StepMix : Modélisation de données hétérogènes et complexes par le biais de modèles de mélange généralisés

Présentation du package et de ses applications en sciences sociales

Clémentine Courdi et Félix Laliberté

Sous la direction de Pr. Éric Lacourse

RQCP

Réseau Québécois COVID - Pandémie

Québec COVID - Pandemic Network (QCPN)

Qu’est-ce que StepMix ?

- Un package de clustering/classification par apprentissage automatique non supervisé
  - Permet de facilement modéliser la famille des modèles de mélange généralisés (*Generalised Mixture Models*), spécifiquement les modèles de mélanges finis (*Finite Mixture Models; FMM*)
    - Analyse de classes latentes (LCA)
    - Analyse de profils latents (LPA)
    - Etc.
- StepMix permet d’identifier des sous-populations homogènes (non observées)
  - Données transversales (p. ex. LCA) et longitudinales (p. ex. RM-LCA)
- StepMix facilite la modélisation en étapes de modèles de mélange structurels
  - Diverses approches en étapes (stepwise)

Modèles de mélange finis et variables latentes catégorielles

Variable latente / *Variable manifeste*	***Catégorielle***	***Continue***
Catégorielle	Analyse de classes latentes (LCA)	Analyse de traits latents (LTrA)
Continue	Analyse de profils latents (LPA)	Analyse factorielle (FA)

- Nomenclature variable au fil de l'histoire et selon les disciplines
- Les frontières entre les différents FMM sont de plus en plus floues avec les avancées techniques permettant de mélanger les types d'indicateurs
- Les FMM postulent l'existence d'une structure latente non observable et représentée par un nombre *fini* d'indicateurs
- Les FMM permettent de classer les individus dans des sous-groupes distinct et mutuellement exclusifs

(Boudon, 1962; Gibson, 1959; Lazarsfled et Henry, 1968; McCutcheon, 1987)

Les modèles de mélange finis en sciences sociales

- Popularité croissante des FMM dans les dernières décennies grâce aux avancées technologiques et informatiques
- Contexte historique
  - Les FMM ont été introduits pour découvrir des sous-groupes cachés (latents) au sein d'une population
  - Les chercheurs souhaitent souvent étudier des phénomènes de nature complexe, mais impossibles à observer directement. Il faut donc utiliser de multiples variables manifestes pour construire une ou plusieurs variables latentes représentant les concepts d'intérêt.
  - FMM valorisés pour leur côté inductif
- Les sous-groupes peuvent représenter soit:
  - 1) De réelles catégories discrètes, qualitativement distinctes
  - 2) Une approximation d'une distribution non-normale

(Boudon, 1962; Lazarsfled et Henry, 1968; McCutcheon, 1987; Nagin, 2005)

Modèle de mesure VS modèle structurel

- En contexte de recherche, on veut généralement lier la variable latente créée par le FMM à des variables externes
  - Prédicteurs
  - Covariables
  - *Distal outcomes*
- Modèle de mesure : estimation de la variable latente
- Modèle structurel : ajout de variables externes
- La manière dont on estime le modèle structurel (simultanément ou séparément du modèle de mesure) est un enjeu majeur qui a mené au développement des approches dites « par étapes »

Estimation algorithmique

- Méthode inductive par nature: aucun a priori sur les paramètres du modèle
- Pour un nombre de classes donné, les paramètres sont estimés avec l'algorithme *expectation-maximisation* jusqu'à l'obtention du *maximum likelihood*
- Deux types de paramètres sont estimés:
  - Prévalence des classes
  - Probabilités conditionnelles de chaque indicateur selon la classe (coefficients)
- Important: les FMM postulent l'indépendance conditionnelle des indicateurs (à l'intérieur de chaque classe)
- Plusieurs possibilités en cas de données manquantes (FIML ou MI)

(Finch et French, 2015; Hsiao et al., 2021; McCutcheon, 1987; Nylund, Asparouhov et Muthén, 2007)

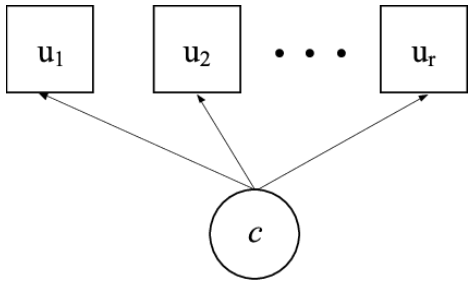
Sélectionner un modèle optimal

- On doit ensuite choisir le bon nombre de classes
- Deux catégories d'indices de *fit* :
  - 1) Les critères d'information *(information criteria)*: BIC, AIC...
  - 2) Les tests de vraisemblance *(likelihood tests)*: LMT-LRT, BLRT...
- Il faut toujours considérer l'interprétation théorique du modèle dans le choix
- Assignation aux classes

(Finch et French, 2015; Hsiao et al., 2021; McCutcheon, 1987; Nylund, Asparouhov et Muthén, 2007)

Exemple: Analyse de classe latente (LCA) simple

- Type de FMM le plus connu et utilisé en sciences sociales
  - Items catégoriels (u_r)
  - Variable latente catégorielle (c)



- Exemple avec données de l'Enquête sociale générale américaine de 1976-77 (N = 3029) (Bakk et al., 2016; Bakk et Kuha, 2021)
  - Objectif: Prédire le statut social des parents à partir de trois indicateurs
    - 1) Score de prestige de la profession du père
      - 3 catégories, où 1 = Faible, 2 = Moyen et 3 = Élevé
    - 2) Niveau d'éducation de la mère
      - 5 catégories, où 1= Inférieur au secondaire et 5 = Doctorat
    - 3) Niveau d'éducation du père
      - 5 catégories, où 1 = Inférieur au secondaire et 5 = Doctorat

Suite exemple: LCA simple

Résultats du modèle à 3 classes

	Faible	Moyen	Élevé
Prévalences de classe	0.70	.23	.07
Prestige de l'emploi du père			
Faible	**0.47**	0.31	0.05
Moyen	**0.53**	**0.67**	**0.46**
Élevé	0.00	0.02	**0.49**
Éducation de la mère			
Inférieur au secondaire	**0.82**	0.14	0.15
Secondaire	0.17	**0.79**	**0.44**
Collège	0.00	0.03	0.01
Baccalauréat	0.01	0.04	**0.30**
Doctorat	0.00	0.01	0.10
Éducation du père			
Inférieur au secondaire	**0.95**	0.06	0.00
Secondaire	0.05	**0.89**	0.11
Collège	0.00	0.00	0.05
Baccalauréat	0.00	0.05	**0.39**
Doctorat	0.00	0.00	**0.44**

Note. Les résultats proviennent provenant de l'Enquête sociale générale américaine de 1976-77 (n = 3029).

Exemple de modèle structurel

- Les revenus des répondants varient-ils en fonction de la classe sociale de leurs parents?

Résultats du modèle à 3 classes

	Faible	Moyen	Élevé
Prévalences de classe	0.70	.23	.07
Prestige de l'emploi du père			
Faible	**0.47**	0.31	0.05
Moyen	**0.53**	**0.67**	**0.46**
Élevé	0.00	0.02	**0.49**
Éducation de la mère			
Inférieur au secondaire	**0.82**	0.14	0.15
Secondaire	0.17	**0.79**	**0.44**
Collège	0.00	0.03	0.01
Baccalauréat	0.01	0.04	**0.30**
Doctorat	0.00	0.01	0.10
Éducation du père			
Inférieur au secondaire	**0.95**	0.06	0.00
Secondaire	0.05	**0.89**	0.11
Collège	0.00	0.00	0.05
Baccalauréat	0.00	0.05	**0.39**
Doctorat	0.00	0.00	**0.44**

Note. Les résultats proviennent provenant de l'Enquête sociale générale américaine de 1976-77 (n = 3029).

Revenu du répondant
(en milliers de dollars)

Résultats du modèle structurel

1) Paramètres estimés du modèle structurel

Revenus moyens estimés selon la classe sociale des parents

	Classe faible	Classe moyenne	Classe élevée
Modèle structurel	27.44 (0.59)	35.94 (1.28)	43.67 (3.85)

Note. Variable indépendante : Classes sociales (LCA); Variable dépendante : Revenu du répondant (en milliers). Erreur-type entre parenthèses.

2) Différences de moyennes

Résultats de la régression

Modèles	Est.	SE	Z	*P*(>|z|)
Classe faible	NA	NA	NA	NA
Classe moyenne	8.50	1.41	6.02	p < .001
Classe élevée	16.22	3.83	4.24	p < .001

Note. Variable indépendante : Classes sociales (LCA). Variable dépendante : Revenu du répondant (en milliers). Est. : Coefficient normalisé; SE : Erreur-type; Z: *Two-tailed Z-test*; *P*(>|z|): valeur-p.

Retour à l'exemple précédent

Rappel

- Lorsqu'on ajoute une ou plusieurs variables externes : 2 modèles
- **Le modèle de mesure**
  - Modèle probabiliste : chaque répondant a une probabilité postérieure d'appartenir à la classe *c*
    - Exemple : le 18ème répondant
      - Probabilité de 0,60 d'appartenir à la classe faible
      - Probabilité de 0,40 d'appartenir à la classe moyenne
- **Le modèle structurel**
  - Modèle linéaire : relation entre le modèle de mesure et la ou les variables externes

Approche utilisée

- Approche utilisée dans l'exemple précédent
  - Première étape : Estimer les paramètres du modèle de mesure
  - Deuxième étape : Assigner les individus aux classes
    - C'est-à-dire : création d'une nouvelle variable
  - Troisième étape : Estimer les paramètres du modèle structurel
    - Relation entre la variable créée et la variable dépendante distale
- Problème :
  - Introduction d'un biais d'assignation (étape 2)
  - Assignation déterministe aux classes latentes
    - On extrait les classes par assignation modale, puis on utilise les classes latentes comme les catégories d'une variable connue
      - P. ex., le 18ème répondant a une probabilité de 1.00 d'appartenir à la classe faible (vs 0,60)
- Approche en 3 étapes dite « naïve »
  - Biaise les paramètres du modèle structurel (étape 3)

Approches par étapes

- Plusieurs approches par étapes (*stepwise approches*)
  - Approche en 1 étape
  - Approche en 2 étapes (Bakk et Kuha, 2018)
  - Approche en 3 étapes naïve
  - Approche en 3 étapes avec correction BCH (Bolck et al., 2004)
  - Approche en 3 étapes avec correction ML (Vermunt, 2010)

Approche en une étape

- Approche en une étape (*one-step approach*)
  - Une autre approche traditionnellement utilisée
  - Principe :
    - Le modèle de mesure et le modèle structurel sont estimés simultanément
    - Tous les paramètres sont libres et estimés en une seule étape
- Problème :
  - Tendance à rendre le modèle de mesure ininterprétable
    - Les prévalences et patrons de probabilités conditionnelles sont différents de ce qui observé sans variable externe
    - Le modèle de mesure est modifié à chaque ajout ou retrait d'une variable externe

Approches en trois étapes avec correction du biais

- Approches en trois étapes avec correction du biais
  - Principe : correction du biais d'assignation aux classes introduite à l'étape 2 (*bias-adjusted three-step approach*)
  - Deux approches populaires :
    - Approche en trois étapes avec correction BCH (Bolck et al., 2004)
    - Approche en trois étapes avec correction ML (Vermunt, 2010)
- Principe
  - Première étape : Estimer les paramètres du modèle de mesure
  - Deuxième étape :
    - Assigner les individus aux classes
    - Estimer les probabilités de mauvaises classifications
  - Troisième étape : Estimer les paramètres du modèle structurel corrigés de l'incertitude d'assignation aux classes

Approche en deux étapes

- Approche en deux étapes (Bakk et Kuha, 2018)
  - Les approches en 3 étapes avec correction demeurent biaisées dans certaines situations
    - P. ex., lorsque le degré de séparation des classes est faible
  - Approche en deux étapes proposée pour remédier au biais d'assignation aux classes (étape 2)
- Principe
  - Première étape :
    - Estimer les paramètres du modèle de mesure
  - Deuxième étape :
    - Fixer les paramètres du modèle de mesure
    - Estimer les paramètres du modèle structurel

Retour à l'exemple présenté

Paramètres des modèles structurels estimés

Revenus moyens estimés selon la classe sociale des parents et l'approche utilisée

Modèles	u_1 Faible (SE)	u_2 Moyenne (SE)	u_3 Élevée (SE)
Approches en étapes			
1 étape	Modèle de mesure distordu		
3 étapes Naïve	27.44 (0.59)	35.94 (1.28)	43.67 (3.85)
3 étapes BCH	26.71 (0.74)	36.81 (1.64)	44.68 (4.26)
3 étapes ML	21.05 (1.52)	44.73 (5.25)	61.26 (11.77)
2 étapes	25.25 (2.48)	38.41 (6.14)	50.66 (7.13)

Note. Variable indépendante : Classes sociales (LCA); Variable dépendante : Revenu du répondant (en milliers).
U_c: Moyenne des revenus selon la classe sociale des parents; SE : Erreur-type.

Résultats de la régression

Différences de revenus moyens estimés

Résultats de la régression selon l'approche par étapes utilisée

Modèles	Est.	SE	Z	*P*(>|z|)
Approches en étapes				
3 étapes Naïve				
Classe moyenne	8.50	1.41	6.02	p < .001
Classe élevée	16.22	3.83	4.24	p < .001
3 étapes BCH				
Classe moyenne	10.11	1.93	5.24	p < .001
Classe élevée	17.97	4.22	4.26	p < .001
3 étapes ML				
Classe moyenne	23.68	6.27	3.77	p < .001
Classe élevée	40.21	11.58	3.47	p < .001
2 étapes				
Classe moyenne	13.16	8.38	1.57	p = .116
Classe élevée	25.41	6.22	4.09	p < .001

Note. Variable indépendante : Classes sociales (LCA). Variable dépendante : Revenu du répondant (en milliers).
Est. : Coefficient normalisé; SE : Erreur-type; Z: *Two-tailed Z-test*; *P*(>|z|): valeur-p.

Variations des approches par étapes

- Objectif :
  - Trouver un compromis entre les biais de l'approche en une étape et de l'approche naïve en trois étapes
- Forces et faiblesses dépendant des conditions spécifiques (types d'indicateurs et distribution, taille d'échantillon, séparation des classes...)
- Consensus :
  - Le modèle de mesure doit être fixé avant l'ajout de variables externes
- Offre :
  - Logiciels commerciaux (Mplus et Latent GOLD)

Nombreuses offres

Package	Version	R	Python	API Scikit-learn	Approche en 2 étapes	Approches en 3 étapes avec correction	Composantes gaussiennes et non gaussiennes	Covariables
StepMix	2.1.3	✓	✓	✓	✓	✓	✓	✓
scikit-learn	1.2		✓	✓				
multilevLCA	1.1	✓			✓		✓	✓
mclust	6.0	✓						
AutoGMM	2.0.1		✓	✓				
MixMod	0.2.0		✓				✓	
Rmixmod	2.1.8	✓					✓	
poLCA	1.6.0.1	✓						✓
depmixS4	1.5	✓					✓	✓
randomLCA	1.1-2	✓						
BayesLCA	1.9	✓						
e1071::lca	1.7-13	✓						
glca	1.3.3	✓						✓
VarSelLCM	2.1.3.1	✓					✓	
FlexMix	2.3-18	✓					✓	✓

Syntaxe StepMix

```
Model = StepMix(n_components=3,
                measurement='categorical_nan’, structural='continuous_nan',
                n_steps=3, correction = 'BCH’,
                random_state=123,
                n_init=20,
                verbose=1)

Model.fit(df_MM, df_SM)

df_boot = Model.bootstrap_stats(df_MM, df_SM, n_repetitions=1000)
```

Exemples disponibles sur Google Colab

Synthèse

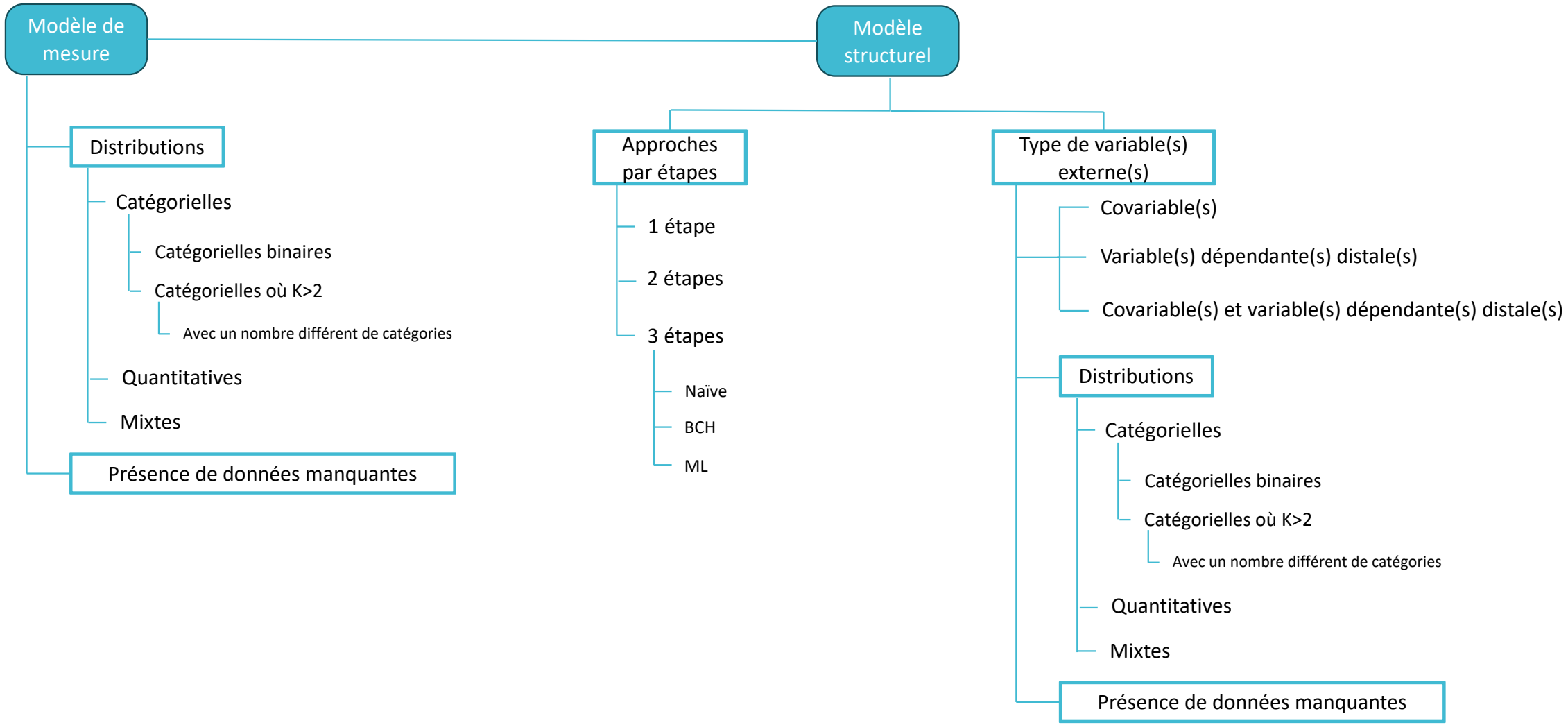
StepMix en bref

- Interface modulable
- Modèle de mélange
  - Plusieurs distributions
  - Avec ou sans variable(s) externe(s)
  - Approches en étapes
  - Données manquantes (FIML)
- Divers utilitaires de simulation
- Bootstrap non paramétrique
  - Intervalles de confiance
  - Erreurs-types, z-tests, p-values

Futurs développements

- Autres distributions
  - Poisson, etc.
- Données longitudinales
  - Analyse de transitions latentes (LTA)
  - Trajectoires par classes latentes (LGMM)

StepMix



(Morin et al., 2023)

Annexes

Références

- Bakk, Z., et Kuha, J. (2018). Two-Step Estimation of Models Between Latent Classes and External Variables. *Psychometrika, 83*(4), 871-892.
- Bakk, Z., et Kuha, J. (2021). Relating Latent Class Membership to External Variables: An Overview. *British Journal of Mathematical and Statistical Psychology, 74*(2), 340–362.
- Bakk, Z., Oberski, D. L., et Vermunt, J. K. (2016). Relating Latent Class Membership to Continuous Distal Outcomes: Improving the Ltb Approach and a Modified Three-Step Implementation. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(2), 278–289.
- Bolck, A., Croon, M., et Hagenaars, J. (2004). Estimating Latent Structure Models with Categorical Variables: One-Step Versus Three-Step Estimators. *Political analysis, 12*(1), 3-27.
- Boudon, R. (1962). Le modèle des classes latentes. *Revue Française de Sociologie, 3*(3), 259.
- Finch, W. H., et French, B. F. (2015). *Latent Variable Modeling with R*. Routledge.
- Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika, 24*(3), 229–252.
- Hsiao, Y.-Y., Kruger, E. S., Lee Van Horn, M., Tofighi, D., MacKinnon, D. P., et Witkiewitz, K. (2021). Latent Class Mediation: A Comparison of Six Approaches. *Multivariate Behavioral Research, 56*(4), 543–557.
- Lazarsfeld, P. F., et Henry, N. W. (1968). *Latent Structure Analysis*. Houghton, Mifflin.
- Nagin, D. S. (2005). *Group-based modeling of development*. Harvard University Press.
- Nylund, K. L., Asparouhov, T., et Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535-569.
- McCutcheon, A. (1987). *Latent Class Analysis*. SAGE Publications, Inc.
- Morin, S., Legault, R., Laliberté, F., Bakk, Z., Giguère, C.-É., de la Sablonnière, R., et Lacourse, É. (2023). StepMix: A Python package for pseudo-likelihood estimation of generalized mixture models with external variables [document soumis pour publication]. *Journal of Statistical Software*.
- Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis, 18*(4), 450-469.