Mesures de distance

Données mixtes

Véronique Tremblay

Plusieurs types de variables

Procédure en présence de plusieurs types de variables

1. Recoder les variables ordinales

2. Déterminer le poids de chaque variable

3. Identifier les variables asymétriques

4. Utiliser l'indice de Gower (1971)

Indice de Gower

C'est un indice qui donne une sorte de moyenne pondérée de toutes les mesures de distance présentées précédemment.

Indice de dissemblance de Gower

$$G(i,j) = \frac{\sum_{k=1}^{K} w_k \gamma_k(i,j) d_k^*(i,j)}{\sum_{k=1}^{K} w_k \gamma_k(i,j)},$$

où w_k est un poids accordé à la variable k

- variable k numérique ou ordinale : $\gamma_k(i,j) = 1$ et $d_k^*(i,j) = |x_{ik} - x_{jk}|/r_k$;
- variable k nominale symétrique : $\gamma_k(i,j) = 1$ et $d_k^*(i,j) = I(x_{ik} \neq x_{jk})$;
- variable k nominale asymétrique : $\gamma_k(i,j) = \{1 - (1 - x_{ik})(1 - x_{jk})\}$ et $d_k^*(i,j) = I(x_{ik} \neq x_{jk})$.

En pratique

Créer la matrice de distance en R

On utilise la fonction `daisy` de la librairie `cluster` (Maechler et al. (2019)).

Préparer les données

```r
musique <- read_delim("donnees/musique2.csv", ";")

musique <- musique %>%
  mutate(genre = as.factor(genre),
         freq_radio = factor(freq_radio,
                             ordered = TRUE), # Variable ordinale
         musicien = as.factor(musicien),
         transport = as.factor(transport))
```

```r
var_dis <- musique %>%
  dplyr::select(-id) %>% # Retirer l'identifiant
  as.data.frame() # Transformer en data.frame
```

Créer des poids

```r
# On créé un data.fame de 1
poids <- as.data.frame(t(rep(1,ncol(var_dis ))))
colnames(poids) <- colnames(var_dis )

# On associe le poids voulu aux variables
poids$m2 <- 2
```

Construire la matrice de dissemblance

```
d <- daisy(var_dis , # Le jeu de données (un data.frame)
           metric = 'gower', # La mesure de distance choisie
           type = list(asymm = 'musicien'), # Identifier les
                                    # variables asymétriques
           weights = poids[1,] # Poids des variables
           )
```

Résumé

- Dissemblance de Gower
- Autre façons de gérer la donnée mixte

Références

Gower, John C. 1971. "A General Coefficient of Similarity and Some of Its Properties." *Biometrics*, 857–71.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2019. *Cluster: Cluster Analysis Basics and Extensions*.