L'analyse de classes/profils latents

Partie 2 : Introduction avec Mplus

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

Partenariat sur la Séparation parentale et Recomposition familiale
Université Laval

15 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

1. Modèle statistique
   - Objectifs de la technique
   - Modèle statistique
   - Version plus "réaliste" (multivariée)
   - Jeu de données
2. Programmation Mplus
   - Pourquoi Mplus?
   - Modèle par défaut
   - Modèles réalistes
   - Modèle avec covariances
   - Modèle avec variances flexibles
   - Comparaison des classes
3. Présentation des résultats
4. Références

Objectifs de la technique

A partir d'un échantillon d'observations multivariées (chaque répondant est mesuré sur plusieurs variables), l'objectif est d'identifier des sous-populations homogènes d'individus (qu'on appellera "classes" ou "profils" selon la nature des variables dépendantes).

En clair, l'analyste connait la valeur des variables pour chaque observation, mais ignore à quelle sous-population chaque observation appartient. On désire donc ici regrouper des observations similaires, contrairement à l'*analyse factorielle* qui regroupe des variables similaires.

Objectifs de la technique

Exemple univarié (une seule variable dépendante) d'une analyse de profils latents (Oberski, 2016):

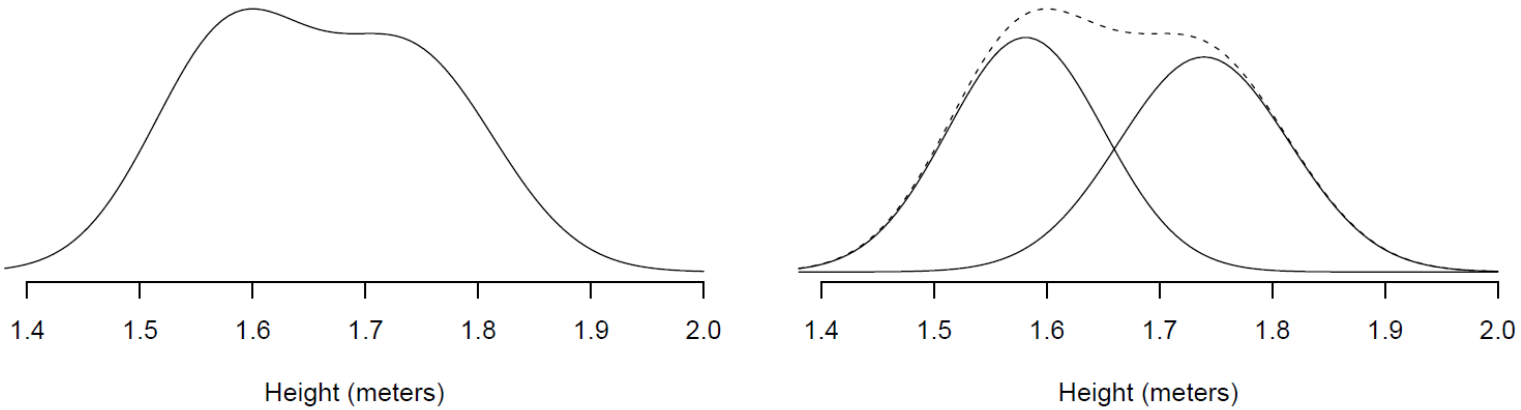


Fig. 1 Peoples' height. Left: observed distribution. Right: men and women separate, with the total shown as a dotted line.

L'objectif est donc d'identifier (1) s'il y a une ou plusieurs sous-population(s), et (2) les caractéristiques de ces sous-populations (prévalence, moyenne et dispersion).

Modèle statistique

Pour P classes et k variables dépendantes, le modèle statistique prend la forme suivante :

$$f(Y) = \sum_{i=1}^{P} \pi_i f_i(Y, \theta_i)$$

avec f la loi de distribution des variables, Y est l'ensemble des observations multivariées (une matrice de dimension n rangées/observations $\times$ k colonnes/variables), π_i le poids de la ie sous-population (probabilité allant de 0 à 1), et θ_i l'ensemble des paramètres associés à la ie sous-population (p.ex., la moyenne et la variance pour une distribution normale).

Modèle statistique

Appliqué à notre problème à $k = 1$ variable (taille) pour $P = 2$ sous-populations, le modèle devient :

$$\text{loi(taille)} = \pi_1 \text{normale}(\mu_1, \sigma_1^2) + \pi_2 \text{normale}(\mu_2, \sigma_2^2)$$

Le but de l'analyse sera donc d'estimer :

- la moyenne et la variance (μ_1, σ_1^2) de la première loi normale (première sous-population)
- la moyenne et la variance (μ_2, σ_2^2) de la seconde loi normale (deuxième sous-population)
- la proportion de la première sous-population π_1 dans le "mélange" (la seconde se déduit de la première, car $\pi_1 + \pi_2 = 1$)

Ce modèle a donc $2 + 2 + 1 = 5$ valeurs inconnues, qu'on appellera des *paramètres*, qui seront estimées à partir du jeu de données.

Modèle statistique multivarié

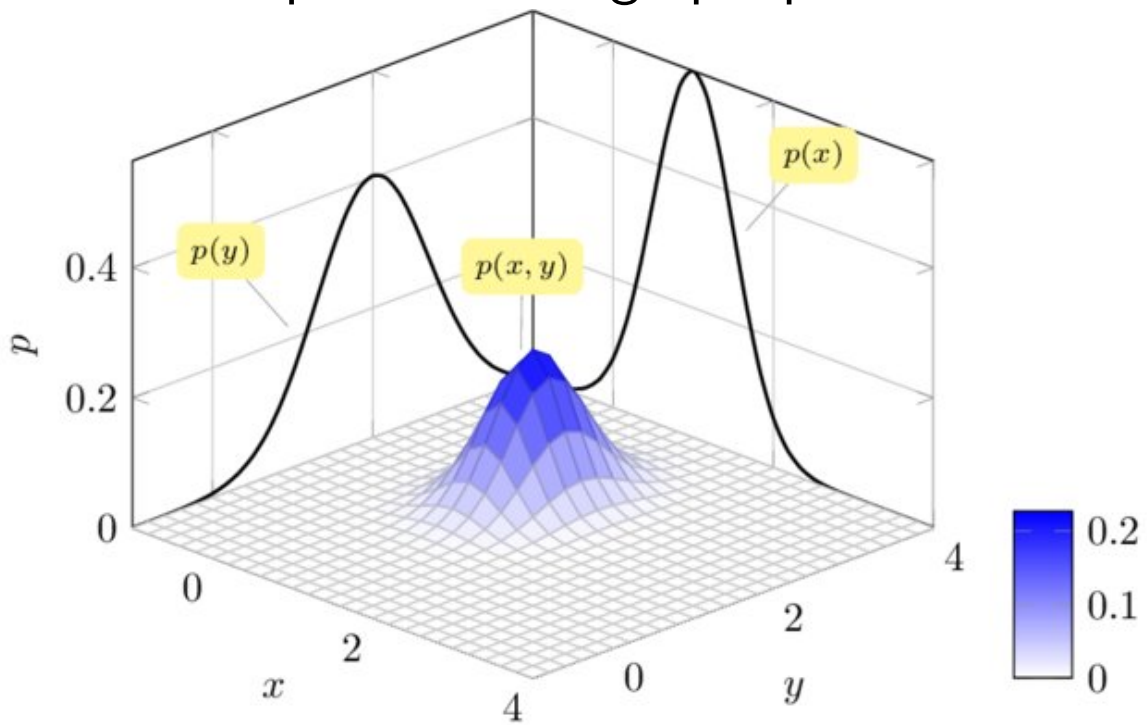
L'exemple précédent est simple à suivre mais il n'est pas réaliste, car nous modélisons rarement une seule variable dépendante à la fois. Voyons maintenant la forme que prendra un modèle réaliste (le plus simple) pour $P = 2$ sous-populations et $k = 2$ variables dépendantes (disons la taille et la santé).

Comme nous avons 2 variables dépendantes, on doit estimer une loi normale *multivariée* pour chaque sous-population.

- loi normale univariée : 1 moyenne et 1 variance pour décrire cette loi
- loi normale bivariée (2 dimensions) : 2 moyennes, 2 variances et 1 covariance pour décrire cette loi.

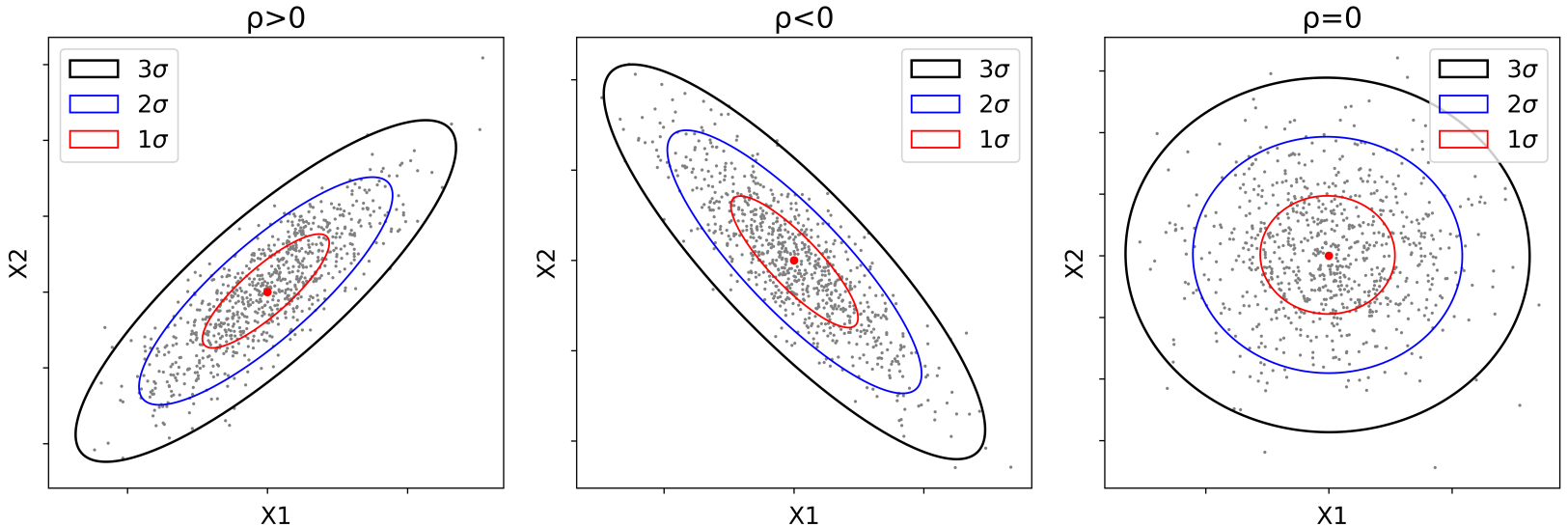
Modèle statistique multivarié

Voici une représentation graphique de la loi normale bivariée



Modèle statistique multivarié

La moyenne et la variance de chaque variable ne permettent pas de caractériser la loi normale bivariée. On doit également tenir compte de la covariance entre les deux variables (dimensions).



Modèle statistique multivarié

Le modèle statistique de profils latents pour la taille (T) et la santé (S) prendra donc la forme d'un mélange de deux lois normales bivariées :

$$\text{loi(taille,santé)} = \pi_1 \mathcal{N}(M_1, \Sigma_1) + \pi_2 \mathcal{N}(M_2, \Sigma_2)$$

où la moyenne bivariée est un vecteur

$$M = \begin{bmatrix} \text{moy(T)} \\ \text{moy(S)} \end{bmatrix}$$

et les variance-covariances sont une matrice

$$\Sigma = \begin{bmatrix} \text{var(T)} & \text{cov(T,S)} \\ \text{cov(S,T)} & \text{var(S)} \end{bmatrix}$$

Modèle statistique multivarié

Le but de ce modèle LPA bivarié sera donc d'estimer :

- les 2 moyennes, 2 variances et 1 covariance de la première loi normale (première sous-population)
- les 2 moyennes, 2 variances et 1 covariance de la seconde loi normale (deuxième sous-population)
- la proportion de la première sous-population π_1 dans le "mélange" (rappel : on peut déduire π_2)

Complexité du modèle : $5 + 5 + 1 = 11$ paramètres à estimer.

Jeu de données

Jeu de données composé de $P = 3$ classes comprenant $n = 200$ observations chacune, tirées d'une loi normale bivariée ($k = 2$ variables):

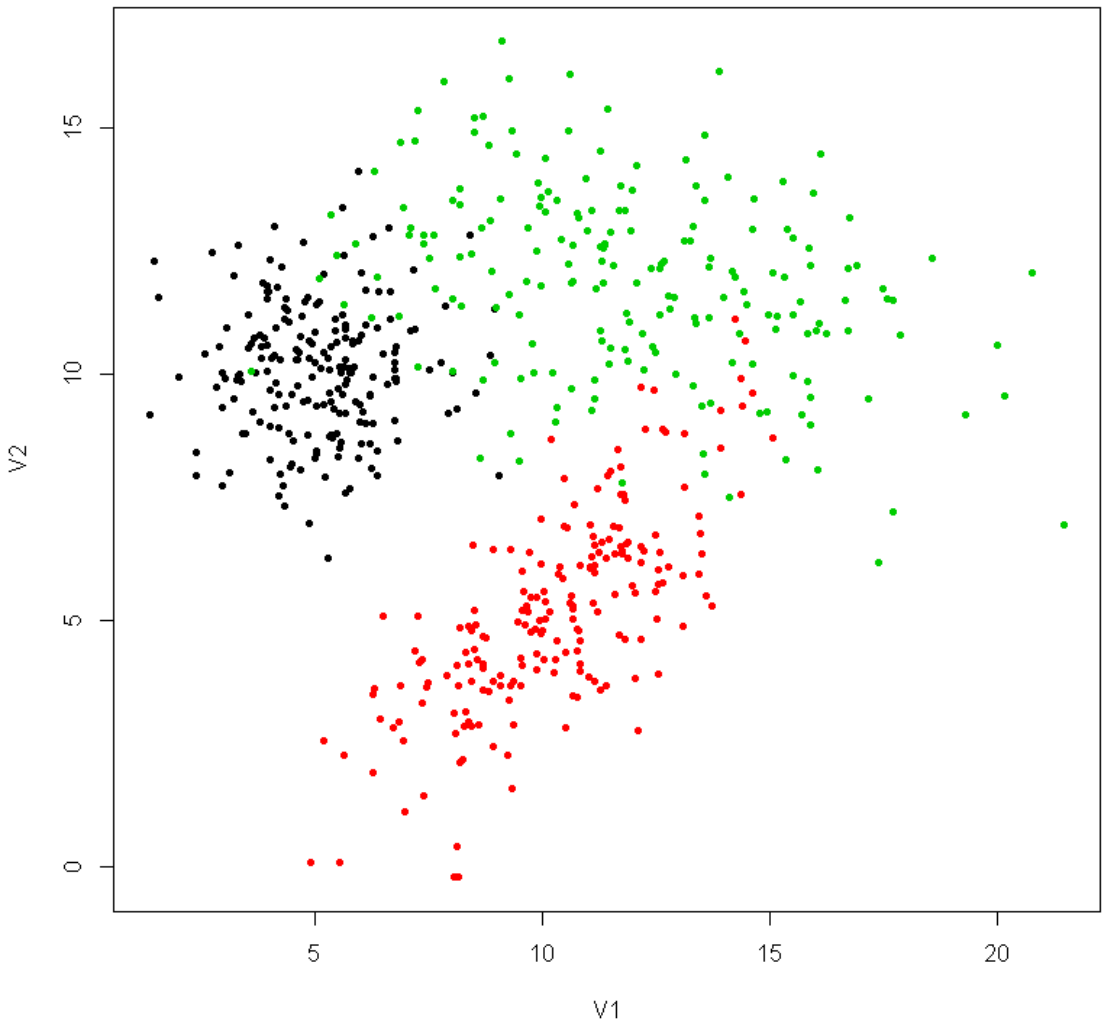
$$\text{Classe 1 : } M_1 = [5, 10], \Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, r_{12} = 0$$

$$\text{Classe 2 : } M_2 = [10, 5], \Sigma_2 = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}, r_{12} = 0.75$$

$$\text{Classe 3 : } M_3 = [12, 12], \Sigma_3 = \begin{bmatrix} 10 & -2 \\ -2 & 4 \end{bmatrix}, r_{12} = -0.32$$

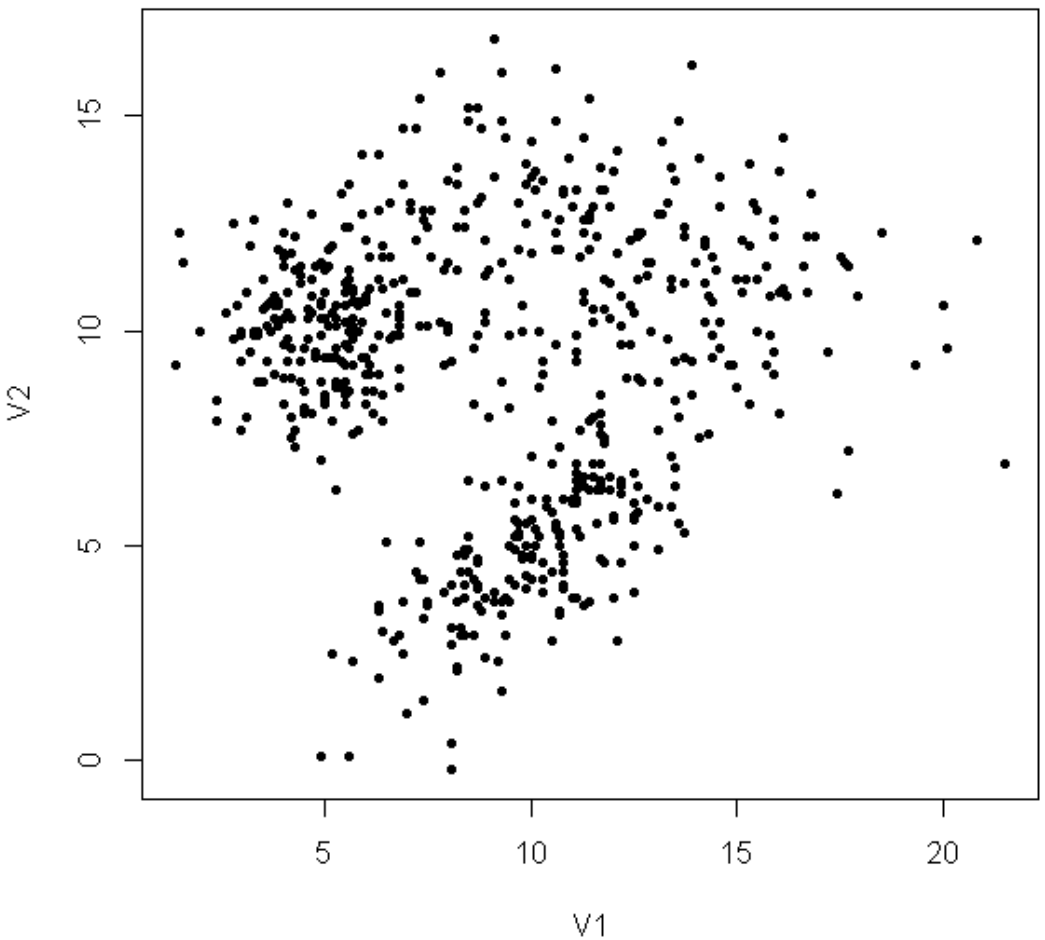
Jeu de données

Illustration du jeu de données (vue théorique):



Jeu de données

Illustration du jeu de données (vue selon le logiciel):

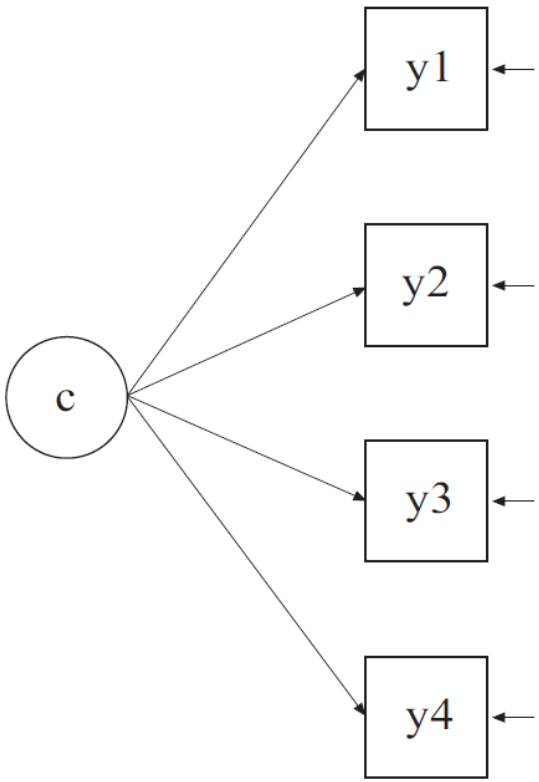


Pourquoi Mplus?

- Logiciel spécialisé dans les structures latentes (EQS, CFA, LPA/LCA, multiniveaux et autres)
- Développement depuis 1998 par une petite équipe d'experts
- Disponible sous Windows, MacOS et Linux
- Permet de personnaliser chaque modèle
- Offre plusieurs approches d'estimation (maximum de vraisemblance standard ou robuste aux DM, bootstrap, etc.)

Modèle par défaut

C'est le modèle LPA/LCA le plus simple estimable en Mplus.



Modèle par défaut

Ce modèle a plusieurs postulats :

- Moyennes peuvent varier entre les variables et entre les classes
- Variances peuvent varier entre les variables mais sont fixes entre les classes
- Covariance nulle entre les variables
- Comme la solution est sensible aux valeurs de départ, Mplus génère (par défaut) 20 valeurs aléatoires des paramètres et vérifie si l'ajustement est reproduit pour les 4 meilleures valeurs. Ce nombre peut être changé avec l'option STARTS.

Modèle par défaut

Comment programmer une LPA ou une LCA? Par la spécification de la distribution des variables.
Par défaut, une variable est considérée comme continue (i.e., modèle normal) dans Mplus. On doit spécifier le type d'une variable uniquement si elle est non-normale : nominale (NOMINAL), binaire/ordinale (CATEGORICAL), dénombrement (COUNT) ou de survie (CENSORED).
Donc, pour réaliser une LPA :

```
...
VARIABLES:
NAMES ARE id v1-v5;
...
```

Modèle par défaut

Pour réaliser une LCA, on devra préciser le type binaire/ordinal des variables (CATEGORICAL):

```
...
VARIABLES:
NAMES ARE id v1-v5;
CATEGORICAL ARE v1-v5;
...
```

IMPORTANT : la distinction LPA vs LCA est inutile en Mplus car on peut mélanger les deux types de variables dans la même analyse!

Modèle par défaut - syntaxe Mplus

Syntaxe Mplus pour le modèle par défaut ($P = 3$ classes):

```
DATA:
FILE IS nomdufichier.dat;
VARIABLE:
NAMES ARE id v1-v2 x1;
USEVARIABLES ARE v1-v2;
MISSING ARE all (999);
CLASSES = c(3);
ANALYSIS:
TYPE = MIXTURE;
OUTPUT:
TECH1 TECH11 RESIDUAL;
```

Modèle par défaut - nombre de paramètres

Voici comment calculer le nombre de paramètres à estimer pour un modèle normal (LPA) à P classes et k variables :

- $P \times k$ paramètres pour les moyennes
- k paramètres pour les variances
- 0 paramètre pour les covariances
- $P - 1$ paramètres pour les π_i (probabilité de classe). Pourquoi $P - 1$? Car $\sum^P \pi_i = 1$.

Exemple : LPA avec 3 classes et 2 variables? Total des paramètres $= 3 \times 2 + 2 + 2 = 10$.

Modèle par défaut - choix du nombre de classes

Mplus calcule un certain nombre d'indices pour qualifier l'ajustement du modèle et aider au choix du nombre de classes à conserver pour le modèle final :

Indices basés sur la vraisemblance

Ces indices permettent de comparer des modèles emboités (mêmes variables mais nombre différent de classes). La log-vraisemblance $\mathcal{LL}$ indique l'ajustement absolu (meilleur = le plus près de zéro) alors que les indices AIC ou BIC/SBC donnent un ajustement "relatif" (selon la complexité du modèle). Le BIC est préférable.

Modèle par défaut - choix du nombre de classes

Autres indices :

Indice basé sur le degré de séparation des classes

L'entropie ε est un indice standardisé (0-1) qui permet d'établir jusqu'à quel point chaque sujet est assigné clairement à une seule classe. Cet indice est calculé à partir des probabilités de classification. Un indice $\varepsilon > 0.80$ est désiré.

Indices subjectifs

On recherche des classes généralisables à la population ($\pi > 5\%$) afin de permettre une bonne compréhension du "mélange". Également, chaque classe/profil devrait être interprétable d'un point de vue théorique/clinique.

Modèle par défaut - choix du nombre de classes

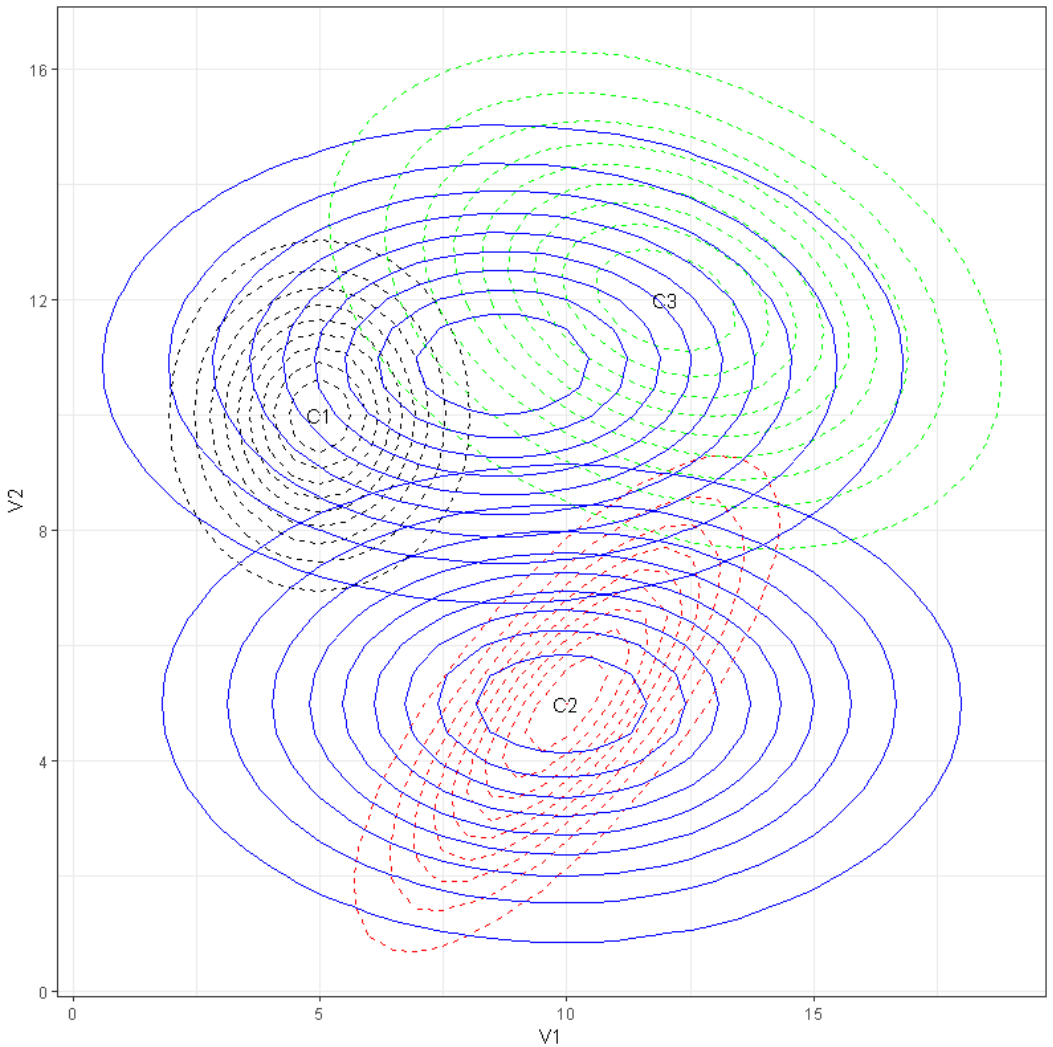
L'examen de ces indices "qualitatifs" peut être complété par des tests statistiques sur le nombre de classes :

1. le Lo-Mendell-Rubin (LMR) likelihood ratio test, qui produit une p-valeur sous Ho : le modèle avec P-1 classes est adéquat (option TECH11)
2. le Parametric Bootstrapped likelihood ratio (BLRT) test (option TECH14)

Pour plus de détails, consulter la note de Asparouhov et Muthen (2012), *Using Mplus TECH11 and TECH14 to test the number of latent classes*, disponible en PDF sur le web.

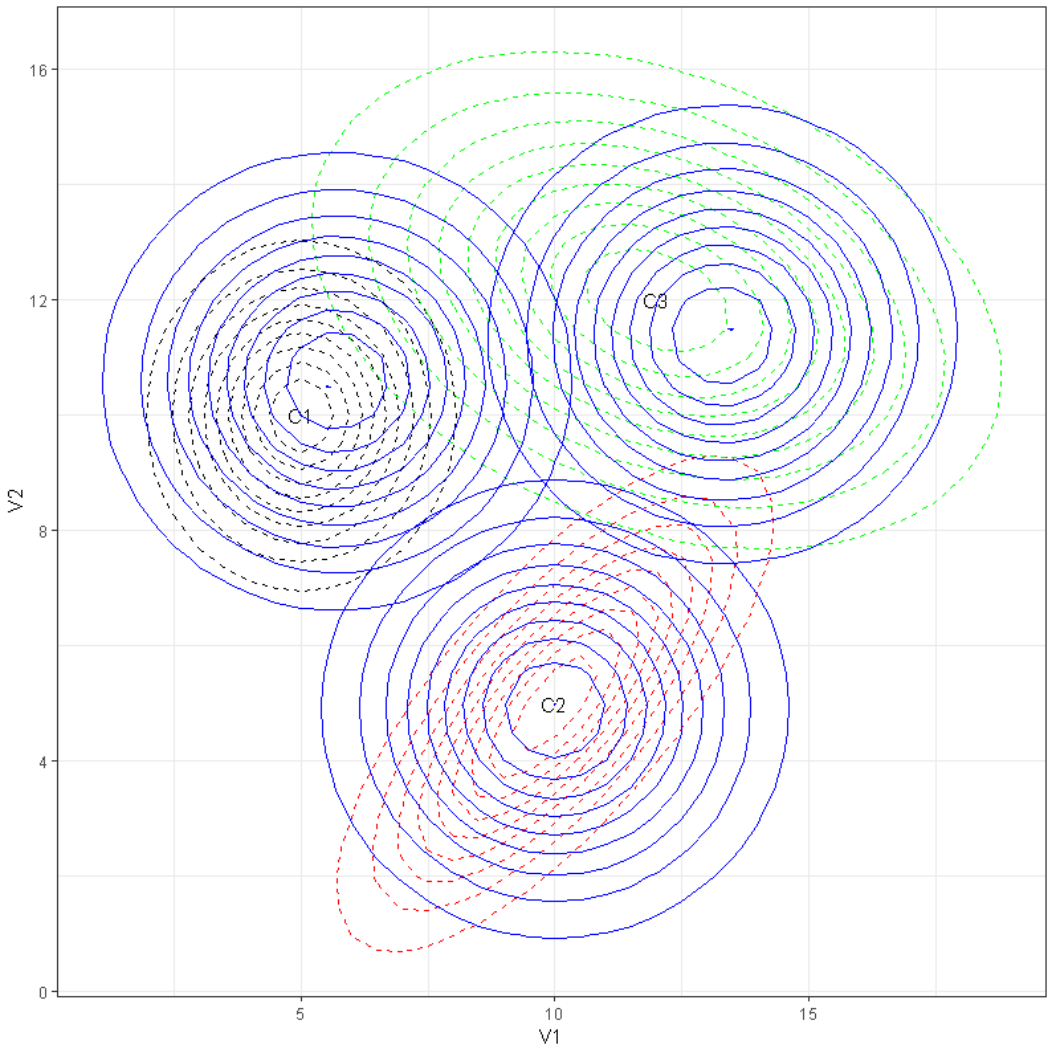
Modèle par défaut - résultats

Solution à deux classes (modèle LPA.A2):



Modèle par défaut - résultats

Solution à trois classes (modèle LPA.A3):



Modèle par défaut - sauvegarde des classes

Option pour sauvegarder les classes assignées et les probabilités a posteriori d'appartenir à chaque classe.

```
...
OUTPUT:
TECH1 TECH11 RESIDUAL;
SAVEDATA:
FILE IS classes.txt;
SAVE = CPROBABILITIES;
```

Modèle par défaut - bilan des indices d'ajustement

Bilan des modèles A (variances fixes, pas de covariance)

Modèle	$\#p$	$\mathcal{LL}$	BIC	ε	π_i	Test LMR
A1	4	-3229	6483	1.000	1.00	n.d.
A2	7	-3184	6413	0.814	.31,.69	$\leq$ .001
A3	10	-3087	6240	0.873	.31,.41,.27	$\leq$ .001
A4	13	-3048	6179	0.865	.31,.15,.37,.17	.002
A5	16	-3013	6130	0.857	.12,.36,.21,.14,.17	.004

Modèles plus réalistes

Le modèle par défaut est simple mais peu réaliste en pratique, en raison de ses nombreuses hypothèses.
Dans un modèle de profils latents*, l'analyste doit avoir une flexibilité pour préciser :

- qu'il existe des relations entre les variables (les variables ne sont pas toutes parfaitement indépendantes)
- que les variances peuvent être différentes entre les classes (une sous-population peut être plus dispersée qu'une autre sur les variables)

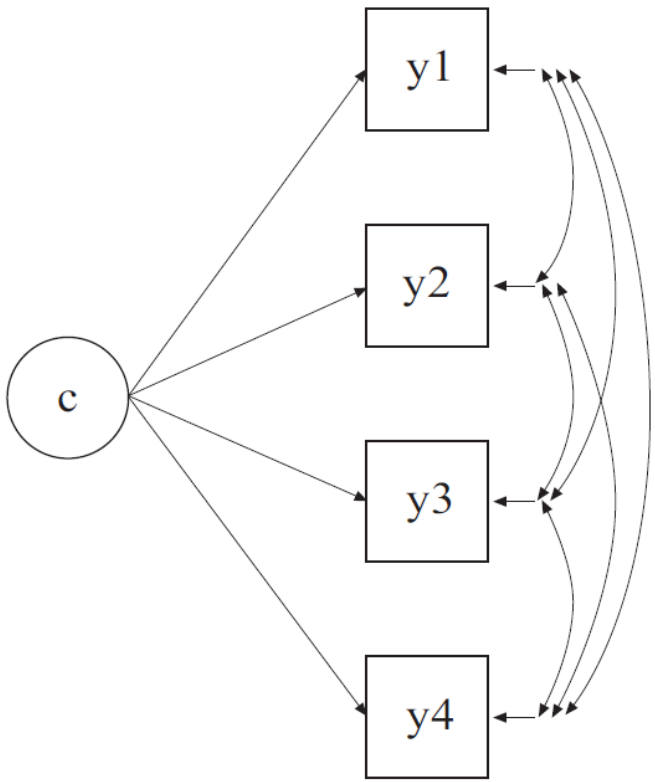
*IMPORTANT : Il n'est pas possible d'estimer des variances ou des covariances dans un modèle de classes latentes.

Modèle avec covariances

- Le modèle par défaut assume une covariance nulle entre chaque paire de variables (hypothèse forte!). Toutefois, cette contrainte permet d'estimer beaucoup moins de paramètres car il y a $k(k-1)/2$ covariances par classe à estimer.
- Avec plusieurs variables, on assiste à une explosion de la complexité du modèle. Par exemple, pour un modèle $P = 3$ classes et $k = 5$ variables, le modèle de base a $(P+1)k + P - 1 = 22$ paramètres mais on ajoute $k(k-1)/2 = 5(4)/2 = 10$ covariances par classe, soit 30 paramètres de plus (total de 52 paramètres!!)

Modèle avec covariances

Une approche plus réaliste est de permettre les covariances entre les variables, mais fixer que ces covariances (et variances) sont les mêmes pour toutes les classes, une condition similaire aux approches de regroupement (méthode hiérarchique de Ward ou par partionnement *k-means*).



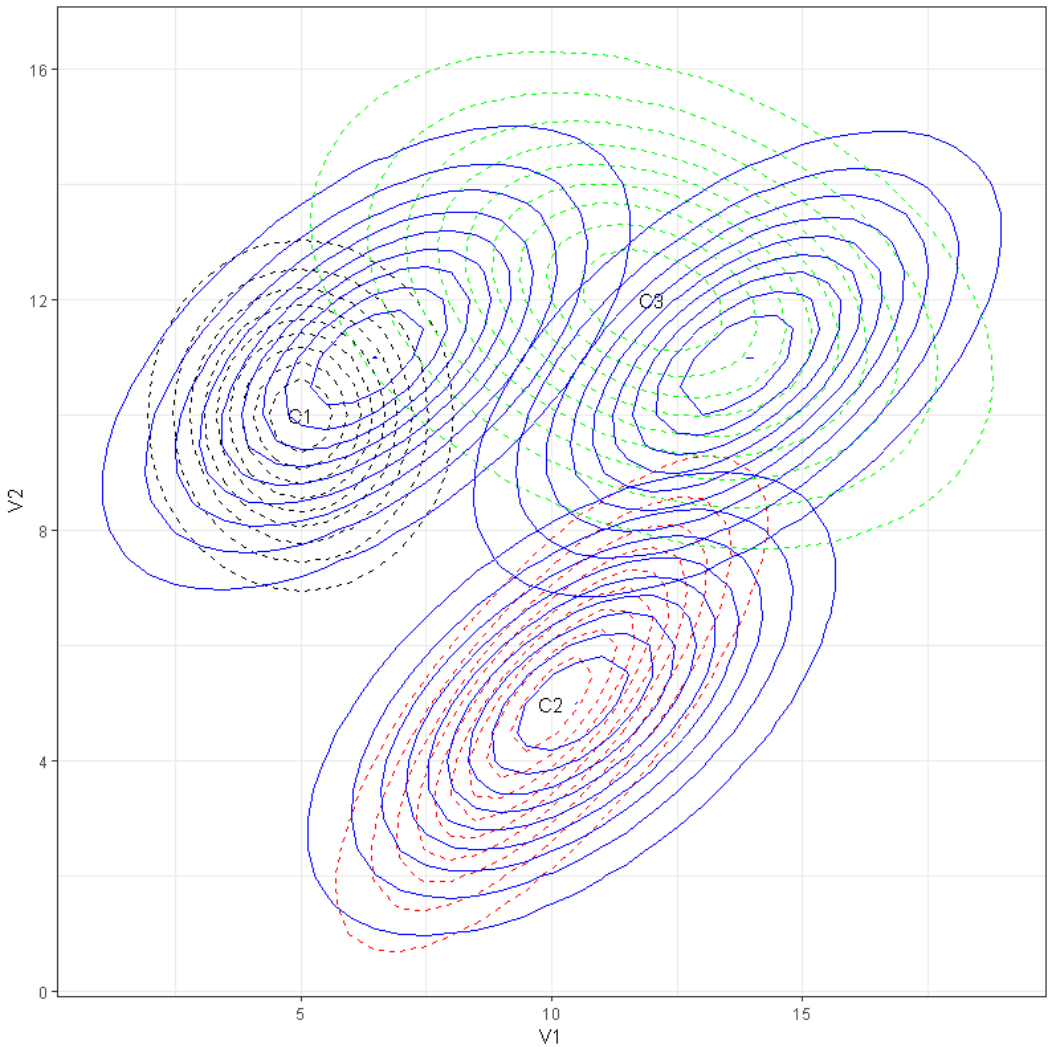
Modèle avec covariances

Addition d'un énoncé MODEL pour ajouter l'estimation des covariances entre les variables (la spécification sous OVERALL force la même valeur pour toutes les classes):

```
...
ANALYSIS:
TYPE = mixture;
MODEL:
%OVERALL%
v1 WITH v2;
...
```

Modèle avec covariances - résultats

Solution à trois classes (modèle LPA.B3):



Modèle avec covariance - bilan

Bilan des modèles B (variances fixes, avec covariances)

Modèle	$\#p$	$\mathcal{LL}$	BIC	ε	π_i	Test LMR
A3	10	-3087	6240	0.873	.31,.41,.27	$\leq .001$
A4	13	-3048	6179	0.865	.31,.15,.37,.17	.002
B3	11	-3049	6170	0.879	.32,.20,.48	$\leq .001$
B4	14	-3022	6133	0.874	.31,.16,.11,.41	.11

Modèle avec variances selon classe

Il est également possible de permettre que les variances de chaque indicateur changent selon la classe (hétérogénéité des variances). Comme il y a k estimés de variance par classe, cette flexibilité ajoute $(P-1)k$ paramètres supplémentaires.
Exemple ($P = 4$ classes et $k = 6$ variables) : modèle par défaut $= (P+1)k + P - 1 = 33$ paramètres. Ajout des variances selon la classe $= (4-1)6 = 18$ paramètres supplémentaires (augmentation de 55% de la taille du modèle).

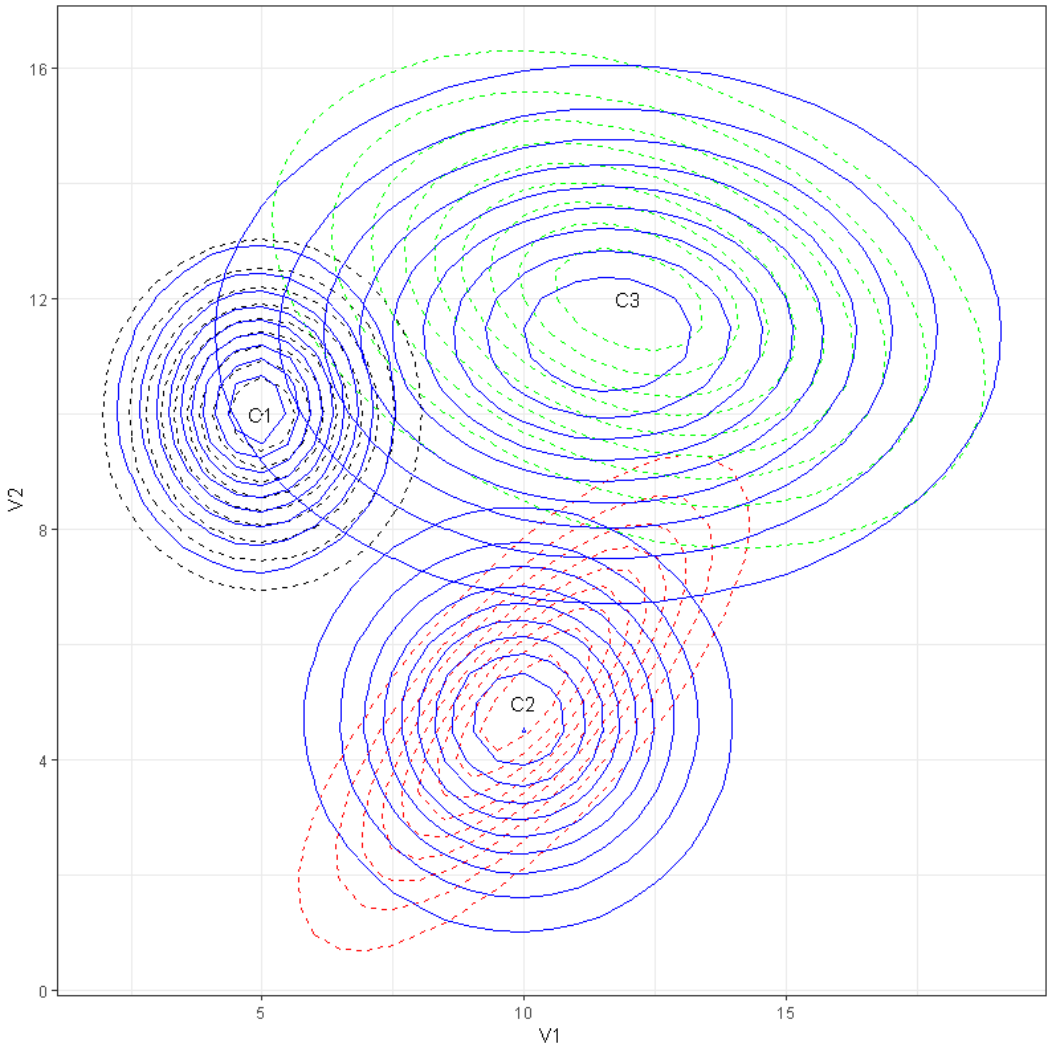
Modèle avec variances selon classe

Addition d'un énoncé MODEL pour spécifier d'estimer les moyennes (e.g., [V1]) ET les variances (e.g., V1) par classe :

```
...
ANALYSIS:
TYPE = mixture;
MODEL:
%c#1%
[v1 v2];
v1 v2;
%c#2%
[v1 v2];
v1 v2;
%c#3%
[v1 v2];
v1 v2;
...
```

Modèle avec variances - résultats

Solution à trois classes, avec variances flexibles (selon la classe), covariance nulle (modèle LPA.C3):



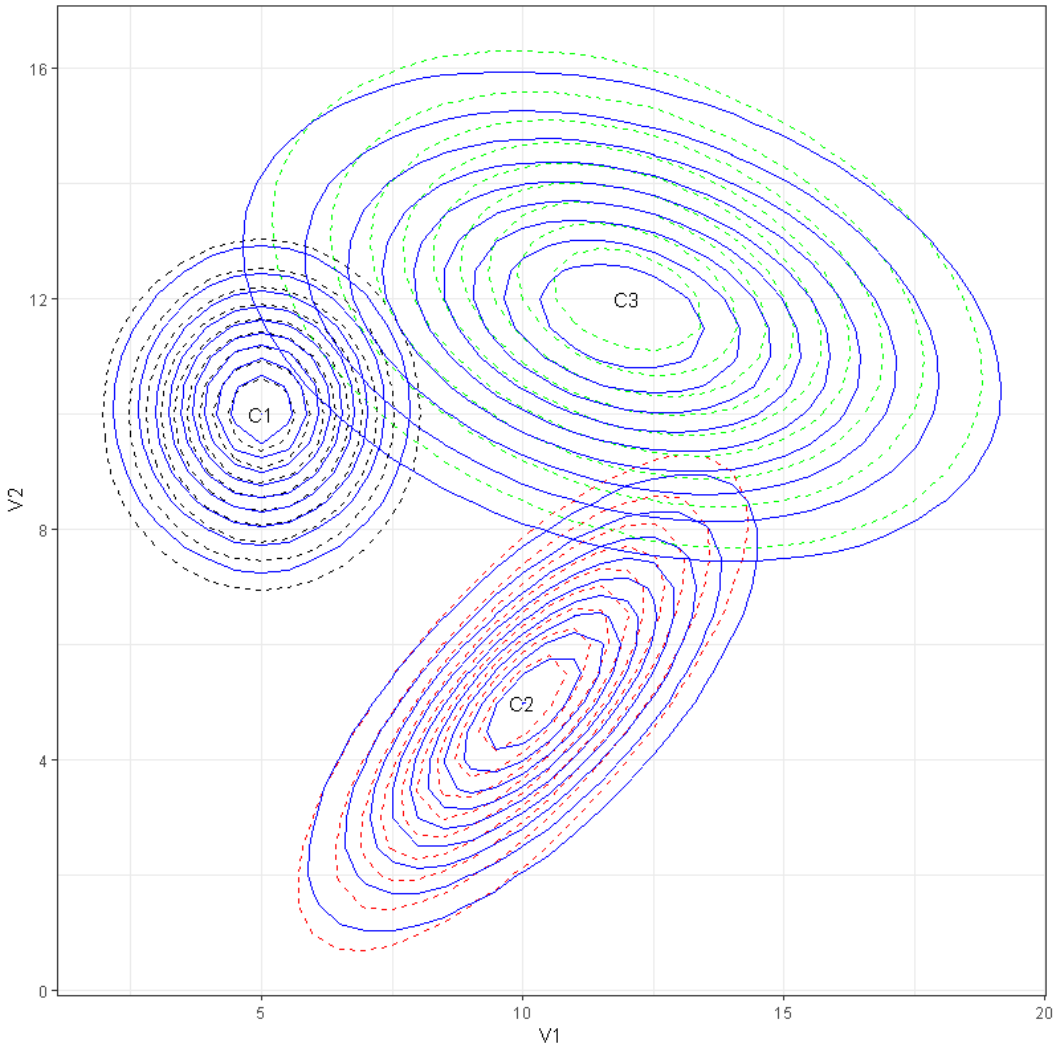
Modèle avec variances - bilan

Bilan des modèles C (variances flexibles, sans covariance)

Modèle	$\#p$	$\mathcal{LL}$	BIC	ε	π_i	Test LMR
A3	10	-3087	6240	0.873	.31,.41,.27	$\leq .001$
B3	11	-3049	6170	0.879	.32,.20,.48	$\leq .001$
C3	14	-3029	6148	0.835	.29,.30,.41	$\leq .001$

Modèle avec variances - résultats

Solution finale à trois classes, avec variances flexibles et covariances flexibles (selon la classe) (modèle LPA.D3):



Modèle avec variances et covariances - bilan

Bilan des modèles D (variances et covariances flexibles)

Modèle	$\#p$	$\mathcal{LL}$	BIC	ε	π_i	Test LMR
A3	10	-3087	6240	0.873	.31,.41,.27	$\leq .001$
B3	11	-3049	6170	0.879	.32,.20,.48	$\leq .001$
C3	14	-3029	6148	0.835	.29,.30,.41	$\leq .001$
D3	17	-2970	6050	0.884	.32,.36,.32	$\leq .001$

Comparaison des classes

Il existe trois approches pour "comparer" les classes (p.ex., sur l'âge) :

1. Importer la classe prédite dans SPSS et réaliser une ANOVA sur l'âge selon la classe (avantage = plus simple, limite = on ne tient pas compte de l'incertitude associée à la classe)
2. Comparer l'âge selon la classe dans Mplus, à l'aide de l'option AUXILIARY (préférable statistiquement)
3. Étudier le pouvoir prédictif de l'âge sur la classe (question distincte)

Comparaison des classes

Dans le modèle LPA.EC, on utilise l'option E pour demander un test d'égalité des moyennes (et ses comparaisons multiples). On peut également utiliser l'option R3STEP pour demander une régression logistique multinominale où la catégorie de réference est tour à tour fixée pour chaque classe :

```
VARIABLE:
...
usevariable are v1 v2 age;
...
auxiliary = age(E);
```

Dans ce modèle, la vraisemblance n'est pas affectée (la comparaison est une étape distincte de l'estimation des classes).

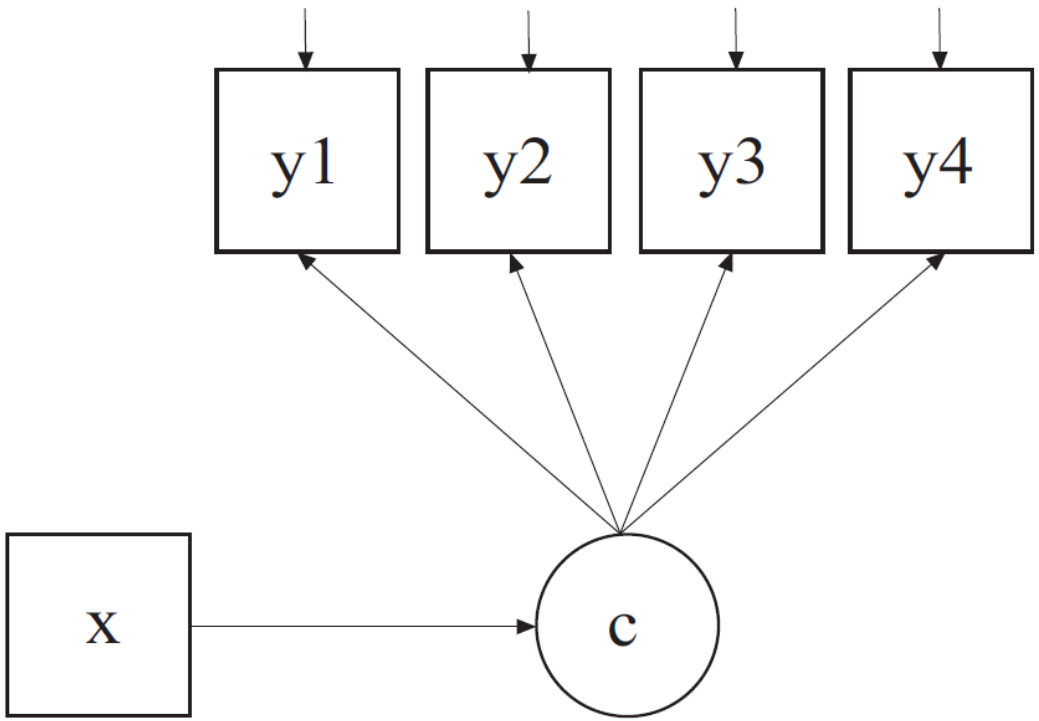
Comparaison des classes

Est-ce que les deux approches (ANOVA dans SPSS vs option AUXILIARY dans Mplus) donnent les mêmes résultats? Voici les moyennes de l'âge (la variable externe) selon l'approche :

Approche	Classe 1	Classe 2	Classe 3	Statistique
ANOVA	20.30_a	20.32_a	20.65_b	$F = 7.62, p = .001$
Mplus	20.31_a	20.32_a	20.64_b	$\chi^2 = 9.74, p = .008$

Ajout d'un prédicteur de classe

Il est possible d'ajouter des variables qui prédisent la variable latente de classe. Ce modèle est plus avantageux que d'étudier ces relations à l'extérieur de Mplus car on tient compte de l'incertitude associée à l'attribution de chaque classe.



Ajout d'un prédicteur de classe

Modèle LPA.EP :

```
...
ANALYSIS:
TYPE = MIXTURE;
MODEL:
%OVERALL%
c ON x1;
...
```

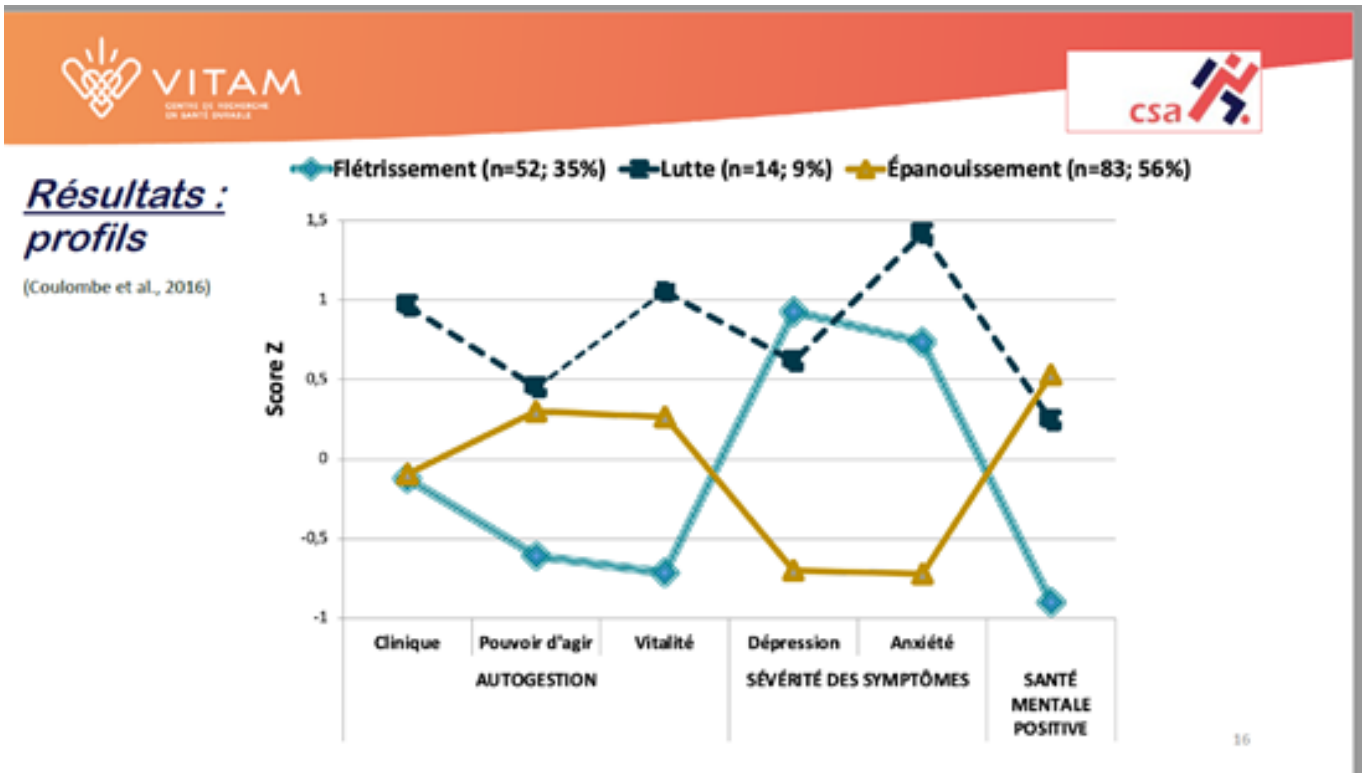
La variable à gauche du ON est prédite (VD, la classe c) alors que celle de droite est le prédicteur (VI, la variable x1). Dans ce modèle, la vraisemblance est affectée car la relation entre le prédicteur et la classe fait partie intégrante du modèle LPA.

Présentation des résultats

1. Décrire l'analyse (type de variable, variances et/ou covariances flexibles) et indiquer le logiciel (version) utilisé
2. Présenter un tableau pour justifier le nombre de classes retenues
3. Présenter un tableau ou une figure pour présenter chaque profil selon les variables indicatrices
4. (optionnel) Présenter les comparaisons entre les classes/profils sur des variables externes

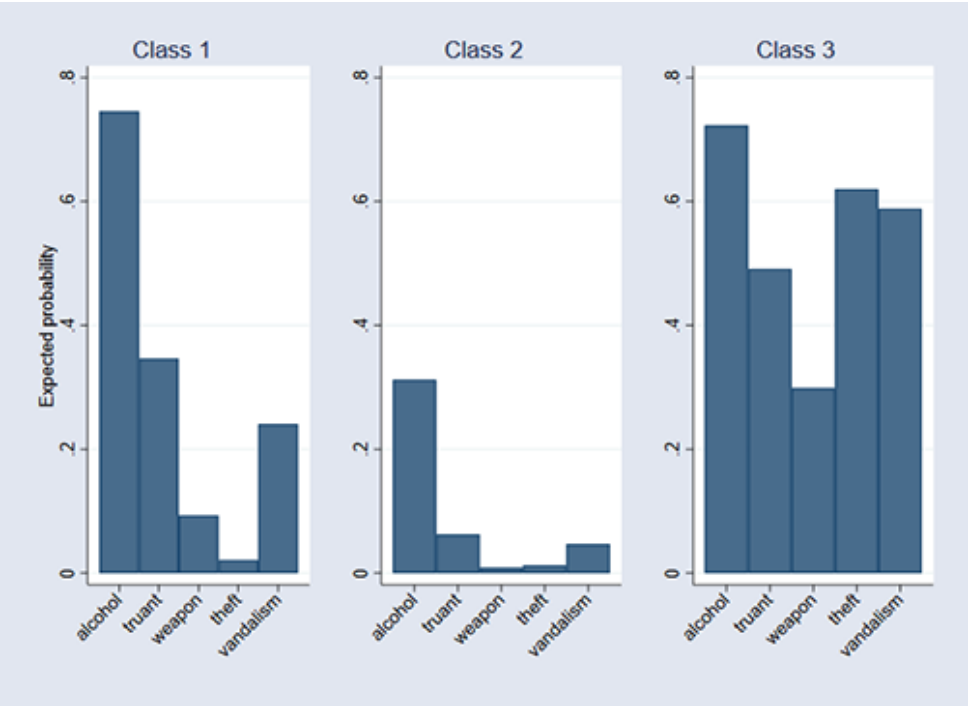
Présentation des résultats d'une LPA

Exemple de résultats pour une analyse de *profils* latents. On peut remarquer que chaque profil se distingue selon la moyenne sur chaque variable indicatrice :



Présentation des résultats d'une LCA

Exemple de résultats pour une analyse de *classes* latentes. On peut remarquer que chaque classe se distingue selon la répartition des modalités de chaque variable indicatrice :



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.
- Muthén, L.K. & Muthén, B.O. (2012). *Mplus User's Guide (7th ed.)* (chapitre 7). Los Angeles, CA.
- 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.