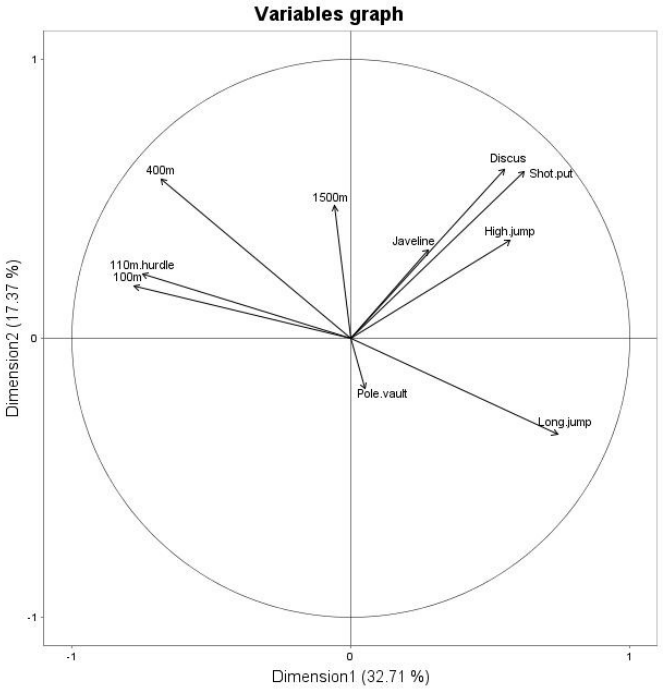
# Analyse de données

Les dernières heures

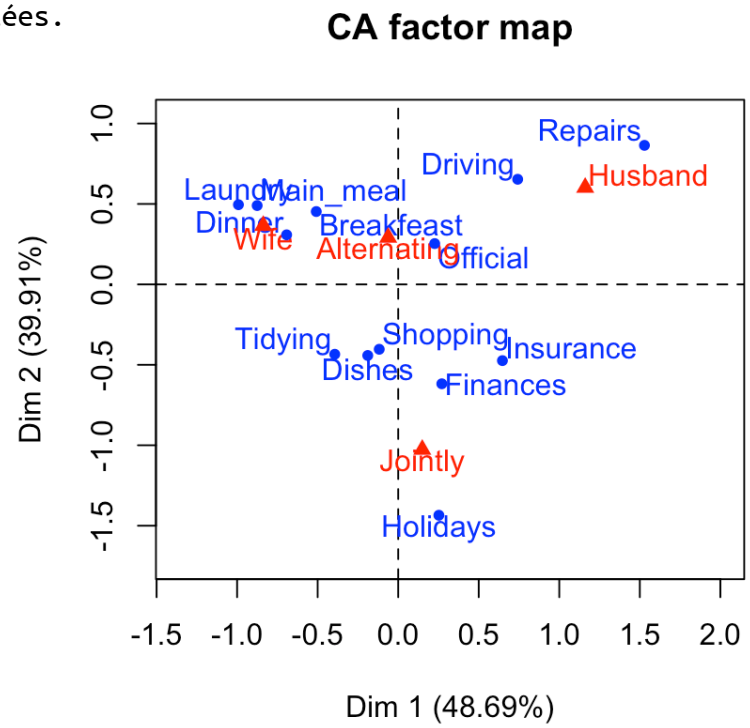
# I ACP – Analyse en Composantes Principales

- Composantes principales = combinaisons linéaires des variables de départ.
- Maximise la variance totale des individus dans un espace de plus faible dimension.
- Les composantes obtenues sont décorrélées.
- Réalisable uniquement si n ≥ p (n = individus, p = variables).
- Différence de longueur entre deux flèches = écart-type.
- Projette les données sur un sous-espace.
- On regarde les proximités entre modalités.
- Le cos est la corrélation entre la variable et la composante.



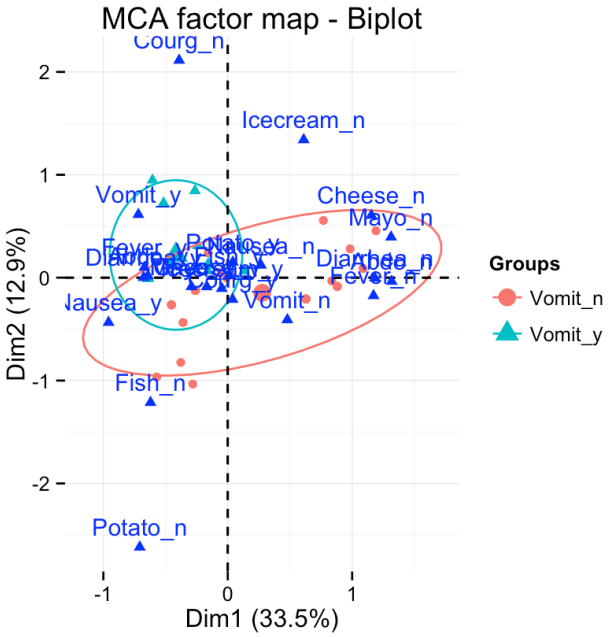

# II AFC – Analyse Factorielle des Correspondances

- Prend en entrée un tableau de contingence.
- Sortie des lignes par rapport aux colonnes.
- C'est une double ACP des profils ligne/colonne.
- Étudier les liaisons entre modalités.
- 2 modalités proches sont liées.



# III AFCM – Analyse Factorielle des Correspondances Multiples

- Analyser les proximités entre les catégories de variables qualitatives et les observations.
- Sur 2 var → valeurs propes ≠ et sans signification statistique.
- S'applique uniquement pour 2 ou plus variables quantitatives.
- Donne des classes à interpréter.
- Si effectué avec un tableau disjonctif complet alors les modalités de toutes les variables et des individus sont représentées.
- Éventuellement on peut interpréter les axes grâce aux contributions des modalités dessus.

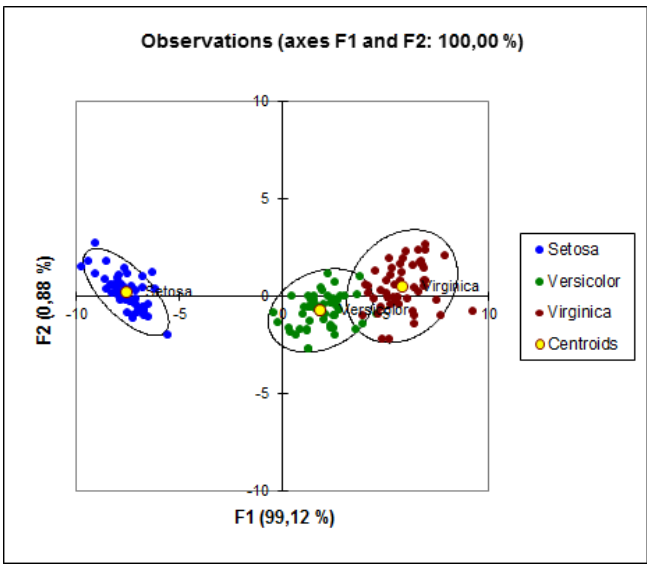



*C LE VOMI LOL*

- On peut transformer les variables qualitatives en variables quantitatives grâce à l'AFCM pour faire une classification.

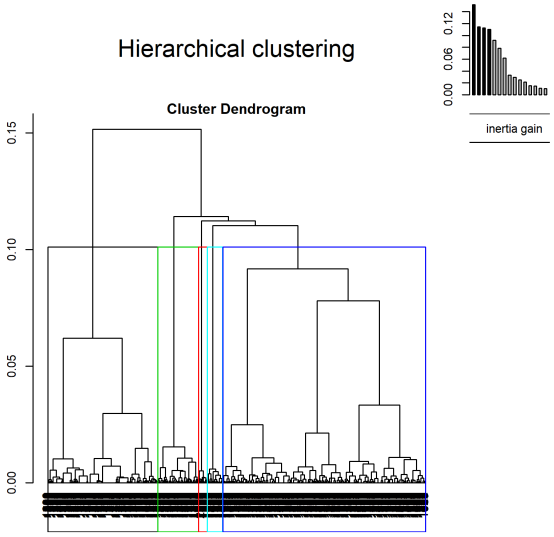
# IV AFD – Analyse Factorielle Discriminante

- Méthode explicative et descriptive.
- Pour p > 2 variables quantitatives et 1 var qualitative (à droite : uniquement des espèces).
- AFC particulière minimisant la variance intraclasse (et maximisant la variance interclasse, donc).
- Choix de dimension = en fonction du nombre de groupes.
- Prédit l'appartenance d'un individu à une classe.



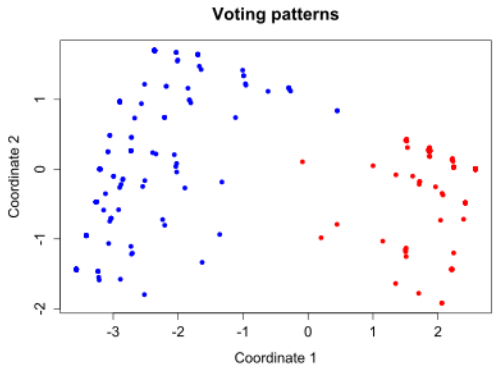
# V CAH – Classification Ascendante Hiérarchique

- Utile pour choisir un nombre de classes.
- Mieux que le Kmeans si nombre de variables inconnu.
- CAH & Kmeans sont itératives.
- Pour n individus et p variables, la matrice sera de taille n*n.
- Classification sur données brutes.
- *Le petit truc en haut à droite c'est la variance intraclasse.*




# VI MDS – Positionnement Multidimensionnel

- Représente les individus ou les variables séparément.
- Peut s'appliquer avec une matrice correspondant aux distances entre variables.
- *À droite: votes à gauche/droite.*




# VII Vrac

- Dans R, **pair(data)** ⇒ compare 2 à 2 les échantillons. S'il y a des représentations à peu près linéaires dans un croisement alors ⇒ fortement corrélées. ⇒ Matrice de variance/covariance.
- En **class**ification, l'objectif est de rassembler les individus en **class**es, c'est-à-dire de minimiser la variance interclasse.
- **P-VALUE > 0.05 ⇒ ACCEPTATION DE L'HYPOTHÈSE D'INDÉPENDANCE**
- La variance peut servir à calculer la dispersion d'une série de valeurs quantitatives.
- Rapport de Corrélation ⇒ pour étudier la liaison entre une variable qualitative et une variable quantitative.
- Modalité = type de [var qualitative]. *Ex : Le sexe. Mâle & femelle sont des modalités.*
- Le boxplot des coordonnées des individus sur chaque composante c'est le diagramme avec plein de boîtes à moustache verticales agencées côte-à-côte.