

ISABEL MOREAU

**Application d'un modèle de classes latentes
avec dépendance familiale à des données de pedigrees**

Essai présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise en biostatistique
pour l'obtention du grade de Maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES ET GÉNIE
UNIVERSITÉ LAVAL
QUÉBEC

2008

Avant-propos

Je tiens tout d'abord à remercier mes codirecteurs, chercheurs au Centre de recherche de l'Université Laval Robert-Giffard (CRULRG), soit Mme Aurélie Labbe, professeure au département de mathématiques et de statistique, Mme Chantal Mérette, professeure au département de psychiatrie, et M. Alexandre Bureau, professeur au département de médecine sociale et préventive. Leur support, leur encadrement, leur flexibilité et leur grande disponibilité tout au long de mon stage et de ma rédaction d'essai ont été grandement appréciés. Je tiens à remercier plus spécialement Mme Aurélie Labbe pour ces précieux conseils et ces corrections tout au long de ma rédaction d'essai qui me furent très utiles.

Je tiens aussi à remercier M. Arafat Tayeb, étudiant au post-doctorat en statistique qui m'a aidé tout au long de mon stage, que ce soit sur le plan de la statistique ou avec les nombreux programmes R.

Finalement, je voulais remercier le CRULRG pour la bourse d'étude qui m'a été octroyée. Ce soutien financier me fut d'une très grande aide tout au long de ma maîtrise. Merci aussi au personnel du Service de consultation statistique de l'Université Laval avec qui j'ai travaillé pendant deux sessions. Ce travail m'a permis de parfaire mes connaissances en statistique dans divers domaines de recherche.

Table des matières

Avant-propos	ii
Table des matières	iii
Liste des tableaux	v
Table des figures	vi
1. Introduction	1
2. Objectifs	3
3. Données	4
3.1 Sujets à l'étude	4
3.2 Variables à l'étude.....	4
3.3 Description des 4 facteurs retenus.....	5
4. Théorie et méthodologie	7
4.1 Modèle de classes latentes classique	7
4.2 Modèle de classes latentes avec dépendance familiale	7
4.2.1 Intégration de la dépendance familiale dans le modèle.....	8
4.2.2 Notations et description du modèle.....	8
4.2.3 Probabilité de classes et de transition.....	9
4.2.4 Algorithme EM	10
4.3 Descriptions des modèles	13
4.3.1 Modèles multinormaux	13
4.3.2 Modèles multinomiaux.....	14
4.4 Extension du modèle	15
4.4.1 Extension à plusieurs générations	15
4.4.2 Extension au cas de symptômes	16
4.5 Valeurs initiales des paramètres.....	18

4.6 Sélection du meilleur modèle	19
4.6.1 Validation croisée de la vraisemblance	20
4.6.2 BIC	20
5. Résultats	22
5.1.1 Objectif 1 : impacts des valeurs initiales des paramètres de la distribution des mesures sur le choix du nombre de classes	22
5.1.2 Conclusion objectif 1	27
5.2.1 Objectif 2 : Déterminer quel outil statistique on doit utiliser pour trouver le meilleur modèle	27
5.2.2 Conclusion objectif 2	29
5.3 Objectif 3 : Déterminer quels sont les meilleurs modèles parmi ceux proposés pour le jeu de données du CRULRG	29
5.3.1 Modèle final pour les données continues	31
5.3.2 Modèle final pour les données discrètes	32
6. Conclusion.....	34
Bibliographie.....	35
Annexe A	36
Annexe B	39
Annexe C	41

Liste des tableaux

5.1 – Taux de concordance des classes des individus, pour les 6 ensembles de valeurs initiales, pour les modèles multinormaux indépendants à 3 et 7 classes.	26
5.2 – Répartition des individus atteints de schizophrénie et de bipolarité en fonction des classes déterminées par le modèle multinomial sans contrainte	32
5.3 – Répartition des individus atteints de schizophrénie et de bipolarité en fonction des classes déterminées par le modèle multinormal indépendant	33
B.1 – Valeurs des log-vraisemblances totales de validation croisée pour le modèle multinomial sans contrainte	39
B.2 – Valeurs des log-vraisemblances totales de validation croisée pour le modèle multinormal indépendant	39
B.3 – Valeur des BIC pour le modèle multinormal indépendant	40
B.4 – Valeurs des log-vraisemblances totales de validation croisée pour les cinq modèles de distributions des mesures	40
C.1 – Valeurs des log-vraisemblances totales de validation croisée pour le modèle multinomial avec contraintes	42

Table des figures

4.1 – Exemple d’une famille nucléaire.....	8
4.2 – Exemple de familles à plusieurs générations	16
4.3 – Exemple de familles à plusieurs générations avec individus atteints.....	17
5.1 – Log-vraisemblances totales de validation croisée pour le modèle multinomial sans contrainte.....	23
5.2 – Maximum des log-vraisemblances totales de validation croisée pour le modèle multinomial sans contrainte	24
5.3 – Log-vraisemblance totale de validation croisée pour le modèle multinormal indépendant..	25
5.4 – Valeurs des BIC pour le modèle multinormal indépendant	28
5.5 – Log-vraisemblances totales de validation croisée pour les cinq distributions de mesures ..	30
5.6 – Moyenne des 4 facteurs pour chaque classe du modèle multinormal indépendant	31
C.1 – Log-vraisemblances totales de validation croisée pour le modèle multinomial avec contraintes	411

1. Introduction

Les diagnostics de maladies psychiatriques telles que la bipolarité et la schizophrénie sont difficiles à poser puisqu'ils découlent d'une série de critères dont certains sont qualitatifs. La présence de symptômes communs aux deux types de maladie rend encore plus difficile la distinction de ceux-ci. Il est donc difficile de diagnostiquer avec certitude ces maladies.

Les troubles bipolaires sont essentiellement caractérisés par les phases manique (exaltation de l'humeur) et dépressive (perte de notion de plaisir). La schizophrénie, quant à elle, est marquée par des délires, des hallucinations, un isolement social ou une désorganisation de la pensée qui conduit parfois à un langage incompréhensible

La prévalence de la schizophrénie dans la population mondiale est de 1%. Le risque augmente de 10% chez les personnes dont leur père, leur mère, leur frère ou leur sœur en est atteint. Chez les jumeaux dizygotes, le risque augmente de 12% lorsque l'un des deux est atteint et le risque augmente de 50% chez les jumeaux monozygotes. On retrouve sensiblement les mêmes risques pour les troubles bipolaires. L'augmentation du risque en fonction de la proximité familiale démontre ainsi une forte composante génétique pour ces maladies. Cependant, puisque chez les jumeaux monozygotes le risque n'est pas de 100%, d'autres composantes peuvent intervenir dans le développement de cette maladie. L'environnement dans lequel un individu évolue a un fort impact sur l'apparition de ces maladies. Le milieu familial et socio-économique d'un individu ainsi que la prise de drogues influencent donc le développement de la schizophrénie.

Puisqu'une même maladie peut être causée par plusieurs gènes distincts, c'est-à-dire qu'il y a présence d'hétérogénéité génétique, on désire regrouper les individus à l'aide de divers symptômes psychiatriques. Ainsi, en regroupant les individus sur la base de leurs symptômes, on espère obtenir des sous-classes de maladie plus homogènes génétiquement. Ces sous-groupes faciliteraient donc la détection des gènes de vulnérabilité des maladies complexes. L'identification de ces gènes aiderait énormément la compréhension de celles-ci, en particulier par le biais de la recherche de nouveaux traitements.

Le modèle utilisé est celui développé par Labbe, Bureau et Mérette (2008) et est basé sur un modèle d'analyse de classes latentes. Contrairement aux autres modèles proposés dans la littérature, celui-ci a pour avantage de prendre en considération la dépendance familiale prescrite dans les données d'études génétiques.

2. Objectifs

Le principal objectif de ce projet de stage est d'évaluer les propriétés du modèle statistique développé par Labbe et *al.* Ce modèle décrit les mesures multivariées de sujets en tenant compte de la dépendance familiale comme une fonction de classes latentes homogènes de maladies.

Cet objectif se redéfinit plus précisément de la manière suivante :

1. Déterminer si les valeurs initiales de l'algorithme EM du modèle ont un impact sur les résultats.
2. Déterminer quel outil statistique on doit utiliser pour trouver le meilleur modèle.
3. Déterminer quels sont les meilleurs modèles parmi ceux proposés pour le jeu de données du CRULRG.

3. Données

3.1 Sujets à l'étude

Depuis plusieurs années, le CRULG (Centre de recherche Université Laval Robert-Giffard) recueille des données sur des sujets atteints de maladies psychiatriques ainsi que sur les membres de leur famille. La base de données utilisée pour cette étude contient 48 familles pour un total de 1273 individus, dont 483 sujets sont diagnostiqués schizophrène ou bipolaire ou s'approchent de l'une de ces deux maladies sans être diagnostiqués. De ce nombre, 363 sont réellement atteints de l'une des deux maladies. Ces sujets proviennent principalement de trois régions du Québec soit de Québec, de la Beauce et du Saguenay/Lac Saint-Jean. Les familles contiennent entre 12 et 78 personnes et, pour chaque individu (affecté ou non par l'une des 2 maladies), un échantillon de sang a été prélevé.

3.2 Variables à l'étude

Les variables récoltées sont le sexe, l'âge, l'identifiant des parents, l'échantillon d'ADN, le statut¹ de l'individu et 82 symptômes mesurés chez les individus atteints. Ces symptômes sont regroupés en 11 dimensions établies par des psychiatres (voir l'annexe 1 pour plus de détails) :

- délire
- hallucination
- comportement bizarre
- anhédonie
- apathie
- catatonie
- retrait affectif
- pensée désorganisée
- alogie
- manie
- dépression.

¹ Individu atteint avec ou sans symptômes mesurés ou individu non-atteint

Parmi les 82 symptômes, 5 représentent une cote globale pour une dimension donnée. Les cotes globales ne feront pas partie de l'analyse. Ainsi, 77 symptômes ont été utilisés pour cette étude. Ces symptômes ont été mesurés par différentes infirmières du CRULRG à partir du dossier des patients et d'une entrevue réalisée avec ceux-ci. Chaque symptôme est coté sur une échelle de 0 (absent) à 5 (sévère) et est mesuré en phases épisode et inter-épisode. La phase épisodique correspond à la phase aigue où les symptômes sont bien apparents. La phase inter-épisode correspond à un moment où les symptômes sont présents (ou non), mais à un niveau moins élevé. Dans le cadre de ce projet, seules les données en phase épisode sont utilisées.

Suite à un accord inter-juges, le nombre de symptômes a été réduit de 77 à 35. Ces 35 symptômes sont bien soutenus dans la littérature et ont une prévalence élevée. Une analyse en composantes principales (ACP) avec *varimax* comme rotation des axes a permis de regrouper les 35 symptômes les plus représentatifs de ces deux maladies en 4 facteurs : manie, dépression, facteur négatif (absence d'un trait de caractère normalement présent) et facteur positif (présence d'un trait de caractère normalement absent). Ces facteurs se retrouvent parmi les onze dimensions susmentionnées. Les symptômes retenus sont ceux dont les poids de l'ACP sont supérieurs à 0,7 et sont forts pour un facteur (un axe) et faibles sur les autres facteurs. Par la suite, pour chaque individu, deux scores ont été associés à chaque facteur : le score continu, soit la moyenne des symptômes d'un facteur, et le score catégorique, soit le maximum des symptômes d'un facteur. On obtient donc deux jeux de données, un continu et un discret, que nous allons analyser par la suite.

3.3 Description des 4 facteurs retenus

La manie se définit comme un état d'euphorie intense tant sur le plan émotionnel que physique. Par exemple, la personne peut devenir hyperactive, n'avoir besoin que de très peu de sommeil, avoir une accélération de la pensée où tout se bouscule dans sa tête, avoir une très forte estime d'elle, devenir hypersensible ou encore, avoir des projets irréalistes. À l'opposé, la dépression est marquée par une perte de notion de plaisir, l'absence de projet, une perte de confiance en soi ou une fatigue accablante. Les symptômes négatifs représentent les comportements qu'un individu

n'a pas mais devrait avoir et ont généralement pour conséquences le repli sur soi et l'isolement. À l'opposé, les symptômes positifs sont ceux qu'un individu a et ne devrait pas avoir tels que le délire et le sentiment de persécution causés par des hallucinations tant auditives, qu'intrapsychiques ou visuelles.

4. Théorie et méthodologie

4.1 Modèle de classes latentes classique

L'approche classique d'analyse de classes latentes, telle que proposée par Clogg (1995), regroupe les individus ayant des patterns communs de différentes variables mesurées. Deux parties sont inhérentes au modèle de classes latentes : le modèle des classes et le modèle des réponses. Le premier détermine la probabilité qu'un individu appartienne à une classe donnée et le second détermine la probabilité des mesures d'un individu sachant sa classe.

Ce modèle suppose l'indépendance des réponses à l'intérieur des sujets et l'indépendance entre les sujets sachant leur classe. De plus, on suppose l'homogénéité intra classe des variables observées, c'est-à-dire que, sachant la classe, les individus ont la même distribution de mesures.

4.2 Modèle de classes latentes avec dépendance familiale

Le modèle de classes latentes avec dépendance familiale développé par Labbe et *al.* (2008) peut utiliser deux types de mesures, soit :

- les mesures cliniques qui sont prises chez tous les individus telles que des mesures d'attention ou de QI
- les symptômes de maladie qui sont des mesures de symptômes prises chez les individus atteints seulement.

Le cas des mesures cliniques sera présenté dans les prochaines sections et par la suite, le cas des symptômes sera brièvement présenté.

On suppose donc pour l'instant que toutes les mesures sont évaluées sur tous les individus. Le modèle considère que chaque individu appartient à une classe latente que l'on désire identifier à partir de ses mesures cliniques.

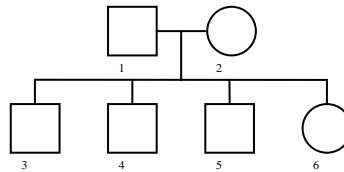
4.2.1 Intégration de la dépendance familiale dans le modèle

Dans le modèle proposé, la dépendance familiale est incorporée au niveau des classes et non au niveau des mesures cliniques. Ainsi, la classe d'un enfant dépend de celle de ses parents et conditionnellement à celles-ci, les classes de frères et de sœurs sont indépendantes.

4.2.2 Notations et description du modèle

Par souci de simplification, le cas d'une seule famille nucléaire constituée de n individus est présenté dans la section 4.2, bien qu'en pratique, on applique le modèle à des familles étendues. Puisque l'indépendance des familles est supposée, la généralisation pour N familles est triviale et est présentée brièvement à la section 4.4.1. La figure 1 présente une famille nucléaire constituée de 6 individus, soit deux parents et quatre enfants. Les hommes sont représentés par des carrés et les femmes par des cercles.

Figure 4.1 - Exemple d'une famille nucléaire



Soit $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ le vecteur de mesures de tous les individus, où $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{id})$ est le vecteur des d mesures de l'individu i , $i = 1, \dots, n$. Les individus sont ordonnés de sorte que \mathbf{Y}_1 et \mathbf{Y}_2 représentent les mesures de la mère et du père respectivement.

On suppose que chaque individu peut être assigné à l'une des K classes latentes. On note C_i la classe du i^e individu. Le vecteur $\mathbf{C} = (C_1, \dots, C_n)$ représente les classes non observées des individus de la famille. On assume le modèle suivant pour les mesures cliniques sachant la classe non observée d'un individu i :

$$\mathbf{Y}_i | C_i = c \sim f(y_i, \theta_c)$$

où θ_c représente les paramètres de la distribution des mesures pour la classe c . La distribution de f dépend du type de mesures (continues ou discrètes). Si les mesures sont continues, on peut utiliser la distribution multinormale, si elles sont catégoriques, on peut utiliser un modèle multinomial. Ces modèles sont présentés en détails à la section 4.3.

4.2.3 Probabilité de classes et de transition

Puisqu'on assume la dépendance entre les individus au niveau de la classe et que la classe d'un enfant dépend uniquement de celles de ses parents, on définit pour la famille:

$$\begin{aligned} P(C_1 = c_1, \dots, C_n = c_n) &= \prod_{i=3}^n P(C_i = c_i | C_1 = c_1, C_2 = c_2) \cdot P(C_1 = c_1) \cdot P(C_2 = c_2) \\ &= \prod_{i=3}^n p_{c_i, c_1, c_2} \cdot p_{c_1} \cdot p_{c_2}. \end{aligned}$$

Ainsi, conditionnellement à la classe des parents, les classes de frères et sœurs sont indépendantes.

La vraisemblance des données sachant $\Delta = (p_{c_1}, p_{c_2}, p_{c, c_1, c_2}, \theta_c)$, soit l'ensemble de tous les paramètres du modèle, est donné par :

$$f(y | \Delta) = \sum_{c_1, c_2, \dots, c_n} \left[\prod_{i=1}^n f(y_i, \theta_{c_i}) \right] \cdot \left[\prod_{i=3}^n p_{c_i, c_1, c_2} \right] \cdot p_{c_1} \cdot p_{c_2}$$

avec les contraintes suivantes sur les paramètres des classes :

$$\sum_{c=1}^K p_c = 1, \tag{4.1}$$

$$\sum_{c=1}^K p_{c, c_1, c_2} = 1 \text{ pour tout } (c_1, c_2) \in \{1, \dots, K\}^2. \tag{4.2}$$

Avec ces seules contraintes sur les paramètres de classes, on doit estimer $(K-1)(K^2+1)$ paramètres pour le modèle de classes latentes. Ainsi, le nombre de paramètres à estimer augmente rapidement avec le nombre de classes dans le modèle. Pour diminuer le nombre de paramètres des classes, on spécifie deux contraintes optionnelles. La première contrainte possible est celle de symétrie sur les probabilités de transition:

$$p_{c,c_1,c_2} = p_{c,c_2,c_1} \text{ pour tout } (c, c_1, c_2) \in \{1, \dots, K\}^3. \quad (4.3)$$

Cette contrainte est acceptable sur le plan de la génétique si on assume que la maladie n'est pas reliée au sexe de l'individu, c'est-à-dire que le gène de maladie n'est situé ni sur le chromosome X ni sur le chromosome Y, ce qui est le cas pour bon nombre de maladies complexes, dont la schizophrénie et la bipolarité.

On peut aussi imposer la contrainte parentale qui implique que la classe d'un enfant est soit celle de son père, soit celle de sa mère :

$$p_{c,c_1,c_2} = 0 \text{ si } c \neq c_1 \text{ et } c \neq c_2. \quad (4.4)$$

4.2.4 Algorithme EM

Les paramètres du modèle peuvent être estimés avec le maximum de vraisemblance, par le biais de l'algorithme EM (McLachlan et Krishnan, 2007). Cette méthode itérative permet de converger vers le maximum local de vraisemblance et s'effectue en deux étapes : une étape E (espérance) et une étape M (maximisation). La première étape consiste à calculer l'espérance de la log-vraisemblance complète (c'est-à-dire en supposant les classes non observées connues) et la seconde, à trouver le maximum local. L'algorithme s'arrête lorsque la différence entre l'espérance de deux log-vraisemblances complètes consécutives $(t, t+1)$ est inférieure à la tolérance fixée. Encore une fois, cette méthode est présentée pour une famille nucléaire et la généralisation à N familles est triviale puisqu'on suppose les familles indépendantes. $\Delta^{(t)}$ est

l'estimation des paramètres après la t^e itération de l'algorithme. Ainsi, l'espérance de la log-vraisemblance sur toutes les données sachant les données observées est :

$$Q[\Delta | \Delta^{(t)}] = \sum_{c=1}^K \left[\sum_{i=1}^n w_{ic}^{(t)} \log f(y_i, \theta_c) + \sum_{c=1}^K (w_{1c}^{(t)} + w_{2c}^{(t)}) \log p_c \right. \\ \left. + \sum_{c_1=1}^K \sum_{c_2=1}^K \sum_{i=3}^n w_i^{(t)}(c, c_1, c_2) \cdot \log(p_{c, c_1, c_2}) \right] \quad (4.5)$$

où les poids, $w_i^{(t)}$, représentent les probabilités postérieures d'appartenir aux classes pour un trio père, mère et enfant à l'itération t et sont donnés par :

$$w_i^{(t)}(c, c_1, c_2) = P(C_i = c, C_1 = c_1, C_2 = c_2 | \mathbf{y}, \Delta^{(t)}) \\ = \frac{P(C_i = c, C_1 = c_1, C_2 = c_2, \mathbf{y} | \Delta^{(t)})}{\sum_{c=1}^K \sum_{c_1=1}^K \sum_{c_2=1}^K P(C_i = c, C_1 = c_1, C_2 = c_2, \mathbf{y} | \Delta^{(t)})} \quad (4.6)$$

pour $(c, c_1, c_2) \in \{1, \dots, K\}^2$, $i = 3, \dots, n$ et où

$$P(C_i = c, C_1 = c_1, C_2 = c_2 | \mathbf{y}, \Delta) = f(\mathbf{y}_i, \theta_{c_i}) \cdot f(\mathbf{y}_1, \theta_{c_1}) \cdot f(\mathbf{y}_2, \theta_{c_2}) \cdot p_{c_1} \cdot p_{c_2} \cdot p_{c, c_1, c_2} \\ \cdot \prod_{\substack{j \neq i \\ j \geq 3}}^K \sum_{c_j=1}^K f(\mathbf{y}_j, \theta_{c_j}) \cdot p_{c_j, c_1, c_2}.$$

Les poids marginaux des enfants à l'itération t , $w_{ic}^{(t)}$ avec $i \geq 3$, sont définis par :

$$w_{ic}^{(t)} = P(C_i = c | \mathbf{y}, \Delta^{(t)}) \\ = \sum_{c_1=1}^K \sum_{c_2=1}^K P(C_i = c, C_1 = c_1, C_2 = c_2 | \mathbf{y}, \Delta^{(t)}) \quad (4.7) \\ = \sum_{c_1=1}^K \sum_{c_2=1}^K w_i^{(t)}(c, c_1, c_2).$$

Les poids des parents, $w_{ic}^{(t)}$ avec $i = 1$ ou 2 , sont définis par les équations suivantes :

$$w_{1c_1}^{(t)} = \sum_{i=3}^n \sum_c \sum_{c_2} w_{ic}^{(t)}(c, c_1, c_2), \text{ pour la mère} \quad (4.8)$$

$$w_{1c_2}^{(t)} = \sum_{i=3}^n \sum_c \sum_{c_1} w_{ic}^{(t)}(c, c_1, c_2), \text{ pour le père} \quad (4.9)$$

Ainsi, on somme les poids marginaux des enfants sur toutes les classes possibles des enfants et sur toutes les classes possibles du conjoint.

Dans l'étape M de l'algorithme, la fonction $Q(\Delta | \Delta^{(t)})$ est maximisée sous les contraintes données par 4.1, 4.2, 4.3 et 4.4 si applicables. Sous ces contraintes, les probabilités de transition de classes à l'itération $(t+1)$ sont :

$$p_{c,c_1,c_2}^{(t+1)} = \frac{\sum_{i=3}^n w_i^{(t)}(c, c_1, c_2)}{\sum_{i=3}^n \sum_{c=1}^K w_i^{(t)}(c, c_1, c_2)} \quad (4.10)$$

Pour les probabilités des fondateurs, on trouve :

$$p_c^{(t+1)} = \frac{w_{1c}^{(t)} + w_{2c}^{(t)}}{2}. \quad (4.11)$$

L'estimation des paramètres θ_c dépend de la densité ou de la distribution de mesures utilisée (voir la section 4.3).

4.3 Descriptions des modèles

Comme il a été mentionné précédemment, le choix de la densité ou de la distribution dépend du type de variables utilisé. Puisque les mesures cliniques sont de type continu ou catégorique, deux modèles sont considérés : le modèle multinormal et le modèle multinomial.

4.3.1 Modèles multinormaux

Avec des mesures continues, la loi multinormale est utilisée. On note $Y_i | C_i = c \sim MN(\mu_c, \Sigma_c)$ où $\mu_c = (\mu_1, \dots, \mu_k)$ et Σ_c est la matrice de variances-covariances de dimension k pour la classe c . Ces paramètres sont estimés par l'algorithme EM.

Certaines contraintes peuvent être ajoutées sur les paramètres pour obtenir différents modèles multinormaux. On peut forcer le modèle à avoir des variances et des covariances différentes pour chaque mesure d'une classe, mais avoir les mêmes valeurs d'une classe à l'autre. C'est-à-dire qu'on a une seule matrice de variance-covariance pour toutes les classes ayant la forme suivante :

$$\Sigma_c = \begin{pmatrix} \sigma_1^2 & \alpha_{12} & \dots & \alpha_{1d} \\ \alpha_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha_{(d-1)d} \\ \alpha_{d1} & \dots & \alpha_{d(d-1)} & \sigma_d^2 \end{pmatrix}$$

où α_{ij} est la covariance entre les mesures i et j .

On peut aussi fixer les mêmes variances et covariances pour chaque mesure d'une classe, mais les faire varier d'une classe à l'autre. C'est-à-dire que peu importe le niveau de mesure, pour une classe donnée, on a les mêmes variances et covariances. Ainsi, pour chaque classe c , on obtient la matrice de variance-covariance suivante :

$$\Sigma_c = \begin{pmatrix} \sigma_c^2 & \alpha_c & \cdots & \alpha_c \\ \alpha_c & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha_c \\ \alpha_c & \cdots & \alpha_c & \sigma_c^2 \end{pmatrix}$$

où α_c est la covariance entre 2 mesures de la classe c .

On peut aussi supposer l'indépendance des mesures et ainsi avoir une matrice de variance-covariance diagonale où la variance varie pour chaque mesure et pour chaque classe. Ainsi, pour chaque classe c , on obtient la matrice de variance-covariance suivante :

$$\Sigma_c = \begin{pmatrix} \sigma_{c1}^2 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_{cd}^2 \end{pmatrix}.$$

4.3.2 Modèles multinomiaux

Si les mesures sont catégoriques, on utilise une distribution multinomiale. On note donc Y_{ij} la j^e mesure du i^e individu, où $Y_{ij} \in \{1, \dots, s\}$. On peut ainsi utiliser le modèle logit multinomial :

$$P(Y_{ij} \leq y | C_i = c) = \frac{e^{\alpha_{c j_1 + \dots + c j_y}}}{1 + e^{\alpha_{c j_1 + \dots + c j_y}}}, y = 1, \dots, s - 1 \quad (4.9)$$

où i représente l'individu, j la mesure et y la valeur de la mesure. Un tel modèle suppose l'indépendance des Y_{ij} conditionnellement à la classe. Ainsi les paramètres à estimer dans chaque étape M et pour chaque classe c sont : $\theta_c = (\theta_{c_1}, \dots, \theta_{c_d})$, où $\theta_{c_j} = (\alpha_{c_{j_1}}, \dots, \alpha_{c_{j(s-1)}})$ et

$$\alpha_{c_{j1}}^{(t+1)} = \log \left[\frac{P^{(t)}(Y_{ij} = 1 | C_i = c)}{1 - P^{(t)}(Y_{ij} = 1 | C_i = c)} \right] = \log \left[\frac{\frac{\sum_{i=1}^n w_{ic}^{(t)} I(Y_{ij} = 1)}{\sum_{i=1}^n w_{ic}^{(t)}}}{1 - \frac{\sum_{i=1}^n w_{ic}^{(t)} I(Y_{ij} = 1)}{\sum_{i=1}^n w_{ic}^{(t)}}}} \right]$$

où les poids w_{ic} sont définis en 4.7, 4.8 et 4.9 et où la fonction I est une fonction indicatrice. On a de plus :

$$\alpha_{c_{jy}}^{(t+1)} = \log \left[\frac{P^{(t)}(Y_{ij} \leq y | C_i = c)}{1 - P^{(t)}(Y_{ij} \leq y | C_i = c)} \right] - \sum_{l=1}^{y-1} \alpha_{c_{jl}}^{(t+1)} = \log \left[\frac{\frac{\sum_{i=1}^n w_{ic}^{(t)} I(Y_{ij} \leq y)}{\sum_{i=1}^n w_{ic}^{(t)}}}{1 - \frac{\sum_{i=1}^n w_{ic}^{(t)} I(Y_{ij} \leq y)}{\sum_{i=1}^n w_{ic}^{(t)}}}} \right] - \sum_{l=1}^{y-1} \alpha_{c_{jl}}^{(t+1)}$$

pour $y = 2, \dots, s-1$.

Pour $y = s$, on a :

$$\alpha_{c_{js}}^{(t+1)} = 1 - \sum_{l=1}^{s-1} \alpha_{c_{jl}}^{(t+1)}.$$

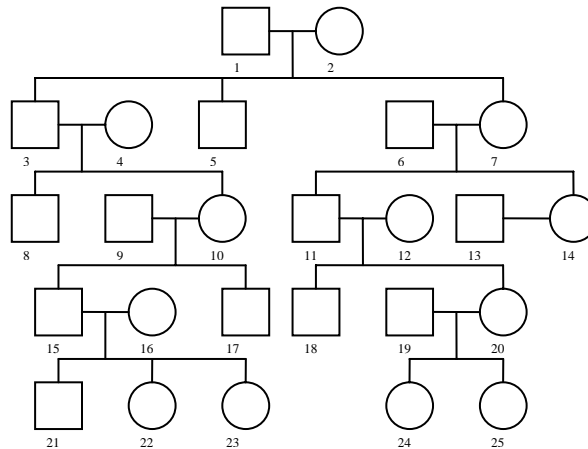
On peut aussi spécifier un modèle probit multinomial avec la contrainte ordinale suivante: $\alpha_{c_{jy}} = \alpha_{c'_{jy}}$ pour tout (c, c') et $y > 1$. Sous cette contrainte, les classes sont ordonnées pour chaque mesure.

4.4 Extension du modèle

4.4.1 Extension à plusieurs générations

Le modèle de dépendance familiale peut être étendu au cas de familles à plusieurs générations. On assume ici que toutes les familles nucléaires d'un pedigree sont des descendants du même couple fondateur et que le pedigree ne contient pas de boucle entre les générations. La figure 3 présente une famille à plusieurs générations où le couple fondateur est constitué des individus 1 et 2.

Figure 4.2 – Exemple de familles à plusieurs générations



Afin de généraliser le modèle de Labbe et *al.*, un nouveau type d'individu a été créé : les connecteurs (par exemple, l'individu 3 de la figure 4.2). Ce sont les sujets qui sont à la fois parents et enfants, c'est-à-dire ceux qui connectent deux familles nucléaires l'une à l'autre. Évidemment, l'extension du modèle augmente le nombre de paramètres à estimer. Ainsi, le principal problème survient lors de l'étape E de l'algorithme EM où le calcul des poids entraîne de longs calculs dont certains sont répétés maintes fois pour chaque individu et pour chaque triplet d'enfants-parents. Le temps d'exécution de l'algorithme, tout comme l'espace mémoire requis lors de chaque itération, est donc élevé pour de grands pedigrees. Pour régler ce problème, une méthode de « peeling » est utilisée. Cette méthode consiste à séparer le pedigree de manière à effectuer les calculs en deux étapes : « descente » et « montée », une technique souvent utilisée en statistique génétique. Le but de cette méthode est de calculer chaque valeur une seule fois et seulement lorsque nécessaire. Les détails de cette méthode ne sont pas présentés ici.

4.4.2 Extension au cas de symptômes

Le modèle peut s'appliquer au cas où l'on mesure des symptômes sur les individus atteints seulement. Ainsi, ce ne sont pas tous les individus d'une famille qui ont des mesures disponibles pour l'analyse. Cependant, certains de ces individus sans mesure clinique doivent être conservés puisqu'ils sont les connecteurs entre les différentes familles nucléaires (par exemple l'individu 11 de la figure 4.3). On doit donc ajouter la variable suivante :

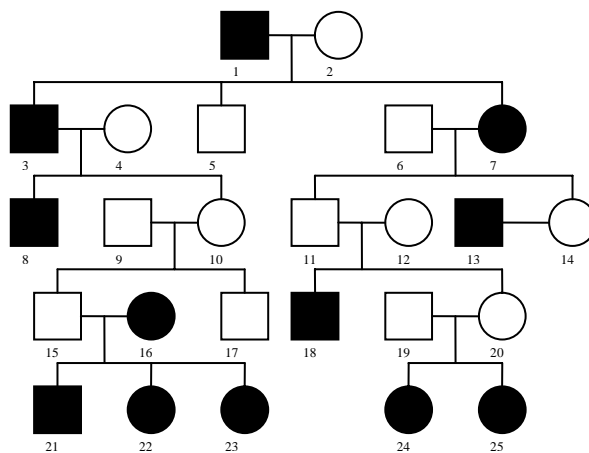
$$S_i = \begin{cases} 1, & \text{si les symptômes de l'individu } i \text{ ont été mesurés} \\ 0, & \text{sinon} \end{cases}$$

On considère ainsi que les individus ayant des mesures manquantes (non mesurées) sont soit asymptomatiques (non-affecté) soit atteint sans que leurs symptômes n'aient été mesurés pour diverses raisons. On suppose que les individus non-atteints peuvent être :

- porteurs d'allèles causant la maladie et peuvent transmettre ces allèles à leurs enfants (de tels individus peuvent donc appartenir à une classe de maladie au même titre qu'une personne atteinte)
- non-porteurs d'allèles causant la maladie et ne peuvent donc pas transmettre d'allèles de maladie à leurs enfants (de tels individus appartiennent à la classe 0, représentant la classe des non-affectés et non-porteurs).

Dans le modèle présenté, on considère que tous les sujets n'ayant pas de mesures cliniques disponibles sont non-atteints par la maladie, c'est-à-dire qu'on ne fait pas la distinction entre les non-atteints et ceux atteints n'ayant pas de symptôme mesuré. De plus, puisque les enfants sans symptôme n'apportent aucune information dans le modèle, ils sont exclus de l'analyse. La figure 4.2 est donc remplacée par la figure 4.3, où les formes noircies représentent des individus atteints de la maladie.

Figure 4.3 – Exemple de familles à plusieurs générations avec individus atteints



4.5 Valeurs initiales des paramètres

Pour utiliser l'algorithme EM, il faut fixer les valeurs initiales des paramètres à estimer à l'itération 0. Certains paramètres, ceux des classes, peuvent être fixés indépendamment du jeu de données tandis que pour d'autres paramètres, ceux des réponses, il est préférable de tenir compte de l'information donnée par les observations.

On considère, par exemple, que la probabilité qu'un fondateur appartienne à la classe c est :

$$p_c^{(0)} = \frac{1}{k+1}, c = \{0, \dots, K\}.$$

En ce qui concerne les probabilités de transition, la probabilité qu'un enfant appartienne à la classe c , sachant celles de ses parents est :

$$p_{c,c_1,c_2}^{(0)} = \begin{cases} \frac{1}{2}, & \text{si } c = c_1 \text{ ou } c = c_2 \\ 1, & \text{si } c_1 = c_2 \\ 0, & \text{sinon} \end{cases}$$

pour $c_1 \neq 0$ et $c_2 \neq 0$

$$p_{c,c_1,c_2}^{(0)} = \begin{cases} 1, & \text{si } c = c_1 \text{ ou } c = c_2 \\ 0, & \text{sinon} \end{cases}$$

pour $c_1 = 0$ ou $c_2 = 0$.

En ce qui concerne les paramètres initiaux de la distribution des symptômes (θ_c), ils sont estimés à partir des données. Pour ce faire, on forme des classes a priori des individus. Trois méthodes sont proposées pour former des groupes initiaux et estimer les paramètres. Ainsi, pour un même ensemble de données, les paramètres initiaux peuvent être différents. On pourra donc évaluer l'impact des valeurs initiales de l'algorithme EM sur la classification finale des individus.

La première méthode (*permut*) consiste à former K classes de même taille (plus ou moins un individu) en répartissant les individus de façon aléatoire dans les classes.

Pour la seconde (*cluster_ind*), on effectue un *clustering* à partir des symptômes de tous les individus pour former K classes. On espère ainsi obtenir des résultats initiaux plus près de ceux finaux car la composition initiale des classes tient compte des symptômes, contrairement à la première méthode.

Pour la troisième méthode (*cluster_fam*), on effectue un *clustering*, mais cette fois, sur un seul individu par famille. Cet individu est choisi aléatoirement. Par la suite, on place tous les individus d'une famille dans la même classe. Cette méthode se rapproche donc du modèle de dépendance familiale proposé dans cet essai.

Par la suite, on estime les paramètres initiaux de la distribution des symptômes pour chaque classe formée avec l'information a priori. Pour les modèles multinormaux, on calcule directement la moyenne et on construit la matrice de variances-covariances à partir des k classes formées par chaque méthode. Pour le modèle multinomial, on utilise le modèle logit ou probit présentés à la section 4.3. Ces paramètres seront utilisés à la première itération de l'algorithme EM.

4.6 Sélection du meilleur modèle

De nombreux modèles peuvent être obtenus à partir d'un même jeu de données en variant le nombre de classes latentes, le type des variables observées ou encore, en ajoutant diverses contraintes sur la paramétrisation de la distribution des symptômes. Pour parvenir à déterminer quel modèle fournit la meilleure approximation de la vraie distribution des symptômes, des méthodes statistiques doivent être utilisées. De nombreuses méthodes ont été proposées dans la littérature. Cependant, due à la structure de dépendance familiale considérée dans le modèle de classes latentes, nombre d'entre elles ne peuvent être utilisées. Ainsi, deux méthodes ont été retenues pour ce projet soient la validation croisée de la vraisemblance (van der Laan et *al.*, 2004) et le *Bayesian Information Criterion*, *BIC*, (Schwarz, 1978).

4.6.1 Validation croisée de la vraisemblance

Avec cette méthode, la vraisemblance d'un modèle est calculée sur un échantillon indépendant de celui servant à l'estimation des paramètres. On peut ainsi comparer les vraisemblances de modèles n'ayant pas le même nombre de paramètres. Dans le modèle développé par Labbe et *al.*, la division du jeu de données en sous-groupes « d'entraînements » et de « tests » se fera en termes de familles et non d'individus, puisque les familles sont considérées indépendantes contrairement aux individus. L'algorithme utilisé est le suivant (on assume que les familles sont dans un ordre aléatoire) :

1. Diviser les N familles en H sous-groupes.
2. Pour chaque sous-groupe h , créer un « échantillon test » en utilisant les familles $(h-1)N/H + 1$ à hN/H . L'« échantillon d'entraînement » est donc créé avec les familles restantes.
3. Pour chaque modèle testé et pour chaque sous-groupe h :
 - a. Estimer les paramètres du modèle à l'aide de l'algorithme EM en utilisant « l'échantillon d'entraînement »
 - b. Calculer la log-vraisemblance, $l_{\text{modèle}}^h$, en utilisant les paramètres estimés en a) sur « l'échantillon test »
4. Pour chaque modèle testé on calcule : $l_{\text{modèle}} = \sum_{h=1}^H l_{\text{modèle}}^h$
5. On choisit le modèle pour lequel $l_{\text{modèle}}$ est maximal
6. On réestime les paramètres du modèle choisi, mais cette fois sur l'ensemble des données.

4.6.2 BIC

Cette méthode a été utilisée puisque la validation croisée ne permettait pas toujours d'obtenir des résultats concluants dû à la présence de plateau dans les graphiques de validation croisée en fonction du nombre de classes (voir la section 5.1.1). La méthode du BIC pénalise les modèles

plus complexes en tenant compte du nombre de paramètres à estimer. Le meilleur modèle est celui pour lequel le BIC est minimal. On a :

$$BIC = -2\log(L) + k\ln(n)$$

où L est la vraisemblance maximisée du modèle,

n est le nombre d'observations et

k est le nombre de paramètres à estimer.

Puisque le BIC assume que les observations sont indépendantes et que dans le modèle proposé les individus ne sont pas indépendants mais les familles le sont, on considère trois cas comme valeurs de la variable n : le nombre total d'individus, le nombre de familles et la moyenne de ces deux nombres.

Peu importe le modèle utilisé, avec les contraintes de symétrie et parentale (équations 4.3 et 4.4), on a toujours le même nombre de paramètres de probabilités à estimer, soit : $2 + K(K + 1)$. Cependant, le nombre de paramètres à estimer pour la distribution des symptômes varie d'un modèle à l'autre. Ainsi, par exemple, pour le modèle multinormal indépendant, on a $2 \cdot K \cdot d$ paramètres de distribution à estimer. Soit, une moyenne et une variance pour chaque classe et pour chaque symptôme. Pour le modèle multinomial avec contraintes, on a $(K + s - 2) \cdot d$ paramètres de distribution à estimer, soit, un paramètre pour le premier niveau de chaque symptôme et de chaque classe, $(K \cdot d)$, et un paramètre pour chaque niveau y , ($y = 2, \dots, s - 1$), de chaque symptôme.

5. Résultats

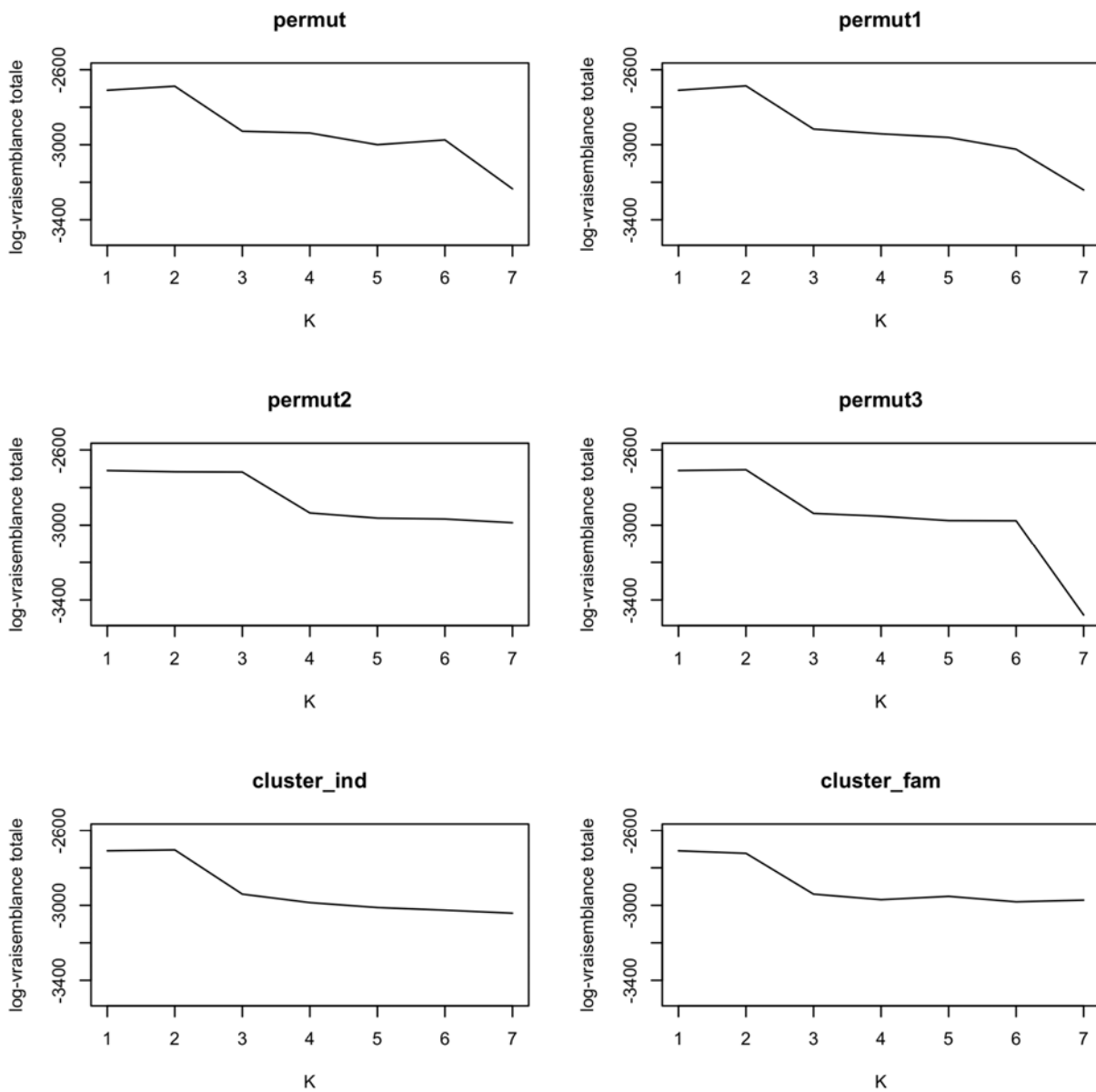
Il est à noter que tous les résultats présentés sont ceux obtenus à partir des symptômes en phase épisodique et avec le modèle ayant la contrainte parentale et celle de symétrie. De plus, lors de la validation croisée, 5 sous-groupes ont été formés et ce sont toujours les mêmes sous-groupes qui sont utilisés peu importe le modèle testé. De plus, on fait varier le nombre de classes (K) de 1 à 7 et les cinq types de distributions de symptômes présentées à la section 4.3 ont été ajustées aux données. Ainsi pour le jeu de données continu, on ajuste les trois types de distributions multinormales et pour le jeu de données discret, on ajuste les deux distributions multinomiales.

5.1.1 Objectif 1 : impacts des valeurs initiales des paramètres de la distribution des mesures sur le choix du nombre de classes

À partir du jeu de données, on a créé 6 ensembles de K classes initiales d'individus (K allant de 1 à 7) à partir desquelles les valeurs initiales de la distribution des symptômes ont été estimées. Pour ces 6 ensembles, quatre ont été créés avec la méthode *permut*, un avec la méthode *cluster_ind* et un avec la méthode *cluster_fam*. Par la suite, pour chacun des 5 types de modèle de réponses (multinormal ou multinomial avec ou sans contraintes), on a évalué l'impact des valeurs initiales de l'algorithme EM à partir de ces ensembles de classes initiales. Par souci de parcimonie, seuls les résultats des modèles multinomial sans contrainte et multinormal indépendant sont présentés.

Les graphiques de la figure 5.1 présentent les résultats de la validation croisée pour les données discrètes, c'est-à-dire pour le modèle multinomial sans contrainte, pour chaque ensemble de classes initiales et pour chaque modèle dont le nombre de classes varie de 1 à 7.

Figure 5.1 – Log-vraisemblances totales de validation croisée² pour le modèle multinomial sans contrainte

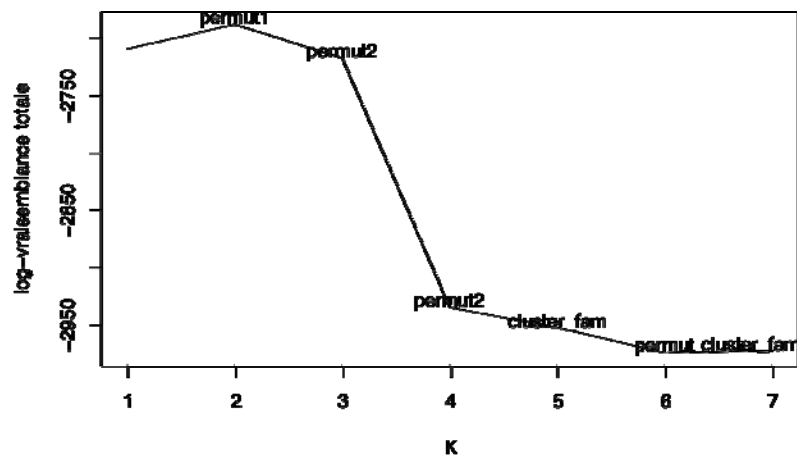


² Le tableau B.1 de l'annexe B présente les valeurs des log-vraisemblances

On constate que 4 des 6 ensembles obtiennent un maximum avec le modèle à 2 classes, tandis que pour l'ensemble *permut2* et *cluster_fam*, le maximum est atteint avec le modèle à 1 classe.

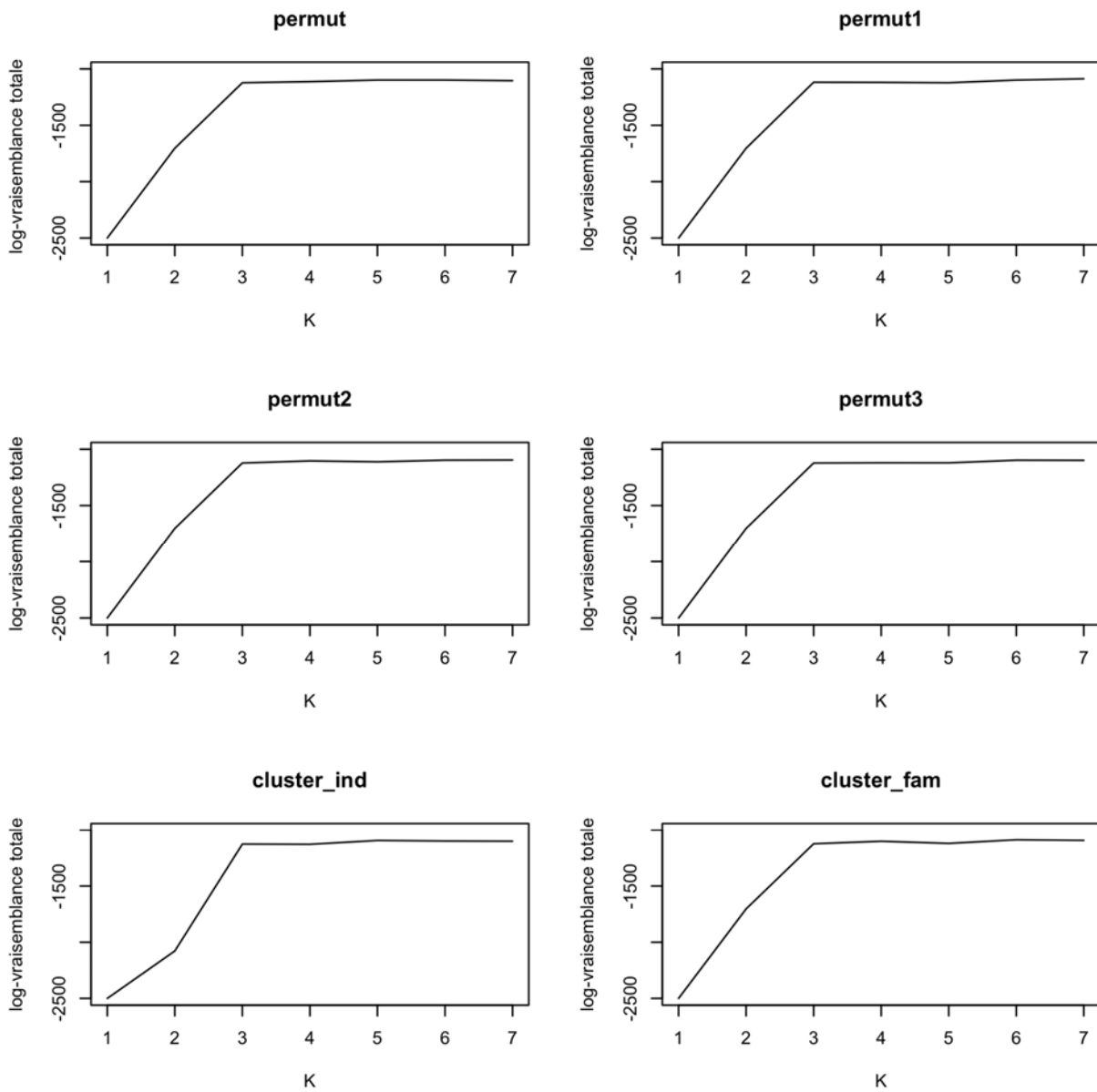
Le graphique de la figure 5.2 présente, parmi les 6 ensembles de valeurs initiales, le maximum de log-vraisemblance totale de validation croisée pour chaque modèle à k classes, $k=1, \dots, 7$.

Figure 5.2 – Maximum des log-vraisemblances totales de validation croisée pour le modèle multinomial sans contrainte



L'ensemble de valeurs initiales pour lequel la log-vraisemblance est maximale varie selon le nombre de classes (K) du modèle ajusté. Ainsi, pour le modèle multinomial sans contrainte, le modèle optimal serait celui à 2 classes avec l'ensemble de valeurs initiales *permut1*.

Figure 5.3 - Log-vraisemblance totale de validation croisée³ pour le modèle multinormal indépendant



Peu importe l'ensemble de valeurs initiales utilisé, les graphiques de la figure 5.3 sont très semblables. On constate que la log-vraisemblance augmente toujours en fonction du nombre de classes: il y a donc présence d'un « plateau ». C'est-à-dire que la log-vraisemblance totale

³ Le tableau B.2 de l'annexe B présente les valeurs des log-vraisemblances

n'atteint pas un maximum pour redescendre par la suite. Il est donc difficile de déterminer le nombre de classes optimal pour ce modèle.

Le tableau 5.1 présente, pour les modèles multinormaux indépendants à 3 et à 7 classes (soit le nombre de classes de début et de fin du plateau), pour chaque combinaison d'ensemble de valeurs initiales, les taux de concordance entre les classes a posteriori des individus.

Tableau 5.1 – Taux de concordance des classes des individus, pour les 6 ensembles de valeurs initiales, pour les modèles multinormaux indépendants à 3 et 7 classes.

	3 classes	7 classes
Ensembles de valeurs initiales	% d'individus qui <u>sont</u> dans la même classe	% d'individus qui <u>sont</u> dans la même classe
<i>Permut – Permut1</i>	100	56.7
<i>Permut – Permut2</i>	100	71.9
<i>Permut – Permut3</i>	100	84.0
<i>Permut – Cluster_ind</i>	100	75.5
<i>Permut – Cluster_fam</i>	100	81.8
<i>Permut1 – Permut2</i>	100	59.0
<i>Permut1 – Permut3</i>	100	58.1
<i>Permut1 – Cluster_ind</i>	100	71.6
<i>Permut1 – Cluster_fam</i>	100	67.2
<i>Permut2 – Permut3</i>	100	67.8
<i>Permut2 – Cluster_ind</i>	100	68.0
<i>Permut2 – Cluster_fam</i>	100	65.8
<i>Permut3 – Cluster_ind</i>	100	79.1
<i>Permut3 – Cluster_fam</i>	100	84.6
<i>Cluster_ind – Cluster_fam</i>	100	89.5

Ainsi, par exemple, qu'on utilise les ensembles de valeurs initiales *permut* ou *permut1*, on obtient la même classification des individus a posteriori (taux de 100%). Il en est de même pour toutes les combinaisons d'ensembles de valeurs initiales de ce modèle. Ainsi, pour le modèle à 3 classes, les valeurs initiales de l'algorithme EM n'ont pas d'impact sur les classes des individus.

Cependant, pour le modèle à 7 classes, les taux de concordance des classes des individus varient entre 56.7 et 89.5 %. Par exemple, près de 60% des individus sont classés dans les mêmes

classes, qu'on utilise les ensembles de valeurs initiales *permut* ou *permut1*. Pour ce modèle, les valeurs initiales ont donc un impact sur les résultats finaux.

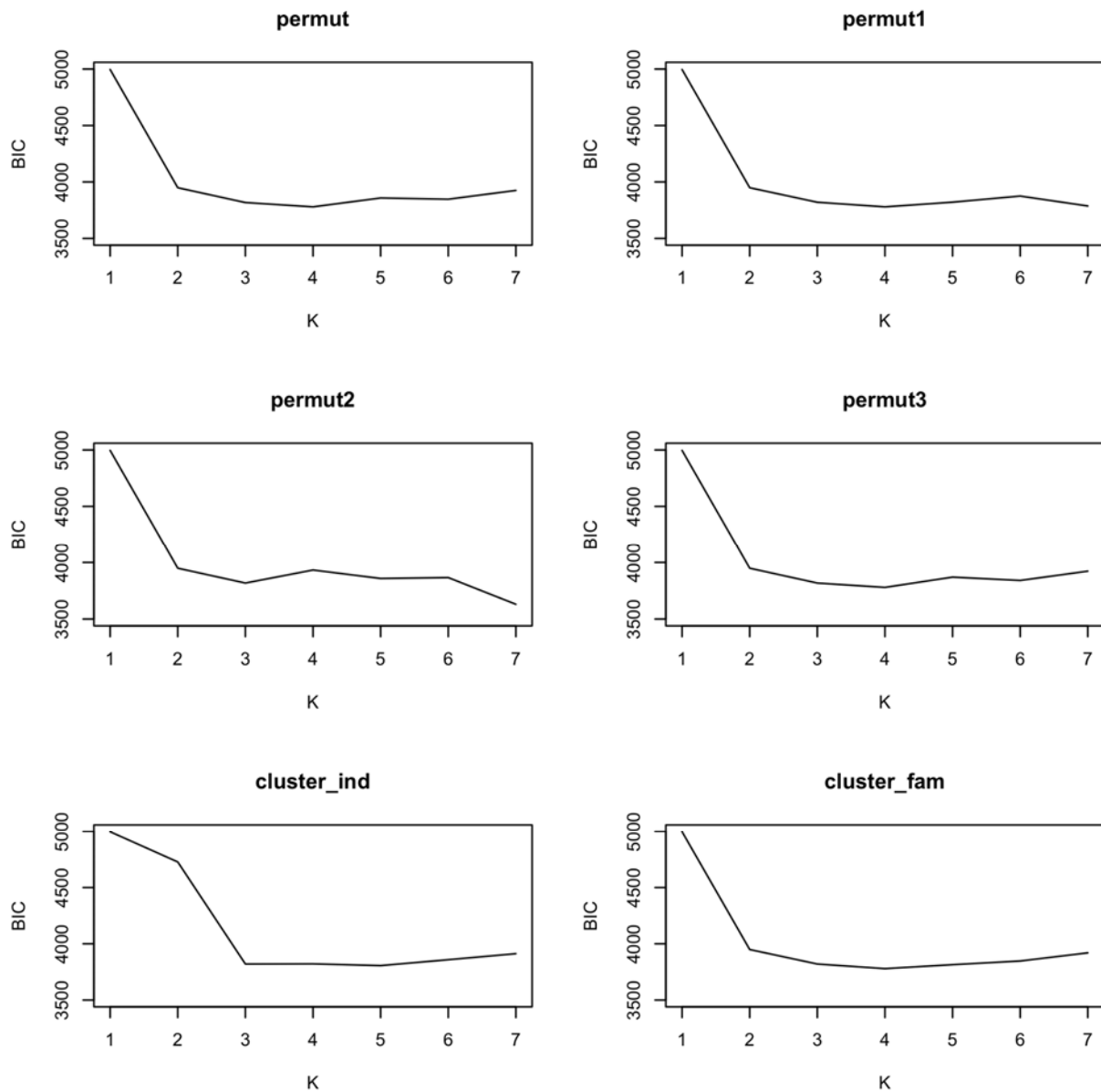
5.1.2 Conclusion objectif 1

Lorsque le nombre de classes du modèle est faible, les valeurs initiales ont peu d'impact sur la classification des individus. Par contre, lorsque ce nombre est plus élevé, il est clair que les valeurs initiales ont un impact sur la classification. Cependant, il est difficile de déterminer quelle méthode il est préférable d'utiliser pour former les groupes à partir desquels les variables initiales seront estimées. En effet, les méthodes de *clustering* ne semblent pas apporter plus d'informations que celles de permutation. Puisque le meilleur ensemble de valeurs initiales varie en fonction de la distribution des mesures utilisée et du nombre de classes fixé, il semble impossible de pouvoir en déterminer un qui donnera le meilleur modèle final. Ainsi, lorsque le nombre de classes est relativement élevé, il est préférable d'essayer plusieurs ensembles de valeurs initiales et de choisir celui pour lequel la log-vraisemblance totale de validation croisée est la plus élevée.

5.2.1 Objectif 2 : Déterminer quel outil statistique on doit utiliser pour trouver le meilleur modèle

Dû à la présence de plateau dans le modèle multinormal indépendant, on utilise le BIC pour obtenir un nombre de classe optimal à partir de ce critère. Peu importe le n utilisé (nombre d'individus, nombre de familles ou la moyenne des deux) on obtient sensiblement les mêmes conclusions. Ainsi, seuls les modèles pour lesquels on utilise le nombre de familles sont présentés. Les graphiques de la figure 5.4 présentent les BIC en fonction du nombre de classes des modèles, pour chaque ensemble de valeurs initiales utilisé.

Figure 5.4 – Valeurs des BIC⁴ pour le modèle multinormal indépendant



Encore une fois, les graphiques sont similaires peu importe l'ensemble de valeurs initiales utilisés et un minimum est atteint pour chacun des ensembles. Cependant, pour *permut2*, le minimum est atteint pour le modèle à 7 classes, contrairement aux autres ensembles où les minimums des BIC sont atteints en début de plateau, soit généralement pour les modèles à 4 classes.

⁴ Le tableau B.3 de l'annexe B présente les valeurs des BIC

5.2.2 Conclusion objectif 2

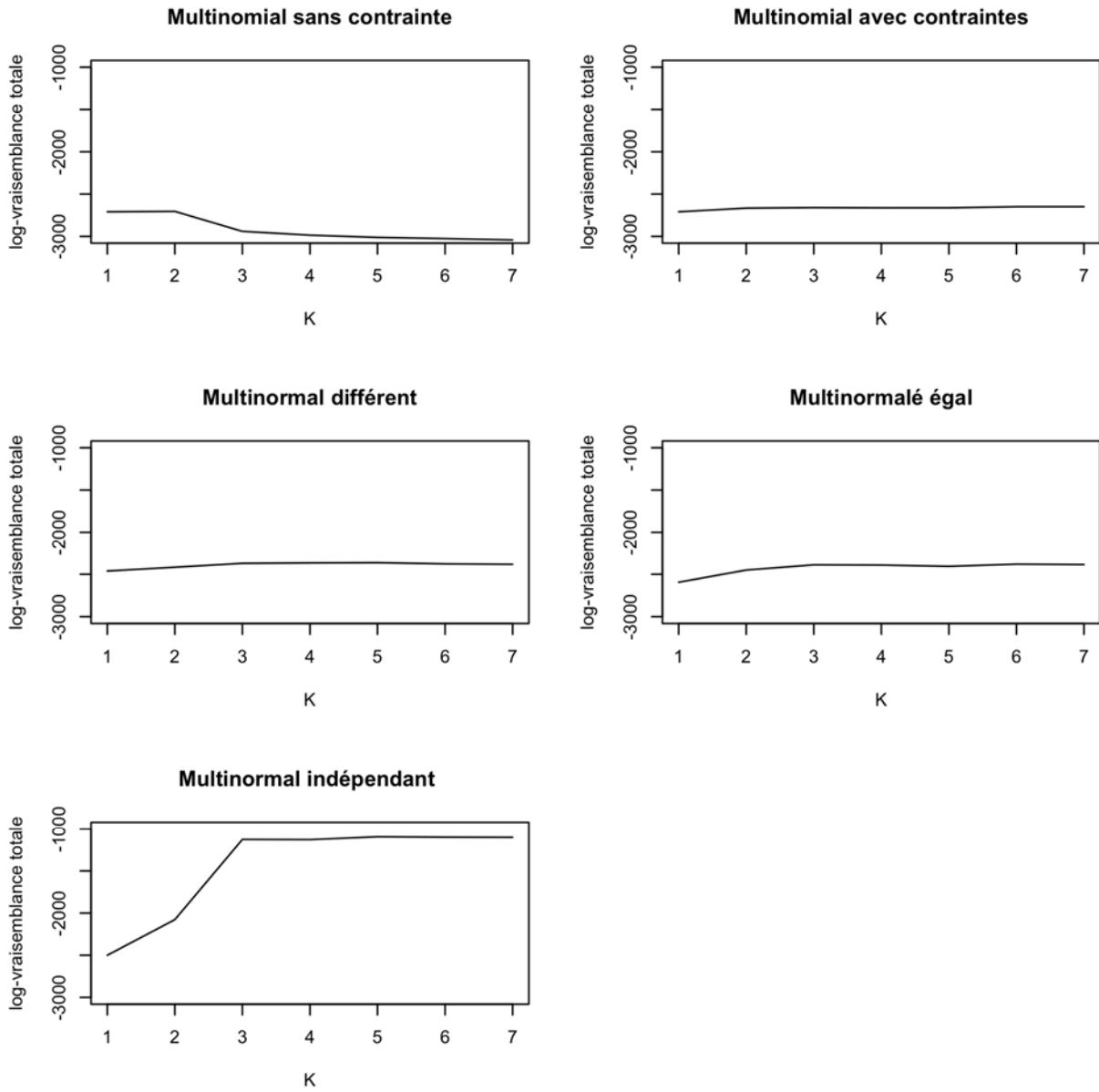
L'utilisation du BIC n'a pas permis de retirer complètement la présence de plateau lors du choix du nombre de classes. Le problème n'est donc pas totalement résolu. Cependant, les résultats obtenus nous permettent de conclure qu'en présence d'un plateau, il est préférable de prendre le modèle ayant le nombre de classes k du début de plateau, car ce sont généralement ceux qui ont la plus petite valeur du BIC. Ainsi, en ayant des modèles avec un plus petit nombre de classes, l'impact des valeurs initiales sur le choix du nombre de classes est faible et les taux de concordance sont meilleurs.

Ainsi, pour la sélection du modèle final, il est préférable d'utiliser la validation croisée et s'il y a présence de plateau, on prend le nombre de classes du début du plateau. Si le nombre de classes est relativement élevé, il est préférable d'utiliser divers ensembles de valeurs initiales pour choisir le modèle final.

5.3 Objectif 3 : Déterminer quels sont les meilleurs modèles parmi ceux proposés pour le jeu de données du CRULRG

Les graphiques de la figure 5.5 présentent les log-vraisemblances totales de validation croisée pour les 5 modèles étudiés. L'ensemble de valeurs initiales utilisé pour tous les modèles est *cluster_ind*. Pour les données continues, le modèle multinormal indépendant est celui pour lequel on a les plus grandes log-vraisemblances pour presque toutes les valeurs de k . Pour ce qui a trait aux données discrètes, il est plus difficile de déterminer quel modèle a la plus grande valeur de log-vraisemblance de validation croisée à partir des graphiques uniquement. En se référant au tableau B.4 de l'annexe B, on constate que le modèle multinomial avec contraintes a les plus grandes valeurs de log-vraisemblance totale de validation croisée.

Figure 5.5 - Log-vraisemblances totales de validation croisée⁵ pour les cinq distributions de mesures

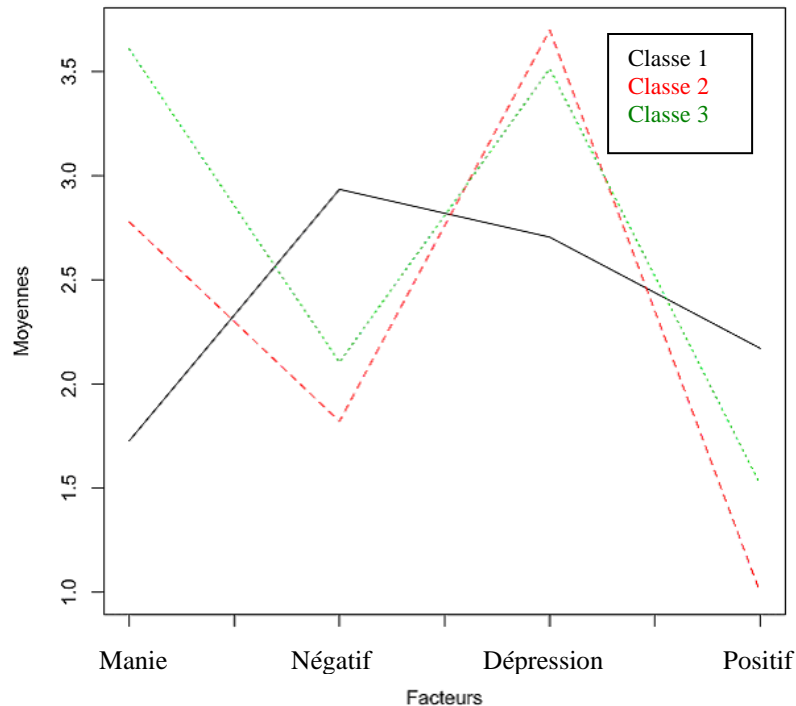


⁵ Le tableau B.4 de l'annexe B présente les valeurs des log-vraisemblances

5.3.1 Modèle final pour les données continues

Pour l'estimation des valeurs initiales du modèle, on choisit, parmi les 6 ensembles, l'ensemble *permut1* puisque c'est celui ayant la plus faible log-vraisemblance de validation croisée (voir le tableau B.4 de l'annexe B). En se référant à la figure 5.3, on constate que le plateau des log-vraisemblances débute au modèle à 3 classes. On ajuste donc ce modèle aux données. Les scores moyens des facteurs pour chaque classe sont présentés à la figure 5.6.

Figure 5.6 – Moyenne des 4 facteurs pour chaque classe du modèle multinormal indépendant



La classe 1 représente en grande partie les schizophrènes dû aux moyennes élevées des facteurs *néгатif* et *positif*. De plus, puisque la moyenne du facteur *dépression* est considérablement élevée, cette classe recoupe probablement certains individus ayant des troubles bipolaires.

La classe 2 représente essentiellement les individus ayant des troubles bipolaires dû aux moyennes élevées des facteurs *manie* et *dépression* qui caractérisent cette maladie.

La classe 3 représente elle aussi des individus ayant des troubles bipolaires mais qui ont une manie plus élevée que ceux de la classe 1. Ainsi, cette classe regroupe les individus ayant des troubles bipolaires qui ressemblent aux schizophrènes puisqu'une manie sévère est souvent accompagnée de délire et que ces deux symptômes caractérisent une grande partie des personnes atteintes de schizophrénie.

Le tableau 5.2 représente la répartition des individus atteints de schizophrénie et de troubles bipolaires parmi les nouvelles classes. La classe 1 contient essentiellement des schizophrènes et la classe 2 contient majoritairement des bipolaires. La classe 3, comme il a été mentionné précédemment, contient majoritairement des individus ayant des troubles bipolaires, mais aussi un bon nombre de schizophrènes.

Tableau 5.2 – Répartition des individus atteints de schizophrénie et de bipolarité en fonction des classes déterminées par le modèle multinomial sans contrainte

	Classes		
	1	2	3
Troubles bipolaires	5	134	57
Schizophrènes	70	27	33

Ainsi, on obtient une variante des définitions cliniques qui viennent nuancer les diagnostics tels qu'on les connaît.

5.3.2 Modèle final données discrètes

En comparant les graphiques de log-vraisemblances totales de validation croisée des divers ensembles de valeurs initiales pour le modèle multinomial avec contraintes (présentés à la figure C.1), on constate qu'un plateau débute à partir du modèle à 5 classes. On utilise donc le modèle à 5 classes et on prend les classes des individus qui ont été déterminées à partir de l'ensemble

permut3 puisque c'est celui ayant la plus grande valeur de log-vraisemblance parmi les modèles à 5 classes (voir le tableau C.1 à l'annexe C).

Tableau 5.3 – Répartition des individus atteints de schizophrénie et de bipolarité en fonction des classes déterminées par le modèle multinormal indépendant

	Classes				
	1	2	3	4	5
Troubles bipolaires	49	16	73	4	54
Schizophrènes	1	74	18	29	8

Le tableau 5.3 donne la répartition des individus atteints en fonction des 5 classes du modèle. Les classes 1, 3 et 5 regroupent essentiellement les bipolaires tandis que les classes 2 et 4 regroupent les schizophrènes. Contrairement au modèle précédent, il n'y a pas de classe qui recoupe les définitions des deux maladies telles qu'on les connaît puisque chaque classe contient majoritairement les individus d'une seule maladie. Ainsi, on obtient des sous-groupes formés essentiellement de schizophrènes ou de bipolaires, soit des sous-groupes similaires à ceux formés par les définitions cliniques.

6. Conclusion

Bien que la méthode de sélection du modèle final soulève encore quelques problèmes, celle-ci donne une très bonne idée du modèle s'ajustant le mieux à un ensemble de données et peut être utilisée pour déterminer le meilleur modèle.

Les nouvelles classifications des individus données par le modèle multinormal indépendant permettent de croire que le modèle développé par Labbe et *al.* sera d'une grande utilité dans la création de sous-groupes plus homogènes génétiquement. La création de ces sous-groupes facilitera ainsi la détection des gènes de maladie lors de l'analyse de liaison.

De plus, une extension du modèle qui tient compte de la différence entre les individus atteints sans symptôme mesuré et ceux non-atteints sera bientôt faite.

Bibliographie

CLOGG, C.C. (1995). Latent class models. In Arminger, G., Clogg, C. C. and Sobel, M. E. Editors, *Handbook of statistical modeling for the social and behavioral sciences*, New York : Penum Pess.

LABBE, A., Bureau A., and Merette C. (2007). Integration of genetic familial structure in latent class models.

MCLACHLAN, G.J. et KRISHNAN, T. (2007). *The EM Algorithm and Extensions*. New York: Wiley.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6(2), 461-464.

VAN DER LAAN, M., DUPONT, S. et KELES, S. (2004). Asymptotic optimality of likelihood-based cross validation. *Statistical Applications in Genetics and Molecular Biology* 3: Article 4.

Annexe A

Description des 11 dimensions et les symptômes s'y rattachant :

- 1. délires :* Fausses croyances personnelles basées sur des inférences incorrectes concernant la réalité extérieure
- référence
 - persécution
 - grandeur
 - somatique
 - culpabilité
 - jalousie
 - nihiliste (refus de l'idéal collectif)
 - religieux
 - contrôle
 - diffusion de la pensée
 - lecture de la pensée
 - insertion de la pensée
 - retrait de la pensée
 - systématisés
 - bizarres
- 2. hallucinations :* Perception sensorielle en l'absence de stimulation externe de l'organe sensoriel concerné
- auditives
 - voix qui commentent
 - voix qui conversent entre elles
 - visuelles
 - tactiles
 - gustatives
- 3. comportement bizarre :*
- habillement et apparence
 - social et sexuel
 - agressif ou agité
 - stéréotypé ou rituel
 - évaluation globale
- 4. anhédonie :* Difficulté à avoir des intérêts et du plaisir
- intérêt et activité
 - intérêt et activité sexuels
 - capacité d'intimité
 - relation avec les pairs/amis
 - évaluation globale
- 5. apathie :* Manque d'énergie, manque de volonté
- toilette et hygiène
 - manque de persistance au travail
 - anergie physique
 - évaluation globale

<i>6. catatonie :</i>	Négativisme, stéréotypes gestuels	
	- évaluation globale - stupeur - rigidité	- flexibilité cirreuse - excitation/agitation - position
<i>7. retrait affectif :</i>	- affect plat - affect émoussé, expression faciale figée - réduction des mouvements spontanés - manque d'expression gestuelle	- pauvre contact visuel - absence de réactivité affective - absence de modulation vocale - évaluation globale
<i>8. pensée désorganisée :</i>	- incohérence - relâchement marqué des associations - discours tangentiel - discours illogique	- discours circonstanciel - logorrhé (flux de parole) - associations par assonance
<i>9. alogie :</i>	Trouble du langage résultant de l'absence d'idées	
	- pauvreté du discours - blocage	- latence de réponse augmentée - persévération (répétition de gestes ou de mots)
<i>10. manie :</i>	- estime de soi exagérée - diminution du besoin de sommeil - parle plus que d'habitude - fuite des idées	- distractibilité - augmentation de l'activité - implication excessive dans des activités à risque - variation de l'humeur
<i>11. dépression :</i>	- humeur dépressive - diminution marquée de l'intérêt ou du plaisir - changement de poids - troubles du sommeil presque tous les jours - agitation ou ralentissement psychomoteur	- fatigue ou perte d'énergie presque tous les jours - culpabilité excessive ou sans fondement - diminution de la capacité de réfléchir ou de se concentrer - idées suicidaires
<i>Autres symptômes :</i>	Symptômes ne faisant pas partie d'un groupe	

- affect grossièrement inapproprié
- labilité émotionnelle (humeur changeante)
- discours distractible
- pauvreté du contenu du discours
- attention/inattention sociale

Annexe B

Tableau B.1 - Valeurs⁶ des log-vraisemblances totales de validation croisée pour le modèle multinomial sans contrainte

Ensembles de valeurs initiales	Nombre de classes						
	1	2	3	4	5	6	7
<i>Permut</i>	-2709.3	-2687.9	-2927.9	-2937.2	-2999.6	-2974.1	-3234.9
<i>Permut1</i>	-2709.3	-2686.0	-2916.3	-2941.6	-2960.8	-3023.9	-3241.0
<i>Permut2</i>	-2709.3	-2715.8	-2717.3	-2934.6	-2962.1	-2966.9	-2986.6
<i>Permut3</i>	-2709.3	-2705.1	-2937.6	-2952.9	-2975.5	-2976.5	-3479.8
<i>Cluster_ind</i>	-2709.3	-2704.4	-2940.2	-2985.6	-3011.8	-3025.8	-3041.9
<i>Cluster_fam</i>	-2709.3	-2722.7	-2940.2	-2969.9	-2952.1	-2981.3	-2972.2

Tableau B.2 – Valeurs⁶ des log-vraisemblances totales de validation croisée pour le modèle multinomial indépendant

Ensembles de valeurs initiales	Nombre de classes						
	1	2	3	4	5	6	7
<i>Permut</i>	-2500.2	-1703.4	-1122.4	-1112.7	-1098.5	-1098.5	-1104.2
<i>Permut1</i>	-2500.2	-1703.4	-1118.6	-1119.2	-1122.4	-1098.8	-1087.3
<i>Permut2</i>	-2500.2	-1703.4	-1122.4	-1103.4	-1111.6	-1096.9	-1095.7
<i>Permut3</i>	-2500.2	-1703.4	-1122.4	-1120.4	-1120.7	-1097.2	-1098.3
<i>Cluster_ind</i>	-2500.2	-2077.8	-1126.3	-1128.5	-1094.1	-1098.9	-1100.6
<i>Cluster_fam</i>	-2500.2	-1703.4	-1123.6	-1101.1	-1120.4	-1087.8	-1093.2

⁶ Les nombres en gras représentent les maximums de log-vraisemblances pour chaque ensemble de valeurs initiales

Tableau B.3 – Valeur⁷ des BIC pour le modèle multinormal indépendant

Ensembles de valeurs initiales	Nombre de classes						
	1	2	3	4	5	6	7
<i>Permut</i>	4996.0	3949.3	3817.7	3779.5	3858.2	3847.2	3925.3
<i>Permut1</i>	4996.0	3949.3	3820.4	3779.5	3821.1	3875.3	3787.3
<i>Permut2</i>	4996.0	3949.3	3817.7	3933.0	3858.2	3866.1	3630.3
<i>Permut3</i>	4996.0	3949.3	3817.7	3779.5	3870.2	3840.7	3923.0
<i>Cluster_ind</i>	4996.0	4727.4	3820.3	3821.8	3806.3	3858.7	3912.3
<i>Cluster_fam</i>	4996.0	3949.3	3820.3	3779.5	3814.5	3847.2	3919.6

Tableau B.4 – Valeurs⁶ des log-vraisemblances totales de validation croisée pour les cinq modèles de distributions des mesures

Modèles	Nombre de classes						
	1	2	3	4	5	6	7
Multinomiale <i>Sans contrainte</i>	-2709.3	-2704.4	-2940.2	-2985.6	-3011.8	-3025.8	-3041.9
Multinomiale <i>Avec contrainte</i>	-2709.3	-2665.2	-2659.0	-2661.7	-2661.4	-2649.0	-2648.7
Multinormal <i>différent</i> ⁸	-2461.3	-2416.5	-2370.9	-2366.1	-2363.4	-2377.5	-2382.3
Multinormal <i>égal</i> ⁹	-2594.2	-2450.0	-2388.6	-2391.3	-2406.0	-2380.6	-2385.1
Multinormal <i>Indépendant</i>	-2500.2	-2077.8	-1126.3	-1128.5	-1094.1	-1098.9	-1100.6

⁷ Les nombres en gras représentent les minimums de BIC pour chaque ensemble de valeurs initiales

⁸ Les matrices de variances-covariances sont différentes d'une classe à l'autre

⁹ Les matrices de variances-covariances sont égales d'une classe à l'autre

Annexe C

Figure C.1 – Log-vraisemblances totales de validation croisée pour le modèle multinomial avec contraintes

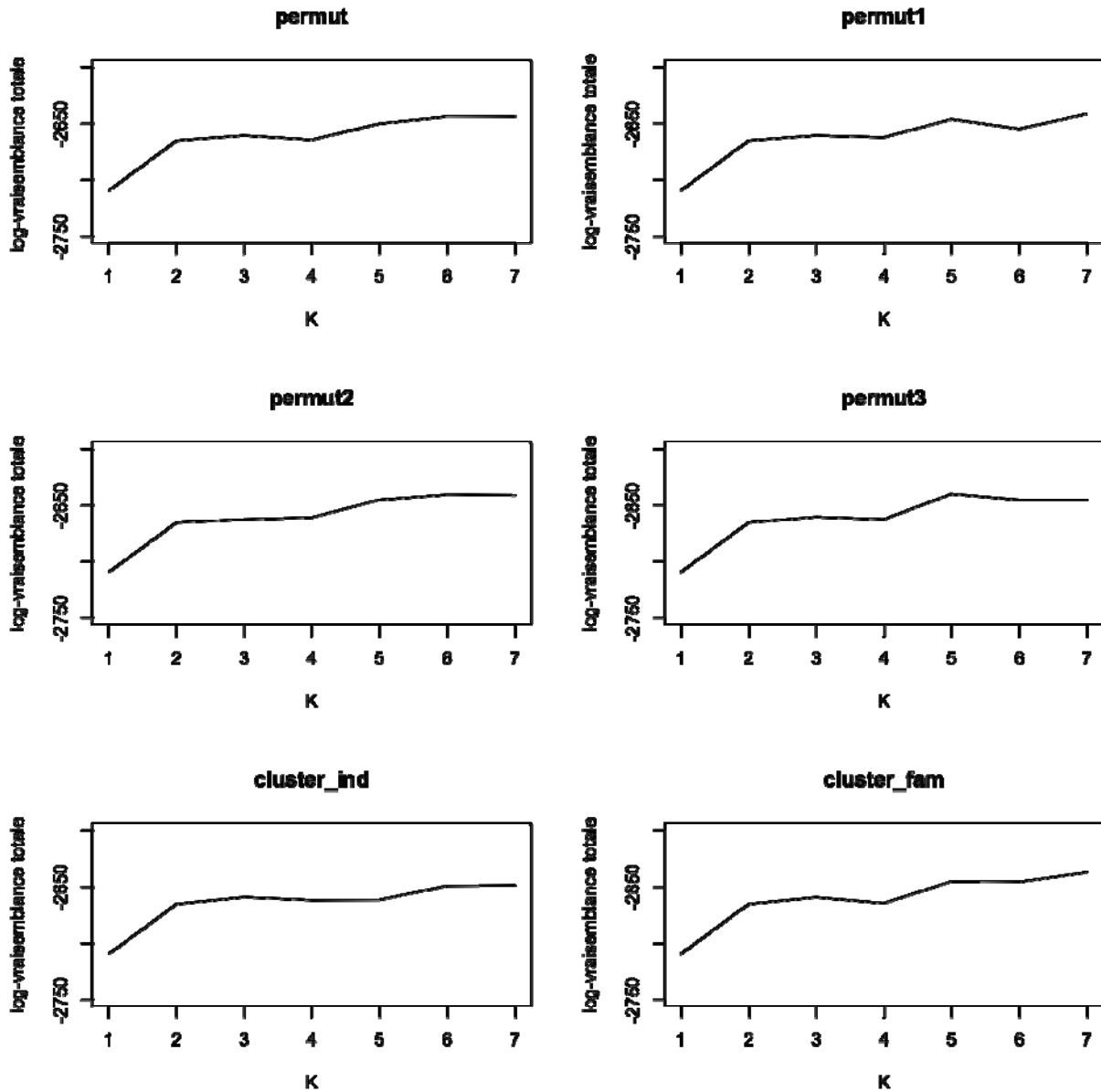


Tableau C.1 – Valeurs des log-vraisemblances totales de validation croisée pour le modèle multinomial avec contraintes

Ensembles de valeurs initiales	Nombre de classes						
	1	2	3	4	5	6	7
<i>Permut</i>	-2709.3	-2665.1	-2660.6	-2664.5	-2650.1	-2643.8	-2644.1
<i>Permut1</i>	-2709.3	-2665.2	-2660.6	-2662.3	-2646.1	-2654.7	-2641.5
<i>Permut2</i>	-2709.3	-2665.2	-2662.4	-2660.7	-2644.9	-2640.4	-2640.8
<i>Permut3</i>	-2709.3	-2665.1	-2660.3	-2662.4	-2639.7	-2645.0	-2645.0
<i>Cluster_ind</i>	-2709.3	-2665.2	-2659.0	-2661.7	-2661.4	-2649.0	-2648.7
<i>Cluster_fam</i>	-2709.3	-2665.1	-2659.1	-2664.5	-2645.1	-2645.6	-2637.2