Modèles statistiques appliqués à l'épidémiologie neuro-comportementale

M. Grzebyk, D. Chouanière, P. Wild, D. Commenges, C. Fabrigoule, H. Jacquim

JUIN 2000

N° ISSN 0397 - 4529

191

Modèles statistiques appliqués à L'épidémiologie neuro-comportementale

Michel Grzebyk-Dominique Chouanière-
Pascal Wild

INSTITUT NATIONAL DE RECHERCHE ET DE SECURITE

SIEGE SOCIAL :
30, RUE OLIVIER-NOYER, 75680 PARIS CEDEX 14

CENTRE DE RECHERCHE :
AVENUE DE BOURGOGNE, 54501 VANDOEUVRE CEDEX

Modèles statistiques appliqués à l'épidémiologie neuro-comportementale

auteurs :

- Michel Grzebyk
- Dominique Chouanière
- Pascal Wild

avec la collaboration à l'U 333 Inserm de

- Daniel Commenges
- Colette Fabrigoule
- Hélène Jacqmin

Exemplaire à demander à

M. Grzebyk, Département Métrologie des Polluants, INRS,
Avenue de Bourgogne - BP 27, F-54501 Vandœuvre Cedex,
tel +33 (0)3 83 50 87 96 - E.mail : grzebyk@inrs.fr

Résumé

L'épidémiologie neuro-comportementale a recours à des tests psychométriques qui génèrent des données particulières comprenant des multi-liaisons entre variables. Ce document fait la synthèse d'une étude qui visait à mettre au point une méthode d'analyse statistique adaptée aux problèmes multi-liaisons. Après une revue de la littérature des méthodes statistiques disponibles pour traiter cette configuration des données, nous avons finalement retenu le modèle linéaire à variables latentes. Il a d'abord été testé et validé sur des données simulées puis appliqué à deux jeux de données réelles, l'un étant issu d'une étude transversale sur la neurotoxicité du toluène réalisée par l'INRS et l'autre étant issu de la cohorte PAQUID de l'Unité 330 de l'INSERM. PAQUID est destinée à estimer l'incidence de la démence chez les personnes âgées et d'en étudier les facteurs de risque ; les données provenant de tests psychométriques, présentent également un modèle de multi-liaisons. Ces différentes mises en SSuvre du modèle à variables latentes ont permis de cerner les conditions de son application et d'en apprécier sa pertinence et ses limites.

Mots-clé : Variables latentes, dépendance conditionnelle, graphe, MCMC, PAQUID, toluène, neurotoxicité, test psychométriques.

Table des matières

Chapitre 1

Introduction

L'épidémiologie neuro-comportementale (ENC) en milieu professionnel a été développée au service Epidémiologie en Entreprises (EE) de l'INRS depuis 1991. L'ENC a pour objet d'étudier les effets sur le système nerveux central (SNC) des produits chimiques auxquels sont exposés les salariés au cours de leur exercice professionnel. Les principaux produits neurotoxiques (NT) reconnus sont les métaux (plomb, mercure, manganèse, aluminium, etc.) les solvants purs (toluène, styrène, xylène, trichloréthylène, chlorure de méthylène, etc.) ou mélangés (white-spirit, essence, etc.), les pesticides, les gaz anesthésiques, etc.

Pour mesurer les effets de ces produits sur le SNC, l'ENC a recours, parmi des explorations paracliniques disponibles, aux tests psycho-comportementaux qui occupent une place privilégiée dans l'évaluation des performances cognitives. Utilisés depuis longtemps pour l'exploration neuropsychologique de patients atteints de lésions cérébrales (traumatisme crânien, tumeur ou accident vasculaire, etc.) ces tests ont été sélectionnés au fil du temps par les neurotoxicologues pour leur capacité à détecter précocement des altérations des fonctions cognitives explorées (bonne sensibilité) et pour leur reproductibilité dans les études d'ENC. Depuis une vingtaine d'années ils ont été regroupés en batterie et informatisés. Parmi la dizaine de batteries NT disponibles, la plus flexible mais aussi la plus répandue est la batterie « Neurobehavioral Evaluation System » (NES) que nous avons retenue au service EE.

Le logiciel NES propose, en sortie, un fichier des résultats des tests, bruts ou synthétisés sous forme de variables. Lors de l'analyse statistique de nos premières études, nous avons été confrontés à une structure de multidépendance des variables des tests, qui nous ont incité à recourir à des modèles statistiques tels que les analyses en composantes principales (ACP), préconisées dans la littérature NT récente. Mais les limites de ces premiers outils sont vite apparues et nous ont contraints à nous tourner vers des modèles jusqu'ici peu utilisés dans le champ de l'épidémiologie professionnelle ; ces modèles mis au point pour d'autres champs disciplinaires, sont, grâce à la puissance de calcul des systèmes informatiques actuels, en plein développement.

Nous allons dans ce premier chapitre, à travers l'exemple d'une étude d'ENC réalisée, expliciter la nature exacte du problème posé par les données et le cheminement de notre réflexion pour le résoudre, puis préciser les collaborations qui ont été établies dans le cadre de ce travail ainsi que ses objectifs.

1.1 Position du problème

L'une des premières études d'ENC avait pour objectif la recherche de la relation entre une exposition professionnelle prolongée au toluène et des altérations des fonctions mentales (FM). Pour les explorer, nous avons proposé à 128 salariés de 2 imprimeries, entre autres explorations, 6 tests de la batterie NES.

L'observation des données recueillies, a montré que les variables générées par le programme

du NES (en général la moyenne et son écart-type) étaient, pour un même test, très corrélées entre elles. Nous avons, dans un premier temps, élaboré, à partir des résultats bruts des tests, de nouvelles variables sous conditions de faible corrélation entre 2 variables issues d'un même test, de normalité de leur distribution et de lisibilité de la variable quant à la FM explorée. Pour chaque test, quand les données brutes le permettaient, 2 types de variables ont été construites :

- une variable de performance au test
- une variable de stabilité de la performance, censée se dégrader plus précocement que la performance elle même dans un processus d'atteinte neurotoxique.

L'analyse statistique classique (régression linéaire univariée) a mis en évidence une relation entre l'exposition au toluène mesurée au poste de travail et une diminution des variables de performance des tests explorant les FM : mémoire et apprentissage. Néanmoins, seules 2 variables, issues d'un test de mémoire à court terme, étaient liées de façon statistiquement significative au niveau d'exposition actuel. Par contre, nous n'avons pas pu mettre en évidence de relations entre les variables des tests et l'indicateur d'exposition qui intégrait la durée et les niveaux passés d'exposition (qui étaient bien supérieurs aux niveaux actuels). Ces résultats suggéraient un processus d'atteinte neurotoxique réversible, différenciée selon les FM qui semblait débuter plus spécifiquement par une détérioration de 2 fonctions : l'apprentissage et la mémoire.

Pour interpréter ces résultats nous avons alors reconsidéré les concepts théoriques qui sous-tendent la construction des tests utilisés dans cette étude. Bien que chaque test soit conçu pour explorer spécifiquement une FM particulière, chaque variable rend compte, en fait, du fonctionnement de plusieurs FM mises en jeu lors de la réalisation du test. Par exemple, pour un test censé explorer la mémoire à court terme, la variable de performance dépendra essentiellement de la capacité mnésique du sujet mais aussi de la concentration avec laquelle le salarié a réalisé cette tâche. A une variable correspond plusieurs FM et inversement une FM est décrite par plusieurs variables. Les variables des tests sont des mesures partielles et indirectes d'une FM qui n'est pas mesurable en tant que telle. Le problème consistait donc à modéliser une variable latente de la FM par une réduction de la dimensionalité de l'information contenue dans les variables des tests. Compte tenu des résultats de l'analyse univariée d'une part et du contexte théorique des tests d'autre part, nous avons décidé d'affiner l'analyse en étudiant l'effet de l'exposition sur les variables latentes des FM.

Quelques publications en ENC proposaient l'analyse en composantes principales (ACP) ou l'analyse factorielle (FA) [34, 21]. Ces techniques modélisent les dépendances et synthétisent l'information contenue dans un grand nombre de variables à partir de la matrice de corrélation des variables initiales. Les "composantes principales" ou axes sont des combinaisons linéaires des variables initiales. Ces axes sont de nouvelles variables censées explorer les FM. Dans [9], les résultats des 2 ACP, menées sur les 2 groupes correspondant aux 2 entreprises de l'étude, n'ont pas permis de retrouver des résultats stables. En effet, les ACP sont très dépendantes des jeux de données sur lesquels elles sont réalisées et la combinaison linéaire des variables sur un axe peut être due à des phénomènes parasites tels que l'effet de l'âge, du niveau socioculturel, etc... De plus, elles mettent sur le même plan, pour un axe donné, toutes les variables quel que soit leur poids dans la détermination de l'axe. Dans [21], l'effet des variables confondantes telles que l'age, le sexe, ... est soustrait des variables des tests avant d'effectuer l'analyse factorielle. Cette approche permet de tenir compte des variables confondantes, mais est critiquable dans la mesure où toutes les variables susceptibles d'influencer les fonctions mentales ne sont pas intégrées au même moment dans l'analyse, alors qu'elles ont un rôle équivalent. Que ce soient l'exposition au solvant ou les autres variables potentiellement confondantes, elles conditionnent directement les FM et non les résultats aux tests. Par ailleurs, une autre faiblesse de ces techniques réside dans le fait que les axes sont interprétés a posteriori. Ceci n'est pas toujours possible et limite l'utilisation des connaissances établies sur les relations entre les FM d'une part et les variables issues des tests d'autre part.

Or, l'état de l'art en neuropsychologie permet de qualifier les liens traduisant l'influence ou non des FM sur les variables. Les connaissances établies permettent de préciser également les dépendances existantes entre les FM elles-mêmes. L'influence des facteurs individuels tels que l'âge, le niveau socioculturel sur les FM a été étudiée. Ces connaissances permettent de reconstituer la structure causale des données qui peut être schématisée sur un graphe de dépendance construit *a priori* dont les flèches traduisent les dépendances entre :

- les variables des tests Y et les variables latentes des FM F , $(F \rightarrow Y)$,
- les variables explicatives (individuelles ou d'exposition professionnelle) X et les variables latentes des FM, $(X \leftarrow F)$,
- les variables latentes entre elles $(F - F)$.

Interrogeant alors le corpus des connaissances en méthodologie statistique, développée dans d'autres contextes : psychométrie, sociométrie, économétrie, etc., nous y avons repéré des modèles capables de tester l'hypothèse d'une structure causale des données représentée par un graphe de dépendance tel que défini plus haut.

1.2 Collaboration

La configuration de multidépendance de données n'est pas l'apanage de l'ENC professionnelle, d'autres thématiques en épidémiologie s'y trouvent confrontées. C'est le cas, en particulier, de l'épidémiologie neurologique ou psychiatrique qui a également recours à des tests psychocomportementaux. Ainsi notre réflexion méthodologique a rencontré l'intérêt de l'équipe de recherche de l'unité 330 de l'INSERM qui a en charge l'exploitation statistique de la cohorte PAQUID constituée pour estimer l'incidence de la démence et en étudier ses facteurs de risque. Cette convergence de problématique a permis une collaboration qui s'est traduit par la mise à notre disposition d'une partie des données de PAQUID accompagnée d'un graphe résumant les dépendances (connues ou à établir) entre les données puis par une expertise commune des premiers résultats.

1.3 Objectifs de l'étude

Le présent rapport relate le développement et les résultats de cette étude méthodologique dont l'objectif général est la mise au point de méthodes d'analyse statistique adaptées au contexte de l'ENC professionnelle et de la démence. L'objectif général s'est ainsi décliné en objectifs secondaires :

- élaborer, pour chacun des jeux de données (INRS - INSERM) « un graphe de dépendance conditionnelle » décrivant les multi-liaisons entre les variables, d'après le corpus des connaissances actuelles en psychologie cognitive et en neuropsychologie.
- répertorier, dans la littérature méthodologique statistique, les outils (modèles statistiques et logiciels) utilisés dans d'autres contextes (psychométrie, sociométrie, économétrie), capables de prendre en compte des données avec des multi-liaisons.
- évaluer la pertinence en regard du problème posé, les conditions d'application, la robustesse, les avantages et les inconvénients de chaque modèle statistique.
- valider le modèle statistique retenu sur des données simulées.
- appliquer le modèle statistique ainsi validé aux deux jeux de données disponibles.

1.4 Plan du rapport

Le chapitre 2 expose les résultats de la recherche bibliographique menée pour mettre au point la méthodologie statistique. Le chapitre 3 donne des précisions sur l'intégration des connaissances neuropsychologiques dans un modèle statistique. Le chapitre 4 détaille des résultats sur un point crucial de ce type de modèles statistique, son identifiabilité. Les chapitres 5, 6 et 7 présentent des résultats d'application du modèle dans 3 situations :

- le modèle a été appliqué à des jeux de données simulées afin de vérifier sa pertinence (chapitre 5),
- le chapitre 6 présente les résultats sur les données TOLUÈNE,
- le chapitre 7 présente les résultats sur les données PAQUID.

Chapitre 2

Revue des méthodes statistiques

2.1 Introduction

Ce chapitre présente une revue bibliographique des modèles et méthodes statistiques utilisables pour aborder ce genre de problème. Il s'agit de modèles multivariés avec ou sans variables latentes. Les deux grands domaines explorés sont d'une part les modèles issus des sciences sociales et humaines tels que les *modèles à équations structurelle* (section 2.2) et d'autre part les *modèles graphiques* (section 2.4) utilisés aussi en sciences sociales et humaines (mais pas par les mêmes équipes !) et dans le domaine de l'intelligence artificielle (en particulier les *réseaux de connaissance bayésiens* [8][1])

2.2 Modèles issus des sciences sociales et humaines

Une revue complète des modèles utilisés dans ces domaines se trouve dans [2] et [6]. La plupart des modèles présentés dans cette section sont issus de ces deux livres.

2.2.1 Les modèles factoriels (*factor analysis*)

S'agissant d'un problème multivarié avec réduction de dimensionnalité, les modèles les plus simples sont les modèles factoriels du type *factor analysis* (FA). Ces modèles sont dédiés à l'analyse de données multivariées continues dans laquelle on souhaite résumer l'information recueillie dans un grand nombre de variable en un nombre limité de facteurs, en se plaçant dans un cadre multigaussien. Toutes les variables mesurées ont le même statut (il n'y a ni variable explicative ni variable à expliquer) et toutes les variables (mesurées et latentes) sont continues. Le vecteur des variables mesurées Z se décompose en une somme de 2 termes en fonction du vecteur des facteurs F et du vecteur des résidus E :

$$Z = \Lambda F + E \tag{2.1}$$

Λ est une matrice des *factor loadings*. Le nombre de facteurs est inférieur au nombre de variables mesurées et les hypothèses de distribution de cette décomposition sont les suivantes :

$$p((F,E)|\Sigma_F,\Sigma_E) = \mathcal{N}(0, \left(\begin{smallmatrix}\Sigma_F & 0\\ 0 & \Sigma_E\end{smallmatrix}\right)) \tag{2.2}$$

Il s'agit d'un modèle de structure de covariance. Dans le modèle le plus «fruste», les matrices Σ_F et Σ_E sont supposées diagonales.

1. traduction de *bayesian belief networks*

Parfois, on peut imposer des hypothèses structurelles (c.-à-d. des 0 dans Σ_F ou dans Λ) ; il s'agit alors d'une analyse factorielle confirmatoire (*confirmatory factor analysis*).

Ces modèles peuvent être ajustés aux données par maximum de vraisemblance ou par moindres carrés, mais tel quel, ces équations sont indéterminées. Si aucune condition de nullité n'est imposée aux coefficients des matrices Σ_F et Σ_E, la solution n'est pas unique ; les solutions sont définies à une rotation (orthogonale ou non) près. Si des contraintes de nullité sont imposées, il faut veiller à ce que l'ensemble conserve sa cohérence algébrique.

2.2.2 Les modèles dits « à équations structurelles » (Structural equation model)

Les modèles à équations structurelles ont été développés comme des généralisations de FA et sont parfois présentés comme des méthodes pour effectuer une inférence causale. Dans ces modèles, les variables mesurées sont séparées en 2 (ou plus) sous-ensembles, les variables explicatives X et les variables à expliquer Y. À chaque sous-ensemble de variables est associé un ensemble de variables latentes, F_x et F_y, comme pour 2 FA. Les 2 ensembles de FA sont reliées par une combinaison linéaire. On se place toujours dans un cadre multigaussien. Les modèles à équations structurelles sont définis par les équations suivantes :

$$Y = \mu_y + \Lambda_y F_y + e_y \tag{2.3}$$
$$X = \mu_x + \Lambda_x F_x + e_x \tag{2.4}$$
$$F_y = BF_y + \Lambda F_x + e \tag{2.5}$$

avec $\mathbf{E}(e_y) = \mathbf{E}(e_x) = \mathbf{E}(e) = 0$, $Cov(F_x, e) = 0$, F_x, e, e_x et e_y sont mutuellement orthogonaux, tandis que

$$Cov(e_x, e_x) = \Theta_x \tag{2.6}$$
$$Cov(e_y, e_y) = \Theta_y \tag{2.7}$$
$$Cov(e, e) = \Phi \tag{2.8}$$
$$Cov(F_x, F_x) = \Phi_x \tag{2.9}$$

La matrice B a une diagonale nulle. La causalité du modèle s'inscrit dans la structure des matrices B, Λ, Λ_x et Λ_y par l'intermédiaire de contraintes sur certains coefficients (soit la nullité, soit une contrainte de signe). Ces modèles ont adopté un code de représentation des relations entre les variables (*i.e.* de la structure causale) sous la forme de graphe. Les variables sont représentées par des nœuds et les coefficients non nuls des matrices B, Λ, Λ_x et Λ_y sont représentés par des flèches rectilignes et les coefficients non nuls des matrices de covariance Θ_x, Θ_y, Φ et Φ_x par des flèches curvilignes.

Techniquement, le modèle est ajusté aux données par maximum de vraisemblance ou par moindres carrés. Plus précisément, on utilise la matrice de covariance des variables mesurées exprimée en fonction des matrices B, Λ, Λ_x, Λ_y, Θ_x, Θ_y, Φ et Φ_x :

$$\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \tag{2.10}$$

avec

$$\Sigma_{yy} = \Lambda_y (I - B)^{-1} (\Lambda \Phi_x \Lambda' + \Phi)(I - B')^{-1} \Lambda_y' + \Theta_y \tag{2.11}$$
$$\Sigma_{yx} = \Lambda_y (I - B)^{-1} \Lambda \Phi_x \Lambda_x' \tag{2.12}$$
$$\Sigma_{xx} = \Lambda_x \Phi_x \Lambda_x' + \Theta_x \tag{2.13}$$

Il s'agit donc d'un modèle de décomposition de la matrice de covariance des données. Récemment, des approches bayésiennes sont apparues (voir [39, 1, 37] par exemple)

Comme pour le FA, se pose le problème de l'indétermination et la cohérence algébrique du modèle. Les modèles à équations structurelles se divisent en plusieurs sous-modèles tels que LISREL, EQS, RAM, MECOSA, ... Ils correspondent à un paramétrage légèrement différent de la structure donnée par les équations (2.3), (2.4) et (2.5). Ces sous-modèles ont généralement donné naissance à un logiciel portant le même nom (cf [24], [3], [38], ...). Certains d'entre eux sont programmés dans la procédure CALIS de SAS [23].

2.2.3 Les modèles conditionnels à équations structurelles

Contrairement à ci-dessus, les variables explicatives X sont utilisées pour modéliser la structure moyenne et sont traitées comme des effets fixes. Le modèle est qualifié de conditionnel dans la littérature car il conduit aux équations :

$$\mathbf{E}(Y|X) = \mu_y + \Pi X \tag{2.14}$$
$$Var(Y|X) = \Sigma_y^x \tag{2.15}$$

Un exemple typique de modèle conditionnel se traduit par le modèle structurel suivant :

$$F_y = BF_y + \Lambda X + e \tag{2.16}$$
$$Y = \mu_y + \Lambda_y F_y + e_y \tag{2.17}$$

avec, une fois de plus, les éléments de la diagonale de B nuls, avec $\mathbf{E}(e_y) = \mathbf{E}(e) = 0$, e, et e_y mutuellement orthogonaux, tandis que

$$Cov(e_y, e_y) = \Theta_y$$
$$Cov(e, e) = \Phi$$

Alors, les structures de moyenne et de covariance des variables Y données par 2.16 s'écrivent :

$$\Pi = \Lambda_y (I - B)^{-1} \Lambda \tag{2.21}$$
$$\Sigma_y^x = \Lambda_y (I - B)^{-1} \Phi (I - B')^{-1} \Lambda_y' + \Theta_y \tag{2.22}$$

L'estimation des paramètres peut, là encore, être obtenu par maximum de vraisemblance. D'après [2], certains logiciels peuvent résoudre ces problèmes (LISCOMP, LISREL, MECOSA, et d'autre) mais ils sont absents d'autres logiciels (EQS en particulier).

Remarque

Il semble que, dans ce domaine, la formulation des modèles associe rarement identification des effets moyens (modélisation de l'espérance mathématique) et identification des (cor)relations entre les variables (modélisation de la structure de covariance). Cela s'accentue quand on fait intervenir des variables latentes. Plus précisément, soit on modélise la moyenne (par des procédures type modèle linéaire, modèle linéaire généralisé, avec ou sans effets mixtes), soit on ne s'intéresse qu'à la structure de covariance, quitte à centrer les variables au préalable. Le *modèle conditionnel à équations structurelles* est la seule exception trouvée dans la littérature ([2], page 221), mais ce modèle est peu décrit et les logiciels utilisés dans ce domaine ne le mentionne pas.

2.2.4 Les modèles à équations structurelles pour des variables ordinales

Comme dans la pratique on rencontre souvent des variables ordinales (dans les tests, questionnaires d'enquêtes, ...) cette question est largement abordée dans la littérature ([33, 28, 26, 5, 2, 6]). La démarche la plus couramment présentée est le *probit multivarié*. À chaque variable ordinale (à expliquer) y ayant n valeurs possibles est associée une variable non-observée (latente) continue gaussienne $\tilde{y}$ ainsi qu'un ensemble de seuils $(\alpha_0 = -\infty, \alpha_i, \alpha_n = +\infty, i = 1, ..., n-1)$ de telle sorte que

$$y = i \Longleftrightarrow \alpha_{i-1} \leq \tilde{Y} < \alpha_i. \tag{2.23}$$

Ce sont alors les variables continues qui sont reliées aux variables latentes.[2]

2.3 Les modèles linéaires inférentiels

2.3.1 Modèles linéaires sans variables latentes

Si on ne tient pas compte des fonctions mentales, on obtient une régression multivariable dans lequel les erreurs sont corrélées, la structure de corrélation étant imputable à (ou contrôlée par, selon les points de vue) l'existence de fonctions mentales. Différentes techniques « classiques » existent pour aborder de tels problèmes :

- les modèles à effets aléatoires (ou mixtes) [19, 2],
- les *Generalized Estimating Equations* ou GEE [19],

Les deux techniques modélisent la moyenne (effets fixes) et la corrélation (effets aléatoires). Mais elles ne sont pas réellement adaptées dans la mesure où la modélisation des effets fixes et des effets aléatoires est disconnectée alors que dans notre problème, il existe une relation algébrique entre la matrice des effets fixes et la structure de corrélation. Ces modèles font souvent appel de façon explicite ou implice au conditionnement des variables. Ils n'introduisent pas la notion de variable latente bien que les effets aléatoires puissent être considérés comme des variables latentes.

2.3.2 Modèles linéaires avec variables latentes

En dehors des approches à équations structurelles, l'introduction des variables latentes dans les modèles inférentiels paraît récente : elle est sans doute liée aux progrès techniques en matière de calculs par ordinateurs. En effet, la façon d'introduire les variables latentes dans ces modèles consiste à simuler les variables latentes pour les utiliser, à leur niveau, comme des variables explicatives. Deux grandes techniques permettent cette stratégie :

- les techniques *bayésiennes* en général, favorisées par l'essor des générateurs pseudo-aléatoires à base de chaînes de Markov et du principe de Monté Carlo (MCMC). L'introduction des variables latentes est assez triviale, la principale difficulté résidant dans le temps de calcul que cela entraîne puisque pour chaque variable latente, on définit une chaîne par unité expérimentale. Comme ces techniques ont été utilisées dans cette étude, des détails supplémentaires sont donnés plus loin. Des exemples peuvent être consultés dans [41] et [42]. L'analyse des jeux de données par l'approche bayésienne a été effectuée à l'aide du logiciel BUGS ([40]).

2. On peut noter que, dans le domaine des modèles à équations structurelles, seul le cas de la variable non mesurée gaussienne est envisagé alors que d'autres cas sont possibles : la distribution logistique et la distribution des valeurs extrêmes. Ces modèles sont connus sous le nom de *proportionnal odds* et *proportionnal hazard* [32].

- les techniques de maximum de vraisemblance pour les données manquantes, dont le développement s'articule autour de l'*algorithme EM* [44]. Cet algorithme, itératif comme la plupart des techniques de maximum de vraisemblance, utilise plusieurs phases de calculs intensifs à chaque itération.

À l'origine, les techniques de maximum de vraisemblance ont été développées pour permettre l'utilisation d'information contenant des données manquantes. Pour généralisre cette approche aux variables latentes, on considère une variable latente comme une variable pour laquelle toutes les données sont manquantes. C'est ce que font Sammel, Ryan et Legler dans une série d'articles ([29], [36] et [35]).

2.4 Les modèles graphiques

Trois livres ont servi à écrire cette section : [11], [16] et [48].

2.4.1 Présentation des modèles graphiques « classiques »

L'objet des modèles graphiques est l'analyse d'un ensemble de variables qui, *a priori*, sont toutes considérées comme aléatoires. L'analyse consiste à établir la structure des dépendances entre les variables. Mais contrairement à tous les modèles exposés ci-dessus, le critère permettant d'établir ou de ne pas établir un lien entre 2 variables est défini en terme de dépendance ou d'indépendance de la loi de probabilité bivariée de ces 2 variables. La structure est représentée symboliquement par un graphe dans lequel les nœuds sont les variables. Les variables sont reliées par des traits sauf dans certains cas : deux variables sont disconnectées si et seulement si elles sont indépendantes. Ainsi, contrairement aux modèles précédents, la structure se construit en considérant d'abord l'absence de dépendance plutôt que l'existence d'une dépendance.

On distingue 2 types de problèmes dans les modèles graphiques :

- l'*apprentissage de structure* (*structural learning*) correspond à l'établissement d'un graphe de dépendance (le meilleur) construit à l'aide d'un jeu de données ; on ne suppose aucune relation *a priori*, l'objectif est de réduire le nombre de connexions.

- la quantification d'une structure (*quantitative learning*) correspond à l'inférence des paramètres d'un graphe donné à l'aide d'un jeu de données.

Notre problématique se situe entre ces deux extrêmes : un certains nombre de relations de dépendance ou d'indépendance sont connues *a priori* et on souhaite déterminer les autres relations.

Deux notions d'indépendance peuvent être utilisées pour la construction du graphe : l'indépendance marginale et l'indépendance conditionnelle dont on rappelle la définition.

Définition 1. *Soit* $\{X_i, i = 1, 2, ...\}$ *un ensemble de variable aléatoires.*

1. *On dit que deux variables* X_1 *et* X_2 *sont (marginalement) indépendantes si et seulement si la loi bivariable* $f_{12}(x_1, x_2)$ *se décompose en* $f_{12}(x_1, x_2) = f_1(x_1)f_2(x_2)$, $f_1(x)$ *et* $f_2(x)$ *étant les lois marginales de* X_1 *et de* X_2. *On écrit alors* $X_1 \perp\!\!\!\perp X_2$.

2. *On dit que deux variables* X_1 *et* X_2 *sont indépendantes conditionnellement aux autres variables si et seulement si la loi conditionnelle bivariable* $f_{12|3,...}(x_1, x_2)$ *se factorise en*

$$f_{12|3,...}(x_1, x_2) = f_{1|3,...}(x_1)f_{2|3,...}(x_2).$$

On écrit alors $X_1 \perp\!\!\!\perp X_2|\{X_3, ...\}$.

FIG. 2.1: *Exemple d'un graphe de covariance pour 4 variables. Il correspond à la relation d'indépendance entre Y_1 et Y_4 ($Y_1 \perp\!\!\!\perp Y_4$). Dans le cas d'un modèle multigaussien il se traduit par un coefficient de corrélation de Y_1 et Y_4 nul.*

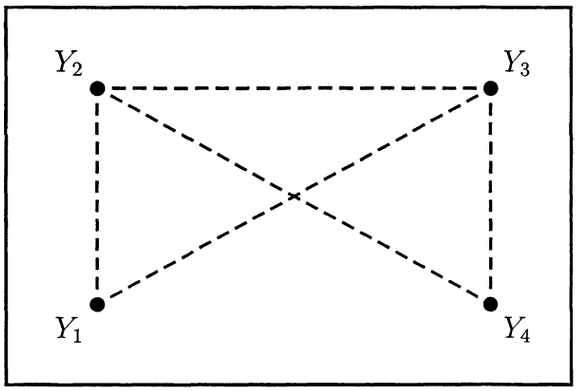
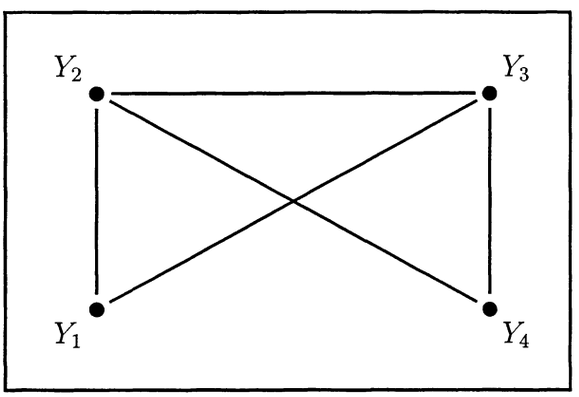


FIG. 2.2: *Exemple d'un graphe de concentration pour 4 variables. Il correspond à la relation d'indépendance « Y_1 indépendant de Y_4 si on connaît Y_2 Y_3 » ($Y_1 \perp\!\!\!\perp Y_4 | \{Y_2, Y_3\}$). Pour une distribution multigaussienne, il se traduit par un coefficient nul dans l'inverse de la matrice de covariance de (Y_1, Y_2, Y_3, Y_4), correspondant aux variables Y_1 et Y_4.*



Dans la notion de dépendance marginale, on «oublie» les autres variables. Dans la notion de dépendance conditionnelle, l'ensemble complet des variables intervient. Un énoncé précis des propriétés statistiques de ces deux notions est présenté dans [13].

Dans le cas de variables continues, multigaussiennes, l'indépendance marginale correspond à la nullité de la covariance (ou de la corrélation) ce qui a donné naissance à la dénomination de *graphe de covariance* au modèle graphique correspondant. Les connexions sont alors représentées par des segments tiretés (Figure 2.1). L'indépendance conditionnelle (aux autres variables) correspond, quant à elle, à la nullité du coefficient de la matrice de concentration (i.e. l'inverse de la matrice de covariance ou l'inverse de la matrice de corrélation) et les graphes associés sont appelés *graphes de concentration*. Les connexions sont alors représentées par des segments en trait continu (Figure 2.2).

Bien que ces définitions s'appuient sur des notions bien différentes de l'indépendance, il faut se rendre compte que ces 2 types de modèle sont, d'une certaine façon, identiques. Dans les 2 cas, l'objectif est de modéliser la loi multivariée de l'ensemble des variables. Avec les graphes de covariance, on s'appuie sur l'indépendance marginale alors qu'avec les graphes de concentration, on s'appuie sur l'indépendance conditionnelle. Dans les deux cas. il ne s'agit que d'**une représen-**

tation du même objet, la loi multivariée. Mais c'est au niveau de l'interprétation de la structure que la différence apparaît. Sous l'impulsion de certains auteurs tels que Cox et Wermuth (cf. [11], [10] par exemple), les graphes de concentration et les formes d'indépendance conditionnelle sont privilégiés car leur interprétation *causale* serait plus riche. Cette préférence semble s'appuyer sur le fait que, dans de nombreux cas pratiques rencontrés, la matrice de concentration contient plus de faibles valeurs (assimilables à une indépendance conditionnelle) que la matrice de variance. Traduit en termes de graphe, le graphe de concentration contient moins de connexions que le graphe de covariance. D'autres raisons sont avancées (plus philosophiques peut-être) : en particulier, l'**interprétation causale** des graphes de concentration, ou plus précisément celle de la notion d'indépendance conditionnelle, serait plus riche, plus forte, ... En effet, cette notion d'indépendance conditionnelle utilise toute l'information disponible pour établir les relations de dépendance ; elle peut se formuler ainsi « *si on connaît telles et telles caractéristiques, les variables* X_1 *et* X_2 *sont indépendantes* ». À l'inverse, la notion de dépendance conditionnelle peut se traduire par « *même si on connaît telles et telles caractéristiques, les variables* X_1 *et* X_2 *sont encore dépendantes* ». Quant à la notion de dépendance marginale , elle considère chaque paire de variable de façon isolée : « *si on ne regarde que les variables* X_1 *et* X_2 *en oubliant l'existence des autres variables, on observe que les variables* X_1 *et* X_2 *sont dépendantes* ». C'est pourquoi, certains auteurs indiquent que l'interprétation causale du lien entre les 2 variables est plus forte en utilisant la notion de dépendance conditionnelle.

2.4.2 Les modèles graphiques à interprétation causale : *covariance selection*

Dempster a sans doute été l'un des pionniers dans l'utilisation de la matrice de concentration pour aborder les données multivariées en introduisant la technique baptisée « sélection de variance de Dempster » (*covariance selection of Dempster*, [11] p. 91 ou [10]), dont l'article « fondateur » est [14].

Dans ce modèle, on fixe *a priori* des éléments de la matrice de concentration à 0, selon les hypothèses que l'on souhaite imposer. Au vu de la littérature, c'est la première fois que dans les modèles graphiques, on impose des conditions à la structure, sur la base de connaissances préétablies.

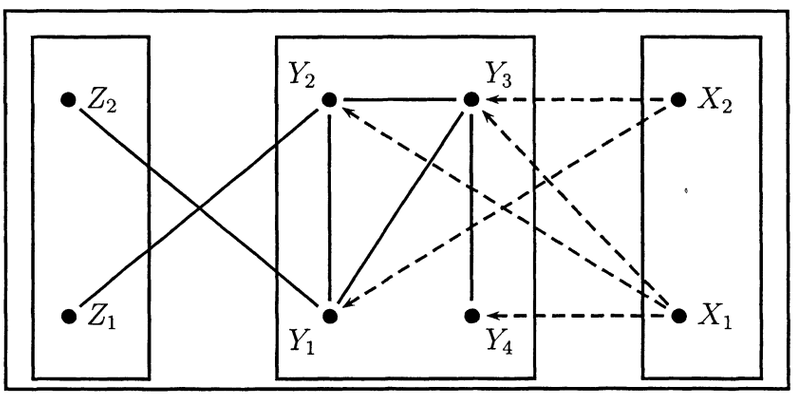
2.4.3 Les modèles graphiques à interprétation causale : les chaînes de graphes

Afin de pouvoir s'adapter à des problématiques plus variées et pour enrichir les interprétations causales des modèles graphiques, différentes extensions des modèles de graphe de concentration ou de covariance (les modèles classiques) sont proposées, en particulier les chaînes de graphes (*chain graphs*) [10], [47] [11]. Ces modèles correspondent aux cas (très nombreux) où l'ensemble des variables peut être décomposé en un certain nombre de sous-ensembles (au moins 2). Ces sous-ensembles sont ordonnés par une relation que l'on peut qualifier d'explicative ou causale. Soient $\{X_i, i \in I_X\}$, $\{Y_j, j \in I_Y\}$, $\{Z_k, k \in I_Z\}$, ... les sous-ensembles de variables considérés. On cherche à les structurer de telle sorte que :

- $\{X_i, i \in I_X\}$ explique $\{Y_j, j \in I_Y\}$,
- $\{X_i, i \in I_X\}$ et $\{Y_j, j \in I_Y\}$ expliquent $\{Z_k, k \in I_Z\}$,
- ...

À cette structure inter-blocs de variables s'ajoute la structure intra-bloc de variables.

FIG. 2.3: *Exemple d'une chaîne de graphes avec 3 blocs de 2, 4 et 2 variables.* (Z_1, Z_2) *sont les variables à expliquer par* (Y_1, Y_2, Y_3, Y_4) *qui elles-mêmes sont des variables à expliquer par* (X_1, X_2). *Le graphe indique les 14 relations d'indépendance suivantes: (1)* $X_1 \perp\!\!\!\perp X_2$ *(comme se sont les seules variables explicatives, il n'y a pas de conditionnement), (2)* $Y_4 \perp\!\!\!\perp X_2|X_1$, *(3)* $Y_2 \perp\!\!\!\perp X_2|X_1$, *(4)* $Y_1 \perp\!\!\!\perp X_1|X_2$, *(5)* $Y_2 \perp\!\!\!\perp Y_4|X_1, X_2, Y_1, Y_3$, *(6)* $Y_1 \perp\!\!\!\perp Y_4|X_1, X_2, Y_2, Y_3$, *(7)* $Z_1 \perp\!\!\!\perp Z_2|(Y_1, Y_2, Y_3, Y_4, X_1, X_2)$, *(8)* $(Z_1, Z_2) \perp\!\!\!\perp (X_1, X_2)|(Y_1, Y_2, Y_3, Y_4)$, *(9)* $Z_1 \perp\!\!\!\perp Y_1|(Y_2, Y_3, Y_4, X_1, X_2)$, *(10)* $Z_1 \perp\!\!\!\perp Y_3|(Y_2, Y_1, Y_4, X_1, X_2)$, *(11)* $Z_1 \perp\!\!\!\perp Y_4|(Y_2, Y_1, Y_3, X_1, X_2)$, *(12)* $Z_2 \perp\!\!\!\perp Y_2|(Y_1, Y_3, Y_4, X_1, X_2)$, *(13)* $Z_2 \perp\!\!\!\perp Y_3|(Y_1, Y_2, Y_4, X_1, X_2)$ *et (14)* $Z_2 \perp\!\!\!\perp Y_4|(Y_1, Y_2, Y_3, X_1, X_2)$. *Dans le cas d'un modèle multi-gaussien, on considère la régression linéaire de* (Z_1, Z_2) *par* (Y_1, Y_2, Y_3, Y_4) *et de* (Y_1, Y_2, Y_3, Y_4) *par* (X_1, X_2) *; la condition 8 indique qu'on n'utilise pas* (X_1, X_2) *comme régresseurs de* (Z_1, Z_2) *; il correspond à des coefficients de régression nuls entre* Z_1 *et* (Y_1, Y_3, Y_4) *(conditions 9, 10 et 11), entre* Z_2 *et* (Y_2, Y_3, Y_4) *(conditions 12, 13 et 14), entre* Y_4 *et* X_2 *(condition 2), entre* Y_2 *et* X_2 *(condition 3) et entre* Y_1 *et* X_1 *(condition 4), à un coefficient de corrélation nul entre* X_1 *et* X_2 *(condition 1), et dans les matrices de concentration des résidus des régressions, à des valeurs nulles, relatives à* Y_4 *et* Y_2 *(condition 5), à* Y_4 *et* Y_1 *(condition 6), et à* Z_1 *et* Z_2 *(condition 7).*



La représentation graphique de ce type de structure consiste à regrouper les variables d'un même bloc dans une boîte et de le représenter de la droite vers la gauche, un bloc explicatif étant à droite d'un bloc à expliquer (voir Figure 2.3). Dans les cas des graphes de concentration ou de covariance, les liens entre les nœuds n'étaient pas orientés. Dans les cas des chaînes de graphes, les liens à l'intérieur des boîtes (structure intra-bloc) ne sont toujours pas orientés, mais les liens entre boîtes, qui représentent des relations explicatives sont représentés par des flèches allant de la variable explicative vers la variable à expliquer.

Dans les graphes de concentration et de covariance, on utilise 2 types de connexion selon la notion d'indépendance utilisée. Dans les chaînes de graphe, on utilise 4 types de liens, selon les notion d'indépendance utilisée :

- à l'intérieur d'un bloc $\{Z_k, k \in I_Z\}$, l'indépendance est toujours (au minimum) conditionnelle aux variables explicatives (des blocs précédents ($\{X_i\}$, $\{Y_j\}$, ...) et :
  - on utilise des traits continus quand l'indépendance est aussi conditionnelle aux autres variables du bloc ; ainsi, l'absence de lien entre les variables Z_1 et Z_2 signifie $Z_1 \perp\!\!\!\perp Z_2|\{Z_{k\neq 1,2}\} \cup \{X_i\} \cup \{Y_j\}$,
  - on utilise des traits tiretés quand l'indépendance n'est pas conditionnelle aux autres variables du bloc ; ainsi, l'absence de lien entre les variables Z_1 et Z_2 signifie $Z_1 \perp\!\!\!\perp Z_2|\{X_i\} \cup \{Y_j\})$,

- entre deux blocs $\{Z_k, k \in I_Z\}$ et $\{Y_j, j \in I_Y\}$, le premier étant expliqué par le second, lui même expliqué par $\{X_i\}$, ... :
  - on utilise des flèches en traits continus quand, pour qualifier la dépendance entre 2 variables Y_1 et Z_1, on utilise les autres variables du bloc à expliquer (Z): ainsi, l'absence de flèche de Y_1 vers Z_1 signifie $(Z_1 \perp\!\!\!\perp Y_1 | \{Z_{j\neq 1}\} \cup \{Y_{j\neq 1}\} \cup \{X_i\})$; en terme de régression, cela signifie qu'on considère la régression de Z_1 par les variables du bloc Y **mais aussi les autres variables du bloc** Z et que le coefficient de régression de Z_1 par Y_1 est alors nul.
  - on utilise des flèches en traits continus quand, pour qualifier la dépendance entre 2 variables Z_1 et Y_1, on n'utilise pas les autres variables du bloc à expliquer Z dans le conditionnement : ainsi, l'absence de flèche de Y_1 vers Z_1 signifie $(Z_1 \perp\!\!\!\perp Y_1 | \{Y_{j\neq 1}\} \cup \{X_i\} \cup ...)$; en terme de régression, cela signifie qu'on considère la régression de Z_1 par les variables du bloc Y **sans inclure les autres variables du bloc** Z et que le coefficient de régression de Z_1 par Y_1 est alors nul.

Pour des raisons de cohérence, on n'utilise généralement qu'un type de flèche (continu ou tireté).

Il faut noter que dans le cas de variables continues, les relations d'explication peuvent se lire en terme de régression linéaire. Par ailleurs, le sous-graphe d'un unique bloc correspond à un graphe (de covariance ou de concentration selon la notion d'indépendance utilisée) mais toujours conditionnellement à l'ensemble des variables explicatives de ce bloc.

2.4.4 Modèles graphiques et variables ordinales

Les modèles graphiques peuvent être utilisées dans le cas de mélange de variables continues et de variables discrètes. La distribution multivariée considérée est alors du type *CG-distribution* ([27, 12, 48, 16, 11]). Cependant, ce type de distribution ne tient pas compte de l'aspect ordinal des variables discrètes et les variables discrètes ne peuvent pas être des variables à expliquer dans les chaînes de graphes.

2.4.5 Introduction des variables latentes dans les modèles graphiques

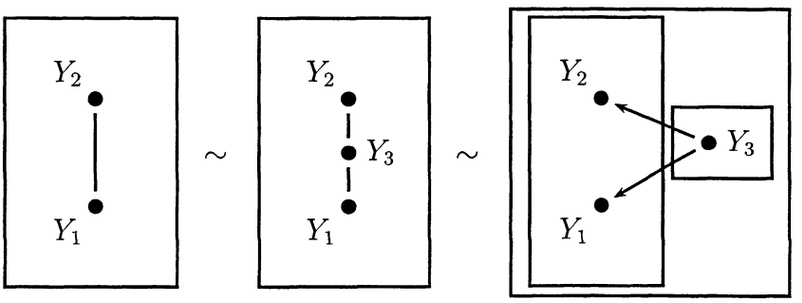
L'introduction des variables latentes dans les modèles graphiques s'est faite de 2 façons. Premièrement, au niveau du modèle probabiliste le plus élémentaire, une relation de dépendance entre 2 variables d'un même bloc peut aboutir à une relation d'indépendance conditionnellement à une variable latente (Figure 2.4). Deuxièmement, au niveau de la pratique (comme notre problème par exemple), de nombreux problèmes font intervenir des variables qui ne sont pas mesurables. En terme de notion d'indépendance, ces 2 niveaux conduisent à la même interprétation : si on connaît les variables cachées, les variables considérées deviennent indépendantes.

Dans la mise en œuvre des modèles graphiques, l'introduction de variables latentes semble rendre les choses difficiles. Ce champ semble peu développé dans la bibliographie, sans doute pour des raisons liées à l'identifiabilité des paramètres. Ce point est abordé plus loin. Le cas pour lequel il existe le plus de résultats est sans doute la formulation de l'analyse factorielle en terme de modèle graphique. Il s'agit alors d'une chaîne de graphe dans laquelle les variables mesurées sont les variables à expliquer et les variables latentes, les variables explicatives.

2.4.6 Liens entre modèles graphiques et modèles à équations structurelles

Comme dans les modèles à équations structurelles, toutes les variables sont supposées être aléatoires ; elles entrent toutes dans la loi de probabilité multivariée, support du modèle graphique

FIG. 2.4: *Exemple d'introduction d'une troisième variable Y_3 pour « briser » la dépendance entre 2 variables Y_1 et Y_2. Les 3 graphes sont équivalents. Le graphe de droite signifie : si on connaît la variable Y_3 et si on fait une régression de Y_1 et Y_2 par Y_3, les résidus sont indépendants. Le graphe du milieu signifie : si on connaît la variable Y_3, les deux variables Y_1 et Y_2 sont indépendantes. Celui de gauche signifie : si on oublie la variable Y_3, Y_1 et Y_2 ne sont plus indépendantes.*



et elles ont toutes une distribution marginale, qu'elles soient continues ou discrètes.

Les modèles graphiques prennent en compte dès leur origine les 2 types de variable avec, cependant, certaines restrictions. Historiquement, les modèles graphiques étaient soit entièrement discrets, soit entièrement continus et les lois supports étaient respectivement les distributions multinomiales et multinormales. Puis ont été développés des modèles pouvant mêler des variables discrètes et continues. Mais, sans doute pour des raisons méthodologiques et algorithmiques, ces modèles sont restreints de telle sorte que toutes les distributions de la factorisation de la loi conjointe doivent appartenir à la famille exponentielle. En particulier, les variable qualitatives (discrètes) ne peuvent pas être expliquées par des variables continues ; seul l'inverse est possible avec l'emploi des distributions conditionnellement gaussiennes (*CG-distribution*).

Dans les modèles à équations structurelles, historiquement toutes les variables sont continues et des extensions faisant intervenir des variables qualitatives ont vu le jour (*latent class models* par exemple [2])

Comme mentionné à plusieurs reprises ci-dessus, pour une distribution multivariée donnée, on peut définir plusieurs modèles graphiques, selon l'interprétation (causale) que l'on souhaite obtenir. N'y a-t-il pas alors de points communs entre les modèles graphiques et les modèles à équations structurelles ? D'après Cox et Wermuth ([11]), dans certaines conditions précises, les modèles à équations structurelles correspondent aux modèles graphiques : il y a équivalence entre les 2 démarches si et seulement si la formulation du modèle à équations structurelles définit de façon correcte une loi de distribution multivariée, ce qui semble ne pas être toujours le cas.

De même, comme cela a déjà été indiqué, il existe une différence entre les 2 modèles :

- dans le cas des modèles graphiques, c'est l'absence de dépendance (avec une certaine signification en terme d'indépendance conditionnelle ou non) qui est recherchée ou qui guide la recherche de la structure,

- dans le cas des modèles à équations structurelles, l'accent porte d'avantage ou uniquement sur les relations pouvant exister et la notion d'indépendance n'est pas clarifiée explicitement. Elle est sous-entendue, mais on ne sait pas véritablement ce qu'elle signifie.

Il en résulte que, les deux représentations graphiques associées aux deux modèles sont totalement différentes. Dans le cas des modèles graphiques, c'est l'absence de lien qui semble avoir le plus d'intérêt alors que dans le cas des équations structurelles, c'est la présence d'un lien qui structure. En outre, la représentation des modèles graphiques peut être qualifiée de plus économe car elle n'introduit que les variables « utiles » et n'a pas besoin d'introduire les variables résiduelles telles

que e_x ou e_y par exemple (cf. eq. (2.3) par exemple) qui sont des *arte facts* dans la mesure où elles ne servent qu'à définir les distributions de certaines variables conditionnellement à d'autres.

En fait, ces deux approches correspondent aussi à deux types de problèmes différents :

- l'apprentissage structural (*structural learning*) qui consiste à découvrir les relations de dépendance/indépendance parmi un ensemble de variables ; en terme d'hypothèse, on teste l'existence des liaisons entre chaque paire de variables ;

- la quantification des structures (*quantitative learning*) qui présuppose l'existence d'une certaine structure, et dont l'objet est de quantifier les relations présupposées.

Cette dichotomie est bien sûre caricaturale dans la mesure où, dans la pratique, l'objectif se situe entre ces 2 extrêmes, selon le poids des connaissances spécifiques au domaine étudié qui sont utilisées pour superviser l'analyse. Dans le cas des modèles graphiques, cette connaissance sert à segmenter les variables en différents blocs et à les ordonner en chaînes de graphes, l'apprentissage structural consistant à établir l'existence de certains liens. Cette segmentation peut être plus ou moins poussée selon le degré de connaissance. Dans le cas des équations structurelles, cette connaissance sert à construire une structure *a priori*. La quantité de connaissances peut se mesurer à la « complexité » et la « quantité de contraintes imposées » à la structure.

2.4.7 Notion d'effet fixe dans les modèles graphiques

La notion d'effet fixe dans les modèles graphiques n'existe pas de façon explicite. Toutes les variables sont supposées aléatoires. Cependant, dans les chaînes de graphes, les valeurs des variables du bloc purement explicatif (le plus à droite) peuvent être considérées comme fixes, c'est à dire qu'on ne cherche pas à spécifier la distribution dans ce bloc (dans ce cas, par convention, le bloc est entouré d'une double frontière). S'agit-il pour autant d'effet fixe ? En tout cas, cette question ne semble pas soulever d'intérêt particulier dans les modèles à équations structurelles, soit parce que cela n'entraîne pas de difficultés supplémentaires, soit parce que cette situation est rarement considérée dans la pratique.

2.4.8 Modèles graphiques et causalité

Les chaînes de graphes ont été développées pour faciliter l'interprétation causale (sous toute les qualifications qu'on peut lui substituer) à l'intérieur d'un ensemble de variables. Mais, comme indiqué ci-dessus pour les graphes de concentration ou de covariance, un graphe associé à un ensemble de variables n'est pas unique : il correspond à une factorisation particulière de la loi multivariée de cet ensemble. Ainsi, tout modèle exprimé par une chaîne de graphe peut être exprimé par un graphe de covariance ou de concentration ou par une autre (plusieurs même) chaîne de graphes. La différence entre toutes ces représentations réside dans l'interprétation *substantive*, c'est à dire le sens que l'on veut donner à la structure entre les variables. Cela se fait en dehors de tout fondement statistique et probabiliste. La recherche d'une chaîne de graphes donnée (plutôt qu'une autre) correspond en quelque sorte à la formulation de l'hypothèse de l'existence d'une certaine structure causale ou explicative ou génératrice du jeu de données. C'est un choix arbitraire au niveau du modèle statistique. Mais il ne l'est pas au niveau du domaine exploré par les données et de la substance que l'on veut donner à l'analyse : il est lié à la structure causale connue ou suspectée.

2.5 Les approches « analyses de données à la française »

Les approches « analyses de données à la française » se distinguent des modèles énoncés ci-dessus par le fait qu'elles ne font référence à aucune loi de probabilité : il s'agit plutôt d'une approche descriptive et géométrique de l'analyse de données. La technique de base est l'analyse en composante principale [15, 18]. Elle a été utilisée dans les premières analyses de l'étude TOLUÈNE [9]. Il a déjà été indiqué pourquoi cette technique n'est pas la plus adaptée dans le présent contexte :

- toutes les variables sont mises sur le même plan, il n'y a ni variables explicatives ni variables à expliquer,

- on ne peut imposer de structure *a priori*, la structure est entièrement interprétée *a posteriori*.

Dans cette famille de techniques, celle qui se rapproche le plus de la problématique étudiée dans ce rapport est l'*approche PLS* ([45]). Le concept de l'approche PLS est proche de celui des modèles à équations structurelles (mais plus simple). Les variables sont séparées en sous-ensembles. Chaque sous-ensemble de variables explore un domaine (vis à vis de la problématique du jeu de données) correspondant à une (ou plusieurs) variable(s) latente(s). Les variables latentes sont reliées par des relations linéaires, certaines étant explicatives, d'autres étant les variables à expliquer. Là encore, il n'y a pas d'hypothèse sur la distribution des variables. Cette technique n'est utilisable que si toutes les variables sont continues, ce qui la rend inadaptée au cas présent.

2.6 Des problèmes pratiques

Vis à vis de la mise en œuvre des modèles graphiques et de modèles à équations structurelles, 2 problèmes pratiques au moins sont soulevés dans la bibliographie :

1. l'identifiabilité des paramètres liée à l'introduction de variables latentes ,

2. la robustesse des résultats.

2.6.1 L'identifiabilité du modèle

Si le problème n'est pas suffisamment spécifié, certains paramètres du modèle ne sont pas identifiables. C'est ce qui apparaît clairement en analyse factorielle (cf. section 2.2.1). Si on ne supervise pas la recherche des facteurs de l'analyse factorielle, il y a une infinité de solutions, toutes équivalentes à une rotation près. On ajoute alors des contraintes algébriques pour lever les indéterminations. Dans les autres modèles, dans lesquels la recherche de structure est supervisée par une connaissance *a priori* de la structure, il peut arriver que cette connaissance ne soit pas suffisante et qu'il demeure des problèmes d'identifiabilité. Cela signifie, en quelque sorte, qu'il n'y a pas assez de contraintes pour définir l'orientation de la structure latente. À l'inverse, la structure proposée par un graphe ne correspond pas toujours à la factorisation d'une loi de probabilité (par exemple, un cycle dans un graphe $X \to Z \to Y \to X$). La question de savoir si un tel modèle est identifiable n'a pas encore de réponse définitive. Dans le cas d'absence d'effet fixe et dans le cadre de l'approche à équation structurelles, Bollen ([6]) passe en revue quelques règles pour vérifier l'identifiabilité dans le cas de modèle à équations structurelles. Quelques résultats sont présentés dans [43] pour l'analyse factorielle.

2.6.2 Robustesse

Plusieurs problèmes de robustesse peuvent se poser. Celui de la normalité des variables mesurées est largement étudié, au moins dans le domaine du *Factor Analysys* et des modèles à équations structurelles (voir [6] ou [7] par exemple, ainsi que les indications données par certains logiciels). Les tests et les estimateurs conservent leurs propriétés asymptotiques dans un grand nombre de cas où cette hypothèse de normalité n'est pas valide. Des procédures d'inférence ont été développées et sont proposées dans les logiciels pour d'autres types de distribution.

Par ailleurs, dans le cadre de recherche de structure (latente ou non) on peut considérer la robustesse par rapport à une mauvaise spécification de la structure *a priori*, et en particulier par rapport à une mauvaise spécification de la structure des variables latentes. Dans la présente étude, comme l'intérêt porte plus particulièrement sur l'effet éventuel de l'exposition sur des performances cognitives choisies, quelle est la robustesse de l'analyse dans le cas où la structure des fonctions mentales que l'on souhaite imposer, n'est pas cohérente avec la « réalité ». Par exemple, il peut y avoir des fonctions mentales qui ne sont pas dans le modèle, mais qui peuvent avoir une influence non négligeable et qui se traduit par le fait que l'indépendance des variables des tests conditionnellement aux fonctions mentales considérées n'est pas vérifiée. Ce problème n'est pas mentionné dans l'approche à équations structurelles. Cela provient sans doute du fait qu'on met un *a priori* fort sur la structure (d'où le nom de ces modèles). Cependant, des procédures sont développées pour tester l'ajustement des modèles sur les données, et pour comparer différents modèles.

Dans le cadre des modèles graphiques, la recherche de « la » structure à l'aide d'un jeu de données est considérée comme une *sélection de modèle* [25, 46]. Plus récemment, le concept d' « incertitude de modèle » (*model uncertainty*) a émergé : l'idée étant qu'il n'y a pas qu'une bonne solution mais qu'un ensemble de solutions «raisonnables» peut être accepté [25, 31]. Dans le cas de chaîne de graphes, la séparation et l'ordonnancement des variables en blocs successifs sont imposés par le corpus de connaissances de la discipline correspondant au jeu de données. Il n'y a pas lieu de la remettre en cause ; tout au plus le corpus de connaissances peut conduire à proposer différentes chaînes de graphes qui sont confrontées aux données par les techniques citées ci-dessus.

Dans le cas faisant intervenir des effets fixes, Sammel et al. [35] soulèvent le problème de robustesse liée au fait que les paramètres des effets fixes interviennent aussi dans la structure de covariance des variables à expliquer.

2.7 Conclusion

Cette étude de la littérature a permis de répertorier un certain nombre de modèles et de techniques pour l'analyse de jeux de données multivariées ou faisant intervenir des variables latentes. Le formalisme des modèles graphiques et des chaînes de graphes a été retenu : la structure des dépendances multivariées connues *a priori* ou recherchées va être représentée par une chaîne de graphes dans laquelle la nature de la dépendance est définie sans ambiguïté ; en particulier, les dépendances inter-fonctions mentales seront des dépendances conditionnelles. Le chapitre suivant donne en détail le cheminement allant de la structuration des données par le corpus de connaissance à la spécification du modèle probabiliste, le trait d'union étant le graphe de dépendance.

Chapitre 3

Des modèles neuropsychologiques aux modèles statistiques

3.1 Introduction

L'objet de ce chapitre est de faire une représentation du problème qui fasse la synthèse de la connaissance en neuropsychologie sur le sujet étudié, qui puisse être transcrit en un modèle probabiliste, et qui soit à la fois compréhensible par les neuropsychologues et les statisticiens. Les modèles graphiques fournissent le cadre de cette représentation. Le problème sera représenté par un graphe qui traduit la connaissance et/ou les hypothèses à tester formulées par les neuropsychologues. Les nœuds sont les variables des tests et d'environnement ainsi que les fonctions mentales. Du point de vue de la neuropsychologie, les flèches entre les nœuds traduisent une influence, les liens (sans flèche) traduisent une co-relation. Du point de vue probabiliste, les flèches représentent des coefficients d'une relation linéaire et les liens (sans flèche) représentent des dépendances entre les variables reliées.

3.2 Structure du graphe symbolisant le point de vue de la neuropsychologie

Dans le problème étudié, on distingue 3 entités :

- les variables des tests (Y), qui sont les variables à expliquer,
- les fonctions mentales (F) qui ne sont pas mesurées et sont donc représentées par des variables latentes,
- les variables d'environnement (X).

Les relations entre ces variables correspondent à un schéma causal (ou générateur ou explicatif) résumé par $Y \leftarrow F \leftarrow X$:

- relations $Y \leftarrow F$: la construction des tests psychométriques définit le nombre de fonctions mentales et les influences que celles-ci ont sur les résultats des tests neurocomportementaux,
- relations $F \leftarrow X$: certains facteurs d'environnement peuvent influencer les fonctions mentales. Par ailleurs, l' éventuel effet de l'exposition aux solvants est l'effet étudié.

Ces relations sont représentées par des flèches. En outre, il peut exister certaines co-relations entre les fonctions mentales. Certaines de ces relations ou co-relations peuvent être des éléments connus ou des relations hypothétiques que l'on veut tester.

Du point de vue de la neuropsychologie, l'ensemble des hypothèses est représenté par graphe. Un exemple de graphe est présenté à la Figure 3.1. Ce graphe contient 11 variables à expliquer - les tests neurocomportementaux, notées de Y_1 à Y_{11} et 5 variables explicatives notées de X_1 à X_5 (dont l'exposition à un produit X_1). On suppose que les tests explorent 4 fonctions mentales notées de F_1 à F_4. Ce graphe correspond aux hypothèses suivantes :

- la fonction mentale F_3 influence toutes les variables de test,
- la fonction mentale F_1 influence les variables de test Y_1, Y_3, Y_4, Y_5, Y_8, et Y_{11},
- la fonction mentale F_2 influence les variables de test Y_2 et Y_{11},
- la fonction mentale F_4 influence les variables de test Y_5, Y_8, Y_{10},
- la fonction mentale F_1 est influencée par les variables explicatives X_1, X_3 et X_4,
- la fonction mentale F_2 est influencée par les variables explicatives X_1, X_2, X_3 et X_5,
- la fonctions mentale F_3 est influencée par les variables explicatives X_1, X_2, X_4 et X_5,
- la fonction mentale F_4 est influencée par les variables explicatives X_1, X_2 et X_4,
- la variable d'exposition peut influencer les 4 fonctions mentales : les quatre relations sont présentes car on cherche à tester l'effet de l'exposition sur les 4 fonctions mentales,
- il y a une co-relation entre les paires de fonctions mentales (F_1, F_2), (F_1, F_3), (F_2, F_3) et (F_3, F_4).

3.3 Définition du modèle probabiliste

3.3.1 Transcription du graphe en modèle probabiliste

La transcription consiste à établir des conventions pour traduire le graphe en modèle probabiliste. Les variables des tests (Y) et les fonctions mentales (F) sont représentées par des variables aléatoires . Comme les fonctions mentales ne sont pas mesurées, ce sont des variables latentes. Pour simplifier l'exposé, on suppose que toutes les variables Y sont continues. Les fonctions mentales sont modélisés par des variables continues elles aussi.

Le graphe correspond alors à la loi de probabilité conjointe de ces variables $\mathcal{F}(y, f, x)$ en suivant les conventions des modèles graphiques qui sont rappelées. Au schéma causal résumé par $Y \leftarrow F \leftarrow X$ correspond la décomposition de $\mathcal{F}(y, f, x)$ en lois conditionnelles :

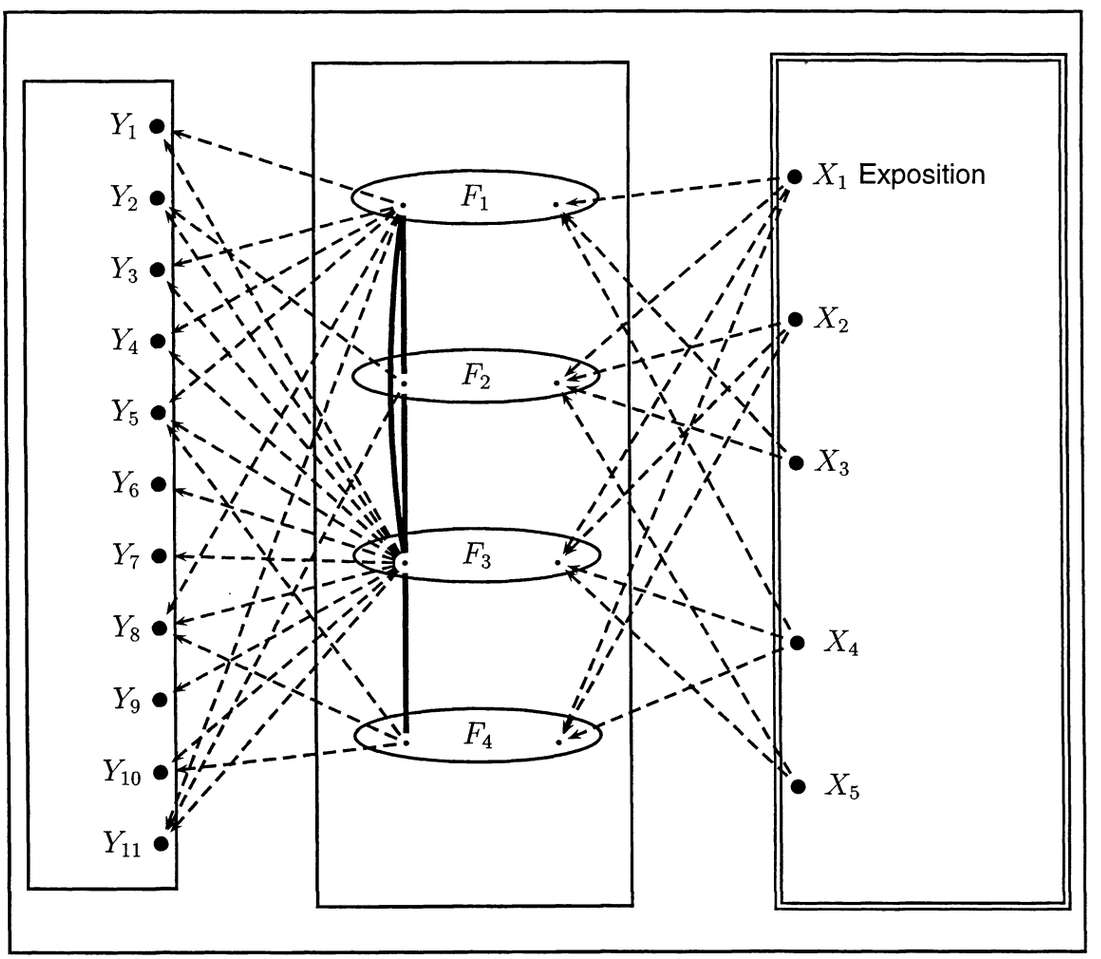
$$\mathcal{F}(y, f, x) = \mathcal{F}(y|f)\mathcal{F}(f|x). \tag{3.1}$$

Les influences des fonctions mentales sur les variables des tests sont modélisées par des régressions linéaires :

$$\mathbf{E}(Y|F) = \Lambda F \tag{3.2}$$

$$\mathbf{var}(Y|F) = \mathrm{diag}(\Sigma_Y). \tag{3.3}$$

FIG. 3.1: *Exemple de la représentation symbolique des connaissances du point de vue de la neuropsychologie. Il y a 11 variables à expliquer - les tests neurocomportementaux, notées de* Y_1 *à* Y_{11} *et 5 variables explicatives notées de* X_1 *à* X_5 *(dont l'exposition à un produit* X_1*). On suppose que les tests explorent 4 fonctions mentales notées de* F_1 *à* F_4*. Le graphe correspond aux hypoyhèses suivantes : la fonction mentale* F_3 *influence toutes les variables de tests, la fonction mentale* F_1 *influence les variables de test* Y_1, Y_3, Y_4, Y_5, Y_8, *et* Y_{11}, *la fonction mentale* F_2 *influence les variables de test* Y_2 *et* Y_{11}, *la fonction mentale* F_4 *influence les variables de test* Y_5, Y_8, Y_{10}, *la fonction mentale* F_1 *est influencée par les variables explicatives* X_1, X_3 *et* X_4, *la fonction mentale* F_2 *est influencée par les variables explicatives* X_1, X_2, X_3 *et* X_5, *la fonction mentale* F_3 *est influencée par les variables explicatives* X_1, X_2, X_4 *et* X_5, *la fonction mentale* F_4 *est influencée par les variables explicatives* X_1, X_2 *et* X_4, *la variable d'exposition peut influencer les 4 fonctions mentales (les quatre relations sont présentes car on cherche à tester l'effet de l'exposition sur les 4 fonctions mentales), il y a une co-relation entre les paires de fonctions mentales* (F_1, F_2), (F_1, F_3), (F_2, F_3) *et* (F_3, F_4).



Les effets des variables d'environnement sur les fonctions mentales sont modélisés par des régressions linéaires :

$$\mathbf{E}(F|X) = \beta X. \tag{3.4}$$

Chaque coefficient de Λ et β correspond à une flèche potentielle entre d'une part une fonction mentale et une variable de test et d'autre part entre une variable explicative et une fonction mentale. Pour prendre en compte la structure du graphe, à chaque flèche absente du graphe correspondra la contrainte de nullité du coefficient. Ainsi, le graphe représenté à la Figure 3.1 est associé aux deux matrices Λ et β suivantes :

$$\Lambda = \begin{pmatrix} ? & 0 & ? & 0 \\ 0 & ? & ? & 0 \\ ? & 0 & ? & 0 \\ ? & 0 & ? & 0 \\ ? & 0 & ? & ? \\ 0 & 0 & ? & 0 \\ 0 & 0 & ? & 0 \\ ? & 0 & ? & ? \\ 0 & 0 & ? & 0 \\ 0 & 0 & ? & ? \\ ? & ? & ? & 0 \end{pmatrix} \tag{3.5}$$

$$\beta = \begin{pmatrix} ? & 0 & ? & ? & ? & 0 \\ ? & ? & ? & 0 & 0 & ? \\ ? & ? & 0 & ? & ? & ? \\ ? & ? & 0 & ? & ? & 0 \end{pmatrix}. \tag{3.6}$$

Les co-relations entre les fonctions mentales sont représentés par des dépendances conditionnelles entre les fonctions mentales : l'absence de lien entre deux fonctions mentales correspond à la nullité du coefficient de régression partiel des résidus du modèle linéaire généralisé :

$$\mathbf{var}(F|X) = \Sigma_F \tag{3.7}$$

avec, pour le graphe de la Figure 3.1

$$\Sigma_F^{-1} = \begin{matrix} ? & ? & ? & 0 \\ ? & ? & ? & 0 \\ ? & ? & ? & ? \\ 0 & 0 & ? & ? \end{matrix} \tag{3.8}$$

Dans le cas de variables de test continues, on suppose que les lois $\mathcal{F}(y|f)$ et $\mathcal{F}(f|x)$ sont multigaussiennes.

3.3.2 Conventions et contraintes liées à l'interprétabilité

Les fonctions mentales sont représentées par des variables de performance

Pour rendre les résultats interprétables, il faut imposer aux fonctions mentales une convention supplémentaire. Les variables latentes sont supposées être des **variables de performance**, c'est à dire qu'à des valeurs élevées correspondent des fonctions mentales performantes (comme c'est le cas pour le «QI»). Cette convention est nécessaire pour l'interprétation du modèle. En particulier, un coefficient de β négatif correspond à une dégradation de la fonction mentale. Un coefficient de

Λ positif correspond à une influence positive de la fonction mentale sur la variable de test. Pour cela, si la neuropsychologie suppose que toutes les fonctions mentales ont une influence positive sur les résultats des tests, alors, dans le modèle statistique, on impose la condition de positivité de tous les coefficients de régression Λ.

Ce type de connaissance est aussi nécessaire pour rendre les calculs réalisables : il s'agit d'une condition d'identifiabilité des paramètres du modèle. Ce point est abordé par la suite.

Choix de l'amplitude des fonctions mentales, définition d'une valeur et d'un individu de référence

Par ailleurs, il existe un certain arbitraire dans la définition probabiliste des fonctions mentales. Celles-ci sont modélisées par des variables continues : mais quelle est alors la signification de l'attribution d'une valeur donnée à une fonction mentale particulière ? Comme on le voit, ce n'est pas une unique valeur qui importe, mais plutôt la différence relative entre deux valeurs d'une même fonction mentale (pour deux individus) ou peut être entre deux valeurs de deux fonctions mentales différentes (pour un même individu). Il est donc nécessaire de fixer des paramètres liés à l'amplitude des fonctions mentales. Ceci peut être fait de plusieurs façons :

- on peut choisir un individu «virtuel» de référence auquel on attribue une valeur **arbitraire** de référence (100 par exemple) à chaque fonction mentale,
- on peut attribuer une valeur **arbitraire** de référence (100 par exemple) à chaque fonction mentale pour la moyenne de la population étudiée,
- on peut fixer l'amplitude de chaque fonction mentale (à des valeurs **arbitraires**) en imposant la variabilité sur la population.

Comme précédemment, ce type de contrainte est aussi nécessaire pour rendre les calculs réalisables : il se traduit en condition d'identifiabilité des paramètres du modèle.

La dernière solution est sans doute la moins intuitive et son interprétation n'est pas évidente. La seconde n'est pas la plus adaptée dans le contexte d'épidémiologie en santé au travail car la référence est propre à la population de l'étude, ce qui rend les comparaisons entre études difficiles. C'est la première qui semble la mieux adaptée. L'individu *virtuel* de référence peut être un individu *virtuel* non exposé, d'un âge donné, ...ce qui permet d'interpréter les effets des paramètres étudiés. Il faut aussi fixer la valeur attribuée à chacune des fonctions mentales de cette référence. Le choix de cette valeur est **arbitraire** ; la valeur 100 permet d'interpréter les coefficients β en pourcentage.

3.4 Le modèle probabiliste

Cette section présente une récapitulation du modèle probabiliste représenté du graphe, dans le cas de variables de test (Y) continues.

1. Les variables des tests (Y) et les variables latentes des fonctions mentales (F) suivent une loi multigaussienne.
2. L'espérance mathématique des variables latentes conditionnellement aux variables explicatives (X) est définie par la relation :

$$\mathbf{E}(F|X) = 100 + \beta X. \tag{3.9}$$

3. La matrice de variance des variables latentes conditionnellement aux variables explicatives (X) est définie la relation :

$$\Sigma_F^{-1} = \begin{pmatrix} ? & ? & ? & 0 \\ ? & ? & ? & 0 \\ ? & ? & ? & ? \\ 0 & 0 & ? & ? \end{pmatrix}. \tag{3.10}$$

On fixe à «zéro» deux des valeurs de la matrice de concentration des fonctions mentales.

4. L'espérance mathématique des variables des tests Y conditionnellement aux variables latentes F est définie par la relation

$$\mathbf{E}(Y|F) = \Lambda F, \tag{3.11}$$

Λ étant une matrice dont les coefficients correspondant à une absence de flèche dans le graphe modélisé sont nuls et dont les coefficients non nuls sont positifs.

5. La matrice de variance des variables conditionnellement aux variables latentes (F) est diagonale.

De cette façon, le loi de (Y, F) conditionnellement à X est entièrement spécifiée par :

$$f(Y, F|X) = \mathcal{N}(\Lambda F, \text{diag}(\Sigma_Y))\mathcal{N}(100 + \beta X, \Sigma_F)) \tag{3.12}$$

3.5 Cas des variables de test ordinales

Un certain nombre des variables de test sont des variables ordinales non continues. Les traiter comme des variables continues est une approximation qui peut entraîner des inexactitudes dans les résultats quand le nombre de modalités de la variable ordinale est limité. Plusieurs possibilités de complexité croissante sont envisageables pour tenir compte de ce caractère ordinal des variables de test :

1. les traiter comme des variables continues,
2. les traiter comme des variables continues tronquées,
3. ajouter un bruit *faible*,
4. suivre une approche type *proportionnal odds* ([32] par exemple ou [41])

La première peut convenir si le nombre de modalités est assez élevé. La seconde a été testée sur les données TOLUÈNE mais des des problèmes techniques ont conduit à abandonner cette option. Pour la troisième, il semble qu'il n'y ait pas de problèmes techniques, mais la justification de la démarche n'est pas très satisfaisante. La quatrième a été testée sur les données PAQUID : les résultats ont été comparés à ceux obtenus par la solution 1 (la comparaison est présentée au Chapitre 7) mais on peut déjà souligner que :

- certains logiciels dédiés aux *modèles à équations structurelles* ne permettent pas de suivre cette démarche, ce qui est le cas d'EQS,
- le logiciel BUGS permet de la mettre en œuvre au prix de temps de calcul exorbitant.

3.6 Discussion, conclusion

Ce chapitre a montré comment on peut passer d'un modèle neuropsychologique à un modèle probabiliste pour analyser des données de l'épidémiologie neurocomportementale.

L'analyse pratique de ce type de modèle nécessite l'emploi de logiciels spécifiques. Le logiciel BUGS est suffisamment souple pour faire cette mise en pratique. Il a été testé sur plusieurs graphes « fictifs » associés à des données simulées, puis sur deux modèles reposant sur les deux jeux de données réels. Les résultats sont présentés dans les chapitres suivants. Pour certaines analyses, nous avons utilisé EQS qui est un logiciel dédié aux *modèles à équations structurelles* afin de comparer les deux approches.

Enfin, il faut rappeler l'importance du poids du modèle neuropsychologique ; c'est lui qui prédétermine tout ou partie de la structure du modèle, l'analyse statistique ne servant qu'à quantifier les relations présupposées.

Chapitre 4

Identifiabilité du modèle probabiliste

4.1 Introduction, définitions

Le chapitre précédent a montré comment on définit le modèle probabiliste à l'aide du graphe indiquant les relations entre les fonctions variables de tests, les fonctions mentales et les variables explicatives ainsi que des conventions sur les fonctions mentales. Cependant, toutes les questions de validité du modèle probabiliste ne sont pas levées quand il s'agit de déterminer la valeur de ses paramètres. Ces valeurs sont les solutions d'un système d'équations dont les inconnues sont les paramètres du modèle. Ces questions de validité peuvent se résumer sous le terme d'identifiabilité selon la définition suivante.

Définition 2. *Un modèle est identifiable si la solution est unique.*

Ce problème est bien réel : les conventions adoptées pour les fonctions mentales (valeur de référence, positivité des coefficients de Λ) permettent de résoudre une partie du problème d'identifiabilité. En leur absence, le modèle ne serait pas identifiable.

On peut distinguer 2 types de problème d'identifiabilité, ou plutôt deux types de problème de non-identifiabilité:

la non identifiabilité structurelle : quelles que soient les données utilisées, le modèle n'est pas identifiable : l'identifiabilité est intrinsèque au modèle.

la non identifiabilité paramétrique : la structure des données est telle que le modèle n'est pas identifiable ; cette notion est comparable aux notions d'*alias* et d'*effets confondus* en terme de plan d'expérience [22].

Seule la première notion est abordée dans la suite. Par rapport à l'équation (3.12), la non-identifiabilité du modèle s'énonce :

Définition 3. *Le modèle n'est pas identifiable si et seulement s'il existe deux jeux de paramètres distincts* $(\Lambda, \Sigma_Y, \beta, \Sigma_F)$ *et* $(\Lambda', \Sigma'_Y, \beta', \Sigma'_F)$ *tels que*

$$\Lambda(100 + \beta) = \Lambda'(100 + \beta') \tag{4.1}$$

$$\Lambda\Sigma_F\Lambda^t + \Sigma'_Y = \Lambda'\Sigma'_F\Lambda'^t + \Sigma'_Y \tag{4.2}$$

Cette définition revient à dire que la décomposition de la loi marginale de Y doit être unique, la première équation étant liée à l'espérance de Y, le seconde à sa variance.

Dans la suite, et pour simplifier les notations, le paramétrage est modifié en incluant la constante 100 dans la matrice β des effets et en ajoutant une variable explicative X_0 valant 1 pour tout les individus de telle sorte que le modèle s'écrit :

$$f(Y, F|X) = \mathcal{N}(\Lambda F, \text{diag}(\Sigma_Y))\mathcal{N}(\beta X, \Sigma_F)). \tag{4.3}$$

Avec ce paramétrage, l'identifiabilité s'écrit :

$$\Lambda\beta = \Lambda'\beta' \tag{4.4}$$

$$\Lambda\Sigma_F\Lambda^t + \Sigma'_Y = \Lambda'\Sigma'_F\Lambda'^t + \Sigma'_Y \tag{4.5}$$

En outre, le problème est décomposé. Dans un premier temps, le cas des variables latentes non corrélées est abordé, puis on généralise au cas des variables latentes ayant une structure de dépendance.

4.2 Identifiabilité, cas de variables latentes non corrélées

4.2.1 Notations et hypothèses

Soient X les variables *design* correspondant aux variables explicatives, F les variables latentes et Y les variables à expliquer. Le nombre de variables *design* explicatives est N_X, le nombre de variables latentes est N_F et le nombre d'Y est N_Y. On suppose que $N_Y \geq N_F$ et $N_X \geq N_F$. La matrice de passage de X à F, β, est une matrice de dimension (N_X, N_F) ; la matrice de passage de F à Y, Λ, est une matrice de dimension (N_F, N_Y). On suppose que leur rang est maximal, soit N_F. On ordonne les X et les Y de telle sorte que les N_F premières colonnes de β et les N_F premières lignes des Λ forment des sous-matrices régulières, notées respectivement α et λ . Alors, il existe une matrice m de dimension $(N_F, N_Y - N_F)$ et une matrice n de dimension $(N_X - N_F, N_F)$ telles que

$$\Lambda = \begin{pmatrix} I_{N_F} \\ m \end{pmatrix} \quad \lambda = M\lambda \tag{4.6}$$

$$\beta = \alpha(I \quad | \quad n) = \alpha N \tag{4.7}$$

La matrice de variance des Y conditionnellement aux F est diagonale et notée D. La matrice de variance des F conditionnellement aux X est diagonale dans cette section notée D_F. Comme il a été indiqué au chapitre précédent, il faut fixer une condition d'identifiabilité relative à l'amplitude des fonctions mentales F. Du point de vue probabiliste, il y a deux manières de fixer l'amplitude des F :

1. en fixant les variances des F conditionnellement aux X (égale à 1 par exemple),

2. en fixant une colonne de β, par exemple celle qui correspond à l'intercept des F.

Ces deux manières sont, bien sûr, algébriquement équivalentes puisqu'on passe de l'une à l'autre par une transformation qui ne fait pas intervenir d'autres paramètres. Pour des raisons d'interprétation, c'est la seconde qui a été retenue dans les applications pratiques comme mentionné à la page 31. Mais dans ce chapitre, c'est la première qui est utilisée car elle permet de ne pas imposer de contraintes sur β. On pose donc :

$$\mathbf{var}(F|X) = I \tag{4.8}$$

Il vient

$$\mathbf{E}(Y) = M\lambda\alpha NX \tag{4.9}$$

$$\mathbf{var}(Y) = M\lambda\lambda^t M^t + D \tag{4.10}$$

De façon plus précise :

$$\mathbf{E}(Y) = \begin{pmatrix} \lambda\alpha & \lambda\alpha n \\ m\lambda\alpha & m\lambda\alpha n \end{pmatrix} \tag{4.11}$$

$$\mathbf{var}(Y) = \begin{pmatrix} \lambda\lambda^t + d & \lambda\lambda^t m^t \\ m\lambda\lambda^t & m\lambda\lambda^t m^t + d' \end{pmatrix} \tag{4.12}$$

La matrice m s'interprète de la façon suivante. Si $\tilde{y} = \lambda F$, alors $y = \tilde{y} + \epsilon$ et $y' = m\tilde{y} + \epsilon'$, y étant le N_F premières variables de Y, y' les $N_Y - N_F$ dernières. On peut aussi écrire :

$$\mathbf{E}(y'|y) = my \tag{4.13}$$

$$\mathbf{var}(y'|y) = mdm^t + d' \tag{4.14}$$

On suppose que D, la matrice de variance de Y conditionnellement aux F est diagonale et qu'aucun des éléments diagonaux est nul. Par ailleurs, les contraintes structurelles imposées au modèle (telles que des 0 dans les matrices Λ et β) restreignent l'ensemble des solutions. Ce sont ces contraintes qui vont rendre le système identifiable ou non. Des contraintes en trop faible nombre ou *mal placées* conduiront à un système non identifiable. On peut se placer dans trois types de problèmes :

1. les contraintes structurelles portent uniquement sur Λ et on cherche la structure $X \to F$, c'est à dire qu'il n'y a aucune contrainte sur β,
2. les contraintes structurelles portent uniquement sur β et on cherche la structure $F \to X$, c'est à dire qu'il n'y a aucune contrainte sur Λ,
3. les contraintes structurelles portent à la fois sur β et Λ.

Soient λ_0, m_0, α_0, n_0, d_0 et d'_0 une solution et Λ_0, M, β_0, N_0 et D les matrices définies par (4.6) et (4.7) . Soient λ, m, α, n, d et d' une autre solution et Λ, M, β, N et D les matrices définies par (4.6) et (4.7). On doit avoir :

$$M_0\lambda_0\alpha_0 N_0 X = M\lambda\alpha N X \tag{4.15}$$

$$M_0\lambda_0\lambda_0^t M_0^t + D_0 = M\lambda\lambda^t M^t + D \tag{4.16}$$

De façon plus précise :

$$\lambda_0\alpha_0 = \lambda\alpha \tag{4.17}$$

$$\lambda_0\alpha_0 n_0 = \lambda\alpha n \tag{4.18}$$

$$m_0\lambda_0\alpha_0 = m\lambda\alpha \tag{4.19}$$

$$m_0\lambda_0\alpha_0 n_0 = m\lambda\alpha n \tag{4.20}$$

$$\lambda_0\lambda_0^t + d_0 = \lambda\lambda^t + d \tag{4.21}$$

$$m_0\lambda_0\lambda_0^t = m\lambda\lambda^t \tag{4.22}$$

$$m_0\lambda_0\lambda_0^t m_0^t + d'_0 = m\lambda\lambda^t m^t + d' \tag{4.23}$$

4.2.2 Trois identités remarquables

Les équations (4.17), (4.18) et (4.19) conduisent aux 2 relations :

$$n_0 = n \quad \text{et} \quad m_0 = m. \tag{4.24}$$

Les équations (4.22) et (4.23) conduisent alors à

$$d'_0 = d'. \tag{4.25}$$

Les problèmes d'identifiabilité portent uniquement sur λ, α et d.

Compte tenu des identités précédentes, les conditions d'identifiabilités deviennent :

$$\lambda_0 \alpha_0 = \lambda \alpha \tag{4.26}$$
$$\lambda_0 \lambda_0^t + d_0 = \lambda \lambda^t + d \tag{4.27}$$
$$\lambda_0 \lambda_0^t m_0^t = \lambda \lambda^t m_0^t \tag{4.28}$$

Si le modèle est identifiable en λ (*i.e.* $\lambda_0 = \lambda$), alors le modèle est complètement identifiable ($\alpha_0 = \alpha$ et $d_0 = d$). Il en est de même si le modèle est identifiable en α.

4.2.3 Une condition suffisante pour simplifier le problème d'identifiabilité

Cette condition vise à lever l'identifiabilité sur d_0 et à réduire à deux (au lieu de trois) le nombre d'équations garantissant l'identifiabilité.

Les équations (4.27) et (4.28) conduisent à l'égalité

$$d_0 m_0^t = d m_0^t. \tag{4.29}$$

Comme les variances des Y conditionnellement aux F sont supposées non nulles, cette égalité est équivalente à

- soit l'égalité $d_0 = d$,
- soit la nullité des colonnes de m_0 qui correspondent au valeurs différentes de d_0 et d.

Il s'agit d'une condition nécessaire pour réduire le problème d'identifiabilité :

Proposition 1. *Quelle que soit l'arrangement des Y (pour définir les N_F premières) et quelles que soient les contraintes, le problème est identifiable en m, n, d et l'identifiabilité se réduit à :*

$$\lambda_0 \alpha_0 = \lambda \alpha$$
$$\lambda_0 \lambda_0^t = \lambda \lambda^t$$

si et seulement si *la matrice m ne contient pas de colonne nulle.*

Dans le cas contraire, i.e. *la matrice m contient des colonnes nulles, alors les éléments de d correspondant aux colonnes* **non nulles** *de m sont identifiables et le problème d'identifiabilité se réduit à :*

$$\lambda_0 \alpha_0 = \lambda \alpha \tag{4.33}$$
$$\lambda_0 \lambda_0^t + \delta_0 = \lambda \lambda^t + \delta, \tag{4.34}$$

δ et δ_0 étant la matrice diagonale des variances de Y conditionnellement aux F correspondant aux colonnes de m nulles.

Cette proposition n'est pas évidente à utiliser pour diagnostiquer l'identifiabilité car elle ne repose pas directement sur les paramètres du modèle mais sur la structure de la matrice m. Comme $m_0 = \lambda'_0 \lambda_0^{-1}$ par définition, où λ'_0 est la matrice des $N_Y - N_F$ dernières lignes de Λ_0, une ligne nulle dans m_0 ne peut pas se produire si le rang de λ'_0 est égal à N_F. Mais ce n'est pas une condition nécessaire. D'où les propositions :

Proposition 2. *Si* $rang(\lambda'_0) = N_F$ *alors* $d_0 = d$ *et quelles que soient les contraintes, l'identifiabilité se réduit à :*

$$\lambda_0 \alpha_0 = \lambda \alpha \tag{4.35}$$
$$\lambda_0 \lambda_0^t = \lambda \lambda^t \tag{4.36}$$

Cette proposition peut être utilisée si $N_Y \geq 2N_F$. Si $N_Y < 2N_F$, c'est à dire si le nombre de ligne de λ' est inférieur strictement à N_F, cette proposition n'est pas utilisable ; la seule condition utilisable est alors la nullité d'une colonne de la matrice m, mais, cette condition est difficilement vérifiable sauf dans quelques cas particuliers.

La nullité de la colonne i de m_0 signifie que la i^{e} colonne des λ'_0 est linéairement indépendante de la i^{e} colonne de λ_0 Une telle situation apparaît dans l'exemple suivant : on peut décomposer λ_0 en $\left(\begin{smallmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{smallmatrix}\right)$ et λ'_0 en $\left(\begin{smallmatrix} 0 & \lambda'_2 \end{smallmatrix}\right)$. C'est un cas « pathologique » car pour l'ensemble des premières variables Y qui définissent λ_1, il n'y a pas de réduction de structure puisque λ_1 est régulière : les variables de cet ensemble sont indépendantes des autres et il y a autant de variables latentes sous-jacentes affectées à λ_1 et dans ce cas, il n'y a pas d'identifiabilité.

4.2.4 Cas où les contraintes ne portent que sur Λ

On se place dans le cas où le rang de λ' est maximal (soit N_F). Alors les équations d'identifiabilité sont données par (4.35) (4.36). S'il n'y a aucune contrainte sur β, (c'est à dire α dans (4.35)), il ne reste qu'une seule équation pour l'identifiabilité (4.36) :

$$\lambda_0 \lambda_0^t = \lambda \lambda^t, \tag{4.37}$$

soit $(N_F + 1)N_F/2$ équations pour N_F^2 inconnues potentielles. L'identifiabilité ne sera garantie que par les contraintes. Du point de vue géométrique, cette équation consiste à rechercher N_F vecteurs (les lignes de λ) dans un espace de dimension N_F, dont la matrice des produits scalaires est donnée et suivant les contraintes éventuelles.

Contraintes obligatoires

Pour garantir l'identifiabilité, il faut fixer le signe des coefficients non nuls de λ. Dans le cas contraire, si λ_0 est une solution, alors $\lambda_0 Diag(-1^j)$ est aussi une solution. Cette contrainte est réalisée pour les données étudiées car on suppose une relation croissante entre les fonctions mentales et les variables de tests utilisées (*i.e.* les coefficients de λ tous positifs).

Du point de vu géométrique, cela revient à fixer le cadran dans lequel on cherche les vecteurs lignes de λ et la multiplication par la matrice diagonale $Diag(-1^j)$ revient à changer de cadran.

Contraintes structurelles

Un fois le cadran fixé, il faut d'autres contraintes pour garantir l'identifiabilité. En effet, le système d'équations (4.36) comprend $N_F(N_F + 1)/2$ équations pour N_F^2 inconnues (les paramètres de λ) potentielles. Il faut donc un minimum de $N_F(N_F - 1)/2$ contraintes d'égalité sur les paramètres de λ. Les contraintes d'inégalité autres que les contraintes obligatoires n'ajoutent rien à l'identifiabilité.

On peut distinguer plusieurs types de contraintes d'égalité.

1. la contrainte de nullité : la contrainte de nullité porte directement sur un paramètre.
2. la contrainte d'égalité non nulle : le paramètre est fixé à une valeur donnée. Il semble que ce type de contrainte soit moins favorable que la contrainte de nullité pour l'identifiabilité (voir avec $N_F = 2$).

3. la contrainte linéaire : une combinaison linéaire de paramètre est nulle. On peut distinguer le cas où la contrainte fait intervenir des paramètres relatifs à la même variable y ou si elle fait intervenir des paramètres relatifs à plusieurs variables y. Le cas le plus favorable d'un point de vue de l'identifiabilité est peut-être celui où la contrainte fait intervenir des paramètres relatifs à plusieurs variables y ; cela peut se voir dans le cas où $N_F = 2$. Mais le cas le plus pratique est peut être l'autre cas pour lequel la contrainte linéaire revient à fixer, *a priori*, le poids relatif des variables latentes sur une variable y donnée.

4. la contrainte affine : une combinaison linéaire de paramètres est égale à une valeur prédéfinie non nulle (même distinction que précédemment).

Dans les graphes TOLUÈNE et PAQUID, on n'utilise que des contraintes de signe (pour fixer le cadran) et des contraintes de nullité. Dans ce cas on peut énoncer la proposition :

Proposition 3. *S'il n'y a que des contraintes de nullité de paramètre sur λ et des contraintes de signe, l'équation* (4.36) *est identifiable si et seulement si il y a au moins $(N_F - 1)N_F/2$ contraintes de nullité et tous les paramètres non nuls de λ ont une contrainte de signe (pour fixer le cadran des solutions).*

4.2.5 Cas où les contraintes ne portent que sur β

Dans ce cas, l'identifiabilité de λ découle de celle de β par

$$\lambda = \lambda_0 \beta_0 \beta^{-1} \tag{4.38}$$

Si $\beta = \beta_0$, alors $\lambda = \lambda_0$. L'équation (4.36) s'écrit alors :

$$\alpha^t \alpha = \alpha_0^t \alpha_0 \tag{4.39}$$

ce qui conduit aux mêmes conclusions que dans le cas où les contraintes sont uniquement sur λ :

Proposition 4. *S'il n'y a que des contraintes de nullité de paramètre sur α et des contraintes de signe, l'équation* (4.39) *est identifiable si et seulement si il y a au moins $(N_F - 1)N_F/2$ contraintes de nullité et tous les paramètres non nuls de α ont une contrainte de signe (pour fixer le cadran des solutions).*

4.3 Identifiabilité, cas de variables latentes corrélées

4.3.1 Notations et hypothèses

Les mêmes notations sont conservées pour les paramètres communs avec le cas des variables latentes non corrélées. Pour modéliser la structure de corrélation entre les fonctions mentales, et pour suivre l'approche retenue pour l'analyse des données TOLUÈNE, on écrit :

$$F = \beta X + Q F_0 \tag{4.40}$$

avec

$$F_0 \sim \mathcal{N}(0, I) \tag{4.41}$$

et

$$\mathbf{var}(F|X) = QQ^t. \tag{4.42}$$

La structure de corrélation est représentée par la matrice carrée Q. Dans le cas non corrélé, $Q = I$ ou plus généralement, Q est diagonale. Toujours dans le cas non corrélé, l'indétermination d'amplitude a été levée en imposant à la **var**$(F|X)$ d'être la matrice identité. Dans le cas corrélé, cette indétermination existe et peut être levée de plusieurs façons :

1. en fixant la diagonale de **var**$(F|X)$ (égale à 1 par exemple),
2. en fixant la diagonale de Q (égale à 1 par exemple),
3. en fixant une colonne de β, par exemple celle qui correspond à l'intercept des F.

C'est cette dernière option qui a été retenue pour l'interprétation des résultats. On peut vérifier que, par transformation linéaire, ces 3 façons de procéder sont équivalentes. Il en résulte qu'il suffit d'établir l'identifiabilité en suivant l'une quelconque de ces options. En effet, imposant un paramétrage différent, elles conduisent à des conditions d'identifiabilité spécifiques, mais ces conditions doivent se déduire l'une de l'autre. Dans la suite, bien que dans la pratique la dernière option est retenue, c'est la seconde approche qui est considérée pour établir les conditions d'identifiabilité car elle permet de ne pas imposer de contraintes sur β.

Avec ces notations, il vient :

$$\mathbf{E}(Y) = M\lambda\alpha NX \tag{4.43}$$
$$\mathbf{var}(Y) = M\lambda QQ^t\lambda^t M^t + D \tag{4.44}$$

De façon plus précise :

$$\mathbf{E}(Y) = \begin{pmatrix} \lambda\alpha & \lambda\alpha n \\ m\lambda\alpha & m\lambda\alpha n \end{pmatrix} \tag{4.45}$$
$$\mathbf{var}(Y) = \begin{pmatrix} \lambda QQ^t\lambda^t + d & \lambda QQ^t\lambda^t m^t \\ m\lambda QQ^t\lambda^t & m\lambda QQ^t\lambda^t m^t + d' \end{pmatrix} \tag{4.46}$$

On suppose toujours que D, la matrice de variance de Y conditionnellement aux F est diagonale et qu'aucun des éléments diagonaux est nul. Par ailleurs, les contraintes structurelles imposées au modèle (telles que des 0 dans les matrices Λ et β) restreignent l'ensemble des solutions. Ce sont ces contraintes qui vont rendre le système identifiable ou non. Des contraintes en trop faible nombre ou *mal placées* conduiront à un système non identifiable. Ne sera traité que le cas où il n'y a pas de contraintes sur β et on suppose que les éléments diagonaux de Q sont égaux à 1.

Soient λ_0, m_0, α_0, n_0, Q_0, d_0 et d'_0 une solution et Λ_0, M, β_0, N_0 et D les matrices définies par (4.6) et (4.7) . Soient λ, m, α, n, Q, d et d' une autre solution et Λ, M, β, N et D les matrices définies par (4.6) et (4.7). On doit avoir :

$$M_0\lambda_0\alpha_0 N_0 X = M\lambda\alpha NX \tag{4.47}$$
$$M_0\lambda_0\lambda_0^t M_0^t + D_0 = M\lambda\lambda^t M^t + D \tag{4.48}$$

De façon plus précise :

$$\lambda_0\alpha_0 = \lambda\alpha \tag{4.49}$$
$$\lambda_0\alpha_0 n_0 = \lambda\alpha n \tag{4.50}$$
$$m_0\lambda_0\alpha_0 = m\lambda\alpha \tag{4.51}$$
$$m_0\lambda_0\alpha_0 n_0 = m\lambda\alpha n \tag{4.52}$$
$$\lambda_0 Q_0 Q_0^t\lambda_0^t + d_0 = \lambda QQ^t\lambda^t + d \tag{4.53}$$
$$m_0\lambda_0 Q_0 Q_0^t\lambda_0^t = m\lambda QQ^t\lambda^t \tag{4.54}$$
$$m_0\lambda_0 Q_0 Q_0^t\lambda_0^t m_0^t + d'_0 = m\lambda QQ^t\lambda^t m^t + d' \tag{4.55}$$

4.3.2 Trois identités remarquables

De la même façon que dans le cas non corrélé, on obtient les trois relations suivantes :

$$n_0 = n \tag{4.56}$$
$$m_0 = m \tag{4.57}$$
$$d'_0 = d' \tag{4.58}$$

et le problème d'identifiabilité se réduit à :

$$\lambda_0 \alpha_0 = \lambda \alpha \tag{4.59}$$
$$\lambda_0 Q_0 Q_0^t \lambda_0^t + d_0 = \lambda Q Q^t \lambda^t + d \tag{4.60}$$
$$\lambda_0 Q_0 Q_0^t \lambda_0^t m_0^t = \lambda Q Q^t \lambda^t m_0^t \tag{4.61}$$

La *première condition suffisante* énoncée dans le cas non corrélé l'est aussi dans le cas corrélé. En effet, les relations (4.60) et (4.61) conduisent à la même égalité :

$$d_0 m_0^t = d m_0^t. \tag{4.62}$$

Le suite du raisonnement exposé à la section 4.2.3 conduit cette fois ci à la proposition :

Proposition 5. *Si* $rang(\lambda'_0) = N_F$ *alors* $d_0 = d$ *et quelles que soient les contraintes, l'identifiabilité se réduit à :*

$$\lambda_0 \alpha_0 = \lambda \alpha \tag{4.63}$$
$$\lambda_0 Q_0 Q_0^t \lambda_0^t = \lambda Q Q^t \lambda^t \tag{4.64}$$

Ainsi, si les contraintes imposées à λ'_0 sont telles que cette matrice ne peut pas être de rang N_F (le rang maximale), ce sont les contraintes sur λ_0 d_0 Q_0 et α_0 qui rendent le problème identifiable à l'aide des équations (4.63), et (4.64). Si $N_Y < 2N_F$, c'est à dire si le nombre de ligne de λ' est inférieur strictement à N_F, cette proposition n'est pas utilisable ; La seule condition utilisable est alors la nullité d'une colonne de la matrice m, mais, comme cela a été souligné, cette condition est difficilement vérifiable sauf dans quelques cas particuliers.

4.3.3 Cas où les contraintes ne portent que sur Λ et Q

On se place dans le cas où le rang de λ' est maximal (soit N_F). Alors les équations d'identifiabilité sont données par (4.35) (4.36). S'il n'y a aucune contrainte sur β, (c'est à dire α dans (4.63)), il ne reste qu'une seule équation pour l'identifiabilité (4.64) :

$$\lambda_0 Q_0 Q_0^t \lambda_0^t = \lambda Q Q^t \lambda^t, \tag{4.65}$$

soit $(N_F + 1)N_F/2$ équations. L'identifiabilité d'α découlera de celle de λ et de Q car si $\lambda = \lambda_0$, alors $\alpha = \lambda^{-1}\lambda_0\alpha_0 = \alpha_0$. L'identifiabilité ne sera garantie que par les contraintes sur λ et sur Q. La matrice Q n'a pas de signification particulière, par rapport au sujet étudier. Il ne s'agit que d'un paramétrage pour garantir la structure de dépendance entre les fonctions mentales. En particulier, on peut toujours se ramener à une matrice Q triangulaire. Comme le nombre de fonctions mentales est fixé, les éléments diagonaux de Q doivent être non nuls. Le nombre de paramètres potentiel de Q est alors $N_F(N_F + 1)/2$. Le nombre total d'inconnues potentielles est alors $N^2 + N_F(N_F + 1)/2$ pour $N_F(N_F + 1)/2$ équations.

Du point de vue géométrique, cette équation consiste à rechercher N_F vecteurs (les lignes de λ) dans un espace de dimension N_F, dont la matrice des produits scalaires est donnée, pour le produit scalaire défini par la matrice définie positive QQ^t et suivant les contraintes éventuelles.

Contraintes obligatoires

Pour garantir l'identifiabilité, il faut fixer le signe des coefficients non nuls de λ et de Q. Dans le cas contraire, si (λ_0, Q_0) est une solution, alors $(\lambda_0 Diag(-1^j), Diag(-1^j)Q_0 Diag(-1^k)$ est aussi une solution. Cette contrainte est réalisée pour les données étudiées car on suppose une relation croissante entre les fonctions mentales et les variables de tests utilisées (*i.e.* les coefficients de λ positifs ou nuls) et des covariances ou covariances conditionnelles positives entres fonctions mentales (*i.e.* les coefficients de Q positifs ou nuls). Du point de vu géométrique, cela revient à fixer le cadran dans lequel on cherche les vecteurs lignes de λ et la multiplication par la matrice diagonale $Diag(-1^j)$ revient à changer de cadran.

Il faut aussi fixer la valeur de N_F éléments supplémentaires de Q (par exemple les éléments diagonaux). Dans le cas contraire, si (λ_0, Q_0) est une solution, alors pour toute matrice diagonale $Diag(d_j)$, $(\lambda_0 Diag(d_j), Diag(1/d_j)Q_0)$ est aussi une solution. Le nombre d'inconnues potentielles est donc ramené à $N^2 + N_F(N_F - 1)/2$.

On remarque qu'il est aussi possible de fixer N_F paramètres non nuls de λ au lieu de fixer la diagonale de Q. Ces deux approches sont équivalentes du point de vue de l'identifiabilité. De même, il est aussi possible de fixer N_F paramètres non nuls de β (par exemple les intercepts des fonctions mentales) au lieu de fixer la diagonale de Q. C'est ce qui est fait en pratique comme indiqué plus haut.

Contraintes structurelles

Un fois le cadran fixé et N_F paramètres de Q, il faut d'autres contraintes pour garantir l'identifiabilité. En effet, le système d'équations (4.64) comprend $N_F(N_F+1)/2$ équations pour $N_F^2+N_F(N_F-1)/2$ inconnues potentielles (les paramètres de λ et Q). Il faut donc un minimum de $N_F(N_F - 1)$ contraintes d'égalité sur les paramètres de λ et de Q. Les contraintes d'inégalité autres que les contraintes obligatoires n'ajoutent rien à l'identifiabilité, en particulier, les contraintes d'inégalité sur les paramètres de Q nécessaires pour imposer éventuellement le signe des covariances ou des concentrations. On peut se demander si ces contraintes peuvent être réparties indépendamment sur λ ou sur Q.

Dans les graphes TOLUÈNE et PAQUID, on n'utilise que des contraintes de signe (pour fixer le cadran) et des contraintes de nullité.

Dans ce cas on peut énoncer la proposition :

Proposition 6. *S'il n'y a que des contraintes de nullité de paramètre et des contraintes de signe, l'équation* (4.64) *est identifiable si la somme du nombre de paramètres non nuls de λ et de Q est au plus égale à $(N_F + 1)N_F/2$ et si tous les paramètres non nuls de λ ont une contrainte de signe (pour fixer le cadran des solutions).*

4.3.4 Cas où les contraintes ne portent que sur β et Q

Dans ce cas, l'identifiabilité de λ découle aussi de celle de β par

$$\lambda = \lambda_0 \beta_0 \beta^{-1} \tag{4.66}$$

Si $\beta = \beta_0$, alors $\lambda = \lambda_0$. L'équation (4.64) s'écrit alors :

$$\alpha^t (QQ^t)^{-1} \alpha = \alpha_0^t (Q_0 Q_0^t)^{-1} \alpha_0 \tag{4.67}$$

soit

$$\alpha^t \mathbf{var}(F|X)^{-1} \alpha = \alpha_0^t \mathbf{var}(F|X)_0^{-1} \alpha_0 \tag{4.68}$$

Cette dernière équation montre que dans cette situation, il est préférable d'utiliser un autre paramétrage de $\mathbf{var}(F|X)$. Si on conserve les éléments de $\mathbf{var}(F|X)^{-1}$ comme paramètres, ce qui permet d'utiliser directement le graphe de concentration pour fixer les contraintes, les résultats sont immédiats, on obtient les mêmes conclusions que dans le cas de contraintes uniquement sur λ et Q :

Proposition 7. *S'il n'y a que des contraintes de nullité de paramètre et des contraintes de signe, l'équation* (4.68) *est identifiable si la somme du nombre de paramètres non non nuls de α et de $\mathbf{var}(F|X)^{-1}$ (ou de Q) est au plus égale à $(N_F + 1)N_F/2$ et si tous les paramètres non nuls de α ont une contrainte de signe (pour fixer le cadran des solutions).*

4.3.5 Applications concrètes

Cette condition permet de statuer sur la non identifiabilité dans quelques cas concrets selon la position et le nombre de contraintes structurelles dans la matrice Λ.

1. En ce qui concerne le nombre de contraintes, si celui-ci est insuffisant pour pouvoir dégager des sous-matrices régulière de rang N_F, on ne peut pas statuer *a priori* sur l'identifiabilité du système.

2. Si, pour tout ordonnement des Y, tel que la matrice des N_F premières lignes de Λ est régulière, la matrice des $N_Y - N_F$ premières lignes de Λ contient une colonne de 0, alors, le système n'est pas identifiable.

3. Si, pour tout ordonnancement des Y, tel que la matrice des N_F premières lignes de Λ est régulière, la somme du nombre de paramètres de cette sous-matrice et du nombre de paramètre de Q est strictement supérieur à $(N_F+1)N_F/2$, alors le système n'est pas identifiable.

Comme l'indique la première condition, dans certains cas, les contraintes structurelles ne permettent pas de dégager des sous-matrices de rang N_F de telle sorte qu'on ne peut pas statuer sur l'identifiabilité *a priori*. On n'a pas une identifiabilité structurelle mais il est possible qu'il y ait une identifiabilité numérique au sens où les valeurs numériques obtenues pour une solution vérifient les conditions d'identifiabilité. Cependant, l'impossibilité de choisir une sous-matrice *a priori* régulière signifie souvent qu'il y a trop de paramètres.

4.4 Solutions apportées par les logiciels

Le problème de l'identifiabilité est cruciale pour l'analyse numérique du jeu de données. L'absence d'identifiabilité se traduit par des problèmes numériques qui rendent les résultats incohérents ou indécidables. Si on effectue l'analyse par maximum de vraisemblance (ou toute autre technique conduisant à la minimisation d'une fonction d'objectif telle que les moindres carrés), comme c'est le cas du logiciel EQS, l'absence d'identifiabilité correspond au cas où l'ensemble des solutions pour lesquelles la vraisemblance est maximale n'est pas réduit à un unique élément mais constitue une variété (affine la plupart du temps). Les techniques numériques utilisées pour rechercher ce maximum sont capables de détecter cette situation et certains logiciels le signalent à l'utilisateur par un message spécifique. Les logiciels sont donc capables de détecter l'absence d'identifiabilité, mais ne sont pas réellement capables d'indiquer avec précision quel sous-ensemble de paramètres est concerné.

Si on effectue l'analyse par une technique bayésienne (comme avec BUGS), l'absence d'identifiabilité se traduit par une distribution *a posteriori* des paramètres singulière. Mais les logiciels d'analyse bayésienne par échantillonnage ne permettent pas de détecter facilement ce problème.

4.5 Discussion

Dans ces démonstrations, on constate que les principales conditions contraignantes d'identifiabilité sont indépendantes du nombre de variables de test (Y) ou de variables explicatives (X)[1] mais plutôt du nombre de variables latentes. Les variables latentes jouent un rôle de goulot d'étranglement.

4.6 Conclusion

Le problème de l'établissement de l'identifiabilité n'a pas été abordé pour tous les types de graphes du type $X \rightarrow F \rightarrow Y$ mais seulement pour ceux pour lesquels toute la connaissance *a priori* porte sur l'une ou l'autre des relations $X \rightarrow F$ d'une part, et $F \rightarrow Y$ d'autre part. C'est le cas pour les deux graphes qui ont été analysés et pour lesquels l'identifiabilité a pu être établie. En revanche, l'identifiabilité dans le cas où la connaissance *a priori* est répartie sur les deux parties du graphe $X \rightarrow F$ et $F \rightarrow Y$ n'a pas été abordée.

Dans le cas PAQUID, on montre que le graphe de la Figure 7.1 est identifiable. Dans le cas TOLUÈNE, le premier graphe formulé (Figure 6.1) n'est pas identifiable : il y a 4 variables latentes et on ne peut pas extraire 4 variables de tests tels que la proposition 6 de la page 43 soit vérifiée. Pour remédier à ce problème, une approximation du graphe a été formulée en éliminant des relations supposées négligeables (Figure 6.2).

Il faut bien insister sur le fait que la non-identifiabilité d'un graphe ne remet pas en cause la validité du modèle neuropsychologique. Elle signifie uniquement qu'il est impossible de l'analyser par une approche numérique. Ainsi, dans le cas du graphe initial de TOLUÈNE, si on ne veut pas modifier le graphe des relations entre F et Y, il est possible de résoudre le problème de non-identifiabilité en ajoutant des variables de test supplémentaires, construites de telle sorte que le graphe complété devienne identifiable. Comme il s'agit souvent d'un problème de nombre de paramètres trop élevé, c'est à dire de nombre de relations $F \rightarrow Y$ trop élevé, c'est à dire encore de tests trop peu spécifiques, on voit l'intérêt de construire des tests très spécifiques à l'une ou l'autre des fonctions mentales. En terme pratique, la non-identifiabilité du premier graphe de TOLUÈNE peu s'interprèter de la façon suivante : la batterie utilisée (ou plutôt le jeu de variables utilisées) n'est pas adaptée pour étudier la structure des 4 fonctions mentales telle que précisée dans le graphe.

1. En ce qui concerne le nombre de variables explicatives (X), il faut émettre des réserves. En effet, dans les deux cas d'étude concernés, ce nombre est supérieur au nombre de fonctions mentales. Dans le cas contraire, on obtiendra des conditions plus souples. Mais ce problème n'est pas à l'ordre du jour de ce rapport.

Chapitre 5

Analyse de jeux de données synthétiques

5.1 Introduction

Afin de tester les procédures, une série de tests a été réalisée à l'aide de jeux de données simulées. Les graphes de jeux de données réels n'étant pas encore établis au moment où ces simulations ont été effectuées, les graphes de jeux de données synthétisées sont fictifs et ne correspondent pas à ces cas réels.

La simulation de jeu de données consiste :

- à choisir une structure sous la forme d'un jeu de paramètres $(\beta, \Lambda, \Phi_{Y|F})$ ainsi qu'un ensemble de variables explicatives (X_i) discrètes à 2 modalités définies pour un certain nombre de sujets $(N = 1000)$,

- puis à simuler successivement les valeurs des variables latentes F_i et des variables à expliquer Y_i à l'aide de leur distribution conditionnelle donnée par l'équation (4.3).

Dans une seconde étape, on applique la procédure Bugs exposée ci-dessus sur les valeurs Y_i et X_i pour reconstituer les paramètres $(\tilde{\beta}, \tilde{\Lambda}, \tilde{\Phi}_{Y|F})$ et inférer les valeurs individuelles de variables latentes.

Plusieurs jeux ont été synthétisés. Leurs paramètres et leurs résultats sont indiqués dans la suite.

5.2 Simulation 1

5.2.1 Structures de la simulation

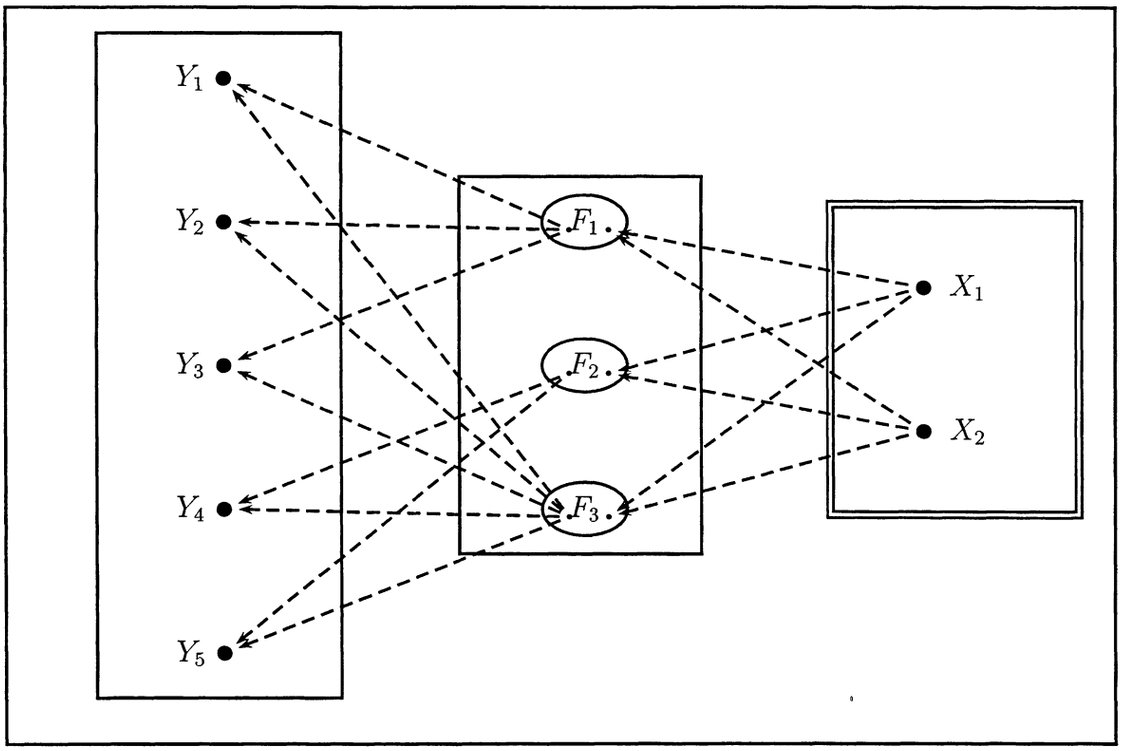
Dans cette simulation, on définit 2 variables d'environnement X, 3 variables latentes F et 5 variables de test Y. Le graphe de dépendance conditionnelle de la simulation est présenté à la Figure 5.1

5.2.2 Paramètres de la simulation

Les paramètres qui définissent la simulation sont :

$$\beta = \begin{pmatrix} 1 & 3 & -5 \\ -10 & 8 & 3 \end{pmatrix} \tag{5.1}$$

FIG. 5.1: *Graphe de dépendance conditionnelle des simulations 1 et 2*



$$Var(F|X) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5.2}$$

$$\lambda = \begin{pmatrix} 1 & 2 & 4 & 0 & 0 \\ 0 & 0 & 0 & 1 & -4 \\ -3 & 2 & 1 & 5 & -2 \end{pmatrix} \tag{5.3}$$

$$Var(Y|F) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 8 \end{pmatrix} \tag{5.4}$$

5.2.3 Résultats obtenus par BUGS pour différents ensembles de valeurs initiales

Les résultats de 3 chaînes sont présentés aux Tableaux 5.1, 5.2, 5.3.

TAB. 5.1: *résultats bugs de la simulation 1, première chaîne (161500 itérations totales, burn in de 136500, une itération sur 10 retenues soit 2500 valeurs.)*

	valeur exacte	valeur initiale	moyenne	écart type	bi	bs	médiane
beta							
beta[1,1]	1	0	0.848	0.152	0.536	1.141	0.848
beta[1,2]	-10	-1	-9.540	0.248	-10.050	-9.082	-9.536
beta[2,1]	3	1	3.117	0.315	2.517	3.683	3.136
beta[2,2]	8	0	8.291	0.517	7.339	9.277	8.299
beta[3,1]	-5	-2	-4.796	0.139	-5.075	-4.537	-4.789
beta[3,2]	3	2	2.924	0.104	2.704	3.129	2.923
lambda							
lambda[1,1]	1	1	1.056	0.034	0.992	1.125	1.054
lambda[1,3]	-3	-1	-3.105	0.090	-3.290	-2.940	-3.108
lambda[2,1]	2	1	2.087	0.055	1.967	2.183	2.091
lambda[2,3]	2	1	2.040	0.086	1.869	2.208	2.044
lambda[3,1]	4	1	4.190	0.102	3.972	4.373	4.197
lambda[3,3]	1	1	0.998	0.131	0.732	1.247	1.001
lambda[4,2]	1	1	0.976	0.060	0.881	1.101	0.967
lambda[4,3]	5	1	5.177	0.144	4.911	5.465	5.180
lambda[5,2]	-4	-1	-3.889	0.245	-4.388	-3.481	-3.863
lambda[5,3]	-2	-1	-2.008	0.167	-2.311	-1.645	-2.016
varyf							
varyf[1]	1	2	0.899	0.175	0.596	1.321	0.889
varyf[2]	2	2	2.482	0.193	2.107	2.867	2.478
varyf[3]	6	2	5.666	0.530	4.619	6.728	5.651
varyf[4]	1	2	1.144	0.335	0.325	1.738	1.152
varyf[5]	8	2	7.427	1.512	4.470	10.190	7.494

TAB. 5.2: *résultats bugs de la simulation 1, seconde chaîne (3600 itérations totales, burn in de 2400, une itération sur 4 retenues soit 3000 valeurs).*

	valeur exacte	valeur initiale	moyenne	écart type	bi	bs	médiane
beta							
beta[1,1]	1	1	0.851	0.175	0.533	1.210	0.840
beta[1,2]	-10	0	-9.569	0.296	-10.110	-8.930	-9.588
beta[2,1]	3	0	3.146	0.261	2.676	3.688	3.135
beta[2,2]	8	1	8.388	0.405	7.596	9.255	8.368
beta[3,1]	-5	-10	-4.843	0.121	-5.090	-4.615	-4.837
beta[3,2]	3	0	2.925	0.125	2.661	3.153	2.930
lambda							
lambda[1,1]	1	1	1.061	0.042	0.981	1.142	1.061
lambda[1,3]	-3	-10	-3.073	0.075	-3.213	-2.916	-3.075
lambda[2,1]	2	5	2.075	0.065	1.961	2.214	2.071
lambda[2,3]	2	1	2.019	0.091	1.851	2.217	2.011
lambda[3,1]	4	1	4.175	0.120	3.964	4.431	4.170
lambda[3,3]	1	6	0.989	0.149	0.723	1.306	0.976
lambda[4,2]	1	1	0.976	0.046	0.894	1.075	0.977
lambda[4,3]	5	1	5.134	0.123	4.906	5.366	5.127
lambda[5,2]	-4	-1	-3.843	0.175	-4.198	-3.496	-3.844
lambda[5,3]	-2	-1	-1.989	0.150	-2.310	-1.706	-1.978
varyf							
varyf[1]	1	2	0.910	0.155	0.620	1.212	0.911
varyf[2]	2	2	2.488	0.185	2.139	2.871	2.480
varyf[3]	6	2	5.679	0.528	4.668	6.734	5.681
varyf[4]	1	4	1.108	0.295	0.543	1.686	1.105
varyf[5]	8	4	7.764	1.261	5.286	10.160	7.753

TAB. 5.3: *résultats bugs de la simulation 1, troisième chaîne (36000 itérations totales, burn in de 24000, une itération sur 4 retenues soit 3000 valeurs).*

	valeur exacte	valeur initiale	moyenne	écart type	bi	bs	médiane
beta							
beta[1,1]	1	1	0.852	0.175	0.536	1.210	0.840
beta[1,2]	-10	-10	-9.569	0.296	-10.110	-8.932	-9.588
beta[2,1]	3	3	3.153	0.262	2.683	3.695	3.140
beta[2,2]	8	8	8.400	0.402	7.603	9.264	8.382
beta[3,1]	-5	-5	-4.843	0.121	-5.090	-4.615	-4.837
beta[3,2]	3	3	2.924	0.124	2.661	3.151	2.930
lambda							
lambda[1,1]	1	1	1.062	0.042	0.981	1.142	1.061
lambda[1,3]	-3	-3	-3.073	0.075	-3.211	-2.916	-3.075
lambda[2,1]	2	2	2.075	0.065	1.961	2.214	2.071
lambda[2,3]	2	2	2.019	0.091	1.852	2.217	2.011
lambda[3,1]	4	4	4.175	0.120	3.964	4.431	4.170
lambda[3,3]	1	1	0.989	0.148	0.726	1.306	0.977
lambda[4,2]	1	1	0.975	0.047	0.892	1.075	0.976
lambda[4,3]	5	5	5.134	0.123	4.906	5.367	5.128
lambda[5,2]	-4	-4	-3.837	0.174	-4.198	-3.493	-3.835
lambda[5,3]	-2	-2	-1.990	0.150	-2.310	-1.706	-1.980
varyf							
varyf[1]	1	1	0.912	0.156	0.620	1.217	0.912
varyf[2]	2	2	2.489	0.185	2.140	2.871	2.481
varyf[3]	6	6	5.677	0.528	4.668	6.729	5.681
varyf[4]	1	1	1.104	0.294	0.542	1.680	1.099
varyf[5]	8	8	7.795	1.253	5.313	10.170	7.794

5.3 Simulation 2

5.3.1 Structures de la simulation

Dans cette simulation, on définit encore 2 variables d'environnement X, 3 variables latentes F et 5 variables de test Y. Le graphe de dépendance conditionnelle de la simulation est le même que celui de la simultation précédente (Figure 5.1).

5.3.2 Paramètres de la simulation

Les paramètres qui définissent la simulation sont :

$$\beta = \begin{pmatrix} 0.25 & 0.75 & -1.25 \\ -2.50 & 2.00 & 0.75 \end{pmatrix} \tag{5.5}$$

$$Var(F|X) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5.6}$$

$$\lambda = \begin{pmatrix} 1 & 2 & 4 & 0 & 0 \\ 0 & 0 & 0 & 1 & -4 \\ -3 & 2 & 1 & 5 & -2 \end{pmatrix} \tag{5.7}$$

$$Var(Y|F) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 8 \end{pmatrix} \tag{5.8}$$

5.3.3 Résultats obtenus par BUGS pour différents ensembles de valeurs initiales

Les résultats de 3 chaînes sont présentés aux Tableaux 5.4, 5.5, 5.6. Il s'agit de 3 chaînes de

TAB. 5.4: Résultats bugs de la simulation 2, chaîne 1 (50000 itérations totales, burn in de 30000, une itération sur 5 retenues soit 4000 valeurs)

	valeur exacte	valeur initiale	mean	sd	bi	bs	median
beta							
beta[1,1]	0.25	1	0.164	0.068	0.036	0.300	0.164
beta[1,2]	-2.5	-10	-2.389	0.091	-2.574	-2.215	-2.387
beta[2,1]	0.75	3	0.844	0.102	0.650	1.056	0.840
beta[2,2]	2.0	8	2.026	0.140	1.761	2.306	2.022
beta[3,1]	-1.25	-5	-1.169	0.059	-1.285	-1.053	-1.169
beta[3,2]	0.75	3	0.778	0.061	0.662	0.900	0.778
lambda							
lambda[1,1]	1	1	1.044	0.045	0.956	1.130	1.044
lambda[1,3]	-3	-3	-3.076	0.085	-3.249	-2.919	-3.072
lambda[2,1]	2	2	2.095	0.067	1.967	2.228	2.096
lambda[2,3]	2	2	2.017	0.093	1.838	2.199	2.016
lambda[3,1]	4	4	4.215	0.121	3.973	4.451	4.217
lambda[3,3]	1	1	0.990	0.152	0.695	1.296	0.989
lambda[4,2]	1	1	0.939	0.058	0.826	1.051	0.941
lambda[4,3]	5	5	5.199	0.125	4.950	5.435	5.202
lambda[5,2]	-4	-4	-3.923	0.242	-4.397	-3.475	-3.922
lambda[5,3]	-2	-2	-2.004	0.164	-2.324	-1.685	-2.002
varyf							
varyf[1]	1	1	1.038	0.214	0.628	1.481	1.030
varyf[2]	2	2	2.549	0.215	2.133	2.969	2.547
varyf[3]	6	6	5.453	0.609	4.283	6.658	5.438
varyf[4]	1	1	0.927	0.410	0.062	1.672	0.950
varyf[5]	8	8	7.118	1.702	3.637	10.260	7.215

TAB. 5.5: résultats bugs de la simulation 2, chaîne 2 (50000 itérations totales, burn in de 30000, une itération sur 5 retenues soit 4000 valeurs)

	valeur exacte	valeur initiale	mean	sd	bi	bs	median
beta							
beta[1,1]	0.25	0	0.164	0.068	0.036	0.300	0.164
beta[1,2]	-2.5	-1	-2.389	0.091	-2.574	-2.215	-2.387
beta[2,1]	0.75	1	0.844	0.102	0.652	1.056	0.841
beta[2,2]	2.0	0	2.027	0.140	1.761	2.306	2.023
beta[3,1]	-1.25	-2	-1.170	0.059	-1.285	-1.053	-1.170
beta[3,2]	0.75	2	0.778	0.061	0.662	0.900	0.778
lambda							
lambda[1,1]	1	1	1.044	0.045	0.956	1.130	1.044
lambda[1,3]	-3	-1	-3.076	0.085	-3.249	-2.919	-3.072
lambda[2,1]	2	1	2.095	0.067	1.967	2.228	2.096
lambda[2,3]	2	1	2.017	0.093	1.838	2.199	2.016
lambda[3,1]	4	1	4.215	0.121	3.973	4.451	4.217
lambda[3,3]	1	1	0.990	0.152	0.696	1.296	0.989
lambda[4,2]	1	1	0.939	0.058	0.826	1.051	0.940
lambda[4,3]	5	1	5.199	0.125	4.950	5.435	5.201
lambda[5,2]	-4	-1	-3.921	0.242	-4.397	-3.475	-3.919
lambda[5,3]	-2	-1	-2.004	0.164	-2.325	-1.685	-2.002
varyf							
varyf[1]	1	2	1.039	0.213	0.633	1.481	1.030
varyf[2]	2	2	2.550	0.215	2.134	2.969	2.547
varyf[3]	6	2	5.451	0.608	4.283	6.654	5.438
varyf[4]	1	2	0.925	0.407	0.062	1.668	0.950
varyf[5]	8	2	7.130	1.703	3.637	10.260	7.235

TAB. 5.6: résultats bugs de la simulation 2, chaîne 3 (50000 itérations totales, burn in de 30000, une itération sur 5 retenues soit 4000 valeurs)

	valeur exacte	valeur initiale	mean	sd	bi	bs	median
beta							
beta[1,1]	0.25	1	0.162	0.068	0.029	0.299	0.163
beta[1,2]	-2.5	0	-2.383	0.087	-2.552	-2.208	-2.384
beta[2,1]	0.75	0	0.830	0.098	0.643	1.023	0.828
beta[2,2]	2.0	1	1.995	0.135	1.740	2.266	1.992
beta[3,1]	-1.25	-10	-1.171	0.059	-1.287	-1.056	-1.171
beta[3,2]	0.75	0	0.784	0.062	0.664	0.905	0.784
lambda							
lambda[1,1]	1	1	1.041	0.046	0.950	1.129	1.041
lambda[1,3]	-3	-10	-3.081	0.081	-3.248	-2.924	-3.081
lambda[2,1]	2	5	2.102	0.063	1.986	2.232	2.100
lambda[2,3]	2	1	2.019	0.091	1.842	2.205	2.018
lambda[3,1]	4	1	4.219	0.116	4.000	4.454	4.213
lambda[3,3]	1	6	0.988	0.148	0.699	1.283	0.988
lambda[4,2]	1	1	0.945	0.057	0.837	1.056	0.944
lambda[4,3]	5	1	5.194	0.124	4.959	5.439	5.189
lambda[5,2]	-4	-1	-3.983	0.236	-4.455	-3.536	-3.974
lambda[5,3]	-2	-1	-2.000	0.167	-2.324	-1.665	-2.001
varyf							
varyf[1]	1	2	1.010	0.181	0.657	1.357	1.009
varyf[2]	2	2	2.529	0.204	2.142	2.941	2.527
varyf[3]	6	2	5.504	0.597	4.378	6.727	5.495
varyf[4]	1	4	1.003	0.340	0.364	1.668	0.997
varyf[5]	8	4	6.722	1.656	3.155	9.756	6.858

5.4 Simulation 3

5.4.1 Structures de la simulation

Dans cette simulation, le graphe générateur de la simulation et le graphe utilisé pour analyser le jeu de données simulé ne sont pas les mêmes. On définit 2 variables d'environnement X, une seule variable latente F et 4 variables de test Y. Le graphe de dépendance conditionnelle générateur de la simulation est présenté à la Figure 5.2 : l'une des variables de test, Y_4, est indépendante de la fonction mentale. Mais, dans le graphe utilisé pour analyser le jeu de données simulé, on fait figurer la relation $F \rightarrow Y_4$: on doit trouver une valeur proche de 0.

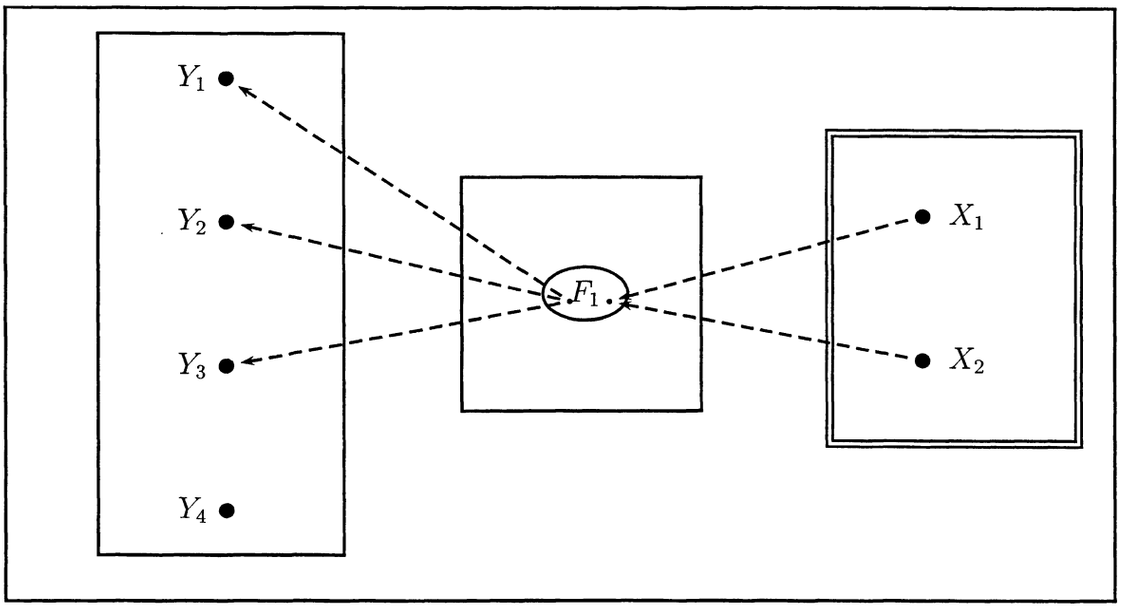
5.4.2 Paramètres de la simulation

Les paramètres qui définissent la simulation sont :

$$\beta = \begin{pmatrix} -1 \\ 3 \end{pmatrix} \tag{5.9}$$

$$Var(F|X) = (1) \tag{5.10}$$

FIG. 5.2: *Graphe de dépendance conditionnelle générateur de la simulation 3*



$$\lambda = \begin{pmatrix} 1 & -2 & 4 & 0 \end{pmatrix} \tag{5.11}$$

$$Var(Y|F) = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.12}$$

5.4.3 Résultats obtenus par BUGS pour différents ensembles de valeurs initiales

Les résultats de 2 chaînes sont présentés aux Tableaux 5.7 et 5.8. La valeur estimée du coefficient $\lambda_{4,1}$ est très proche de 0 dans les 2 chaînes.

5.5 Conclusion

Pour les trois simulations, on constate une grande similitude entre les résultats des trois chaînes et une bonne adéquation des résultats avec les valeurs «vraies» des paramètres des simulations.

Pour chaque simulation, les résultats inter-chaînes sont pratiquement identiques alors qu'il existe une sorte de biais entre les résultats et les paramètres des simulations. Ce biais est sans doute le résultat de la procédure suivie : on n'a utilisé qu'une seule simulation pour chaque jeu de paramètres de sorte que les résultats rendent compte de la structure de cette unique simulation. Il s'agit sans doute d'un biais d'échantillonnage. Pour le vérifier, il faudrait, pour chaque ensemble de paramètres, effectuer plusieurs simulations, et pour chacune de ces simulations, effectuer plusieurs chaînes puis en analyser les résultats. Le fait que les trois chaînes donnent des résultats similaires permet de valider partiellement l'approche MCMC pour laquelle il faut effectuer des vérifications *a posteriori* sur la convergence des chaînes. Plusieurs tests de convergence sont possibles

TAB. 5.7: *résultats bugs de la simulation 3, chaîne 1 (50000 itérations totales, burn in de 20000, une itération sur 10 retenues soit 3000 valeurs)*

	valeur exacte	valeur initiale	mean	sd	bi	bs	median
beta							
beta[1,1]	-1	1	-1.012	0.063	-1.136	-0.889	-1.011
beta[1,2]	3	-10	2.872	0.096	2.688	3.061	2.872
lambda							
lambda[1,1]	1	1	1.029	0.035	0.963	1.097	1.029
lambda[2,1]	-2	-3	-2.066	0.058	-2.181	-1.956	-2.066
lambda[3,1]	4	2	4.149	0.117	3.924	4.384	4.148
lambda[4,1]	0	2	0.020	0.012	0.002	0.048	0.019
varyf							
varyf[1]	2	1	1.931	0.095	1.757	2.122	1.927
varyf[2]	2	1	2.323	0.161	2.021	2.651	2.318
varyf[3]	6	1	5.771	0.564	4.703	6.963	5.758
varyf[4]	1	1	0.946	0.043	0.865	1.032	0.944

TAB. 5.8: *résultats bugs de la simulation 3, chaîne 2 (50000 itérations totales, burn in de 20000, une itération sur 10 retenues soit 3000 valeurs)*

	valeur exacte	valeur initiale	mean	sd	bi	bs	median
beta							
beta[1,1]	0	1	-1.013	0.062	-1.137	-0.894	-1.011
beta[1,2]	0	-10	2.872	0.094	2.693	3.063	2.872
lambda							
lambda[1,1]	10	1	1.029	0.035	0.959	1.097	1.029
lambda[2,1]	-3	-3	-2.066	0.058	-2.179	-1.955	-2.066
lambda[3,1]	1	2	4.150	0.115	3.924	4.379	4.151
lambda[4,1]	0.5	2	0.020	0.012	0.001	0.046	0.019
varyf							
varyf[1]	2	3.33	1.931	0.098	1.750	2.129	1.928
varyf[2]	2	2	2.328	0.167	2.011	2.670	2.324
varyf[3]	6	100	5.782	0.566	4.694	6.910	5.770
varyf[4]	1	100	0.944	0.042	0.864	1.033	0.942

([4, 20]). Dans les simulations des deux jeux de données qui sont analysés, le nombre de paramètres à analyser est assez élevé ce qui rend les tests proposés délicats à réaliser. C'est pourquoi ils n'ont pas été réalisés. Pour s'assurer de la convergence, il a été décidé, dans le cadre de ce premier travail, de toujours faire plusieurs longues chaînes (plusieurs dizaines de milliers d'itérations), de se limiter à une vérification de la convergence au vu de l'historique de chaque chaîne et, pour chaque paramètre, de se limiter à une comparaison de visu des résultats inter-chaînes.

Chapitre 6

Étude toluène

6.1 Présentation de l'étude TOLUÈNE

Les risques neurotoxiques du toluène ont déjà fait l'objet d'une abondante littérature qui établit, chez l'homme, un risque d'atteinte des fonctions cognitives pour une exposition aiguë supérieure à 100 ppm et pour une exposition au long cours, supérieure à 50 ppm. L'objectif de l'étude rapportée ici était d'étudier les effets du toluène sur le SNC pour une exposition au long cours, inférieure à 50 ppm et en dehors d'une exposition aiguë.

L'exposition au toluène a été mesurée aux postes de travail occupés au moment du recueil des données ; elle a fait l'objet de 2 campagnes de collecte de 231 prélèvements d'air ambiant dans les ateliers des 2 imprimeries de l'étude et a mis en évidence des niveaux d'exposition faibles, compris entre 3 et 18 ppm pour l'entreprise I et entre 2 et 27 ppm pour l'entreprise II. La durée de l'exposition a été calculée, pour chaque salarié, d'après les données de sa carrière professionnelle recueillie par 2 sources différentes (auprès du service du personnel et de l'intéressé lui-même) et confrontées entre elles. Pour les 2 entreprises de l'étude, les ateliers avaient été déménagés dans des locaux moins pollués (en 1990 pour l'entreprise I et 1987 pour l'entreprise II). Nous disposions pour 4 des postes de l'ancienne unité de l'entreprise II de mesures montrant une exposition élevée (>100 ppm). Pour estimer le niveau passé d'exposition dans les autres ateliers (ceux de la même usine et ceux de l'usine I) nous avons extrapolé les niveaux à partir des coefficients de proportionnalité entre les ateliers, obtenus d'après les données des campagnes actuelles. Nous disposions ainsi de niveaux passés d'exposition (mesurés ou estimés).

Les effets sur le système nerveux qui ont été pris en compte sont :

- les atteintes des fonctions cognitives (boucle perceptivo-motrice, mémoire, apprentissage, concentration-attention) mesurées à l'aide des tests psycho-comportementaux de la batterie NES que nous détaillerons dans le § suivant,

- les symptômes neurologiques et psychologiques allégués par les salariés recueillis au moyen d'un auto-questionnaire de fréquence des symptômes : EUROQUEST

Vingt-quatre variables de contrôle ont été prises en compte. Certaines sont liées :

- à des caractéristiques individuelles des salariés : sexe, âge, niveau scolaire, etc.

- à leur mode de vie : consommation de produits ayant des effets connus sur les SNC (alcool, médicaments psycho-actifs, etc.)

- à des antécédents médicaux ayant ou pouvant avoir des effets sur le SNC,

- à leur état le jour du test (fatigue, quantité de sommeil, etc...).

Dans le but de ne mesurer que les effets chroniques et non aigus du toluène, questionnaire et tests psycho-comportementaux se sont déroulés après au moins 48 heures de soustraction à l'exposition au toluène, ceci ayant été contrôlé, en cas de doute, par la mesure du toluène dans l'air expiré.

Une mise au point méthodologique a porté sur la construction de variables pertinentes pour les tests, les symptômes et les facteurs de contrôle. Les critères retenus pour élaborer des nouvelles variables de tests ont été définis dans le chapitre d'introduction. Les variables des tests et de contrôle retenues pour étudier le graphe de dépendance conditionnelle sont décrites dans le paragraphe suivant.

La première analyse statistique a testé, grâce à une régression linéaire univariée, 2 hypothèses :

- un effet cumulatif et irréversible du toluène sur le SNC en intégrant sous la forme d'un index cumulé, la durée et l'intensité de l'exposition tout au long de la carrière professionnelle,
- un effet réversible en prenant en compte le niveau d'exposition actuelle par atelier.

Les résultats n'ont pas confirmé la première hypothèse. Par contre, on a observé, après ajustement sur les facteurs de confusion, une diminution des performances mnésiques et d'apprentissage quand l'exposition actuelle au poste de travail augmente, les relations étant statistiquement significatives pour 2 variables de performance d'un test de mémoire à court terme (voir tableau 6.1).

La deuxième analyse est une ACP dont les résultats sont détaillés dans le rapport [9]. Bien que les tableaux de valeurs propres soient très comparables entre les 2 entreprises puisque les 4 premiers axes expliquent 69% de la variance pour l'entreprise I et 65% pour l'entreprise II, la comparaison, entre les 2 usines, des coefficients de corrélation de Pearson des variables des tests sur ces 4 premiers axes, n'est pas satisfaisante ; en effet, nous ne retrouvons pas des axes stables toujours définis par les mêmes variables.

Nous n'avons pas retrouvé de relation entre l'index cumulé d'exposition et les axes ACP. Par contre, les fonctions mesurées par les deux premiers axes ACP se dégradent quand le niveau actuel de toluène augmente, se rapprochant du seuil de significativité :

- l'axe 1 ($p1 = 0,08$ pour la 1ère entreprise et $p2 = 0,39$ pour la 2ème),
- l'axe 2 ($p1 = 0,10$ pour la 1ère entreprise et $p2 = 0,24$ pour la 2ème).

Pour les axes 3 et 4, il n'y a pas d'effet de l'exposition.

L'interprétation des axes ACP est difficile compte tenu de l'instabilité des 2 analyses : l'axe 1 semble, comme souvent, l'axe des bonnes performances générales aux tests. L'axe 2 est défini par les variables de stabilité. L'axe 3 oppose les réponses aux tests justes mais apportées avec lenteur aux réponses rapides mais plus souvent inexactes. L'axe 4 oppose les variables de la mémoire à court terme et de la mémoire à long terme. Néanmoins malgré ces imperfections, cette analyse a confirmé des liaisons fortes entre variables. Certaines de ces relations étaient connues, d'autres étaient liées à la spécificité de ce jeu de données. Elle suggérait également un effet de l'exposition actuelle sur certaines fonctions mentales.

6.2 Données utilisées pour le modèle à variables latentes

6.2.1 La population de l'étude

Compte tenu du faible effectif des femmes et de l'effet « sexe » sur les performances des tests, l'analyse avec ce modèle n'a concerné que la population masculine de l'étude ($N = 114$).

6.2.2 Les variables explicatives X

Deux types de variables sont à considérer, celles qui font l'objet de l'hypothèse de l'étude (l'exposition au toluène) et les variables individuelles ayant un effet sur les fonctions mentales.

L'exposition au toluène

La variable qui est liée à l'hypothèse de l'étude est l'exposition professionnelle au toluène ; compte tenu des résultats de l'analyse univariée, nous avons retenu dans le modèle, l'exposition **actuelle** au poste de travail. Les valeurs par poste de travail et entreprise sont rapportées dans le tableau 6.2. La moyenne est à 15 ppm et l'écart type 9.8 ppm. L'effet recherché est une diminution des performances pour les 4 fonctions mentales quand l'exposition augmente.

Les variables individuelles

Parmi les 24 variables de contrôle recueillies, nous avons retenu pour le modèle, celles qui étaient des facteurs de confusion dans le modèle de régression linéaire et/ou celles dont les effets sur les fonctions mentales sont bien documentées dans la littérature.

L'âge : dans le groupe d'étude, la moyenne d'âge est de 38,5 ans (écart-type = 9 ans). Compte tenu de la tranche d'âge de notre population, nous avons considéré que l'effet « âge » qui agit en dégradant les performances, ne pouvait être sensible que pour 3 fonctions mentales : apprentissage, attention, vitesse de réaction.

Le niveau scolaire : parmi les 3 variables candidates pour le décrire, nous avons retenu l'âge de fin d'études car il est lié à presque toutes les variables des tests. L'âge moyen de fin d'études est de 16 ans, compris entre 11 et 24 ans, l'écart-type est de 2,3 ans. Pour des raisons numériques, cette variable a été transformée en variable discrète à 3 modalités :

modalité 1 : de 11 à 14 ans

modalité 2 : de 15 à 17 ans

modalité 3 : de 18 à 24 ans

Le niveau scolaire est, d'après la littérature, lié aux 4 fonctions mentales : plus le niveau scolaire augmente, meilleures sont les performances.

La consommation d'alcool ou de produits médicamenteux ayant un effet sur le SNC : nous avons combiné 2 variables initialement séparées. La consommation d'alcool avait été codée d'après l'auto-déclaration des salariés selon 4 modalités :

modalité 1 : consommation quotidienne

modalité 2 : consommation occasionnelle (de temps en temps)

modalité 3 : absence de consommation

modalité 4 : non réponse (mais il n'y en a pas dans cette étude)

Compte tenu du faible pourcentage de buveurs quotidiens (13% contre 55% attendus), nous avons contourné une sous-déclaration éventuelle en réunissant tous les buveurs dans une seule classe. La consommation d'alcool est devenue une variable dichotomique (oui/non). Par ailleurs, la prise de médicaments au

moment de l'étude et pouvant avoir un effet sur le SNC avait été codée de façon dichotomique (oui/non).

La variable finale est la somme logique de la consommation d'alcool et/ou de médicaments psycho-actifs. 89% du groupe d'étude consomment l'un ou l'autre de ces produits. Les effets attendus sont une diminution des performances pour les 4 fonctions mentales.

Les antécédents médicaux pouvant ou ayant un effet sur les fonctions mentales : ils ont été codés en oui/non. 23% des 114 salariés avaient au moins un antécédent potentiellement délétère pour le SNC. Il s'agissait le plus souvent de traumatismes crâniens. Les effets attendus sont la diminution des performances pour les 4 fonctions mentales.

6.2.3 Les fonctions mentales

Six des 7 tests retenus dans cette étude explorent, selon le concepteur de la batterie, 4 fonctions mentales :

- la boucle perceptivo-motrice, par le test du temps de réaction simple (simple reaction time)
- la concentration-attention par le test du codage de symboles (symbol digit test),
- l'apprentissage par le test d'apprentissage d'associations de 2 mots (associate learning test)
- la mémoire :
  - à court terme par 2 tests :
    - ◇ empan discontinu antérograde et rétrograde (digit span forwards et backwards)
    - ◇ mémoire des formes (pattern memory test)
  - à moyen terme par le test de rappel des associations (learning recall ou associate learning delayed recognition test)

6.2.4 Les relations entre fonctions mentales

- Toutes les performances des fonctions mentales sont dépendantes de la capacité à se concentrer.
- A capacité de concentration constante, les performances d'apprentissage sont directement dépendantes des performances mnésiques à court terme.

6.2.5 Les tests et leurs variables Y

Les tests utilisés dans cette étude et les variables que nous avons construites sont les suivants :

Simple Reaction Time Test

Ce test mesure le temps de réaction du sujet à un message sensoriel élémentaire. A l'apparition aléatoire d'un carré rouge au centre de l'écran, le sujet tape sur une touche de joystick le plus rapidement possible. Il y a 6 séries de 10 stimulus ; si les temps de réponse sont trop longs plusieurs fois de suite, le test s'arrête. Deux variables sont dans le modèle :

- une variable de performance brute : RT_PERF ou simple reaction time perf (Y_{11}),

– une variable de stabilité de la performance: RT_STAB ou simple reaction time stab (Y_{10}) .

Mode de calcul :

– *Lecture des 60 temps (tableau* $RD(1) - RD(60)$*)*

– *Epuration dans ce tableau des valeurs aberrantes (réf. : manuels NES2 4.25 et 4.63) si* $[RD(i) < 100]$ *ou* $[RD(i) > 1000]$ *ou* $[RD(i) > moyenne + 3\sigma]$ $RD(i)$ *est remplacé par la valeur manquante*

– *Calcul des variables de performance et de stabilité pures :*

- *tri du tableau* RD *par valeur croissante,*
- *(*Y_{11}*) :* $1/Log_{10}(quantile10\%(RD))$
- *(*Y_{10}*) :* $1/Log_{10}(100 * (quantile40\% - quantile10\%)/quantile10\%)$

Symbol Digit Test

Emprunté à l'échelle clinique de mémoire de Wechsler (Wechsler Memory Scale), il en est une version informatisée ; il est très sensible aux NT [67] et à toute atteinte cérébrale [68]. En haut de l'écran apparaît une grille avec 9 symboles appariés à 9 chiffres. Dans une deuxième grille, en bas de l'écran, le sujet doit taper, sur le pavé numérique, le plus rapidement possible, les chiffres correspondant aux symboles selon le modèle de la grille supérieure. Il y a 5 grilles de 9 cases à remplir. La première grille est considérée comme un essai. L'appariement symbole-chiffre est aléatoire et nouveau à chaque grille. Deux variables sont dans le modèle :

– une variable de performance brute : SD_PERF ou symbol digit perf (Y_9),

– une variable de stabilité de la performance: SD_STAB ou symbol stab (Y_8).

Mode de calcul :

– *Lecture des 4 séries de temps : tableau* $SD(ij) i = 1,4 (n^\circ série) j = 1,8 (n^\circ symbole)$

– *Elimination des sujets n'ayant pas au moins deux séries justes*

– *Calcul du temps moyen pour chaque série (réf. : manuel NES 4.25 et 4.63)*

$$SD\_MOY = \frac{\sum_j |SD(ij)|}{8 - nombred'erreursdanslasérie i}$$

– *tri des SD_MOY par ordre croissant*

– *(*Y_9*) =* $1/\sqrt{(SD\_MOY1 + SD\_MOY2)/2}$

– *(*Y_8*) =* $1/\sqrt{SD\_MOY2 - SD\_MOY1}$

Pattern Memory Test

C'est un test de mémoire à court terme de reconnaissance de figures géométriques, équivalent informatisé du test de Benton. Une figure géométrique complexe apparaît quelques secondes sur l'écran. Le sujet doit la reconnaître parmi les 3 qui lui sont ensuite proposées. Le nombre d'essais est de 25. Deux variables sont dans le modèle :

- une variable de performance brute : la performance à ce test (qui est le nombre de réponses justes) n'est pas retenue car elle manque de sensibilité (le nombre moyen de figures reconnues est 20, il est trop proche du maximum, 24) alors que le temps mis à trouver la figure quand la réponse est exacte est mieux distribué : PM_TEMPS ou *patt.mem.temp.moy.juste* (Y_6),
- une variable de stabilité de la performance : PM_STAB ou *patt.mem.stab* (Y_7)

Mode de calcul :

- *Lecture des 24 temps (on élimine le premier) : tableau PM(24)*
- *Calcul (réf. : manuels NES2 4.25 et 4.63) de :*
- *(Y_6)= inverse du temps moyen pour une forme juste (PM(i) >0)*
- *(Y_7) = inverse de l'écart type du PM_TEMPS*

Test du Vocabulaire

Le test du vocabulaire est un test de synonymes. On présente un mot cible pour lequel le sujet doit trouver un synonyme parmi 5 proposés. Ce test est théoriquement utilisé pour vérifier la comparabilité socio-culturelle des groupes exposés/non-exposés ; mais il est, en fait, lié à toutes les autres performances des tests et dans une autre étude nous l'avons même trouvé corrélé aux performances de motricité pure. Il semble traduire plus qu'un niveau d'études, une aptitude globale aux apprentissages scolaires relevant peut être d'une compliance à l'acquisition scolaire ou à l'épreuve scolaire, celle qui est en jeu dans la passation des tests. Nous avons reconsidéré cette performance de vocabulaire et son statut est passé de variable X à celui de Y. La variable utilisée est une performance brute au test : NB_REP ou performance vocabulaire (Y_5).

Mode de calcul :

- *Lecture des 34 réponses : tableau REP(34)*
- *(Y_5) = nombre de réponses justes*

Digit Span Forwards et Backwards

C'est un test de mémoire à court terme de rappel de chiffres. Le sujet doit immédiatement taper sur le pavé numérique, les chiffres qui viennent de défiler à l'écran, dans leur ordre d'apparition à l'écran dans la première partie de tests puis dans l'ordre inverse de leur apparition dans la deuxième partie. Le nombre de chiffres proposés augmente tant que le sujet réussit. Au bout de deux échecs successifs pour une série avec même longueur, le test s'arrête. Les 2 variables sont les performances maximales dans l'ordre et dans l'ordre inverse : DS_FORW ou digit span forwards (Y_3) et DS_BACK ou digit span backwards (Y_3).

Mode de calcul :

- *(Y_4) = nombre maximal de chiffres retenus dans l'ordre de leur apparition,*
- *(Y_3) = nombre maximal de chiffres retenus dans l'ordre inverse de leur apparition.*

Associate Learning Test

C'est un équivalent du test d'apprentissage des paires de la WMS, 7 couples de mots du type «*Eric est médecin*» sont présentés une première fois ; le sujet doit ensuite répondre «*Eric est .....*» en choisissant parmi les 7 possibilités. Quels que soient ses résultats, le test est présenté 3 fois de suite. La variable est la meilleure performance pondérée par le rang de la performance maximale : AI_SCORE ou associate learning perf (Y_2).

Mode de calcul :

- *Lecture des 3 résultats* AL_1, AL_2, AL_3,
- *Calcul du score combinant performance maximale et rang du maximum :*

$$Y_2 = 3\max(AL1, AL2, AL3) + [3 - rang(MAX(AL1, AL2, AL3))].$$

Learning Recall ou Associate Learning Delayed Recognition Test

C'est un test de mémoire à moyen terme. Une demi-heure après l'apprentissage des mots couplés du test précédent, le sujet doit rappeler le plus grand nombre de paires. La variable est la performance brute au test : LR_BRUT ou *associate recall perf* (Y_1)

Mode de calcul :

- *(Y_1) = score brut du nombre de couples de mots restitués*

6.2.6 Les relations entre les fonctions mentales et les variables Y du graphe 1

Le premier graphe élaboré est reporté dans la Figure 6.1. Il décrit la structure suivante.

- La boucle perceptivo-motrice influence les 3 variables qui nécessitent la mise en jeu de la perception pour repérer une information et d'un acte moteur pour la réponse ; ce sont les 3 variables qui intègrent un temps de réaction : Y_{11}, Y_9 et Y_6. Plus la fonction mentale est performante plus les temps de réaction sont courts et plus les variables Y_{11}, Y_9 et Y_6 sont élevées.

- L'apprentissage rend compte des variables du test spécifique d'apprentissage mais aussi de la performance au vocabulaire. Plus la fonction mentale est bonne plus les performances sont hautes.

- La mémoire influence les performances des tests spécifiques de mémoire (Associate Recall, Digit Span, Pattern Memory) mais il y a également une composante mémoire dans le test du symbol digit dans sa version informatisée ([30]) et dans la performance au vocabulaire. Plus la fonction mentale est bonne, plus les performances (Associate Recall, Digit Span) sont élevées, plus les temps de réaction (Pattern Memory et Symbol Digit) sont courts et plus les variables Y_9 et Y_6 sont élevées.

- L'attention influence toutes les variables : plus la capacité attentionnelle est bonne plus les variables de performance (d'apprentissage ou de mémoire) sont élevées, plus les variables correspondant aux temps de réaction sont élevées et plus les variables de stabilité sont grandes.

6.2.7 Les relations entre les fonctions mentales et les variables Y du graphe 2

Or comme nous l'avons déjà indiqué, le modèle statistique correspondant à ce premier graphe n'est pas identifiable. Afin de poursuivre notre démarche, nous avons proposé un second graphe (Figure 6.2). Il contient un nombre plus faible de relations entre les fonctions mentales et les variables de tests. Le second graphe doit être considéré comme une *approximation* du premier : les flèches enlevées sont celles que l'on suppose être négligeables. Les modifications sont les suivantes :

- Deux fonctions mentales sont inchangées : boucle perceptivo-motrice et apprentissage
- Pour la mémoire, la flèche en relation avec la performance au vocabulaire a été enlevée car les autres variables des tests de mémoire mesuraient soit la mémoire à court terme soit la mémoire à moyen terme mais pas la mémoire à long terme qui est en jeu dans le test des synonymes.
- Pour l'attention, 3 flèches ont été supprimées soit parce que les autres fonctions sont plus massivement impliquées dans le résultat de la variable (performance Vocabulaire, *patt.mem.-temps.moy.juste*) ou que les méthodes « d'épuration » de ces variables mises en œuvre se sont attachées à dépister et à supprimer les valeurs aberrantes liées à un manque d'attention (Simple Reaction Time perf).

TAB. 6.1: *Means of test scores adjusted according to the current exposure levels in ppm.*

Current Exposure (in ppm)		No exposure	<5	<10	<15	<20	<25	≥ 25	Slope/100 ppm (p)
	plant A	N = 9	N = 22	N = 4	N = -	N = 4	N = -	N = -	
	plant B	N = -	N = 14	N = 14	N = 6	N = -	N = 32	N = 23	
Age	mean	43	38	41	46	43	36	37	
	SD	(9)	(10)	(9)	(10)	(9)	(7)	(8)	
Age of the end	mean	17	17	17	18	15	16	16	
of schooling	SD	(2)	(2)	(3)	(2)	(3)	(2)	(3)	
Synonym Score		23	24	21	29	23	20	20	
Reaction Time	plant A	2.28	2.27	2.27	-	2.26	-	-	+0 .01 (0.93)
Performance	plant B	-	2.34	2.29	2.31	-	2.32	2.31	-0.05 (0.41)
(in milliseconds)									
Reaction Time	plant A	2.26	2.39	2.56	-	2.30	-	-	+0.17 (0.89)
Variability Index	plant B	-	2.41	2.42	2.37	-	2.35	2.40	-0.17 (0.67)
Digit Symbol		2.33	2.80	2.64	2.53	3.30	2.71	2.84	-0.26 (0.09)
Latency (in seconds)									
Pattern Memory		19.7	20.9	20.9	22.2	20	19.7	20.2	-1.58 (0.46)
Performance									
Digit Span Forwards		7.44	6.19	6.06	7.17	5.75	5.84	5.52	-2.66 (0.02)
Performance									
Digit Span Backwards		6.33	5.89	5.17	6.67	6.00	4.91	4.57	-4.09 (0.00)
Performance									
Associate Learning Score		15.7	16.6	15.3	18.5	14.2	14.2	13.9	-6.12 (0.17)
Learning Recall		4.44	4.78	4.28	5.17	3.75	3.81	3.78	-2.09 (0.27)
Performance									

TAB. 6.2: *Exposure measurements and estimation*

Factory A					
Workshop	1	2	3	4	5
Mean measurements[a] (N)	4.6 (50)	16.3 (12)	4.0 (9)	4.8 (9)	3.5 (6)
Estimate current exposure[b]	4.6	18.1	3.2	5.0	4.0
Estimate former exposure[c]	30	118	21	33	26

Factory B									
Workshop	1	2	3	4	5	6	7	8	9
Mean 1996 exposure [1] (N)	2.0 (10)	16.8 (5)	5.6 (16)	27.6 (32)	23.4 (19)	28.6 (17)	20.7 (22)	24.5 (9)	20.9 (15)
Mean 1981 exposure [1] (N)	- (0)	- (0)	174 (6)	91 (6)	110 (8)	79 (4)	- (0)	- (0)	- (0)
Estimated 1996 exposure [2]	2.2	14.3	8.3	27.4	22.3	25.7	23.0	20.6	23.6
Estimated 1981 exposure [3]	15	94	54	179	146	168	150	135	155

[a] arithmetic mean of n measurements

[b] mean exposure assuming a constant geometric standard deviation across all workshops

[c] mean exposure assuming a constant geometric standard deviation and a constant ratio between present and former exposure.

FIG. 6.1: *Premier graphe*

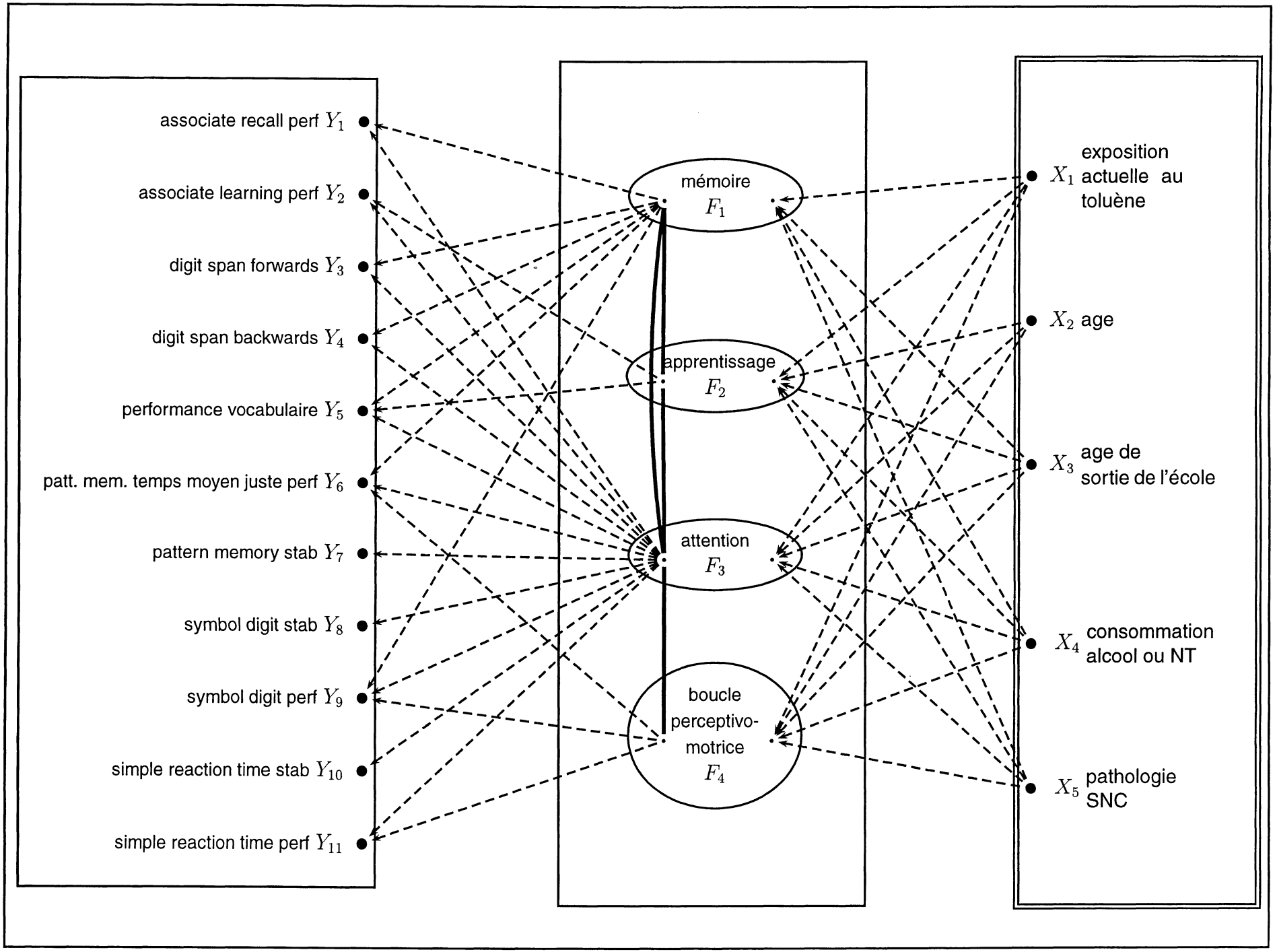
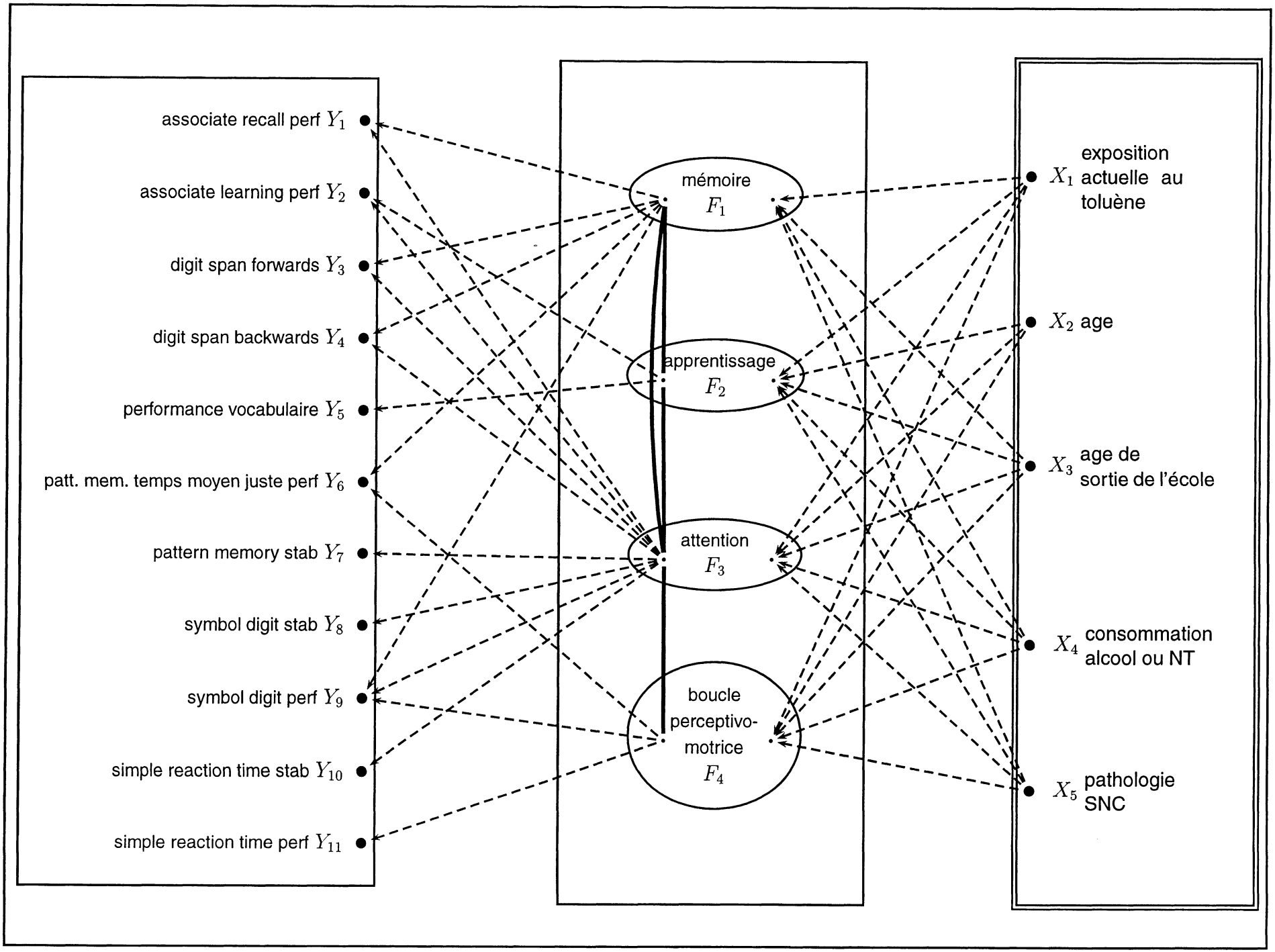


FIG. 6.2: *Second graphe*



6.3 Modélisation de la structure de dépendance entre les fonctions mentales

Dans le graphe retenu, les fonctions mentales sont dépendantes les unes des autres selon une structure bien particulière : il y a 2 «zéros» dans la matrice de concentration des variables latentes. Pour tenir compte de cette contrainte dans les logiciels, nous sommes passés par l'intermédiaire de quatre nouvelles variables aléatoires gaussiennes, (F_1', F_2', F_3', F_4'), de variance égale à 1 et mutuellement indépendantes telles que :

$$\begin{pmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{pmatrix} = Q \begin{pmatrix} F_1' \\ F_2' \\ F_3' \\ F_4' \end{pmatrix} \tag{6.1}$$

où Q est une matrice. Avec Q de la forme :

$$Q = \begin{pmatrix} c_1 & c_2 & c_3 & 0 \\ 0 & c_4 & c_5 & 0 \\ 0 & 0 & c_6 & 0 \\ 0 & 0 & c_7 & c_8 \end{pmatrix}, \tag{6.2}$$

la matrice de concentration de (F_1, F_2, F_3, F_3) à la forme voulue. En outre, si on souhaite que les corrélations partielles des fonctions mentales soient positives, il faut imposer les contraintes suivantes :

$$\frac{c_4 c_3}{2c_5} \leq c_1 \tag{6.3}$$

$$0 \leq c_2 \leq \frac{c_4 c_3}{c_5} \tag{6.4}$$

$$\frac{c_5 c_2}{c_4} \leq c_3 \leq \frac{2c_5 c_1}{c_4} \tag{6.5}$$

$$\frac{c_5 c_2}{c_3} \leq c_4 \leq \frac{2c_5 c_1}{c_3} \tag{6.6}$$

$$\frac{c_4 c_3}{2c_1} \leq c_5 \leq \frac{c_4 c_3}{c_2} \tag{6.7}$$

6.4 Résultats numériques du graphe identifiable

6.4.1 Modèle statistique

Le paramétrage est fait conformément à l'équation (3.12). La matrice Σ_Y est diagonale. Les options suivantes ont été suivies :

- les variables ordinales sont traitées comme des variables continues,
- le niveau moyen des fonctions mentales de la référence est fixé à 100.

La référence choisie est définie ainsi :

- exposition actuelle : nulle,
- âge de référence : 20 ans,
- âge de fin d'étude : compris entre 11 et 14 ans,

- absence de consommation d'alcool ou de médicaments psycho-actifs,
- absence d'antécédents médicaux ou de pathologies actuelles ayant ou pouvant avoir un effet sur les fonctions mentales.

6.4.2 Commentaires sur les chaînes

Trois chaînes ont été calculées. Comme le nombre de paramètres est élevé (63) et qu'il est nécessaire d'effectuer des chaînes relativement longues (au minimum 30000 itérations), la synthèse numérique commune des trois chaînes n'a pas été réalisée, en particulier en ce qui concerne l'évaluation quantitative de la convergence et la stabilité inter-chaîne et intra-chaîne.

D'un point de vue qualitatif, l'examen de l'historique de chaque chaîne permet de conclure :

- chaque chaîne paraît stabilisée, ce qui indique leur convergence,
- d'une chaîne à l'autre, les résultats se recoupent : pour chaque paramètre, les distributions échantillonnées par chaque chaîne sont similaires.

Des moyens numériques plus importants permettraient de confirmer ces observations par un analyse quantitative de la convergence et la stabilité inter-chaîne et intra-chaîne.

Comme les chaînes sont équivalentes, les résultats présentés dans la suite sont issus d'une seule chaîne. Il s'agit d'une chaîne de 30000 itérations dont seulement une itération sur 10 parmi les 25000 dernières sont utilisées dans les calculs (2500 itérations retenues).

6.4.3 Résultats d'une chaîne

Les résultats sont présentés dans les Tableaux 6.3, 6.4, 6.5, 6.6, 6.7 et 6.8 . Ces tableaux présentent les valeurs des paramètres qui rendent compte des hypothèses causales des modèles et des paramètres desciptifs de la population :

- les paramètres de passage des variables latentes aux variables à expliquer (paramètres Λ au Tableau 6.3) qui traduisent l'effet des fonctions mentales sur les résultats aux tests,
- les paramètres de passage des variables explicatives aux variables latentes (paramètres β au Tableau 6.5), qui indiquent les effets des variables explicatives sur les fonctions mentales,
- les corrélations partielles des fonctions mentales au Tableau 6.8,

Par ailleurs, la matrice Q qui sert à construire les variations des fonctions mentales est présentée au Tableau 6.7. Enfin, deux tableaux (6.4 et 6.6) donnent des statistiques descriptives des variables de test et des variables latentes. Dans le Tableau 6.6, l'écart-type de chaque variable latente conditionnellement aux variables explicatives est obtenu en prenant la racine carrée de la somme des carrés de ces coefficients dans la matrice Q.

Dans chaque tableau, on indique la valeur moyenne des paramètres ainsi qu'un intervalle de confiance à 95 %. Quand un paramètre est imposé et nul (correspondant à l'absence d'un lien dans le graphe), la valeur indiquée est «0» et il n'y a pas d'intervalle de confiance.

6.5 Interprétation des résultats

6.5.1 Analyse de la matrice β

Elle montre que l'exposition actuelle au toluène est liée à une diminution des variables latentes représentant les deux fonctions mentales *mémoire* et *apprentissage*. Il faut noter l'importance de

TAB. 6.3: *Matrice Λ des coefficients de régression des variables de test par les fonctions mentales et écarts types résiduels des variables de test.*

	mémoire	apprentissage	attention	boucle perceptivo-motrice
Assoc. recall perf.	1.2e-2 6.2e-4 ; 3.3e-2	0	2.9e-2 8.3e-3 ; 4.2e-2	0
assoc. learning perf	0	6.6e-2 1.3e-2 ; 1.2e-1	6.4e-2 7.1e-3 ; 1.3e-1	0
digit span forwards	5.1e-2 3.6e-2 ; 6.4e-2	0	1.3e-2 1.1e-3 ; 2.7e-2	0
digit span backwards	5.5e-2 4.5e-2 ; 6.5e-2	0	2.7e-3 8.2e-5 ; 9.2e-3	0
perf. vocabulaire	0	1.7e-1 1.2e-1 ; 2.2e-1	0	0
pattern memory temps moyen juste	4.3e-4 3.4e-5 ; 1.0e-3	0	0	3.9e-3 3.2e-3 ; 4.4e-3
pattern memory stab.	0	0	5.2e-3 4.4e-3 ; 6.1e-3	0
symbol digit stab.	0	0	2.1e-2 1.8e-2 ; 2.5e-2	0
symbol digit perf.	7.4e-4 6.7e-5 ; 1.7e-3	0	2.8e-3 1.5e-3 ; 3.8e-3	3.1e-4 7.8e-6 ; 1.1e-3
simple reaction time, stab.	0	0	3.8e-3 3.3e-3 ; 4.4e-3	0
simple reaction time, perf.	0	0	0	7.0e-3 6.5e-3 ; 7.5e-3

TAB. 6.4:

	moyenne	écart type	écart type résiduel
Assoc. recall perf.	4.12	*1.9*	*1.9e+0* 1.6e+0 ; 2.2e+0
assoc. learning perf	14.98	*4.79*	*4.6e+0* 3.9e0 ; 5.0e+0
digit span forwards	6.06	**1.37**	**6.8e-1** 2.6e-1 ; 9.3e-1
digit span backwards	5.26	**1.70**	**9.9e-1** 7.1e-1 ; 1.2e+0
perf. vocabulaire	21.98	**5.57**	**4.7e+0** 3.9e+0 ; 5.6e+0
pattern memory temps moyen juste	0.43	*0.053*	*5.2e-2* 4.5e-2 ; 5.9e-02
pattern memory stab.	0.54	**0.25**	**5.5e-2** 4.2e-2 ; 7.2e-2
symbol digit stab.	2.19	*1.05*	*9.9e-1* 8.7e-1 ; 1.1e+0
symbol digit perf.	0.39	**0.082**	**6.6e-2** 5.6e-2 ; 7.7e-2
simple reaction time, stab.	0.40	*0.062*	*6.3e-2* 5.5e-2 ; 7.4e-2
simple reaction time, perf.	0.70	*0.049*	*4.8e-2* 4.0e-2 ; 5.7e-2

TAB. 6.5: *Matrice β des effets des variables explicatives sur les fonctions mentales*

	mémoire	apprentissage	attention	boucle perceptivo-motrice
exposition actuelle	-8.2e-1 -1.2e+0 ; -4.3e-1	-7.5e-1 -1.2e+0 ; -2.3e-1	7.2e-2 -2.0e-1 ; 4.0e-1	-2.2e-2 -1.5e-1 ; 1.1e-1
age	0	8.0e-1 2.4e-2 ; 2.0e+0	-3.1e-1 -6.5e-1 ; 9.0e-2	-8.8e-3 -1.8e-1 ; 1.8e-1
age de sortie de l'école (3 modalité) (15-17)/(11-14)	1.2e+1 2.1e+0 ; 2.3e+1	2.5e+1 8.0e+0 ; 4.9e+1	4.4e+0 -3.1e+0 ; 1.2e+1	8.7e-1 -2.5e+0 ; 4.5e+0
(18-24)/(11-14)	1.6e+1 5.1e+0 ; 2.9e+1	3.1e+1 1.2e+1 ; 6.0e+1	6.9e+0 -1.4e+0 ; 1.7e+1	2.5e-1 -3.6e+0 ; 4.3e+0
consommation alcool ou NT	-5.9e+0 -1.9e+1 ; 6.5e+0	7.7e+0 -8.1e+0 ; 2.9e+1	7.6e+0 -1.0e+0 ; 1.7e+1	-1.6e-1 -4.1e+0 ; 3.7e+0
pathologies cérébrales	-4.4e+0 -1.4e+1 ; 5.8e+0	1.6e+0 -1.2e+1 ; 1.5e+1	-8.0e+0 -1.5e+1 ; -1.0e+0	1.2e+0 -2.3e+0 ; 4.5e+0

TAB. 6.6: *Statistiques descriptives des fonctions mentales*

	mémoire	apprentissage	attention	boucle perceptivo-motrice
moyenne	92	131	104	100
écart type expliqué	10.7	14.5	6.8	0.6
écart type résiduel	17.3	8.5	6.3	2.0

TAB. 6.7: *Matrice Q de construction des fonctions mentales.*

	F_1'	F_2'	F_3'	F_4'	écart type résiduel
mémoire	1.4e+1 6.0e+0 ; 1.9e+1	4.5e+0 1.2e-1 ; 1.3e+1	9.2e+0 3.4e+0 ; 1.6e+1	0	17.3
apprentissage	0	5.2e+0 9.0e-1 ; 1.0e+1	6.7e+0 1.6e+0 ; 1.2e+1	0	8.5
attention	0	0	6.3e+0 2.8e+0 ; 9.5e+0	0	6.3
boucle perceptivo-motrice	0	0	1.4e+0 1.3e-1 ;3.0e+0	1.4e+0 5.1e-2 ; 3.5e+0	2.0

TAB. 6.8: *Matrice des corrélations partielles des fonctions mentales*

mémoire	3.1e-1 8.0e-3 ; 8.0e-1	1.3e-1 2.1e-3 ; 5.1e-1	0
	apprentissage	5.6e-1 3.1e-2 ; 9.9e-1	0
		attention	5.2e-1 2.7e-2 ; 1.0e+0
			boucle perceptivo-motrice

l'effet de l'exposition actuelle : pour 10 ppm de toluène, la performance de la variable latente *mémoire* se dégrade de 8.2 % [4,3% ; 12%] et celle de la variable latente *apprentissage* de 7.5 % [2,3% ; 12%]. Par ailleurs, on retrouve un effet des facteurs potentiels de confusion, cohérent avec les hypothèses : plus l'âge de sortie de l'école est précoce, plus les variables latentes des fonctions *mémoire* et *apprentissage* sont faibles. Les valeurs de la variable latente *mémoire* sont supérieures de 12 % [2,1% ; 23%] chez les salariés sortis de l'école à 15-17 ans et de 16 % [5,1% ; 29%] chez les salariés sortis de l'école à 18-24 ans, par rapport à ceux sortis à 11-14 ans. La même observation concerne la variable latente *apprentissage* : si on a fini l'école entre 15-17 ans, les valeurs sont supérieures de 25 % [8,0% ; 49 %] et de 31 % [12% ; 60%], en cas d'arrêt de l'école entre 18-24 ans. Ces deux résultats peuvent être assimilés à une relation dose-effet : plus on sort tard de l'école, meilleures sont les compétences mnésiques et d'apprentissage. L'existence, dans les antécédents des sujets, d'un problème ayant affecté le SNC (comme un traumatisme crânien avec perte de connaissance), est comme attendu, associé à une diminution de la variable latente représentant la fonction *attentionnelle* : l'existence d'une pathologie diminue de 8% [1% ; 15%] les valeurs de la variable latente).

L'effet de l'âge n'est significatif que sur la fonction *apprentissage*. Le vieillissement d'une année est lié à une augmentation de 0.8 % [0,024% ; 2%] de la variable latente représentant la fonction *apprentissage*.

6.5.2 Analyse de la matrice Λ

Les valeurs **0** correspondent à l'absence de relation entre la fonction mentale et la variable du test, structurellement défini a priori. Comme par construction, les relations entre fonctions mentales et les variables de test sont positives, l'intervalle de confiance ne peut contenir 0 ; ce tableau ne permet pas de tester la nullité d'un paramètre, c'est à dire l'absence d'une relation présupposée. [1]

On ne peut pas, pour une variable latente donnée, comparer la contribution à chaque variable de test car les variables de test ont des échelles différentes. Pour chaque variable de test, on peut comparer la contribution respective de chaque variable latente.

Les résultats attendus sont :

- Pour les variables du *Digit Span Forwards* et *Backwards*, la part très dominante de la variable latente *mémoire* par rapport à celle de la variable latente *attention* (respectivement 0,051 et 0,056 contre 0,013 et 0,003).

1. La nullité d'un paramètre pourrait être testée en comparant la qualité d'estimation de 2 modèles, l'un avec et l'autre sans la flèche. La comparaison peut se faire à l'aide d'un test de vraisemblance par exemple. Cette approche peut aussi être suivie pour tester les contraintes, pourvu que le graphe augmenté de la flèche à tester soit identifiable. Cet aspect n'a pas pu être traité dans le cadre de ce premier travail.

- Pour la performance au *Symbol Digit*, la répartition des contributions des trois fonctions mentales : 0,0028 (environ 70%) pour l'*attention*, 0,00074 (environ 20%) pour la *mémoire* et 0,00034 (environ 9%) pour la *boucle perceptivo-motrice*.

- Dans la performance du test *Associate Learning*, l'équivalence des contributions des fonctions mentales *apprentissage* et *attention* (0,066 et 0,064).

Les résultats moins attendus concernent :

- La performance du test *Associate Recall* qui représente une performance de mémoire à moyen terme mais qui est davantage liée à la variable latente *attention* qu'à la variable latente *mémoire* (0,029 contre 0,012). Cela peut s'expliquer par le fait que la variable latente *mémoire* regroupe deux fonctions mentales conceptuellement séparées (mémoire à court terme et mémoire à moyen terme) que l'on n'a pas pu distinguer dans le modèle (à cause du faible nombre de variables censées explorer la mémoire à moyen terme).

- Pour le temps moyen juste du test *Pattern Memory*, la contribution de la variable latente *boucle perceptivo-motrice* est dix fois plus importante que celle de la variable latente *mémoire* (0,0039 contre 0,00043) alors que le test est défini comme un test de mémoire. Ceci peut s'expliquer par le fait que la tâche de mémoire s'applique à la reconnaissance visuelle d'une forme géométrique pour laquelle la perception est fortement impliquée.

6.5.3 Analyse du tableau 6.4

Ce tableau permet d'apprécier la part de variance de chaque variable de test expliquée par le modèle en comparant l'écart-type de la population à l'écart-type résiduel. Pour une même variable, une faible différence entre les 2 écart-types signifie que la variance de cette variable est peu expliquée par les quatre fonctions mentales. Ceci est le cas pour six d'entre elles (en italique dans le tableau). Les cinq autres (en gras dans le tableau) présentent au contraire une différence importante, ce qui traduit que leur variance est mieux expliquée par les variations des fonctions mentales.

6.5.4 Analyse de la matrice Q de construction des variables latentes

Elle permet de calculer l'écart-type de chaque variable latente qui n'est pas prise en compte par les variables explicatives. Ces valeurs correspondent, aux coefficients de Λ près, à la part de variabilité des variables de test qui est expliquée par le modèle sans l'être par les variables explicatives. On peut comparer cet écart-type par rapport à la valeur *100* attribuée arbitrairement à l'individu de référence. On constate que c'est la variable latente *mémoire* qui est la plus variable. On constate aussi que l'écart-type de la variable latente *boucle perceptivo-motrice* est très faible.

6.5.5 Analyse de la matrice des corrélations partielles des variables latentes

Comme précédemment, par construction, les corrélations partielles entre deux variables latentes sont positives. Donc il n'y a pas non plus de test de significativité dans le tableau. Les valeurs *0* correspondent également à des *0* structurels. Les intervalles de confiance sont larges, surtout pour la corrélation partielle entre les variables latentes *attention* et *apprentissage* d'une part, et entre les variables latentes *attention* et *boucle perceptivo-motrice* d'autre part.

Ceci traduit que les données ne contiennent pas d'information sur les corrélations entre les fonctions mentales.

6.5.6 Analyse descriptive des variables latentes

Pour les trois fonctions mentales *mémoire*, *apprentissage* et *attention*, les variances expliquées et résiduelles sont d'un même ordre de grandeur et représentent environ 10% de la valeur de référence (qui est égale à 100). Pour la fonction mentale *boucle perceptivo-motrice*, la moyenne est pratiquement égale à la valeur de référence et les écart-types sont faibles : la variable latente de la *boucle perceptivo-motrice* est quasiment constante dans la population étudiée.

6.6 Discussion des résultats

Les résultats de la matrice β sont cohérents avec les hypothèses. Ils confirment et précisent ceux de l'analyse univariée, mettant en évidence un effet de l'exposition actuelle sur des variables latentes représentant la mémoire et l'apprentissage.

L'effet de l'âge est néanmoins inattendu sur la fonction apprentissage. On s'attendait à une dégradation de cette variable latente avec l'âge. Ceci est peut-être dû à la construction de la variable latente *apprentissage*, à partir de deux tests dont le *test du vocabulaire*, dont on sait qu'il s'améliore avec l'âge. Le *test du vocabulaire* teste le résultat d'un apprentissage ancien, tandis que l'autre test qui participe à la définition de cette variable latente, teste l'apprentissage actuel. La définition de cette variable latente est sûrement à revoir. Nous avons, par construction, éliminé l'effet de l'âge sur la mémoire, sur des arguments théoriques liés à la tranche d'âge limitée de la population (dont l'écart-type n'est que de 9 ans). Ceci est discutable car l'analyse univariée mettait en évidence un effet *âge*. L'effet de l'existence d'une pathologie du SNC n'est significative que pour la variable latente *attention*. Ceci tient probablement à la nature de la pathologie codée *oui* qui comportait essentiellement des traumatismes crâniens avec perte de connaissance. La consommation *alcool et/ou médicaments psycho-actifs* n'a pas d'effet sur les quatre variables latentes. Cette variable est probablement de mauvaise qualité car on a regroupé des buveurs occasionnels d'alcool dont certaines performances cognitives pourraient être améliorées par l'alcool d'après [17], des buveurs plus réguliers et les consommateurs de médicaments dont les performances cognitives peuvent être plutôt dégradées.

6.7 Conclusion

Pour remédier au problème d'identifiabilité, la solution envisagée a été de simplifier le graphe. Cela n'est pas entièrement satisfaisant pour les raisons déjà évoquées. Une manière plus satisfaisant envisageable pour la suite consiste à ajouter des variables de test (ou des tests) supplémentaires qui soient plus spécifiques et en particulier qui ne soient pas liées à l'attention.

Cette première analyse apporte néanmoins des éléments de réponse sur le rôle de l'exposition au toluène sur les fonctions mentales pour un graphe donné (en l'occurrence le graphe 2).

Le graphe 2 utilisé est sûrement perfectible et différentes améliorations ont été annoncées dans le paragraphe précédent.

Chapitre 7

Étude PAQUID

7.1 Présentation de l'étude PAQUID

La détection des sujets à risque de développer une démence est un enjeu majeur de santé publique, en particulier pour réaliser des essais thérapeutiques visant à mettre au point chez ces sujets un traitement préventif de l'évolution vers un stade démentiel.

Seules les études longitudinales peuvent permettre de déterminer les outils cognitifs nécessaires à cette détection. Dans ces études, les performances cognitives de sujets considérés comme normaux à un temps donné peuvent en effet être mises en relation avec la survenue d'une démence au cours du suivi ultérieur.

Une analyse de ce type avait été effectuée sur les données de la cohorte Paquid. Elle avait été effectuée sur les 1142 sujets normaux de cette étude qui avaient passé tous les tests neuropsychologiques au suivi T1 de cette enquête et qui étaient devenus déments au suivi T3, deux ans après.

Les tests neuropsychologiques analysés au suivi T1 étaient :

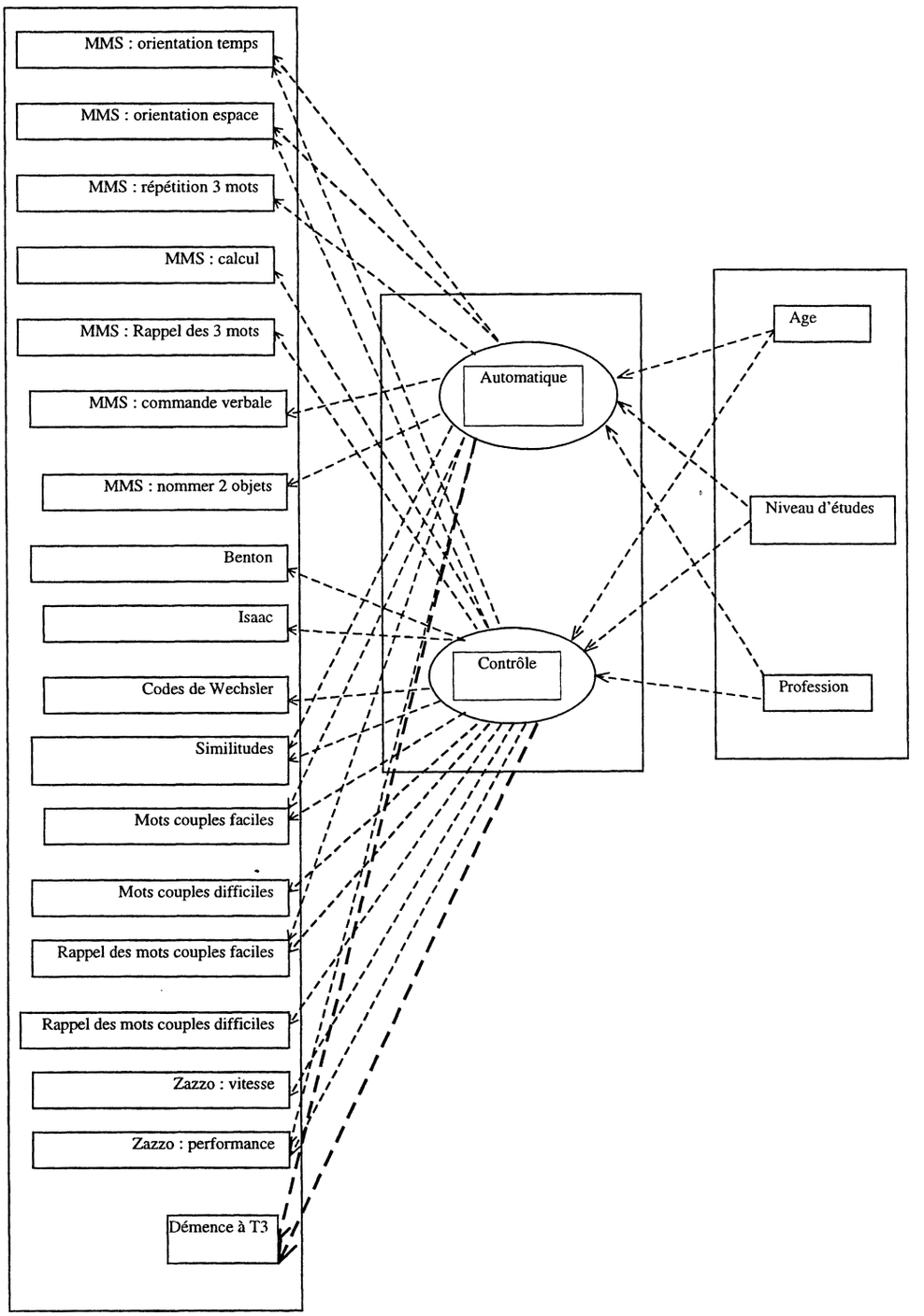
- Une échelle composite d'évaluation des fonctions cognitives: le MMS (Mini-Mental-State Evaluation) de Folstein.

- Des tests psychométriques plus spécifiques. Le test de rétention visuelle de Benton explore la mémoire visuelle à court terme et possède également une composante attentionelle importante. La mémoire verbale associative est évaluée par le test des mots couplés de Wechsler. Le set test d'Isaac évalue la préservation des répertoires sémantiques par les capacités du sujet à générer une liste de mots appartenant à une catégorie spécifique : couleurs, animaux, fruits, villes ; il nécessite également une recherche consciente et active en mémoire et une bonne gestion de la mémoire de travail. Les troubles de l'attention visuo-spatiale sont mis en évidence par le test des barrages de Zazzo. Les perturbations du raisonnement et de l'attention sont étudiés par le test des codes de Wechsler. La conceptualisation est évaluée par le test des similitudes de la WAIS.

Une analyse en composantes principales avait été réalisée en prenant en compte l'ensemble des résultats aux tests du suivi à un an de la cohorte. Le poids des individus sur chacun de ces facteurs a par la suite été utilisé dans une analyse permettant de déterminer le risque de démence deux ans plus tard. Seul le facteur 1 de l'ACP s'est révélé associé à un risque de future démence. L'analyse des tests les plus caractéristiques du facteur 1 tendait à montrer que ce facteur représentait les composantes attentionelles ou contrôlées des différents tests. Ce résultat tendait une hypothèse de la littérature selon laquelle la préservation des processus automatiques par rapport à la détérioration des processus contrôlés représentait l'un des premiers signes d'un processus démentiel.

C'est cette hypothèse que nous avons voulu tester dans le cadre de ce travail.

FIG. 7.1: *Graphe de* PAQUID



7.2 Fonctions mentales et variables utilisées

7.2.1 Les fonctions mentales

Le graphe proposé (Figure 7.1) comporte la composante contrôlée des tests cognitifs (contrôle) et la composante automatique de ces tests (automatique).

7.2.2 Les variables explicatives (X)

L'age, le niveau d'études et profession principale exercée ont une influence sur les deux fonctions mentales, bien qu'elles agissent plus fortement sur le contrôle que sur l'automatisme. Le niveau d'études est codée en 2 modalités selon que le sujet possède ou non un diplôme. La profession principale exercée est codée en 3 modalités:

1. ouvrier agricole,
2. manuel,
3. non manuel.

7.2.3 Les variables des test et leurs relations avec les fonctions mentales

Les flèches reliant les variables des tests aux composantes *contrôle* et *automatisme* reflètent notre conception a priori de la charge des différents tests en composantes attentionelles ou au contraire en composantes plus automatiques.

Le MMS est un test composite comportant un certain nombre d'items évaluant des processus cognitifs différents. Pour notre analyse nous avons pris en compte séparément différentes parties du MMS.

MMS orientation temps

Le sujet doit répondre à cinq questions concernant l'orientation dans le temps (date, jour de la semaine, mois, saison, année). Score de 0 à 5. Cette variable comporte à la fois des composantes automatiques et des composantes contrôlées.

MMS orientation espace

Le sujet doit répondre à cinq questions concernant l'orientation dans l'espace (lieu, étage, ville, département, pays). Score de 0 à 5. Cette variable comporte à la fois des composantes automatiques et des composantes contrôlées.

MMS répétition des trois mots

Le sujet doit répéter 3 mots énoncés par l'expérimentateur. Score de 0 à 3. Cette variable comporte essentiellement des composantes automatiques. Dans cette cohorte, tous les sujets ont obtenu le score de 3 et ce test n'est pas utilisé.

MMS calcul

Le sujet doit effectuer successivement cinq soustractions (soustraire le chiffre 7 du chiffre 100 et recommencer à partir du reste quatre fois). Score de 0 à 5. La soustraction nécessite une grande charge attentionelle et cette variable n'est reliée qu'aux processus contrôlés.

MMS rappel des trois mots

Le sujet doit rappeler les trois mots qu'il avait répétés avant la soustraction. Il s'agir en terme neuropsychologique de "mémoire à long terme", puisque au cours de la soustraction, le sujet a dû vider sa mémoire de travail des trois mots et il doit donc aller les rechercher activement en mémoire. Score de 0 à 3. Cette variable comporte essentiellement des composantes automatiques.

MMS commande verbale

Le sujet doit effectuer trois commandes verbales (prenez cette feuille de papier, pliez la en deux et poser la par terre). Score de 0 à 3. Cette variable comporte essentiellement des composantes automatiques.

MMS nommer deux objets

Le sujet doit donner le nom de deux objets (une montre et un crayon). Score de 0 à 2. Cette variable comporte essentiellement des composantes automatiques. Dans cette cohorte, tous les sujets ont obtenu le score de 2 et ce test n'est pas utilisé.

Benton

Dans le test de rétention visuelle de Benton le sujet doit regarder pendant 10 secondes une figure géométrique, puis il doit la reconnaître parmi quatre. Ce test explore la mémoire visuelle à court terme et possède également une composante attentionelle importante. Score de 0 à 15. Cette variable est essentiellement reliée au contrôle à cause de sa charge attentionelle.

Isaac

Dans le set test d'Isaac, le sujet doit générer en une minute les plus possible de mots appartenant à une catégorie sémantique, et ceci successivement pour quatre catégories (couleur, animaux, fruits et villes). Ce test évalue la préservation des répertoires sémantiques ; il nécessite également une recherche consciente et active en mémoire et une bonne gestion de la mémoire de travail. Score non plafonné. Cette variable est essentiellement reliée au contrôle à cause de sa charge attentionelle.

Codes de Wechsler

Dans ce test qui fait partie de l'échelle d'intelligence de Wechsler le sujet doit utiliser une échelle de correspondance entre chiffres et symboles pour reporter les symboles corrects en dessous de lignes de chiffres. Ce test suppose un raisonnement logique puisque le sujet doit comprendre la correspondance entre chiffre et symboles, mais il présente aussi une composante attentionelle importante. Score de 0 à 93. Cette variable est essentiellement reliée au contrôle à cause de sa charge attentionelle.

Similitudes

Dans le test des similitudes, le sujet doit dire en quoi deux éléments se ressemblent, ce qu'ils ont en commun (par exemple un chien et un lion). Les sujets ont deux points s'ils donnent la catégorie (animaux), 1 point s'ils donnent un caractère commun (peau) et 0 s'ils donnent la différence. Nous avons utilisé une version raccourcie aux cinq premières questions. Ce test mesure

les capacités de conceptualisation du sujet. Score de 0 à 10. Cette variable comporte à la fois des composantes automatiques et des composantes contrôlées.

Le test des mots couplés fait partie de l'échelle de mémoire de Wechsler. Le sujet écoute une liste de 10 paires de deux mots (par ex. école-épicerie) et il doit ensuite rappeler le deuxième mot de chaque paire quand on lui énonce le premier mot de la paire. Le test comporte 3 essais d'apprentissage et un rappel différé après un test interférant. Nous avons distingué les six paires de mots comportant une association facile parce que les mots sont reliés sémantiquement (par exmple rose-fleur) des quatre paires de mots comportant une association difficile parce que les mots ne sont pas reliés sémantiquement (par exemple choux-plume). Quatre variables caractérisent les scores à ce test.

Mots couplés faciles

Il s'agit du score réalisé aux paires faciles lors du premier essai d'apprentissage. Scores de 0 à 6. Cette variable comporte à la fois des composantes automatiques et des composantes contrôlées.

Mots couplés difficiles

Il s'agit du score réalisé aux paires difficiles lors du premier essai d'apprentissage . Score de 1 à 4. Cette variable est essentiellement reliée au contrôle.

Rappel des mots couplés faciles

Il s'agit du score réalisé lors du rappel différé des paires faciles. Score de 0 à 6. Cette variable comporte à la fois des composantes automatiques et des composantes contrôlées.

Rappel des mots couplés difficiles

Il s'agit du score réalisé lors du rappel des paires difficiles. Score de 0 à 4. Cette variable est essentiellement reliée au contrôle.

Le test de Barrage de Zazzo mesure l'attention sélective visuo-spatiale par la capacité du sujet à sélectionner et à barrer une cible parmi des distracteurs. Ce test est chronométré et nous avons distingué la vitesse de réalisation du test et le nombre de bonnes cibles barrées. Deux variables caractérisent les performances à ce test :

Zazzo vitesse

Il s'agit de la vitesse de réalisation du test. Score non plafonné. La vitesse à ce test est supposé refléter les capacités d'inhibition du sujet et cette variable est donc seulement reliée aux processus de contrôle.

Zazzo performance

Il s'agit du nombre de bonnes cibles barrées sur les 29. Score de 0 à 29. Cette variable de performance présente à la fois des aspects automatiques et des aspects contrôlés.

7.3 Résultats numériques du graphe 1 de PAQUID

7.3.1 Modèle statistique

Le paramétrage est fait conformément à l'équation (3.12). Les matrices Σ_F et Σ_Y sont diagonale. Le niveau moyen des fonctions mentales de la référence est fixé à 100. La référence choisie est définie ainsi :

- âge de référence : l'âge moyen de la population analysée (*74,3 ans*),
- niveau scolaire : *sans diplôme*,
- niveau professionnel : *ouvrier agricole*.

Ces choix constituent l'analyse de référence. Comme le nombre de sujets est assez élevé, d'autres analyses ont été effectuées en faisant d'autres choix afin de vérifier leur équivalence et d'évaluer leurs avantages ou inconvénients. Ainsi, trois[1] analyses sont présentées :

1. l'analyse de référence, faite avec BUGS,
2. la même analyse, faite avec EQS,
3. une analyse en modifiant le paramétrage de telle sorte que le niveau moyen des fonctions mentales de la référence ne soit plus fixé à 100, mais que la variance des fonctions mentales soit fixée à la valeur 1,

En ce qui concerne la *démence à T3*, qui est une variable ordinale, un lien logistique a été utilisé pour modéliser l'influence des fonctions mentales dans les trois analyses faites avec BUGS (1, 3, 4). Comme EQS ne permet pas d'analyser des modèles où sont présents simultanément des variables ordinales à expliquer et des variables explicatives, la *démence à T3* est codée en $0-1$ dans la seconde analyse et est traitée comme une variable continue.

7.3.2 Commentaires sur les chaînes

Comme le nombre de paramètres est élevé (46 au minimum, plus selon les analyses) et qu'il est nécessaire d'effectuer des chaînes relativement longues (au minimum 30000 itérations), la synthèse numérique commune des cinq chaînes n'a pas été réalisée, en particulier en ce qui concerne l'évaluation quantitative de la convergence et la stabilité inter-chaîne et intra-chaîne.

D'un point de vue qualitatif, l'examen de l'historique de chaque chaîne permet de conclure :

- chaque chaîne paraît stabilisée, ce qui indique leur convergence,
- d'une chaîne à l'autre, les résultats se recoupent : pour chaque paramètre, les distributions échantillonnées par chaque chaîne sont similaires.

Des moyens numériques plus importants permettraient de confirmer ces observations par un analyse quantitative de la convergence et la stabilité inter-chaîne et intra-chaîne. Comme les chaînes sont équivalentes, les résultats présentés dans la suite sont issus d'une seule chaîne.

1. Une analyse en modélisant les variables ordinales par une approche du type *proportionnal odds* a été tentée mais elle n'a pas abouti suite à la survenue de problèmes numériques. On peut noter toutefois que, comme le montre l'utilisation d'un lien logistique pour la démence, les variables (ordinales) à 2 modalités pourraient être modélisées par une régression logistique.

TAB. 7.1: *Matrice β*

	automatique	contrôle
age	1.414e-2 -1.450e-2 4.194e-2	-1.613e+0 -1.815e+0 -1.416e+0
niveau scolaire	2.011e-1 -4.598e-2 4.802e-1	1.612e+1 1.333e+1 1.893e+1
profession en 3 niveaux	 -3.604e-1 -8.553e-1 8.051e-2 -3.945e-1 -8.889e-1 1.088e-1	 2.045e+0 -3.596e+0 7.839e+0 9.020e+0 3.013e+0 1.507e+1

TAB. 7.2: *Écarts types des fonctions mentales*

	moyenne	intervalle de confiance à 95%
automatique	0.21	0.060 ; 0.506
contrôle	14.9	13.8 ; 16.1

7.3.3 Solution 1 : l'analyse *de référence*

Il s'agit d'une chaîne de 73000 itérations dont seulement une itération sur 10 parmi les 35000 dernières sont utilisées dans les calculs (3500 itérations retenues). Les résultats sont présentés dans les Tableaux 7.1, 7.2,7.3, . Ces tableaux comportent les valeurs des paramètres qui rendent compte des hypothèses causales du modèle :

- les paramètres de passage des variables latentes aux variables à expliquer (paramètres Λ au Tableau 7.3) qui traduisent l'effet des fonctions mentales sur les résultats aux tests et sur le statut de démence à T3,
- les paramètres de passage des variables explicatives aux variables latentes (paramètres β au Tableau 7.1),
- les variances des fonctions mentales (au Tableau 7.2) qui traduisent la variabilité des fonctions mentales qui n'est pas expliquée par les variables explicatives.

Dans chaque tableau, on indique la valeur moyenne des paramètres ainsi qu'un intervalle de confiance à 95 %. Quand un paramètre est imposé et nul (correspondant à l'absence d'un lien dans le graphe), la valeur indiquée est «0» et il n'y a pas d'intervalle de confiance.

TAB. 7.3: *Matrice Λ des coefficients de régression des variables de test par les fonctions mentales*

	automatique	contrôle
benton	0	9.55e-2
		9.07e-2 ; 1.01e-1
isaac	0	2.48e-1
		2.34e-1 ; 2.62e-1
codes de wechsler	0	2.66e-1
		2.52e-1 ; 2.81e-1
similitudes	1.75e-3	6.08e-2
	4.83e-5 ; 6.05e-3	5.62e-2 ; 6.49e-2
zazzo vitesse	0	1.22e-2
		1.16e-2 ; 1.29e-2
zazzo perf	2.48e-1	2.60e-2
	2.42e-1 ; 2.55e-1	2.05e-2 ; 3.16e-2
mms		
orientation temps	4.01e-2	6.50e-3
	3.84e-2 ; 4.20e-2	4.93e-3 ; 8.04e-3
orientation espace	4.87e-2	1.09e-3
	4.80e-2 ; 4.93e-2	5.40e-4 ; 1.66e-3
calcul	0	3.61e-2
		3.43e-2 ; 3.82e-2
rappel 3 mots	0	1.53e-2
		1.44e-2 ; 1.62e-2
commande verbale	2.98e-2	0
	2.96e-2 ; 2.99e-2	
mots couples faciles	2.26e-2	2.52e-2
	1.89e-2 ; 2.64e-2	2.17e-2 ; 2.86e-2
mots couples difficiles	0	1.44e-2
		1.35e-2 ; 1.54e-2
rappels mots	4.57e-2	1.05e-2
couples faciles	4.38e-2 ; 4.77e-2	8.78e-3 ; 1.22e-2
rappels mots	0	2.40e-2
couples difficiles		2.27e-2 ; 2.54e-2
démence	5.86e-2	-9.68e-2
à T3	3.49e-2 ; 8.43e-2	-1.28e-1 ; -6.97e-2

TAB. 7.4: *Matrice β obtenue par maximum de vraisemblance avec le logiciel* EQS. *Ces résultats sont à comparer au Tableau 7.1.*

	automatique	contrôle
age	0.014 -0.013 ; 0.040	-1.617 -1.81 ; -1.42
niveau scolaire	1.87e-1 -1.80e-1 ; 5.55e-1	1.64e+1 1.35e+1 ; 1.93e+1
profession en 3 niveaux	-6.74e-2 -8.01e-1 ; 6.66e-1 -1.10e-1 -8.56e-1 ; 6.35e-1	2.75e+0 -3.03e+0 ; 8.54e+0 9.71e+0 3.53e+0 ; 1.59e+1

TAB. 7.5: *Écarts types des fonctions mentales obtenus par maximum de vraisemblance avec le logiciel* EQS. *Ces résultats sont à comparer au Tableau 7.2.*

	estimation	intervalle de confiance à 95%
automatique	0.0	0.00 ; 0.078
contrôle	14.9	13.7 ; 16.1

7.3.4 Solution 2 : comparaison avec le maximum de vraisemblance

Une estimation par maximum de vraisemblance a été obtenue à l'aide du logiciel EQS. Les variables ordinales à expliquer sont traitées comme des variables continues et le niveau moyen de référence des fonctions mentales est fixée à 100. Mais, le logiciel ne permet pas d'utiliser un lien logistique pour modéliser l'influence des fonctions mentales sur la démence : la démence a été considérée comme une variable continue (à valeur 0 ou 1) de sorte que ces paramètres ne sont pas directement comparables.

Les résultats numériques sont présentés aux Tableaux 7.4, 7.5, 7.6. Ces tableaux montrent les valeurs estimées par les deux méthodes. On constate une grande similitude, sauf bien entendu pour la démence. Cependant, les signes des paramètres de l'influence des fonctions mentales sur la démence sont les mêmes.

TAB. 7.6: *Matrice Λ obtenue par maximum de vraisemblance avec le logiciel* EQS. *Ces résultats sont à comparer au Tableau 7.3.*

	automatique	contrôle
benton	0	0.0948
		0.0897; 0.1000
isaac	0	0.2457
		0.2323;0.2591
codes de wechsler	0	0.2641
		0.2493; 0.2789
similitudes	0.00	0.0617
	-0.3674; 0.3674	0.0582; 0.0653
zazzo vitesse	0	0.0121
		0.0114; 0.0128
zazzo perf	0.2475	0.0260
	0.2410; 0.2541	0.0204; 0.0316
mms		
orientation temps	0.0399	0.0065
	0.0381; 0.0418	0.0049; 0.0081
orientation espace	0.0485	0.0011
	0.0479; 0.0492	0.0006; 0.0016
calcul	0	0.0359
		0.0339; 0.0379
rappel 3 mots	0	0.0152
		0.0143; 0.0161
commande verbale	0.0297	0
	0.0295; 0.0299	
mots couples faciles	0.0223	0.0252
	0.0186; 0.0260	0.0219; 0.0286
mots couples difficiles	0	0.0143
		0.0133; 0.0152
rappels mots	0.0456	0.0105
couples faciles	0.0436; 0.0475	0.0087; 0.0122
rappels mots	0	0.0238
couples difficiles		0.0225; 0.0252
démence	0.0026	-0.0020
à T3	0.0021; 0.0032	-0.0025; -0.0016

TAB. 7.7: *Matrice β obtenue en fixant la variance des variables latentes à « 1». Les valeurs sont corrigées pour être comparables à celles du Tableau 7.1.*

	automatique	contrôle
age	8.94e-3 -1.90e-2 ; 3.68e-2	-1.63e+0 -1.84e+0 ; -1.44e+0
niveau scolaire	2.86e-1 -4.57e-2 ; 6.49e-1	1.63e+1 1.35e+1 ; 1.90e+1
profession en 3 niveaux	1.30e-1 -4.41e-1 ; 6.33e-1 6.42e-2 -4.56e-1 ; 5.64e-1	2.82e+0 -2.74e+0 ; 8.75e+0 9.88e+0 4.11e+0 ; 1.61e+1

TAB. 7.8: *Écarts types des fonctions mentales obtenus en fixant la variance des variables latentes à « 1». Les valeurs sont corrigées pour être comparables à celles du Tableau 7.2.*

	estimation
automatique	9.59e-2
contrôle	2.27e+2

7.3.5 Solution 3 : comparaison des résultats quand on fixe les variances des fonctions mentales à 1

Des essais ont été faits en fixant les variances des fonctions mentales à 1 et en laissant libre le niveau moyen de référence. D'un point de vue algébrique, cette option est équivalente à celle de l'analyse de référence qui fixe le niveau moyen de référence à 100. D'un point de vue numérique, on constate que cette option est moins favorable : il y a une très grande dépendance entre la valeur du niveau moyen et les valeurs des paramètres de Λ et β. En conséquence la stabilité et la convergence est beaucoup plus lente et il faut des chaînes beaucoup plus longues pour obtenir la stabilité.

Néanmoins, les résultats numériques sont présentés aux Tableaux 7.7, 7.8, 7.9. Ces tableaux montrent les valeurs estimées par les deux méthodes. Pour l'option variance fixée, on présente les résultats obtenus par une chaîne, après avoir fait la transformation linéaire qui permet de comparer directement les valeurs obtenues par les deux approches.

On constate une grande similitude des résultats. Mais, du fait de la mauvaise stabilité des chaînes, la préférence est donnée à l'analyse de référence et à ses options.

TAB. 7.9: *Matrice Λ obtenue en fixant la variance des variables latentes à «1». Les valeurs sont corrigées pour être comparables à celles du Tableau 7.3.*

	automatique	contrôle
benton	0	9.48e-2
		8.96e-2 ; 9.98e-2
isaac	0	2.46e-1
		2.33e-1 ; 2.58e-1
codes de wechsler	0	2.64e-1
		2.50e-1 ; 2.79e-1
similitudes	1.80e-3	6.02e-2
	5.57e-5 ; 5.73e-3	5.57e-2 ; 6.42e-2
zazzo vitesse	0	1.21e-2
		1.15e-2 ; 1.27e-2
zazzo perf	2.47e-1	2.55e-2
	2.40e-1 ; 2.55e-1	2.03e-2 ; 3.11e-2
mms		
orientation temps	3.99e-2	6.48e-3
	3.81e-2 ; 4.18e-2	4.86e-3 ; 8.01e-3
orientation espace	4.86e-2	9.24e-4
	4.79e-2 ; 4.94e-2	3.48e-4 ; 1.50e-3
calcul	0	3.59e-2
		3.40e-2 ; 3.78e-2
rappel 3 mots	0	1.52e-2
		1.43e-2 ; 1.61e-2
commande verbale	2.96e-2	0
	2.93e-2 ; 3.00e-2	
mots couples faciles	2.29e-2	2.47e-2
	1.85e-2 ; 2.63e-2	2.16e-2 ; 2.85e-2
mots couples difficiles	0	1.43e-2
		1.33e-2 ; 1.52e-2
rappels mots	4.58e-2	1.02e-2
couples faciles	4.37e-2 ; 4.78e-2	8.52e-3 ; 1.21e-2
rappels mots	0	2.38e-2
couples difficiles		2.25e-2 ; 2.52e-2
démence	5.74e-2	-9.51e-2
à T3	3.16e-2 ; 8.45e-2	-1.29e-1 ; -6.60e-2

7.4 Interprétation des résultats

L'interprétation est basée sur les résultats numériques de l'analyse de référence (Tableaux 7.3, 7.1, 7.2). Vu les grandes similitudes avec les autres analyses, ce choix n'a pas de grandes conséquences sur l'interprétation.

- La variance de la variable latente *automatisme* est très faible, les effets des variables explicatives sur la variable latente *automatisme* sont presque nuls : cette VL est quasiment constante.

- De ce fait, les valeurs des coefficients de régression des variables de test et en particulier de la démence par la variable latente *automatisme* sont difficilement interprétables.

- Le coefficient de régression de la démence à T3 par la variable latente *contrôle* est négatif : une baisse de la performance de la variable latente représentant la fonction mentale *contrôle* à T2 correspond à une incidence plus élevée de l'apparition d'une démence à T3 . Au contraire, le coefficient de regression de la démence sur la VL automatisme est positif. Ceci pourrait conforter l'hypothèse d'un maintien des processus automatiques dans la phase-préclinique de la démence où seuls les processus contrôlés se détérioreraient.

7.5 Discussion

La très faible valeur de la variance résiduelle de la fonction mentale *automatique* et des paramètres mesurant l'association entre les variables explicatives X et cette fonction mentale jettent un doute sur la validité de ce modèle. Peut-on interpréter une «variable latente» dont la variabilité est nulle? Plus précisément, l'interprétation des relations entre la démence ou les tests et la variable latente *automatique* est-elle licite?

La faible valeur de la variance de la fonction mentale *automatisme* est probablement la conséquence de la faible variance de la variable de test qui est considéré dans le graphe comme entièrement *automatique*. D'ailleurs, les deux autres variables qui sont déclarées entièrement *automatiques* dans le graphe sont constantes et ont été supprimées de l'analyse. En effet, ces 3 variables sont des sous-tests très faciles du MMS qui sont réussis par la plupart des sujets (*répétition de 3 mots*, *commande verbale* et *nommer 2 objets*). De manière plus générale, les tests automatiques sont par définition des tests faciles (donc de faible variance). Il paraît donc illusoire d'étudier les processus automatiques et contrôlés dans une population peu ou pas détériorée. L'estimation de ce modèle sur un échantillon de sujets déjà déments serait probablement plus instructive.

La compréhension des résultats obtenus sur l'échantillon Paquid et plus largement de l'influence des différentes hypothèses sur les résultats de ce type de modèle nécessitera des analyses complémentaires :

- Il serait utile d'essayer d'estimer le modèle « complet » incluant les 2 variables latentes *contrôle* et *automatique* et toutes les flèches entres les variables de test Y et les fonctions mentales F, ainsi qu'un modèle incluant une seule variable latente (le *contrôle*) et toutes les flèches entres les variables de test Y et cette variable latente. La comparaison de la vraisemblance de ces deux modèles permettrait de tester si la deuxième fonction mentale est utile à l'ajustement des données.

- Si l'analyse précédente amène à conclure que la deuxième fonction mentale n'est pas utile, la question reste ouverte de savoir si un autre graphe incluant plusieurs fonctions mentales mieux choisies pourrait constituer un meilleur modèle. Il serait par exemple intéressant d'estimer le modèle correspondant au deuxième graphe (Figure 7.2). Ce graphe inclut les mêmes

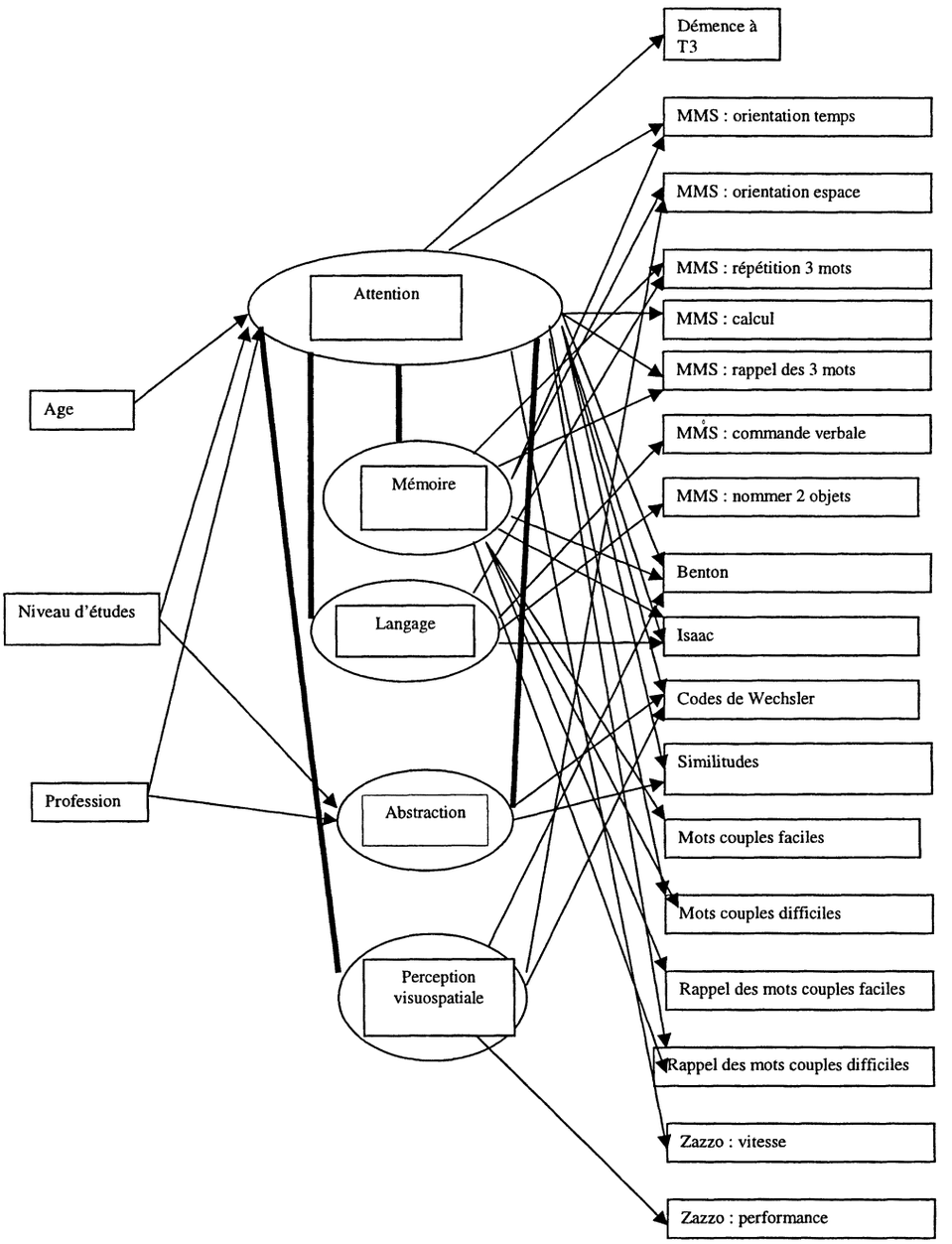
variables de test et les mêmes variables explicatives mais il fait intervenir 5 fonctions mentales qui sont supposées plus spécifiques de certains tests que les notions «contrôle» et «automatique».

- Dans les deux graphes, il serait intéressant de supprimer le sous-test facile du MMS *commande verbale*[2] pour évaluer dans quelle mesure il influence les résultats concernant notamment les variables latentes *automatique* dans le graphe 7.1 et *langage* dans le graphe 7.2.

- Par ailleurs, la prise en compte du facteur *démence à trois ans* fait problème. Dans le graphe 7.1, la variable démence est considérée comme les variables de test Y. La démence étant liée à toutes les variables latentes, il est probable que sa présence n'influe pas sur l'estimation des relations entre variables latentes et tests. Cependant, il parait plus logique d'utiliser une modélisation en deux étapes : estimation du modèle n'incluant pas la variable démence puis estimation *a posteriori* des valeurs des variables latentes pour chaque sujet et inclusion de ces variables comme variables explicatives dans un modèle de régression logistique dont la variable dépendante serait la *démence à trois ans*.

2. Les 2 autres sous-tests faciles du MMS (*répétition de 3 mots nommer 2 objets*) sont supprimés de fait car les résultats sont constants sur l'échantillon PAQUID utilisé

FIG. 7.2: *Graphe 2 de* PAQUID



Chapitre 8

Conclusion

La faisabilité du modèle statistique a été vérifiée sur des données simulées. Son application à des données réelles permet de confirmer son utilité et de le valider. En effet, pour les deux études, des analyses antérieures classiques suggéraient des hypothèses qui se trouvent confirmer par le modèle statistique mis en œuvre ici :

- pour la cohorte PAQUID, la fonction *contrôle* est la première atteinte dans l'évolution de la démence alors que la fonction *automatique* est préservée plus longtemps, avec les réserves exprimées dans la discussion des résultats.

- pour les données TOLUENE, le processus d'atteinte neurotoxique semble d'abord concerner l'apprentissage et la mémoire.

Les coefficients de régression et leurs intervalles de confiance pour la matrice β valident l'existence des flèches des graphes.

Par ailleurs, ce modèle statistique est utilisable que les données multivariées soient ordinales ou continues. Cependant, on peut noter quelques réserves.

- Le modèle statistique est très dépendant du graphe de dépendance conditionnelle établi *a priori*. La construction d'un tel graphe doit donc s'appuyer sur des connaissances établies du domaine étudié. En effet, s'il est possible de tester l'existence ou l'inexistence d'une flèche $X \rightarrow F$ figurant dans le graphe, l'absence de flèche $F \rightarrow Y$ est rédibitoire : ce sont les connaissances établies du domaine étudié qui imposent cette contrainte. Si les connaissances établies ne précisent pas l'absence d'une flèche, celle-ci doit figurer dans le graphe.

- Même si un graphe est satisfaisant sur le plan de la connaissance, il peut se révéler non identifiable. Certaines des conditions pour rendre identifiable un graphe ont été détaillées. Dans le cas du graphe TOLUENE, il a été nécessaire de faire une approximation en enlevant plusieurs flèches.

- Cette approche nécessite des logiciels spécifiques. EQS ne peut être utilisé que dans certains cas. Parmi les logiciels testés, Bugs semble le plus adapté au problème posé mais son utilisation a les inconvénients suivants :

  - le temps de calcul est élevé,
  - les résultats ne sont pas présentés sous une forme directement interprétable,
  - il faut obtenir une convergence pour une chaîne donnée et une stabilité des chaînes entre elles, ce qui représente un étape délicate vu le nombre de paramètres dans chacun des modèles.

- Enfin, il y a deux limites à l'utilisation de ce modèle :
  - la distribution des variables doit être compatible avec les hypothèses du modèle de normalité pour les variables continues ; pour les variables ordinales, le modèle des *proportional-odds* est une réponse théorique dont la mise en pratique s'est révélée délicate puisque les analyses n'ont pu être menées à terme.
  - les flèches absentes sur le graphe doivent l'être dans la réalité. Dans le cas contraire, les variables latentes ne représentent pas les fonctions mentales que l'on souhaite modéliser et les résultats sur les hypothèses (effet de l'exposition par exemple) ne peuvent pas être interprétés en termes d'effets sur ces fonctions mentales.

Les objectifs initiaux de l'étude ont été atteints. Nous disposons d'une méthode statistique adaptée aux problèmes posés par l'ENC. Les conditions d'application du modèle sont cernées, les avantages et inconvénients sont repérés. Des progrès sont encore à envisager dans le développement de cette méthode :

- par la réalisation des tests de vraisemblance qui permettront de valider, une par une, les flèches d'un graphe,
- par l'approfondissement des conditions de l'identifiabilité structurelle d'un graphe.
- dans le cas de l'utilisation du logiciel BUGS, par l'élaboration d'une procédure de vérification de la convergence et de la répétabilité des chaînes,
- dans l'amélioration de la prise en compte du caractère ordinal de certaines variables de test,
- dans l'approfondissement de l'analyse sur le jeu de données PAQUID tel qu'il est envisagé dans la discussion des résultats.

Cette méthode peut être adaptée à l'analyse des données longitudinales comme c'est le cas de données dont nous disposons déjà :

- de la cohorte des apprentis peintres ; il s'agit d'une étude longitudinale prospective d'ENC destinée à étudier les effets NT des solvants de la peinture.
- de la cohorte PAQUID avec les données des 5 visites,

En outre, cette méthode pourra s'appliquer dans tout autre domaine de l'épidémiologie lorsque les données présentent les caractéristiques suivantes : données multivariées, avec présence de fonctions latentes, nécessitant une réduction de dimensionalité comme, par exemple, en l'épidémiologie psychosociale.

Bibliographie

[1] G. Arminger. A bayesian approach to nonlinear latent variable models using gibbs sampler and metropolis-hastings algorithm. *Psychometrika*, 63(3):271–300, 1998.

[2] G. Arminger, C. C. Clogg, and M. E. Sobel, editors. *Handbook of statistical modeling for social and behavioral sciences*. Plenum Press, New York, 1995.

[3] P. M. Bentler. *EQS 4.0, Structural Equations Program Manual*. Multivariate Software, Inc., Encino, CA, 1995.

[4] J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors. *Bayesian statistics 4*, Oxford, 1992. Clarendon Press.

[5] R. D. Bock and R. D. Gibbons. High-dimensional multivariate probit analysis. *Biometrics*, 52(4):1183–1194, 1996.

[6] K. A. Bollen. *Structural equations with latent variables*. series in probability and mathematical statistics. Wiley, New York, 1989.

[7] M. W. Browne. Robustness of statistical inference in factor analysis and related models. *Biometrika*, 74(2):375–384, 1987.

[8] W. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8(2), 4 1996.

[9] D. Chouanière. Les troubles neuro-comportementaux liés à l'exposition professionneelle prolongée au toluène. Rapport d'étude, INRS, 1997.

[10] D. Cox and N. Wermuth. Linear dependencies represented by chain graphs (with discussion). *Statistical Science*, 8:204–284, 1993.

[11] D. Cox and N. Wermuth. *Multivariate Dependencies. Models, analysis and interpretation*. Chapman & All, London, 1996.

[12] D. R. Cox and N. Wermuth. Response models for mixed binary and quantitative variables. *Biometrika*, 79(3):441–461, 1992.

[13] A. P. Dawid. Conditional independence in statistical theory. *J. R. Statist. Soc. B*, 41(1):1–31, 1979.

[14] A. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[15] E. Diday, J. Lemaire, J. Pouget, and F. Testu. *Éléments d'ananlyse de données*. Dunod, Paris, 1982.

[16] D. Edwards. *Introduction to graphical modelling*. Springer Texts in Statistics. Springer-Verlag, New York, 1995.

[17] P. K. Elias, M. F. Elias, R. B. D'Agostino, H. Silbershatz, and P. A. Wolf. Alcohol consumption and cognitive preformance in the framingham heart study. *American Journal of Epidemiology*, 150(6), 1999.

[18] B. S. Everitt and G. Dunn. *Applied multivariate data analysis*. Arnold, London, 1991.

[19] L. Fahrmeir and G. Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer Texts in Statistics. Springer-Verlag, New York, 1994.

[20] W. J. Gilks, S. Richardson, and D. J. Spiegelalter. *Markov Chain Monte Carlo in practice*. Chapman & All, London, 1996.

[21] N. J. Heyer, A. C. Bittner, and D. Echeverria. Analysing multivariante neurobehavioral outcomes in occupational studies: a comparison of approaches. *Neurotoxicology and Teratology*, 18(4):401–406, 1996.

[22] K. Hinkelmann and O. Kempthorne. *Design and analysis of experiments, volume I. Introduction to experimental design*. series in probability and mathematical statistics. John Wiley & Sons Ltd, New York, 1994.

[23] S. I. Inc. *SAS/STAT® User's Guide, Version 6*. SaS Institute Inc., Cary, NC, fourth edition, 1989.

[24] K. G. Jörestog and D. Sörbom. *LISREL 8 : Structural Equation modeling with with SIMPLIS command language*.

[25] R. E. Kaas and A. E. Raftery. Bayes factors and model uncertainty. *J. Am. Statist. Ass.*, 90:773–795, 1995.

[26] K. Kim. A bivariate cumulative probit regression model for ordered categorical data. *Statistics in Medicine*, 14(12):1341–1352, 1995.

[27] S. L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annales of Statistics*, 17(1):31–57, 1989.

[28] S.-Y. Lee, W.-Y. Poon, and P. M. Bentler. Structural equation models with continuous and polytomous varaibles. *Psychometrika*, 57(1):89–105, 1992.

[29] J. M. Legler and L. M. Ryan. Latent variable models for multiple birth outcomes. *J. Am. Statist. Ass.*, 1997.

[30] M. D. Lezak. *Neuropsychological assessment*. Oxford University Press, Oxford, 1983.

[31] D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical modeles using occam's window. *J. Am. Statist. Ass.*, 89:1535–1546, 1994.

[32] P. McCullagh and J. N. FRS. *Generalized Linear Models*. Chaman and Hall, London, second edition, 1989.

[33] B. Muthén. A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, 49(1):115–132, 1984.

[34] B. Muthén. Latent variable modeling in epidemiology. *Alcohol Health & Research World*, 16(4):286–292, 1992.

[35] M. D. Sammel and L. M. Ryan. Latent variable models with fixed effects. *Biometrika*, 52(4):650–663, 1996.

[36] M. D. Sammel, L. M. Ryan, and J. M. Legler. Latent variable models for mixed discrete and continuous outcome. *J. R. Statist. Soc. B*, 59(3):667–678, 1997.

[37] R. Scheines, H. Hoiltink, and A. Boomsma. Bayesian estimation and testing of structural equation models. *Psychometrika*, 64(1):37–52, 1999.

[38] A. Schepers and G. Arminger. *MECOSA: A programm for the analysis of general mean- and covariance structures with non-metric variables, Users guide*. Frauenfeld: SLI-AG, Zürcher Str. 300, CH-8500 Frauenfeld, Switzerland, 1992.

[39] J.-Q. Shi and S.-Y. Lee. A bayesian estimation of factor score in confirmatory factor model with polytomous, censored or truncated data. *Psychometrika*, 62(1):29–50, 1997.

[40] D. Spiegelhalter, A. Thomas, N. Best, and W. Gilks. *BUGS 0.5, Bayesian inference using Gibbs sampling manual (version* ii*)*. Medical Research Council Biostatistics Unit, Cambridge, 1996.

[41] D. Spiegelhalter, A. Thomas, N. Best, and W. Gilks. *BUGS 0.5 Examples, volume 1*. Medical Research Council Biostatistics Unit, Cambridge, 1996.

[42] D. Spiegelhalter, A. Thomas, N. Best, and W. Gilks. *BUGS 0.5 Examples, volume 2*. Medical Research Council Biostatistics Unit, Cambridge, 1996.

[43] E. Stanghellini. Identification of a single-factor model using graphical gaussian rules. *Biometrika*, 84(1):241–244, 1997.

[44] M. A. Tanner. *Tools for statistical inference*. Springer Texts in Statistics. Springer-Verlag, New York, 1993.

[45] M. Tenenhaus. L'approche pls. Les cahiers de recherche 643, groupe HEC, 1998.

[46] D. Wedelin. Efficient estimation and model selection in large graphical models. *Statistics and Computing*, 6(4):313–323, 1996.

[47] N. Wermuth and S. L. Lauritzen. On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. R. Statist. Soc. B*, 52(1):21–50, 1990.

[48] J. Whittaker. *Graphical models in applied multivariate statistics*. John Wiley & Sons Ltd, Chichester, 1990.