Démarche statistique

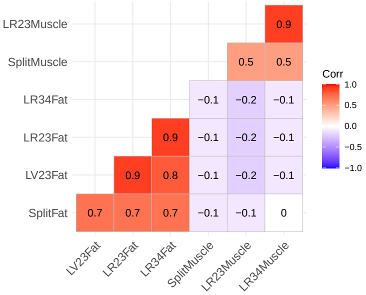
Session 7 - Réduction de données pour la visualisation

David Causeur
Institut Agro Rennes Angers
IRMAR UMR 6625 CNRS

12 janvier, 2025

Objectif : visualiser de manère simplifiée des profils décrits pas plusieurs variables (quantitatives)

Illustration : pour un échantillon de 354 carcasses de porcs, on dispose d'un profil de 7 mesures d'épaisseurs de gras et de muscle en differents sites anatomiques



Objectif : identifier des scores permettant une description simplifiée des profils complets.

Par exemple, au vu de la structure de corrélation ci-dessus,

- Un score synthétisant les épaisseurs de gras
- Un score synthétisant les épaisseurs de muscle

Réduction de la dimension des données

Hypothèse : Chaque profil $(x_1, \ldots, x_p)$ de p variables peut être approché de manière satisfaisante par $q \leq p$ **scores** quantitatifs $z_j, \ j = 1, \ldots, q$

- les scores z_j sont dits **latents** car ils ne sont pas mesurés directement sur chaque individu.
- les scores z_j sont calculés à partir des variables $(x_1, \ldots, x_p)$ sans a priori sur les contributions de chaque variable

Un type particulier de score : la composante principale

- **Synthèse par combinaison linéaire** des variables centrées-réduites :

$$z_1 = a_1 \frac{x_1 - \bar{x}_1}{s_1} + \ldots a_p \frac{x_p - \bar{x}_p}{s_p}$$

- **Choix des coefficients** a_k : la variance de z_1 est **la plus grande possible** parmi toutes les combinaisons linéaires (en imposant $a_1^2 + \ldots + a_p^2 = 1$).

Remarques :

- les coefficients a_k s'appellent les **loadings** (en anglais), ou **coordonnées des variables** (en français).
- la variance de z_1 s'appelle l'**inertie** $I(z_1)$ de la composante principale

Illustration pour deux variables fortement corrélées

Illustration : Composante principale pour synthétiser deux épaisseurs de gras x_1 et x_2 fortement corrélées

z_1 définit un nouvel axe (**axe principal**) synthétique de x_1 et x_2 : si la valeur de z_1 est grande, les valeurs de x_1 et x_2 sont aussi élevées (et inversement)

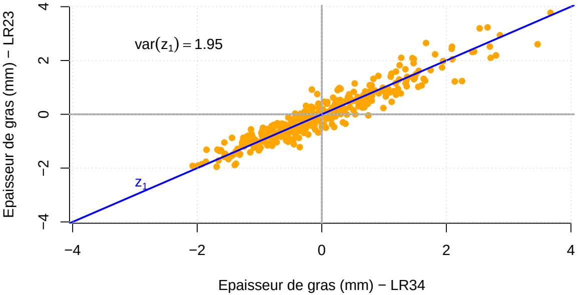
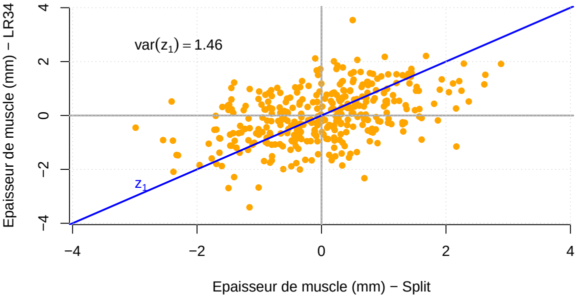


Illustration pour deux variables modérément corrélées

Illustration : Composante principale pour synthétiser deux épaisseurs de muscle x_1 et x_2 modérément corrélées

z_1 synthétise mal (x_1, x_2) : une valeur élevée de z_1 peut par exemple être associée à une valeur forte de x_1 et modérée de x_2



Inertie de la composante principale : $\mathrm{I}(z_1) = \mathrm{Var}(z_1)$

- $0 \leq \mathrm{I}(z_1) \leq p$
- Si $(x_1, \ldots, x_p)$ sont parfaitement corrélées, $\mathrm{Var}(z_1) = \mathrm{I}_{\max}(z_1) = p$

Inertie d'une composante principale

Ilustration (pour deux variables fortement corrélées)

```
pca <- PCA(dta[,c("LR23Fat","LR34Fat")],graph=FALSE)
round(pca$eig,3)
```

```
       eigenvalue percentage of variance cumulative percentage of variance
comp 1      1.949                 97.457                            97.457
comp 2      0.051                  2.543                           100.000
```

Mesure de la capacité de synthèse d'une composante principale

$$100\,\frac{\mathrm{I}(z_1)}{I_{\max(z_1)}} = 100\,\frac{\mathrm{I}(z_1)}{p}$$

Inertie non-expliquée par z_1 : construction d'une deuxième composante principale z_2

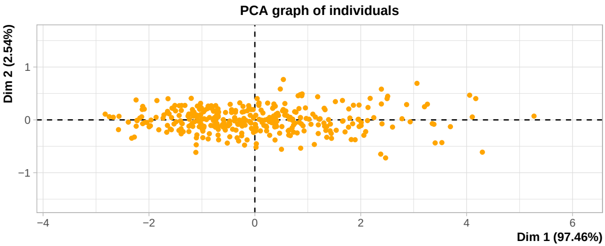
- z_2 est une combinaison linéaire des variables explicatives centrées-réduites
- La variance de z_2 est la plus grande possible
- **La corrélation entre z_1 et z_2 est nulle**

(z_1, z_2) synthétisent parfaitement (x_1, x_2) : inertie cumulée de z_1 et $z_2 = 100\%$

Analyse en Composantes Principales (ACP) de deux variables

Analyse en composantes principales : représentation des individus

```
plot(pca,choix="ind",pch=16,col.ind="orange",label="none")
```



Dans le cas $p = 2$, la représentation (z_1, z_2) est obtenue par rotation de la représentation (x_1, x_2)

Interprétation des composantes principales par les scores extrêmes

Identification des individus extrêmes sur le 1er axe principal

```r
z1 <- pca$ind$coord[,1]
select_high <- which(z1>4)
select_low <- which(z1< -2.5)
cbind(dta[,c(1,2,6,8,10)],PC1=z1)[c(select_high,select_low),]
```

```
    GENOTYPE SEX LR23Fat LR34Fat      LMP       PC1
53        P0   M  24.260  25.300 70.20833  5.272338
62       P25   M  22.435  21.620 74.75886  4.056788
142       P0   M  20.570  24.630 75.11300  4.296818
143       P0   M  21.630  22.675 73.25228  4.106929
163      P25   M  22.550  22.030 73.37780  4.172376
29        P0   F   6.220   7.960 87.72287 -2.575837
121       P0   F   6.315   6.715 87.97819 -2.827052
211      P25   F   6.380   7.005 87.06073 -2.748965
322      P50   F   6.525   7.200 88.14694 -2.673707
346      P50   F   6.805   7.405 87.23998 -2.565935
```

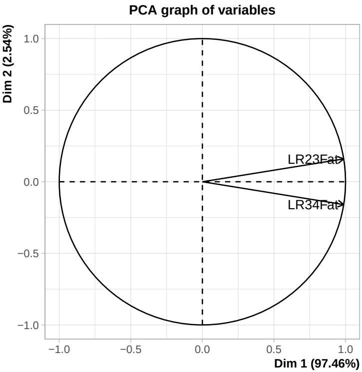
La 1ère composante principale oppose les carcasses maigres (à gauche) aux carcasses grasses (à droite)

Interprétation des composantes principales par le cercle des corrélations

Cercle des corrélations : représentation des variables x_j dans le plan formé par les composantes principales

- Chaque x_j est représentée par une flèche de longeur 1 partant de l'origine du plan
- Les coordonnées de l'extrémité de la flèche sont la corrélation avec z_1 (abscisse) et la corrélation avec z_2 (ordonnée)

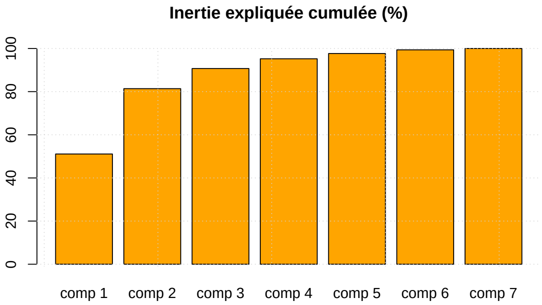
```
plot(pca,choix="var")
```



Analyse en Composantes Principales (ACP) de p variables

Illustration : Deux composantes principales synthétisent 81% de la variabilité des épaisseurs de gras et de muscle

```
pca <- PCA(dta[,3:9],graph=FALSE)
barplot(pca$eig[,3],col="orange",main="Inertie expliquée cumulée (%)")
grid()
```

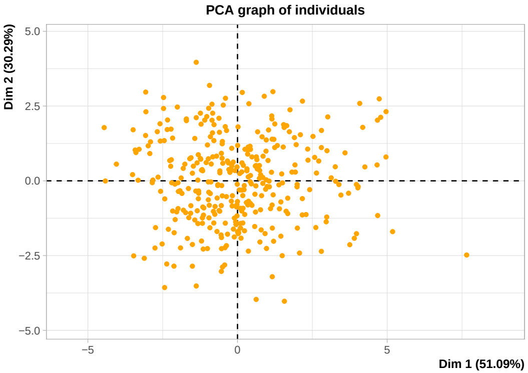


Remarque : l'inertie des composantes principales de rang ≥ 3 est plus petite que 1, soit plus petite que la variance de chaque x_j (centrée-réduite)

Plan factoriel

Plan factoriel : espace de représentation défini par deux composantes principales

```
plot(pca,choix="ind",pch=16,col.ind="orange",label="none",axes=c(1,2))
```



Attention : chaque point (z_1, z_2), représentant un individu, est la projection de son profil complet $(x_1, \ldots, x_p)$ dans le plan factoriel

Qualité de représentation d'un individu dans le plan factoriel

Illustration : un individu a une valeur anormalement élevée de z_1

```
outlier <- which.max(pca$ind$coord[,1])
dta[outlier,-(1:2)]
```

```
   SplitFat SplitMuscle LV23Fat LR23Fat LR23Muscle LR34Fat LR34Muscle      LMP
53    23.08      66.625  28.055   24.26     46.935    25.3      44.43 70.20833
```

Peut-on considérer que sa position dans le plan (z_1, z_2) est conforme à la réalité de ce profil ?

Mesure de la qualité de représentation d'un individu

- $M^* = (z_1, z_2)$ est la projection sur le plan factoriel du point $M = (x_1, \ldots, x_p)$ correspondant au profil complet
- si M^* est proche de M, le profil complet est bien représenté dans le plan (z_1, z_2)
- Critère de qualité : si la mesure de l'angle θ entre $\overrightarrow{OM^*}$ et $\overrightarrow{OM}$ est proche de 0, alors M^* est proche de M

```
pca$ind$cos2[outlier,]
```

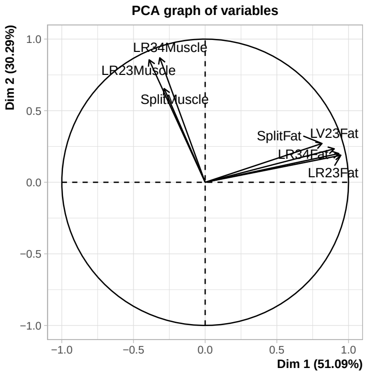
```
      Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
0.876333556 0.092012322 0.028197696 0.001559481 0.001047756
```

Pour cet individu, $\cos^2\theta = 0.968$, donc θ est très proche de 0

Interprétation des composantes principales

Illustration : Deux composantes principales synthétisent 81% de la variabilité des épaisseurs de gras et de muscle

```r
plot(pca,choix="var")
```



- l'axe 1 oppose les carcasses ayant des épaisseurs de gras élevées (droite) à celles ayant des épaisseurs de gras faibles (gauche)
- l'axe 2 oppose les carcasses ayant des épaisseurs de muscle en LR34 et LR23 élevées (droite) à celles ayant des épaisseurs de muscle en LR34 et LR23 faibles (bas)

Qualité de représentation d'une variable dans le plan factoriel

Illustration : l'épaisseur de muscle SplitMuscle est représentée par une flèche de longueur sensiblement inférieure à 1

```
round(pca$var$coord["SplitMuscle",],digits=3)
```

```
 Dim.1  Dim.2  Dim.3  Dim.4  Dim.5
-0.286  0.655  0.691  0.107  0.024
```

Mesure de la qualité de représentation d'une variable

- Pour chaque variable x_j, la flèche visible dans le plan factoriel est la projection de la flèche de longueur 1 dans l'espace formé par toutes les composantes principales
- Critère de qualité : si la mesure de l'angle θ entre la flèche dans l'espace formé par toutes les composantes principales et la flèche projetée est proche de 0, alors la variable est bien représentée par la flèche projetée

```
pca$var$cos2["SplitMuscle",]
```

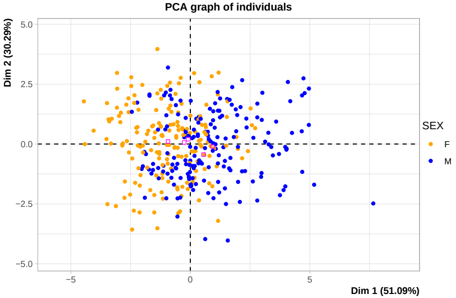
```
       Dim.1        Dim.2        Dim.3        Dim.4        Dim.5
0.0818682295 0.4285098242 0.4769829055 0.0115231106 0.0005940454
```

La variable SplitMuscle est mieux représentée dans le plan formé par (z_2, z_3)

Interprétation à l'aide de variables catégorielles supplémentaires

Illustration : représentation de groupes dans un plan factoriel

```
pca <- PCA(dta,quali.sup=1:2,quanti.sup=10,graph=FALSE)
plot(pca,choix="ind",label="none",col.hab=c("orange","blue"),habillage=2)
```



Lien entre chaque composante principale et les variables catégorielles supplémentaires (R^2)

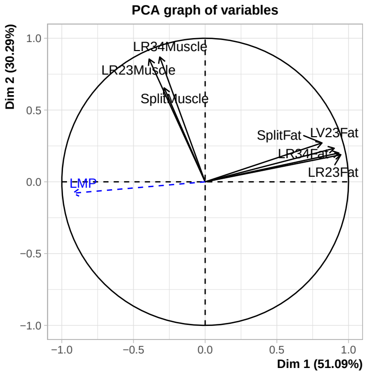
```
round(pca$quali.sup$eta2,3)
```

```
         Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
GENOTYPE 0.028 0.030 0.037 0.027 0.009
SEX      0.242 0.005 0.007 0.014 0.003
```

Interprétation à l'aide de variables quantitatives supplémentaires

Illustration : représentation d'une variable quantitative dans un plan factoriel

```
plot(pca,choix="var")
```



Lien entre chaque composante principale et la variable quantitative supplémentaire

```
round(pca$quanti.sup$cor,3)
```

```
     Dim.1  Dim.2 Dim.3 Dim.4 Dim.5
LMP -0.923 -0.079 0.081 0.002 0.028
```

Ce qu'il faut retenir

L'Analyse en Composantes Principales (ACP) permet de décrire des profils multivariés de données par l'extraction de scores synthétiques (les composantes principales)

- **Mesure du niveau de dépendance entre les variables** : la dépendance est d'autant plus forte que le nombre de composantes principales synthétisant les profils complets est limité
- **Exploration de la structure de dépendance entre les variables** : identification de groupes de variables corrélées entre elles
- **Analyse simplifiée et description de la répartition des individus** : selon des gradients le long des composantes principales, ou par agrégats dans les plans factoriels

Attention

- L'ACP ne vise pas à expliquer la relation entre une variable à expliquer et des variables explicatives
  - Voir modèle linéaire
- L'ACP ne vise pas à décrire des différences entre groupes de données
  - Voir analyse discriminante, modèle linéaire généralisé