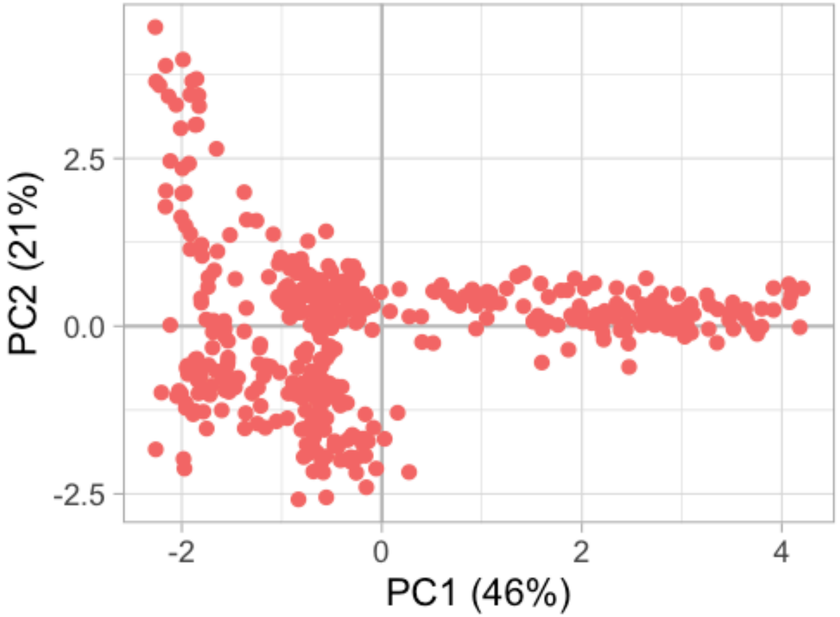
En pD:

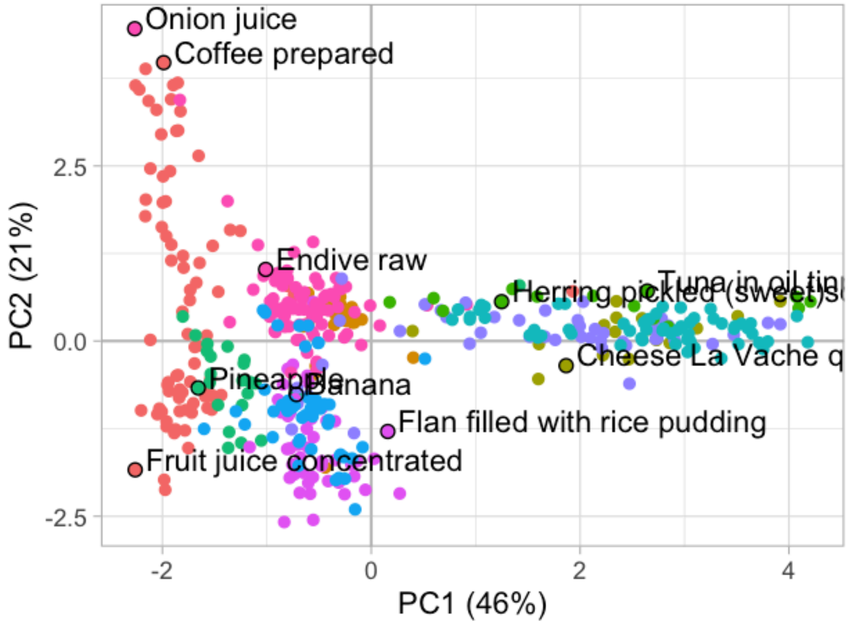


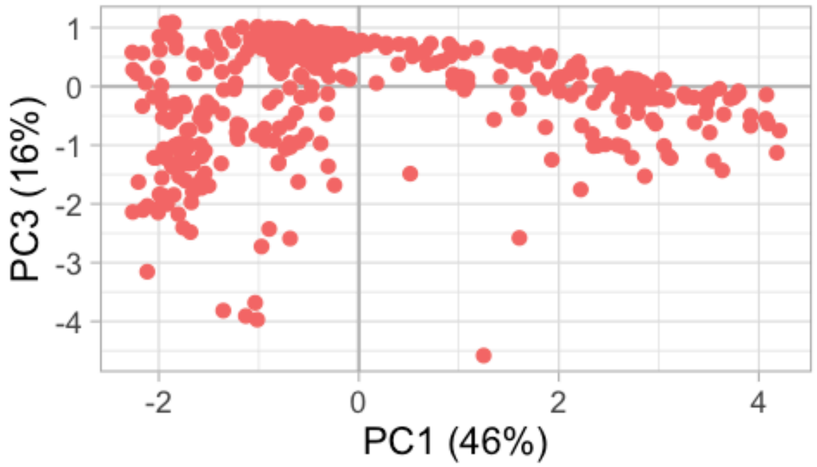
On garde 3 CPs
(80+% de variance)

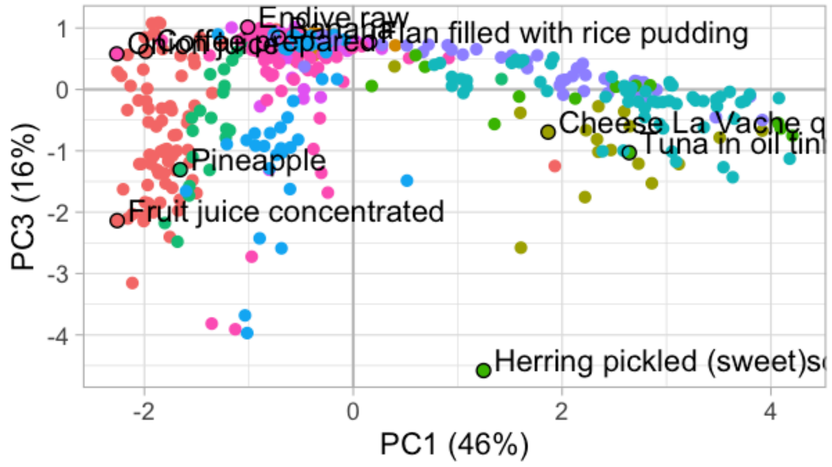


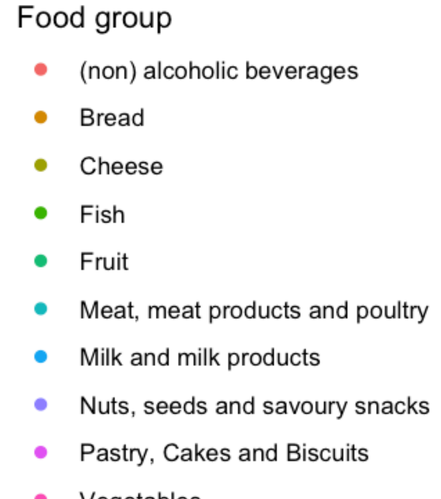
Représentation des échantillons dans l'espace des CPs





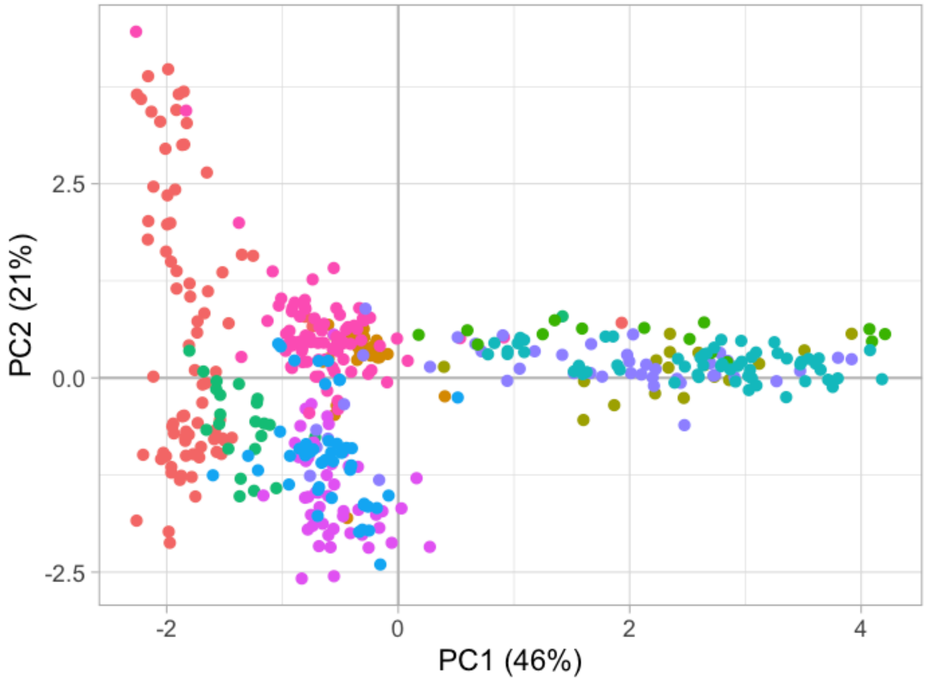


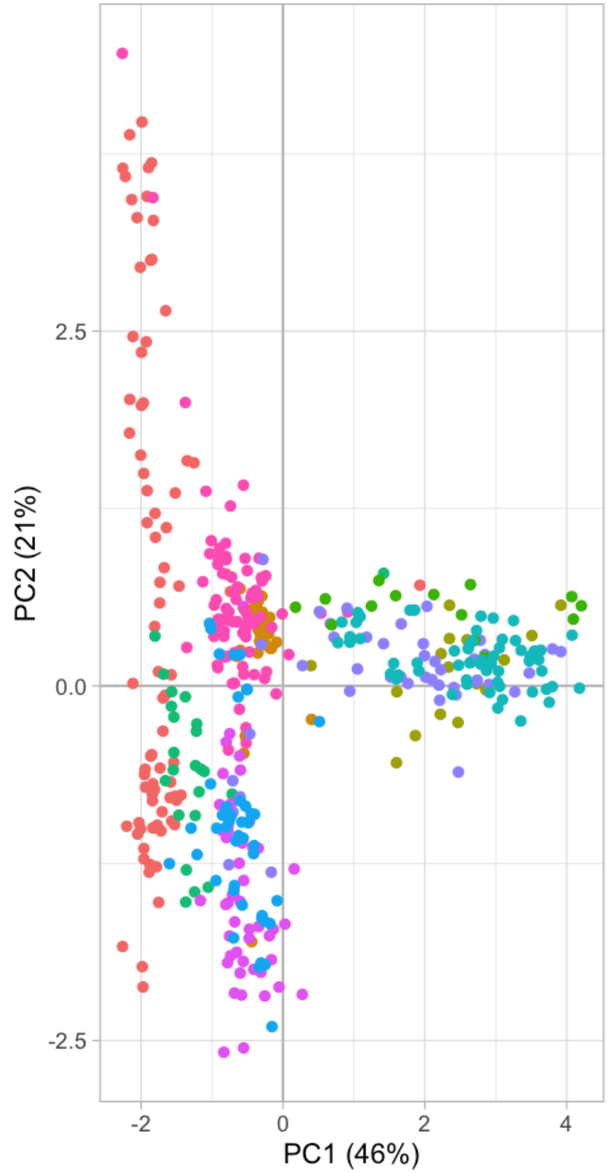




Représentation des échantillons dans l'espace des CPs

Quelle version de ces projections dans CP1-CP2 préférez-vous?

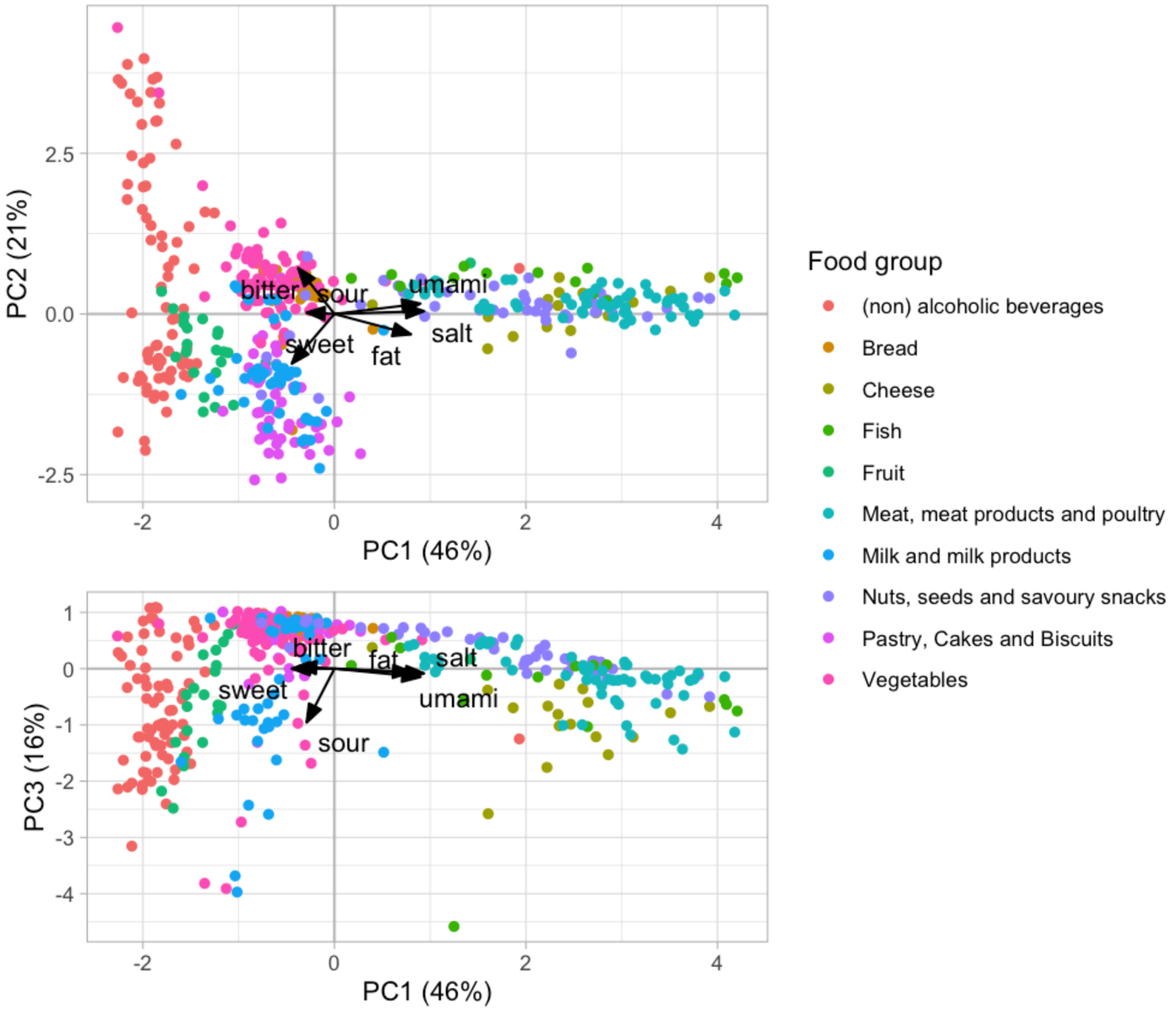




Pour donner une représentation "fidèle" de la variance expliquée par chaque CP, il est préférable de contraindre l'aspect des axes à 1:1 ou à $\lambda_1 : \lambda_2$

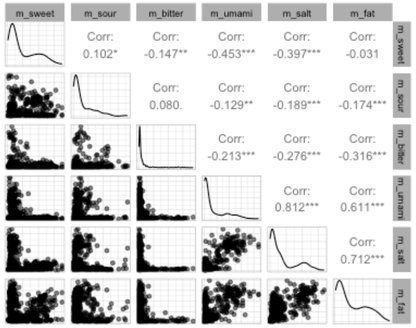
“Biplots”

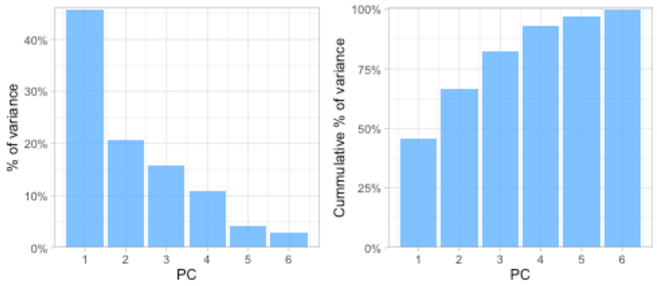
Il est parfois utile pour se “rappeler” de l’interprétation des axes principaux de représenter les projections des n échantillons et des p variables sur le même graphique.

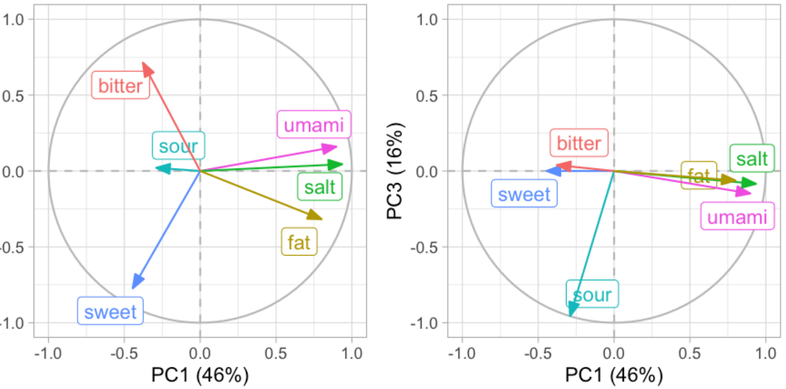


Comparaison avec des données simulées

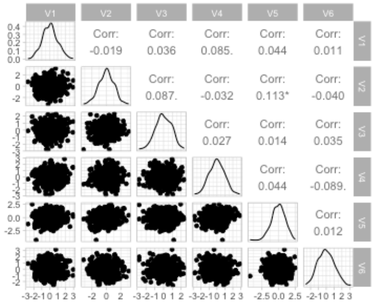
Les données de saveurs

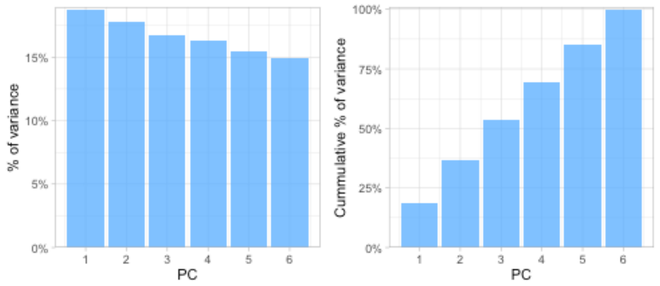


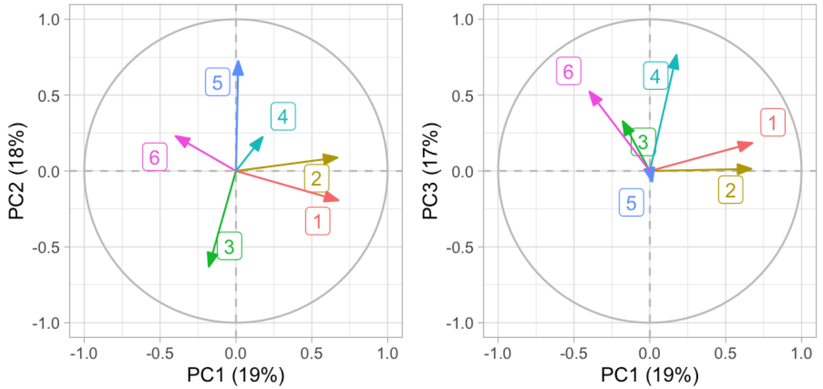




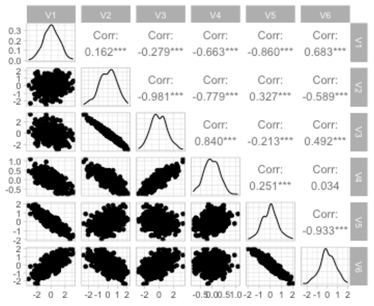
Données simulées sans “structure latente” (toutes les variables sont indépendantes)

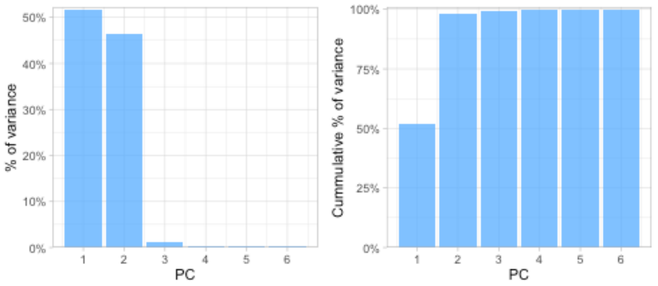


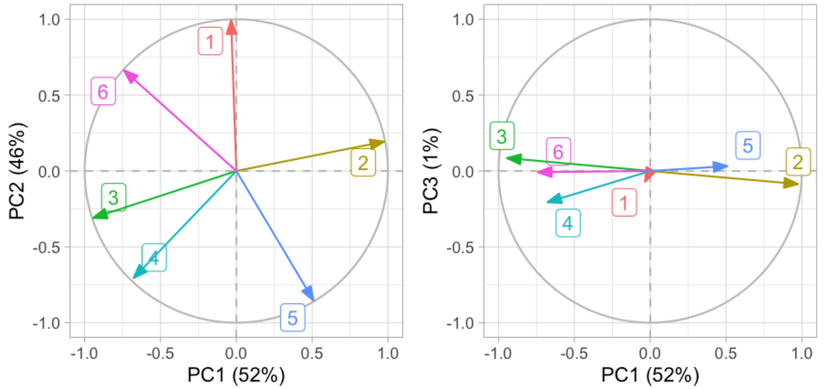




Données simulées avec “structure latente” (les variables dépendent de deux V.A. indép.)







Résumé

L’analyse en composantes principales est une méthode d’exploration de jeux de données multivariés.

Elle est particulièrement utile **s’il existe des corrélations entre les variables**.
Dans ce cas, il est probable que certaines variables soient redondantes et que les échantillons soient dispersé autour d’un sous-espace de dimension moindre ($k < p$)

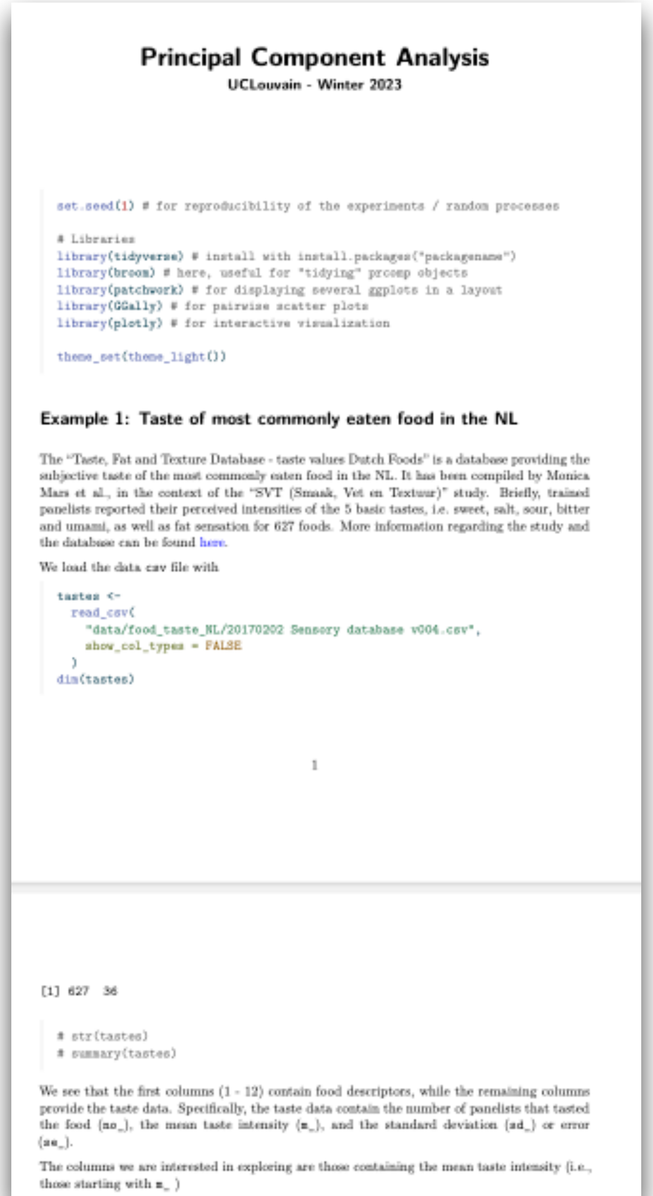
La PCA identifie un nouveau système d’axes orthogonaux (**les composantes principales**) qui, de manière décroissante/itérative, sont alignés avec les directions de variation maximale des données.

C’est une méthode **linéaire** (basée sur la covariance qui mesure des relations linéaires entre 2 VA)

Exercices pour la prochaine classe

- Télécharger le fichier de code (PCA_1.Qmd)
- Exécuter et essayer de comprendre le code “chunck-by-chunk”.
- Poster vos questions éventuelles sur le forum du cours.
- Lire la documentation de la fonction `prcomp`
- Vérifier que vous pouvez “render” le fichier au format pdf et/ou html (en vue de la préparation au projet / exercices notés)
- Répondre aux exercices à la fin du fichier de code

- In the "Food taste" dataset, how do you interpret that the first 3 PCs are well aligned with 5 of the 6 flavors? What does it say about our taste perception?
- Increase the variance of the random noise in the second simulation (the one with two latent variables). What do you observe?
- Compare the coefficients for the linear combination of the latent variables with the corrleation (circle of correlation), and with the PC1 and PC2 rotations. What do you observe? What makes them different or similar? How could we make the same?



Principal Component Analysis

UCLouvain - Winter 2023

```
set.seed(1) # for reproducibility of the experiments / random processes

# Libraries
library(tidyverse) # install with install.packages("packagename")
library(broom) # here, useful for "tidying" prcomp objects
library(patchwork) # for displaying several ggplots in a layout
library(GGally) # for pairwise scatter plots
library(plotly) # for interactive visualization

theme_set(theme_light())
```

Example 1: Taste of most commonly eaten food in the NL

The “Taste, Fat and Texture Database - taste values Dutch Foods” is a database providing the subjective taste of the most commonly eaten food in the NL. It has been compiled by Monica Mars et al., in the context of the “SVT (Smaak, Vet en Textuur)” study. Briefly, trained panelists reported their perceived intensities of the 5 basic tastes, i.e. sweet, salt, sour, bitter and umami, as well as fat sensation for 627 foods. More information regarding the study and the database can be found here.

We load the data csv file with

```
tastes <-
  read_csv(
    "data/food_taste_NL/20170202 Sensory database v004.csv",
    show_col_types = FALSE
  )
dim(tastes)
```

1

```
[1] 627  36
```

```
# str(tastes)
# summary(tastes)
```

We see that the first columns (1 - 12) contain food descriptors, while the remaining columns provide the taste data. Specifically, the taste data contain the number of panelists that tasted the food (`no_`), the mean taste intensity (`m_`), and the standard deviation (`sd_`) or error (`se_`).

The columns we are interested in exploring are those containing the mean taste intensity (i.e., those starting with `m_`)

Références additionnelles

Linear Algebra “refresher”

Chapter 7 of Strang, Gilbert. 2009. *Introduction to Linear Algebra*. Fourth. Wellesley-Cambridge Press.

Principal Component Analysis (and multivariate methods)

Mardia, Kanti, John T Kent, and John M Bibby. 1979. *Multiariate Analysis*. New York: Academic Press.

Jolliffe, Ian. 2002. *Principal Component Analysis*. Wiley Online Library.

With applications to biology

Chapter 7 of “Modern Statistics for Modern Biology”, Holmes and Hubert, 2019 (or online)
https://web.stanford.edu/class/bios221/book/07-chap.html