L'analyse de classes/profils latents

Partie 1 : Un survol conceptuel

Hans Ivers, Ph.D., Stat.ASSQ
hans.ivers@psy.ulaval.ca

Partenariat sur la Séparation parentale et Recomposition familiale
Université Laval

2 février 2022

Cette communication a été rendue possible grâce aux fonds du Conseil de recherches en sciences humaines (CRSH) du Canada.

PARTENARIAT DE RECHERCHE
Séparation parentale, recomposition familiale

Plan de la présentation

Pourquoi cette technique?

A partir d'un échantillon d'observations multivariées (chaque répondant étant évalué sur plusieurs variables), l'objectif est d'identifier des *sous-populations homogènes* de répondants (qu'on appellera "classes" ou "profils").

On désire donc ici regrouper des observations "similaires", contrairement à l'*analyse factorielle* qui vise à regrouper des variables "similaires".

On dira que cette analyse est *centrée sur les personnes*, à la différence d'analyses centrées sur les variables..

Pourquoi cette technique?

Illustrons avec l'exemple le plus simple, réalisé sur une seule variable dépendante : la taille des répondant(e)s (Oberski, 2016). On assume ici la présence de deux sous-populations:

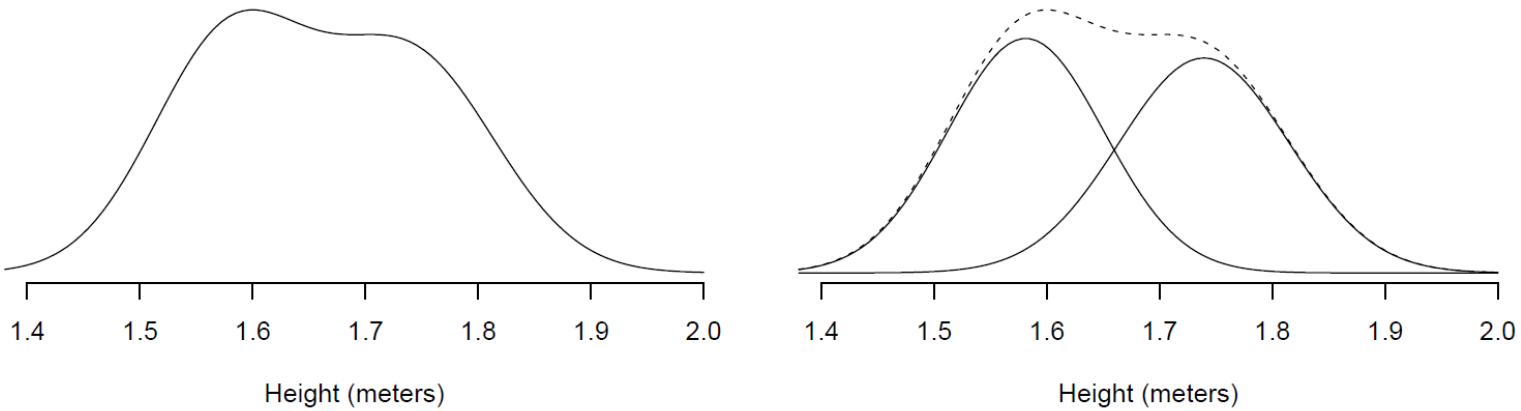


Fig. 1 Peoples' height. Left: observed distribution. Right: men and women separate, with the total shown as a dotted line.

Le modèle postule que la forme de la distribution complète des observations corresponds à la *somme pondérée* des distributions de chacune des sous-populations.

Objectifs de cette technique

L'objectif général est d'étudier l'hétérogénéité d'une population, afin de développer une *typologie empirique*.
Concrètement,

1. Estimer le nombre de sous-populations dans le jeu de données qui permet d'isoler des sous-populations homogènes et distinctes les unes des autres
2. Identifier les caractéristiques de chaque sous-population (i.e, proportion de la population, moyennes et écarts-type/variances sur chaque attribut)
3. Étudier les prédicteurs de l'appartenance à ces sous-populations

Objectifs de cette technique

Reprenons notre exemple sur la taille.

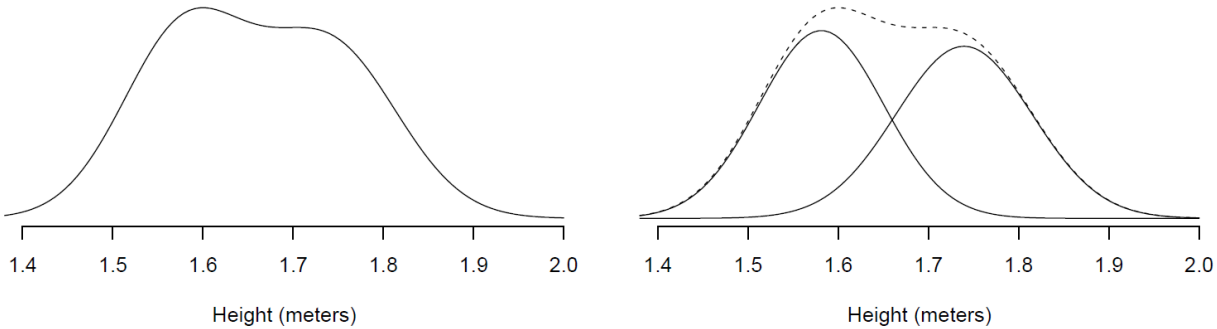


Fig. 1 Peoples' height. Left: observed distribution. Right: men and women separate, with the total shown as a dotted line.

Avec cette technique, nos objectifs seront donc :

1. De vérifier si un modèle postulant 2 sous-populations s'ajuste mieux aux données qu'un modèle à une sous-population (ou 3, ou plus);
2. Pour chaque sous-population, l'analyse pourra estimer la proportion de répondants qu'on y retrouve, la taille moyenne et sa dispersion (écart-type ou variance);

Définitions

On retrouve divers appellations pour ces techniques dans la littérature :

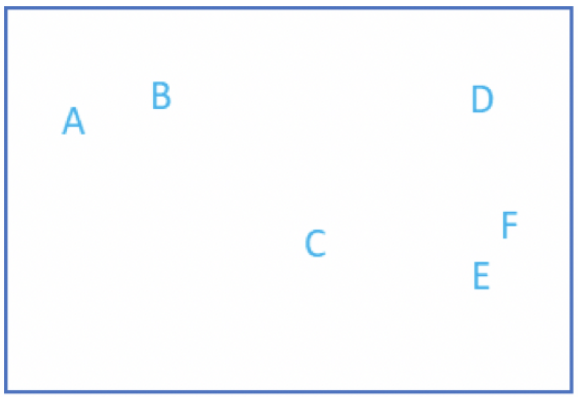
- Latent Class Analysis (LCA) - Analyse de classes latentes, pour des variables catégorielles
- Latent Profile Analysis (LPA) - Analyse de profils latents pour des variables continues
- Model-based Clustering

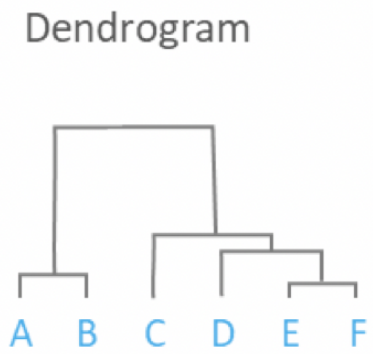
Toutes ces techniques appartiennent à la grande famille statistique des "modèles de mélange" (*mixture modeling*) - l'idée ici étant qu'on tente d'identifier les ingrédients qui ont permis d'obtenir le mélange observé dans notre jeu de données.

Avantages

On distingue cette approche (basée sur un modèle) d'autres approches connues comme analyses de regroupement (*cluster analysis*) :

1. *hierarchical clustering*





2. *K-means clustering*

Ces deux approches sont limitées aux variables normales, car basées sur une mesure de distance, et n'offrent pas de tests statistiques sur le nombre de sous-populations.

Comment trouver les sous-populations?

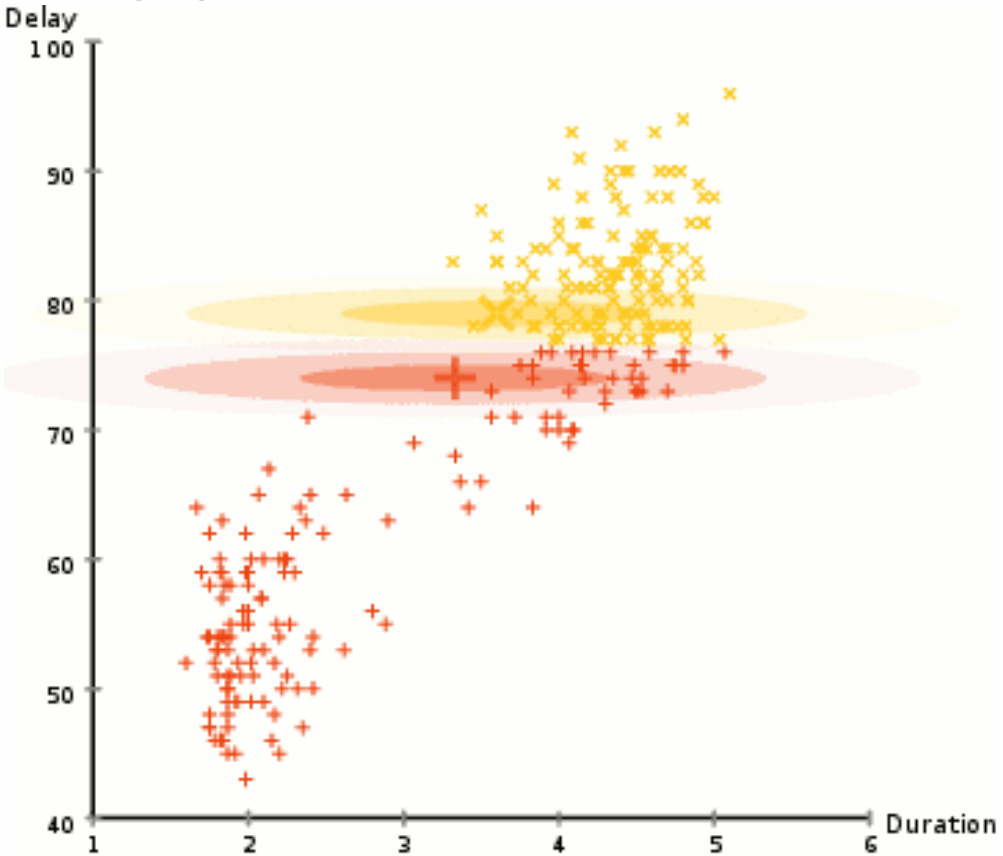
On parle d'analyse de classes/profils **latent(e)s**. Que signifie ce dernier mot? On dira qu'une variable est *latente* quand on ne connait pas sa valeur, seulement sa distribution (ici catégorielle/nominale).

Statistiquement, notre problème est un problème de *données manquantes* : l'analyste a les observations mais il lui manque les classes/profils! Solution? Utiliser un outil d'estimation pour les données manquantes, soit l'algorithme EM (*Expectation Maximization*).

C'est une approche par itération pour estimer les paramètres (prévalence, moyenne et variance de chaque classe) qui permettent de trouver le nombre pré-fixé de sous-populations les plus distinctes possibles.

Comment trouver les sous-populations?

Pour illustrer cette approche par étapes, voici une animation de l'algorithme EM (source : Wikipedia) pour identifier deux sous-populations selon deux variables dépendantes :



Nature des variables

Historiquement, l'analyse de regroupement était réservée aux variables normales, qu'on devait standardiser (Z, moyenne = 0, ET = 1) pour les exigences des mesures de distance.

- **Analyse de profils latents** : variables "continues", distribuées selon loi normale, dénombrement ou ordinale à plusieurs modalités (disons 5 ou plus);
- **Analyse de classes latentes** : variables binaires, nominales (catégories non-ordonnées) ou ordinales (2 à 4 modalités)

Avec le logiciel Mplus, cette distinction n'existe plus car ce logiciel supporte ces divers types de variables dans une même analyse. Il est donc possible de "mélanger" les types sans besoin de standardisation d'échelle.

Taille d'échantillon

Pas de solution simple.. Selon Tein et al. (2013):

> *Facteurs les plus importants sont : (1) le nombre de classes et (2) le degré de séparation des classes. La taille des classes n'est pas très importante.*

Donc, un petit échantillon ne pose pas problème si on a peu de classes et qu'elles sont bien distinctes.

Taille d'échantillon

Berg O. Muthén (auteur de Mplus) suggère que :

> *"A common question asked by researchers is, "What sample size do I need for my study?" Over the years, several rules of thumb have been proposed, such as 5 to 10 observations per parameter, 50 observations per variable, no less than 100, and so on. In reality, there is no rule of thumb that applies to all situations. " (Muthén & Muthén, 2002)*
>
> *"It all depends on how well the classes are separated. I have done successful mixture modeling with only 30 subjects [...]. General rules of thumbs are not worth much for mixtures because results depend so much on the specifics of your situation. (Muthén, 2013)*

Exemple d'article

Children and Youth Services Review 125 (2021) 106006

Contents lists available at ScienceDirect

Children and Youth Services Review

journal homepage: www.elsevier.com/locate/childyouth

The latent profile analysis of Chinese adolescents' depression: Examination and validation

Runting Chen, Yueyi Huang, Meng Yu *

Guangdong Provincial Key Laboratory of Social Cognitive Neuroscience and Mental Health, Department of Psychology, Sun-Yat Sen University, Guangzhou 510006, PR China

Exemple d'article

Éléments importants pour comprendre une analyse de classes/profils latent(e)s :

1. Description de l'analyse et du logiciel retenu
2. Indices statistiques pour justifier le choix du *nombre* de classes/profils
3. Figure pour *illustrer* les classes/profils
4. Prévalence (et spécifications?) de chaque classe/profil
5. Comparaison statistique des profils (optionnel)

Prochaine formation

Voici les thèmes qui seront abordés dans la prochaine formation (15 février) :
- présentation du modèle statistique de mélange
- introduction à Mplus
- exemple complet d'une analyse de profils latents avec Mplus
- interprétation des sorties
- programmation des modèles plus complexes (variances selon les classes, ajout de covariances)

Quelques références

- Caron, P.-O. (2018). *La Modélisation par Équations Structurelles avec Mplus*. PUQ.
- Collins, L. M., & Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Wiley.
- Oberski, D.L. (2016). Mixture models : latent profile and latent class analysis. Dans J. Robertson & M. Kaptsein (Eds.) *Modern Statistical Methods for HCI* (pp. 275-287). Springer.

Quelques références (suite)

- Muthén BO. *Sample size for LCA*. MPlus 2013; http://www.statmodel.com/discussion/messages/23/12750.html?1370464379. Accessed 05-25, 2018.
- Muthén LK, Muthén BO. How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal*. 2002;9(4):599-620.
- Tein J-Y, Coxe S, Cham H. Statistical Power to Detect the Correct Number of Classes in Latent Profile Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*. 2013;20(4):640-657.