Nombre de groupes

Choix du nombre de groupes

Véronique Tremblay

Stratégie pour le choix du nombre de groupes

Pour la segmentation de clientèle

1. Regarder le dendogramme
2. Couper à un endroit qui a du sens
3. Analyser les segments
4. Modifier les groupes

Les autres indicateurs

Il y a plusieurs autres indicateurs[1].

La librairie `NbClust`[2] en contient une trentaine.

[1] Surtout si les variables sont continues

[2] Pour variables continues

```
     KL CH Hartigan CCC Scott Marriot
[1,]  6  6        3   6     3       3

     TrCovW TraceW Friedman Rubin Cindex DB
[1,]      3      3        5     6      5  7

     Silhouette Duda PseudoT2 Beale Ratkowsky Ball
[1,]          6    6        6     6         3    3

     PtBiserial Frey McClain Dunn Hubert SDindex
[1,]          6    1       2    6      0       5
```

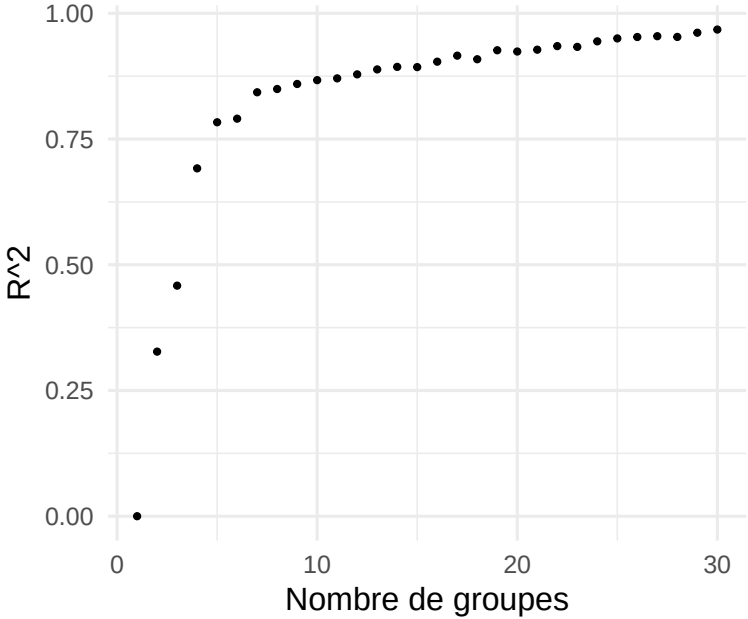
Les indicateurs basés sur l'inertie

$$I_{totale} = I_{intra-groupe} + I_{inter-groupe}$$

- Ces indicateurs sont plus pertinents avec des variables continues
- Prendre garde au poids des variables et à la standardisation

Pseudo R^2

$$\text{Pseudo-}R^2 = \frac{I_{inter-groupe}}{I_{total}}$$



R^2 selon le nombre de groupes

La statistique de Caliliski - Harabasz (CH)

$$CH = \frac{I_{inter-groupe}/(k-1)}{I_{intra-groupe}/(n-k)}$$

Indice de Dunn

On maximise l'indice suivant:

$$D = \frac{\text{Distance minimale entre 2 groupes}}{\text{Distance maximale dans un groupe}}$$

L'indice de Dunn cherche donc à créer des groupes denses et bien séparés.

L'indice de Silhouette

La silhouette de l'observation i mesure la confiance dans le choix du groupe pour l'observation i:

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)}$$

où

- a_i est la distance moyenne entre l'observation i et les autres observations de son groupe
- b_i est la distance moyenne entre l'observation i et les observations du groupe le plus proche de i

On souhaite maximiser la silhouette moyenne des observations.

L'indice de Silhouette

```
s_moy <- silhouette(gr_m_moy,d)

mean(s_moy[,'sil_width'])
```

```
## [1] 0.4012718
```

```
s_ppv <- silhouette(gr_m_ppv,d)

mean(s_ppv[,'sil_width'])
```

```
## [1] 0.210907
```

```
s_vpd <- silhouette(gr_m_vpd,d)

mean(s_vpd[,'sil_width'])
```

```
## [1] 0.39881
```

Tests de stabilité

1. Tester différentes approches
2. Avec différents sous-échantillons
3. Vérifier si les groupes sont stables d'un essai à l'autre

```r
table(gr_m_moy, gr_m_ppv)
```

```
##          gr_m_ppv
## gr_m_moy  1  2  3  4  5  6
##        1 11  0  0  0  0  0
##        2  0 11  0  0  0  0
##        3  0  0  3  0  0  0
##        4  0  0 11  0  0  1
##        5  0  6  0  1  0  0
##        6  0  0  0  0  3  0
```

Quelques conseils pour le choix des groupes

- Privilégiez l'interprétabilité et l'utilité des groupes à un critère quelconque.
- Évitez les critères basés sur l'inertie ou la variance pour des groupes de taille et d'étendue inégales.

Attention

- Est-ce qu'un petit groupe est nécessairement inintéressant?

Pour les mélange de modèle de densités

Indices liés aux probabilités estimées

- la probabilité a posteriori d'appartenir à chaque classe, $\widehat{\gamma}_{ki}$
- l'indice d'entropie calculé à partir des $\widehat{\gamma}_{ki}$

Pour les modèles de mélange

AIC, BIC, Tests d'hypothèses