Latent Class Analysis & Gold Standard Imparfait

gilles.hunault@univ-angers.fr

Sommaire :

Version 1.05 du 6 décembre 2014 vers 22 : 18.

1. Variables et classes latentes

A coté des variables observées (dites aussi **manifestes**) dans un tableau de données, on postule des variables **latentes** (on dit aussi hypothétiques, cachées, abstraites, inobservables) telles que, selon G.Saporta (CNAM, 2006) :

1. hypothèse fondamentale : les covariations entre variables observées s'expliquent par la dépendance de chaque variable observée avec les variables latentes.
2. principe d'indépendance conditionnelle : les variables observées sont indépendantes conditionnellement aux variables latentes.

Ainsi pour des tests quantitatifs d'aptitude à l'embauche, on peut postuler une variable latente quantitative discutable **"intelligence"** qui s'exprimerait en points d'intelligence, détectée à partir de tests quantitatifs logiques, spatiaux, lexicaux...

De même, pour une suite de questions à réponses binaires en oui/non sur des attitudes vis à vis d'une communauté donnée on peut trouver une variable qualitative **"tolérance"** ou de "permissivité" à plusieurs modalités comme par exemple : "très tolérant", "peu tolérant", "totalement intolérant"...

En fonction des données observées, les modèles latents portent différents noms :

	V. observées quantitatives	**V. observées qualitatives**
V. latentes quantitatives	Facteurs latents Analyses factorielles (Equations structurales)	Traits latents Réponses à l'item Modèles de Rasch
V. latentes qualitatives	Profils latents (Modèles de mélange particuliers)	Classes latentes

Même si les méthodes et calculs différent, on trouve souvent comme calculs des appproximations par la méthode de Newton-Raphson, des inférences Bayésiennes ou des estimations paramétriques par l'algorithme EM (Expectation Maximisation) qui s'inscrit dans le cadre général du maximum de vraisemblance.

Une fois le modèle statistique fixé et quelques paramètres choisis (choix du nombre de classes, choix de ρ...) on dispose d'une "vraie" modélisation par estimation avec

- des indicateurs de qualité (genre AIC, BIC..)

- des intervalles de confiance (sous certaines hypothèses)

Si on estime une **variable qualitative binaire latente** associée à la caractérisation d'une pathologie, on peut alors estimer des valeurs de *Sensibilité* et de *Spécificité* comme avec un "vrai" *gold standard* et en estimer aussi *la "vraie" prévalence.*

On trouve donc souvent dans la littérature des articles qui utilisent les méthodes statistiques **LCA** (Latent Class Analysis) pour comparer des tests diagnostics. Les autres méthodes comme **DA** (Discrepant Analysis) et **CRS** (Composite Reference Standard) semblent être abandonnées car trop imparfaites.

2. Implémentations

Il existe des sites et des logiciels pour les modèles latents comme par exemple :

- le logiciel " Latent Gold " distribué par Statistical Innovations

- la procédure LCA du logiciel SAS

- le logiciel "Latent Class Analysis" de J. Uebersax

- le package poLCA du logiciel R

- le logiciel TAGS de Pouillot, Gerbier et Gardner

Le package poLCA pour Polytomous variable Latent Class Analysis implémente des calculs *d'analyse de structure latente* nommés aussi *d'analyse de classe latente* et de *régression de classe latente* pour des variables de sortie polytomiques.

3. Applications parisiennes au gold standard imparfait

Les trois articles parisiens fournis par PCA sont assez différents quant aux modèles latents utilisés et aux buts visés.

Dans l'article de 2008 qui traite de l'identification des facteurs de variabilité via la concordance entre LSM et FT, le logiciel TAGS est utilisé avec différents *cutoffs* pour obtenir l'*accuracy* (Pre, Se et Sp) de LSM et FT.

L'article de 2012 présente des résultats liés à la *relative accuracy* de LSM, FT et de la biopsie dans l'hépatite C à l'aide d'un modèle LCM-R (*latent class model with random effects*).

Pour l'article de 2013 qui s'intéresse à l'utilisation de SWE, FT et LSM, des modèles LCM classiques sans effets aléatoires ont été appliqués pour le diagnostic de cirrhose et, pour la fibrose significative – comme aucun LCM ne convenait – des modèles LCM-R ont été employés, avec des "résultats hétérogènes".

4. Discussion et conclusion personnelle

Dans la littérature médicale (humaine et animale) on retrouve souvent des modèles latents car il y a de très nombreuses situations où soit il n'y a pas de *gold standard* (la "vérité" càd la "maladie" n'est pas directement observable) soit il n'existe qu'un *grey standard* càd un *gold standard* imparfait.

Ces articles présentent des analyses statistiques avec des données réelles et des choix de paramètres parfois discutables (ou simplement non étayés scientifiquement) et obtiennent en général des résultats cohérents.

Toutefois, épistémologiquement, il est certainement insoutenable, indéfendable qu'une "vérité cachée" soit révélée par des méthode statistiques.

De plus, la convergence de méthodes d'estimations vers des valeurs compatibles avec "ce qu'on aimerait trouver ou démontrer" présente un attrait considérable surtout si l'analyse de covariables corrobore le choix du nombre de classes et la structure des classes avec des statistiques descriptives interprétables. On se sent alors "justifié" et la notion d'*explication* au sens statistique du terme devient alors une "auto confirmation" des modèles.