Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'**Université de Montpellier**

Préparée au sein de l'école doctorale **I2S**
Et de l'unité de recherche **IMAG**

Spécialité: **Biostatistique**

Présentée par **Myriam Tami**

Approche EM pour modèles multi-blocs à facteurs à une équation structurelle

Soutenue le 12 Juillet 2016 devant le jury composé de

M. X. Bry	Maître de Conférences	Université de Montpellier	Co-directeur de thèse
Mme B. de Saporta	Professeure	Université de Montpellier	Présidente du jury
M. C. Lavergne	Professeur	Université Paul-Valéry	Directeur de thèse
M. G. Saporta	Professeur émérite	CNAM	Rapporteur
M. J. Saracco	Professeur	Université de Bordeaux 1	Rapporteur
Mme L. Trinchera	Maîtresse de Conférences	NEOMA Business School	Examinatrice

Université Montpellier

IMAG
INSTITUT MONTPELLIERAIN ALEXANDER GROTHENDIECK

École doctorale I2S : Information Structures Systèmes

THÈSE

pour obtenir le grade de docteur délivré par

l'Université de Montpellier

Spécialité doctorale "Biostatistique"

présentée et soutenue publiquement par

Myriam Tami

le 12 Juillet 2016

Approche EM pour modèles multi-blocs à facteurs à une équation structurelle

sous la direction de

Directeur de thèse : **Christian Lavergne**
Co-directeur de thèse : **Xavier Bry**

Quelques mots de remerciements...

Pour commencer, merci aux rapporteurs ainsi qu'aux membres du jury pour leurs lectures attentives et la pertinence de leurs remarques.

Merci à mes directeurs de thèse pour avoir été complémentaires et pour la liberté de travail de recherche qu'ils m'ont laissée.

Je tiens également à remercier les différents membres du laboratoire de l'IMAG pour le climat de travail agréable auquel ils contribuent : les permanents pour le partage de leurs expériences, leurs conseils mais aussi aux doctorants pour leur amitié bienveillante, leur disponibilité à l'échange de nos connaissances ainsi qu'au partage de nos premières expériences d'enseignement qui aboutirent à de longues réflexions didactiques.

Merci aussi à Bernadette Lacan, Sophie Cazanave Pin et leurs collègues du secrétariat de l'IMAG pour leur précieuse aide administrative et la patience dont ils font preuve.

Enfin, je remercie mes proches et amis pour leur écoute, leur présence constante et pour m'avoir toujours poussé à dépasser les obstacles rencontrés.

Pour finir, de manière générale, merci à toutes celles et tous ceux qui m'ont soutenu que ce soit de manière brève ou de manière inconditionnelle.

Introduction

L'intérêt des modèles d'équations structurelles à variables latentes est de pouvoir modéliser des relations entre des variables observables et non observables. Leur formalisme mathématique est un système d'équations organisé en deux parties. La première est constituée des équations modélisant les relations de causalité entre les variables latentes (non observables) seules. La seconde est constituée des équations décrivant les relations de causalité entre les variables latentes et les variables observables. Ces modèles semblent alors adaptés à la modélisation et la quantification de systèmes de concepts complexes et non mesurables directement (Bacher, 1987).

Ces modèles font l'objet de nombreux travaux de recherche en statistique et en analyse de données depuis le début du XX^{e} siècle : que ce soit au niveau des relations de causalité comme au niveau des variables latentes. Les praticiens et chercheurs de nombreux domaines d'application s'y sont intéressés, notamment en sociologie, psychologie et plus récemment en marketing, gestion et management. Par exemple, pour l'étude et la modélisation de la satisfaction des consommateurs les travaux de Stan and Saporta (2006); Jakobowicz (2007) peuvent être consultés. Il existe deux familles de méthodes d'estimation de ces modèles : d'une part, l'approche PLS (Partial Least Squares) (Wold, 1966; Karl G. Jöreskog, 1982; Wold, 1985) fondée sur la régression et une estimation par moindres carrées partiels ; d'autre part, l'analyse de la structure de covariance connue sous le nom de LISREL (LInear Structural RELations) (Jöreskog, 1970; Jöreskog and Sörbom, 1982). Originellement, LISREL n'estime que les coefficients de équations reliant les variables latentes en se fondant sur une estimation du maximum de vraisemblance de la structure de covariance des variables observables induite par le modèle. Concernant l'estimation des variables latentes, PLS leur impose d'être des composantes (i.e. : combinaison linéaire des variables observées), ce qui réduit l'espace des solutions. Cette technique a pendant un certain temps complété l'approche LISREL afin de permettre une estimation des variables latentes (McDonald, 1996; Tenenhaus, 2007). Ce n'est que depuis peu (Jöreskog, 2000) que LISREL propose une méthode d'estimation des variables latentes moins contrainte. Cette technique est du type moindres carrés mais n'utilise qu'une partie des équations du modèle. Les équations modélisant les relations entre les variables latentes seules, n'y sont pas utilisées. De manière globale, LISREL et PLS sont à la fois concurrentes et complémentaires : PLS est plutôt employée pour faire de la prévision alors que LISREL l'est plutôt pour de la validation de modèle (Stan and Saporta, 2006).

Dans notre travail, nous nous plaçons dans le paradigme du maximum de vraisemblance en montrant qu'il permet une estimation des variables latentes. Dans le cas où les variables latentes ne sont astreintes qu'à suivre une loi gaussienne (on les nomme "facteurs"), l'algorithme Expectation-Maximization (EM) (Dempster et al., 1977) s'avère être l'outil adéquat. Il permet à la fois l'estimation des paramètres d'un modèle à équations structurelles et celles de ses facteurs latents. Afin de développer l'approche d'estimation via l'algorithme EM et de mettre en place les notations, un modèle à une seule équation structurelle a été choisi. Bien

entendu, cette approche peut se généraliser à des modèles plus riches et complexes. D'ailleurs, l'algorithme EM peut s'appliquer à de nombreux modèles.

C'est une démarche qui a par exemple été appliquée dans le cas des modèles linéaires mixtes (Andrade and Helms, 1984; Dempster et al., 1981). En effet, dans les années 50, les méthodes d'estimations issues des travaux de Fisher, étaient du type Analyse de Variance (ANOVA) à effets aléatoires. Dans les années 70, Rao a émis plusieurs critiques sur les faiblesses computationnelles de la méthode ANOVA et proposa la méthode MINQUE (MInimum-Norm Quadratic Unbiased Equation). Plus tard, les méthodes numériques proposées par Newton et Raphson et la méthode de Scoring de Fisher ont été développées mais restent difficiles à implémenter à certains modèles, notamment à variables latentes. Dans les années 80 l'algorithme EM est proposé comme une alternative à la méthode MINQUE. Cela a permis à la fois de proposer la reconstruction des effets aléatoires et de rester dans le cadre du maximum de vraisemblance gaussien. Aujourd'hui, dans le cas des modèles linéaires mixtes, EM s'est imposé car il permet un travail d'estimation riche, sophistiqué et généralisable. Ainsi, dans ce travail se trouve la même idée d'injecter l'algorithme EM dans l'estimation des modèles d'équations structurelles à facteurs latents, ce qui permettra de proposer de multiples généralisations. Actuellement, dans le paradigme des modèles à équations structurelles, l'algorithme EM est utilisé de manière classique pour le cas de problèmes de données manquantes i.e. : lorsque pour certaines observations, un nombre fixé de données associées à une ou plusieurs variables ne sont pas disponibles. À la différence de l'algorithme EM, LISREL propose la méthode FIML (Full Information Maximum Likelihood) (Arbuckle et al., 1996) où afin d'estimer les paramètres du modèle, seules les données disponibles de chaque observation sont utilisées lors de l'estimation par maximum de vraisemblance. Cette méthode est moins brutale que les méthodes de type *deletion*. On peut faire le parallèle avec l'algorithme NIPALS pour PLS. Pour un certain type de données manquantes, des méthodes d'imputation nécessitant la connaissance du processus conduisant à l'absence des données sont employées. Elles sont associées aux modèles de mélanges et aux modèles dits de sélection d'échantillon. Cependant, pour les modèles d'équations structurelles à variables latentes, la littérature se limite au cas LISREL sur lequel se sont penchés Muthén et al. (1987); Tang and Lee (1998); Lee and Tang (2006). Ils supposent le processus connu et les modèles de mélanges ainsi que de sélection d'échantillon sont employés. Néanmoins, EM fait son apparition dans les travaux de Tang and Lee (1998). Ils reformulent le processus d'absence des données en estimant sa distribution qu'ils utilisent dans un algorithme EM. EM est aussi utilisé dans le cadre des modèles de mélanges où les données observées sont complétées par des classes inobservables et que l'on cherche à reconstruire (Moosbrugger et al., 1997). L'ensemble de ces méthodes est disponible à travers le logiciel commercial Mplus de Muthén and Muthén (1998). Plusieurs types de données manquantes et de méthodes d'imputation existent, mais seules les plus classiques et celles employant l'algorithme EM seront abordées dans le cadre de ce travail afin de clarifier les différences entre ces approches et celle que nous proposons. En effet, les problématiques sont différentes. L'approche que nous proposons ne répond pas aux problèmes de données partiellement manquantes mais complète les données observées par les facteurs afin de les estimer en plus des paramètres du modèle à travers l'outil : algorithme EM. Pour appliquer l'algorithme EM aux modèles à facteurs et à équations structurelles, un modèle simple et non restrictif sinon qu'il ne contient qu'une équation structurelle reliant les variables latentes a été choisi. Il est composé d'un facteur dépendant et plusieurs facteurs explicatifs, où chacun des facteurs est lié à un groupe de variables observables. Le système d'équations formé par le modèle comporte une seule équation structurelle mais aussi des équations linéaires reliant chaque groupe de variables observables à son facteur. Nous le formaliserons au premier chapitre. Nous y présenterons également les modèles à facteurs avant d'introduire les notations choisies pour les différentes approches d'estimation qui seront abordées. Parmi celles-ci, les deux familles d'approches d'estimation LISREL et PLS classiquement utilisées dans la littérature seront introduites et illustrées. Au chapitre 3 nous

poursuivrons sur les récents développements méthodologiques de LISREL et PLS ainsi que leurs réponses à la question de la reconstruction des concepts latents. Ensuite, nous présenterons l'algorithme EM ainsi que son application pour estimer le maximum de vraisemblance et plus particulièrement les facteurs dans le cas d'un premier modèle élémentaire à une équation structurelle. Ce dernier sera généralisé au chapitre suivant à un modèle multi-bloc plus riche avec adjonction de covariables. Nous y aborderons la question des performances de l'approche proposée. Ces dernières seront étudiées au travers de données simulées et d'une analyse de sensibilité. Puis nous illustrerons cette approche sur des données environnementales. Nous terminerons par un chapitre où cette approche est utilisée dans le contexte d'un essai clinique en cancérologie comme étape préliminaire lors d'une étude longitudinale de la qualité de vie relative à la santé sur les facteurs reconstruits à plusieurs temps de suivis.

Table des matières

Introduction **iii**

1 Les modèles à équations structurelles et variables latentes, et leurs méthodes d'estimation **1**
1.1 Les modèles étudiés . . . 2
1.1.1 Les modèles à facteurs . . . 2
1.1.2 Les modèles à équations structurelles . . . 4
1.1.3 Les modèles à équations structurelles à facteurs . . . 6
1.1.4 Formalisme . . . 6
1.1.5 Commentaires : lien entre modèle à facteurs et ACP . . . 9
1.2 Les méthodes d'estimation des modèles à équations structurelles et variables latentes . . . 13
1.2.1 Modèle à composantes : l'approche de Wold (PLS) . . . 13
1.2.2 Modèles à facteurs : l'approche de Jöreskog (LISREL) . . . 18
1.3 Conclusion et discussion . . . 29

2 L'estimation par algorithme EM **31**
2.1 Utilisation actuelle de l'algorithme EM par les approches PLS et LISREL . . . 32
2.1.1 Introduction et motivations . . . 32
2.1.2 La question des données manquantes . . . 32
2.1.3 La question de la reconstruction de concepts latents . . . 37
2.1.4 Conclusion . . . 40
2.2 L'algorithme EM . . . 41
2.2.1 Structure générale de l'algorithme EM . . . 41
2.2.2 Preuve de la convergence de l'algorithme EM . . . 43
2.3 Méthode d'estimation par algorithme EM pour un modèle à une équation structurelle et un facteur par bloc de variables observées . . . 44
2.3.1 Modèle à deux blocs : l'un dépendant et l'autre explicatif . . . 44
2.3.2 L'algorithme EM pour le modèle à deux blocs : l'un dépendant et l'autre explicatif . . . 45
2.4 Conclusion et discussion . . . 50

3 Algorithme EM et modèles multi-blocs à facteurs **53**
3.1 Introduction et motivations . . . 54
3.2 Estimation par algorithme EM d'un modèle structurel multi-blocs à facteurs 54
3.2.1 Introduction . . . 54
3.2.2 Structure générale du modèle . . . 54

3.2.3 Estimation par algorithme EM . . . 55
3.2.4 L'algorithme . . . 56
3.2.5 Performances de l'approche . . . 57
3.2.6 Application de l'approche EM à des données réelles environnementales 57
3.2.7 L'article soumis . . . 60
3.3 Résultats complémentaires de l'application environnementale : le cas du modèle sans la covariable *géologie* . . . 82
3.3.1 Le modèle sans covariables . . . 82
3.3.2 Tableaux supplémentaires de l'application aux données environnementales . . . 84
3.4 Perspectives et discussion sur les questions du nombre de blocs, de facteurs, de parcimonie et d'unicité des solutions . . . 86

4 Analyse longitudinale de la qualité de vie sur des facteurs reconstruits 89
4.1 Introduction . . . 90
4.2 Contexte . . . 90
4.2.1 Essai clinique . . . 90
4.2.2 Critères d'évaluation du bénéfice d'un traitement lors d'un essai clinique 91
4.2.3 La qualité de vie : un critère d'évaluation alternatif . . . 92
4.3 Le critère de qualité de vie (QdV) relative à la santé (HRQoL) . . . 92
4.3.1 Mesure de la HRQoL par auto-questionnaires . . . 93
4.3.2 Évaluation de la HRQoL . . . 93
4.3.3 Analyse longitudinale classique de la HRQoL . . . 95
4.4 Une analyse longitudinale de la HRQoL en deux étapes . . . 97
4.4.1 Introduction . . . 97
4.4.2 Première étape : analyse transversale . . . 97
4.4.3 Seconde étape : analyse longitudinale par modèle linéaire mixte . . . 99
4.4.4 Application à des données réelles issues de l'essai clinique CO-HO-RT 99
4.4.5 L'article soumis . . . 100
4.5 Discussion, commentaires et perspectives . . . 120

Conclusion 121

A Démonstrations aboutissant aux formules des estimateurs des paramètres du modèle de LISREL 123

B Compléments à la deuxième étape de l'algorithme de Jöreskog 129

C Questionnaire EORTC QLQ-C30 spécifique au cancer 131

D Questionnaire EORTC QLQ-BR23 spécifique au cancer du sein 133

Références 135
Bibliographie . . . 135

Résumé / Abstract 144

Liste des tableaux

1.1 Recoupement des notations utilisées pour les différentes approches d'estimation. 8

3.1 Application aux données *genus* sans covariables T : estimations des paramètres D et b, et des corrélations de $\widetilde{g}$ avec Y. 83
3.2 Application aux données *genus* sans la covariable *géologie* : estimations des paramètres scalaires. 83
3.3 Application aux données *genus* sans covariables T : estimation des paramètres d^1 et a^1, et des corrélations de $\widetilde{f^1}$ avec les variables X^1. 84
3.4 Application aux données *genus* sans covariables T : estimation des paramètres d^1 et a^1, et des corrélations de $\widetilde{f^1}$ avec les variables X^1. 84
3.5 Application aux données *genus* sans covariables T : estimation des paramètres d^2 et a^2, et des corrélations de $\widetilde{f^2}$ avec les variables X^2. 85
3.6 Application aux données *genus* sans covariables T : estimations des paramètres d^2 et a^2, et des corrélations de $\widetilde{f^2}$ avec les variables X^2. 86

Liste des figures

1.1 Les trois différents types de modèles de mesure. 5
1.2 Exemple de modèle structurel comportant les deux types de variables latentes : exogène et endogène. 6
1.3 Exemple de modèle à une équation structurelle à variables latentes. 7
1.4 Diagramme du modèle (1.5) pour l'approche PLS. 14
1.5 Schéma de l'approche PLS 16
1.6 Schéma de la correction apportée au vecteur $b'_{2[1]}$ par sa projection orthogonale sur la direction du vecteur $b'_{1[2]}$ au sens de Ψ^{-1}. 22
1.7 Diagramme du modèle (1.5) pour l'approche LISREL. 27

2.1 Tableau de données X de dimension $n \times q$. 33
2.2 Trois familles de données manquantes : MNAR, MAR et MCAR ordonnées par l'intensité d'aléa de leur absence et décrites mathématiquement. 34
2.3 Un modèle à deux groupes de VO : un dépendant et un explicatif, à une équation structurelle. 44

3.1 Procédure itérative de la méthode d'estimation par algorithme EM. 58
3.2 Diagramme structurel du peuplement arboré expliqué par deux blocs de variables observées liées aux conditions environnementales locales et à l'activité végétale. 59

4.1 Les différentes étapes d'un essai clinique pour un traitement anti-cancéreux. . 91
4.2 Structure du questionnaire EORTC QLQ-C30. 94
4.3 Diagramme du modèle structurel induit par la décomposition du QLQ-C30 à chaque temps de suivi. 99

CHAPITRE 1

Les modèles à équations structurelles et variables latentes, et leurs méthodes d'estimation

Sommaire

1.1 Les modèles étudiés 2
1.1.1 Les modèles à facteurs 2
1.1.2 Les modèles à équations structurelles 4
1.1.3 Les modèles à équations structurelles à facteurs 6
1.1.4 Formalisme 6
1.1.5 Commentaires : lien entre modèle à facteurs et ACP 9
1.2 Les méthodes d'estimation des modèles à équations structurelles et variables latentes 13
1.2.1 Modèle à composantes : l'approche de Wold (PLS) 13
1.2.2 Modèles à facteurs : l'approche de Jöreskog (LISREL) 18
1.3 Conclusion et discussion 29

1.1 Les modèles étudiés

Un modèle à équations structurelles et à variables latentes consiste en un système d'équations : de mesure d'une part et structurelles d'autre part. Chacune a un rôle bien particulier : les équations structurelles modélisent uniquement les relations entre les variables non observables, dites latentes alors que les équations de mesure décrivent les relations entre les variables observables et les latentes.

Lorsque la modélisation des relations de causalité entre des variables observables et latentes est posée, vient le choix de certaines hypothèses restrictives et celui de la méthode d'estimation. Les coefficients ne sont pas les seuls éléments inconnus des modèles structurels : les variables latentes le sont aussi. La question de leur estimation se pose alors. Les deux familles de méthodes d'estimation des modèles structurels à variables latentes que sont LISREL et PLS, proposent des techniques d'estimation des variables latentes au niveau individuel (nommées "scores"). Pour y parvenir, elles font des hypothèses plus ou moins contraignantes sur la nature des variables latentes. La technique d'estimation est ainsi dépendante des hypothèses contraignant les variables latentes. LISREL suppose qu'elles suivent une loi gaussienne (on les nomme "facteurs") et PLS fait l'hypothèse qu'elles sont des combinaisons linéaires des variables observées (on les nomme "composantes").

L'approche EM développée dans ce travail de thèse, fait quant à elle l'hypothèse que les variables latentes sont des facteurs. Pour mettre en place cette approche, un modèle simple à une équation structurelle où les variables latentes sont des facteurs est formalisé dans la première partie de ce chapitre. Ensuite, les notations des approches EM, PLS et LISREL étant différentes, elles seront synthétisées dans un tableau pour faciliter le passage d'une méthode à l'autre. Enfin, ce chapitre présente les approches PLS et LISREL ainsi que les méthodes numériques associées.

1.1.1 Les modèles à facteurs

Les modèles à facteurs et leur analyse ont été introduits par Spearman (1904), Kelley (1928) et Thurstone (1931) dans le cadre du domaine des sciences humaines. L'objectif était de répondre au besoin de condensation des données statistiques multivariées en un plus petit nombre d'éléments ou facteurs synthétisant les variables. Depuis lors ces méthodes n'ont cessé de se développer et diversifier. Ces facteurs sont considérés comme des variables latentes, résumant chacune un bloc de variables observées suffisamment corrélées entre elles. Ainsi, par exemple l'analyse factorielle exprime la corrélation entre un grand nombre de variables observées par un petit nombre de facteurs décorrélés. Les variables sont mathématiquement décrites comme des combinaisons linéaires des facteurs auxquels on ajoute un paramètre de moyenne et une erreur de mesure. Les facteurs étant non directement observés, ils doivent être reconstruits en parallèle de l'estimation des paramètres du modèle. Les différentes techniques développées pour l'estimation des modèles à facteurs sont pour la plupart fondées sur le critère du maximum de vraisemblance avec des contraintes d'identification sur les paramètres.

Écriture des modèles à facteurs

Soit q le nombre de variables et n le nombre d'observations sur lesquelles les variables sont recueillies. On note y_i^j la valeur de la j-ème variable, $j \in [\![1, q]\!]$ pour la i-ème observation, $i \in [\![1, n]\!]$. Le modèle décrivant les variables $y^1, \ldots, y^q$ en fonction de K facteurs $g_1, \ldots, g_k, \ldots, g_K$ où $K < q$ peut être formulé au niveau élémentaire d'une variable pour une observation de la manière suivante :

$$y_i^j = \mu_j + g_{i1} b_{1j} + \ldots + g_{iK} b_{Kj} + \varepsilon_i^j$$

où μ_j est le paramètre de moyenne et b_{kj} sont des coefficients pondérateurs (nommés parfois "loadings"). Ce qui se généralise à l'écriture matricielle suivante :

$$Y = \mathbb{1}_n \mu' + GB + \varepsilon$$

où $Y = [y^1, \ldots, y^q]$ est la matrice des données, $\mu' = (\mu_1, \ldots, \mu_q)$ le vecteur des moyennes, G la matrice des facteurs de dimension $(n \times K)$, B une matrice déterministe de dimension $(K \times q)$ à coefficients inconnus dite matrice des pondérations ou "loadings" et ε de dimension $(n \times q)$ est la matrice constituée des erreurs de mesures ε_i^j supposées indépendantes entre elles et indépendantes des facteurs. On suppose également que les n observations d'une même variable sont indépendantes. Avant de présenter les hypothèses classiques à cette modélisation, notons que pour toute observation i, le modèle peut être formulé comme suit :

$$y_i{}' = \mu' + g_i{}' B + \varepsilon_i{}' \tag{1.1}$$

où $g_i{}'$ est un vecteur aléatoire de dimension $(1 \times K)$ et $\varepsilon_i{}'$ un vecteur aléatoire de dimension $(1 \times q)$ dont la variance représente la variabilité des observations non expliquée par les facteurs. Les hypothèses classiques de ces modèles sont :

— $\varepsilon_i \sim \mathcal{N}(0, \Psi)$ où $\Psi = diag(\sigma_j^2)_{j \in [\![1,q]\!]}$;
— $g_i \sim \mathcal{N}(0, I_K)$;
— ε_i et g_ℓ sont mutuellement indépendants pour tout i, ℓ.

Ces hypothèses impliquent que le modèle est construit de telle manière que toute la corrélation entre les variables observables passe par les K facteurs. Ainsi, conditionnellement aux facteurs, les variables observables sont indépendantes et leur distribution conditionnellement aux facteurs se factorise comme suit :

$$p(y|g;\theta) = \prod_{i=1}^{n} p(y_i|g_i;\theta) \tag{1.2}$$

où, $g = (g_1, \ldots, g_n)$, $y = (y_1, \ldots, y_n)$ et $\theta = (\mu, B, \Psi)$ l'ensemble des paramètres. Ce modèle consiste donc à chercher le nombre minimal de facteurs K tel que cette propriété soit satisfaite. Dans le cadre des hypothèses de distributions gaussiennes du modèle (1.1), $\forall i,\ g_i \sim \mathcal{N}(0, I_K)$ et $y_i \sim \mathcal{N}(\mu, B'B + \Psi)$. De plus,

$$\begin{aligned} Cov\left(g_i, y_i{}'\right) &= E\left[g_i y_i{}'\right] \\ &= E\left[g_i \mu' + g_i g_i{}' B + g_i \varepsilon_i{}'\right] \\ &= E\left[g_i g_i{}' B\right] \\ &= V\left[g_i\right] B \\ &= B \end{aligned}$$

On en déduit,

$$\begin{pmatrix} y_i \\ g_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} B'B + \Psi & B' \\ B & I_K \end{pmatrix}\right).$$

Or, nous savons que si deux variables X_1 and X_2 sont distribuées suivant une loi gaussienne telle que, $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$
où, μ_1 $(r \times 1)$, μ_2 $(s \times 1)$, Σ_{11} $(r \times r)$, Σ_{12} $(r \times s)$, Σ_{21} $(s \times r)$ et Σ_{22} $(s \times s)$;
alors,

$$(X_1|X_2 = x_2) \sim \mathcal{N}\left(M = \mu_1 + \Sigma_{12}{\Sigma_{22}}^{-1}(x_2 - \mu_2), \phi = \Sigma_{11} - \Sigma_{12}{\Sigma_{22}}^{-1}\Sigma_{21}\right) \tag{1.3}$$

Par conséquent, la distribution de y_i conditionnelle à g_i est de la forme $\mathcal{N}(\mu + B'g_i, \Psi)$.
Puisque $y_i|g_i$ est un vecteur gaussien de longueur q alors pour toute variable $j \in [\![1, q]\!]$, $y_i^j|g_i$

sont indépendantes si et seulement si elles sont non corrélées. Or, Ψ est diagonale, la propriété précédente (1.2) se caractérise donc par :

$$\prod_{i=1}^{n} p\left(y_i|g_i;\theta\right) = \prod_{i=1}^{n}\prod_{j=1}^{q} p\left(y_i^j|g_i;\theta\right)$$

Écriture des modèles à facteurs par condition sur matrice de variance-covariance

Dans un cadre plus général que celui du cas gaussien, d'après la définition de Knott and Bartholomew (1999),

Définition 1. On dit que $Y = [y^{1\prime}, \ldots, y^{q\prime}]$ est un vecteur aléatoire vérifiant un modèle à K facteurs, si et seulement s'il existe un vecteur aléatoire g à valeur dans $\mathbb{R}^K$ tel que, conditionnellement à g, les variables aléatoires $y^1, \ldots, y^q$ soient indépendantes.
Pour deux éléments quelconques y_i^j et y_i^l de y_i, les moments sont :

$$\begin{cases} Var(y_i^j|\mu, B, g, \Psi) &= \sigma_j^2 \\ Var(y_i^j|\mu, B, \Psi) &= \sum_{k=1}^{K} {b_k^j}^2 + \sigma_j^2 \quad \forall j \end{cases}$$

D'autre part on a :

$$\begin{cases} Cov(y_i^j, y_i^l|\mu, B, g, \Psi) &= 0 \\ Cov(y_i^j, y_i^l|\mu, B, \Psi) &= \sum_{k=1}^{K} b_k^j b_k^l, \quad \forall j, l \; ; \quad j \neq l. \end{cases}$$

En se fondant sur ces propriétés, il devient possible d'écrire autrement le modèle à K facteurs par condition sur la matrice de variance-covariance Σ^Y mise sous la forme :

$$\Sigma^Y = B'B + \Psi. \tag{1.4}$$

1.1.2 Les modèles à équations structurelles

Les modèles à équations structurelles ont été en partie développés par Bollen (2014) et Kaplan (2008). À travers les modèles à équations structurelles, on conjecture que l'ensemble des relations de cause à effet représente la complexité du phénomène que l'on cherche à étudier ou comprendre. Cette complexité est structurée en une complexité interne et externe. La complexité interne (resp. externe) du phénomène correspond à celle des liens entre les variables latentes (resp. entre les variables observables et latentes) du modèle. Si chaque relation de cause à effet est représentée par une flèche et chaque variable par un symbole, un schéma permet d'illustrer la structure de la complexité du phénomène étudié.
Pour les modèles d'équations structurelles il existe un formalisme permettant leur illustration sous forme de diagramme. La convention suivante est classique et adaptée : les ellipses correspondent aux variables latentes et les rectangles aux variables observables. Les flèches représentent les relations entre les variables et leur sens indique celui de l'action de l'une sur l'autre. En outre, chaque modèle est subdivisé en deux sous-modèles : l'un est le modèle de mesure, constitué des équations de mesure et le second est le modèle structurel, constitué des équations structurelles.

Le sous-modèle de mesure (équations de mesure)

Les équations de mesures décrivent la manière dont les variables observables sont liées aux variables latentes. Trois types de modèles de mesures sont envisagés classiquement (cf. 1.1) et sont schématisés comme suit :

— **mode formatif** : chaque Variable Observable (VO) représente une dimension de la Variable Latente (VL), dite alors "exogène". Elle est alors décrite comme la combinaison linéaire des VO. La VL est alors contrainte à être une composante.

— **mode réflectif** : toutes les VO unidimensionnelles contribuent à mesurer une seule VL qui est alors dite "endogène". Cela reproduit le cas des modèles d'analyse factorielle dans lesquels chaque variable est fonction d'un facteur. Ce qui est moins contraignant que le cas formatif.

— **mode mixte** : un mélange du mode formatif et du mode réflectif.

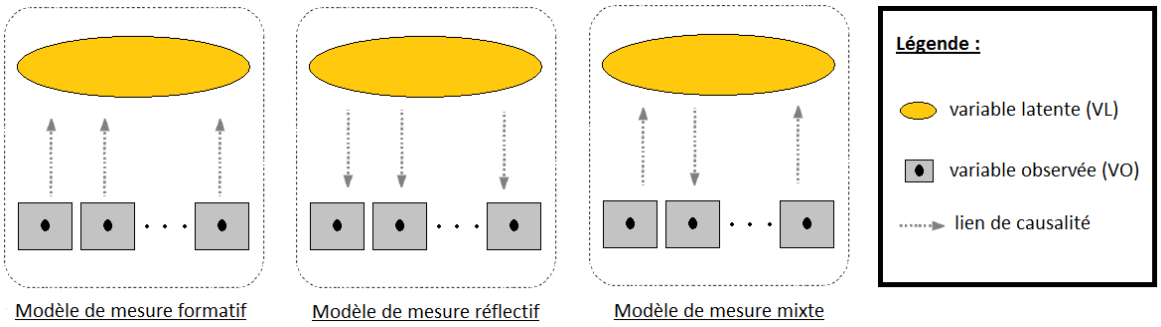


FIGURE 1.1 – Les trois différents types de modèles de mesure.

Par exemple, le modèle de mesure réflectif schématisé plus haut peut être formalisé mathématiquement par l'équation suivante :

$$Y = \mathbb{1}_n \mu' + gb + \varepsilon^Y$$

où l'on note g la VL (vecteur de longueur n qui correspond au nombre d'individus), Y la matrice de toutes les VO (autant de VO de longueur n que de rectangles contenant un point), b le vecteur des coefficients scalaires de liaisons (autant que de flèches), μ' un vecteur ligne constitué des paramètres moyennes de chaque VO et ε^Y une matrice aléatoire des erreurs dont la variance de chaque ligne représente la variabilité des individus non expliquée par la VL.

Le sous-modèle structurel (équations structurelles)

Quant aux équations structurelles, elles décrivent la manière dont les VL sont liées entre elles. Leur analyse revient à celle des relations entre les VL. Il existe plusieurs natures de VL (cf. figure 1.2). Une VL est dite :

— **exogène** : lorsqu'elle ne dépend d'aucune autre VL. Schématiquement ce sera une VL à partir de laquelle toutes les flèches partent vers d'autres VL.

— **endogène** : lorsqu'elle est expliquée par d'autres VL. Schématiquement elle reçoit au moins une flèche mais peut aussi en envoyer.

Si on nomme g la VL endogène et f^1, f^2 les VL exogènes du modèle structurel schématisé ci-dessus, ce dernier peut s'écrire :

$$g = c^1 f^1 + c^2 f^2 + \varepsilon^g$$

où les scalaires c^1, c^2 sont des coefficients structurels et ε^g le vecteur des erreurs de même longueur n que les VL.

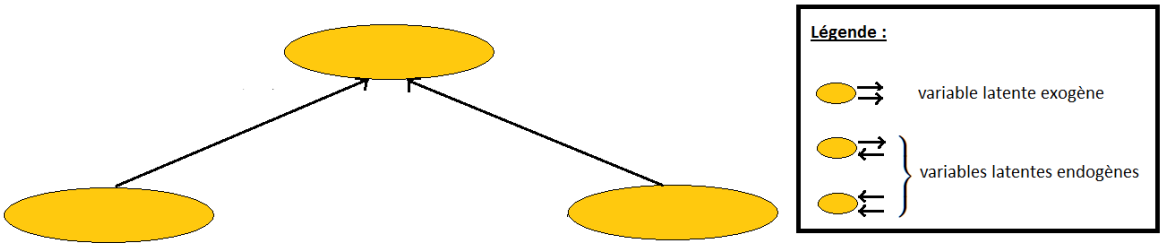


FIGURE 1.2 – Exemple de modèle structurel comportant les deux types de variables latentes : exogène et endogène.

Le modèle complet : modèle à équations structurelles et à variables latentes

Un modèle à équations structurelles à variables latentes est la fusion de deux sous-modèles : l'un de mesure et l'autre structurel. Par exemple, si l'on fusionne le modèle structurel de la section précédente avec le modèle de mesure réflectif formalisé plus haut, on obtient le modèle à une équation structurelle suivant et schématisé figure 1.3 où les VL sont des facteurs :

$$\begin{cases} Y &= \mathbb{1}_n {\mu_Y}' + gb + \varepsilon^Y \\ X^1 &= \mathbb{1}_n {\mu_1}' + f^1 {a^1}' + \varepsilon^1 \\ X^2 &= \mathbb{1}_n {\mu_2}' + f^2 {a^2}' + \varepsilon^2 \\ g &= f^1 c^1 + f^2 c^2 + \varepsilon^g \end{cases} \tag{1.5}$$

Notons que dans "modèle à équations structurelles", "équations" est au pluriel. Il peut en effet a priori contenir plusieurs équations structurelles. Le modèle que nous venons d'écrire n'en contient qu'une seule. Par la suite, sans perte de généralité nous proposerons une méthode d'estimation par algorithme EM d'un modèle à une seule équation structurelle où les VL sont des facteurs liés aux VO en mode réflectif. En effet, l'objectif de ce travail est essentiellement de présenter cette nouvelle approche d'estimation. En revanche, il est tout à fait possible d'enrichir et complexifier le modèle. D'ailleurs étendre cette approche à un modèle à plusieurs équations structurelles est une des perspectives de ce travail.

1.1.3 Les modèles à équations structurelles à facteurs

Il s'agit de modèles à équations structurelles dont les VL sont des facteurs. Classiquement, on cherche à estimer les paramètres du modèle, notamment ses coefficients. Or, ici la présence de VL pose aussi le problème de leur propre estimation. Dans la littérature, il existe deux grandes familles de méthodes : les méthodes de type PLS, qui supposent que les VL sont contraintes à la nature de composantes (i.e : chaque VL est la combinaison linéaire des VO auxquelles elle est liée) et LISREL reposant sur la seule hypothèse de normalité des VL et pour laquelle, ces dernières ne sont estimées que depuis (Jöreskog, 2000) par une procédure lourde, complexe et séparée de l'estimation des paramètres du modèle.
Ainsi, considérer que les VL sont des facteurs, est moins contraignant pour le modèle, ce qui peut inciter à travailler avec ce type de modèle. Le modèle à composantes, partant d'une hypothèse plus restrictive, possède toutefois un avantage sur le modèle à facteurs pour la prévision, par le fait que les VL y sont formulées en fonction des VO. Nous présentons les deux grandes familles d'estimation des modèles à équations structurelles : PLS et LISREL ci-après.

1.1.4 Formalisme

PLS est une approche à composantes contrairement à LISREL. Toutes les deux ont été développées dans les années 1970 mais avec un formalisme différent et sur des hypothèses différentes. Par souci de clarté, nous présenterons chacune d'elle avec ses propres notations et

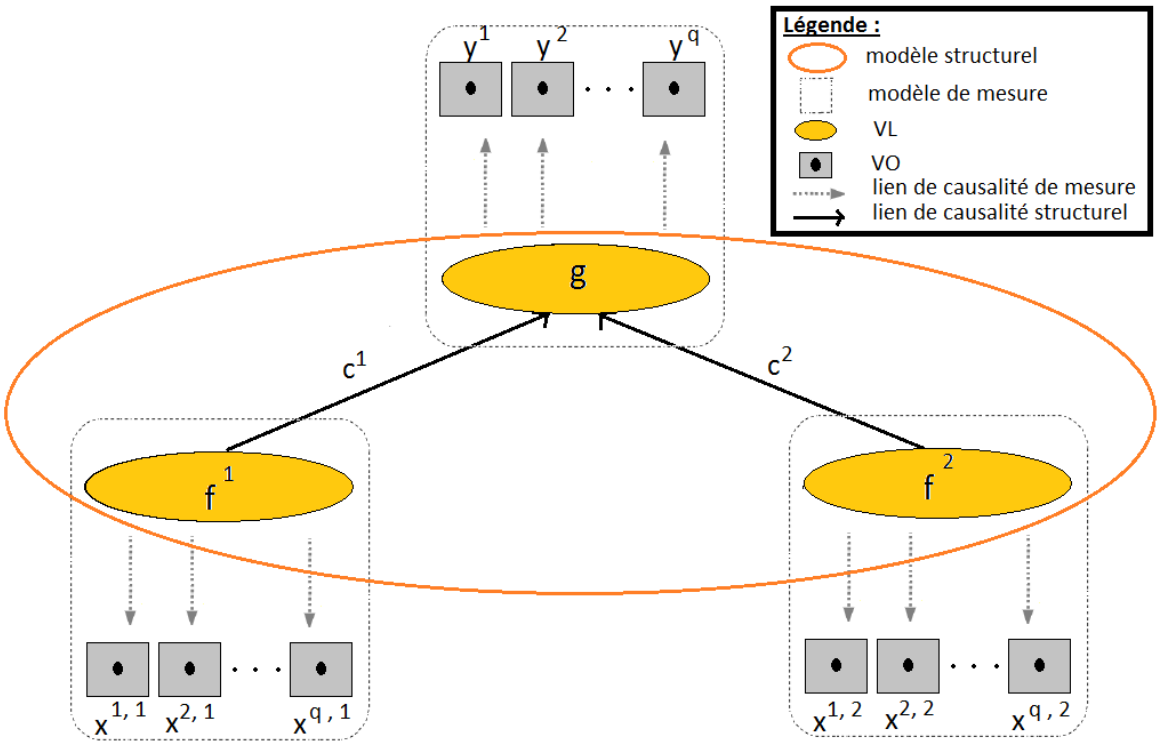


FIGURE 1.3 – Exemple de modèle à une équation structurelle à variables latentes.

ses conventions graphiques habituelles. En revanche, les notations choisies pour développer notre approche d'estimation des Modèles à Équations Structurelles (SEM) à facteurs sont du type de celles associées aux modèles à facteurs présentés dans ce chapitre. Pour faciliter la lecture des différentes approches et le passage de l'une à l'autre, nous présentons dans le tableau 1.1 suivant l'essentiel des notations propres à chacune. Des compléments de notations seront faits si nécessaire dans chacune des différentes sections présentant les approches.

Éléments du modèle	Notations propres à l'approche LISREL	Notations propres à l'approche PLS	Notations propres à l'approche EM pour les SEM à facteurs
VL (endogènes, exogènes)	(η,ξ)	(ξ,ξ)	(g,f)
VO liées aux VL (endogènes, exogènes)	(y,x)	(x,x)	$(Y,\ X)$
Termes d'erreurs	$\varepsilon,\ \delta,\ \zeta$	$\varepsilon,\ \zeta$	$\varepsilon^Y,\ \varepsilon^m,\ \varepsilon^g$
Coefficients de mesure liés aux VL (endogènes, exogènes)	$(\Lambda_y = [\lambda_y],\ \Lambda_x = [\lambda_x])$	π	$(b,\ a^m)$
Coefficients structurels liant les VL (endogènes, exogènes) aux endogènes	$(B = [\beta],\ \Gamma = [\gamma])$	$(\beta,\ \beta)$	$(-,\ c^m)$
Indice des blocs de VO liées aux VL endogènes	$-$	$k_{endo} \in [\![1, K_{endo}]\!]$	$-$
Indice des blocs de VO liées aux VL exogènes	$-$	$k_{exo} \in [\![1, K_{exo}]\!]$	$m \in [\![1, p]\!]$
Indice des VO du bloc lié à la VL endogène	$j \in [\![1, p]\!]$	$j \in [\![1, p_{k_{endo}}]\!]$	$j \in [\![1, q_Y]\!]$
Indice des VO des m blocs associés aux VL exogènes	$j \in [\![1, q]\!]$	$j \in [\![1, p_{k_{exo}}]\!]$	$j \in [\![1, q_m]\!]$
Indice des VL (endogènes, exogènes)	$(j \in [\![1, m]\!],\ j \in [\![1, n]\!])$	(k_{endo}, k_{exo})	$(-,\ -)$
Nombre d'individus	N	N	n
Nombre de VL (endogènes, exogènes)	$(m,\ n)$	$(K_{endo},\ K_{exo})$	$(1,\ p)$
Nombre de VO liées aux VL (endogènes, exogènes)	$(p,\ q)$	$\left(\sum_{k_{endo}} p_{k_{endo}},\ \sum_{k_{exo}} p_{k_{exo}}\right)$	$(q_Y,\ q_p)$
Nombre total de VL	$m + n$	K	$1 + p$
Nombre total de VO	$p + q$	$\sum_{k_{endo},\ k_{exo}} p_{k_{endo}} + p_{k_{exo}}$	$q_Y + \sum_{m=1}^{p} q_m$
L'ensemble des paramètres	Θ	Θ	θ

Tableau 1.1 – Recoupement des notations utilisées pour les différentes approches d'estimation.

1.1.5 Commentaires : lien entre modèle à facteurs et ACP

L'analyse factorielle et la méthode d'Analyse en Composantes Principales (ACP) sont liées par leur objectif de réduction des données. En effet, dans le cadre de la statistique descriptive multidimensionnelle et dans le cas de variables quantitatives, l'ACP est très utilisée en pratique pour réduire le grand nombre de variables observées en un plus petit nombre d'éléments nommés composantes. Cet objectif est atteint en combinant les variables fortement corrélées dans une même composante. On retrouve ici l'esprit de réduction du nombre de variables observées fondant l'analyse factorielle (Lawley and Maxwell, 1963). La différence entre l'ACP et l'analyse factorielle se situe au niveau de la nature des éléments synthétisant les nombreuses variables observées de départ. En effet, dans le cadre de l'ACP, ce sont des composantes, c'est à dire des combinaisons linéaires des variables observées, alors que dans le cas de l'analyse factorielle, ce sont des facteurs, sur lesquels seule une hypothèse de distribution est faite. De plus, dans les modèles factoriels, ce sont les variables observées qui sont des combinaisons linéaires des facteurs. Ainsi, la méthode descriptive ACP se distingue de l'analyse factorielle qui repose sur un modèle probabiliste. En effet, le modèle initial est semblable mais le traitement des données par ces deux approches est différent.
Avant de présenter l'ACP comme un cas particulier du modèle à facteurs, nous en présentons la théorie dans la section suivante. Pour des raisons de simplicité, on se limitera à présenter cette théorie pour la métrique identité.

Théorie de l'Analyse en Composantes Principales (ACP)

Étant donné un vecteur $Y = [y^{1\prime}, \ldots, y^{q\prime}]$ de q variables, l'objectif de l'ACP est de construire des variables $g_1, \ldots, g_K$ linéaires en Y, deux à deux non corrélées et de variance (dite aussi inertie) maximale. Avec une métrique identité [1], le critère sous-jacent se formalise comme suit,

$$\max_{\alpha'\alpha=1} \{Var(g)\} = \max \left\{ \frac{\alpha' \Sigma^Y \alpha}{\alpha'\alpha} \right\} \tag{1.6}$$

où, $g = Y\alpha = \sum_{j=1}^{q} \alpha_j y^j \in \mathbb{R}^n$, $\alpha \in \mathbb{R}^q$ et Σ^Y la matrice de variance-covariance de Y. Par satisfaction de ce critère, les représentations graphiques des données dans un espace de dimension q seront réduites par une projection de celles-ci dans un plus petit espace de dimension K de la manière la plus fidèle. Ces représentations graphiques seront alors plus commodes à interpréter.
La première solution α_1 de (1) est fourni par une diagonalisation Σ^Y. C'est à dire par la résolution de $\Sigma^Y \alpha_1 = \lambda_1 \alpha_1$ avec ${\alpha_1}'\alpha_1 = 1$ et $Var(g_1) = \lambda_1$. En effet, si
$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K \geq \ldots \geq \lambda_q$ sont les valeurs propres de Σ^Y, on a :

$$\forall k \in [\![1, K]\!], \quad Var(g_k) = Var(Y\alpha_k) = \lambda_k$$

α_1 sera l'axe principal d'inertie mais aussi le vecteur propre (normé [2] à l'unité au sens ici de la métrique identité) associé à la plus grande valeur propre obtenue λ_1 par la diagonalisation de $\Sigma^Y I_q$. Une fois que le vecteur α_1 est obtenu, la première composante principale [3] est
$g_1 = \frac{1}{\sqrt{\lambda_1}} Y\alpha_1$. Les autres composantes principales vont être cherchées orthogonales aux précédentes de façon analogue. [4]

1. Pour une métrique quelconque Q, ce critère s'écrit :

$$\max_{\alpha'Q\alpha=1} \{Var(g)\} = \max \left\{ \frac{\alpha' Q \Sigma^Y Q \alpha}{\alpha' Q \alpha} \right\}$$

2. Dans le cas général, normé au sens de la métrique Q choisie et par la diagonalisation de $\Sigma^Y Q$.
3. La première composante principale est dans le cas de la métrique Q : $g_1 = \frac{1}{\sqrt{\lambda_1}} YQ\alpha_1$.
4. Les composantes sont normées à la valeur propre associée car on veut que la distance entre deux observations dans le plan, soit la meilleure approximation de la vraie distance.

Pour satisfaire ce critère, on construit d'abord $g_1 = Y\alpha_1$ de variance maximale sous la contrainte ${\alpha_1}'\alpha_1 = 1$, puis $g_2 = Y\alpha_2$ de variance maximale sous la contrainte ${\alpha_2}'\alpha_2 = 1$ et $Cov(g_1, g_2) = 0$ et, de façon générale $g_k = Y\alpha_k$ de variance maximale sous la contrainte ${\alpha_k}'\alpha_k = 1$ et $Cov(g_l, g_k) = 0 \quad \forall l < k$ et $k \in [\![1, K]\!]$ où K est le nombre de composantes principales retenues. K est le nombre de plus grandes valeurs propres $\lambda_1, \ldots, \lambda_K$ de Σ^Y tel que $\sum_{j=1}^{q} Var(y^j) - \sum_{k=1}^{K} Var(g_k) \leq \epsilon$ avec ϵ fixé d'avance et proche de zéro. On dit alors que le choix de K est fait de manière à ce que la part de la variance expliquée par les facteurs $g_1, \ldots, g_K$ soit grande et la plus proche possible de celle des données originelles. Ce qui s'écrit également :

$$tr(\Sigma^Y) - \sum_{k=1}^{K} \lambda_k \leq \epsilon \;\; tel\ que \;\; \sum_{k=K+1}^{q} \lambda_k \leq \epsilon$$

Ainsi, l'ACP peut être vue comme une approche cherchant à approximer la matrice Σ^Y de rang q par une matrice de rang K plus petit.
En pratique, K, le nombre de composantes retenues est déterminé par l'utilisation de trois règles empiriques alternatives :

- K est choisi tel que le pourcentage de l'inertie expliquée ou cumulée $tr(\Sigma^Y) = \sum_{k=1}^{K} \lambda_k$ est proche de 100 % ;
- K correspond au nombre de bandes de l'histogramme des valeurs propres avant la première rupture. On parle de rupture dans l'éboulis ;
- K correspond à l'indice k de la dernière valeur propre λ_k supérieure à 1, dans le cas de données centrées réduites d'après le critère de Kaiser.

Suivant la règle, la valeur de K peut être différente. Il faut alors faire un choix.
Pour détailler ce qui fait la différence entre ACP et estimation d'un modèle à facteurs, nous allons compléter par quelques notions la théorie de l'approche ACP présentée dans la section précédente.

Notions théoriques supplémentaires sur l'ACP

D'après la section précédente, si on note $A = [\alpha_1, \ldots, \alpha_q]$ la matrice orthogonale dont les colonnes sont constituées par une base orthonormée de vecteurs propres de Σ^Y, on peut écrire,

$$\Sigma^Y = ALA' \tag{1.7}$$

avec $L = diag(\lambda_1, \ldots, \lambda_q)$.
Introduisons quelques notations nécessaires à la comparaison. On considère,
$A_1 = [\alpha_1, \ldots, \alpha_K]$, $A_2 = [\alpha_{K+1}, \ldots, \alpha_q]$ et $L = diag(L_1, L_2)$ où $L_1 = diag(\lambda_1, \ldots, \lambda_K)$ et $L_2 = diag(\lambda_{K+1}, \ldots, \lambda_q)$. Alors (1.7) peut se décomposer comme suit,

$$\Sigma^Y = A_1 L_1 A_1' + A_2 L_2 A_2'.$$

Si à nouveau on note, $\Lambda = A_1 L_1^{\frac{1}{2}}$ et $D_2 = A_2 L_2 {A_2}'$, (1.7) peut également se décomposer comme suit,

$$\Sigma^Y = \Lambda\Lambda' + D_2. \tag{1.8}$$

On retrouve alors une écriture similaire à celle de la structure de la matrice de variance-covariance (1.4) du modèle à facteurs. Nous pouvons aussi remarquer que si les $(q - K)$ dernières valeurs propres sont proches de zéro (c'est à dire $tr(L_2) < \epsilon$, on peut approximer Σ^Y par $A_1 L_1 A_1'$. Il devient alors facile de voir la liaison entre l'approche ACP et le modèle à facteurs. En effet, si on note $g = [g_1, \ldots, g_K]' = A_1' Y$ le vecteur des K premières composantes et $\Lambda = A_1 L_1^{\frac{1}{2}}$ et $\phi = L_1^{-\frac{1}{2}} g$, on obtient la décomposition de Y sous la forme d'un modèle à facteurs :

$$Y = A_1 g + u = \Lambda\phi + u \tag{1.9}$$

où, $u = Y - A_1 A_1' Y$, ϕ a le rôle de facteur et correspond à g pour le modèle (1.1) (Λ correspond à B, u à ε et le paramètre de moyenne est nul). De plus, il est facile de retrouver les hypothèses du modèle à facteurs.

— L'hypothèse du modèle à facteur $Var(g) = I$, se retrouve par,

$$\begin{aligned} Var(\phi) &= L_1^{-\frac{1}{2}} Var(g) L_1^{-\frac{1}{2}} \\ &= L_1^{-\frac{1}{2}} L_1 L_1^{-\frac{1}{2}} \quad \textit{selon la diagonalisation de } \Sigma^Y \\ &= I. \end{aligned}$$

— L'hypothèse du modèle à facteur $Cov(\epsilon, g') = 0$, se retrouve comme suit, où pour des raisons de simplification de notations et de calculs nous ferons l'hypothèse que $E(Y) = 0$:

$$\begin{aligned} Cov\left(u, \phi'\right) &= E\left(u\phi'\right) - E(u) E\left(\phi'\right) \\ &= E\left(u\phi'\right) \\ &= E\left(A_2 A_2' Y Y' A_1 L_1^{-\frac{1}{2}}\right) \\ &= A_2 A_2' \Sigma^Y A_1 L_1^{-\frac{1}{2}} \\ &= 0. \end{aligned}$$

— L'hypothèse du modèle à facteur $Var(\epsilon) = \Psi$, se retrouve par :

$$\begin{aligned} Var(u) &= Var(Y - A_1 A_1' Y) \\ &= Var(A_2 A_2' Y) \\ &= A_2 A_2' \Sigma^Y \left(A_2 A_2'\right)' \\ &= A_2 L_2 A_2' \\ &= D_2. \end{aligned}$$

en effet, D_2 correspond à Ψ dans la décomposition de la matrice de variance Σ^Y dans (1.8).

La méthode ACP : un cas particulier du modèle à facteurs

Dans le cadre de l'ACP, la décomposition de Y sous la forme d'un modèle à facteurs (1.9) montre que les deux approches (ACP et analyse factorielle) ont un même modèle initial. La vision de l'ACP comme cas particulier du modèle à facteur devient explicite lorsque nous supposons que les variables observées vérifient un modèle à K facteurs, ce qui revient à ce que $\Sigma^Y = BB' + \Psi$ avec B de dimension $(q \times K)$ de rang K et Ψ diagonale définie positive. Si la matrice Ψ était connue, on aurait $\Sigma^Y - \Psi = BB'$ et la matrice $\Sigma^Y - \Psi$ serait de rang K et admettrait $(q - K)$ valeurs propres nulles. Alors, la diagonalisation $\Sigma^Y - \Psi$ fournirait exactement la matrice B. En effet, on aurait,

$$\Sigma^Y - \Psi = ALA' \tag{1.10}$$

où, $L = diag(L_1, 0_{(q-K)\times(q-K)})$. Ainsi, $\Sigma^Y - \Psi = A_1 L_1 A_1' = (A_1 L_1^{\frac{1}{2}})(L_1^{\frac{1}{2}} A_1) = \Lambda\Lambda'$. De plus, lorsque Σ^Y est la matrice des corrélations, $\Sigma^Y - \Psi$ est dite matrice des corrélations réduites. Par conséquent, on en conclut que l'analyse factorielle est équivalente à une ACP sur la matrice des corrélations réduites.
Ainsi, en pratique, il est recommandé de démarrer l'estimation d'un modèle à facteurs, par une ACP. Cela permet d'obtenir une estimation du nombre de facteurs à retenir. Ces approches s'avèrent donc complémentaires en pratique.

La différence entre l'ACP et l'analyse factorielle se situe au niveau de la décomposition de la matrice de variance-covariance Σ^Y. Les contraintes sont différentes selon les approches. Pour l'ACP, lors de la décomposition $\Sigma^Y = \Lambda\Lambda' + D_2$, la matrice Σ^Y de rang q est approximée par $\Lambda\Lambda'$ avec Λ de rang K et D_2 de rang $(q-K)$, alors que pour l'analyse factorielle, la décomposition $\Sigma^Y = BB' + \Psi$ est telle que le rang de Ψ est $q \neq q-K$. Ainsi, bien que ces deux approches ont un même modèle initial et que lorsque les variables vérifient un modèle à K facteurs des similitudes se retrouvent, il existe une différence de traitement des données. De plus, l'ACP est descriptive alors que l'analyse factorielle est probabiliste et procède à une estimation du maximum de vraisemblance.

1.2 Les méthodes d'estimation des modèles à équations structurelles et variables latentes

Il existe deux grandes familles de méthodes : PLS développée par Wold (1966) et LISREL issue des travaux de Jöreskog (1967, 1969). Pour chacune d'elle, les notations sont différentes. En effet, dans la littérature les notations relatives aux VL, VO et paramètres ne sont pas fixées. Nous faisons le choix de présenter chacune des approches dans des sections indépendantes selon leur propre formalisme mathématique. Pour faciliter la compréhension, nous allons introduire en complément du tableau 1.1 quelques notations supplémentaires.
L'objectif de ces méthodes est d'estimer les paramètres du modèle : coefficients de mesure, coefficients structuraux mais aussi les matrices de variances-covariances pour l'approche LISREL. Pour y parvenir, des méthodes numériques ont été développées et implémentées. La présentation de chacune des approches PLS et LISREL est accompagnée des principaux algorithmes associés. Nous commençons par l'approche PLS qui est plus simple.

1.2.1 Modèle à composantes : l'approche de Wold (PLS)

L'approche PLS (Partial Least Squares) a été introduite par Wold (1966, 1982). Ses travaux de recherche ont d'abord traité de l'ACP avec la présentation de l'algorithme NILES (*Nonlinear estimation by Iterative LEast Squares*, Wold (1966)). Ensuite, il a proposé l'algorithme NIPALS (*Nonlinear Iterative PArtial LEast Squares*, Wold (1973)). Puis ses travaux aboutirent à l'approche PLS (Wold, 1985; Lohmöller, 2013), issue de l'estimation des moindres carrés. Elle est fondée sur des régressions simples et multiples. Son avantage est qu'elle nécessite peu d'hypothèses.

Écriture du modèle général et de ses hypothèses dans la littérature

Notons x_k le k-ème bloc de VO de dimension $n \times p_k$ et ξ_k la k-ème VL de dimension $n \times 1$ (cf. 1.1). L'écriture du modèle dans la littérature est la suivante :

$$\begin{cases} x_{kj} = \pi_{kj}\xi_k + \epsilon_{kj} & \text{(1.11a)} \\ \xi_k = \sum_{i:\xi_i \to \xi_k} \beta_{ik}\xi_i + \zeta_k & \text{(1.11b)} \end{cases}$$

où, (1.11a) et (1.11b) sont respectivement le sous-modèle de mesure et le sous-modèle structurel, $\xi_i \to \xi_k$ signifie que ξ_i fait partie des déterminants (variable explicative) de ξ_k ; π représente les coefficients de mesure (nommés aussi "loadings") ; β les coefficients structurels et ϵ et ζ les erreurs de mesures.
Les hypothèses de ce modèle sont :

- $Corr(\epsilon_{kj}, \zeta_k) = 0$, $\forall$ i et k ;
- $Corr(\xi_i, \zeta_k) = 0$, $\forall$ i tel que $i \neq k$;
- $Corr(\epsilon_{kj}, \xi_k) = 0$, $\forall$ j et k ;
- $Corr(\epsilon_{kj}, \epsilon_{lm}) = 0$, $\forall$ k, j, l, m tels que $(k, j) \neq (l, m)$.

Illustration graphique

Par exemple, selon les notations et les conventions graphiques de PLS, le schéma 1.3 donne la figure 1.4 et se modélise :

$$\begin{cases} x_{kj} = \pi_{kj}\xi_k + \epsilon_{kj}, \quad \forall k \in [\![1, 3]\!] \; et \; \forall j \in [\![1, q]\!] \\ \xi_3 = \beta_{13}\xi_1 + \beta_{23}\xi_2 + \zeta_3 \end{cases}$$

où, ξ_1, ξ_2 et ξ_3 correspondent respectivement à f^1, f^2 et g ; β_{13} et β_{23} à respectivement c^1 et c^2 ; $\forall j$ π_{1j}, π_{2j} et π_{3j} correspondent respectivement à a^1, a^2 et b ; ζ_3 correspond à ϵ_g et $\forall j$, les variables x_{1j}, x_{2j}, x_{3j} correspondent respectivement aux variables X^1, X^2, Y et ϵ_{1q}, ϵ_{2q} correspondent respectivement aux erreurs de mesures ϵ^1, ϵ^2.

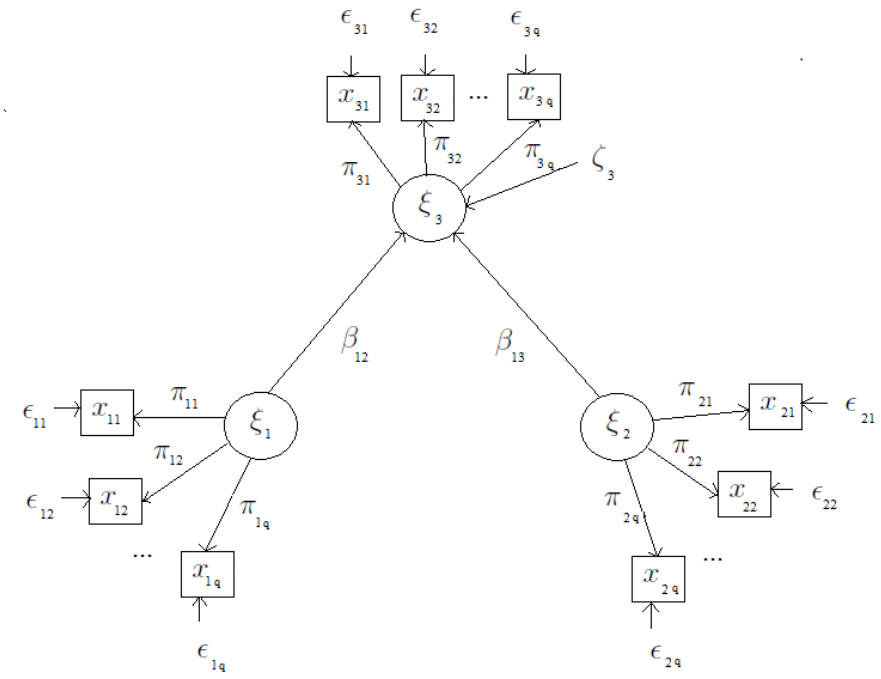


FIGURE 1.4 – Diagramme du modèle (1.5) pour l'approche PLS.

L'algorithme PLS

En pratique, l'approche PLS résout les équations du modèle de manière itérative. Elle utilise un algorithme qui estime d'une part les VL et d'autre part les coefficients du modèle. Après initialisation de poids externes w_{kj}, il alterne la construction des VL, en se basant tantôt sur le modèle de mesure (externe) et tantôt sur le modèle structurel (interne). Dans l'estimation externe, les VL standardisées sont notées y_k et estimées comme combinaison linéaire de leurs VO centrées i.e. : $y_k \propto \pm x_k w_k$, où $\propto$ signifie que le terme de gauche est égal au terme de droite standardisé et $\pm$ que le signe est choisi tel que y_k est positivement corrélée au maximum de variables de x_k. L'estimation interne, quant à elle, tient compte des relations de causalité qui lient les VL dont la notation devient z_k et les approxime par $z_k = \sum_{j=1}^{K} c_{kj} e_{kj} y_j$, où $c_{kj} = 1$ si ξ_k et ξ_j ont un lien de causalité et 0 sinon. Et $e_{kj} = corr(y_k, y_j)$ si la relation de causalité va de ξ_k vers ξ_j, sinon lorsque la relation de causalité va de ξ_j vers ξ_k, e_{kj} est égal au coefficient de régression, de la régression de y_k sur y_j. Lorsque la convergence est atteinte, les coefficients du modèle sont estimés par régression simple ou multiple (OLS) suivant le nombre de VL et celles-ci sont remplacées par leurs estimations. Dans cette approche, les VL sont estimées par des composantes construites à partir des VO qui leur sont respectivement associées, ce qui fait qu'elles ne satisfont plus la définition classique d'une VL (Bollen, 2014).

La procédure de l'algorithme PLS

L'algorithme PLS procède comme suit :

1. Initialisation des poids externes w_{kj}.
2. Estimation externe fondée sur le modèle de mesure des VL ξ_k :
   Les VL standardisées sont approximées par une combinaison linéaire y_k de leurs VO centrées x_k telle que :
   $$y_k \propto \pm x_k w_k. \tag{1.13}$$
3. Estimation interne fondée sur le modèle structurel des VL ξ_k :
   Les variables standardisées sont approximées à nouveau par z_k telle que :
   $$z_k = \sum_{j=1}^{K} c_{kj} e_{kj} y_j \tag{1.14}$$
4. Actualisation des poids externes :
   Il existe deux façons de faire nommées **mode A** et **mode B**. Le mode A qui correspond au schéma de mesure réflectif (Wold, 1982) est le plus utilisé dans la littérature. Il propose d'actualiser les poids externes par :
   $$w_k = \frac{1}{{z_k}' z_k} {x_k}' z_k. \tag{1.15}$$
   Le mode B correspond au schéma de mesure formatif (Wold, 1982) et actualise les poids externes par :
   $$w_k = ({x_k}' x_k)^{-1} {x_k}' z_k \tag{1.16}$$
   sous la contrainte ${w_k}' {x_k}' x_k w_k = N$. On reconnait la solution de régression de $z_k = x_k w_k$.
5. Si la convergence est atteinte on passe à la dernière étape 6. Sinon, on retourne à la deuxième étape de l'algorithme.
6. Après convergence, les coefficients structurels $\hat{\beta}$ sont calculés par régression simple ou multiple suivant le nombre de VL du modèle.

L'algorithme de l'approche PLS présenté est plus précisément celui de l'approche dite PLS Path-Modeling (PLSPM) détaillée dans Lohmöller (2013).
Pour chacun des deux modes A et B, il existe trois variantes (ou "schémas") nommées "centroïde" (cf. figure 1.5), "factorielle" et "structurelle". Elles correspondent à de légères modifications des expressions (1.15) et (1.16) actualisant les poids externes. Cela conduit à six stratégies possibles de détermination des VL dans l'approche PLS. Pour chacune de ces six stratégies, le calcul des poids externes w_k est réalisée par une procédure itérative proposée par Wold et détaillée dans Lohmöller (2013).
Usuellement, lors de l'étape d'initialisation, les poids externes sont choisis tel que $w_{kj} = sign(corr(x_{kj}, y_k))$ si $k = 1$ et 0 sinon. Puis, les poids externes sont standardisés de manière à ce que la variance des VL soit de 1. Cette procédure n'est pas unique, Tenenhaus et al. (2005) propose une autre méthode qui consiste à prendre pour chaque bloc, les éléments du premier vecteur propre issu de l'ACP avec une majorité de signes positifs. Dans le cas où il y a égalité entre les signes positifs et négatifs, la variable avec la plus grande corrélation en valeur absolue prend le signe positif.
Il existe aussi plusieurs variante à la procédure PLS présentée. Par exemple lors de l'étape 3, Lohmöller (2013) propose de remplacer les termes $c_{kj} e_{kj}$ par les coefficients de régression de y_k sur les y_j qui en sont explicatifs.

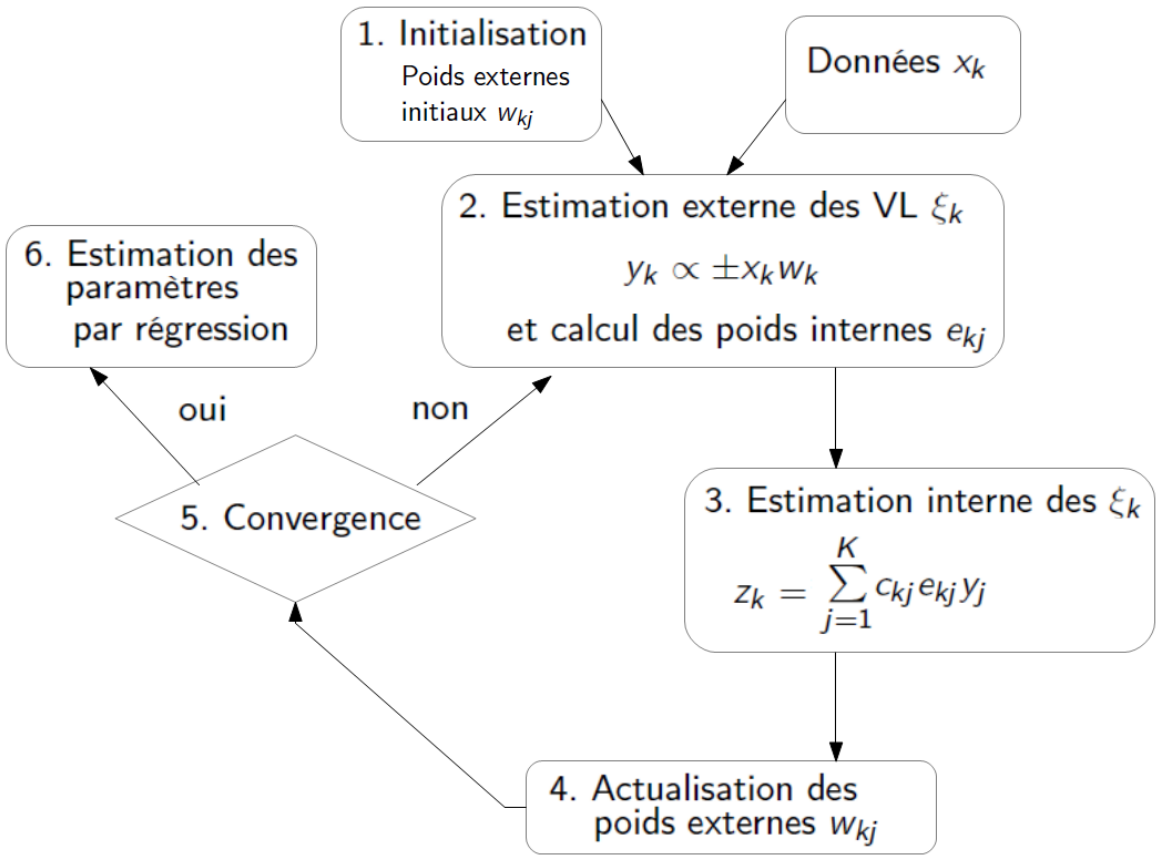


FIGURE 1.5 – Schéma de l'approche PLS

Propriétés de l'algorithme PLS

Actuellement, peu de propriétés ont été démontrées mais en pratique, plusieurs d'entre elles sont supposées car généralement observées. Par exemple, la propriété de convergence avec une probabilité 1 de l'algorithme PLS a été démontrée par Lyttkens et al. (1975) pour un nombre de bloc de variables $K \leq 2$. À partir de trois blocs, la convergence est seulement observée en pratique. D'autres chercheurs se sont par la suite penchés sur les propriétés de PLS et ont cherché à en démontrer. Mais beaucoup d'entre elles sont restées au stade de la conjecture, en voici quelques unes non exhaustives :

Conjecture 1. Les estimations des coefficients du modèle sont consistantes au sens large (Wold, 1982).
Cette conjecture a été étudiée par Dijkstra (1981) qui a montré que les valeurs estimées sont proches de la valeur réelle lorsque le nombre de VO par bloc et le nombre d'unités statistiques tendent vers l'infini. Cependant, il mit en évidence que rien ne pouvait prouver qu'elles tendaient vers cette valeur à la limite. Un exemple d'analyse de sensibilité des estimations des paramètres de la méthode PLS a été réalisé par Jakobowicz (2007) et illustre cette conjecture.

Les critères suivants illustrent que contrairement à l'algorithme de l'approche PLSPM, dans certains cas, PLS consiste en l'optimisation d'une fonction à travers un critère à satisfaire. En fonction du nombre de blocs de variables les critères sont différents. Pour les cas de plus de deux blocs, nous avons les deux critères suivants :

Critère 1. Pour l'actualisation des poids externes par le mode B, sous le choix du schéma centroïde, les VL sont obtenues en maximisant le critère :

$$\sum_{k,l:\xi_l \leftrightarrow \xi_k} |corr\,(x_k w_k, x_l w_l)\,|. \tag{1.17}$$

où, $x_k w_k$ et $x_l w_l$ correspondent aux deux composantes associées aux blocs de VO x_k et x_l et, $\xi_l \leftrightarrow \xi_k$ signifie l'ensemble des relations de causalité entre les VL ξ. C'est à dire, l'ensemble des relations où pour k et l fixés :

— soit, ξ_l est une VL explicative de ξ_k $(\xi_l \rightarrow \xi_k)$;

— soit, ξ_l est une VL dépendante de ξ_k $(\xi_k \rightarrow \xi_l)$.

Critère 2. Pour l'actualisation des poids externes par le mode B, sous le choix du schéma factoriel, les VL sont obtenues en maximisant le critère :

$$\sum_{k,l:\xi_l \leftrightarrow \xi_k} corr^2 \left(x_k w_k, x_l w_l\right). \tag{1.18}$$

Pour le cas de deux blocs de variables, on peut se référer aux travaux de Tenenhaus (1999). Néanmoins, des propriétés de PLS ont été démontrées. En voici quelques unes que nous ne développerons pas :

Propriété 1. Au niveau de l'estimation du modèle interne, les variantes "centroïde" et "factorielle" traitent de façon symétrique l'ensemble des VL contrairement à la variante structurelle qui différencie les VL explicatives des VL dépendantes.

Propriété 2. L'estimation du modèle est sensible "à la mise à l'échelle" (i.e : au nombre de VO) uniquement lors de l'utilisation du mode A.
Cette propriété a été démontrée par Dijkstra (1981).

Propriété 3. Les estimations obtenues par les modes A et B sont similaires lorsque la matrice traitée est orthogonale.
Cela se justifie par le fait que le mode A est fondé sur des régressions simples et le mode B sur des régressions multiples. Lorsque la matrice est orthogonale, peu importe que les régressions soient simples ou multiples. Cela revient au même.
D'après Wold (1982) le choix du mode est suscité par la logique du modèle de mesure :

— le mode A correspond au modèle de mesure réflectif ;

— le mode B correspond au modèle de mesure formatif.

Mais cette formalisation n'est fondée sur aucune démonstration et est discutée par (Jakobowicz, 2007). Selon lui, la différence entre les modes n'est pas associée au choix du schéma mais à l'objet d'étude. Il propose de motiver le choix du mode par les points suivants :

— On choisit le mode A si on veut donner un plus grand poids au modèle externe ou si on a beaucoup de VO par bloc. Car l'application du mode A sur un seul bloc de VO revient à faire une ACP. Ainsi, lors de l'estimation avec le choix du mode A, le modèle externe sera favorisé. En outre, lorsqu'il y a beaucoup de VO dans un bloc, il est préférable de choisir le mode A car avec le mode B, les risques de multicolinéarité sont plus élevés.

— On choisit le mode B si on veut donner un plus grand poids au modèle interne. Car pour le cas de deux blocs de VO, le mode B revient à faire une analyse canonique qui maximise les corrélations entre les deux facteurs. Ainsi, dans le cas d'une modélisation comportant plusieurs blocs de VO, si le mode B est choisit lors de l'estimation, le modèle interne est favorisé.

Enfin, pour une revue plus complète de la littérature concernant l'algorithme PLS et sa convergence, les travaux de Henseler (2010), Krämer (2005) et Hanafi (2004) peuvent être consultés.

Derniers développements liés à PLS

L'approche PLS étant fondée sur des régressions ordinaires, elle peut être appliquée au delà de VO continues. Il arrive que les praticiens l'utilisent pour traiter des VO binaires ou des VO catégorielles ordonnées. Pour cela, ils s'autorisent à les supposer continues puisque la méthode PLS n'est fondée sur aucune hypothèse de distribution des données. Néanmoins, dans le cas de variables catégorielles à peu de modalités ou non ordonnées, il n'est pas possible de supposer leurs distributions continues, alors l'approche PLS ne pourra être appliquée.
Cependant, dans le cas de données dont on ne peut faire l'hypothèse de distribution continue, plusieurs adaptations de l'approche PLS ont été proposées tel que Partial Maximum Likelihood de (Derquenne, 2005; Jakobowicz and Derquenne, 2007).
Les travaux de Betzin and Henseler (2005) présentent dans le cadre de la méthode PLS, l'utilisation des moindres carrés alternés pour "quantifier" les variables catégorielles. Et dans la même ligne directrice, Jakobowicz (2007) a travaillé sur la question de la non linéarité dans les modèles à équations structurelles.
D'autres techniques de *path modeling* ont été développées. L'approche ACT (*Analyse en composante thématique*) (Bry, 2003) est une variante de PLSPM (Lohmöller, 2013) : elle ne cherche qu'une composante par bloc de VO. L'avantage de l'ACT est qu'elle s'affranchit des hypothèses sur le sens des liaisons entre les VL des groupes explicatifs et le groupe dépendant. Elle fournit une base de facteurs dans chaque groupe. Jusqu'alors PLSPM et l'ACT n'optimisent pas de critère. Deux méthodes plus récentes sont fondées sur l'optimisation d'un critère : RGCCA (Tenenhaus and Tenenhaus, 2011) et THEME (Bry and Verron, 2015). Elles consistent en la régression linéaire multivariée multi-blocs sur composantes.
RGCCA est elle aussi une technique de *path modeling* qui estime qu'une composante par bloc. Elle cherche des composantes corrélées dont les liaisons peuvent être des relations bivariées, symétriques. Au niveau des diagrammes cela va se traduire par des flèches à double sens. Contrairement à PLSPM, elle optimise un critère :

$$\max_{w_1,\ldots,w_K} \sum_{k,l=1,k\neq l}^{K} c_{kl} g\left(corr\left(x_k w_k, x_l w_l\right)\right) \tag{1.19}$$

où,

- $c_{kl} = 1$ si x_k et x_l sont liées et 0 sinon ;
- g est une fonction pouvant être de la forme :
  - $g(x) = x$ et on retrouve l'approche PLS ;
  - $g(x) = |x|$ et on retrouve la variante centroïde de PLS (Wold, 1985) ;
  - $g(x) = x^2$ et on retrouve la variante factorielle de PLS (Lohmöller, 2013).

Elle présente donc l'avantage d'optimiser un critère et de réunir différentes variantes de l'approche PLS.
THEME (Bry and Verron, 2015) est quant à elle une technique où les liaisons sont partielles et non globales (comme les correlations et covariances). (Bry and Verron, 2015) montre qu'il est possible d'avoir une absence de corrélation entre deux variables ayant pourtant une influence l'une sur l'autre lorsqu'une troisième entre en jeu. Les avantages de THEME sont qu'elle permet de :

- séparer les effets explicatifs pour un même groupe dépendant ;
- extraire plusieurs composantes par bloc dans un ordre hiérarchique.

Le modèle "thématique" est donc plus général que celui de RGCCA.

1.2.2 Modèles à facteurs : l'approche de Jöreskog (LISREL)

Cette approche est fondée sur l'analyse de la structure de la matrice de variance-covariance du modèle (cf. (1.4) section 2.2.2). Depuis les années 1940, plusieurs méthodes d'estimation

fondées sur l'analyse de corrélation canonique ont été développées par Lawley (1940, 1942, 1943, 1967); Rao (1955); Howe (1955); Bargmann (1957) sur lesquelles Jöreskog s'est en partie appuyé pour construire son approche. La méthode de Jöreskog utilise un système d'équations structurelles et se concentre sur l'estimation de la matrice de variance-covariance. Cette approche a été implémentée dans le logiciel LISREL 8 par Joreskog and Sorbom (1996). D'où le nom couramment donnée à l'approche de Jöreskog : LISREL. Cependant, d'autres noms peuvent être rencontrés dans la littérature, tels que "*Covariance Structure Analysis*" (CSA) ou SEM...
Avant de présenter l'approche LISREL dans son formalisme le plus général, nous débuterons par le calcul des estimateurs de maximum de vraisemblance des paramètres du modèle dans le cadre simplifié des notations du modèle à facteurs présenté au chapitre 1. Ce choix est fait dans l'objectif de faciliter la compréhension théorique de cette approche.

La fonction de vraisemblance

Dans le cadre du modèle à facteur (1.1) et des hypothèses d'indépendance et de loi gaussienne des facteurs faites section (2.2.1), les y_i suivent indépendamment une loi $\mathcal{N}(\mu, \Sigma^Y)$ tel que $\Sigma^Y = B'B + \Psi$. La vraisemblance d'une séquence d'observations $Y = [y_1, \ldots, y_n]$ et $\theta = (\mu, B, \Psi)$ l'ensemble des paramètres à estimer est :

$$\begin{aligned} l(\theta; Y) &= p(y_1, \ldots, y_n; \theta) \\ &= (2\pi)^{-nq/2} |\Sigma^Y|^{-n/2} exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)' {\Sigma^Y}^{-1} (y_i - \mu) \right\} \\ &= (2\pi)^{-nq/2} |\Sigma^Y|^{-n/2} exp\left\{ -\frac{n}{2} tr\left(S {\Sigma^Y}^{-1} \right) \right\} \end{aligned}$$

où la matrice de variance-covariance empirique est notée :

$$S = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)'(y_i - \mu). \tag{1.20}$$

On en déduit la log-vraisemblance :

$$\mathcal{L}(\theta; Y) = -\frac{n}{2} \left(log|\Sigma^Y| + tr\left(S {\Sigma^Y}^{-1} \right) \right). \tag{1.21}$$

La maximisation de cette fonction revient à la minimisation de la suivante :

$$\mathcal{L}(\theta; Y) = log|\Sigma^Y| + tr\left(S {\Sigma^Y}^{-1} \right) \tag{1.22}$$

Et de manière triviale, la maximisation de cette fonction par rapport à μ donne :

$$\hat{\mu} = \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{1.23}$$

Pour obtenir les formules des autres estimateurs, on poursuit l'optimisation de cette fonction. Maximiser la log-vraisemblance (1.21) revient à minimiser la fonction F suivante [5] :

$$F(B, \Psi) = log|\Sigma^Y| + tr\left(S {\Sigma^Y}^{-1} \right) - log|S| - q \tag{1.24}$$

5. F étant définie par la différence entre la log-vraisemblance et la constante $log|S| + q$, elle va mesurer la proximité entre la matrice de variance covariance théorique Σ^Y et la matrice de variance covariance empirique S. Notons aussi que puisque les variables observables sont supposées gaussiennes, les éléments de la matrice S suivent une loi de Wishart à n degrés de liberté.

où $log|S| + q$ est une constante. La procédure de minimisation consiste à résoudre $\frac{\partial F}{\partial (B,\Psi)} = 0$, ce qui revient à chercher le minimum pour Ψ donné et ensuite le minimum global pour B donné. La dérivée partielle de F par rapport à B est :

$$\frac{\partial F}{\partial B} = 2\Sigma^{Y\,-1}\left(\Sigma^{Y} - S\right)\Sigma^{Y\,-1}B \tag{1.25}$$

et la dérivée partielle par rapport à Ψ est :

$$\frac{\partial F}{\partial \Psi} = diag\left(\Sigma^{Y\,-1}\left(\Sigma^{Y} - S\right)\Sigma^{Y\,-1}\right). \tag{1.26}$$

Les formules des estimateurs sont alors obtenues par annulation de ces dérivées partielles et l'utilisation des identités de Lawley and Maxwell (1963) suivantes (cf. Jöreskog (1969)] :

$$\begin{aligned} \Sigma^{Y\,-1} &= \Psi^{-1} - \Psi^{-1}B'(I + B\Psi^{-1}B')^{-1}B\Psi^{-1} \\ \Sigma^{Y\,-1}B' &= \Psi^{-1}B'(I + B\Psi^{-1}B')^{-1} \end{aligned} \tag{1.27}$$

Enfin, suite à la résolution du système formé des équations (1.25) et (1.26) les formules des estimateurs obtenues sont :

$$\begin{aligned} &\hat{\mu} = \overline{Y} \\ &\hat{B}' = J^{-1}B'\Psi^{-1}\left(S - \Psi\right), \quad tel\ que\ J = B'\Psi^{-1}B \\ &\hat{\Psi} = diag\left(S - \hat{B}'\hat{B}\right) \end{aligned} \tag{1.28}$$

Remarque. La consultation des démonstrations des résultats (1.25), (1.26) et (1.28) est possible dans l'annexe A.
Les formules solutions (1.28) ne représentent ni le minimum de F selon Ψ pour B donné, ni le minimum de F selon B pour Ψ donné. Elles constituent seulement les relations entre les paramètres au minimum global de F. D'ailleurs, l'équation de $\hat{\Psi}$ dans (1.28) dépend de B qui dépend lui même de Ψ. Les estimations du maximum de vraisemblance de B et Ψ doivent satisfaire les équations (1.28) obtenues ou des équations matricielles équivalentes. En pratique, la résolution de ces équations se fait de manière itérative. Des méthodes numériques permettant le calcul des estimations sont présentées dans la section qui suit.

Méthodes numériques pour le calcul des estimations du maximum de vraisemblance

Les équations matricielles (1.28), solutions de l'estimation du maximum de vraisemblance ne pouvant être résolues algébriquement, les procédures itératives constituent alors un recours possible. Différentes méthodes itératives sont présentes dans la littérature telles que celles proposées par Lawley (1942, 1943); Rao (1955); Howe (1955); Bargmann (1957). Lawley fut le premier à proposer une méthode itérative pour un nombre de facteurs $K = 1$ en 1941 puis de $K = 2$ en 1943. Son approche fut ensuite utilisée par Emmett (1949) qui la développa jusqu'au nombre de $K = 3$ facteurs. Ces procédures ont ensuite été exposées dans un livre de Lawley and Maxwell (1963) et aussi par Howe (1955) qui en propose des modifications et cite d'autres approches telles que celle de Rao puis celle de Gauss-Seidel qu'il recommande. Jöreskog (1967) est du même avis et affirme que parmi toutes les approches citées plus haut, une modification de l'approche de Gauss-Seidel utilisée par Browne (1968) est selon lui la plus compétitive vis à vis de celle qu'il propose dans son article et que nous présenterons plus loin. Mais avant, nous présentons l'incontournable approche de Lawley and Maxwell (1963).

Algorithme itératif proposé par Lawley

L'approche classique, repose sur des propriétés d'algèbre linéaire de base. Après avoir observé que la deuxième équation de (1.28) peut se réécrire, $B'\Psi^{-1}(S-\Psi) = JB'$, il vient que la matrice J contient les K valeurs propres de la matrice $\Psi^{-1}(S-\Psi)$ et que les lignes de B' correspondent aux vecteurs propres. En effet, par soucis d'identifiabilité du modèle, J est contrainte à être diagonale. En considérant que les éléments de B' sont réels, il peut être alors démontré que le minimum de F pour une matrice Ψ donnée est obtenu lorsque les vecteurs ligne de B' sont choisis tels qu'ils correspondent aux K plus grandes valeurs propres de $\Psi^{-1}(S-\Psi)$ (cf. Jöreskog (1967) pour plus de détails).
Lawley and Maxwell (1963) proposent la variante qui va suivre où ils font les hypothèses que :

— le nombre de facteurs est $K = 3$ pour simplifier les développements (Emmett, 1949) ;

— J est diagonale.

Nous noterons $\forall k \in [\![1, K]\!]$, b'_k la ligne k de B' et $B_{[t]}$ la matrice B à la t-ième itération.
L'algorithme procède comme suit :

1. Initialisation : $B'_{[1]}$ est initialisé tel que ses K lignes correspondent aux K premières composantes principales de l'ACP des variables observées Y et $\Psi_{[1]} = s_{11} - \sum_{k=1}^{K} b_{1k}$.
2. Actualisation des paramètres B' et Ψ à l'itération [2] suivante :
   (a) Actualisation de la ligne 1 de B' :
   On calcule les vecteurs ligne suivants,
   $$\begin{cases} w'_1 = b'_{1[1]}\Psi^{-1}_{[1]} \\ u'_1 = w'_1 S - b'_{1[1]} \\ h_1 = u'_1 w_1 \end{cases}$$
   puis on obtient, $b'_{1[2]} = \dfrac{1}{\sqrt{h_1}} u'_1$.
   (b) Actualisation de la ligne 2 de B' :
   On calcule les vecteurs ligne suivants,
   $$\begin{cases} w'_2 = b'_{2[1]}\Psi^{-1}_{[1]} \\ j_{21} = w'_2 b'_{1[2]} \\ u'_2 = w'_2 S - b'_{2[1]} - j_{21} b'_{1[2]} \\ h_2 = u'_2 w_2 \end{cases}$$
   puis on obtient, $b'_{2[2]} = \dfrac{1}{\sqrt{h_2}} u'_2$.
   (c) Actualisation de la dernière ligne $K = 3$ de B' :
   On calcule les vecteurs ligne suivants,
   $$\begin{cases} w'_3 = b'_{3[1]}\Psi^{-1}_{[1]} \\ j_{31} = w'_3 b'_{1[2]} \\ j_{32} = w'_3 b'_{2[2]} \\ u'_3 = w'_3 S - b'_{3[1]} - j_{31} b'_{1[2]} - j_{32} b'_{2[2]} \\ h_3 = u'_3 w_3 \end{cases}$$
   puis on obtient, $b'_{3[2]} = \dfrac{1}{\sqrt{h_3}} u'_3$.
   (d) Actualisation de Ψ : $\Psi_{[2]} = S - B_{[2]} B'_{[2]}$.
3. Tant que la convergence n'a pas lieu on passe à la deuxième étape de l'algorithme avec actualisation de B' à l'itération suivante et ainsi de suite jusqu'à obtention de valeurs stationnaires. On considère alors que l'algorithme a convergé [6].

6. Les conditions de convergence de cet algorithme n'ont pas été établies mais la convergence est observée en pratique.

À l'étape 2.(a), lorsqu'on explicite les termes h_1 et u'_1 dans l'expression de $b'_{1[2]}$, on reconnait la ligne 1 de l'équation de $\hat{B}$ dans (1.28). Cependant, l'apparition des termes j_{21}, j_{31} et j_{32} aux étapes 2.(b) et 2.(c) rendent cette reconnaissance moins évidente. Ces termes correspondent aux éléments hors diagonale de la matrice J, laquelle est supposée diagonale. Or, après initialisation, les valeurs des paramètres ne sont pas encore suffisamment proches de leurs valeurs au maximum de vraisemblance. Ainsi, leurs valeurs ne sont pas telles que l'hypothèse de J diagonale est respectée. Alors, à chaque actualisation des paramètres, leurs valeurs sont corrigées par des termes de la matrice J, car les formules solutions (1.28) sont obtenues sous cette hypothèse. En revanche, lorsque les paramètres tendent vers les valeurs stationnaires, les éléments hors diagonale de J tendent vers zéro. La correction sera de plus en plus faible au fur et à mesure que l'on se rapproche du maximum de vraisemblance.
Cela peut s'interpréter géométriquement. Prenons l'exemple de l'étape 2.(b) où

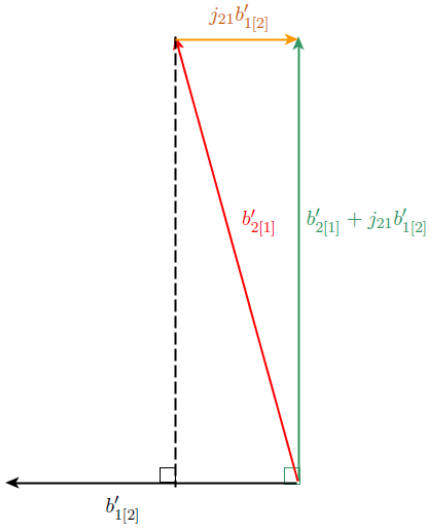


FIGURE 1.6 – Schéma de la correction apportée au vecteur $b'_{2[1]}$ par sa projection orthogonale sur la direction du vecteur $b'_{1[2]}$ au sens de Ψ^{-1}.

$u'_2 = w'_2 S - b'_{2[1]} - j_{21} b'_{1[2]}$. À l'aide du schéma 1.6, on voit que $j_{21}b'_{1[2]}$ correspond à la projection orthogonale du vecteur $b'_{2[1]}$ sur la droite de même direction que le vecteur $b'_{1[2]}$ au sens de Ψ^{-1}. Ainsi dans u'_2, lorsqu'on retranche $b'_{2[1]} + j_{21}b'_{1[2]}$, on "amène" le vecteur $b'_{2[1]}$ sur la direction orthogonale à $b'_{1[2]}$ afin de respecter l'hypothèse J diagonale. On peut également faire le rapprochement avec la procédure classique utilisant les valeurs propres et vecteurs propres de la matrice $\Psi^{-1}(S-\Psi)$ citée en début de section. En effet, si au maximum de vraisemblance la matrice B' contient en ligne les vecteurs propres, alors ceux ci sont Ψ^{-1} orthogonaux et donc, à chaque actualisation, la ligne $k+1$ doit être Ψ^{-1} orthogonale à la ligne précédente k.

Cette procédure n'est pas unique et Howe (1955) en a proposé quelques modifications. À itération fixée, la première ne procède pas à l'orthogonalisation de chaque vecteur ligne de B' par rapport au précédent, ce qui permet d'économiser du temps de calcul.

Modifications de l'algorithme de Lawley proposées par Howe

La première modification est évidente : il suffit d'étendre naïvement l'étape 2.(a) de l'algorithme de Lawley aux étapes d'actualisation de chacune des lignes suivantes de B'. Ce qui s'écrit :

1. Initialisation : $B'_{[1]}$ est initialisé tel que ses K lignes correspondent aux K premières composantes principales de l'ACP des variables observées Y et $\Psi_{[1]} = s_{11} - \sum_{k=1}^{K} b_{1k}$.

2. Actualisation des paramètres B' et Ψ à l'itération $[t+1]$:
   (a) Actualisation de B' : $B'_{[t+1]} = \left(B'_{[t]}{\Psi_{[t]}}^{-1}B_{[t]}\right)^{-1}\left(B'_{[t]}{\Psi_{[t]}}^{-1}S - B'_{[t]}\right)$;
   (b) Actualisation de Ψ : $\Psi_{[t+1]} = S - B_{[t]}B'_{[t]}$.
3. Tant que la convergence n'a pas lieu on passe l'étape 2.(a) puis 2.(b) et ainsi de suite jusqu'à obtention de valeurs stationnaires.

Cet algorithme ne tenant pas compte de l'hypothèse J diagonale, implique que les résultats obtenus sont différents de ceux de l'algorithme de Lawley.

La deuxième consiste à imposer à J l'hypothèse de forme triangulaire inférieure, ce qui s'écrit :

1. Initialisation : $B'_{[1]}$ est initialisé tel que ses K lignes correspondent aux K premières composantes principales de l'ACP des variables observées Y et $\Psi_{[1]} = s_{11} - \sum_{k=1}^{K} b_{1k}$.
2. Actualisation des paramètres B' et Ψ à l'itération $[t+1]$:
   (a) Actualisation de B' : $B'_{[t+1]} = J - [t]^{-1}\left(B'_{[t]}{\Psi_{[t]}}^{-1}S - B'_{[t]}\right)$;
   (b) Actualisation de Ψ : $\Psi_{[t+1]} = S - B_{[t]}B'_{[t]}$.
3. Tant que la convergence n'a pas lieu on passe l'étape 2.(a) puis 2.(b) et ainsi de suite jusqu'à obtention de valeurs stationnaires.

La dernière modification proposée est la suivante :

1. Initialisation : $B'_{[1]}$ est initialisé tel que ses K lignes correspondent aux K premières composantes principales de l'ACP des variables observées Y et $\Psi_{[1]} = s_{11} - \sum_{k=1}^{K} b_{1k}$.
2. Actualisation des paramètres B' et Ψ à l'itération $[t+1]$:
   (a) Actualisation de B' : $B'_{[t+1]} = \left(I + B'_{[t]}{\Psi_{[t]}}^{-1}B_{[t]}\right)^{-1}\left(B'_{[t]}{\Psi_{[t]}}^{-1}S\right)$;
   (b) Actualisation de Ψ : $\Psi_{[t+1]} = S - B_{[t]}B'_{[t]}$.
3. Tant que la convergence n'a pas lieu on passe l'étape 2.(a) puis 2.(b) et ainsi de suite jusqu'à obtention de valeurs stationnaires.

Le seul avantage de ces modifications serait une convergence plus rapide. Mais si le gain en temps de calcul à chaque itération semble évident, l'augmentation de la vitesse de convergence n'a pas été prouvée. Or, Howe (1955) les recommande malgré tout par rapport à l'algorithme de Lawley. Cependant, Jöreskog (1966) montre les limites de ces dernières. Comme par exemple la convergence non systématique ou la lenteur de convergence.
Howe (1955) référence d'autres approches dont par exemple la méthode de Gauss-Seidel (c.f. la section 3.3 de Howe (1955)) qui présente selon lui et Jöreskog (1967) plus d'avantages que celle de Lawley présentée plus haut. Dans le but d'améliorer la vitesse de convergence, par la suite Jöreskog (1967) a développé une procédure reposant sur d'anciennes approches telles que celle de Fletcher and Powell (1963) et celles de Newton Raphson. Ces dernières avaient été délaissées parce qu'elles nécessitaient des calculs lourds de la matrice hessienne des dérivées secondes de F et de son inverse. Elles présentent alors plusieurs inconvénients. Par exemple, pour une des méthodes Newton-Raphson, la matrice hessienne nommée E ainsi que son inverse étaient calculés à chaque itération, ce qui est très gourmand en temps de calcul. Gourmandise qui est d'autant importante que le nombre de paramètres est élevé. Une autre approche Newton Raphson propose quant à elle de ne calculer la matrice E et son inverse qu'une seule fois et de les utiliser tels quels à chaque itération. À première vue, cela peut sembler être moins gourmand en temps de calcul. Mais la conséquence est alors l'augmentation du nombre d'itérations nécessaire pour arriver à convergence, lequel est d'autant plus élevé que l'initialisation est éloignée de l'optimum. L'approche de Fletcher et Powell est un compromis entre ces deux cas de méthode Newton Raphson. Elle calcule une matrice inverse à chaque itération convergeant vers l'inverse de E avec un faible coût en temps de calcul. C'est pour son efficacité que Jöreskog va la combiner à la méthode du pas descendant dans la procédure qu'il propose (Jöreskog, 1967).

Algorithme proposé par Jöreskog

Jöreskog (1967, 1970) fonde sa procédure sur une modification de la méthode computationnelle de Fletcher and Powell (1963). Pour les 5 premières itérations elle procède selon la méthode du pas descendant et ensuite utilise les idées de l'approche de Fletcher et Powell. Celle-ci cherche à minimiser le gradient et la matrice hessienne de F. Ainsi, elle nécessite le calcul des dérivées secondes de F en plus des dérivées premières de F. Lors de sa présentation, nous verrons qu'elle calcule aussi la valeur de la fonction F à chaque itération.
La minimisation de F par l'algorithme de Jöreskog, consiste après un choix initial de valeurs $\theta^{[0]}$ à calculer des estimations $\theta^{[t]}$, $\forall t$ où $[t]$ est la t-ième itération tel que :

$$F\left(S, \Sigma\left(\theta^{[t+1]}\right)\right) \leq F\left(S, \Sigma\left(\theta^{[t]}\right)\right)$$

où $\theta = \{\mu, B, \Psi\}$. Or, d'après les formules solutions (1.28), les paramètres Ψ et B sont liés, contrairement à μ qui peut être estimé de manière indépendante. Donc $\theta^{[t]}$ peut se résumer à $\Psi^{[t]}$. Alors, le critère à satisfaire devient :

$$F\left(S, \Sigma\left(\Psi^{[t+1]}\right)\right) \leq F\left(S, \Sigma\left(\Psi^{[t]}\right)\right)$$

Pour ce faire, des hypothèses sont faites et des notations introduites. On note et définit :

- $x := (x_1, \ldots, x_q)$ le vecteur des éléments diagonaux de Ψ ;
- e le vecteur gradient de F tel que $e_i^{[t]} := \left(\frac{\partial F}{\partial \Psi_{ii}}\right)_{\Psi=\Psi^{[t]}}$;
- E la matrice hessienne de F telle que $E_{ij}^{[t]} := \left(\frac{\partial^2 F}{\partial \Psi_{ii} \partial \Psi_{jj}}\right)_{\Psi=\Psi^{[t]}}$;
- E^* la matrice information de Fisher utilisée pour approximer l'inverse de la matrice E telle que $E_{ij}^{[t]} := \frac{n}{2} E\left[\left(\frac{\partial F}{\partial \Psi_{ii}} \frac{\partial F}{\partial \Psi_{jj}}\right)_{\Psi=\Psi^{[t]}}\right]$;
- $d^{[t]}$ la direction selon laquelle on va chercher $x^{[t+1]}$;
- $\alpha^{[t]}$ la distance du point $x^{[t]}$ suivant la direction $d^{[t]}$ pour obtenir $x^{[t+1]}$;
- $s(\alpha)$ la pente qui correspond au coefficient directeur de la tangente en la localisation $x^{[t]}$.

Les hypothèses sont :

- $\Psi = diag\,(\Psi_{ii})$ est diagonale ;
- $\forall i \in [\![1, q]\!]$, $\Psi_{ii} \geq \epsilon$;
- E^* est définie positive ;
- la procédure est exécutée pour un nombre de facteurs $K \ll q$ fixé.

L'écriture du critère peut donc encore se simplifier :

$$F\left(x^{[t+1]}\right) \leq F\left(x^{[t]}\right)$$

et l'algorithme procède comme suit :

1. Initialisation :
   - Il est possible d'initialiser $\forall i \in [\![1, q]\!]$, $x_i^{[0]} = 1$; $E^{*[0]} = I_q$;
   - mais pour diminuer le nombre d'itérations nécessaire à la convergence, il est recommandé d'initialiser $\forall i \in [\![1, q]\!]$, $x_i^{[0]} = \left(1 - \frac{k}{2p}\right)\left(\frac{1}{s^{ii}}\right)$ où s^{ii} est le i-ème élément diagonal de S^{-1} ; $E^{*[0]} = \left[E^{[0]}\right]^{-1}$;
   - l'initialisation des paramètres $\theta^{[0]}$ et notamment celle de B (à partir de laquelle on peut retrouver Ψ, μ étant estimé dès le départ par l'équation associée en (1.28)) se fait par une méthode de type ACP.

2. À l'itération $[t]$:
   (a) On approxime $E_{ij}^{[t]}$ par $\frac{n}{2}E\left[\left(\frac{\partial F}{\partial \Psi_{ii}}\frac{\partial F}{\partial \Psi_{jj}}\right)_{\Psi=\Psi^{[t]}}\right]$ (les formules utilisées pour le calcul de cette matrice sont présentées dans l'annexe B pour ne pas alourdir la description de l'algorithme) ;
   (b) On calcule le gradient e de F (sa méthode de calcul est elle aussi décrite dans l'annexe B pour ne pas alourdir la description de l'algorithme) ;
   (c) On calcule la direction $d^{[t]} = -E^{*[t]}e^{[t]}$;
   (d) On choisit α le meilleur possible c'est à dire tel que la pente $s(\alpha)$ soit la plus proche possible de zéro. En effet à l'optimum la pente est nulle. La procédure n'est pas détaillée ici mais l'idée consiste à choisir une valeur d'essai α_1 et à calculer les $F^*(\alpha_1)$ et $s^*(\alpha_1)$ ainsi que $F^*(0)$ et $s^*(0)$ tels que décrits aux deux points suivants. Puis en fonction du signe de la pente $s^*(\alpha_1)$ interpoler (quand il est positif) ou extrapoler (quand il est négatif) le minimum en utilisant les deux valeurs de la pente $s^*(\alpha)$ pour $\alpha = 0$ et $\alpha = \alpha_1$;
   (e) On calcule $F^*(\alpha) := F\left[x^{[t]} + \alpha d^{[t]}\right]$ où $\alpha > 0$;
   (f) On calcule la pente $s^*(\alpha) := {d^{[t]}}' e_\alpha$ où e_α est le vecteur gradient au point $x^{[t]} + \alpha d^{[t]}$;
   (g) On appelle $\alpha^{[t]}$ la valeur courante de α la plus minimisante de F.
3. À l'itération $[t+1]$:
   (a) On actualise $x^{[t+1]} = x^{[t]} + \alpha^{[t]} d^{[t]}$;
   (b) On calcule la valeur $F(x^{[t+1]})$ et le gradient $e^{[t+1]}$;
   (c) On actualise $E^{*[t+1]} = E^{*[t]} + \frac{1}{\beta^{[t]}} y^{[t]} {y^{[t]}}' - \frac{1}{\gamma^{[t]}} z^{[t]} {z^{[t]}}'$ telle que,

$$\begin{aligned} y^{[t]} &= x^{[t+1]} - x^{[t]} \\ h^{[t]} &= g^{[t+1]} - g^{[t]} \\ z^{[t]} &= E^{*[t]} h^{[t]} \\ \beta^{[t]} &= {y^{[t]}}' h^{[t]} \\ \gamma^{[t]} &= {h^{[t]}}' z^{[t]} \end{aligned}$$

4. On repasse à l'étape 2 puis 3 et ainsi de suite jusqu'à satisfaire un critère d'arrêt. Par exemple, lorsque les valeurs absolues des dérivées du premier ordre de F par rapport aux paramètres sont toutes inférieures à une valeur positive proche de zéro.

La dernière valeur de E^* obtenue est multipliée par $\frac{2}{n}$ et donne une estimation de la matrice de variance-covariance au maximum de la log-vraisemblance. Même si F n'est pas quadratique, on peut considérer qu'elle l'est au voisinage de l'optimum. Sur la diagonale de E^* on a alors une estimation de la variance des paramètres estimés $\hat{\theta}$ et si on note cette estimation e_{ii}^*. Un intervalle de confiance à 95% peut être approximé par :

$$\hat{\theta}_i - 2\sqrt{(2/n)e_{ii}^*} < \theta_i < \hat{\theta}_i + 2\sqrt{(2/n)e_{ii}^*}$$

Écriture du modèle général et de ses hypothèses

L'écriture du modèle dans la littérature est la suivante (cf. le tableau des notations 1.1) :

$$\begin{cases} y = \Lambda_y \eta + \epsilon \\ x = \Lambda_x \xi + \delta \\ \eta = B\eta + \Gamma\xi + \zeta \end{cases}$$

où, Λ_y, Λ_x, B, Γ sont les matrices des coefficients ; η les variables latentes endogènes, ξ les variables latentes exogènes et ϵ, δ, ζ les erreurs de mesure. On note également $V(\xi) = \Phi$, $V(\zeta) = \Psi$ et $\Theta = \{\Lambda_y,\Lambda_x,\text{B},\Gamma,\Phi,\Psi,\Theta_\epsilon,\Theta_\delta\}$ l'ensemble des paramètres.
Les hypothèses de ce modèle sont :

- Les VO (y, x) sont normales multivariées ;
- ζ et ξ sont non corrélées ;
- $I - B$ est non singulière ;
- $\text{E}(\zeta)=\text{E}(\xi)=\text{E}(\eta)=0$;
- ϵ et η sont non corrélées ;
- δ et ξ sont non corrélées ;
- ϵ, δ et ζ sont mutuellement non corrélées ;
- $\text{E}(\epsilon)=\text{E}(\delta)=0$;
- $\text{V}(\epsilon)= \Theta_\epsilon$, $\text{V}(\delta)=\Theta_\delta$.

Souvent, les conditions d'identifiabilité choisies se traduisent par les hypothèses suplémentaires suivantes :

- $V(\xi) = \Phi = I$;
- $V(\zeta) = \Psi$ est diagonale.

Dans la section 3.1.1., le développement de la fonction de vraisemblance et le calcul des formules solutions, sont faits sous l'hypothèse : $\Gamma = B\Psi^{-1}B'$ (correspondant à la matrice J) est diagonale et la contrainte d'identifiabilité : $\Phi = I$ (cf. Jöreskog (1967b) et Lawley et Maxwell (1963)) pour pouvoir optimiser F et simplifier les calculs. On peut remarquer que $\Phi = I$ correspond à notre hypothèse : les facteurs f suivent une gaussienne centrée et de variance-covariance la matrice identité.

Illustration graphique

Par exemple, le diagramme 1.3 peut, avec les notations et les conventions graphiques de LISREL, être illustré par le diagramme figure 1.7 et s'écrire comme suit :

$$\begin{cases} y = \Lambda_y\eta + \epsilon \\ x = \Lambda_x\xi + \delta \\ \eta = \Gamma\xi + \zeta \end{cases}$$

où, η correspond à g', Λ_y correspond à B', x correspond $[X^1, X^2]$, ξ correspond à $\begin{pmatrix} f^{1\prime} \\ f^{2\prime} \end{pmatrix}$, Λ_x correspond à $[a^1, a^2]'$, Γ correspond à (c^1, c^2), $Var(\xi) = \Phi$ est fixée à I_n,
$Var(\zeta) = \Psi$ est fixée à $(c^1)^2+(c^2)^2$, $Var(\epsilon) = \Theta_\epsilon$ correspond à Ψ_Y et $Var(\delta) = \Theta_\delta$ correspond à $diag(diag(\Psi_1), diag(\Phi_2))$.

L'approche LISREL

Les VO étant normales multivariées, l'ensemble de l'information des données est présente dans les moments d'ordre 1 et 2 : la moyenne et la matrice de variance-covariance. D'où l'intérêt d'estimer la matrice de variance-covariance à partir du modèle à équations structurelles. L'idée fondamentale de l'approche LISREL repose sur l'hypothèse que la matrice de variance-covariance Σ des VO (y, x) est égale à la matrice de variance-covariance théorique $\Sigma(\Theta)$ exprimée en fonction de l'ensemble des paramètres Θ à estimer. Ce qui s'écrit :

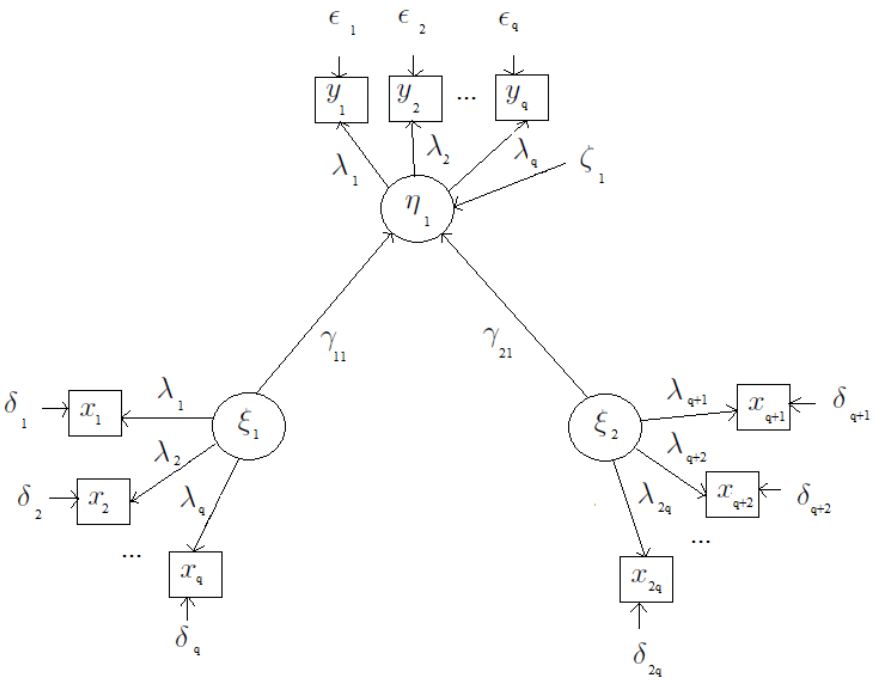


FIGURE 1.7 – Diagramme du modèle (1.5) pour l'approche LISREL.

$$\Sigma = \Sigma(\Theta)$$

où Σ la matrice des colonnes de x est supposée définie positive.
En utilisant les hypothèses de non corrélation on obtient que $\Sigma(\Theta)$ s'écrit :

$$\begin{pmatrix} V(y) & COV(y,x) \\ COV(x,y) & V(x) \end{pmatrix} = \begin{pmatrix} \Lambda_y(I-B)^{-1}(\Gamma\Phi\Gamma' + \Psi)[(I-B)^{-1}]'\Lambda_y' + \Theta_\epsilon & \Lambda_y(I-B)^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'[(I-B)^{-1}]'\Lambda_y' & \Lambda_x\Phi\Lambda_x' + \Theta_\delta \end{pmatrix}$$

Pour un jeu de données fixé, la matrice Σ observée correspond à la matrice de variance-covariance empirique, notée S, qui est toujours supposée définie positive. Alors, en pratique, à l'aide de S, on explicite $\hat{\Sigma}(\Theta)$ en minimisant la fonction basée sur le maximum de vraisemblance (cf. (1.24)),

$$F(S, \Sigma(\Theta)) = log|\Sigma(\Theta)| + tr(S\Sigma(\Theta)^{-1}) - log|S| - q$$

$\Sigma(\Theta)$ est estimée de manière à minimiser l'écart entre elle et S au sens de la fonction F à optimiser. C'est en cela et par les hypothèses et propriétés de distribution que l'approche de Jöreskog permet d'évaluer la qualité d'ajustement du modèle aux données. On note $\hat{\Sigma}(\Theta)$ l'estimation obtenue de $\Sigma(\Theta)$. Cette minimisation se fait par l'optimisation de la fonction F dont nous avons décrit la procédure plus haut.
Dans la littérature d'autres fonctions F associées à d'autres approches sont proposées. En voici quelques une :

— L'approche des moindres carrés généralisés (GLS) :

$$F(S, \Sigma(\Theta)) = \frac{1}{2} tr(S^{-1}(S - \Sigma(\Theta))^2).$$

— L'approche des moindres carrés non pondérés (ULS) McDonald (1996) :

$$F(S, \Sigma(\Theta)) = \frac{1}{2} tr((S - \Sigma(\Theta))^2).$$

— L'approche des moindres carrés pondérés (WLS ou ADF) :

$$F(S, \Sigma(\Theta)) = (S - \Sigma(\Theta))' W^{-1} (S - \Sigma(\Theta)).$$

où W est une matrice des poids.

Cependant, l'approche WLS a l'avantage de pouvoir être utilisée pour des données non normales ou ordinales mais présente l'inconvénient de nécessiter de grands jeux de données. Les références Bollen (2014) et McDonald (1996) présentent de manière complète les approches GLS et ULS. L'approche ULS possède le grand avantage de proposer une réponse à la question importante de l'estimation des facteurs dont leurs estimations sont nommées "scores". L'optimisation de la fonction F de ULS est complétée par le calcul de chaque score comme combinaison linéaire de ses VO. Sans ce complément, ni l'optimisation de la fonction F de ULS, ni de celles de GLS ou WLS ne permet l'estimation des facteurs. Nous aborderons à nouveau l'approche ULS dans le chapitre suivant lorsque nous traiterons de la question de l'estimation des facteurs.

1.3 Conclusion et discussion

À l'exception de quelques cas d'approches PLS (cf. les critères 2 et 3 précédents) et de RGCCA et THEME, PLSPM n'optimise pas de fonction contrairement à LISREL et ne fait pas d'hypothèse sur les distributions des VL. LISREL quant à elle, optimise une fonction F, ce qui revient à maximiser la fonction de vraisemblance du modèle pré-établi. Cela se paye par la nécessité d'hypothèses de distributions gaussiennes des variables (rares dans certains domaines). De plus, LISREL repose sur un problème d'optimisation dont la résolution itérative est lourde. Ceci donne un avantage computationnel à PLSPM. Un autre avantage pour PLSPM est que les VL standardisées sont estimées par des combinaisons linaires de leurs VO centrées et leurs estimations nommées "scores", sont obtenues directement après convergence de l'algorithme. *A contrario*, dans le cadre de l'approche LISREL l'estimation des VL n'est pas automatique. De plus, les méthodes proposant cette estimation n'ont été proposées que récemment. Ces scores sont calculables a posteriori par deux méthodes : l'une développée par Jöreskog (2000) et l'autre par Tenenhaus et al. (2005). La méthode proposée par Jöreskog (2000) est du type moindres carrés mais n'utilise pas les équations structurelles du modèle. L'autre méthode, proposée par Tenenhaus et al. (2005), s'inspire de l'approche PLS en calculant les scores comme combinaison linéaire de toutes les VO centrées où les poids sont les estimations des coefficients du modèle. Il en résulte que les estimations obtenues sont proches de celles des sorties de l'approche PLS. Un autre inconvénient pour l'approche LISREL est qu'elle nécessite l'inversion de matrices de covariances S ou Σ^Y qui ne sont pas forcément définies positives. Si elles ne le sont pas, on est confronté à un problème de non identifiabilité du modèle. De plus, si ces matrices sont singulières, on obtient en pratique que leur déterminant est nul ou négatif. Or, le déterminant d'une matrice de covariance est une variance généralisée et ne peut donc pas être négative. D'après Anderson and Gerbing (1984) cela peut arriver quand l'échantillon est de petite taille. Chen et al. (2001) et Wothke (1993) détaillent ces problèmes d'identification du modèle et de matrices singulières liés à l'approche LISREL. Ces méthodes sont donc multiples et complexes. Bien que PLS semble présenter un avantage pratique par rapport à LISREL (résultant en partie de la contrainte des VL à être des composantes), celui ci se paye par un espace des solutions réduit. De plus, LISREL traite mieux les relations partielles que PLS. Mais PLS n'est plus la seule méthode à composantes pour estimer les SEM. RGCCA et THEME le font aussi via l'optimisation d'un critère. En outre, THEME a en plus de l'avantage de traiter les relations partielles, celui d'extraire plusieurs VL par groupe dans un ordre hiérarchique. Enfin, les méthodes PLS et LISREL ont été comparées dans plusieurs travaux (Jakobowicz, 2007; Stan and Saporta, 2006; Karl G. Jöreskog, 1982) et il en résulte qu'elles sont à la fois concurrentes et complémentaires : PLS est plus adaptée pour faire de la prévision alors que LISREL l'est plus pour de la validation de modèle.

CHAPITRE 2

L'estimation par algorithme EM

Sommaire

2.1 Utilisation actuelle de l'algorithme EM par les approches PLS et LISREL . 32
2.1.1 Introduction et motivations 32
2.1.2 La question des données manquantes 32
2.1.3 La question de la reconstruction de concepts latents 37
2.1.4 Conclusion 40
2.2 L'algorithme EM 41
2.2.1 Structure générale de l'algorithme EM 41
2.2.2 Preuve de la convergence de l'algorithme EM 43
2.3 Méthode d'estimation par algorithme EM pour un modèle à une équation structurelle et un facteur par bloc de variables observées 44
2.3.1 Modèle à deux blocs : l'un dépendant et l'autre explicatif 44
2.3.2 L'algorithme EM pour le modèle à deux blocs : l'un dépendant et l'autre explicatif 45
2.4 Conclusion et discussion 50

2.1 Utilisation actuelle de l'algorithme EM par les approches PLS et LISREL

2.1.1 Introduction et motivations

Depuis la création des approches LISREL et PLS, de multiples travaux de recherches ont contribué à les améliorer, leur adjoindre des fonctionnalités complémentaires et à les étendre à des modèles plus généraux. Les travaux de Muthén et al. (1987); Muthén (1989); Arbuckle et al. (1996) se sont par exemple penchés sur la question de la gestion des données manquantes. Quant à McDonald (1996); Jöreskog (2000); Tenenhaus (2007), ils se sont penchés sur la question de l'estimation des variables latentes bien que, Bartlett (1937, 1938) et Thomson (1948) avaient déjà abordé cette question. Ils proposent des méthodes pour l'estimation des variables latentes (nommées scores) dans l'analyse en facteurs communs et spécifique. Enfin, Moosbrugger et al. (1997); Klein and Moosbrugger (2000); Bollen (1995, 1996); Joreskog and Sorbom (1996); Jöreskog and Yang (1997) et plus récemment Jakobowicz and Saporta (2007) sont une liste non exhaustive de chercheurs qui ont contribué à développer des techniques d'estimation de modèles plus généraux tenant compte de relations non linéaires entre les VL. Par exemple, les modèles à équations structurelles non linéaires et à classes latentes peuvent être estimés par un algorithme EM. D'ailleurs, l'utilisation de l'algorithme EM par Goodman (1974a,b) a fourni la première méthode cohérente d'estimation du modèle des classes latentes de Lazarsfeld (1950, 1954) considéré comme un modèle de mélange. Les contributions étant nombreuses, la première partie de ce chapitre se focalise uniquement sur celles utilisant l'algorithme EM ou traitant de la question de l'estimation des scores. Cela permet de situer la contribution de ce travail de thèse vis à vis des travaux existants utilisant également l'algorithme EM ou ayant le même objectif d'estimation des concepts latents. Par la suite, l'algorithme EM est présenté puis utilisé pour un modèle structurel à facteurs simple afin de mettre en place l'approche d'estimation par algorithme EM et les notations associées, que nous utiliserons ultérieurement.

2.1.2 La question des données manquantes

En statistique, on parle de donnée manquante lorsque pour une variable donnée et une observation donnée, on n'a pas de valeur. La gestion de ces données manquantes est un problème qui ne peut être ignoré. La vaste littérature sur ce sujet nous apprend qu'en fonction de la quantité des données manquantes et de leur type, diverses solutions peuvent être choisies pour les traiter. On peut par exemple supprimer les observations pour lesquelles une ou plusieurs variables ont des valeurs manquantes, imputer des valeurs à ces données manquantes ou encore utiliser des algorithmes qui permettent de réaliser des analyses statistiques malgré l'absence de ces données.
Lors d'une étude statistique sur un jeu de données réelles, la première étape est la préparation des données. Le praticien doit alors procéder au "nettoyage" de la base de données qui consiste à identifier les données aberrantes, les observations atypiques et gérer les données manquantes. Différentes méthodes de traitement de celles-ci existent dans la littérature. Le choix de la méthode adéquate ne peut s'effectuer qu'après avoir identifié le processus sous-jacent à l'absence des données. Il existe trois familles de données manquantes décrites par la suite et illustrées figure 2.2.

Par ailleurs, nous notons X le tableau de données (cf. figure 2.1) où $x_{j,i}$ est l'observation située en colonne j et ligne i associée à la variable X_j, $j \in [\![1, q]\!]$, pour l'unité statistique $i \in [\![1, n]\!]$. On note également X_{-j} l'ensemble des variables du jeu de données X privé de la variable X_j.
Les trois types de données manquantes peuvent être ordonnés par l'intensité d'aléa liée à leur absence. Nous allons les décrire ici, du degré d'aléa le plus élevé au plus faible :

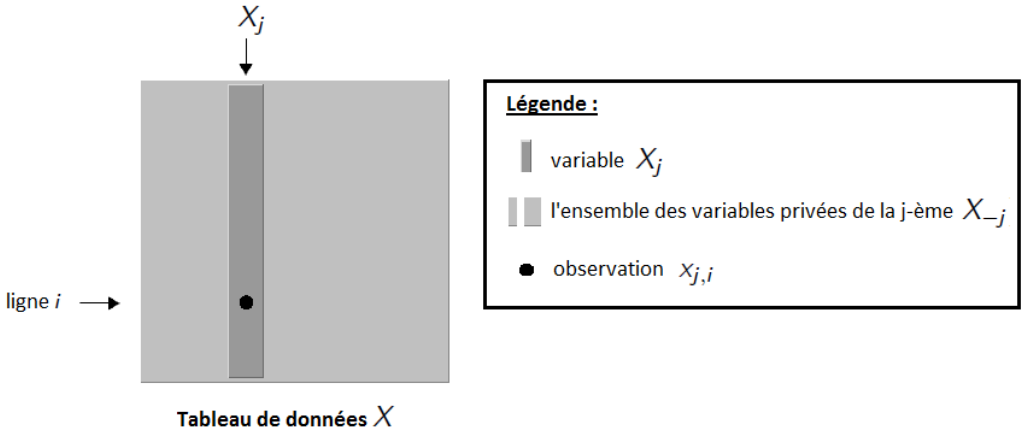


FIGURE 2.1 – Tableau de données X de dimension $n \times q$.

— **Les données manquantes du type MCAR** (Missing Completly At Random) sont celles qui manquent de manière complètement aléatoire. C'est à dire que la probabilité qu'une donnée associée à une variable soit manquante ne dépend ni des valeurs prises par la variable, ni de celles prises par les autres variables du jeu de données.

— **Les données manquantes du type MAR** (Missing At Random) sont celles qui manquent de manière aléatoire. C'est à dire que la probabilité qu'une donnée associée à une variable soit manquante ne dépend pas de la variable en question mais peut dépendre des autres variables.

— **Les données manquantes du type MNAR** (Missing Not At Random) sont celles qui manquent de manière non aléatoire. Pour celles-ci, la probabilité qu'une donnée associée à une variable soit manquante dépend à la fois de la variable en question et des autres variables du jeu de données. Les données manquantes suivent alors un processus pouvant être défini par une fonction. Les données manquantes sont dites censurées, tronquées ou sélectionnées.

Dans le cas de ce dernier type de données manquantes non aléatoirement (MNAR), celles-ci ne peuvent être traitées que si le processus est connu. Pour une revue de la littérature dans le cadre des modèles à équations structurelles, on peut lire les travaux de :

— Gold and Bentler (2000) pour les cas de données manquantes MAR et MCAR ;

— Muthén et al. (1987) pour les problèmes de données manquantes du type MNAR.

Pour commencer, nous présenterons rapidement d'une part les méthodes classiques de traitement des données manquantes du type MCAR et MAR puis d'autre part, celles liées aux approches LISREL et PLS. Ensuite nous exposerons le cas des données manquantes MNAR dans le cadre des modèles à équations structurelles.

Les données manquantes aléatoirement et complètement aléatoirement : MCAR et MAR

Nous exposons les méthodes habituellement employées pour traiter les données manquantes MCAR et MAR ainsi que leur principe. Afin de les illustrer, nous nous plaçons dans le cadre d'un tableau de données X dont les q variables X_j sont en colonnes et les n observations indicées par i sont en ligne.

Les méthodes usuelles

— **La méthode par suppression (ou de type *deletion*** : C'est une méthode radicale qui consiste à retirer du tableau de données X toutes les lignes i pour lesquelles au moins

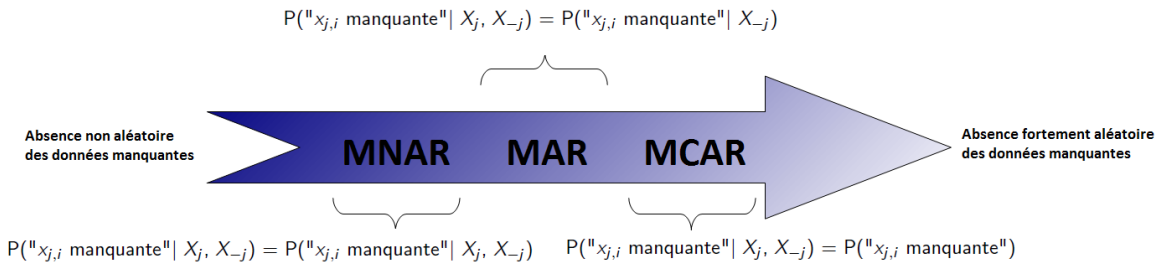


FIGURE 2.2 – Trois familles de données manquantes : MNAR, MAR et MCAR ordonnées par l'intensité d'aléa de leur absence et décrites mathématiquement.

une des variables X_j possède une observation manquante. La conséquence fréquemment observée en pratique est que beaucoup de lignes du tableau sont supprimées ce qui engendre des biais et une importante perte d'information.

— **La méthode de remplacement par la moyenne** : Comme son nom l'indique, chaque observation $x_{j,i}$ manquante est substituée par la moyenne des observations disponibles de la variable X_j. Cette technique est souvent utilisée mais engendre une réduction de la variabilité des variables et produit donc des biais sur les variances et covariances. Remarquons qu'il est possible de procéder de manière similaire avec la médiane, le mode etc.

— **La méthode d'imputation Hot-Deck** : Très utilisée dans la pratique des enquêtes (Ford, 1983), cette technique consiste à chercher pour chaque ligne i où se situe une observation manquante, une ligne i' similaire. Alors, on impute les observations $x_{j,i'}$ aux positions manquantes $x_{j,i}$.

— **La méthode d'imputation simple par algorithme EM** : Très populaire, l'algorithme EM procède en deux étapes :

Étape E (Expectation) La moyenne conditionnelle de la log-vraisemblance des données complètes (observées et manquantes) sachant les données observées (non manquantes) est calculée ;

Étape M (Maximization) Les paramètres sont actualisés par maximisation de la fonction obtenue à l'étape E. Ainsi, dans le cadre de l'approche LISREL, la matrice de covariance est mise à jour.

Après une initialisation adéquate, ces deux étapes sont réitérés jusqu'à satisfaire un critère d'arrêt. Par exemple, jusqu'à obtention lors de deux itérations successives, des matrices de covariance quasi-identiques. Cet algorithme, outil principal de la contribution de ce travail de thèse, sera décrit plus en détail dans les chapitres suivants. Il a été développé par Dempster et al. (1977) et sa version la plus utilisée pour le traitement des données manquantes est la suivante :

— **La méthode d'imputation multiple via algorithme EM** : Elle a été proposée par Rubin (1987) et consiste à répéter la méthode d'imputation simple par algorithme EM.

— **Les méthodes d'imputation multiples** : Celles-ci utilisent les méthodes d'imputations usuelles fondées sur le maximum de vraisemblance (modèles à effets mixtes). Elles répètent ces procédures d'imputation simples et présentent elles aussi l'avantage d'augmenter la variance des paramètres estimés.

Différents travaux de comparaison des méthodes de traitement des données manquantes ont été effectués et il en ressort qu'en général les plus adéquates sont celles fondées sur l'algorithme EM et l'imputation multiple. Pour une revue de ces travaux, on peut consulter (Brown, 1994; Gold and Bentler, 2000) et (Olinsky et al., 2003).

Après application de ces méthodes, le tableau de données qui en résulte est dit complet. On peut alors passer au choix d'une procédure d'estimation du modèle mathématique établi. Dans le cadre d'une modélisation à équations structurelles à VL les paramètres peuvent être estimés par les exemples d'approches LISREL ou PLS. Mais notons que les différentes procédures présentées ne répondent pas à la question de l'estimation des variables latentes. Avant de traiter de la problématique d'estimation des variables latentes et des facteurs, nous allons présenter des procédures complémentaires développées au cours des 20 dernières années.

La méthode *Full Information Maximum Likelihood* (FIML) pour LISREL

Dans un objectif de gestion des données manquantes, l'approche LISREL offre la possibilité d'utiliser la fonction FIML (Arbuckle et al., 1996). Elle permet l'estimation des paramètres d'un modèle à équations structurelles de manière moins brutale que la méthode par délétion. Elle propose qu'à chaque ligne i du tableau de données, seules les observations disponibles soient utilisées lors de l'estimation du maximum de vraisemblance en faisant abstraction des valeurs manquantes. Plus précisément, l'estimation du maximum de vraisemblance est obtenue en procédant aux modifications suivantes :
À chaque variable X_j à n observations dont l sont manquantes, on considère une variable X_j^* à $n-l$ observations telle que,

$$X_j^* = I_s X_j$$

où I_s est une matrice identité amputée des lignes i associées aux valeurs manquantes de X_j. On suppose alors que toutes les variables du tableau X ainsi modifiées suivent une loi normale de moyenne $I_s\overline{x_j^*}$ et de matrice de variance-covariance $I_s \Sigma {I_s}'$ avec Σ la matrice de variance-covariance de X. On appellera X^* le nouveau tableau de données.
Alors, pour chaque observation i de X^* la fonction de log-vraisemblance est définie par :

$$\mathcal{L}(\theta_i) = -\frac{1}{2}\left\{ ln\left|\Sigma_i\right| + \left(X_i^* - \overline{x_i^*}\right)' {\Sigma_i}^{-1} \left(X_i^* - \overline{x_i^*}\right)\right\} + \lambda_i$$

où,

- X_i^* est le vecteur contenant uniquement les valeurs observées disponibles de la ligne i de X ;
- $\overline{x_i^*}$ est le vecteur des moyennes des variables dont les observations sont disponibles pour le cas de la ligne i ;
- λ_i une constante liée au nombre de valeurs manquantes à la ligne i ;
- Σ_i est la matrice de variance covariance des variables dont les observations sont disponibles à la ligne i ;
- $\theta_i = \left\{\overline{x_i}, \Sigma_i\right\}, \forall i \in [\![1, n]\!]$.

Bien sûr par la suite on a :

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \mathcal{L}(\theta_i)$$

L'avantage de FIML est que les variances et covariances sont obtenues sans biais associé à la taille de l'échantillon.
Dans le cas de l'approche PLS, lors de données manquantes, l'algorithme NIPALS (Non Linear Iterative PArtial Least Squares) est utilisé. Il a été développé bien avant l'algorithme PLS appliqué aux modèles à équations structurelles et s'intègre au niveau des premières étapes

(1.13) et (1.15) de l'algorithme. Ces dernières sont modifiées de telle manière à n'utiliser que les données disponibles à chaque ligne i. Ainsi les variables latentes sont estimées en ne tenant compte que des données non manquantes.
Des travaux de comparaison entre les méthodes de traitement des données manquantes dans le cadre des modèles à équations structurelles aboutissent aux conclusions que lorsque les données sont normales, FIML obtient de meilleurs résultats que NIPALS et que les méthodes usuelles (Olinsky et al., 2003) et (Verleye, 1997). NIPALS a tendance à rendre les écarts types petits. Cependant, lorsque les données ne sont pas normales, PLS est parfois préféré pour son absence d'hypothèse distributionnelle.

Les données manquantes non aléatoirement : MNAR

Pour traiter les données manquantes du type MNAR, il est nécessaire de connaître le processus d'absence sous-jacent. Ce processus peut prendre la forme d'une fonction ou d'un modèle. Il y a les modèles de mélanges et les modèles "sample selection models". Dans le cas des modèles à équations structurelles, on peut rencontrer dans la littératures plusieurs méthodes de traitement des données manquantes MNAR. Nous discuterons de deux d'entre elles : la méthode d'Heckman et celle de Tobit. Les modèles Tobit (1958) et d'Heckman (1979) sont des modèles économétriques. Celui de Tobit a été proposé par James Tobin (1958) et est utilisé pour décrire une relation entre une variable dépendante censurée ou tronquée et une variable indépendante (exogène).
Exemples de données manquantes non aléatoirement :

— une femme qui sort d'un essai longitudinal sur la chute de cheveux parce qu'elle en a perdu ;

— la non déclaration de revenus par une personne parce que celui ci est faible.

Muthén et al. (1987); Muthén (1989), utilisent les principes du modèle Tobit et proposent une "analyse factorielle Tobit". Mais celle-ci étant limitée aux cas des données censurées, tronquées ou sélectionnées, il est préférable d'utiliser le modèle Heckit de Heckmann fondé sur les "sample selection models". Cependant, ces deux techniques reposent sur des hypothèses de distributions normales des observations, ce qui permet l'utilisation du maximum de vraisemblance pour les estimations mais limite l'utilisation de ces méthodes à LISREL. Ainsi, le traitement des données manquantes du type MNAR n'est effectué que sous l'approche LISREL et non sous PLS.
Par exemple, si on nomme θ_1 la concaténation des paramètres liés au modèle et θ_2 celui lié aux processus d'absence des données, ce dernier étant supposé connu, il suffit d'estimer θ_1 en supposant θ_2 connu. Muthén et al. (1987) utilisent une fonction de vraisemblance adaptée et ont validé cette approche sur des données simulées.
Tang and Lee (1998); Lee and Tang (2006) se sont aussi penchés sur la problématique des données manquantes non aléatoirement pour les modèles à équations structurelles. Mais la nécessité de connaître le processus d'absence des données manquantes reste nécessaire. Ils ont proposé (ref. Tang et Lee 1998) une méthode fondée sur l'algorithme EM pour des données MNAR tronquées ou censurées. Quelques années plus tard (Lee and Tang, 2006), ils se sont tournés vers l'utilisation d'outils bayésiens pour l'analyse des modèles d'absence des données, en supposant que l'absence des données suit un modèle de régression logistique.

Pour finir, bien que la question de la gestion des données manquantes et de leur traitement reste la source de beaucoup de travaux de recherche et stimule encore nombre de chercheurs, l'application des méthodes proposées jusqu'à présent reste faible. En effet, en pratique, il est rare de connaître le processus d'absence des données. Et puis lorsqu'on le suppose connu, sa modélisation reste difficile et son estimation délicate. Ceci s'explique par la sensibilité au modèle des estimations des coefficients. Par exemple, l'emploi de la méthode de Heckmann, s'accompagne d'une validation par analyse de sensibilité. En outre, ces méthodes sont sen-

sibles aux déviations de la normalité des données, elles nécessitent une bonne spécification du processus d'absence. Enfin, l'implémentation de la plupart d'entre elles uniquement sur le logiciel commercial Mplus, ne favorise pas leur utilisation.

2.1.3 La question de la reconstruction de concepts latents

Les variables latentes sont des concepts non directement observables du modèle dont on suppose l'existence. Lorsque le modèle est établi on suppose par ailleurs les relations de cause à effet entre les VL à travers les équations structurelles puis entre elles et les variables observables à travers les équations de mesures. On pourrait naïvement considérer les VL comme des données manquantes et chercher à les estimer avec les méthodes décrites précédemment. Cependant, les variables latentes sont des données manquantes particulières. Elles n'appartiennent à aucune des familles de données manquantes présentées. Les VL, sont des variables pour lesquelles toutes les observations $i \in [\![1, n]\!]$ sont indisponibles. Ainsi, les méthodes de traitement des données manquantes exposées à la section précédente ne sont pas adaptées ou du moins non applicables directement. Or la question de leur estimation vient tout naturellement et serait une source d'information pertinente. Cela permettrait de les quantifier et d'enrichir l'étape d'interprétation lors d'une étude statistique.
Parmi les approches d'estimation des modèles à équations structurelles, PLS permet une reconstruction des VL depuis de nombreuses années. PLS ne faisant pas d'hypothèse de distribution sur les variables, permet de réaliser une reconstruction des VL en leur imposant d'être des composantes. Comme cela est décrit aux étapes (1.13), (1.14) et (1.15) de l'algorithme PLS au chapitre précédent, les VL sont tout simplement reconstruites comme des combinaisons linéaires des VO. Dans un objectif de prévision, cela représente un grand avantage par rapport à l'approche LISREL qui a tardé à proposer une méthode d'estimation des VL. Ainsi, avant que Jöreskog (2000), complète LISREL8 par une technique d'estimation des VL nommées scores, McDonald (1996) propose une fonction ULS permettant de calculer chaque score comme combinaison linéaire de ses VO. Ceci est contraignant vis-à-vis des hypothèses de distributions normales des variables chez LISREL. En imposant comme dans PLS cette nature de composante aux VL de LISREL, l'espace des solutions s'en trouve en effet réduit. La technique d'estimation des scores de Jöreskog (2000) respecte davantage la nature des VL de LISREL.

Les scores des variables latentes dans LISREL

Il a fallu attendre la version LISREL 8.30 pour que l'approche LISREL permette de quantifier les VL du modèle. La technique repose sur la théorie de l'analyse factorielle (Lawley and Maxwell, 1963). Nous allons présenter l'approche et le calcul des formules des scores.
Nous nous plaçons dans le cadre des notations du modèle LISREL tel qu'il est présenté au chapitre précédent (section 1.2.2) :

$$\begin{cases} y = \tau_y + \Lambda_y \eta + \epsilon & \text{(2.1a)} \\ x = \tau_x + \Lambda_x \xi + \delta & \text{(2.1b)} \\ \eta = \alpha + \beta\eta + \Gamma\xi + \zeta & \text{(2.1c)} \end{cases}$$

où,

- l'équation structurelle (2.1c) a été complétée par l'ordonnée à l'origine (nommée aussi "intercept") α ;
- l'équation structurelle (2.1c) a été enrichie de relations entre les variables endogènes η ;
- les équations de mesures (2.1a) et (2.1b) sont respectivement complétées par les ordonnées à l'origine τ_y et τ_x.

Les équations de mesures peuvent être fusionnées sous l'écriture matricielle :

$$x^\star = \tau + \Lambda\xi^\star + \delta^\star \tag{2.2}$$

équivalente à :

$$\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} \tau_y \\ \tau_x \end{pmatrix} + \begin{pmatrix} \Lambda_y & 0 \\ 0 & \Lambda_x \end{pmatrix} + \begin{pmatrix} \eta \\ \xi \end{pmatrix} + \begin{pmatrix} \epsilon \\ \delta \end{pmatrix}. \tag{2.3}$$

On note, $E(\xi) = \kappa$ et $V(\xi) = \Phi$ puis de manière similaire, on note $E\left(\xi^\star\right) = \kappa^\star$ et $V\left(\xi^\star\right) = \Phi^\star$ tels que :

$$\Phi^\star = \begin{pmatrix} A\left(\Gamma\Phi\Gamma' + \Psi\right)A' & A\Gamma\Phi \\ \Phi\Gamma' A' & \Phi \end{pmatrix}$$

et,

$$\kappa^\star = \begin{pmatrix} (I-B)^{-1}(\alpha\Gamma\kappa) \\ \kappa \end{pmatrix}.$$

En réutilisant les notations ξ, x et δ pour reformuler l'équation de mesure (2.2), on obtient :

$$x = \Lambda\xi + \delta \tag{2.4}$$

tel que, $\xi = \xi^\star - \kappa^\star$, $x = x^\star - \tau - \Lambda\kappa^\star$ et $\delta = \delta^\star$. Alors, il suffit de calculer soit la formule de l'estimateur $\hat{\xi^\star}$ soit celle de $\hat{\xi}$ pour obtenir la formule de l'autre par la relation $\hat{\xi} = \hat{\xi^\star} - \hat{\kappa^\star}$ les unissant.
Soient N observations x_i, d'après Anderson and Rubin (1956), on estime ξ en minimisant :

$$\sum_{i=1}^{N} (x_i - \Lambda\xi_i)'\Theta^{-1}(x_i - \Lambda\xi_i) \tag{2.5}$$

sous la contrainte,

$$\frac{1}{N}\sum_{i=1}^{N} \xi_i {\xi_i}' = \Phi$$

où $\Theta = \begin{pmatrix} \Theta_\epsilon & {\Theta_{\delta\epsilon}}' \\ \Theta_{\delta\epsilon} & \Theta_\delta \end{pmatrix}$ et sous les hypothèses de positivité des matrices Θ et Φ et l'hypothèse que Λ est de rang plein.
Cette minimisation est effectuée par la méthode du Lagrangien où on minimise :

$$\sum_{i=1}^{N} (x_i - \Lambda\xi_i)'\Theta^{-1}(x_i - \Lambda\xi_i) + tr\left\{L\left(\sum_{i=1}^{N} \xi_i {\xi_i}' - N\Phi\right)\right\} \tag{2.6}$$

avec L la matrice symétrique multiplicatrice de Lagrange. Cette minimisation permet d'obtenir la formule de l'estimateur de ξ et donc des VL η_i et ξ_i pour une observation i :

$$\hat{\xi_i} = \left(\Lambda'\Theta^{-1}\Lambda + L\right)^{-1}\Lambda'\Theta^{-1}x_i \tag{2.7}$$

où L est telle que,

$$\begin{cases} \left(\Lambda'\Theta^{-1}\Lambda + L\right)\Phi\left(\Lambda'\Theta^{-1}\Lambda + L\right) = \Lambda'\Theta^{-1}A\Theta^{-1}\Lambda \\ A = \sum_{i=1}^{N} \xi_i {\xi_i}' \end{cases}$$

Dans (Jöreskog, 2000), on peut trouver la formulation de l'estimateur de $\xi^\star$ après plusieurs décompositions matricielles. Il montre également que les estimateurs obtenus sont sans biais.

Jöreskog propose ensuite d'estimer à nouveau les paramètres β et Γ et de vérifier si l'on retrouve les mêmes estimations que celles obtenues par l'CSA de LISREL. En effet, la minimisation de l'équation (2.5) correspond à l'estimation du maximum de vraisemblance pour le modèle privé des équations structurelles. Ce qui revient à utiliser des techniques de moindres carrées uniquement sur les équations de mesure. Les VL sont alors considérées comme des paramètres fixes à estimer. On peut donc discuter le fait que pour l'estimation des VL le modèle ne soit pas utilisé au complet lors de l'étape d'optimisation. Enfin, bien que cette technique soit lourde, elle est une avancée pour LISREL vis à vis de PLS.

Les classes latentes de modèles à équations structurelles non linéaires

Les modèles à équations structurelles classiques considèrent les relations entre les VL comme linéaires. Les équations structurelles décrivent les VL endogènes par une fonction linéaire des VL exogènes. Cette modélisation peut parfois se révéler simpliste pour le théoricien comme le praticien qui peut vouloir passer à des relations plus complexes. On peut ainsi compléter les relations linéaires par un produit de VL exogènes. On considère l'exemple simple d'équation structurelle avec interaction suivant :

$$\eta = \alpha + \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_1\xi_2 + \zeta \tag{2.9}$$

où les notations correspondent à celles de la littérature.
Proposée par Kenny and Judd (1984), ce type d'équation structurelle avec interaction élémentaire de deux VL fut la source du développement de plusieurs approches. Les méthodes d'analyses implémentées pour ce type de modélisation structurelle sont détaillées dans Hayduk (1988) pour LISREL 7 - ML, Moosbrugger et al. (1993) et Ping Jr (1996a,b) pour LISREL 7 - ML (2-Step-T). Puis pour LISREL 8 - ML, on peut consulter Jaccard and Wan (1995, 1996) et Jaccard et al. (1990), Jöreskog et al. (1996); Jöreskog and Yang (1997); Schermelleh-Engel et al. (1998) et enfin Yang Jonsson (1997). Cependant, ces méthodes sont fondées sur des hypothèses de distributions normales uniquement et considèrent le produit des VL exogènes comme une seule VL. Or, il est évident que le fait d'enrichir l'équation structurelle par une interaction de type produit de VL exogènes aura pour conséquence une distribution non normale des VL endogènes η. Il peut être envisagé de tester la performance de méthodes sans hypothèse de distribution telles que PLS, RGCCA ou THEME. Mais elles restent inadaptées pour les modèles à relations non linéaires. De plus, d'après Klein and Moosbrugger (2000), cela rend ces techniques (PLS, RGCCA et THEME) lentes d'autant plus que le nombre d'observations est petit. D'autres approches se sont développées tenant compte de cette non linéarité dont par exemple LME-ULS (Moosbrugger et al., 1993, 1996) ; PLS (Chin et al., 2003) ; LISREL 8-WLS et LISREL8-WLSA (Jöreskog et al., 1996; Jöreskog and Yang, 1997) et (Yang Jonsson, 1997; Schermelleh-Engel et al., 1998) qui nécessitent tout de même un nombre d'observations élevé. Par ailleurs, la méthode 2SLS, proposée par (Bollen, 1995, 1996) et (Schermelleh-Engel et al., 1998) se révèle lente et adaptée seulement à des modèles simples. Il a fallut attendre les travaux de Moosbrugger et al. (1997); Klein and Moosbrugger (2000) et Schermelleh-Engel et al. (1998) pour la création de la méthode LMS-ML qui considère enfin la non-linéarité des VL endogènes en utilisant aussi des distributions non normales. Les auteurs proposent de généraliser le type d'interaction simple (2.9) à plusieurs interactions simultanées et formalise l'équation structurelle par :

$$\eta = \alpha + \Gamma\xi + \xi'\Omega\xi + \zeta \tag{2.10}$$

où Ω est une matrice de dimension $(n \times n)$ triangulaire supérieure avec des zéros sur la diagonale.
L'approche LMS-ML est fondée sur les modèles de mélanges. Elle partitionne les n VL exogènes en deux groupes $z_1^\star = (z_1, ..., z_k)$ et $z_2^\star = (z_{k+1}, ..., z_n)$ en fonction respectivement de leur nature linéaire ou non linéaire. Ces deux groupes constituent des classes latentes et la

distribution normale des observations (x, y) sera considérée comme la marginale du vecteur "augmenté" $\left(z_1^\star, x, y\right)$ où :

$$z_1^\star \sim \mathcal{N}\left(0, I_k\right)$$

et

$$(x, y)\,|z_1^\star \sim \mathcal{N}\left(\mu\left(z_1^\star\right), \Sigma\left(z_1^\star\right)\right)$$

où $\mu\left(z_1^\star\right)$ et $\Sigma\left(z_1^\star\right)$ sont les concaténations respectives des μ_k associés aux $\mu\left(z_k\right)$ et des Σ_k associés aux $\Sigma\left(z_k\right)$. La méthode repose sur l'utilisation de l'algorithme EM pour l'estimation par maximum de vraisemblance des paramètres μ_k et Σ_k du modèle de mélange. La log-vraisemblance utilisée est celle des observations (x, y) complétées par les classes latentes auxquelles appartiennent les VL exogènes. Ils estiment alors les classes de chacune des VL du vecteur ξ mais il n'est pas question dans cette approche de l'estimation des scores des VL exogènes ou endogènes.

2.1.4 Conclusion

La littérature nous apprend que les deux familles de méthodes d'estimation des modèles à équations structurelles que sont LISREL et PLS ont constamment été la source de travaux de recherche ayant pour but l'amélioration de la gestion des données manquantes, l'adjonction de fonctionnalités permettant l'estimation des concepts latents tels que les VL ou de nouvelles interactions latentes nées de la généralisation du modèle à équations structurelles. L'ensemble des contributions présentées dans ce chapitre sont fondées sur l'algorithme EM. Ce dernier, très populaire par ses performances et sa simplicité d'implémentation, a depuis sa création contribué à améliorer de nombreuses techniques statistiques. D'ailleurs, l'un des premier champs de la statistique à en bénéficier fut celui des modèles linéaires mixtes à effets aléatoires. Nous avons pour objectif de contribuer aux modèles à équations structurelles par une technique d'estimation utilisant l'algorithme EM. Avant d'introduire cette utilisation, la section suivante présente l'algorithme EM.

2.2 L'algorithme EM

L'algorithme EM de Dempster, Laird et Rubin (1977) est une procédure générale pour maximiser la vraisemblance en présence de données manquantes. Dans le cadre des modèles structurels à facteurs, les données manquantes correspondent aux facteurs. Afin d'estimer les paramètres du modèle, cet algorithme procède en deux étapes nommées E (pour "Expectation") et M (pour "Maximization").

2.2.1 Structure générale de l'algorithme EM

Soit y un vecteur aléatoire pour lequel les éléments sont les données observées y_i, $i \in [\![1, n]\!]$, dont la densité notée $p(y;\theta)$ dépend du vecteur de paramètres inconnus $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ dans l'espace Θ. On note f un vecteur de variables aléatoires latentes qui représente les données manquantes. La concaténation de y (les données observées incomplètes) avec f (les données non observables ou manquantes) en un vecteur (y, f) correspond aux données complétées.
Pour estimer les paramètres par maximum de vraisemblance, l'objectif est alors de maximiser la log-vraisemblance marginale correspondant aux données observées (c'est à dire $p(y;\theta)$). Mais cette maximisation est souvent difficile. L'algorithme EM est une manière indirecte et plus simple de maximiser la vraisemblance marginale. L'idée sous-jacente d'EM est de maximiser l'espérance conditionnelle de $\mathcal{L}(\theta; y, f)$ prise par rapport à la distribution conditionnelle des données non observables f sachant les données observées y. On note cette dernière $p(f|y;\theta)$, ce qui permet d'obtenir une quantité déterministe dépendante de θ.
Le principe de l'algorithme EM est le suivant :

- Étape E :
  - on calcule la densité conditionnelle $p(f|y;\theta)$;
  - on intègre la log-vraisemblance jointe $\mathcal{L}(\theta; y, f)$ sur f à l'aide de la densité conditionnelle $p(f|y;\theta)$;
- Étape M :
  - on maximise le résultat sur θ ;
  - on met à jour la valeur de θ par sa valeur estimée.

En effet,

$$\begin{aligned}
\mathcal{L}(\theta; y) &= ln\left[p(y;\theta)\right] \\
&= ln\left[p(y, f;\Theta)\right] \\
&= ln \int_f p(y, f;\theta)\, df \\
&= ln \int_f \frac{p(y, f;\theta)}{p(f|y;\theta)} p(f|y;\theta)\, df \\
&\geq \int_f ln\left\{\frac{p(y, f;\theta)}{p(f|y;\theta)}\right\} p(f|y;\theta)\, df
\end{aligned} \tag{2.11}$$

où nous avons appliqué l'inégalité de Jensen en nous basant sur la concavité de la fonction logarithme. Nous définissons :

$$\mathcal{L}(\theta; q) := \int_f ln\left\{\frac{p(y, f;\theta)}{p(f|y;\theta)}\right\} p(f|y;\theta)\, df. \tag{2.12}$$

En outre,

$$\begin{aligned}
\mathcal{L}(\theta; q) &= \int_f ln\left[p(y, f;\theta)\right] p(f|y;\theta)\, df - \int_f ln\left[p(f|y;\theta)\right] p(f|y;\theta)\, df \\
&= \mathbb{E}_y^f\left[ln\left[p(y, f;\theta)\right]\right] - \int_f ln\left[p(f|y;\theta)\right] p(f|y;\theta)\, df
\end{aligned} \tag{2.13}$$

où $\mathbb{E}_y^f[.]$ est l'espérance par rapport à la distribution conditionnelle des données f sachant y. Pour $\theta = \theta^{[t]}$, nous avons $\int_f ln\left[p\left(f|y;\theta\right)\right] p\left(f|y;\theta\right) df = K\left(p\left(f|y;\theta\right)|p\left(f|y;\theta\right)\right)$ la divergence de Kullback-Leibler. Or,

$$\forall p_1, p_2,\ p_1 = p_2 \Longleftrightarrow K\left(p_1|p_2\right) = 0.$$

Ceci se justifie par :

$$\begin{aligned}\forall p_1, p_2,\ K\left(p_1|p_2\right) &= -\int_z ln\frac{p_1(z)}{p_2(z)} p_2(z)dz \\ &\geq -ln\left(\int_z \frac{p_1(z)}{p_2(z)} p_2(z)dz\right) \\ &= 0\end{aligned}$$

car la fonction $-ln(.)$ est strictement convexe. On obtient donc l'égalité que pour $p_1 = p_2$. Ainsi, maximiser $\mathcal{L}\left(\theta; q\right)$ revient à maximiser $\mathcal{L}\left(\theta; y\right)$. Alors, une vision du maximum de vraisemblance peut être la suivante :

$$\begin{aligned}&\frac{\partial}{\partial\theta}\mathcal{L}\left(\theta; y\right) = 0 \\ \Longleftrightarrow\ &\frac{\partial}{\partial\theta}\mathbb{E}_y^f\left[ln\left[p\left(y, f; \theta\right)\right]\right] = 0 \\ \Longleftrightarrow\ &\frac{\partial}{\partial\theta}\mathbb{E}_y^f\left[\mathcal{L}\left(\theta; y, f\right)\right] = 0.\end{aligned}$$

D'autre part, à chaque itération $[t]$, avec la valeur courante $\theta^{[t]}$, les bornes d'intégration et $p\left(f|y;\theta = \theta^{[t]}\right)$ ne dépendant pas de θ, on peut donc permuter "intégration sur f" et "dérivation par rapport à θ", ce qui permet de poursuivre par les équivalences :

$$\begin{aligned}\Longleftrightarrow\ &\frac{\partial}{\partial\theta}\int_f ln\left[p\left(y, f; \theta\right)\right] p\left(f|y;\theta = \theta^{[t]}\right) df = 0 \\ \Longleftrightarrow\ &\int_f \frac{\partial}{\partial\theta} ln\left[p\left(y, f; \theta\right)\right] p\left(f|y;\theta = \theta^{[t]}\right) df = 0\end{aligned}$$

$$\Longleftrightarrow \mathbb{E}_y^f\left[\frac{\partial}{\partial\theta}\mathcal{L}\left(\theta; y, f\right)\right] = 0. \tag{2.14}$$

Sur cette base, à chaque itération [t] l'algorithme EM réalise les deux étapes suivantes :

1. **Étape E** : Avec la valeur courante $\theta^{[t]}$ on calcule :

$$\mathcal{Q}\left(\theta, \theta^{[t]}\right) := \mathbb{E}_y^f\left[ln\left[p\left(y, f; \theta = \theta^{[t]}\right)\right]\right] = \int_f \mathcal{L}\left(\theta; y, f\right) p\left(f|y;\theta = \theta^{[t]}\right) df \tag{2.15}$$

2. **Étape M** : On actualise la valeur courante par maximisation de la fonction obtenue à l'étape E par rapport à θ :

$$\theta^{[t+1]} = \underset{\theta}{\text{argmax}}\mathcal{Q}\left(\theta, \theta^{[t]}\right) \tag{2.16}$$

La nouvelle valeur $\theta^{[t+1]}$ va permettre de mettre à jour (2.15) dans l'étape E. On repasse alors à l'étape M, et ainsi de suite jusqu'à convergence de l'algorithme.

Remarque. Pour l'étape M, la maximisation de

$$\mathcal{L}\left(\theta; q\right) = \mathbb{E}_y^f\left[\mathcal{L}\left(\theta; y, f\right)\right] - \int_f ln\left[p\left(f|y;\theta\right)\right] p\left(f|y;\theta\right) df$$

revient à la maximisation de $\mathcal{L}\left(\theta; y\right)$ non pas parce que $\int_f ln\left[p\left(f|y;\theta\right)\right] p\left(f|y;\theta\right) df = 0$ comme cela est le cas pour l'étape E avec $\theta = \theta^{[t]}$, mais parce que $\int_f ln\left[p\left(f|y;\theta\right)\right] p\left(f|y;\theta\right) df$ devient indépendant de θ. En effet, à l'étape M la fonction $argmax(.)$ est effectuée sur θ et non sur $\theta = \theta^{[t]}$.

2.2.2 Preuve de la convergence de l'algorithme EM

La convergence monotone de EM s'obtient en montrant que :

$$\mathcal{L}\left(\theta^{[t+1]}; y\right) \geq \mathcal{L}\left(\theta^{[t]}; y\right).$$

Nous avons,

$$\begin{aligned} & p\left(y, f; \theta\right) = p\left(y; \theta\right) p\left(f|y; \theta\right) \\ \Longrightarrow & \mathcal{L}\left(\theta; y, f\right) = \mathcal{L}\left(\theta; y\right) + \mathcal{L}\left(\theta; f|y\right). \end{aligned}$$

En prenant l'espérance conditionnelle correspondant à la valeur précédente du paramètre θ, nous obtenons :

$$\begin{aligned} & \mathbb{E}_y^f\left[\mathcal{L}\left(\theta = \theta^{[t]}; y, f\right)\right] = \mathbb{E}_y^f\left[\mathcal{L}\left(\theta = \theta^{[t]}; y\right)\right] + \mathbb{E}_y^f\left[\mathcal{L}\left(\theta = \theta^{[t]}; f|y\right)\right] \\ \Longleftrightarrow & \mathcal{Q}\left(\theta, \theta^{[t]}\right) = \mathcal{L}\left(\theta = \theta^{[t]}; y\right) + \mathcal{J}\left(\theta, \theta^{[t]}\right). \end{aligned}$$

En effet, $\mathbb{E}_y^f\left[\mathcal{L}\left(\theta = \theta^{[t]}; y\right)\right] = \mathcal{L}\left(\theta = \theta^{[t]}; y\right) \int_f p\left(f|y; \theta\right) df = \mathcal{L}\left(\theta = \theta^{[t]}; y\right)$.
À l'étape M, nous avons :

$$\begin{aligned} \theta^{[t+1]} &= \underset{\theta}{\operatorname{argmax}} \mathcal{Q}\left(\theta, \theta^{[t]}\right) \\ \Longrightarrow & \mathcal{Q}\left(\theta, \theta^{[t+1]}\right) \geq \mathcal{Q}\left(\theta, \theta^{[t]}\right) \end{aligned}$$

donc

$$\mathcal{L}\left(\theta = \theta^{[t+1]}; y\right) + \mathcal{J}\left(\theta; \theta^{[t+1]}\right) \geq \mathcal{L}\left(\theta = \theta^{[t]}; y\right) + \mathcal{J}\left(\theta; \theta^{[t]}\right). \tag{2.17}$$

D'autre part,

$$\begin{aligned} & \mathcal{J}\left(\theta; \theta^{[t+1]}\right) - \mathcal{J}\left(\theta; \theta^{[t]}\right) = \mathbb{E}_y^f\left[\mathcal{L}\left(\theta = \theta^{[t+1]}; f|y\right)\right] - \mathbb{E}_y^f\left[\mathcal{L}\left(\theta = \theta^{[t]}; f|y\right)\right] \\ & = \mathbb{E}_y^f\left[ln\left[p\left(f|y; \theta = \theta^{[t+1]}\right)\right] - ln\left[p\left(f|y; \theta = \theta^{[t]}\right)\right]\right] \\ & = \mathbb{E}_y^f\left[ln\left\{\frac{p\left(f|y; \theta = \theta^{[t+1]}\right)}{p\left(f|y; \theta = \theta^{[t]}\right)}\right\}\right] \\ & \leq ln\left\{\mathbb{E}_y^f\left[\frac{p\left(f|y; \theta = \theta^{[t+1]}\right)}{p\left(f|y; \theta = \theta^{[t]}\right)}\right]\right\} \\ & = ln(1) = 0. \end{aligned} \tag{2.18}$$

D'où d'après (2.17), nous obtenons :

$$\mathcal{L}\left(\theta = \theta^{[t+1]}; y\right) \geq \mathcal{L}\left(\theta = \theta^{[t]}; y\right).$$

Nous en déduisons que la vraisemblance marginale des observations est croissante majorée. Elle converge donc vers un maximum local ou un point-selle de la vraisemblance.

Remarque. Contrairement au cas où l'on observe en même temps y et f, il peut exister des maxima locaux qui bloquent la convergence de l'algorithme vers le maximum global. EM peut s'avérer très sensible aux valeurs initiales choisies.

Remarque. Le nombre d'itérations nécessaires pour approcher de manière satisfaisante un maximum local n'est pas connu au préalable. Il n'y a pas de procédure générale pour prévoir le nombre d'itérations à opérer. Pour finir, l'algorithme EM est très apprécié pour sa simplicité d'implémentation, ses itérations peu gourmandes en temps de calcul et le peu de mémoire (de stockage) nécessaire lors de son utilisation.

2.3 Méthode d'estimation par algorithme EM pour un modèle à une équation structurelle et un facteur par bloc de variables observées

Pour exposer progressivement la méthode d'estimation par algorithme EM, nous la présentons dans un premier temps sur un modèle à une équation structurelle comportant seulement deux groupes de VO. L'un des deux groupes est lié à un facteur explicatif et l'autre à un facteur dépendant (cf. figure 2.3).

2.3.1 Modèle à deux blocs : l'un dépendant et l'autre explicatif

Le groupe dépendant Y (resp. explicatif X) est structuré autour du facteur latent g (resp. f). Au niveau du modèle interne, le facteur g dépend du facteur f. Cette relation de cause à effet est formalisée par une unique équation structurelle. Quant aux deux modèles externes, chacun formalise un groupe de variables Y (resp. X) comme dépendant du facteur g (resp. f).

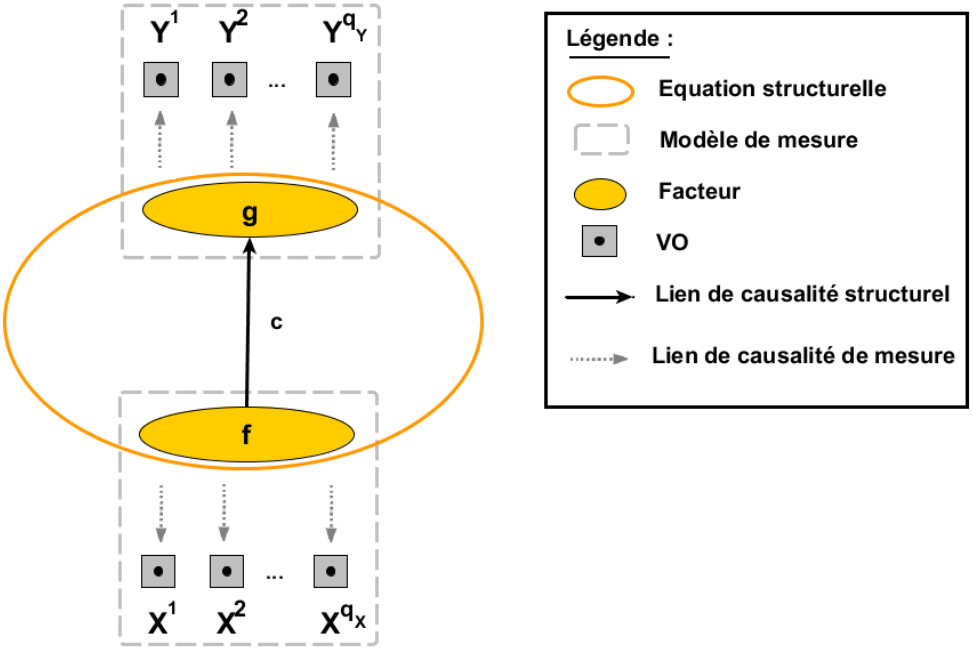


FIGURE 2.3 – Un modèle à deux groupes de VO : un dépendant et un explicatif, à une équation structurelle.

Soit q_Y (resp. q_X) le nombre de variables étudiées Y_j (resp.X_j). Les facteurs g et f sont de longueur n, le nombre d'observations dont on dispose pour chacune des variables. μ_j^Y (resp. μ_j^X) est le vecteur des paramètres moyennes de Y_j (resp. X_j), b_j (resp. a_j) sont les vecteurs de coefficients pondérateurs associés à Y_j (resp. X_j) et ε_{ij}^Y (resp. ε_{ij}^X) les erreurs associées à la variable Y_j (resp. X_j). Le coefficient pondérateur c et les erreurs ε_i^g sont associés au facteur dépendant g. Le modèle peut alors être formulé par le système d'équations :

$$\begin{cases} Y &= \mathbb{1}_n {\mu_Y}' + g b' + \varepsilon^Y \\ X &= \mathbb{1}_n {\mu_X}' + f a' + \varepsilon^X \\ g &= fc + \varepsilon^g \end{cases} \tag{2.19}$$

où l'ensemble des paramètres du modèle est
$\theta = \left\{\mu^Y, \mu^X, b, a, c, \{\forall j \in [\![1, q_Y]\!]; \sigma^2_{Yj}\}, \{\forall j \in [\![1, q_X]\!]; \sigma^2_{Xj}\}\right\}$, avec $b' = \left(b_1, ..., b_{q_Y}\right)$, $a' = \left(a_1, ..., a_{q_X}\right)$ et c un scalaire.
Les hypothèses de ce modèle sont :

— $\varepsilon_i^X \sim \mathcal{N}(0, \psi_X)$ où $\psi_X = diag\left\{\sigma^2_{X,1}, \ldots, \sigma^2_{X,q_X}\right\}$ est une matrice de dimension $q_X \times q_X$;
— $\varepsilon_i^Y \sim \mathcal{N}(0, \psi_Y)$ où $\psi_Y = diag\left\{\sigma^2_{Y,1}, \ldots, \sigma^2_{Y,q_Y}\right\}$ est une matrice de dimension $q_Y \times q_Y$;
— $f_i \sim \mathcal{N}(0, 1)$;
— $\varepsilon_i^g \sim \mathcal{N}(0, 1)$;
— $g_i \sim \mathcal{N}\left(0, c^2 + 1\right)$;
— ε_i^X, $\varepsilon_{i'}^Y$,$\varepsilon_{i''}^g$ et f_ℓ sont mutuellement indépendants pour tout i, i', i'', ℓ.

En effet, d'après la contrainte d'identifiabilité :

$$\begin{aligned}\mathbb{V}(g) &= c^2\mathbb{V}(f) + \mathbb{V}(\varepsilon_i^g)\\ &= c^2 + 1\end{aligned}$$

2.3.2 L'algorithme EM pour le modèle à deux blocs : l'un dépendant et l'autre explicatif

Pour i une observation, le modèle (2.19) peut être formulé selon le système d'équations suivant :

$$\left\{\begin{array}{lcl} y_i' &=& \mu_Y{}' + g_i b' + \varepsilon_i{}^{Y'}\\ x_i' &=& \mu_X{}' + f_i a' + \varepsilon_i{}^{X'}\\ g_i &=& f_i c + \varepsilon_i{}^g \end{array}\right. \tag{2.20}$$

Soient les concaténations $z = (Y, X)$ et $h = (g, f)$. Alors, la log vraisemblance s'écrit :

$$\begin{aligned}\mathcal{L}(\theta; z, h) = -\frac{1}{2}\sum_{i=1}^{n} & ln|\psi_Y| + ln|\psi_X|\\ &+ \left(y_i - \mu^Y - g_i b\right)'\psi_Y^{-1}\left(y_i - \mu^Y - g_i b\right)\\ &+ \left(x_i - \mu^X - f_i a\right)'\psi_X^{-1}\left(x_i - \mu^X - f_i a\right)\\ &+ \left(g_i - f_i c\right)^2\\ &+ f_i^2 + \lambda\end{aligned} \tag{2.21}$$

où λ est une constante. Cette log vraisemblance est dite "log vraisemblance complétée de z". Elle correspond à la log-vraisemblance des données observées z complétées par les données manquantes h. Avant d'expliciter les étapes de calcul permettant son obtention, dans le cadre des hypothèses du modèle (2.19), on peut développer la log-vraisemblance complétée comme suit :

$$\begin{aligned}\mathcal{L}(\theta; z, h) = -\frac{1}{2}\sum_{i=1}^{n} & ln\left(\prod_{j=1}^{q_Y}\sigma^2_{Y\,j}\right) + ln\left(\prod_{j=1}^{q_X}\sigma^2_{X\,j}\right) +\\ &\sum_{j=1}^{q_Y}\left\{\left(y_{ij} - b_j g_i - \mu_j^Y\right)^2\sigma^{-2}_{Y\,j}\right\} + \sum_{j=1}^{q_X}\left\{\left(x_{ij} - a_j f_i - \mu_j^X\right)^2\sigma^{-2}_{X\,j}\right\} +\\ &\left(g_i - f_i c\right)^2 + f_i^2 + \lambda\end{aligned}$$

$$\begin{aligned}\mathcal{L}(\theta; z, h) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{q_Y} & ln\left(\sigma^2_{Y\,j}\right) + \sum_{j=1}^{q_X} ln\left(\sigma^2_{X\,j}\right) +\\ &\sum_{j=1}^{q_Y}\left\{\left(y_{ij} - b_j g_i - \mu_j^Y\right)^2\sigma^{-2}_{Y\,j}\right\} + \sum_{j=1}^{q_X}\left\{\left(x_{ij} - a_j f_i - \mu_j^X\right)^2\sigma^{-2}_{X\,j}\right\} +\\ &\left(g_i - f_i c\right)^2 + f_i^2 + \lambda\end{aligned}$$

$$\mathcal{L}(\theta; z, h) = -\frac{n}{2}\left\{\sum_{j=1}^{q_Y} ln\left(\sigma^2_{Y\,j}\right) + \sum_{j=1}^{q_X} ln\left(\sigma^2_{X\,j}\right)\right\} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{q_Y}\left\{\left(y_{ij} - b_j g_i - \mu_j^Y\right)^2 \sigma_{Y}^{-2}{}_{j}\right\} + \sum_{j=1}^{q_X}\left\{\left(x_{ij} - a_j f_i - \mu_j^X\right)^2 \sigma_{X}^{-2}{}_{j}\right\} + (g_i - f_i c)^2 + f_i^2 + \lambda$$

où λ est une constante et l'ensemble des paramètres du modèle est $\theta = \left\{\mu^Y, \mu^X, b, a, c, \{\forall j \in [\![1, q_Y]\!]; \sigma^2_{Y\,j}\}, \{\forall j \in [\![1, q_X]\!]; \sigma^2_{X\,j}\}\right\}$. Par la suite les écritures ne seront pas autant développées (expression des ψ en les σ associés) afin d'être plus concis et de limiter la redondance. En effet, l'idée du développement est toujours la même.

La log vraisemblance complétée de z dans le cas du modèle à deux groupes

Nous avons,

$$\begin{aligned} p(z_i, h_i; \theta) &= p(y_i, x_i, g_i, f_i; \theta) \\ &= p(y_i, x_i | g_i, f_i; \theta)\, p(g_i, f_i; \theta) \\ &= p(y_i, x_i | g_i, f_i; \theta)\, p(g_i, f_i; \theta)\, p(f_i) \\ &= p(x_i | y_i, g_i, f_i; \theta)\, p(y_i | g_i, f_i; \theta)\, p(g_i | f_i; \theta)\, p(f_i) \\ &= p(x_i | f_i; \theta)\, p(y_i | g_i; \theta)\, p(g_i | f_i; \theta)\, p(f_i) \end{aligned}$$

d'où,

$$\mathcal{L}(\theta; z_i, h_i) = \mathcal{L}(\theta; x_i | f_i) + \mathcal{L}(\theta; y_i | g_i) + \mathcal{L}(\theta; g_i | f_i) + \mathcal{L}(f_i).$$

D'après l'écriture du modèle et compte tenu des propriétés des distributions gaussiennes on obtient :
$x_i \sim \mathcal{N}\left(\mu^X, aa' + \psi_X\right)$
$x_i | f_i \sim \mathcal{N}\left(\mu^X + f_i a', \psi_X\right)$
$y_i \sim \mathcal{N}\left(\mu^Y, b(c^2+1)b' + \psi_Y\right)$
$y_i | g_i \sim \mathcal{N}\left(\mu^Y + g_i b, \psi_Y\right)$
$g_i \sim \mathcal{N}\left(0, c^2 + 1\right)$
$g_i | f_i \sim \mathcal{N}(f_i c, 1)$.
D'où l'expression de la log vraisemblance complétée (2.21).

Estimateurs des paramètres : formules explicites caractérisant les solutions

Pour obtenir les estimateurs des paramètres il faut résoudre (2.14) en utilisant la fonction de vraisemblance complétée (2.21). Ce qui revient à résoudre le système suivant :

$$\begin{cases} \mathbb{E}_{z_i}^{h_i}\left[\frac{\partial}{\partial \mu^Y}\mathcal{L}(z,h)\right] &= 0 \\ \mathbb{E}_{z_i}^{h_i}\left[\frac{\partial}{\partial \mu^X}\mathcal{L}(z,h)\right] &= 0 \\ \mathbb{E}_{z_i}^{h_i}\left[\frac{\partial}{\partial a}\mathcal{L}(z,h)\right] &= 0 \\ \mathbb{E}_{z_i}^{h_i}\left[\frac{\partial}{\partial b}\mathcal{L}(z,h)\right] &= 0 \\ \mathbb{E}_{z_i}^{h_i}\left[\frac{\partial}{\partial c}\mathcal{L}(z,h)\right] &= 0 \\ \mathbb{E}_{z_i}^{h_i}\left[\frac{\partial}{\partial \sigma^2_Y}\mathcal{L}(z,h)\right] &= 0 \\ \mathbb{E}_{z_i}^{h_i}\left[\frac{\partial}{\partial \sigma^2_X}\mathcal{L}(z,h)\right] &= 0 \end{cases} \tag{2.22}$$

Dans le cas où les matrices de variances des erreurs sont telles que $\psi_Y = \sigma_Y^2 Id_{q_Y}$ et $\psi_X = \sigma_X^2 Id_{q_X}$, la résolution de ce système (2.22) est équivalente à celle du suivant :

$$\begin{cases} \mathbb{E}_{z_i}^{h_i}\left[\sum_{i=1}^{n} \psi_Y^{-1}\left(y_i - \mu^Y - g_i b\right)\right] &= 0 \\ \mathbb{E}_{z_i}^{h_i}\left[\sum_{i=1}^{n} \psi_X^{-1}\left(x_i - \mu^X - f_i a\right)\right] &= 0 \\ \mathbb{E}_{z_i}^{h_i}\left[\sum_{i=1}^{n} f_i \psi_X^{-1}\left(x_i - \mu^X - f_i a\right)\right] &= 0 \\ \mathbb{E}_{z_i}^{h_i}\left[\sum_{i=1}^{n} g_i \psi_Y^{-1}\left(y_i - \mu^Y - g_i b\right)\right] &= 0 \\ \mathbb{E}_{z_i}^{h_i}\left[\sum_{i=1}^{n} f_i\left(g_i - f_i c\right)\right] &= 0 \\ \mathbb{E}_{z_i}^{h_i}[n\ q_Y\ \sigma_Y^{-2} - \sigma_Y^{-4}\sum_{i=1}^{n} ||y_i - \mu^Y - g_i b||^2] &= 0 \\ \mathbb{E}_{z_i}^{h_i}[n\ q_X\ \sigma_X^{-2} - \sigma_X^{-4}\sum_{i=1}^{n} ||x_i - \mu^X - f_i a||^2] &= 0 \end{cases} \tag{2.23}$$

Or, pour chaque observation i, $h_i|z_i \sim \mathcal{N}\left(m_i = \begin{pmatrix} m_{1i} \\ m_{2i} \end{pmatrix}, \Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix}\right)$ tel que,

$$m_i = \begin{pmatrix} \left(c^2+1\right) b' & ca' \\ cb' & a' \end{pmatrix} \left(\begin{pmatrix} \left(c^2+1\right) bb' + \Psi_Y & cba' \\ cab' & aa' + \Psi_X \end{pmatrix} \right)^{-1} \begin{pmatrix} y_i - \mu^Y \\ x_i - \mu^X \end{pmatrix}$$

$$\Gamma = \begin{pmatrix} c^2+1 & c \\ c & 1 \end{pmatrix} - \begin{pmatrix} \left(c^2+1\right) b' & ca' \\ cb' & a' \end{pmatrix} \left(\begin{pmatrix} \left(c^2+1\right) bb' + \Psi_Y & cba' \\ cab' & aa' + \Psi_X \end{pmatrix} \right)^{-1} \begin{pmatrix} \left(c^2+1\right) b & cb \\ ca & a \end{pmatrix}$$

Ainsi, en notant pour tout $i \in [\![1, n]\!]$,
$\widetilde{\gamma_i} = \mathbb{E}_{z_i}^{h_i}\left[g_i^2\right] = \left(\mathbb{E}_{z_i}^{h_i}\left[g_i\right]\right)^2 + \mathbb{V}_{z_i}^{h_i}\left[g_i\right] = m_{1i}^2 + \Gamma_{11}$;
$\widetilde{\phi_i} = \mathbb{E}_{z_i}^{h_i}[f_i^2] = \left(\mathbb{E}_{z_i}^{h_i}\left[f_i\right]\right)^2 + \mathbb{V}_{z_i}^{h_i}[f_i] = m_{2i}^2 + \Gamma_{22}$;
$\widetilde{g_i} = \mathbb{E}_{z_i}^{h_i}[g_i] = m_{1i}$;
$\widetilde{f_i} = \mathbb{E}_{z_i}^{h_i}[f_i] = m_{2i}$,

le système à résoudre devient :

$$\begin{cases} \sum_{i=1}^{n}\left(y_i - \mu^Y - \widetilde{g_i} b\right) &= 0 \\ \sum_{i=1}^{n}\left(x_i - \mu^X - \widetilde{f_i} a\right) &= 0 \\ \sum_{i=1}^{n} \widetilde{g_i} y_i - \widetilde{g_i}\mu^Y - \widetilde{\gamma_i} b &= 0 \\ \sum_{i=1}^{n} \widetilde{f_i} x_i - \widetilde{f_i}\mu^X - \widetilde{\phi_i} a &= 0 \\ \sum_{i=1}^{n} \Gamma_{12} + \widetilde{f_i}\widetilde{g_i} - \widetilde{\phi_i} c &= 0 \\ n\ q_Y\ \sigma_Y^{-2} - \sigma_Y^{-4}\sum_{i=1}^{n} ||y_i - \mu^Y||^2 + ||b||^2\widetilde{\gamma_i} - 2\left(y_i - \mu^Y\right)\widetilde{g_i} b &= 0 \\ n\ q_X\ \sigma_X^{-2} - \sigma_X^{-4}\sum_{i=1}^{n} ||x_i - \mu^X||^2 + ||a||^2\widetilde{\phi_i} - 2\left(x_i - \mu^X\right)\widetilde{f_i} a &= 0 \end{cases} \tag{2.24}$$

car $\mathbb{E}_{z_i}^{h_i}[f_i g_i] = Cov_{z_i}^{h_i}[f_i, g_i] + \mathbb{E}_{z_i}^{h_i}[f_i]\mathbb{E}_{z_i}^{h_i}[g_i] = \Gamma_{12} + \widetilde{f_i}\widetilde{g_i}$. D'où les formules explicites des paramètres caractérisant les solutions de ce système (2.24) (et donc de (2.14)) :

$$
\begin{aligned}
&\widehat{\mu^Y} = \overline{y} - \hat{b}\overline{\widetilde{g}} \\
&\widehat{\mu^X} = \overline{x} - \hat{a}\overline{\widetilde{f}} \\
&\hat{b} = \frac{\overline{\widetilde{g}y} - \overline{y}\,\overline{\widetilde{g}}}{\overline{\widetilde{\gamma}} - \overline{\widetilde{g}}^2} \\
&\hat{a} = \frac{\overline{\widetilde{f}x} - \overline{x}\overline{\widetilde{f}}}{\overline{\widetilde{\phi}} - \overline{\widetilde{f}}^2} \\
&\hat{c} = \frac{\Gamma_{12} + \overline{\widetilde{f}_i\widetilde{g}_i}}{\overline{\widetilde{\phi}_i}} \\
&\widehat{\sigma_Y^2} = \frac{1}{nq_Y}\sum_{i=1}^{n}\left\{||y_i - \widehat{\mu^Y}||^2 + ||\hat{b}||^2\widetilde{\gamma}_i - 2\left(y_i - \widehat{\mu^Y}\right)'\hat{b}\widetilde{g}_i\right\} \\
&\widehat{\sigma_X^2} = \frac{1}{nq_X}\sum_{i=1}^{n}\left\{||x_i - \widehat{\mu^X}||^2 + ||\hat{a}||^2\widetilde{\phi}_i - 2\left(x_i - \widehat{\mu^X}\right)'\hat{a}\widetilde{f}_i\right\}
\end{aligned}
\tag{2.25}
$$

Estimateurs des paramètres obtenus dans le cas ψ_Y et ψ_X diagonales

Dans le cas où les de matrices de variances des erreurs sont diagonales tel que dans les hypothèses du modèle, les estimateurs de b, a, c, μ^Y et μ^X restent inchangés. Néanmoins, $\forall j \in [\![1, q_Y]\!]$, ${\sigma_Y}_j^2$ et $\forall j \in [\![1, q_X]\!]$, ${\sigma_X}_j^2$ sont des paramètres à estimer.
Or,

$$
\begin{aligned}
&\frac{\partial}{\partial {\sigma_X}_j^2}\left[\mathcal{L}(z,h)\right] = -\frac{n}{2}\,{\sigma_X}_j^{-2} + \frac{1}{2}\sum_{i=1}^{n}{\sigma_X}_j^{-4}\left(x_{ij} - \mu_j^X - f_i a_j\right)^2 \\
&\Rightarrow \mathbb{E}_{z_i}^{h_i}\left[\frac{\partial}{\partial {\sigma_X}_j^2}\left[\mathcal{L}(z,h)\right]\right] = 0 \\
&\Leftrightarrow \widehat{{\sigma_X}_j^2} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(x_{ij} - \widehat{\mu_j^X}\right)^2 + \hat{a_j}^2\widetilde{\phi}_i - 2\left(x_{ij} - \widehat{\mu_j^X}\right)\hat{a_j}\widetilde{f}_i\right\}.
\end{aligned}
$$

De manière similaire,

$$
\widehat{{\sigma_Y}_j^2} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(y_{ij} - \widehat{\mu_j^Y}\right)^2 + \hat{b_j}^2\widetilde{\gamma}_i - 2\left(y_{ij} - \widehat{\mu_j^Y}\right)\hat{b_j}\widetilde{g}_i\right\}
$$

Par conséquent, dans ce cas les formules explicites des paramètres caractérisant les solutions sont :

$$
\begin{aligned}
\widehat{\mu^Y} &= \bar{y} - \hat{b}\bar{\widetilde{g}} \\
\widehat{\mu^X} &= \bar{x} - \hat{a}\bar{\widetilde{f}} \\
\hat{b} &= \frac{\overline{\widetilde{g}y} - \bar{y}\,\bar{\widetilde{g}}}{\bar{\widetilde{\gamma}} - \bar{\widetilde{g}}^2} \\
\hat{a} &= \frac{\overline{\widetilde{f}x} - \bar{x}\bar{\widetilde{f}}}{\bar{\widetilde{\phi}} - \bar{\widetilde{f}}^2} \\
\hat{c} &= \frac{\Gamma_{12} + \overline{\widetilde{f}_i\widetilde{g}_i}}{\overline{\widetilde{\phi}_i}} \\
\widehat{\sigma_{Y}{}_j^2} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(y_{ij} - \widehat{\mu_j^Y}\right)^2 + \hat{b_j}^2\widetilde{\gamma}_i - 2\left(y_{ij} - \widehat{\mu_j^Y}\right)\hat{b_j}\widetilde{g}_i\right\} \\
\widehat{\sigma_{X}{}_j^2} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(x_{ij} - \widehat{\mu_j^X}\right)^2 + \hat{a_j}^2\widetilde{\phi}_i - 2\left(x_{ij} - \widehat{\mu_j^X}\right)\hat{a_j}\widetilde{f}_i\right\}
\end{aligned}
\tag{2.26}
$$

À partir des formules des estimateurs des paramètres obtenus, il devient possible de construire une procédure itérative permettant, à partir d'une initialisation pertinente de l'ensemble des paramètres, d'actualiser le paramètre courant $\theta^{[t-1]}$ en $\theta^{[t]}$ à la t-ième itération. Cette procédure est stoppée sur satisfaction d'un critère d'arrêt et la dernière valeur courante retenue comme estimation des paramètres. L'estimation par algorithme EM procédera en estimant respectivement les facteurs g et f par les paramètres de moyenne $\widetilde{g} = {m_1}^{[t]}$ et $\widetilde{f} = {m_2}^{[t]}$ de la distribution conditionnelle $h|z$ obtenus à l'itération précédant la satisfaction du critère d'arrêt.

L'algorithme de l'estimation par EM

Après initialisation des valeurs des paramètres θ, l'algorithme procède en deux étapes à l'itération courante $[t-1]$:

1. Avec la valeur courante $\theta^{[t-1]}$ on calcule pour chaque observation $i \in [\![1, n]\!]$:

$$
\begin{aligned}
\widetilde{\gamma}_i &= {m_1}_i^{[t-1]^2} + \Gamma_{11}^{[t-1]} \\
\widetilde{\phi}_i &= {m_2}_i^{[t-1]^2} + \Gamma_{22}^{[t-1]} \\
\widetilde{g}_i &= {m_1}_i^{[t-1]} \\
\widetilde{f}_i &= {m_2}_i^{[t-1]}
\end{aligned}
$$

2. On actualise la valeur courante $\theta^{[t]}$ en utilisant les formules solutions adéquates ((2.25) ou (2.26)) suivant les hypothèses plus ou moins contraignantes sur les matrices de variance des erreurs. Pour le cas (2.25) on obtient :

$$
\begin{aligned}
\mu^{Y^{[t]}} &= \overline{y} - b^{[t-1]}\overline{\widetilde{g}} \\
\mu^{X^{[t]}} &= \overline{x} - a^{[t-1]}\overline{\widetilde{f}} \\
b^{[t]} &= \frac{\overline{\widetilde{g}y} - \overline{y}\,\overline{\widetilde{g}}}{\overline{\widetilde{\gamma}} - \overline{\widetilde{g}}^2} \\
a^{[t]} &= \frac{\overline{\widetilde{f}x} - \overline{x}\overline{\widetilde{f}}}{\overline{\widetilde{\phi}} - \overline{\widetilde{f}}^2} \\
c^{[t]} &= \frac{\Gamma_{12}^{[t]} + \overline{\widetilde{f}_i\widetilde{g}_i}}{\overline{\widetilde{\phi}_i}} \\
{\sigma_Y}^{2^{[t]}} &= \frac{1}{nq_Y}\sum_{i=1}^{n}\left\{||y_i - \mu^{Y^{[t-1]}}||^2 + ||b^{[t-1]}||^2\widetilde{\gamma}_i - 2\left(y_i - \mu^{Y^{[t-1]}}\right)' b^{[t-1]}\widetilde{g}_i\right\} \\
{\sigma_X}^{2^{[t]}} &= \frac{1}{nq_X}\sum_{i=1}^{n}\left\{||x_i - \mu^{X^{[t-1]}}||^2 + ||a^{[t-1]}||^2\widetilde{\phi}_i - 2\left(x_i - \mu^{X^{[t-1]}}\right)' a^{[t-1]}\widetilde{f}_i\right\}
\end{aligned}
\tag{2.27}
$$

La nouvelle valeur $\theta^{[t]} = \left\{\mu^{Y^{[t]}}, \mu^{X^{[t]}}, b^{[t]}, a^{[t]}, c^{[t]}, \{\forall j \in [\![1, q_Y]\!]; {\sigma^2_Y}_j^{[t]}\}, \{\forall j \in [\![1, q_X]\!]; {\sigma^2_X}_j^{[t]}\}\right\}$ va permettre de mettre à jour $\widetilde{\gamma}_i$, $\widetilde{\phi}_i$, $\widetilde{g}_i$, et $\widetilde{f}_i$ dans l'étape 1. On repasse alors à l'étape 2, et ainsi de suite jusqu'à convergence de l'algorithme.
Dans le cas où ψ_Y et ψ_X sont diagonales, l'algorithme procède de la même manière. Il suffit de changer respectivement les formules des estimateurs de ${\sigma_Y}^{2^{[t+1]}}$ et ${\sigma_X}^{2^{[t+1]}}$ (cf. (2.27)) par les suivantes pour respectivement tout $j \in [\![1, q_Y]\!]$ et $j \in [\![1, q_X]\!]$:

$$
\begin{aligned}
{\sigma_Y}_j^{2^{[t]}} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(y_{ij} - \widehat{\mu_j^Y}^{[t-1]}\right)^2 + \left(\hat{b_j}^{[t-1]}\right)^2\widetilde{\gamma}_i - 2\left(y_{ij} - \widehat{\mu_j^Y}^{[t-1]}\right)\hat{b_j}^{[t-1]}\widetilde{g}_i\right\} \\
{\sigma_X}_j^{2^{[t]}} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(x_{ij} - \widehat{\mu_j^X}^{[t-1]}\right)^2 + \left(\hat{a_j}^{[t-1]}\right)^2\widetilde{\phi}_i - 2\left(x_{ij} - \widehat{\mu_j^X}^{[t-1]}\right)\hat{a_j}^{[t-1]}\widetilde{f}_i\right\}
\end{aligned}
$$

2.4 Conclusion et discussion

L'utilisation de l'algorithme EM permet à la fois une estimation de concepts latents moins contraints (facteurs plutôt que composantes) tout en tenant compte de toutes les équations du modèle lors de la maximisation de la vraisemblance. En effet, les concepts latents conjecturés sous forme de facteurs ne peuvent être estimés par les méthodes de type PLS, RGCCA ou THEME qui, en les contraignant à des composantes, réduisent l'espace des solutions. Quant à LISREL (tout comme pour PLS), les différentes contributions pour l'estimation de concepts latents présentées au début de ce chapitre rencontrent des limites. Bien que, parmi elles, la méthode la plus adaptée à l'estimation des concepts latents soit celle de l'estimation des scores de Jöreskog (2000), la technique d'estimation reste lourde et n'exploite pas la totalité des équations du modèle. Seules les équations de mesures sont prises en compte lors de l'étape de minimisation des moindres carrés. L'équation structurelle semble être négligée : elle n'est pas prise en compte, alors qu'elle a servi à contraindre la variance des VO et son estimation. En revanche, l'approche d'estimation par algorithme EM a l'avantage de maximiser la log-vraisemblance issue du modèle structurel complet. Ainsi, lors de l'estimation des paramètres et des facteurs latents, toutes les équations sont utilisées lors de la maximisation.
Au chapitre suivant, la méthode est étendue à un modèle plus riche. Le nombre de groupes explicatifs va être augmenté et des covariables seront adjointes. On parlera de modèle structurel multi-blocs à facteurs. Les performances de la méthode d'estimation par algorithme EM seront étudiées dans ce cadre par une analyse de sensibilité.

D'autres généralisations sont possibles. Par exemple, augmenter le nombre de facteurs par sous-modèle de mesure. Dans le cas du modèle à deux groupes présenté dans ce chapitre, il suffit de prendre un groupe de $K_G < q_Y$ facteurs G (resp. $F < q_X$) dépendant (resp. explicatifs) à la place de g (resp. f). Alors le modèle se généralise à :

$$\left\{ \begin{array}{lcl} Y & = & \mathbb{1}_n {\mu_Y}' + GB + \varepsilon^Y \\ X & = & \mathbb{1}_n {\mu_X}' + FA + \varepsilon^X \\ G & = & FC + \varepsilon^G \end{array} \right. \tag{2.28}$$

où G (resp. F) est la matrice de dimension $n \times K_G$ (resp. $n \times K_F$) des K_G facteurs communs en fonction desquels le modèle exprime les variables $Y_1, \ldots, Y_{q_Y}$ (resp. $X_1, \ldots, X_{q_X}$). Les coefficients pondérateurs sont sous forme de matrices :

— A de dimension $K_F \times q_X$;

— B de dimension $K_G \times q_Y$;

— C de dimension $K_F \times K_G$.

Enfin la matrice des erreurs ε^g associées au vecteur g devient la matrice des erreurs ε^G de dimension $n \times K_G$ associée à la matrice G.
Les hypothèses du modèle se généralisent aux suivantes :

— $\varepsilon_i^X \sim \mathcal{N}(0, \psi_X)$ où $\psi_X = diag\left\{\sigma^2_{X,1}, \ldots, \sigma^2_{X,q_X}\right\}$ est une matrice de dimension $q_X \times q_X$;

— $\varepsilon_i^Y \sim \mathcal{N}(0, \psi_Y)$ où $\psi_Y = diag\left\{\sigma^2_{Y,1}, \ldots, \sigma^2_{Y,q_Y}\right\}$ est une matrice de dimension $q_Y \times q_Y$;

— $F_i \sim \mathcal{N}(0, I_{K_F})$;

— $\varepsilon_i^G \sim \mathcal{N}(0, I_{K_G})$;

— $G_i \sim \mathcal{N}(0, C'C + I_{K_G})$;

— ε_i^X, $\varepsilon_{i'}^Y$,$\varepsilon_{i''}^G$ et f_ℓ sont mutuellement indépendants pour tout i, i', i'', ℓ.

En effet, la condition d'identifiabilité devient :

$$\begin{aligned} \mathbb{V}(G) &= C'\mathbb{V}(F)\, C + \mathbb{V}\left(\varepsilon_i^G\right) \\ &= C'C + I_{K_G} \end{aligned}$$

Pour une observation i, le modèle peut se réécrire :

$$\left\{ \begin{array}{lcl} x_i' & = & \mu^{X\prime} + F_i' A + {\varepsilon_i}^{X\prime} \\ y_i' & = & \mu^{Y\prime} + G_i' B + {\varepsilon_i}^{Y\prime} \\ G_i' & = & F_i' C + {\varepsilon_i}^{G\prime} \end{array} \right. \tag{2.29}$$

La log vraisemblance s'obtient par des développements similaires aux précédents :

$$\begin{aligned} \mathcal{L}(\theta; z_i, H_i) = -\frac{1}{2} \sum_{i=1}^{n} & ln|\psi_Y| + ln|\psi_X| \\ & + \left(y_i - \mu^Y - G_i B\right)' \psi_Y^{-1} \left(y_i - \mu^Y - G_i B\right) \\ & + \left(x_i - \mu^X - F_i A\right)' \psi_X^{-1} \left(x_i - \mu^X - F_i A\right) \\ & + \left(g_i - F_i C\right)' \left(g_i - F_i C\right) \\ & + F_i' F_i + \lambda \end{aligned}$$

où λ est une constante, $ln|I_{K_G}| = 0$ et compte tenu des propriétés des distributions gaussiennes on obtient :
$x_i \sim \mathcal{N}\left(\mu^X, A'A + \psi_X\right)$;
$x_i|F_i \sim \mathcal{N}\left(\mu^X + F_i A, \psi_X\right)$;

$y_i \sim \mathcal{N}\left(\mu^Y, B'\left(C'C + I_{K_G}\right)B + \psi_Y\right)$;
$y_i|G_i \sim \mathcal{N}\left(\mu^Y + G_iB, \psi_Y\right)$;
$G_i \sim \mathcal{N}\left(0, C'C + I_{K_G}\right)$;
$G_i|F_i \sim \mathcal{N}\left(F_iC, I_{K_G}\right)$.

Quant à la distribution conditionnelle $h|z$, elle se généralise à :

$$H_i|z_i \sim \mathcal{N}\left(M_i = \begin{pmatrix} M_{1i} \\ M_{2i} \end{pmatrix}, \Gamma^* = \begin{pmatrix} \Gamma^*_{11} & \Gamma^*_{12} \\ \Gamma^*_{21} & \Gamma^*_{22} \end{pmatrix}\right)$$

où $M_i = \Gamma^*_b \Gamma^{*-1}_c \mu^\star_i$ et $\Gamma^* = \Gamma^*_a - \Gamma^*_b \Gamma^{*-1}_c \Gamma^{*\prime}_b$ tel que,
$\Gamma^*_a = \begin{pmatrix} C'C + Id_{k_G} & C' \\ C & Id_{k_F} \end{pmatrix}$, $\Gamma^*_b = \begin{pmatrix} (C'C + Id_{k_G})B & C'A \\ CB & A \end{pmatrix}$,
$\Gamma^*_c = \begin{pmatrix} B'(C'C + Id_{k_G})B + \Psi_Y & B'C'A \\ A'CB & A'A + \Psi_X \end{pmatrix}$ et $\mu^\star_i = \begin{pmatrix} y_i - \mu^Y \\ x_i - \mu^X \end{pmatrix}$.
Pour finir, la reconstruction des facteurs par l'approche est obtenue à une rotation orthogonale près. On a ainsi une infinité de solutions possibles basées sur des transformations orthogonales près des facteurs. Ainsi, l'extension aux facteurs multiples par groupe de variables pose un problème supplémentaire d'identifiabilité, puisque les facteurs d'un groupe ne sont alors identifiables qu'à une transformation orthogonale arbitraire près, ce qui impose l'adjonction d'un nombre de contraintes supplémentaires (Saidane, 2006).

CHAPITRE 3

Algorithme EM et modèles multi-blocs à facteurs

Sommaire

3.1 Introduction et motivations 54
3.2 Estimation par algorithme EM d'un modèle structurel multi-blocs à facteurs 54
3.2.1 Introduction 54
3.2.2 Structure générale du modèle 54
3.2.3 Estimation par algorithme EM 55
3.2.4 L'algorithme 56
3.2.5 Performances de l'approche 57
3.2.6 Application de l'approche EM à des données réelles environnementales 57
3.2.7 L'article soumis 60
3.3 Résultats complémentaires de l'application environnementale : le cas du modèle sans la covariable *géologie* 82
3.3.1 Le modèle sans covariables 82
3.3.2 Tableaux supplémentaires de l'application aux données environnementales 84
3.4 Perspectives et discussion sur les questions du nombre de blocs, de facteurs, de parcimonie et d'unicité des solutions 86

3.1 Introduction et motivations

La méthode d'estimation par algorithme EM présentée au chapitre précédent pour un modèle simple est généralisée dans ce chapitre à un modèle multi-blocs :

— le nombre de groupe explicatif va passer de un à p ;

— des covariables sont adjointes à chaque groupe, explicatif comme dépendant.

Les apports sont alors que :

— plus de relations de causalité entre groupes de VOs via leurs facteurs vont être possibles à modéliser ;

— des effets de covariables vont pouvoir être isolés de ceux des facteurs dans chaque groupe.

Cette généralisation offre au praticien la possibilité de réaliser une étude plus riche lors de l'utilisation de la méthode d'estimation par algorithme EM. Actuellement, dans le cadre de ce modèle, la méthode a été soumise à publication. Après une brève introduction, l'article soumis "EM estimation of a Structural Equation Model" est inséré dans ce chapitre. Il présente le modèle à p groupes explicatifs et un dépendant avec adjonction de covariables, ainsi que l'extension de la méthode à ce dernier. La méthode implémentée sur R y est ensuite illustrée sur des données simulées et appliquée sur données réelles environnementales. Les performances de cette approche y sont évaluées par une analyse de sensibilité. L'article est suivi de résultats complémentaires et pour finir, ce chapitre ouvre une discussion et aborde des perspectives possibles.

3.2 Estimation par algorithme EM d'un modèle structurel multi-blocs à facteurs

3.2.1 Introduction

Dans le contexte des modèles à équations structurelles, on suppose avoir observé plusieurs groupes de variables Y, X^1, ..., X^p sur les mêmes unités. Chacun de ces groupes est supposé structuré autour d'un facteur latent. Ces facteurs sont alors liés par un modèle linéaire. Chaque groupe peut aussi être enrichi de variables explicatives additionnelles T, T^1, ..., T^p, également observées et nommées "covariables". Actuellement, lorsque le choix de la nature de facteur des VLs est fait, seule LISREL (Jöreskog, 2000) permet leur estimation en plus de celle des paramètres. Cependant, comme nous l'avons dit au chapitre précédent, cette technique n'utilise pas toutes les équations du modèle pour obtenir les estimations des facteurs. *A contrario*, dans le paradigme de l'estimation par maximum de vraisemblance, l'estimation par algorithme EM ne néglige aucune équation du modèle lors de l'estimation simultanée des facteurs et des paramètres. De plus, grâce à EM, cette méthode est aussi efficace en terme de temps de calcul. Nous avons en effet constaté qu'un faible nombre d'itérations suffisait lors de l'application de la méthode à la fois sur des données simulées et sur un jeu de données réelles.

3.2.2 Structure générale du modèle

Les notations du modèle sont similaires à celles du modèle (2.19) mais étendues à un nombre p de groupes explicatifs. Ainsi chacun des groupes de variables Y, X^1,..., X^p dépend d'une variable latente (respectivement g, f^1, ..., f^p). Au niveau du modèle interne, le facteur latent g dépend de f^1, ..., f^p (cf. figure 1 de l'article soumis et inséré en section 3.2.7). Quant aux modèles externes, chacun formalise un groupe de variables $X^m = (X_1^m, ...X_{q_m}^m)$ (resp. $Y = (Y_1, ...Y_{q_Y})$) comme dépendant du facteur f^m (resp. g). Chacune de ces dépendances est enrichie par une dépendance supplémentaire aux covariables T^m (resp. T). Soient, D (resp. D^m) une matrice $r_T \times q_Y$ (resp. $r_m \times q_m$) de coefficients pondérateurs des covariables, b (resp.

a^m) un vecteur $1 \times q_Y$ (resp. $1 \times q_m$) de coefficients pondérateurs des facteurs et ε^Y (resp. ε^m) une matrice d'erreurs $n \times q_Y$ (resp. $n \times q_m$), associées au groupe de variable Y (resp. X^m). On notera également ε^g la matrice des erreurs associées à g. Le modèle peut alors être formulé ainsi :

$$\left\{ \begin{array}{ccc} Y & = & TD + gb' + \varepsilon^Y \\ \forall m \in [\![1,p]\!], X^m & = & T^m D^m + f^m a^{m\prime} + \varepsilon^m \\ g & = & f^1 c^1 + \cdots + f^p c^p + \varepsilon^g \end{array} \right. \tag{3.1}$$

où les éléments de la première colonne des matrices de covariables T et T^m sont fixés à 1 pour pouvoir capturer les ordonnées à l'origine (i.e. : moyennes des variables du groupe). Ainsi, la première ligne de D et de chaque matrice D^m correspondent aux espérances des variables de Y et X^m respectivement.. Sous contraintes d'identifiabilité, pour toute observation $i \in [\![1,n]\!]$ et $\forall m \in [\![1,p]\!]$ les hypothèses du modèle (3.1) sont les suivantes :

- $f^m \sim \mathcal{N}(0,1)$;
- $\varepsilon_i^m \sim \mathcal{N}(0,\psi_m)$ où $\psi_m = diag(\sigma^2_{m,j})_{j\in[\![1,q_m]\!]}$ de dimension $q_m \times q_m$;
- $\varepsilon_i^Y \sim \mathcal{N}(0,\psi_Y)$ où $\psi_Y = diag(\sigma^2_{Y,j})_{j\in[\![1,q_Y]\!]}$ de dimension $q_Y \times q_Y$;
- $\varepsilon_i^g \sim \mathcal{N}(0,1)$;
- $g \sim \mathcal{N}(0,(c^1)^2 + \ldots + (c^p)^2 + 1)$;
- ε_i^g est indépendant des f^m pour toute observation i ;
- ε^Y et ε^m sont indépendants.

Remarque. L'adjonction de covariables (ou variables explicatives) notées T (resp. T^m) dans le modèle (3.1) permet d'ajouter d'autres déterminants des blocs de variables observées Y (resp. X^m). Cela laisse la possibilité aux blocs Y et X^m de ne pas être déterminés uniquement par les facteurs g et f^m mais de recevoir d'autres influences que celle des facteurs. Ainsi, chaque modèle de mesure est plus riche qu'un modèle de mesure réflectif classique. Dans le cadre de données où les observations sont organisées en différents groupes, un exemple d'influence pouvant être exercée par une covariable pourrait être celle d'"effets groupes". Par exemple, pour un échantillon d'une population d'individus, un exemple de variable explicative T pourrait être leurs différents niveaux d'étude, ce qui apporterait de l'information vis-à-vis de l'influence de ce determinant sociologique.

3.2.3 Estimation par algorithme EM

Afin de simplifier les développements, l'estimation par algorithme EM est présentée avec une restriction à $p = 2$ groupes explicatifs dans l'article. Alors, pour i une observation, le modèle peut être formulé selon le système d'équations suivant :

$$\left\{ \begin{array}{ccc} y_i' & = & t_i{}'D + g_i b' + \varepsilon_i{}^{y\prime} \\ x_i^{1\prime} & = & t_i^{1\prime} D^1 + f_i^1 a^{1\prime} + \varepsilon_i{}^{1\prime} \\ x_i^{2\prime} & = & t_i^{2\prime} D^2 + f_i^2 a^{2\prime} + \varepsilon_i{}^{2\prime} \\ g_i & = & f_i^1 c^1 + f_i^2 c^2 + \varepsilon_i{}^g \end{array} \right. \tag{3.2}$$

où nous faisons les simplifications suivantes $\psi_Y = \sigma_Y^2 Id_{q_Y}$, $\psi_1 = \sigma_1^2 Id_{q_1}$,$\psi_2 = \sigma_2^2 Id_{q_2}$. On note $\theta = \{D, D^1, D^2, b, a^1, a^2, c^1, c^2, \sigma_Y^2, \sigma_1^2, \sigma_2^2\}$ l'ensemble des paramètres du modèle. Ainsi, la dimension de θ est :

$$K^\star = 5 + q_Y(r_T + 1) + \sum_{m=1}^{2} q_m(r_m + 1)$$

Les hypothèses de ce modèle sont celles du modèle à p groupes explicatifs et un groupe dépendant.
Pour $z = (y, x^1, x^2)$ et $h = (g, f^1, f^2)$ la log vraisemblance s'écrit :

$$\begin{aligned}\mathcal{L}(\theta; z, h) = -\frac{1}{2}\sum_{i=1}^{n} & ln|\psi_Y| + ln|\psi_1| + ln|\psi_2| \\ & + (y_i - D't_i - g_i b)'\psi_Y^{-1}(y_i - D't_i - g_i b) \\ & + (x_i^1 - D^{1\prime}t_i^1 - f_i^1 a^1)'\psi_1^{-1}(x_i^1 - D^{1\prime}t_i^1 - f_i^1 a^1) \\ & + (x_i^2 - D^{2\prime}t_i^2 - f_i^2 a^2)'\psi_2^{-1}(x_i^2 - D^{2\prime}t_i^2 - f_i^2 a^2) \\ & + (g_i - c^1\ f_i^1 - c^2\ f_i^2)^2 + (f_i^1)^2 + (f_i^2)^2 + \lambda\end{aligned}$$

où λ une constante et dans notre cas (3.2), $\theta = \{D, D^1, D^2, b, a^1, a^2, c^1, c^2, \sigma_Y^2, \sigma_1^2, \sigma_2^2\}$.
Pour maximiser la fonction log-vraisemblance par algorithme EM le système (2.14) doit être résolu. Pour ce faire, la distribution conditionnelle $h|z$ utilisée pour chaque observation i notée :

$$h_i|z_i \sim \mathcal{N}\left(m_i = \begin{pmatrix} m_{1i} \\ m_{2i} \\ m_{3i}\end{pmatrix}, \Sigma_i = \begin{pmatrix} \sigma_{11_i} & \sigma_{12_i} & \sigma_{13_i} \\ \sigma_{21_i} & \sigma_{22_i} & \sigma_{23_i} \\ \sigma_{31_i} & \sigma_{32_i} & \sigma_{33_i}\end{pmatrix}\right).$$

Son expression est explicitée dans l'article et son obtention dans les annexes de ce dernier.
Les notations suivantes sont ensuite introduites :

$$\begin{array}{ll}
\widetilde{\gamma_i} = \mathbb{E}_{z_i}^{h_i}\left[g_i^2\right] = \left(\mathbb{E}_{z_i}^{h_i}\left[g_i\right]\right)^2 + \mathbb{V}_{z_i}^{h_i}\left[g_i\right] = {m_1}_i^2 + \sigma_{11_i}\,; & \widetilde{g_i} = \mathbb{E}_{z_i}^{h_i}\left[g_i\right] = m_{1i}\,; \\
\widetilde{\phi_i^1} = \mathbb{E}_{z_i}^{h_i}\left[\left(f_i^1\right)^2\right] = \left(\mathbb{E}_{z_i}^{h_i}\left[f_i^1\right]\right)^2 + \mathbb{V}_{z_i}^{h_i}\left[f_i^1\right] = {m_2}_i^2 + \sigma_{22_i}\,; & \widetilde{f_i^1} = \mathbb{E}_{z_i}^{h_i}\left[f_i^1\right] = m_{2i}\,; \\
\widetilde{\phi_i^2} = \mathbb{E}_{z_i}^{h_i}\left[\left(f_i^2\right)^2\right] = \left(\mathbb{E}_{z_i}^{h_i}\left[f_i^2\right]\right)^2 + \mathbb{V}_{z_i}^{h_i}\left[f_i^2\right] = {m_3}_i^2 + \sigma_{33_i}\,; & \widetilde{f_i^2} = \mathbb{E}_{z_i}^{h_i}\left[f_i^2\right] = m_{3i}.
\end{array}$$

Puis, à partir des notations précédentes, les formules solutions de (2.14) sont :

$$\begin{aligned}
\hat{b} &= \frac{(\overline{\widetilde{g}y} - \overline{yt'})(\overline{tt'})^{-1}\ \overline{\widetilde{g}t}}{\overline{\widetilde{\gamma}} - \overline{\widetilde{g}t'}(\overline{tt'})^{-1}\overline{\widetilde{g}t}} \\
\widehat{a^m} &= \frac{\overline{\widetilde{f^m}x^m} - \overline{x^m t^{m\prime}}(\overline{t^m t^{m\prime}})^{-1}\overline{\widetilde{f^m}t^m}}{\overline{\widetilde{\phi^m}} - \overline{\widetilde{f^m}t^{m\prime}}(\overline{t^m t^{m\prime}})^{-1}\overline{\widetilde{f^m}t^m}} \\
\widehat{c^1} &= \frac{(\sigma_{12} + \overline{\widetilde{f^1}\widetilde{g}})\overline{\widetilde{\phi^2}} - (\sigma_{13} + \overline{\widetilde{f^2}\widetilde{g}})(\sigma_{23} + \overline{\widetilde{f^1}\widetilde{f^2}})}{\overline{\widetilde{\phi^1}\widetilde{\phi^2}} - (\sigma_{23} + \overline{\widetilde{f^1}\widetilde{f^2}})^2} \\
\widehat{c^2} &= \frac{(\sigma_{13} + \overline{\widetilde{f^2}\widetilde{g}})\overline{\widetilde{\phi^1}} - (\sigma_{12} + \overline{\widetilde{f^1}\widetilde{g}})(\sigma_{23} + \overline{\widetilde{f^1}\widetilde{f^2}})}{\overline{\widetilde{\phi^1}\widetilde{\phi^2}} - (\sigma_{23} + \overline{\widetilde{f^1}\widetilde{f^2}})^2} \\
\widehat{D'} &= (\overline{yt'} - \widehat{b}\ \overline{\widetilde{g}t'})(\overline{tt'})^{-1} \\
\widehat{D^{m\prime}} &= (\overline{x^m t^{m\prime}} - \widehat{a^m}\overline{\widetilde{f^m}t^{m\prime}})(\overline{t^m t^{m\prime}})^{-1} \\
\widehat{\sigma_Y^2} &= \frac{1}{nq_Y}\sum_{i=1}^{n}\{||y_i - \widehat{D'}t_i||^2 + ||\hat{b}||^2\widetilde{\gamma_i} - 2(y_i - \widehat{D'}t_i)'\hat{b}\widetilde{g_i}\} \\
\widehat{\sigma_m^2} &= \frac{1}{nq_m}\sum_{i=1}^{n}\{||x_i^m - \widehat{D^{m\prime}}t_i^m||^2 + ||\widehat{a^m}||^2\widetilde{\phi_i^m} - 2(x_i^m - \widehat{D^{m\prime}}t_i^m)'\widehat{a^m}\widetilde{f_i^m}\}
\end{aligned} \tag{3.3}$$

3.2.4 L'algorithme

Pour estimer les paramètres de θ et les facteurs g, f^m, la procédure itérative qui suit est employée où $[t]$ correspond à la t-ème itération.

1. Initialisation = choix des valeurs initiales des paramètres $\theta^{[0]}$.

2. À l'itération courante $t \geq 1$, jusqu'à satisfaction du critère d'arrêt on procède comme suit :
   (a) **E-step** : Avec $\theta^{[t-1]}$,
      i. On calcule explicitement la distribution $h_i|z_i$ pour tout $i \in [\![1, n]\!]$.
      ii. On estime les valeurs des facteurs $\widetilde{g}^{[t]}$, $\widetilde{f^m}^{[t]}$, $m \in \{1, 2\}$.
      iii. On calcule $\widetilde{\gamma}^{[t]}$ and $\widetilde{\phi^m}^{[t]}$, $m \in \{1, 2\}$.
   (b) **M-step** :
      i. On actualise θ à $\theta^{[t]}$ en introduisant $\widetilde{g}^{[t]}$, $\widetilde{\gamma}^{[t]}$ et $\widetilde{f^m}^{[t]}$, $\widetilde{\phi^m}^{[t]}$, $m \in \{1, 2\}$ dans les formules solutions (3.3).
3. Le critère d'arrêt suivant avec ϵ le plus petit possible est utilisé :

$$\sum_{k^\star=1}^{K^\star} \frac{|\theta^{\star[t+1]}[k^\star] - \theta^{\star[t]}[k^\star]|}{\theta^{\star[t+1]}[k^\star]} < \epsilon$$

où $\theta^\star$ est le vecteur de dimension $K^\star$ contenant tous les paramètres scalaires de l'ensemble des paramètres de θ.
L'algorithme de cette procédure d'estimation est illustré par le schéma 3.1.

3.2.5 Performances de l'approche

À l'aide d'une analyse de sensibilité [1], nous montrons les performances de la méthode. Sur R, la méthode implémentée a été appliquée à plusieurs groupes de 100 jeux de données simulés. Chaque groupe de 100 jeux de données a été simulé pour $r_T = r_1 = r_2 = 2$, c'est à dire que deux covariables sont simulées pour chaque matrice de covariables T, T^1, T^2. Pour évaluer la qualité des estimations obtenues en fonction des paramètres de dimension des données, chaque groupe de 100 jeux de données a été simulé pour différentes valeurs de :

— $n \in \{50, 100, 200, 400\}$ et un nombre de $q = 40$ VOs pour chaque groupe explicatif ou dépendant ;

— $q \in \{5, 10, 20, 40\}$ et un même nombre de $n = 400$ unités.

Pour étudier la qualité des estimations obtenues pour $\varepsilon = 10^{-2}$, sur chacun des groupes de 100 jeux de données, des statistiques ont été calculées. Pour les facteurs (resp. les paramètres) les moyennes des corrélations entre les facteurs (resp. paramètres) simulés et estimés ont été calculées et illustrées par des box plots. Les variances des corrélations sont quant à elles illustrées par des intervalles de confiances à 95% pour les paramètres structurels c^1 et c^2 et un paramètre de variance. Ce choix est justifié par le fait que les paramètres les plus sensibles sont les paramètres structurels et les paramètres de variance. L'intervalle de confiance d'un seul des paramètres de variance est présenté car pour les autres, on a obtenu des résultats similaires. Les conclusions sont qu'un nombre minimal de $n = 100$ observations est nécessaire et qu'il est préférable d'avoir un nombre minimal de $q = 10$ VO par groupe.

3.2.6 Application de l'approche EM à des données réelles environnementales

La méthode est aussi illustrée sur un jeu de données réelles nommé *genus*. Il provient du package R SCGLR (Mortier et al., 2014) qui répertorie les abondances de 27 espèces d'arbres d'une forêt tropicale du Congo ainsi que 40 variables géologiques pour $n = 1000$ parcelles de 8 km^2. Les variables géologiques comprennent :

1. dans un sens différent de la "sensitivity analysis" introduite par Saltelli et al. (2000).

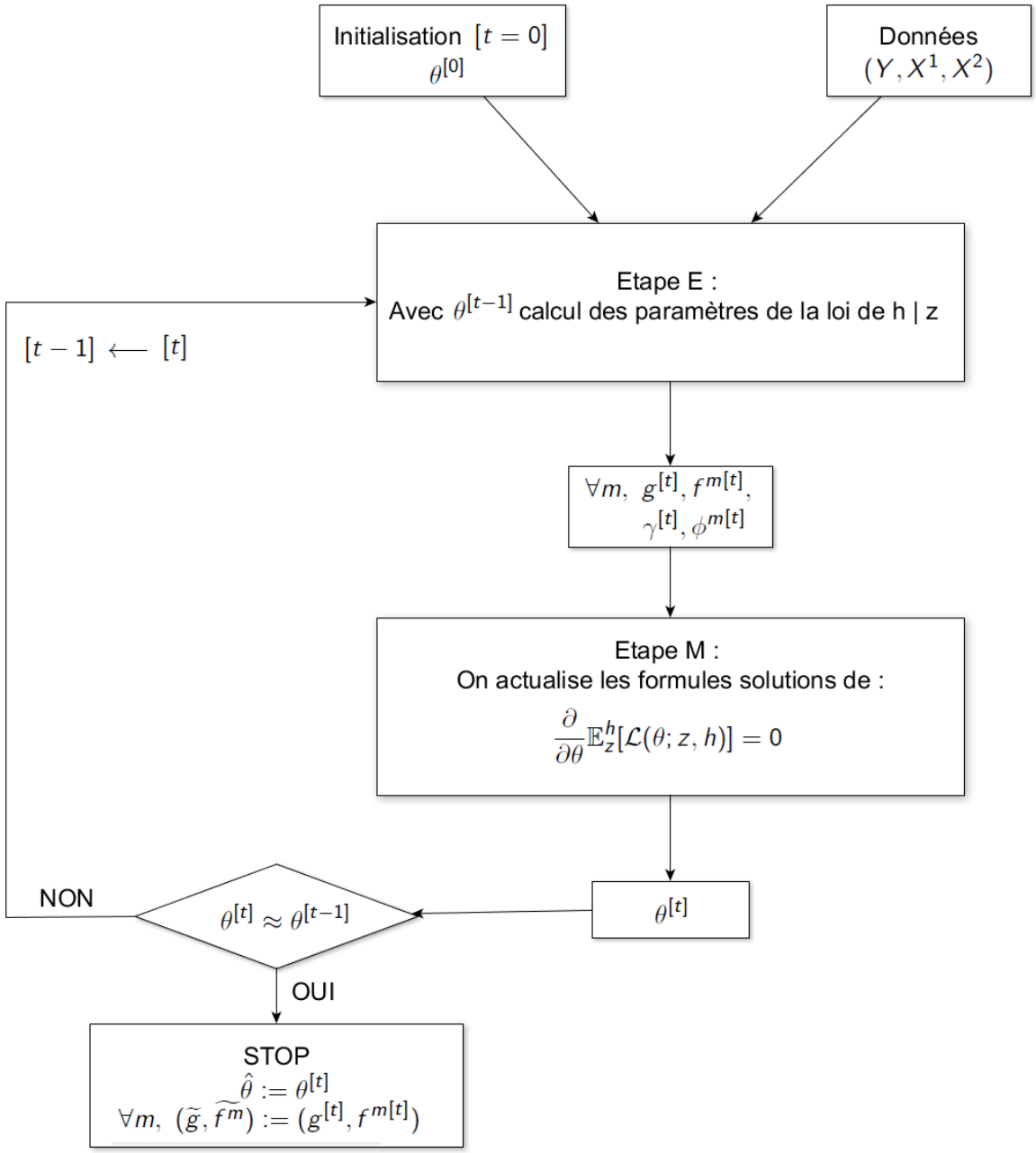


FIGURE 3.1 – Procédure itérative de la méthode d'estimation par algorithme EM.

— 16 variables pluviométriques et de localisation (longitude, latitude et altitude), regroupées sous la matrice X^1 ;

— 23 variables mesurant l'activité de photosynthèse, regroupées sous la matrice X^2 ;

— une variable nominale *géologie* associant à chacune des parcelles la nature de son sol. Elle comporte 5 modalités.

Le but de cette application est de modéliser les abondances d'arbres Y à partir des autres groupes de VO X^1 et X^2 en procédant à de la réduction de dimension (cf. le diagramme structurel 3.2). En effet, une ACP des variables X^1 et X^2 confirme que ces deux groupes sont clairement séparés. Ainsi l'hypothèse que les blocs explicatifs sont indépendants est faite. Le groupe de VO dépendantes est Y de dimension $n \times q_Y$ où $q_Y = 27$ espèces d'arbres. Pour chacune des observations (i.e : des parcelles), le nombre d'arbres d'une même espèce répertoriés sur la parcelle est divisé par la surface de celle-ci.

Remarque. Ce pré-traitement de Y a pour objectif d'enlever l'effet de taille des parcelles dans les variables dépendantes, puisque les variables explicatives sont des variables par unité de surface. Le modèle ne concerne que des variables invariantes par changement de taille.

Remarque. Une conséquence de ce pré-traitement est la conversion des variables de comptage de Y en des variables "pseudo-continues". On se rapproche des hypothèses de la méthode d'estimation par EM. Un modèle structurel à deux groupes explicatifs (X^1 et X^2) et un groupe dépendant (Y) est établi. En revanche, au vu du rôle que peut jouer la variable *géologie* sur l'abondance des arbres Y, il semble intéressant de compléter le modèle par une matrice de covariables T de dimension $n \times q_T$ où $q_T = 5$ pour la moyenne et les 5 modalités moins une qui est choisie comme référence. Pour tester l'effet que la nature du sol pourrait avoir sur l'abondance des espèces d'arbres (Y), il est intéressant de faire deux modélisations, l'une avec les covariables T et l'autre sans T.

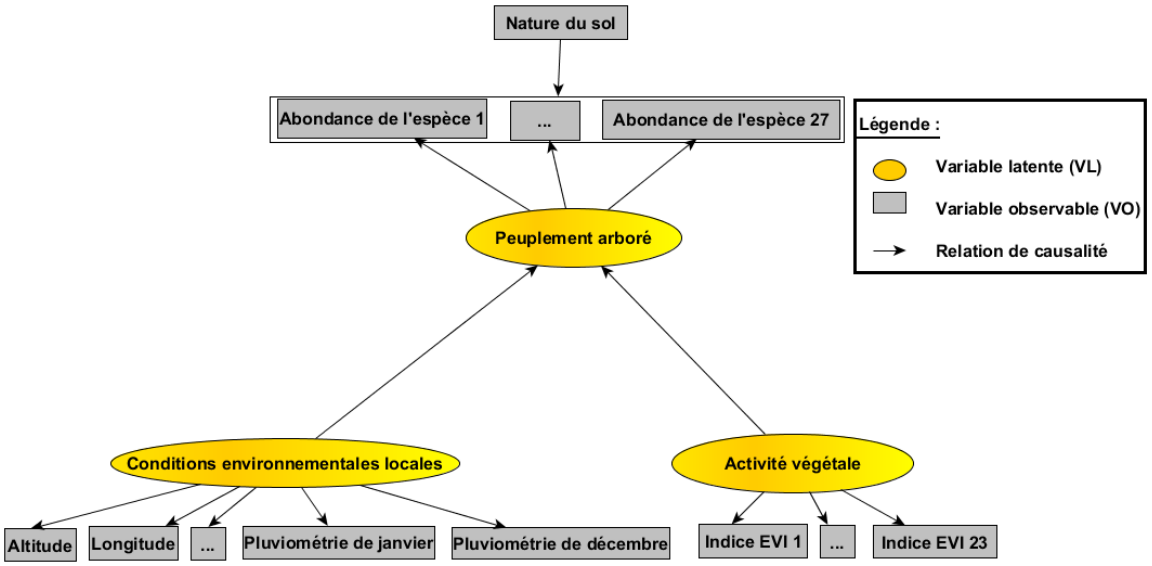


FIGURE 3.2 – Diagramme structurel du peuplement arboré expliqué par deux blocs de variables observées liées aux conditions environnementales locales et à l'activité végétale.

Remarque. Dans le modèle les T ne sont pas liées aux VO X^1 et X^2 car la variable *géologie* ne semble pas être conceptuellement liée à la pluviométrie ou à l'activité de photosynthèse. Le choix de présenter seulement le modèle avec T a été fait dans l'article faute de place. En revanche, des résultats liés aux deux modèles sont présentés et complétés dans la section "Résultats complémentaires" de ce chapitre. Pour le modèle avec T, pour $\varepsilon = 10^{-3}$, la convergence a eu lieu en 58 itérations. Il résulte des différentes estimations que pour certaines espèces d'arbres, un effet "nature du sol" semble exister et se distinguer de celui du facteur

g. Pour étudier la qualité du modèle avec T, la méthode d'estimation par algorithme EM lui a été appliquée pour différents sous échantillons du jeu de données *genus*. Ces différents échantillons ont été choisis de même taille. Puis, pour chacun d'entre eux, les estimations des paramètres et des facteurs ont été calculés. Les corrélations entre les différentes estimations des paramètres d'une part et des facteurs d'autre part ont été étudiées. Les valeurs obtenues étaient proches de 1 en valeur absolue. Les paramètres, comme les facteurs estimés sont donc fortement corrélées pour les différents échantillon, ce qui est en faveur de la modélisation avec T.

3.2.7 L'article soumis

EM estimation of a Structural Equation Model

Xavier Bry[1,2], **Christian Lavergne**[1,3] **and Myriam Tami**[1,2]

[1] *Institut Montpelliérain Alexander Grothendieck (IMAG), CNRS, Montpellier, France*
[2] *University of Montpellier, Montpellier, France*
[3] *University Paul-Valéry Montpellier 3, Montpellier, France*

Address for correspondence

Myriam Tami, Department of Mathematics, Faculty of Sciences, University of Montpellier, 4, Place Eugène Bataillon, CC437, 34095 Montpellier cedex 5, France
E-mail : myriam.tami@umontpellier.fr
Phone : (+330) 467 143 951
Fax : (+330) 467 143 558

Abstract

In this work, we propose a new estimation method of a Structural Equation Model (SEM). Contrasting with the classical SEM approach, our method is not based on the constrained estimation of the covariance structure of the data. It consists in viewing the Latent Variables (LV's) as missing data and using the EM algorithm to maximize the whole model's likelihood, which simultaneously provides estimators not only of the model's coefficients, but also of the values of LV's. Through a simulation study, we investigate how fast and accurate the method is, and eventually apply it to real environmental data.

Key words

EM algorithm ; Factor model ; Latent Variable ; Structural Equation Model.

1 Introduction

When it comes to modeling phenomena involving indirect measurements, SEMs are handy and, as such, widely used. These allow to formalize statistical links between LV's on the one hand, and between LV's and observed ones, on the other hand. Such models originate in psychometry, and have now been under attention for a century. Several ways were proposed to handle their estimation. Historically, the first proposed approach was that of factor analysis. This approach views the LV's as unknown unconstrained factors having a known distribution (typically standard normal). This factoring approach is then chiefly concerned by the analysis of the covariance structure of the Observed Variables (OV's), this structure having a particular design constrained by the LV's underlying the observed ones. Under a multivariate normal assumption for all random variables in the model, the variance-covariance matrix of the OV's follows a Wishart distribution. The proposed estimation technique is a constrained maximum likelihood estimation of this variance-covariance matrix. This factoring approach was extended by Jöreskog to more general SEM involving LInear Structural RELations (LISREL) between LV's (Jöreskog, 1970; Jöreskog and Sörbom, 1982), and later culminated in works by Muthén, who addressed several issues, such as General Linear Modeling, and also dealing with missing data in OV's by means of the EM algorithm (Muthén et al., 1987; Muthén and Muthén, 1998). This approach is theoretically well grounded, but

has at least three drawbacks. The first one is that the direct maximization of the likelihood of the variance-covariance matrix is technically tricky, and all the more so as the structural equation system is complex. The second one is that, concentrating on the variance-covariance matrix, this approach does not directly provide estimates of the factor-values at unit level (called individual scores). The third one is that factors are not constrained enough to allow predicting the values of the OV's for statistical units that do not belong to the estimation sample. To overcome such drawbacks, an alternative approach to structural modeling has been proposed by Wold : the Partial Least squares Path-Modeling (PLSPM) approach. This approach views LV's as components, i.e. unknown combinations of the OV's they are related to. No assumption is made as to their distribution, and only least-square technology is involved in the estimation. Computation is faster and convergence is said to be faster and more robust. Moreover, components can be predicted from OV's and so, scores are a direct output of this approach. Another consequence is the ability to predict the values of dependent variables for new statistical units. In spite of such commodities, the PLSPM approach does not deal appropriately with the partial relationships between components involved in a multiple regression equation, as shown in Bry and Verron (2015), who lately proposed an extended method to extract components suiting a multiple equation model : THEME. Besides, a LISREL-related method, Unweighted Least Squares (ULS) has been proposed by McDonald (1996), that adds to Jöreskog's assumptions the constraint that LV's are linear combinations of the observed ones. Now, constraining LV's to be components may be considered unnecessarily limiting when one is not interested in prediction, and the factoring approach still makes a strong point in this respect. Later, Jöreskog (2000) completed the LISREL approach with a method to estimate LV's scores. However, this method is based on a least squares technique performed on the mere measurement equations, overlooking the structural equation. All one needs is then to overcome LISREL's drawbacks, i.e. to make estimation easier and more direct for both parameters and scores. The present works proposes to tackle this issue by viewing the factors' values as missing data and using the EM algorithm (initially designed by Dempster et al. (1977) to deal with missing data) to maximize the likelihood of the whole data. Extending the use of EM in a related way was proposed by Dempster et al. (1981) and Andrade and Helms (1984) in the framework of mixed linear models. The great advantage of EM over the classical Newton-Raphson, Fisher Scoring and Fletcher and Powell algorithms is that EM automatically allows to keep parameters in their space and does not require computing the hessian matrix on each step. Although LISREL only considers Full-Information Maximum Likelihood (FIML, (Arbuckle et al., 1996)), introducing the use of EM in it has been considered by Lee and Tang (2006); Muthén et al. (1987); Tang and Lee (1998) in order to deal with censoring when it occurs according to a known process. Our approach is different from all previous ones, in that we neither consider the missing data to be partial or censored. Nor do we consider the likelihood of the mere constrained sample variance-covariance matrix, but that of the whole data and the complete model.

In order to keep the developments simple in the paper, we restricted our SEM to only one structural equation, which does not lessen the generality of our method. The paper is organized as follows. Section 2 formally introduces the equations of the SEM we deal with. Section 3 applies the EM algorithm to the SEM and derives the estimation formulas. Section 4 first presents a simulation-based study of the performance of the method, with comparison to more classical methods, and then an application to environmental data.

2 The model

2.1 Notations

Data notations

The data consists in blocks of OV's describing the same n units. We consider the following data-matrices and notations :
$\mathrm{Y} = \{y_i^j\}$; $i \in [\![1,n]\!]$, $j \in [\![1,q_\mathrm{Y}]\!]$ is the $n \times q_\mathrm{Y}$ matrix coding the dependent block of OV's $y^1,...,y^{q_\mathrm{Y}}$, identified with its column-vectors.
$\mathrm{X}^m = \{x_i^{j,m}\}$; $i \in [\![1,n]\!]$, $j \in [\![1,q_m]\!]$, $m \in [\![1,p]\!]$ is the $n \times q_m$ matrix coding the m^{ieth}-explanatory block of OV's $x^{1,m},...,x^{q_m,m}$. Value of variable $x^{j,m}$ for unit i is denoted $x_i^{j,m}$. Variable-blocks will be referred to through the corresponding matrices.
T (resp. $\mathrm{T}^1,...,\mathrm{T}^p$) refers to a $n \times r_\mathrm{T}$ (resp. $n \times r_1$, ..., $n \times r_p$) matrix of covariates.
We assume that the units :

— Hence the rows of matrices $\mathrm{Y},\mathrm{X}^1,...,\mathrm{X}^p$ are independent multivariate normal vectors.

Model notations

For the sake of simplicity, the SEM we handle here is a restricted one, in that it contains only one structural equation, relating a dependent latent factor g, underlying block Y, to p explanatory latent factors $f^1,...,f^p$ respectively underlying blocks $\mathrm{X}^1,...,\mathrm{X}^p$ (cf. figure 1). The main assumptions of this model are the following :

— Factors $f^1,...,f^p$ are standard normal, i.e. $\forall m \in [\![1,p]\!], \mathbb{E}(f^m) = 0$ and $\mathbb{V}(f^m) = \mathrm{I}_n$.

— In each block (e.g. X^p), the OV's (e.g. $x_j^m, j \in [\![1,q_p]\!]$) depend linearly on the block's factor (e.g. f^m) and a block of extra-covariates (e.g. T^m), conditional on which they are independent. Including extra-covariates allows to remove their effect on the OV's from that of the factor, and thus, look for a better focused factor, which is very important in many practical situations.

— Factor g is normal with zero-mean, and its expectation conditional on $f^1,...,f^p$ is a linear combination of them.

The SEM consists of $p+1$ measurement equations and one structural equation. It is graphed on figure 1.

2.2 Measurement equations

As formerly mentioned, each measurement equation relates the variables in a block X^m (respectively Y) to the block's factor f^m (resp. g). This link may also involve covariates T^m (resp. T) : each OV is expressed as a linear combination of the factor, the covariates and some noise. Hence the $p+1$ measurement equations :

$$\begin{cases} \mathrm{Y} &= \mathrm{TD} + gb' + \varepsilon^\mathrm{Y} \\ \forall m \in [\![1,p]\!], \mathrm{X}^m &= \mathrm{T}^m\mathrm{D}^m + f^m a^{m\prime} + \varepsilon^m \end{cases}$$

where D (resp. D^m) is a $r_\mathrm{T} \times q_\mathrm{Y}$ (resp. $r_m \times q_m$) parameter matrix, b (resp. a^m) a $1 \times q_\mathrm{Y}$ (resp. $1 \times q_m$) parameter vector, and ε^Y (resp. ε^m) an $n \times q_\mathrm{Y}$ (resp. $n \times q_m$) measurement-error matrix.
We impose that the first column of T (resp. T^m) matrix is equal to constant vector having all elements equal to 1. Thus, the first row of D (resp. D^m) contains mean-parameters.
As far as distributions are concerned, we assume that :

— $\varepsilon_i^\mathrm{Y} \sim \mathcal{N}(0, \psi_\mathrm{Y})$, where $\psi_\mathrm{Y} = diag(\sigma^2_{\mathrm{Y},j})_{j \in [\![1,q_\mathrm{Y}]\!]}$;

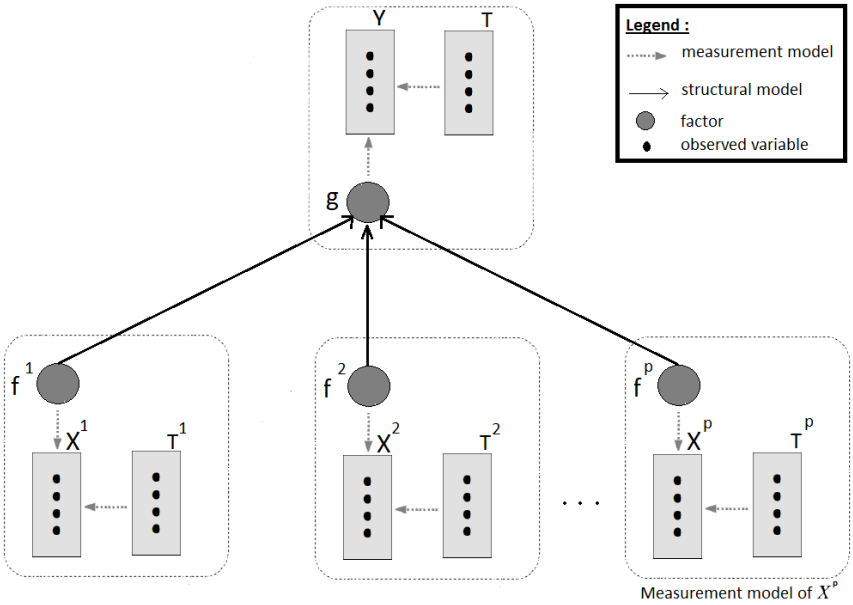


FIGURE 1 – The Structural Equation Model

— $\forall m \in [\![1,p]\!] : \varepsilon_i^m \sim \mathcal{N}(0, \psi_m)$, where $\psi_m = diag(\sigma^2_{m,j})_{j\in[\![1,q_m]\!]}$;

— ε^Y and ε^m, $\forall m \in [\![1,p]\!]$ are independent.

As to the factors, we assume that :

— $\forall m \in [\![1,p]\!] : f^m \sim \mathcal{N}(0, Id_n)$ with $f^1, ..., f^m$ independent.

2.3 Structural equations

The structural equation we consider relates dependent factor g to explanatory factors $f^1, ..., f^p$ (cf. figure 1) through a linear model :

$$\{ \; g \;\; = \;\; f^1 c^1 + \cdots + f^p c^p + \varepsilon^g$$

where $\forall m \in [\![1,p]\!]$, c^m is a scalar parameter, and $\varepsilon^g \in \mathbb{R}^n$ is a disturbance vector.
We assume that :

— $\varepsilon^g \sim \mathcal{N}(0,1)$;

— ε^g is independent of ε^Y and ε^m, $\forall m \in [\![1,p]\!]$.

N.B. The unit-variance of disturbance ε^g serves an identification purpose.
Hence the overall model :

$$\left\{ \begin{array}{ccc} Y & = & TD + gb' + \varepsilon^Y \\ \forall m \in [\![1,p]\!], X^m & = & T^m D^m + f^m a^{m\prime} + \varepsilon^m \\ g & = & f^1 c^1 + \cdots + f^p c^p + \varepsilon^g \end{array} \right. \tag{1}$$

where the K-dimensional vector of parameters is
$\theta = \{D, D^1, ..., D^p, b, a^1, ..., a^p, c^1, c^2, \psi_Y, \psi_1, ..., \psi_p\}$.
Thus, when all ψ matrices are diagonal, we have :

$$K = 2 + q_Y(r_T + 2) + \sum_{m=1}^{p} q_m(r_m + 2) \tag{2}$$

2.4 A simplified model

In order to avoid heavy formulas in the development of the algorithm, we shall use in the sequel, with no loss of generality, a simplified model involving $p = 2$ explanatory blocks X^1 and X^2. The corresponding equation set, for a given unit i, reads :

$$\left\{\begin{array}{rcl} y_i' & = & t_i{}'D + g_i b' + \varepsilon_i{}^{y'} \\ x_i^{1\prime} & = & t_i^{1\prime} D^1 + f_i^1 a^{1\prime} + \varepsilon_i{}^{1\prime} \\ x_i^{2\prime} & = & t_i^{2\prime} D^2 + f_i^2 a^{2\prime} + \varepsilon_i{}^{2\prime} \\ g_i & = & f_i^1 c^1 + f_i^2 c^2 + \varepsilon_i{}^{g} \end{array}\right. \tag{3}$$

Then, $\theta = \{D, D^1, D^2, b, a^1, a^2, c^1, c^2, \sigma_Y^2, \sigma_1^2, \sigma_2^2\}$, and (cf. (2)), the dimension of θ is :

$$K = 5 + q_Y(r_T + 1) + \sum_{m=1}^{2} q_m(r_m + 1)$$

3 Estimation using the EM algorithm

We propose to carry out likelihood maximization through an iterative EM algorithm (Dempster et al. (1977), section 4.7). Each iteration of the algorithm involves an Expectation (E)-step followed by a Maximization (M)-step. Dempster et al. (1977) prove that the EM algorithm yields maximum likelihood estimates. Moreover, they proved that even if the starting point is one where the likelihood is not convex, if an instance of the algorithm converges, it will converge to a (local) maximum of the likelihood. Another major advantage of the EM algorithm is that it can be used to "estimate" missing values through their expectation conditional on the observed data. Thus, if we consider LV's as missing data, the EM algorithm is an adequate tool to maximize the likelihood of a statistical model involving LV's, but also to estimate these LV's. In our SEM framework, the LV's are the factors. Thus, EM enables us to estimate the factors at unit-level. We shall present the algorithm on the simplified model (3) with no loss of generality.

3.1 The EM algorithm

Let $z = (y, x^1, x^2)$ be the OV's and $h = (g, f^1, f^2)$ the LV's. The EM algorithm is based on the log-likelihood associated with the complete data (z, h).

The complete log-likelihood function

Let $p(z, h; \theta)$ denote the probability density of the complete data. The corresponding log-likelihood function is :

$$\begin{aligned} \mathscr{L}(\theta; z, h) = -\frac{1}{2}\sum_{i=1}^{n} \{ & ln|\psi_Y| + ln|\psi_1| + ln|\psi_2| \\ & + (y_i - D' t_i - g_i b)' \psi_Y^{-1} (y_i - D' t_i - g_i b) \\ & + (x_i^1 - D^{1\prime} t_i^1 - f_i^1 a^1)' \psi_1^{-1} (x_i^1 - D^{1\prime} t_i^1 - f_i^1 a^1) \\ & + (x_i^2 - D^{2\prime} t_i^2 - f_i^2 a^2)' \psi_2^{-1} (x_i^2 - D^{2\prime} t_i^2 - f_i^2 a^2) \\ & + (g_i - c^1 f_i^1 - c^2 f_i^2)^2 + (f_i^1)^2 + (f_i^2)^2 \} + \lambda \end{aligned} \tag{4}$$

Where λ is a constant. Because of the simplification made in the section 2.4, here, $\theta = \{D, D^1, D^2, b, a^1, a^2, c^1, c^2, \sigma_Y^2, \sigma_1^2, \sigma_2^2\}$. Indeed, $\psi_Y = \sigma_Y^2 Id_{q_Y}$, $\psi_1 = \sigma_1^2 Id_{q_1}$ and $\psi_2 = \sigma_2^2 Id_{q_2}$.

Estimation of the SEM

To maximize this function, in the EM framework (Foulley, 2002), we must solve :

$$\mathbb{E}_z^h\left[\frac{\partial}{\partial\theta}\mathscr{L}(\theta;z,h)\right]=0. \tag{5}$$

This demands that we know the derivatives of the log-likelihood function and the distribution $p_{z_i}^{h_i}$ of h_i conditional on z_i for each observation $i\in[\![1,n]\!]$. Let us introduce the following notations :

$$p_{z_i}^{h_i}=\mathcal{N}\left(\mathrm{M}_i=\begin{pmatrix}m_{1\,i}\\ m_{2\,i}\\ m_{3\,i}\end{pmatrix},\Sigma=\begin{pmatrix}\sigma_{11}&\sigma_{12}&\sigma_{13}\\ \sigma_{21}&\sigma_{22}&\sigma_{23}\\ \sigma_{31}&\sigma_{32}&\sigma_{33}\end{pmatrix}\right)$$

$$\begin{array}{ll}\widetilde{g_i}=\mathbb{E}_{z_i}^{h_i}[g_i]=m_{1\,i}\,; & \widetilde{\gamma_i}=\mathbb{E}_{z_i}^{h_i}[g_i^2]=(\mathbb{E}_{z_i}^{h_i}[g_i])^2+\mathbb{V}_{z_i}^{h_i}[g_i]={m_1}_i^2+\sigma_{11}\\ \widetilde{f_i^1}=\mathbb{E}_{z_i}^{h_i}[f_i^1]=m_{2\,i}\,; & \widetilde{\phi_i^1}=\mathbb{E}_{z_i}^{h_i}[(f_i^1)^2]=(\mathbb{E}_{z_i}^{h_i}[f_i^1])^2+\mathbb{V}_{z_i}^{h_i}[f_i^1]={m_2}_i^2+\sigma_{22}\\ \widetilde{f_i^2}=\mathbb{E}_{z_i}^{h_i}[f_i^2]=m_{3\,i}\,; & \widetilde{\phi_i^2}=\mathbb{E}_{z_i}^{h_i}[(f_i^2)^2]=(\mathbb{E}_{z_i}^{h_i}[f_i^2])^2+\mathbb{V}_{z_i}^{h_i}[f_i^2]={m_3}_i^2+\sigma_{33}\end{array}$$

For all $\widetilde{\xi}\in\{\widetilde{g},\widetilde{f^1},\widetilde{f^2},\widetilde{\gamma},\widetilde{\phi^1},\widetilde{\phi^2}\}$, we denote $\widetilde{\xi}=(\widetilde{\xi}_i)_{i=1,\dots,n}\in\mathbb{R}^n$.
The parameters of the normal distribution $p_{z_i}^{h_i}$ are explicit and have the following form :

$$\mathrm{M}_i=\Sigma_2^*{\Sigma_3^*}^{-1}\mu^*\text{ and }\Sigma=\Sigma_1^*-\Sigma_2^*{\Sigma_3^*}^{-1}{\Sigma_2^*}'\text{ where :}$$

$$\Sigma_1^*=\begin{pmatrix}(c^1)^2+(c^2)^2+1&c^1&c^2\\ c^1&1&0\\ c^2&0&1\end{pmatrix};\ \Sigma_2^*=\begin{pmatrix}((c^1)^2+(c^2)^2+1)b'&c^1a^{1\prime}&c^2a^{2\prime}\\ c^1b'&a^{1\prime}&0_{(1,q_2)}\\ c^2b'&0_{(1,q_1)}&a^{2\prime}\end{pmatrix}$$

$$\Sigma_3^*=\begin{pmatrix}((c^1)^2+(c^2)^2+1)bb'+\Psi_{\mathrm{Y}}&c^1ba^{1\prime}&c^2ba^{2\prime}\\ c^1a^1b'&a^1a^{1\prime}+\Psi_1&0_{(q_1,q_2)}\\ c^2a^2b'&0_{(q_2,q_1)}&a^2a^{2\prime}+\Psi_2\end{pmatrix};\ \mu_i^*=\begin{pmatrix}y_i-\mathrm{D}'t_i\\ x_i^1-\mathrm{D}^{1\prime}t_i^1\\ x_i^2-\mathrm{D}^{2\prime}t_i^2\end{pmatrix}$$

These results are demonstrated in Appendix B. Expressions of the first-order derivatives of $\mathscr{L}$ with respect to θ are established in Appendix C and written in the following form with $m\in\{1,2\}$:

$$\left\{\begin{array}{rcl}\frac{\partial}{\partial\mathrm{D}'}\mathscr{L}(z,h)&=&\sum_{i=1}^n\psi_{\mathrm{Y}}^{-1}(y_i-\mathrm{D}'t_i-g_ib)\,t_i'\\ \frac{\partial}{\partial\mathrm{D}^{m\prime}}\mathscr{L}(z,h)&=&\sum_{i=1}^n\psi_m^{-1}(x_i^m-\mathrm{D}^{m\prime}t_i^m-f_i^ma^m)\,t_i^{m\prime}\\ \frac{\partial}{\partial b}\mathscr{L}(z,h)&=&\sum_{i=1}^n g_i\psi_{\mathrm{Y}}^{-1}(y_i-\mathrm{D}'t_i-g_ib)\\ \frac{\partial}{\partial a^m}\mathscr{L}(z,h)&=&\sum_{i=1}^n f_i^m\psi_m^{-1}(x_i^m-\mathrm{D}^{m\prime}t_i^m-f_i^ma^m)\\ \frac{\partial}{\partial c^m}\mathscr{L}(z,h)&=&\sum_{i=1}^n f_i^m(g_i-c^2\,f_i^2-c^1\,f_i^1)\\ \frac{\partial}{\partial\sigma_{\mathrm{Y}}^2}\mathscr{L}(z,h)&=&n\,q_{\mathrm{Y}}\,\sigma_{\mathrm{Y}}^{-2}-\sigma_{\mathrm{Y}}^{-4}\sum_{i=1}^n||y_i-\mathrm{D}'t_i-g_ib||^2\\ \frac{\partial}{\partial\sigma_m^2}\mathscr{L}(z,h)&=&n\,q_m\,\sigma_m^{-2}-\sigma_m^{-4}\sum_{i=1}^n||x_i^m-\mathrm{D}^{m\prime}t_i^m-f_i^ma^m||^2\end{array}\right. \tag{6}$$

So, here formula (5) (and also (6)) develops into :

$$\begin{cases} \sum_{i=1}^{n} (y_i - \mathrm{D}' t_i - \widetilde{g}_i b)\, t_i' &= 0 \\ \sum_{i=1}^{n} (x_i^m - \mathrm{D}^{m\prime} t_i^m - \widetilde{f_i^m} a^m)\, t_i^{m\prime} &= 0 \\ \sum_{i=1}^{n} \widetilde{g}_i y_i - \widetilde{g}_i \mathrm{D}' t_i - \widetilde{\gamma}_i b &= 0 \\ \sum_{i=1}^{n} \widetilde{f_i^m} x_i^m - \widetilde{f_i^m} \mathrm{D}^{m\prime} t_i^m - \widetilde{\phi_i^m} a^m &= 0 \\ \sum_{i=1}^{n} \sigma_{12} + \widetilde{f_i^1}\widetilde{g}_i - c^2\,\sigma_{23} - c^2\,\widetilde{f_i^1 f_i^2} - \widetilde{\phi_i^1} c^1 &= 0 \\ \sum_{i=1}^{n} \sigma_{31} + \widetilde{f_i^2}\widetilde{g}_i - c^2\,\widetilde{\phi_i^2} - c^1\sigma_{32} - c^1\widetilde{f_i^1 f_i^2} &= 0 \\ n q_{\mathrm{Y}} \sigma_{\mathrm{Y}}^{-2} - \sigma_{\mathrm{Y}}^{-4} \sum_{i=1}^{n} ||y_i - \mathrm{D}' t_i||^2 + ||b||^2 \widetilde{\gamma}_i - 2(y_i - \mathrm{D}' t_i)' \widetilde{g}_i b &= 0 \\ n q_m \sigma_m^{-2} - \sigma_m^{-4} \sum_{i=1}^{n} ||x_i^m - \mathrm{D}^{m\prime} t_i^m||^2 + ||a^m||^2 \widetilde{\phi_i^m} - 2(x_i^m - \mathrm{D}^{m\prime} t_i^m)' \widetilde{f_i^m} a^m &= 0 \end{cases} \tag{7}$$

System (7) is easy to solve and the solution is given in the next section.

Results

The explicit solution of system (7) (and also of (5)) is the following :

$$\begin{aligned} \hat{b} &= \frac{(\overline{\widetilde{g}y} - \overline{yt'})(\overline{tt'})^{-1}\,\overline{\widetilde{g}t}}{\overline{\widetilde{\gamma}} - \overline{\widetilde{g}t'}(\overline{tt'})^{-1}\overline{\widetilde{g}t}} \\ \widehat{a^m} &= \frac{\overline{\widetilde{f^m} x^m} - \overline{x^m t^{m\prime}}(\overline{t^m t^{m\prime}})^{-1}\overline{\widetilde{f^m} t^m}}{\overline{\widetilde{\phi^m}} - \overline{\widetilde{f^m} t^{m\prime}}(\overline{t^m t^{m\prime}})^{-1}\overline{\widetilde{f^m} t^m}} \\ \widehat{c^1} &= \frac{(\sigma_{12} + \overline{\widetilde{f^1}\widetilde{g}})\overline{\widetilde{\phi}^2} - (\sigma_{13} + \overline{\widetilde{f^2}\widetilde{g}})(\sigma_{23} + \overline{\widetilde{f^1}\widetilde{f^2}})}{\overline{\widetilde{\phi}^1\widetilde{\phi}^2} - (\sigma_{23} + \overline{\widetilde{f^1}\widetilde{f^2}})^2} \\ \widehat{c^2} &= \frac{(\sigma_{13} + \overline{\widetilde{f^2}\widetilde{g}})\overline{\widetilde{\phi}^1} - (\sigma_{12} + \overline{\widetilde{f^1}\widetilde{g}})(\sigma_{23} + \overline{\widetilde{f^1}\widetilde{f^2}})}{\overline{\widetilde{\phi}^1\widetilde{\phi}^2} - (\sigma_{23} + \overline{\widetilde{f^1}\widetilde{f^2}})^2} \\ \widehat{\mathrm{D}'} &= (\overline{yt'} - \hat{b}\,\overline{\widetilde{g}t'})(\overline{tt'})^{-1} \\ \widehat{\mathrm{D}^{m\prime}} &= (\overline{x^m t^{m\prime}} - \widehat{a^m}\overline{\widetilde{f^m} t^{m\prime}})(\overline{t^m t^{m\prime}})^{-1} \\ \widehat{\sigma_{\mathrm{Y}}^2} &= \frac{1}{n q_{\mathrm{Y}}} \sum_{i=1}^{n} \{||y_i - \widehat{\mathrm{D}'} t_i||^2 + ||\hat{b}||^2 \widetilde{\gamma}_i - 2(y_i - \widehat{\mathrm{D}'} t_i)' \hat{b} \widetilde{g}_i\} \\ \widehat{\sigma_m^2} &= \frac{1}{n q_m} \sum_{i=1}^{n} \{||x_i^m - \widehat{\mathrm{D}^{m\prime}} t_i^m||^2 + ||\widehat{a^m}||^2 \widetilde{\phi_i^m} - 2(x_i^m - \widehat{\mathrm{D}^{m\prime}} t_i^m)' \widehat{a^m} \widetilde{f_i^m}\} \end{aligned} \tag{8}$$

The algorithm

We denote $[t]$ the t^{ieth}-iteration of the EM algorithm.

1. Initialization [1] = choice of the initial parameter values $\theta^{[0]}$.

1. In the initialization step, $\forall m \in [\![1, p]\!]$ we propose to obtain $\mathrm{D}^{m\,[0]}$ by multiple linear regression of X^m on T^m. Then, to initialize the others parameter values, we compute each approximated factor $\widetilde{f^m}^{[0]}$ and $\widetilde{g}^{[0]}$ as first principal component of $\mathrm{X}^m - \mathrm{T}^m\mathrm{D}^{m\,[0]}$ and $\mathrm{Y} - \mathrm{T}\mathrm{D}^{[0]}$. Then, we initialize a^m, σ_m^2 (resp. b, σ_y^2) through a multiple linear regression of $\mathrm{X}^m - \mathrm{T}^m\mathrm{D}^{m\,[0]}$ on $\widetilde{f^m}^{[0]}$ (resp. of $\mathrm{Y} - \mathrm{T}\mathrm{D}^{[0]}$ on $\widetilde{g}^{[0]}$). Finally, each $c^{m\,[0]}$ can be obtained by multiple linear regression of $\widetilde{g}^{[0]}$ on the p factors $\widetilde{f^m}^{[0]}$.

2. Current iteration $t \geq 1$, until stopping condition is met :
   (a) **E-step** : with $\theta^{[t-1]}$,
      i. Calculate explicitly distribution $p_{z_i}^{h_i}$ for each $i \in [\![1,n]\!]$.
      ii. Estimate the factor-values $\widetilde{g}^{[t]}$, $\widetilde{f^m}^{[t]}$, $m \in \{1,2\}$.
      iii. Calculate $\widetilde{\gamma}^{[t]}$ and $\widetilde{\phi^m}^{[t]}$, $m \in \{1,2\}$.
   (b) **M-step** :
   Update θ to $\theta^{[t]}$ by injecting $\widetilde{g}^{[t]}$, $\widetilde{\gamma}^{[t]}$ and $\widetilde{f^m}^{[t]}$, $\widetilde{\phi^m}^{[t]}$, $m \in \{1,2\}$ into the formulas in (8).
3. We used the following stopping condition with the smallest possible ϵ :

$$\sum_{k=1}^{K} \frac{|\theta^{[t+1]}[k] - \theta^{[t]}[k]|}{\theta^{[t+1]}[k]} < \epsilon \tag{9}$$

where $\theta[k]$ is the k^{ieth}-scalar element of parameter-vector θ.

4 Numerical results on simulated data

4.1 Data generation

We consider $n = 400$ units and $q_Y = q_1 = q_2 = 40$. Therefore, the 120 OV's Y, X^1, X^2 are simulated so as to be structured respectively around three factors g, f^1, f^2. Factors f^1 and f^2 are explanatory of g. Besides, we consider $r_T = r_1 = r_2 = 2$, i.e 2 covariates are simulated for each covariate matrix T, T^1 and T^2. The data is simulated as follows :

1. Choice of θ :
   (a) $D = D^1 = D^2$ = matrices filled row-wise with the ordered integer sequence ranging from 1 to 80 (indeed : $r_T * q_Y = r_1 * q_1 = r_2 * q_2 = 2 * 40$) ;
   (b) $b = a^1 = a^2$ = ordered integer sequence ranging from 1 to 40 ;
   (c) $c^1 = c^2 = 1$;
   (d) $\sigma_Y^2 = \sigma_1^2 = \sigma_2^2 = 1$.
2. Simulation of factors g, f^1, f^2 :
   (a) Simulate vectors f^1 and f^2 of $n = 400$ normally distributed random numbers with mean 0 and variance 1 (abbreviated $\forall m, \in \{1,2\}$
   $f^m \sim \mathcal{N}(0, Id_{400})$) ;
   (b) Simulate ε^g according to distribution $\varepsilon^g \sim \mathcal{N}(0, Id_{400})$;
   (c) Calculate $g = f^1 c^1 + f^2 c^2 + \varepsilon^g$.
3. Simulation of noises ε^Y, ε^1, ε^2 :
   Each element of matrix ε^Y, (resp. ε^1, ε^2) is simulated independently from distribution $\mathcal{N}(0, \sigma_Y^2 = 1)$ (resp. $\sigma_1^2 = 1$, $\sigma_2^2 = 1$).
4. Simulation of covariate matrices T, T^1, T^2 :
   Each element of matrices T, T^1, T^2 is simulated according to the standard normal distribution.
5. Simulation of Y, X^1, X^2 :
   Y, X^1, X^2 are eventually calculated through formulas in the model (1).

This simulation scheme was performed 100 times, each time yielding a set of simulated data matrices (Y, X^1, X^2). Then for each simulated data, we ran the estimation algorithm with a threshold value $\epsilon = 10^{-2}$, yielding the average results presented in section 4.2. Thus from $400 * 120 = 48000$ scalar elements of data, we estimated $3 * n = 1200$ scalar elements of factors plus
$K = 5 + 3 * 40(2 + 1) = 365$ scalar parameters, i.e : 1565 scalars.

4.2 Results

Convergence was observed in almost all cases in less than five iterations. We assess the quality of the estimations as follows.

— On the one hand, we calculate the absolute relative deviation between each simulated scalar parameter in θ and its estimate, and then average these deviations over the 100 simulations. We then produce a box-plot of the average absolute relative deviations (cf. figure 2). This makes the interpretation easier, since we only need to look at the box-plot's values and check that they are positive (because of the absolute value) and close to 0.

— On the other hand, to assess the quality of the factor estimations, we compute the 300 values of square correlations between the simulated concatenated factors (g, f^1, f^2) (respectively) and the corresponding estimates $(\widetilde{g}, \widetilde{f^1}, \widetilde{f^2})$. Once again, we produce a box-plot of these correlations (cf. figure 3) and check that it indicates values close to 1.

Figures 2 and 3 show clearly that the estimates are very close to the actual quantities. Indeed, on figure 2, the median of average absolute relative deviations is 0.018, first and third quartiles being 0.015 and 0.023 respectively. On figure 3, the median of square correlations is 0.998, first and third quartiles being 0.997 and 0.999 respectively. So, factor g (respectively f^1 and f^2) turn out to be drawn towards the principal direction underlying the bundles made up by OV's Y (respectively X^1 and X^2). Now, we may legitimately wonder how the quality of estimations could be affected by the number of observations and the number of OV's in each block. In the following section we provide a sensitivity analysis performed to investigate this issue.

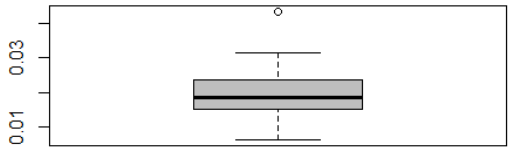


FIGURE 2 – Box plot of the average absolute relative deviations of the simulated parameters and their estimates.

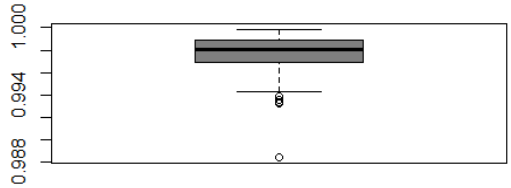


FIGURE 3 – Box plot of the correlations of the simulated factors and their estimates

4.3 Sensitivity analysis of estimations

We performed a sensitivity analysis on the simulated data presented in section 4.1. The purpose was to investigate the influence of the block-sizes (n, q_Y, q_1, q_2) on the quality of estimation, both of parameters and factors. To simplify the analysis, we imposed $q_Y = q_1 = q_2 = q$ and varied n and q separately, i.e. studied the cases $n = 50, 100, 200, 400$ with $q = 40$ and $q = 5, 10, 20, 40$ with $n = 400$. Each case was simulated 100 times. Therefore, we simulated 800 data-sets.

Sensitivity with respect to the number n of observations

In this section, we study the evolution with n of the average estimation of structural coefficients c^1 and c^2 and parameter σ_Y^2 with respect to their actual values, all equal to 1, and that of the correlations of factors with their estimates. The number of OV's is set to $q = 40$ in each block. Figure 6 graphs these evolutions (average value of estimate in plain line), including a 95% confidence-interval about each average estimate (dotted line). This figure shows that the biases and the standard deviations are, as expected, more important for little values of n, but also that the quality of estimation is already quite good for $n = 50$. As for the correlations of factors with their estimates, figure 4 shows that they increase and get close to one as n increases, with a dispersion decreasing to 0. However, even for $n = 50$, the correlations are mostly above 0.95, indicating that the factors are correctly reconstructed.

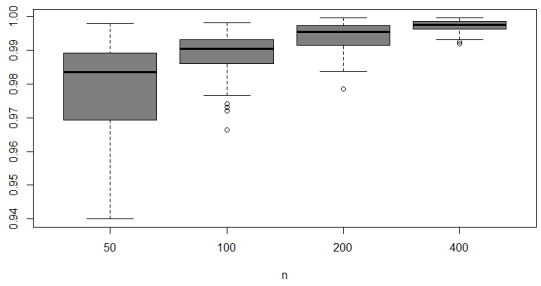


FIGURE 4 – Box plots of the correlations of simulated factors and their estimates for various values of n.

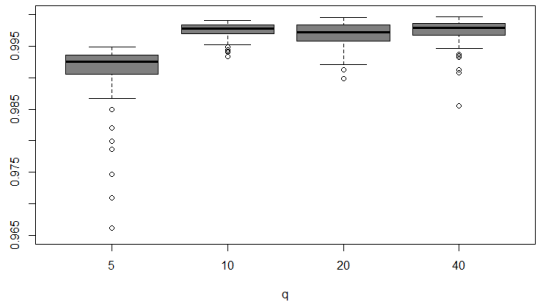


FIGURE 5 – Box plots of the correlations of simulated factors and their estimates for various values of q.

Sensitivity with respect to the number q of OV's in each block

Likewise, we study the evolution of the average estimates of c^1, c^2, σ_Y^2 and the correlation of factors with their estimates for different values of q, with n set to 400. We observe that, unsurprisingly, the biases and the standard deviations decrease as q increases (cf. figure 6). We observe that they stabilize even faster with q than with n, particularly σ_Y^2. Indeed, from $q = 10$ on, the confidence interval is narrow enough. As for the factors, figure 5 shows that their correlations with their estimates are already very close to 1 for $q = 5$, with a very small variance, and keep increasing with q. To sum things up, the sample size n proved to have more impact on the quality of parameter estimation and factor reconstruction than the number of OV's. Now, the quality of factor reconstruction remains high for rather small values of n or q. We advise to use a minimal sample size of

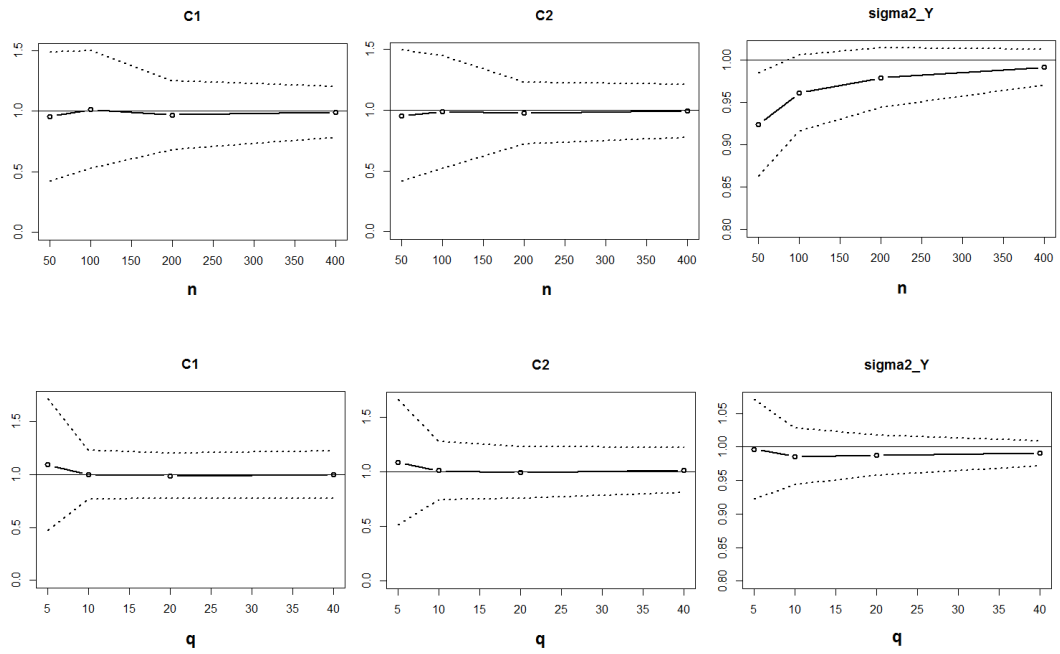


FIGURE 6 – Average estimates of c^1, c^2, σ^2_Y and 95% confidence intervals as a function of n and as a function of q.

$n = 100$ to obtain really stable structural coefficients. Above this threshold, n has but little impact on the biases and standard deviations of estimates.

5 An application to environmental data

5.1 Data presentation

We apply our model to the data-set *genus*, built from the CoForChange database and provided in the R-package SCGLR by Mortier et al. (2014). It gives the abundances of 27 common tree genera present in the tropical moist forest of the Congo-Basin, and the measurements of 40 geo-referenced environmental variables, for $n = 1000$ inventory plots (observations). Some of the geo-referenced environmental variables describe 16 physical factors pertaining to topography, geology and rainfall description. The remaining variables characterize vegetation through the enhanced vegetation index (EVI) measured on 16 dates.
In this section, we aim at modeling the tree abundances from the other variables, while reducing the dimension of data. The dependent block of variables Y therefore consists of the $q_Y = 27$ tree species counts divided by the plot-surface. A PCA of the geo-referenced environmental variables and the photosynthetic activity variables confirms that EVI measures are clearly separated from the other variables (cf. figure 7). Indeed, figure 7 shows two variable-bundles with almost orthogonal central directions. This justifies using our model (cf. section 5.2) with $p = 2$ explanatory groups, one of them (X^1) gathering $q_1 = 16$ rainfall measures and location variables (longitude, latitude and altitude), and the second one (X^2), the $q_2 = 23$ EVI measures. Besides, in view of the importance of the geological substrate on the spatial distribution of tree species in the Congo Basin, showed by Fayolle et al. (2012), we chose to put nominal variable *geology* in a block T. This block therefore contains constant 1 plus all the indicator variables of geology but one, which will therefore be the reference value. Geology having 5 levels, T has 5 columns.

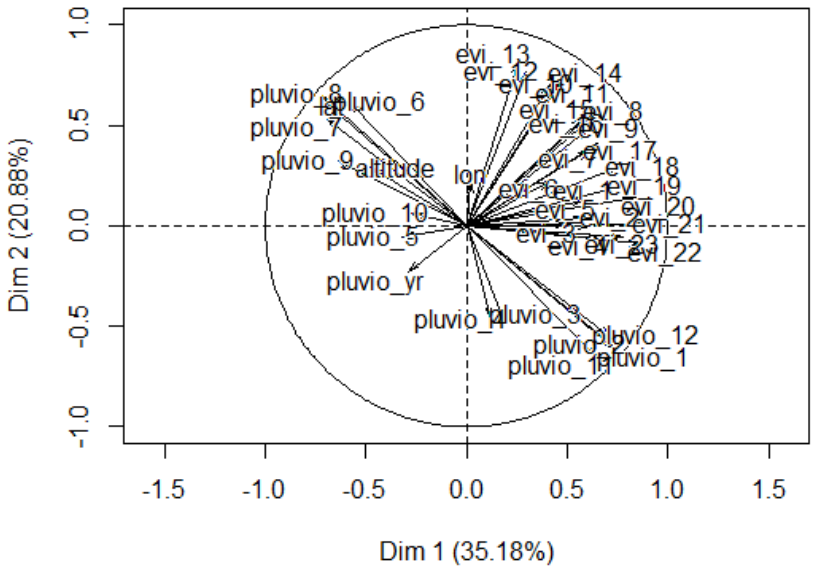


FIGURE 7 – Correlation-scatterplot yielded by the PCA of the X^1 and X^2 geo-referenced environmental variables (obtained with the FactoMineR R-package).

5.2 Model specification with geologic covariates

Here is the model used with the variable-blocks designed in section 5.1. :

$$\begin{cases} Y &= TD + gb' + \varepsilon^Y \\ X^1 &= \mathbb{1}_n d^{1\prime} + f^1 a^{1\prime} + \varepsilon^1 \\ X^2 &= \mathbb{1}_n d^{2\prime} + f^2 a^{2\prime} + \varepsilon^2 \\ g &= f^1 c^1 + f^2 c^2 + \varepsilon^g \end{cases}$$

where $n = 1000$, $q_Y = 27$, $q_1 = 16$, $q_2 = 23$ and $r_T = 5$. The first row of D is a parameter vector that contains the means of the Y's, and the other rows contain the overall effects of the geological substrates with respect to the reference one. Next section presents the model's parameter-estimates where, in Table 1, each row r of D is noted $D[r,]$.

5.3 Results

With a threshold value $\varepsilon = 10^{-3}$, convergence was reached after 58 iterations. Some parameter-estimates are presented in Tables 1 and 2. For practical reasons, the remaining tables of parameter-estimates are given in the supplementary material.

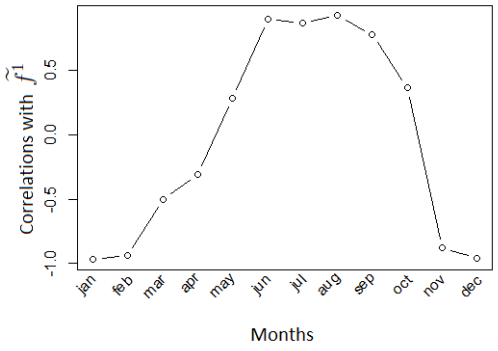


FIGURE 8 – Correlations of $\widetilde{f^1}$ with the monthly variables of X^1 : two rainfall regimes.

	Parameter-estimates						
Variables	D[1,]	D[2,]	D[3,]	D[4,]	D[5,]	b	Correlations with $\widetilde{g}$
gen1	0.76	0.16	0.06	0.68	-0.12	-0.13	-0.14
gen2	0.54	-0.28	-0.03	-0.03	-0.28	0.47	0.58
gen3	0.41	-0.23	-0.02	0.25	-0.37	0.29	0.36
gen4	0.12	0.14	0.03	0.52	0.30	0.09	0.15
gen5	0.31	0.15	0.19	-0.20	0.84	0.09	0.16
gen6	0.55	-0.12	-0.26	0.06	-0.02	0.14	0.18
gen7	0.46	0.06	-0.04	-0.37	0.43	0.14	0.18
gen8	0.55	0.04	-0.09	-0.16	0.04	0.42	0.52
gen9	0.92	-0.54	0.26	-0.66	-0.61	0.07	0.03
gen10	0.68	0.40	0.20	0.37	0.06	-0.32	-0.39
gen11	1.74	-0.50	-0.21	0	-0.67	0.33	0.39
gen12	0.87	0.14	0.73	-0.51	-0.21	0.24	0.26
gen13	1.08	-0.09	-0.37	-0.02	-0.53	0.26	0.29
gen14	0.41	-0.16	-0.10	0.12	-0.36	-0.05	-0.07
gen15	0.51	0.01	-0.11	0.27	-0.18	0.29	0.37
gen16	0.50	-0.19	-0.01	0.55	-0.27	0.1	0.14
gen17	0.79	-0.54	-0.20	-0.52	-0.45	0.39	0.45
gen18	0.16	-0.05	0.20	0.03	-0.03	0.18	0.23
gen19	0.34	0.06	0.41	-0.11	0.38	0.23	0.31
gen20	0.49	0.02	-0.21	0.08	0.14	-0.2	-0.24
gen21	0.79	-0.30	-0.12	0.71	-0.13	0.12	0.19
gen22	0.32	-0.07	-0.07	0.38	-0.11	0.23	0.3
gen23	1.02	-0.28	-0.31	0	-0.07	0.46	0.58
gen24	0.80	-0.23	-0.08	0.22	-0.47	0.57	0.7
gen25	0.60	-0.16	-0.04	0.97	-0.49	0.41	0.53
gen26	0.84	0.22	0.27	-0.70	0.82	0.04	0.07
gen27	0.27	0.41	0.69	-0.24	0.56	0.08	0.11

TABLEAU 1 – Application to the *genus* data with geologic covariate : estimates of parameters D and b, and correlations of $\widetilde{g}$ with the variables in Y.

Scalar parameter-estimates				
c^1	c^2	σ_1^2	σ_2^2	σ_Y^2
0.35	0.01	0.50	0.53	0.84

TABLEAU 2 – Application to *genus* data with geologic covariate : scalar parameter-estimates.

It can be seen in Tables 1 and 3 that for certain species, the geologic substrate seems to be of great importance (e.g. for gen1, gen5, gen7, gen9, gen12, gen16, gen21, gen25, gen26, gen27), whereas for others, it only has a small impact on the abundances (e.g. for gen2, gen6, gen8, gen10, gen18, gen20, gen23). Moreover, Table 1 shows that the correlations of $\widetilde{g}$ with Y are high in absolute value only for few variables : *gen2*, *gen23*, *gen24* and *gen25*. Therefore, only these are well accounted for by our model. Although we have carried out the analysis with variables *gen2*, *gen3*, *gen8*, *gen10*, *gen11*, *gen15*, *gen17*, *gen23*, *gen24* and *gen25*, the results are practically the same as when

we take all variables. The correlations of $\widetilde{f^1}$ with variables *pluvio_1* to *pluvio_12* of X^1 show two rainfall regimes (cf. figure 8). Indeed, *pluvio_1* corresponds to january, *pluvio_2*, to february, ..., *pluvio_12* to december. The Central African Republic has a tropical climate : the dry season ranges from November to April and the rainy season from June to September. Figure 8 shows that $\widetilde{f^1}$ is positively correlated to the rainfalls of the rainy season and negatively to those of the dry one.

5.4 Assessing the model quality through re-sampling

To assess the stability of results and thus, validate the models with covariate, we used a 5-fold re-sampling technique : 5 separate 200 units-samples were randomly extracted from the complete *genus*. For each, we obtained estimated parameters and factors. Then, for each sample, we computed an average Mean Square Error (MSE) and an average correlation of the parameter-estimates obtained on the sample with those obtained on the complete data. Finally, on each sample, we calculated an average MSE and correlation of the factor-estimates obtained on the sample with the corresponding ones obtained on the complete data for the units belonging to the sample.

Figure 9 (resp. 10) shows the average MSE (resp. the correlation) between parameters estimated on 5 data samples $\theta_{s\in[\![1,5]\!]}$ and parameters estimated on the complete data θ. More precisely, for these average MSE (respectively correlations), the median is $3.85 * 10^{-3}$ (resp. 0.99), the first quartile is $1.95 * 10^{-3}$ (resp. 0.99) and the third quartile is $6.17 * 10^{-3}$ (resp. 0.99). These values are close to 0 (resp. 1). So, we can be rather confident in the estimates of parameters obtained in the previous section.

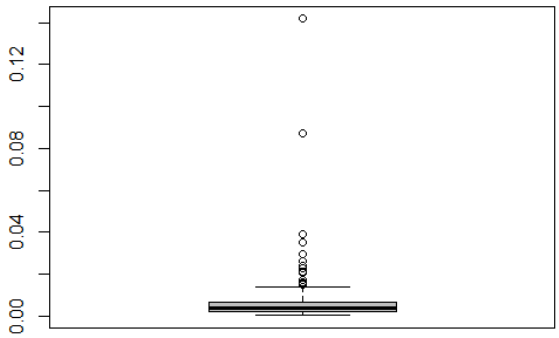


FIGURE 9 – Box plot of the average MSE's of the parameter-estimates obtained on the 5 *genus* data sub-samples and those obtained on the complete data.

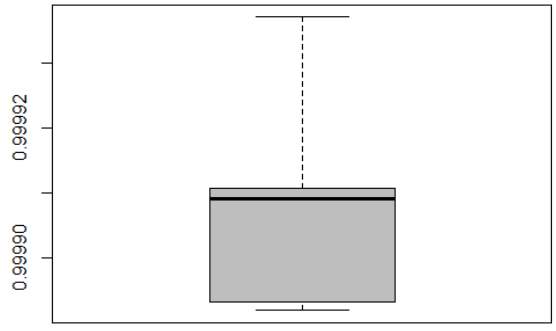


FIGURE 10 – Box plot of the average correlations of the parameter-estimates obtained on the 5 *genus* data sub-samples and those obtained on the complete data.

Figure 11 and figure 12 respectively give the box-plot of the factors' average MSE and correlation for each of the 5 samples. More precisely, for these average MSE's (respectively correlations), the median is $1.15 * 10^{-2}$ (resp. 0.98), the first quartile is $7.44 * 10^{-3}$ (resp. 0.98) and the third quartile is $3.53 * 10^{-2}$ (resp. 0.99). These values are close enough to 0 (resp. 1) to allow us to be confident in the estimates obtained on the complete data.

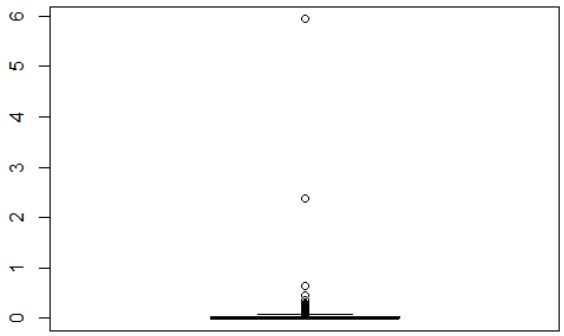


FIGURE 11 – Box-plot of the average MSE of factor-estimates.

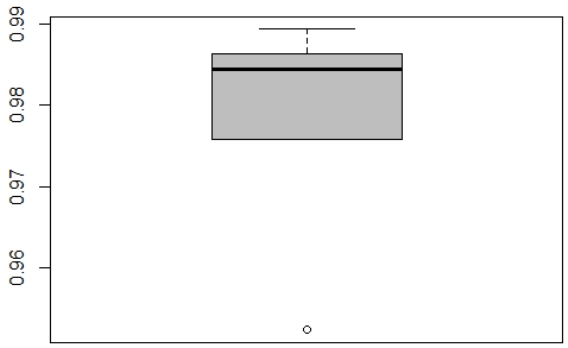


FIGURE 12 – Box-plot of the average correlation of factor-estimates.

6 Conclusion

The maximum-likelihood estimation method is known to be a stringent method of estimation having nice properties. In the context of estimation methods of an SEM, the LISREL approach is based on likelihood maximization, contrary to PLS, THEME, and other component-based methods. However, LISREL mainly focuses on the variance-covariance structure, the likelihood of which it maximizes under constraints. The LV's scores estimation is based on a least squares technique performed on the mere measurement equations. To estimate both parameters and scores in a row, we proposed to carry out likelihood maximization of the complete model through the EM algorithm. This approach assumes that LV's are factors, which is less constraining than assuming they are components. Therefore, this approach has clear assets over the more classical one. Sensitivity analysis allowed to assess its performances. Eventually, the application on environmental data proved satisfactory and demonstrated how to practically use this method.

Acknowledgments

The forest inventories were funded by the French Agency for Development (AFD) through the PARPAF project. We wish to thank S. Chong (TCA), A. Banos (SCAF), and the ‘"Ministère des Eaux, Forêts, Chasse et Pêche" of the Central African Republic for authorizing access to the inventory data, and the field teams who drew up these inventories. This inventory is part of the ErA Net BiodivERsA CoForChange project, funded by the National Research Agency (ANR) and the Natural Environment Research Council (NERC), involving 16 European, African and international partners and a number of timber companies (see the list on the website, http ://www.coforchange.eu).

Appendix A. Calculation of the complete data log-likelihood function $\mathscr{L}$

In the case of the simplified model (3), $p = 2$, $\psi_Y = \sigma_Y^2 \mathrm{I}d_{q_Y}$, $\psi_1 = \sigma_1^2 \mathrm{I}d_{q_1}$ and $\psi_2 = \sigma_2^2 \mathrm{I}d_{q_2}$, and for observation i we have,

$$\begin{aligned}
p(z_i, h_i; \theta) &= p(y_i, x_i^1, x_i^2, g_i, f_i^1, f_i^2; \theta) \\
&= p(y_i, x_i^1, x_i^2 | g_i, f_i^1, f_i^2; \theta) p(g_i, f_i^1, f_i^2; \theta) \\
&= p(y_i, x_i^1, x_i^2 | g_i, f_i^1, f_i^2; \theta) p(g_i | f_i^1, f_i^2; \theta) p(f_i^1, f_i^2; \theta) \\
&= p(y_i, x_i^1, x_i^2 | g_i, f_i^1, f_i^2; \theta) p(g_i | f_i^1, f_i^2; \theta) p(f_i^1; \theta) p(f_i^2; \theta) \\
&= p(y_i, x_i^1, x_i^2 | g_i, f_i^1, f_i^2; \theta) p(g_i | f_i^1, f_i^2; \theta) p(f_i^1) p(f_i^2) \\
&= p(x_i^1, x_i^2 | y_i, g_i, f_i^1, f_i^2; \theta) p(y_i | g_i, f_i^1, f_i^2; \theta) p(g_i | f_i^1, f_i^2; \theta) p(f_i^1) p(f_i^2) \\
&= p(x_i^1, x_i^2 | f_i^1, f_i^2; \theta) p(y_i | g_i; \theta) p(g_i | f_i^1, f_i^2; \theta) p(f_i^1) p(f_i^2) \\
&= p(x_i^1 | x_i^2, f_i^1, f_i^2; \theta) p(x_i^2 | f_i^1, f_i^2; \theta) p(y_i | g_i; \theta) p(g_i | f_i^1, f_i^2; \theta) p(f_i^1) p(f_i^2) \\
&= p(x_i^1 | f_i^1; \theta) p(x_i^2 | f_i^2; \theta) p(y_i | g_i; \theta) p(g_i | f_i^1, f_i^2; \theta) p(f_i^1) p(f_i^2)
\end{aligned}$$

where $\theta = \{\mathrm{D}, \mathrm{D}^1, \mathrm{D}^2, b, a^1, a^2, c^1, c^2, \psi_Y, \psi_1, \psi_2\}$ is the set of model parameters. Therefore,

$$\mathscr{L}(\theta; z_i, h_i) = \mathscr{L}(\theta; x_i^1 | f_i^1) + \mathscr{L}(\theta; x_i^2 | f_i^2) + \mathscr{L}(\theta; y_i | g_i) + \mathscr{L}(\theta; g_i | f_i^1, f_i^2) + \mathscr{L}(f_i^1) + \mathscr{L}(f_i^2)$$

Because of the model and the normal distribution properties we obtain :
$x_i^m | f_i^m \sim \mathscr{N}(t_i^{m\prime} \mathrm{D}^m + f_i^m a^{m\prime}, \psi_{\mathrm{X}^m})$
$y_i | g_i \sim \mathscr{N}(t_i{}' \mathrm{D} + g_i b', \psi_Y)$
$g_i | f_i^1, f_i^2 \sim \mathscr{N}(f_i^1 c^1 + f_i^2 c^2, 1)$
$f_i^m \sim \mathscr{N}(0, 1)$
Then, we get the complete data log-likelihood function (4), where λ is a constant. Also, the set of model parameters $\theta = \{\mathrm{D}, \mathrm{D}^1, \mathrm{D}^2, b, a^1, a^2, c^1, c^2, \psi_Y, \psi_1, \psi_2\}$ in our case corresponds to
$\theta = \{\mathrm{D}, \mathrm{D}^1, \mathrm{D}^2, b, a^1, a^2, c^1, c^2, \sigma_Y^2, \sigma_1^2, \sigma_2^2\}$ because of the simplification made in the section 2.3. Indeed, $\psi_Y = \sigma_Y^2 \mathrm{I}d_{q_Y}$, $\psi_1 = \sigma_1^2 \mathrm{I}d_{q_1}$ and $\psi_2 = \sigma_2^2 \mathrm{I}d_{q_2}$.
Therefore, we can also write the complete data log-likelihood function with replacing $ln|\psi_Y|$ (resp. $\forall m \in \{1,2\}$, $ln|\psi_m|$) by $q_Y ln(\sigma_Y^2)$ (resp. $\forall m \in \{1,2\}$, $q_m ln(\sigma_m^2)$).

Appendix B. Distribution of $h_i|z_i$

In the case of the simplified model (3), $p=2$, $\psi_Y = \sigma_Y^2 Id_{q_Y}$, $\psi_1 = \sigma_1^2 Id_{q_1}$ and $\psi_2 = \sigma_2^2 Id_{q_2}$, and for observation i, the normality of the distribution of $h_i|z_i$ presented in section 3.1.2. derives from the classical result [2] about the conditioning of normally distributed variables. Before using this result, we calculate the joint distribution of $(g_i, f_i^1, f_i^2, y_i, x_i^1, x_i^2)$.

We know that, for observation i,
$y_i \sim \mathcal{N}(D' t_i, b((c^1)^2 + (c^2)^2 + 1)b' + \Psi_Y)$
$x_i^m \sim \mathcal{N}(D^{m\prime} t_i^m, a^m a^{m\prime} + \Psi_m)$
$g_i \sim \mathcal{N}(0, (c^1)^2 + (c^2)^2 + 1)$
$f_i^m \sim \mathcal{N}(0,1)$
Then, after calculating the required covariances we obtain,

$$(g_i, f_i^1, f_i^2) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}(c^1)^2 + (c^2)^2 + 1 & c^1 & c^2\\ c^1 & 1 & 0\\ c^2 & 0 & 1\end{pmatrix}\right)$$

and,

$$(y_i, x_i^1, x_i^2) \sim \mathcal{N}\left(\begin{pmatrix}D' t_i\\ D^{1\prime} t_i^1\\ D^{2\prime} t_i^2\end{pmatrix}, \begin{pmatrix}((c^1)^2 + (c^2)^2 + 1)bb' + \Psi_Y & c^1 b a^{1\prime} & c^2 b a^{2\prime}\\ c^1 a^1 b' & a^1 a^{1\prime} + \Psi_1 & 0_{(q_1,q_2)}\\ c^2 a^2 b' & 0_{(q_2,q_1)} & a^2 a^{2\prime} + \Psi_2\end{pmatrix}\right)$$

Then, after calculating the required covariances we obtain the joint distribution, $(g_i, f_i^1, f_i^2, y_i, x_i^1, x_i^2) \sim \mathcal{N}(M_i^*, \Sigma^*)$ such that,

$$M_i^* = \begin{pmatrix}0_{(3,1)}\\ D' t_i\\ D^{1\prime} t_i^1\\ D^{2\prime} t_i^2\end{pmatrix} \text{ and } \Sigma^* = \begin{pmatrix}\Sigma_1^* & \Sigma_2^*\\ \Sigma_2^{*\prime} & \Sigma_3^*\end{pmatrix}.$$

Where, $\Sigma_1^* = \begin{pmatrix}(c^1)^2 + (c^2)^2 + 1 & c^1 & c^2\\ c^1 & 1 & 0\\ c^2 & 0 & 1\end{pmatrix}$; $\Sigma_2^* = \begin{pmatrix}((c^1)^2 + (c^2)^2 + 1)b' & c^1 a^{1\prime} & c^2 a^{2\prime}\\ c^1 b' & a^{1\prime} & 0_{(1,q_2)}\\ c^2 b' & 0_{(1,q_1)} & a^{2\prime}\end{pmatrix}$;

$$\Sigma_3^* = \begin{pmatrix}((c^1)^2 + (c^2)^2 + 1)bb' + \Psi_Y & c^1 b a^{1\prime} & c^2 b a^{2\prime}\\ c^1 a^1 b' & a^1 a^{1\prime} + \Psi_1 & 0_{(q_1,q_2)}\\ c^2 a^2 b' & 0_{(q_2,q_1)} & a^2 a^{2\prime} + \Psi_2\end{pmatrix}.$$

Finally, we use result (10) and obtain the distribution, $h_i|z_i \sim \mathcal{N}(M_i, \Sigma)$ where, $M_i = \Sigma_2^* \Sigma_3^{*-1} \mu_i^*$ and $\Sigma = \Sigma_1^* - \Sigma_2^* \Sigma_3^{*-1} \Sigma_2^{*\prime}$, such that $\mu_i^* = \begin{pmatrix}y_i - D' t_i\\ x_i^1 - D^{1\prime} t_i^1\\ x_i^2 - D^{2\prime} t_i^2\end{pmatrix}$.

2. If two variables X_1 and X_2 are normally distributed such that,
$\begin{pmatrix}X_1\\X_2\end{pmatrix} \sim \mathcal{N}\left(\mu = \begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}, \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\right)$
where, μ_1 $(r \times 1)$, μ_2 $(s \times 1)$, Σ_{11} $(r \times r)$, $\Sigma_{12}$$(r \times s)$, Σ_{21} $(s \times r)$ and Σ_{22} $(s \times s)$;
then,

$$(X_1|X_2 = x_2) \sim \mathcal{N}(M = \mu_1 + \Sigma_{12}{\Sigma_{22}}^{-1}(x_2 - \mu_2), \phi = \Sigma_{11} - \Sigma_{12}{\Sigma_{22}}^{-1}\Sigma_{21}) \quad (10)$$

Appendix C. Calculation of the first-order derivatives of $\mathscr{L}$

We calculate the first-order derivatives of the complete data log-likelihood function (4), where $\theta = \{D, D^1, D^2, b, a^1, a^2, c^1, c^2, \psi_Y, \psi_1, \psi_2\}$, $\psi_Y = \sigma_Y^2 I d_{q_Y}$, $\psi_1 = \sigma_1^2 I d_{q_1}$ and $\psi_2 = \sigma_2^2 I d_{q_2}$.
There are matrix-parameters (D, D^1, D^2), vector-parameters (b, a^1, a^2) and scalar parameters $(c^1, c^2, \sigma_Y^2, \sigma_1^2, \sigma_2^2)$. Then, $\mathscr{L}$ is a sum of three types of functions : the logarithm, the square function and a quadratic form function $(w - X\beta)'\Gamma(w - X\beta)$, where Γ is symmetric and w $(q \times 1)$, X $(q \times m)$, β $(m \times 1)$ and Γ $(q \times q)$. The first-order derivatives of the logarithm function and the square function are in our case trivial. The first-order derivative of $(w - X\beta)'\Gamma(w - X\beta)$ with respect to X is less trivial but necessary.

$$\begin{aligned} d_X[(w - X\beta)'\Gamma(w - X\beta)] &= (w - X\beta)'\Gamma(-dX\beta) + (-dX\beta)'\Gamma(w - X\beta) \\ &= -2(w - X\beta)'\Gamma(dX\beta) \\ &= tr[-2(w - X\beta)'\Gamma(dX\beta)] \\ &= tr[-2\beta(w - X\beta)'\Gamma dX] \\ &=< -2\beta(w - X\beta)'\Gamma | dX > \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{d}{dX}[(w - X\beta)'\Gamma(w - X\beta)] &= (-2\beta(w - X\beta)'\Gamma)' \\ &= -2(\beta(w - X\beta)'\Gamma)' \\ &= -2\Gamma(w - X\beta)\beta' \end{aligned}$$

Likewise, we establish that :

$$\frac{\partial}{\partial D'}\mathscr{L}(z, h) = \sum_{i=1}^{n} \psi_Y^{-1}(y_i - D' t_i - g_i b)\, t_i{}'$$

Similar reasoning can be applied to D^m and allows to obtain the second row of (6). Concerning the third and the fourth row of (6), we use the classical result :

$$\frac{\partial}{\partial \beta}[(w - X\beta)'\Gamma(w - X\beta)] = -2X'\Gamma(w - X\beta)$$

Eventually, the fifth, the sixth and the eighth rows of (6) are obtained in a trivial way.

Appendix D. Table (3) of section 5.

Variables	Differences	Variables	Differences
gen1	0.80	gen15	0.45
gen2	*0.28*	gen16	0.82
gen3	0.62	gen17	0.54
gen4	0.52	gen18	*0.25*
gen5	1.04	gen19	0.52
gen6	*0.32*	gen20	*0.35*
gen7	0.80	gen21	1.01
gen8	*0.20*	gen22	0.49
gen9	0.92	gen23	*0.31*
gen10	*0.40*	gen24	0.69
gen11	0.67	gen25	1.46
gen12	1.24	gen26	1.52
gen13	0.53	gen27	0.93
gen14	0.48		

TABLEAU 3 – Application to the *genus* data with geologic covariate : Differences between maximal and minimal values of geologic effects D[1,], D[1,]+D[2,], D[1,]+D[3,], D[1,]+D[4,], D[1,]+D[5,] (highlights on the greater differences, italics on the smaller).

Références

Andrade, D. t. F. and Helms, R. W. (1984). Maximum likelihood estimates in the multivariate normal with patterned mean and covariance via the em algorithm. *Communications in Statistics - Theory and Methods*, 13(18) :2239–2251. 62

Arbuckle, J. L., Marcoulides, G. A., and Schumacker, R. E. (1996). Full information estimation in the presence of incomplete data. *Advanced structural equation modeling : Issues and techniques*, 243 :277. 62

Bry, X. and Verron, T. (2015). THEME : THEmatic Model Exploration through Multiple Co-Structure maximization. *Journal of Chemometrics*, 29(12) :637–647. 62

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38. 62, 65

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in Covariance Components Models. *Journal of the American Statistical Association*, 76(374) :341–353. 62

Fayolle, A., Engelbrecht, B., Freycon, V., Mortier, F., Swaine, M., Réjou-Méchain, M., Doucet, J.-L., Fauvet, N., Cornu, G., and Gourlet-Fleury, S. (2012). Geological Substrates Shape Tree Species and Trait Distributions in African Moist Forests. *PLoS ONE*, 7(8) :e42381. 71

Foulley, J.-L. (2002). Algorithme EM : Théorie et application au modèle mixte. *Journal de la Société française de statistique*, 143(3-4) :57–109. 66

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, 57(2) :239–251. 61

Jöreskog, K. G. (2000). Latent Variable Scores and their uses. Scientific Software International. 62

Jöreskog, K. G. and Sörbom, D. (1982). Recent Developments in Structural Equation Modeling. *Journal of Marketing Research*, 19(4) :404–416. 61

Lee, S.-Y. and Tang, N.-S. (2006). Bayesian Analysis of Nonlinear Structural Equation Models with Nonignorable Missing Data. *Psychometrika*, 71(3) :541–564. 62

McDonald, R. P. (1996). Path Analysis with Composite Variables. *Multivariate Behavioral Research*, 31(2) :239–270. 62

Mortier, F., Trottier, C., Cornu, G., and Bry, X. (2014). SCGLR-An R Package for Supervised Component Generalized Linear Regression. *Journal of Statistical Software*. 71

Muthén, B., Kaplan, D., and Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52(3) :431–462. 61, 62

Muthén, L. and Muthén, B. (1998). MPlus User's Guide. 61

Tang, M.-L. and Lee, S.-Y. (1998). Analysis of structural equation models with censored or truncated data via EM algorithm. *Computational statistics & data analysis*, 27(1) :33–46. 62

3.3 Résultats complémentaires de l'application environnementale : le cas du modèle sans la covariable *géologie*

3.3.1 Le modèle sans covariables

Dans cette section l'intérêt est porté sur les conséquences d'une modélisation sans covariables T. Le modèle est alors :

$$\begin{cases} Y &= \mathbb{1}_n d' + gb' + \varepsilon^Y \\ X^1 &= \mathbb{1}_n d^{1\prime} + f^1 a^{1\prime} + \varepsilon^1 \\ X^2 &= \mathbb{1}_n d^{2\prime} + f^2 a^{2\prime} + \varepsilon^2 \\ g &= f^1 c^1 + f^2 c^2 + \varepsilon^g \end{cases}$$

où les covariables T sont réduites au vecteur unitaire $\mathbb{1}_n$. Les dimensions $n = 1000$, $q_Y = 27$, $q_1 = 16$, $q_2 = 23$ restent inchangées et d' (resp. $d^{1\prime}$ et $d^{2\prime}$) est un vecteur paramètre contenant les moyennes des variables Y (resp. X^1 and X^2).

Résultats

Pour $\varepsilon = 10^{-3}$, la convergence est observée après 49 itérations. Les estimations des paramètres sont présentées dans les tableaux 3.1, 3.2, 3.4 and 3.6.

	Estimations des paramètres		
Variables	d	b	Corrélation avec $\widetilde{g}$
gen1	0.95	0.16	0.19
gen2	0.45	0.34	0.41
gen3	0.38	0.34	0.42
gen4	0.35	0.18	0.22
gen5	0.45	-0.15	-0.18
gen6	0.54	0.18	0.22
gen7	0.44	-0.16	-0.2
gen8	0.51	0.29	0.35
gen9	0.56	-0.01	-0.01
gen10	0.86	-0.2	-0.24
gen11	1.54	0.35	0.43
gen12	0.74	-0.04	-0.04
gen13	0.94	0.28	0.35
gen14	0.35	0.1	0.12
gen15	0.54	0.39	0.47
gen16	0.58	0.32	0.4
gen17	0.48	0.17	0.21
gen18	0.17	0.16	0.2
gen19	0.41	0.04	0.05
gen20	0.53	-0.13	-0.16
gen21	0.93	0.41	0.5
gen22	0.40	0.39	0.48
gen23	0.96	0.45	0.56
gen24	0.74	0.54	0.66
gen25	0.76	0.64	0.78
gen26	0.85	-0.33	-0.41
gen27	0.40	-0.19	-0.24

TABLEAU 3.1 – Application aux données *genus* sans covariables T : estimations des paramètres D et b, et des corrélations de $\widetilde{g}$ avec Y.

Estimations des paramètres scalaires				
c^1	c^2	σ_1^2	σ_2^2	σ_Y^2
0.27	-0.07	0.50	0.53	0.90

TABLEAU 3.2 – Application aux données *genus* sans la covariable *géologie* : estimations des paramètres scalaires.

Le tableau 3.2 et le tableau similaire dans le cas de la modélisation avec covariables T (cf. l'article), montrent que peu importe la présence des covaribles T, les variables pluviométriques et géographiques X^1 jouent un rôle plus important que les variables de photosynthèse X^2 sur l'abondance des espèces d'arbres Y. Les estimations de σ_1^2, σ_2^2, σ_Y^2 ne changent pas significativement. Dans les deux cas, les régimes pluviométriques sont mis en évidence. Cependant, le tableau en annexe D de l'article montre l'impact de la nature du sol sur l'abondance des espèces d'arbres gen1, gen5, gen7, gen9, gen12, gen16, gen21, gen25, gen26 et gen27. Par conséquent la présence des covariables T liées à la nature du sol (*geologie*) dans le modèle semble importante et avoir du sens.

3.3.2 Tableaux supplémentaires de l'application aux données environnementales

	Estimations des paramètres		
Variables	d^1	a^1	Corrélations avec $\widetilde{f^1}$
altitude	4.43	0.62	0.66
pluvio_yr	44.45	0.16	0.17
pluvio_1	2.48	-0.91	-0.97
pluvio_2	4.32	-0.88	-0.94
pluvio_3	9.65	-0.47	-0.5
pluvio_4	8.56	-0.28	-0.3
pluvio_5	6.68	0.26	0.28
pluvio_6	5.98	0.83	0.89
pluvio_7	4.78	0.81	0.86
pluvio_8	4.17	0.86	0.91
pluvio_9	11.46	0.72	0.77
pluvio_10	10.17	0.34	0.36
pluvio_11	4.36	-0.83	-0.88
pluvio_12	2.13	-0.9	-0.96
lon	14.57	0.04	0.04
lat	2.49	0.92	0.98

TABLEAU 3.3 – Application aux données *genus* sans covariables T : estimation des paramètres d^1 et a^1, et des corrélations de $\widetilde{f^1}$ avec les variables X^1.

	Estimations des paramètres		
Variables	d^1	a^1	Corrélations avec $\widetilde{f^1}$
altitude	4.43	0.63	0.66
pluvio_yr	44.45	0.16	0.17
pluvio_1	2.48	-0.92	-0.97
pluvio_2	4.32	-0.89	-0.94
pluvio_3	9.65	-0.48	-0.5
pluvio_4	8.56	-0.29	-0.31
pluvio_5	6.68	0.26	0.28
pluvio_6	5.98	0.84	0.89
pluvio_7	4.78	0.82	0.86
pluvio_8	4.17	0.87	0.92
pluvio_9	11.46	0.73	0.77
pluvio_10	10.17	0.34	0.36
pluvio_11	4.36	-0.84	-0.88
pluvio_12	2.13	-0.91	-0.96
lon	14.57	0.04	0.05
lat	2.49	0.93	0.98

TABLEAU 3.4 – Application aux données *genus* sans covariables T : estimation des paramètres d^1 et a^1, et des corrélations de $\widetilde{f^1}$ avec les variables X^1.

	Estimations des paramètres		
Variables	d^2	a^2	Corrélations avec $\widetilde{f^2}$
evi_1	15.51	0.63	0.65
evi_2	13.47	0.59	0.6
evi_3	14.83	0.51	0.52
evi_4	14.67	0.58	0.6
evi_5	16.44	0.56	0.57
evi_6	18.74	0.51	0.52
evi_7	18.44	0.75	0.76
evi_8	20.59	0.8	0.82
evi_9	21.83	0.76	0.78
evi_10	19.19	0.74	0.76
evi_11	18.22	0.67	0.69
evi_12	15.92	0.61	0.63
evi_13	15.4	0.58	0.6
evi_14	13.51	0.7	0.72
evi_15	14.57	0.69	0.71
evi_16	14.95	0.76	0.78
evi_17	16.09	0.73	0.75
evi_18	15.95	0.77	0.79
evi_19	17.12	0.73	0.75
evi_20	15.02	0.75	0.77
evi_21	15.87	0.75	0.77
evi_22	14.21	0.71	0.73
evi_23	15.26	0.68	0.69

TABLEAU 3.5 – Application aux données *genus* sans covariables T : estimation des paramètres d^2 et a^2, et des corrélations de $\widetilde{f^2}$ avec les variables X^2.

	Estimations des paramètres		
Variables	d^2	a^2	Corrélations avec $\widetilde{f^2}$
evi_1	15.51	0.63	0.65
evi_2	13.47	0.59	0.6
evi_3	14.83	0.51	0.52
evi_4	14.67	0.58	0.6
evi_5	16.44	0.56	0.57
evi_6	18.74	0.5	0.52
evi_7	18.44	0.74	0.76
evi_8	20.59	0.8	0.82
evi_9	21.83	0.76	0.78
evi_10	19.19	0.74	0.76
evi_11	18.22	0.67	0.69
evi_12	15.92	0.61	0.63
evi_13	15.4	0.58	0.6
evi_14	13.51	0.7	0.72
evi_15	14.57	0.69	0.71
evi_16	14.95	0.76	0.78
evi_17	16.09	0.73	0.75
evi_18	15.95	0.77	0.79
evi_19	17.12	0.73	0.75
evi_20	15.02	0.75	0.77
evi_21	15.87	0.75	0.77
evi_22	14.21	0.71	0.73
evi_23	15.26	0.67	0.69

TABLEAU 3.6 – Application aux données *genus* sans covariables T : estimations des paramètres d^2 et a^2, et des corrélations de $\widetilde{f^2}$ avec les variables X^2.

3.4 Perspectives et discussion sur les questions du nombre de blocs, de facteurs, de parcimonie et d'unicité des solutions

L'application de l'approche EM sur les données *genus* suggère à la fois une manière de construire un modèle structurel à la lumière d'une ACP préliminaire et une façon d'évaluer sa qualité par un procédé de ré-échantillonnage. L'ACP a permis de distinguer deux groupes de VO (X^1 et X^2) pouvant expliquer le groupe de VO (Y). Les facteurs étant supposés indépendants, cette étape est très utile : les groupes de variables doivent être proches de l'orthogonalité. Cependant, comme cela a été discuté à la fin du chapitre 2, il pourrait être envisagé que le nombre de facteurs par groupe ne se limite pas à un seul. Il viendrait alors tout naturellement la question de comment déterminer le nombre de facteurs par groupe. En effet, la méthode d'estimation par algorithme EM, tout comme les méthodes PLS et LISREL, n'est appliquée qu'après avoir pré-établi un modèle structurel avec un nombre fixé de facteurs. En pratique, plusieurs experts des disciplines des domaines d'applications concernés échangent avec le statisticien jusqu'à converger vers une modélisation pertinente. Mais cela n'implique pas que cette modélisation soit la meilleure. Il serait intéressant de pouvoir évaluer objectivement la qualité du choix de modèle. Ainsi, des statistiques évaluant la qualité d'ajustement d'un modèle doivent être développées pour la méthode d'estimation par algorithme EM.
Concernant la généralisation du modèle à plusieurs facteurs par groupe, nous avons mentionné plus haut la nécessité de contraintes d'identification supplémentaires. Si de plus pour un groupe le nombre de facteurs choisi est trop élevé, la matrice des coefficients pondérateurs associée risque de ne pas être de plein rang. Il est alors recommandé d'ajouter des contraintes

sur la matrice des pondérations pour la contraindre à être de plein rang.
Pour éviter les problèmes de multiplicité des solutions, il faut aussi que le modèle soit parcimonieux. Quelque soit le nombre de facteurs communs par groupe, il faut que la différence entre le nombre d'équations du modèle et le nombre d'inconnues à estimer soit positive. Prenons l'exemple de l'équation de mesure :

$$Y = TD + GB + \varepsilon^Y \tag{3.4}$$

Si Ψ_Y est supposée diagonale alors $\Sigma^Y = B' \left(C'C + I_{K_G}\right) B + \Psi_Y$ contient $q_Y \times q_Y + q_Y = q_Y \left(q_Y + 1\right)$ éléments. Puisque Σ^Y est symétrique, elle est constituée de $\frac{1}{2} q_Y \left(q_Y + 1\right)$ éléments distincts. En revanche, le nombre de paramètres libres de l'équation (3.4) est $q_T \times q_Y + K_G \times q_Y + q_Y$ appartenant à D, B et Ψ_Y auquel est retranché le nombre de contraintes $\frac{1}{2} K_G \left(K_G - 1\right)$ nécessaires à l'identification des pondérations et des facteurs associés. En effet, $\frac{1}{2} K_G \left(K_G - 1\right)$ correspond au nombre paramètres fixés qui ne sont donc pas libres. Plus de détails sont disponibles dans Saidane (2006). Le nombre de paramètres libres de l'équation (3.4) est alors $q_Y \left(q_T + K_G + 1\right) - \frac{1}{2} K_G \left(K_G - 1\right)$. Pour obtenir une solution unique, la différence d_Y entre le nombre d'équations et le nombre d'inconnues (i.e. : paramètres libres) doit être positive. En effet,

- $d_Y < 0$ signifie qu'il y a plus de paramètres à estimer que d'équation et il y a donc une infinité de solutions ;
- $d_Y \geq 0$ signifie qu'il y a soit autant soit plus de paramètres à estimer que d'équations, il y a donc une solution.

Par conséquent, pour avoir une unique solution lors de la procédure d'estimation, il faut imposer un nombre de contraintes $d_Y \geq 0$ tel que :

$$d_Y = \frac{1}{2} q_Y \left(q_Y + 1\right) - \left\{ q_Y \left(q_T + K_G + 1\right) - \frac{1}{2} K_G \left(K_G - 1\right) \right\}.$$

Ainsi, dans le cas du modèle à un facteur par groupe étudié (cf. le modèle (3.2)), il faut $d_Y + d_1 + d_2$ contraintes telles que,

$$\left\{ \begin{array}{rcl} d_Y & \geq & 0 \\ d_1 & \geq & 0 \\ d_2 & \geq & 0 \end{array} \right.$$

ce qui équivaut à,

$$\left\{ \begin{array}{rcl} d_Y & \geq & \frac{1}{2} q_Y \left(q_Y - 3 - 2q_T\right) \\ d_1 & \geq & \frac{1}{2} q_1 \left(q_1 - 3 - 2q_T\right) \\ d_2 & \geq & \frac{1}{2} q_2 \left(q_2 - 3 - 2q_T\right) \end{array} \right.$$

Le modèle étudié est donc parcimonieux si à la fois $q_Y \geq 3$, $q_1 \geq 3$ et $q_2 \geq 3$. Par exemple, dans le cas du modèle étudié, si celui-ci était sans covariables (i.e. : $q_T = 0$), il serait donc nécessaire d'avoir un minimum de 3 VO par groupe pour que la procédure itérative de la méthode d'estimation par algorithme EM converge vers une unique solution.
Tout comme dans le cas du modèle à deux groupes (un dépendant et un explicatif) présenté au chapitre précédent, le problème d'identifiabilité lié à la reconstruction des facteurs persiste : les facteurs d'un groupe ne sont identifiables qu'à une transformation orthogonale arbitraire près.
Pour finir, des perspectives autres que le développement de statistiques d'ajustement de modèle ou d'indice de validation sont à envisager. Cette approche pourrait en effet être étendue aux modèles linéaires généralisés. Les variables ne seraient alors plus gaussiennes mais pourraient être catégorielles.

CHAPITRE 4

Analyse longitudinale de la qualité de vie sur des facteurs reconstruits

Sommaire

4.1 Introduction 90
4.2 Contexte 90
4.2.1 Essai clinique 90
4.2.2 Critères d'évaluation du bénéfice d'un traitement lors d'un essai clinique 91
4.2.3 La qualité de vie : un critère d'évaluation alternatif 92
4.3 Le critère de qualité de vie (QdV) relative à la santé (HRQoL) 92
4.3.1 Mesure de la HRQoL par auto-questionnaires 93
4.3.2 Évaluation de la HRQoL 93
4.3.3 Analyse longitudinale classique de la HRQoL 95
4.4 Une analyse longitudinale de la HRQoL en deux étapes 97
4.4.1 Introduction 97
4.4.2 Première étape : analyse transversale 97
4.4.3 Seconde étape : analyse longitudinale par modèle linéaire mixte . . . 99
4.4.4 Application à des données réelles issues de l'essai clinique CO-HO-RT 99
4.4.5 L'article soumis 100
4.5 Discussion, commentaires et perspectives 120

Le travail présenté dans ce chapitre a été réalisé conjointement avec Barbieri Antoine et Caroline Bascoul Mollevi. Les données utilisées sont issues de l'essai clinique CO-HO-RT (Azria et al., 2010).

4.1 Introduction

La méthode d'estimation par algorithme EM a été appliquée au domaine de la qualité de vie relative à la santé (HRQoL) en cancérologie, en collaboration avec des membres de l'Unité de Biostatistique à l'Institut régional du Cancer de Montpellier (ICM). En oncologie, la (HRQoL) est devenue un critère essentiel dans les essais cliniques mais son analyse longitudinale reste complexe. En effet, un des freins conceptuels de ce critère est son aspect multidimensionnel et la multiplicité des tests qu'il engendre. L'emploi de l'approche EM comme étape préliminaire à l'analyse longitudinale de ce critère a pour objectif de contribuer à la simplification de l'analyse. Cette contribution est actuellement soumise à publication. Avant de présenter l'article en question, ce chapitre va dans un premier temps situer le contexte et les motivations de ce travail. Il mettra également en place toutes les définitions des termes techniques du domaine oncologique nécessaires à la compréhension des développements présentés. Un accent particulier est porté sur la définition du critère de la HRQoL, son évaluation par auto-questionnaires et l'analyse longitudinale classique dont il fait l'objet. Ensuite, l'article est brièvement introduit par une courte description de l'analyse longitudinale utilisant l'approche EM ainsi que par la présentation de quelques résultats obtenus sur des données réelles. L'article soumis à publication suivra afin de présenter l'application complète et l'ensemble des résultats liés à l'analyse longitudinale proposée. Pour finir, ce chapitre ouvre le travail sur de possibles perspectives et une discussion.

4.2 Contexte

Le critère de HRQoL est souvent étudié afin de mettre en avant le bénéfice d'une nouvelle molécule dans le cadre de la conception d'un futur traitement. En fonction de la pathologie ciblée et de ses traitements actuels, un nouveau médicament peut trouver une place sur le marché lorsqu'il améliore par exemple la Qualité de Vie (QdV) des patients. Alors, son laboratoire de conception doit réaliser des essais sur l'Homme avec l'accord de l'ANSM (*Agence Nationale de Sécurité du Médicament et des produits de santé*). Ces essais s'effectuent selon une succession de 4 phases, ce que l'on appelle un "essai clinique". Le bénéfice clinique du traitement et la QdV sont étudiés lors de la troisième phase, mais il arrive aussi que la QdV soit étudiée lors de la deuxième phase.

4.2.1 Essai clinique

Avant de mettre sur le marché un nouveau traitement, le laboratoire pharmaceutique procède à un essai clinique. Suivant le traitement mis en place, ce dernier peut légèrement différer. La figure 4.1 illustre le processus d'un essai clinique dans le cadre d'un traitement anti-cancéreux. L'essai clinique est compartimenté : il y a une première partie pré-clinique où des tests sont effectués soit en laboratoire, soit sur des animaux (in vitro, in vivo), puis une partie liée au développement thérapeutique. Cette seconde partie est subdivisée en 3 phases : la phase I, la phase II et la phase III.
La phase I a pour objectif d'évaluer la tolérance à la molécule active du traitement et d'établir la dose recommandée à administrer. Pour y parvenir on cherche la dose maximale pouvant être tolérée. Dans le domaine de la cancérologie, cette phase I est réalisée sur une cohorte de 15 à 50 patients avec un cancer méta-statique, c'est à dire sur des patients qui ne sont pas en rémission ou pour lesquels aucun traitement n'existe.
La phase II vient alors évaluer l'efficacité de la molécule. Pour le cas du domaine de la cancérologie, cette efficacité est évaluée en terme d'activité anti-tumorale. On cherche à étudier :

— la diminution de la tumeur ;

— la tolérance à moyen terme ;

— l'allongement de la survie ;

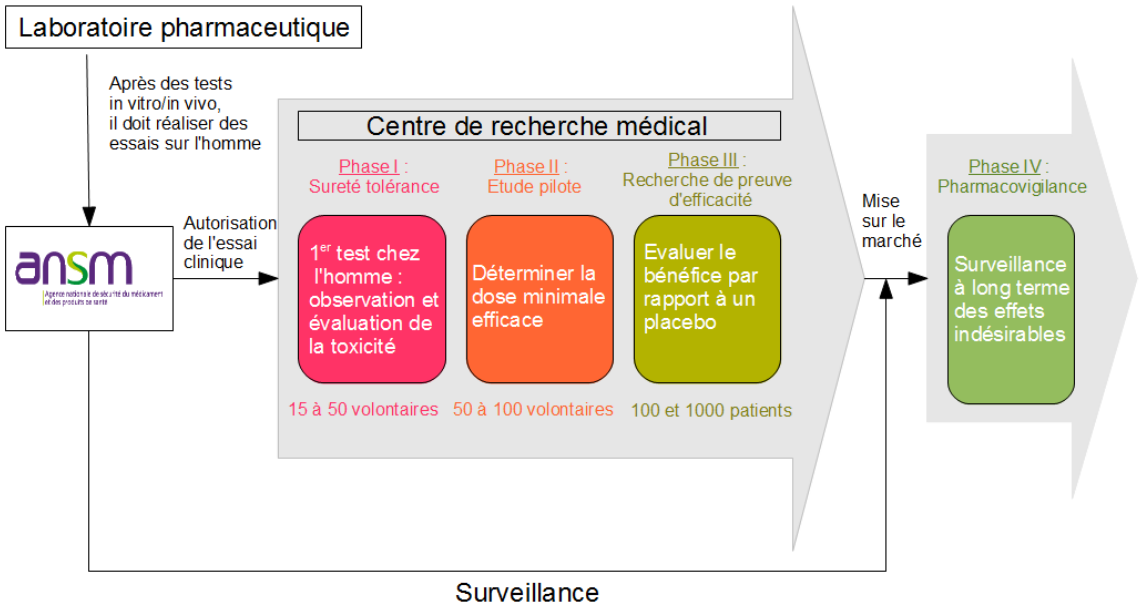


FIGURE 4.1 – Les différentes étapes d'un essai clinique pour un traitement anti-cancéreux.

— l'amélioration de la Qualité de Vie (QdV).

Cette phase est accomplie sur des patients en échecs thérapeutique dont le nombre est supérieur à celui de la phase I mais inférieur à 100 patients.
La phase III a quant à elle pour objectifs d'étudier l'efficacité du traitement en situation réelle, et aussi d'évaluer son bénéfice clinique. La cohorte de patients est alors choisie de manière à être plus représentative de la population de patients susceptible de bénéficier du traitement. Lors de cette phase, le nombre de patients est en général compris entre 100 et 1000. Enfin, pour évaluer le bénéfice clinique, les patients sont séparés en deux groupes : un groupe de référence $G0$ et un groupe $G1$. On procède ensuite à deux types de comparaisons :

— soit on compare les résultats du traitement testé sur $G1$ à ceux du traitement standard (actuellement prescrit) administré à $G0$;

— soit on compare les résultats du traitement testé sur $G1$ à ceux avec absence de traitement (un placebo est administré à $G0$).

Pour que les résultats soient en faveur de la nouvelle molécule, il faut obtenir soit une prolongation de la survie, soit une amélioration de la QdV.

4.2.2 Critères d'évaluation du bénéfice d'un traitement lors d'un essai clinique

Pour évaluer le bénéfice clinique d'un nouveau traitement, le choix du critère de jugement est essentiel. Il existe plusieurs types de critères de jugement dont voici une liste non exhaustive :

— la survie globale ;

— la QdV ;

— la survie sans progression ;

— la maladie sans progression.

Bien que les deux derniers critères relèvent de mesures quantitatives objectives car centrées sur la tumeur, le critère de survie globale est considéré en cancérologie comme le critère de jugement de référence en phase III. Il est objectif, universel et clairement défini par l'intervalle de temps jusqu'au décès, toutes causes confondues. Cependant, aujourd'hui ce critère rencontre des limites. Les traitements sont de plus en plus efficaces et allongent considérablement la durée de vie. Le cas du cancer du sein est un exemple. Après une chirurgie, il est de plus en plus difficile d'obtenir des différences significatives avec le critère de survie globale lors d'une durée d'étude fixe. Or, la durée d'étude du bénéfice d'un traitement en phase III doit rester raisonnable. Pour pallier l'allongement des durées d'études, des critères de jugement alternatifs peuvent être utilisés.

4.2.3 La qualité de vie : un critère d'évaluation alternatif

Les critères de jugement alternatifs sont objectifs et centrés sur la tumeur. Ils relèvent de mesures quantitatives. Mais avant d'utiliser un critère de jugement alternatif en tant que critère de jugement principal, il faut soit valider ce critère comme critère de jugement substitutif à la survie globale, soit s'assurer du bénéfice clinique pour le patient. Pour cela, il devient essentiel de joindre à l'utilisation de ce critère alternatif, un critère secondaire centré sur le patient. La HRQoL est un exemple de critère de jugement secondaire. Ainsi, dans le contexte de la cancérologie, la HRQoL est un critère de jugement à la fois alternatif et secondaire utile pour s'assurer du bénéfice d'un traitement. Actuellement, lors d'un test clinique, en l'absence d'effet sur la survie globale, la HRQoL est considérée comme second critère de jugement principal par l'Amercan Society of Clinical Oncology (ASCO) et la Food and Drug Administration (FDA) (Beitz et al., 1995). Il arrive même, sous certaines conditions, qu'elle soit citée comme critère principal lors d'un essai clinique (Fiteni et al., 2015). Les situations gériatriques et palliatives étant de plus en plus fréquentes, ce critère a aussi l'avantage d'être pertinent pour assurer confort et bien-être aux patients. Cependant, l'utilisation de ce critère se heurte à plusieurs inconvénients. D'abord, au niveau conceptuel, l'évaluation de le HRQoL est réalisée par le biais d'auto-questionnaires complétés par les patients. Elle est donc subjective mais aussi dynamique car la HRQoL peut varier tout au long du traitement pour un même patient. Par conséquent, il faudrait s'interroger sur le sens clinique d'un changement de niveau de HRQoL. De plus la définition de la HRQoL n'est pas clairement établie bien qu'elle soit présentée comme sous-jacentes à plusieurs dimensions à la fois fonctionnelles et symptomatiques telles que la douleur et la fatigue (cf. section 2). Par ailleurs, au niveau méthodologique, on peut rencontrer des problèmes de données manquantes ou liés à la nature multidimensionnelle de la HRQoL. En effet, actuellement la qualité de vie est étudiée dimension par dimension, ce qui engendre une multiplicité des tests. L'ensemble de ces inconvénients constituent des limites à l'évaluation de la HRQoL mais aussi des limites à l'utilisation des résultats qu'elle engendre, et donc à l'aide à la décision quant à l'évaluation du bénéfice thérapeutique. Par conséquent, les cliniciens oncologues restent réticents au changement de leurs pratiques et à se fonder uniquement sur les résultats du critère de HRQoL en phase III pour évaluer le bénéfice clinique d'un nouveau traitement. D'autant plus que les analyses longitudinales réalisées sur les données de la HRQoL et leur interprétation restent complexes.

4.3 Le critère de qualité de vie (QdV) relative à la santé (HRQoL)

Le concept de QdV est défini par l'Organisation Mondiale de la Santé (OMS) comme "un état de bien-être physique, mental et social complet, et pas seulement l'absence de maladie ou d'infirmité" (WHO, 1948). Le critère de QdV est donc par définition :

— subjectif : la notion de bien-être est propre à chacun ;

— dynamique : pour un même individu la notion de "bien-être" peut varier au cours du temps ;

— multidimensionnel, regroupant au moins trois types de familles de dimensions : certaines liées au fonctionnement physique, d'autres d'ordre psychologique et les dernières d'ordre sociale.

Le caractère multidimensionnel de la HRQoL

Lorsque la QdV est étudiée du point de vu médical, sa définition est orientée de manière à ce que le critère soit spécifique au domaine de la santé. Des dimensions informatives sur l'impact direct ou indirect d'une maladie donnée ou d'un traitement donné sont aussi considérées. Les dimensions "difficultés financières", "constipation", "symptômes au bras" sont des exemples. On parle de qualité de vie relative à la santé (HRQoL).

Le caractère subjectif de la HRQoL

La variabilité inter-individuelle de ce critère est incontestable. Elle peut dépendre de la culture de l'individu, de son vécu, de ses attentes, de ses valeurs sociales, de son environnement, etc.

Le caractère dynamique de la HRQoL

Ce critère est aussi caractérisé par une variabilité intra-individuelle. Après un intervalle de temps suffisant un même individu peut évoluer et évaluer son niveau de QdV de façon différente même s'il n'a pas objectivement changé. En effet, l'état moral, l'intensité des symptômes subis et les expériences de vie durant les dernières semaines peuvent influencer la sensibilité de l'individu et son jugement vis-à-vis de sa QdV.

4.3.1 Mesure de la HRQoL par auto-questionnaires

Les données de QdV sont obtenues à partir de questionnaires (cf. les annexes C et D) remplis par les patients eux mêmes à différentes visites durant leur traitement. L'outil de mesure de la QdV est donc un auto-questionnaire. Il peut comporter plusieurs questions pour une même dimension du concept. Techniquement, on dit qu'une dimension est "indirectement mesurée par plusieurs items". Ces items sont élaborés de manière à être clairement compréhensibles et à limiter la variabilité d'interprétation. Chaque item comporte un ensemble de réponses binaires ou polytomiques parmi lesquelles le patient doit cocher. Pour le cas des réponses à choix polytomique, les catégories de réponses sont ordonnées d'un extrême à l'autre. Le nombre de catégories n'excède jamais 7 et peut donc être pair ou impair. La construction est réalisée de manière à ce que lorsque l'on souhaite forcer le sujet à prendre position, le nombre de modalités que propose l'item est pair. Lorsqu'il est impair, la réponse médiane est alors neutre. Ces différentes catégories de réponses sont associées à des valeurs numériques entières à partir desquelles un score est établi pour chacune des dimensions du questionnaire.
Plusieurs questionnaires ont été construits, mais dans le domaine de l'oncologie, seuls deux types d'entre eux se distinguent. Pour l'Europe, il existe ceux développés par l'EORTC (European Organization of Research ant Treatment of Cancer) et pour l'Amérique, ceux du groupe FACT.

4.3.2 Évaluation de la HRQoL

En Europe, le questionnaire de référence dans le domaine de la cancérologie est l'EORTC QLQ-C30 (nommé simplement QLQ-C30 pour la suite de la rédaction). Il en existe plusieurs versions. Celle qui est recommandée par le groupe EORTC aujourd'hui est la 3.0 (cf. Annexe C). Elle a été vérifiée et validée au niveau psychométrique dans 81 langues afin d'accroître son utilisation et de permettre une comparaison des résultats issus de différentes études. Ses items et dimensions sont respectivement au nombre de 30 et 15. Ses dimensions sont organisées en

une dimension *statut global de santé* et deux familles de dimensions : l'une liée aux symptômes et l'autre aux fonctions physiques. Sa structure est illustrée figure 4.2.

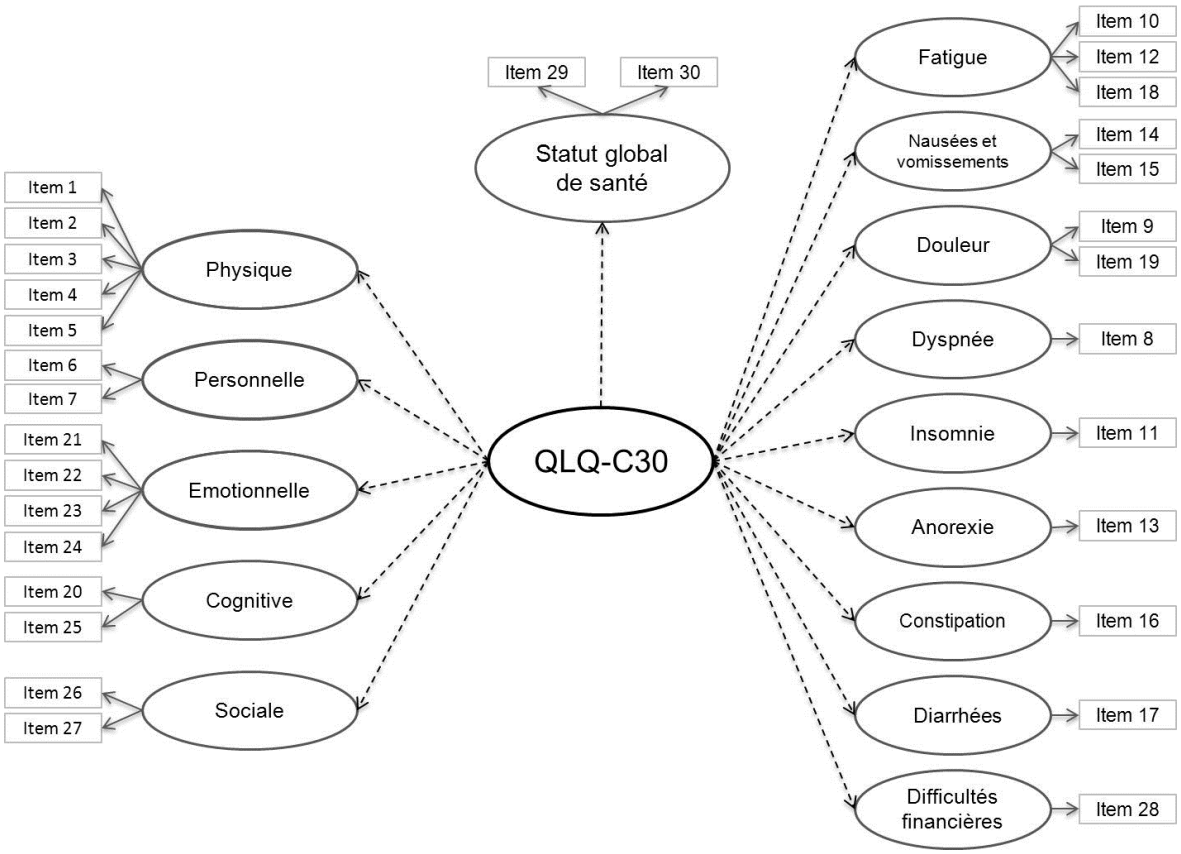


FIGURE 4.2 – Structure du questionnaire EORTC QLQ-C30.

Les différentes dimensions fonctionnelles sont les suivantes : "physique", "personnelle" (concernant les activités quotidiennes, professionnelles ou de loisir), "émotionnelle", "cognitive" (liée à la concentration et la mémoire) et "sociale". Parmi elles, celle qui compte le plus d'items est la dimension physique. Le nombre d'items a été choisi lors de la construction du questionnaire en fonction de la complexité de la notion voulant être mesurée et du degré de précision attendu.
La famille de dimensions symptomatiques comporte les suivantes : "fatigue", "nausée et vomissements", "douleur", "dyspnée" (difficulté respiratoire, essoufflement), "insomnie", "anorexie", "constipation", "diarrhées", "difficultés financières".
Les items 1 à 28 comportent 4 modalités de réponses, telles que : 1 pour "pas du tout", 2 pour "un peu", 3 pour "assez" et 4 pour "beaucoup". Les items 29 et 30 dont est composée la dimension *statut global de santé*, sont constitués de 7 modalités de réponses allant de 1 pour "très mauvaise" jusqu'à 7 pour "excellente".
Suivant les cancers, le questionnaire QLQ-C30 dit "générique" est complété par un questionnaire supplémentaire. Pour le cancer du sein, le QLQ-C30 est complété par le questionnaire spécifique EORTC QLQ-BR23 (cf. Annexe D). Il permet l'évaluation de 8 dimensions supplémentaires que sont d'une part : "image corporelle", "fonctionnement sexuel", "plaisir sexuel" et "perspectives futures", pour la famille de dimensions fonctionnelles ; et d'autre part : "symptômes liés au traitement", "symptômes liés au aux seins", "symptômes liés au bras" et "l'inquiétude liée à la perte de cheveux", pour la famille de dimensions symptomatiques.
Pour permettre l'étude de la QdV, les réponses aux questionnaires sont récoltées et les données obtenues sont pré-traitées. Tout d'abord, le pré-traitement a lieu à l'échelle des dimensions. À partir des réponses aux items de chacune d'entre elles, un score est obtenu. Il est calculé par la somme des valeurs des modalités cochées par le patient pour la dimension en question.

Pour chacun de ces groupements de J items, on parle "score-somme" S calculé comme suit :

$$S = \sum_{j=1}^{J} Y^{(j)} \tag{4.1}$$

où $Y^{(j)}$ est la réponse à l'item $j \in [\![1, J]\!]$. Pour permettre la comparaison de score-somme provenant de questionnaires différents, celui-ci est standardisé de manière à ce que les différentes valeurs qu'il puisse prendre soient comprises entre 0 et 100.
Bien sûr les problèmes classiques de données manquantes peuvent être rencontrés, mais ils ne seront pas abordés ici. Dans le cadre de ce type de données de QdV récoltées au cours du temps de suivi des patients, un phénomène dit de "response shift" peut avoir lieu. Il traduit le processus d'adaptation du patient à la toxicité des traitements mais aussi son acceptation de la maladie. La conséquence est qu'un biais peut apparaître lors de l'analyse. Mais de manière similaire, ce problème ne sera pas développé. D'ailleurs, il ne sera pas non plus pris en compte dans la modélisation proposée par la suite lors de l'analyse de la QdV.

4.3.3 Analyse longitudinale classique de la HRQoL

En cancérologie, actuellement l'analyse longitudinale de la HRQoL est réalisée sur des données brutes pré-traitées. Pour un individu et une dimension fixés, les données catégorielles récoltées sont transformées en une valeur numérique allant de 0 et 100, nommée score de HRQoL. Cette technique suppose que le score est une bonne représentation de la dimension latente à un erreur de mesure ε près. Elle se nomme CTT (Classical Test Theory) ou l'"approche du vrai score" et est fondée sur l'étude des scores S obtenus par l'intermédiaire des réponses aux items. Le modèle linaire mixte étant l'un des outils les plus classiques utilisés, il sera présenté.

Pré-traitement des données par procédure de scoring

Selon les recommandations de l'EORTC, en Europe, lors des essais cliniques des scores-sommes standardisés S sont calculés pour chaque dimension des auto-questionnaires. La formule utilisée dépend de la famille à laquelle appartient la dimension. Lorsque sa famille est symptomatique, les scores-sommes standardisés sont calculés par la formule :

$$S = \frac{\left(\frac{1}{J}\sum_{j=1}^{J} Y^{(j)}\right) - 1}{M - 1} \times 100 \tag{4.2}$$

où J est le nombre d'items à M modalités de réponses pour une dimension donnée et $Y^{(j)}$ la j-ème réponse cochée par le patient associée au j-ème item de la dimension en question.
En revanche, pour une dimension appartenant à la famille fonctionnelle, la formule de calcul des scores-sommes est :

$$S = \left[1 - \frac{\left(\frac{1}{J}\sum_{j=1}^{J} Y^{(j)}\right) - 1}{M - 1}\right] \times 100 \tag{4.3}$$

Cette dernière formule permet de "symétriser" les échelles de réponses des dimensions fonctionnelles vis-à-vis des symptomatiques. Le but est qu'un score élevé (resp. faible) corresponde à une bonne (resp. faible) capacité fonctionnelle (physique, sociale...) mais aussi à un faible (resp. haut) niveau de symptômes (fatigue, douleur...) subis. Ainsi, un score élevé reflète un bon niveau de QdV et vice et versa. Cette méthode de pré-traitement des données brutes des auto-questionnaires s'appelle procédure de scoring.
La réalisation de cette dernière sur des données brutes de QdV permet de donner à la nature catégorielle des dimensions une allure de variable quantitative continue. Les questionnaires

étant complétés à différents temps de suivi pour un même patient, les jeux de données obtenus donnent accès à des données répétées permettant une analyse longitudinale de la HRQoL.

L'outil Modèle linéaire mixte (L2M)

Actuellement, les modèles linéaires mixtes (Linear Mixed Models, LMM ou L2M) font partie des outils les plus classiquement utilisés pour l'analyse longitudinale de la HRQoL. Ils permettent deux niveaux de lecture des données répétées : un niveau global commun à tous les individus et un niveau individuel. Ces niveaux se traduisent respectivement par les effets fixes et les effets aléatoires. Ainsi, les paramètres des effets fixes sont communs à l'ensemble des individus et ceux des effets aléatoires varient d'un individu à l'autre. La puissance de cet outil est dans la séparation de la variance totale en deux parties : la variance due aux effets aléatoires et celle affectée aux erreurs de mesure. De manière générale, ce type de modèle se formalise par :

$$Y = X\beta + U\xi + \varepsilon \tag{4.4}$$

où,

- Y est le vecteur aléatoire de longueur n à expliquer ;
- $X\beta$ est la partie fixe du modèle avec X la matrice d'incidence connue de dimension $n \times p$ et β le vecteur de longueur p des paramètres inconnus des effets fixes ;
- $U\xi$ est la partie aléatoire telle que U est la matrice d'incidence connue de dimension $n \times q$ et ξ est le vecteur de longueur q des effets aléatoires. Ce vecteur se décompose en L effets aléatoires considérés dans le modèle tel que $\xi' = \left({\xi_1}', \ldots, {\xi_\ell}', \ldots, {\xi_L}'\right)$. Pour tout $\ell \in [\![1, L]\!]$, l'effet aléatoire ξ_ℓ est de longueur q_ℓ et on obtient $q = \sum_{\ell=1}^{L} q_\ell$. La matrice U est quant à elle pour tout $\ell \in [\![1, L]\!]$, la concaténation de matrices d'incidence sous-jacentes U_ℓ de dimension $n \times q_\ell$ tel que $U = \left[U_1, \ldots, U_\ell, \ldots, U_L\right]$.
- ε est le vecteur aléatoire de longueur n.

Les hypothèses de ce modèle sont :

- les effets aléatoires ξ_ℓ suivent une distribution gaussienne centrée de matrice de variance-covariance $\sigma_\ell^2 \Gamma_\ell$ où Γ_ℓ est une matrice connue et de dimension $q_\ell \times q_\ell$;
- les effets aléatoires sont deux à deux indépendant et ξ suit une gaussienne centrée de matrice de variance-covariance $\Gamma = diag\left\{\sigma_\ell^2 \Gamma_\ell\right\}$;
- pour i une observation, $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_{\varepsilon,i}^2\right)$;
- ε et les effets aléatoires ξ_ℓ sont mutuellement indépendants.

Pour i un individu et v le numéro de la visite au cours du temps de suivi, le modèle linéaire mixte pour l'étude longitudinale d'une dimension de HRQoL représentée par son score S s'écrit :

$$\begin{cases} S_{iv} &= \alpha_{iv} + \varepsilon_{iv} \\ \alpha_{iv} &= X_{iv}\beta + U_{iv}\xi_i \end{cases} \tag{4.5}$$

où, α_{iv} est la dimension latente de HRQoL associée à son score S_{iv} pour l'individu i et la visite v et ξ_i sont les effets fixes spécifiques aux individus.
Au delà de la décomposition de la variance totale du modèle en deux parties, l'avantage de cette modélisation est de permettre d'étudier l'évolution de la variable réponse S au cours du temps (i.e : les visites v), et de la comparer entre deux groupes (par exemple un groupe sous le traitement testé et un groupe sous placebo). La partie fixe du modèle décrit une tendance moyenne de l'évolution au cours du temps alors que la partie aléatoire représente les tendances spécifiques à chaque individu i. Plus particulièrement à l'analyse longitudinale de la HRQoL, une modélisation du score S où deux groupes G0 et G1 sont différenciés par une fonction indicatrice peut être la suivante :

$$\left\{ \begin{array}{lcl} S_{iv} & = & \alpha_{iv} + \varepsilon_{iv} \\ \alpha_{iv} & = & \beta_0 + \xi_{i,0} + \beta_1 \mathbb{1}_{i \in G0} + t(\beta_2 + \beta_3 \mathbb{1}_{i \in G0} + \xi_{i,1} \end{array} \right. \tag{4.6}$$

où,

- t est le temps en abscisse dont les valeurs disponibles sont associées aux visites v ;
- β_0(resp. $\beta_0 + \beta_1$) est l'intercept du groupe de référence G0 (resp. du groupe G1) ;
- β_2 (resp. $\beta_2 + \beta_3$) est la pente du groupe de référence $G0$ (resp. du groupe G1) ;
- pour chaque individu i, ce modèle distingue deux effets aléatoires $\xi_{i,0}$ et $\xi_{i,1}$ tels que $\xi_i = (\xi_{i,0}, \xi_{i,1})'$ afin de décrire les tendances spécifiques à chaque individu i par deux paramètres respectivement associés à l'ordonnée à l'origine et la pente.

Ainsi, $\xi_{i,0}$ permet de tenir compte du fait que les individus n'ont pas le même niveau de HRQoL au début de l'étude ($t = 0$) alors que $\xi_{i,1}$ représente les évolutions de la HRQoL au cours du temps de chaque individu. Les évolutions peuvent en effet être différentes suivant les individus ou les groupes d'individus. Les évolutions de la HRQoL sont supposées linaires par ce modèle (4.6) mais le caractère dynamique de la HRQoL n'est pas nécessairement linéaire. Des modélisations polynomiales ou linéaires par morceaux peuvent être envisagées. En terme d'interprétation clinique, une modélisation linéaire par morceaux présenterait l'avantage de mettre en évidence des périodes de changement d'évolution de la HRQoL.
Parmi les limites de la méthode d'analyse longitudinale de la HRQoL présentée il y a la modélisation du score par dimension. On aboutit alors à 15 modélisations par L2M et 15 analyses longitudinales. Le nombre de tests en est alors d'autant multiplié en pratique. Bien qu'il arrive que des connaissances a priori sur le traitement testé permettent de se focaliser sur une dimension ou un groupe de dimensions lors de l'analyse principale de la HRQoL, une approche d'étude globale de la HRQoL est proposée dans la section suivante pour pallier l'inconvénient de multiplicité des tests.

4.4 Une analyse longitudinale de la HRQoL en deux étapes

4.4.1 Introduction

L'objectif étant de tenir compte de l'aspect multidimensionnel et longitudinale de la HRQoL, l'approche présentée combine à la fois une modélisation par équation structurelle à facteurs et un modèle linéaire mixte. Au vu de la structure du questionnaire, dans une première étape la méthode estime à chaque visite un même modèle à une équation structurelle à facteurs avec un facteur résumant la famille de dimensions fonctionnelles et un second résumant la famille de dimensions symptomatiques. Puis dans une seconde étape, l'analyse longitudinale de la HRQoL par L2M est réalisée, où la dimension *statut global de santé* est choisie comme variable réponse représentative de la HRQoL globale des patients. Les facteurs reconstruits pour chacune des visites lors de la première étape sont concaténés et injectés dans le L2M, ce qui contribue à la globalité de l'analyse longitudinale de la HRQoL. Le modèle linéaire mixte prendra en compte l'effet individu (effet aléatoire) et l'effet traitement (effet fixe). Cette approche est possible par la maximisation de la vraisemblance de chaque modèle structurel par l'approche EM proposée dans ce travail de thèse. Elle donne une estimation des facteurs de manière efficace et permet ainsi de les introduire dans le modèle linéaire mixte de la seconde étape. Cette technique en 2 étapes est développée sous le logiciel R et illustrée sur des données d'un essai clinique en cancérologie. Bien que cette approche soit détaillée dans l'article soumis et inséré à la fin de la section, elle est brièvement introduite.

4.4.2 Première étape : analyse transversale

Le modèle structurel proposé est issu de la décomposition du questionnaire QLQ-C30 qui distingue un groupe de dimensions fonctionnelles, un groupe de dimensions symptomatiques

et une dimension *statut global de santé*. La structure de ce questionnaire suggère une modélisation à deux facteurs f^1 et f^2 qui quantifient et résument respectivement les statuts fonctionnels et symptomatiques pour tous les individus. f^1 et f^2 sont respectivement liés au groupe de variables observables fonctionnelles X^1 et symptomatiques X^2. Ces liaisons sont formalisées par des équations dites de mesures (cf. (4.7a), (4.7b)). Les facteurs f^1 et f^2 sont aussi liés à la variable *statut global de santé*, notée y. Elle est expliquée par les facteurs, ce qui se traduit par des liaisons formalisées par une équation structurelle (cf.(4.7c)). La concaténation des équations de mesures et de l'équation structurelle forme un système d'équations qui aboutit au modèle structurel construit à chaque temps de suivi (cf. (4.7)). Chacune de ces relations de dépendance peut être enrichie par une dépendance supplémentaire aux covariables T, T^1 et T^2 telles que le traitement par exemple. En outre, l'ensemble des variables observables sont quantitatives. Les données résultant des questionnaires subissent un pré-traitement selon la procédure de scoring (Fayers et al., 2001).

Un modèle multi-blocs à facteurs à une équation structurelle

Les données sont organisées sous forme de groupes de variables observables décrivant les mêmes n individus :
$X^m = \{x_i^{j,m}\}\,;\, i \in [\![1, n]\!],\, j \in [\![1, q_m]\!],\, m \in [\![1, 2]\!]$ la m-ième matrice explicative de dimension $n \times q_m$ constituée des variables $x^{1,m}, \ldots, x^{q_m,m}$. La valeur de la variable $x^{j,m}$ pour l'individu i est notée $x_i^{j,m}$.
$y = \{y_i\}\,;\, i \in [\![1, n]\!]$, est le vecteur de longueur n représentant la variable observable dépendante des facteurs f^1 et f^2.
T (resp. T^1, T^2) de dimension $n \times r_T$ (resp. $n \times r_1$, $n \times r_2$) sont les matrices de covariables.
Soient d (resp. D^m) un vecteur $r_T \times 1$ (resp. une matrice $r_m \times q_m$) de coefficients pondérateurs, a^m un vecteur $q_m \times 1$ de coefficients pondérateurs, et ε^y (resp. ε^m) un vecteur des erreurs $n \times 1$ (resp. une matrice des erreurs $n \times q_m$) associées à y (resp. X^m). Le modèle peut alors être formulé ainsi :

$$\begin{cases} X^1 = T^1 D^1 + f^1 a^{1\prime} + \varepsilon^1 & \text{(4.7a)} \\ X^2 = T^2 D^2 + f^2 a^{2\prime} + \varepsilon^2 & \text{(4.7b)} \\ y \;\; = Td + f^1 c^1 + f^2 c^2 + \varepsilon^y & \text{(4.7c)} \end{cases}$$

où les éléments de la première colonne des matrices de covariables T et T^m sont fixées à 1. Ainsi, d et la première ligne de chaque matrice D^m correspondent aux paramètres de moyennes. On y adjoint sous contraintes d'identifiablité, les hypothèses suivantes :
Les n observations sont indépendantes ; $\varepsilon_i^y \sim \mathcal{N}(0,\, \sigma_y^2)$; $\forall m \in [\![1, 2]\!] : \varepsilon_i^m \sim \mathcal{N}(0,\, \psi_m)$, où $\psi_m = diag(\sigma_m^2)$; ε^y et ε^m sont indépendants $\forall m \in [\![1, 2]\!]$; $\forall m \in [\![1, 2]\!] : f^m \sim \mathcal{N}(0, I_n)$ avec f^1, f^2 indépendants et ε^y, ε^m, f^m, $\forall m \in [\![1, 2]\!]$ sont mutuellement indépendants. Nous faisons aussi les hypothèses que $\forall m \in [\![1, 2]\!]$ les variables observées X^m (par exemple $x_j^m, j \in [\![1, q_p]\!]$) dépendent linéairement du facteur f^m et des covariables T^m, conditionnellement auxquelles ils sont indépendants ; y dépend linéairement des facteurs f^1 et f^2 et de la covariable T.
Ce modèle à une équation structurelle est établi pour chaque temps de suivi et pour i une observation, le modèle peut être formulé selon le système d'équations suivant :

$$\begin{cases} x_i^{1\prime} &= t_i^{1\prime} D^1 + f_i^1 a^{1\prime} + {\varepsilon_i}^{1\prime} \\ x_i^{2\prime} &= t_i^{2\prime} D^2 + f_i^2 a^{2\prime} + {\varepsilon_i}^{2\prime} \\ y_i' &= t_i' d + f_i^1 c^1 + f_i^2 c^2 + {\varepsilon_i}^{y\prime} \end{cases} \tag{4.8}$$

La figure 4.3 représente son diagramme.

Estimation du modèle et des facteurs à chaque visite par l'approche EM

La méthode par algorithme EM présentée au chapitre précédent est utilisée pour estimer les facteurs $\widetilde{f_i^m}$ pour chaque individu i en plus des paramètres θ du modèle, où

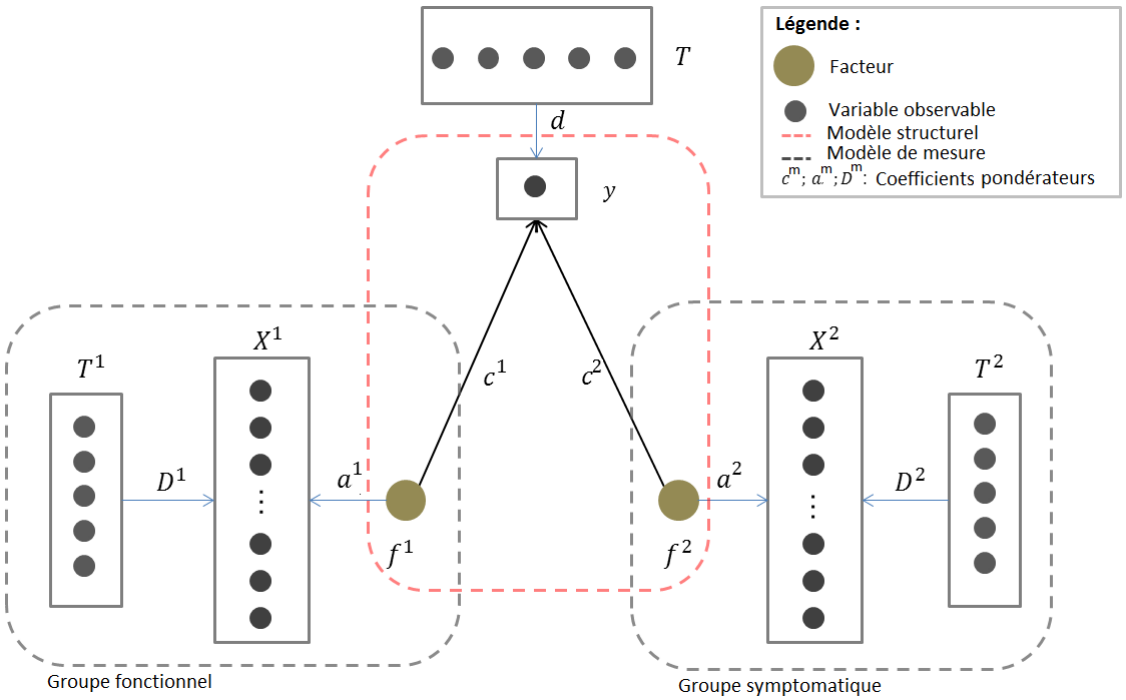


FIGURE 4.3 – Diagramme du modèle structurel induit par la décomposition du QLQ-C30 à chaque temps de suivi.

$\theta = \{d, D^1, D^2, a^1, a^2, c^1, c^2, \sigma_y^2, \sigma_1^2, \sigma_2^2\}$.
Dans le cadre de l'analyse transversale, le modèle (4.7) est établi à chaque visite v. Puis par la méthode d'estimation fondée sur EM, pour chacune des visites v, les estimations des facteurs $\widetilde{f}_v^1$ et $\widetilde{f}_v^2$ en plus des estimations des paramètres sont obtenues. Alors, la concaténation de la variable *statut global de santé* et des estimations des deux facteurs sur l'ensemble des temps de suivi permet de passer à l'étude longitudinale des données. Ainsi, au lieu d'analyser la HRQoL dimension par dimension de manière indépendante, on injecte les deux facteurs estimés dans le modèle linéaire mixte. L'analyse longitudinale devient alors globale.

4.4.3 Seconde étape : analyse longitudinale par modèle linéaire mixte

L'analyse longitudinale est réalisée via un modèle linéaire mixte. Le but est d'étudier l'évolution au cours du temps de suivi de la HRQoL représentée par les mesures répétées y_{iv} de la variable *statut global de santé* lors de la visite v pour un individu i. Celles-ci sont décrites par un modèle linéaire mixte standard formalisé comme suit :

$$y_{iv} \quad = \quad \mu + \boldsymbol{x}_{iv}'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\xi}_i + \varepsilon_{iv}, \tag{4.9}$$

où, μ est l'intercept ; β le vecteur des effets fixes ; $\boldsymbol{x}_{iv}$ le vecteur "design" contenant les estimations des facteurs $\widetilde{f}_v^m$ obtenues par la procédure précédente et d'autres covariables ; ξ_i le vecteur des effets aléatoires individuels tel que $\xi \sim \mathcal{N}(\mathbf{0}, \Gamma)$ où Γ est la matrice de covariance, $\boldsymbol{u}_i$ le vecteur "design" et $\varepsilon_{iv} \sim \mathcal{N}\left(0, \sigma_{\varepsilon,i}^2\right)$ sont les termes d'erreurs.
Les deux principaux avantages de cet outil sont de prendre en compte la variabilité induite par les données répétées dans le temps pour un même patient i et de quantifier la part d'information apportée par les variables explicatives.

4.4.4 Application à des données réelles issues de l'essai clinique CO-HO-RT

Une application sur des données réelles de QdV issues de l'essai clinique CO-HO-RT (Azria et al., 2010) est présentée dans l'article et complétée par son analyse longitudinale. Le

contexte est celui d'une étude de phase 2 randomisée évaluant les toxicités cutanées d'un traitement par radiothérapie-létrozole concomitant ou radiothérapie suivie par létrozole en situation adjuvante (postopératoire) de cancer du sein. Le nombre d'observations par visites (questionnaires entièrement remplis) pour un total de 121 patientes restantes est variable au cours du temps de suivi. Le tableau suivant récapitule le nombre n_v d'observations i disponibles en fonction des 8 visites v.

Visite v (mois)	0	3	6	12	15	18	21	24
n_v	113	106	102	100	102	84	91	90

Dans le cadre de ces données, l'effet du type de traitement est étudié sur l'évolution de la HRQoL au cours du temps de suivi des patientes. Pour ce faire, dans un premier temps des modèles structurels avec covariables sont établis puis estimés. Ensuite à chaque visite les facteurs estimés $\widetilde{f}_v^1$ sont comparés entre eux puis $\widetilde{f}_v^2$ entre eux. Les résultats obtenus sont qu'aucun des deux types de traitements reçus par les patientes n'a d'effet significatif sur leur QdV.
Lors de la seconde étape, une procédure de sélection de modèle au sens du BIC a aussi été réalisée. La procédure a retenu le modèle suivant :

$$y_{iv} \quad = \quad \alpha + \beta_1 \widetilde{f_{iv}^1} + \beta_2 \widetilde{f_{iv}^2} + \xi_i + \varepsilon_{iv}$$

À partir de ce modèle un intérêt a été porté à la part d'information contenue par les différents éléments le composant. Le résultat est que le facteur fonctionnel $\widetilde{f^1}$ explique deux fois plus le *statut global de santé* que le facteur symptomatique $\widetilde{f^2}$. Mais cela est à pondérer par la forte corrélation existante entre les deux facteurs : la présence des deux facteurs dans le modèle a du sens.

4.4.5 L'article soumis

EM algorithm estimation of a structural equation model for the longitudinal study of the quality of life

Antoine Barbieri[1,2], Myriam Tami[1], Xavier Bry[1], David Azria[2], Sophie Gourgou[2], Caroline Bascoul-Mollevi[2] and Christian Lavergne[1,3]

Abstract :
The health-related quality of life data is measured through self-questionnaires filled up at different times. We focused on the oncology data reported through the EORTC questionnaires which decompose the health-related quality of life into several functioning dimensions, several symptomatic dimensions and the Global Health Status. The aim is to explain the latter, which represents the most general concept, through the other dimensions. First, a similar structural equation model is used at each time, in which the global health status is explained by two latent variables. Each latent variable is a factor which summarizes respectively the functional dimensions and the symptomatic dimensions. This is achieved through the maximization of the likelihood of each structural equation model using the EM algorithm, with the advantage to give an estimation of the subject-specific factors. Then, to consider the longitudinal aspect, the global health status variable and the two factors are concatenated for each visit. The global health status can be then explained by the two factors estimated in the first step and additional explanatory variables using a linear mixed model. This model takes into account the inter-subject variability via specific-subject random effects and other covariates such as the treatment.

Key words : EM algorithm, Structural equation modeling, Longitudinal analysis, Mixed models, Health-related quality of life, Oncology data

[1] *Institut Montpelliérain Alexander Grothendieck, University of Montpellier, Montpellier, France*
[2] *Unity of biometry, Institut régional du Cancer Montpellier - Val d'Aurelle, Montpellier, France*
[3] *University Paul-Valéry Montpellier 3, Montpellier, France*

Corresponding author :
Antoine Barbieri or Myriam Tami
Institut montpelliérain Alexander Grothendieck, University of Montpellier, Place Eugène bataillon, 34098 Montpellier, France
E-mail : `Antoine.Barbieri@umontpellier.fr` or `Myriam.Tami@umontpellier.fr`

1 Introduction

The study of the quality of life concerns a lot of research domains such as social and medical sciences. The subjective, multidimensional and dynamic features of this concept makes it difficult to measure and analyze. Nowadays, the measure of the concept has become relatively standard through the use of specific psychometric questionnaires which take all the features into account. Indeed, this kind of questionnaire decomposes the concept into several (sub-)dimensions measured through a grouping of one or several objective questions, called items. The patient-reported outcomes are increasingly used to support decision making aiming at clinical benefits. In oncology, the health-related quality of life (HRQoL) endpoint is essentially to improve patients' care and better evaluate the impact of treatments on their everyday life[1]. The HRQoL could be considered a primary endpoint in certain situations like geriatric or palliative situations. In Europe, the questionnaires are developed by the European Organization for Research and Treatment of Cancer (EORTC). The standard one is the EORTC QLQ-C30[2] and it decomposes the HRQoL into five functional dimensions, nine symptomatic dimensions and the Global Health Status (GHS). In clinical trials, patients fill out the generic HRQoL questionnaire on every visit : inclusion, during the treatment and the follow-up. The repeated measurements over time give the longitudinal aspect of their HRQoL level for each dimension.

If the measurement tools are similar across the application fields, there is a strong heterogeneity as to the statistical model used for the longitudinal analysis[3–5]. The item responses are built from a Likert scale and then the analyses are performed on dimensions (multiple-item ordinal data). In literature, two of the key statistical approaches are : the item response theory (IRT) and the classical test theory (CTT)[6;7]. In the former approach, the IRT models take into account the raw data directly[8–11]. The corresponding models are generalized linear mixed models for ordinal data with specific parametrization of the linear predictor. The main assumption is that there exists an unidimensional latent variable that contains all of the information which is common to each item[12] (multidimensional responses). The distinction between the responses relative to different items is made through item parameters in the linear predictor. This kind of models has lot of success in the psychometric and social fields and its interest is growing in the medical domain. Not only does it offer the advantage of taking into account the raw data, but it also allows to further analyze every specific item[13]. The second approach is based on score-study. The use of a scoring procedure by dimensions is required. The underlying philosophy is that the score is close to the real concept level up to some error. The use of a quantitative summary variable is common practice because so far, classical statistical methods for quantitative variables are more powerful and easier to implement and interpret than those conventionally applied to qualitative variables[14]. The associated classical model is the Linear Mixed Model (LMM) which allows to take into account the influence of both covariates and the variability within data through random effects. The IRT approach is scientifically more rigorous and informative than the score-study one[5;13], but the comparison studies did not show different conclusions across the different kind of mixed regression models used[4]. In this paper, we focused on score-study for the HRQoL analysis in oncology medical research.

Nowadays, the HRQoL study is being carried out through an independent analysis on each HRQoL dimension. However, this leads to a problem of multiple testing and even more so if supplementary modules by cancer location are used adding other specific functional and symptomatic dimensions. Indeed, the influence of each covariate is tested for each HRQoL dimension separately. A way to get around that in the clinical trials is to take

into account only some specific HRQoL dimensions and not all of them, the choice of the dimensions depending from pathology and their clinical pertinence for the trial. The GHS is always included in the chosen dimensions.

In questionnaire-data study in medical research, the use of models involving latent variables is common practice[15;16]. This began with IRT models and structural equation modeling has picked up momentum over time[17;18], in particular for the HRQoL study[19;20]. A structural equation model formalizes the dependence links of observed numeric variables through fewer unobserved ones, call latent variables. Every latent variable is assumed to be underlying a specific set of observed variables and summarize its information. Literature widely presents two competing families of methods that deal with structural equation models : factor-methods and component-methods. The first family is based on maximum likelihood estimation and was first developed by Jöreskog[21] and implemented in the LISREL software. This method terms factors as latent variables merely assumed to be standardized normal and reflected by the observed variables. Besides, they base the estimation of their parameters (linear predictor coefficients and error-variances) on the structure of the covariance matrix of the data according to the model. But, they don't estimate the values of the factors. The component-methods family was initially developed by Wold and named PLS (for Partial Least Squares)[22–24]. This approach assumes that each latent variable is a component, i.e a linear combination of the corresponding observed variables. This assumption, stronger than the mere distribution assumption on factors, allows to easily estimate the latent variable. In many areas, it is of essence to be able to estimate the value of latent variables on subject units, since these values allow to efficiently analyze their disparities on a reduced number of dimensions. Therefore, we are interested in estimating these values but without the component constraint. So, we developed a new approach[25] based on an EM-algorithm estimation of a structural equation model. In this approach, the structural equation model parameters and the factors values are estimated with only a normal-distribution assumption on factors.

In this work, we propose a global approach taking into account the multidimensional nature of data at each time and the longitudinal aspect induced by these repeated measures. The aim is to explain the GHS (representing the general concept of the HRQoL) using the information from all HRQoL functional and symptomatic dimensions as well as additional explanatory variables. To link all observations made at a given time, a structural equation model is used on data after a pre-treatment carrying out the scoring-procedure proposed by EORTC[26]. The EM algorithm[27] is used to estimate the model parameters and factors. On the other hand, the longitudinal aspect is treated through the LMM in the same way as for the above-mentioned CTT approach. The latent variables estimated through the previous structural equation models at all times are taken into account in the same way as other observable variables which can influence the HRQoL evolution. The first section presents the methodology details of the first step, associated with the structural equation models. Then, the second section shows the LMM in this specific context. Finally, this new approach is demonstrated on real data from cancer clinical data.

2 A specific structural equation model for transversal way

A structural equation model has two types of equations, named measurement equations and structural equations. Each measurement equation relates a latent variable to the corresponding set of observed variables. A structural equation formalizes an assumed relationship between latent variables. In this part, the data feeding the structural model

are taken at a specific time. The construct of the structural model proposed in this section results from the EORTC QLQ-C30[26] decomposition, which differentiates the functional dimension family, the symptomatic dimension family and the GHS.

This decomposition suggests a model built upon two factors f^1 and f^2 that quantify and summarize respectively the individual's overall functioning health (functional status) and the individual's overall symptomatic health (symptomatic status). Thus, f^1 and f^2 are respectively linked to the observed functional variables family X^1 and the observed symptomatic variables family X^2, which respectively reflect the functional status and the symptomatic status. More precisely, X^1 and X^2 are quantitative observed variables obtained from the responses to the questionnaire after performing a scoring procedure. Besides, f^1 and f^2 explain the quantitative observed variable y (GHS) obtained after the scoring procedure.

2.1 The model and notations

The data consists in blocks of observed variables describing the same n subjects. We consider the following data-matrices and denote :
$\mathrm{X}^m = \{x_i^{j,m}\}$; $i \in [\![1,n]\!]$, $j \in [\![1,q_m]\!]$, $m \in [\![1,2]\!]$ is the $n \times q_m$ matrix coding the m^{ieth} explanatory block of observed variables $x^{1,m},...,x^{q_m,m}$. The value of variable $x^{j,m}$ for subject i is denoted $x_i^{j,m}$. Variable-blocks refer the corresponding matrices.
$y = \{y_i\}$; $i \in [\![1,n]\!]$, is the n-length vector [1] coding the observed variable depending on factors f^1 and f^2.
T (resp. $\mathrm{T}^1, \mathrm{T}^2$) refers to a $n \times r_\mathrm{T}$ (resp. $n \times r_1$, $n \times r_2$) matrix of covariates.
We assume that :

— The n subjects, hence the rows of vector y and matrices $\mathrm{X}^1, \mathrm{X}^2$ are independent and are multivariate normal vectors.

The structural equation model we handle here consists of three equations and is used to construct the factors $\widetilde{f^1}, \widetilde{f^2}$ for each visit. It is graphed on Figure 1.

As formerly mentioned, for each equation of this model, each observed variable in a block is expressed as a linear combination of the corresponding factor (both factors for y), the covariates and some noise. Hence, the model :

$$\begin{cases} \mathrm{X}^1 = \mathrm{T}^1\mathrm{D}^1 + f^1 a^{1\prime} + \varepsilon^1 & \text{(1a)} \\ \mathrm{X}^2 = \mathrm{T}^2\mathrm{D}^2 + f^2 a^{2\prime} + \varepsilon^2 & \text{(1b)} \\ y = \mathrm{T}d + f^1 c^1 + f^2 c^2 + \varepsilon^y & \text{(1c)} \end{cases}$$

The corresponding equation set, for a given subject i, reads :

$$\begin{cases} x_i^{1\prime} = t_i^{1\prime}\mathrm{D}^1 + f_i^1 a^{1\prime} + {\varepsilon_i}^{1\prime} \\ x_i^{2\prime} = t_i^{2\prime}\mathrm{D}^2 + f_i^2 a^{2\prime} + {\varepsilon_i}^{2\prime} \\ y_i' = t_i{}'d + f_i^1 c^1 + f_i^2 c^2 + {\varepsilon_i}^{y\prime} \end{cases} \quad (2)$$

where d (resp. D^m) is a $r_\mathrm{T} \times 1$ (resp. $r_m \times q_m$) parameter matrix, a^m a $q_m \times 1$ parameter vector, and ε^y (resp. ε^m) an $n \times 1$ (resp. $n \times q_m$) measurement-error matrix.
We impose that the first column of T as well as of each T^m matrix is equal to the constant

1. It is possible to easily generalize the n-length vector y to a matrix Y ($n \times q_\mathrm{Y}$) of q_Y observed variables explained by the factors.

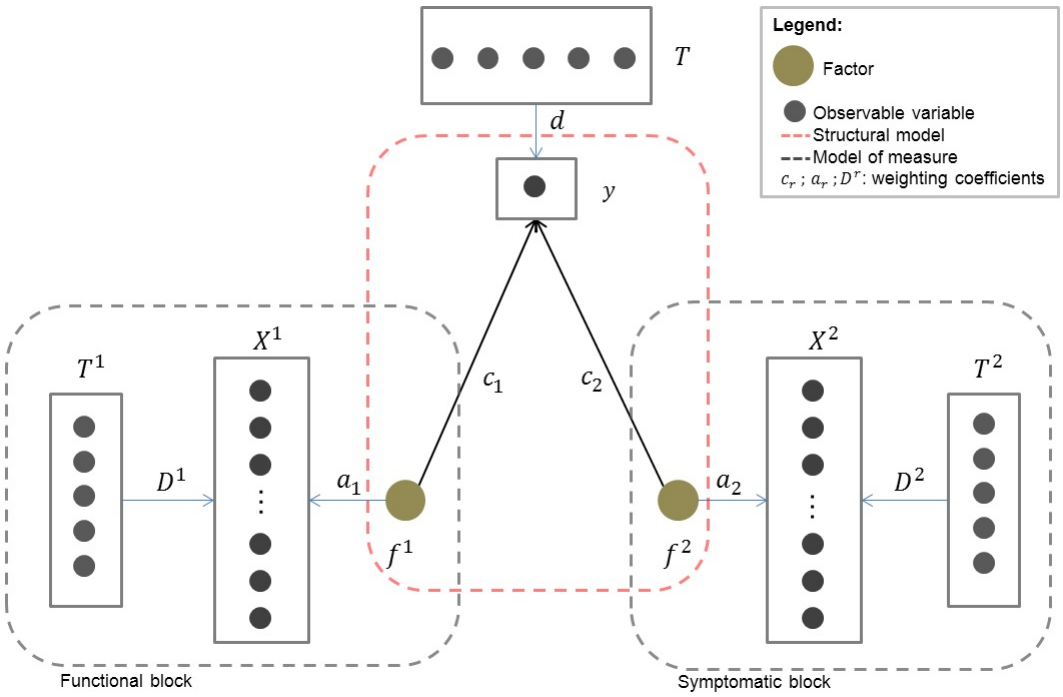


FIGURE 1 – Diagram of the structural equation model considered from the QLQ-C30 decomposition

vector having all elements equal to 1. Thus, d and the first row of each D^m contains mean-parameters.
The main assumptions of this model are the following :
As far as distributions are concerned, we assume that :

— $\varepsilon_i^y \sim \mathcal{N}(0, \sigma_y^2)$;

— $\forall m \in [\![1,2]\!] : \varepsilon_i^m \sim \mathcal{N}(0, \psi_m)$, where $\psi_m = diag(\sigma_m^2)$;

— ε^y and ε^m are independent $\forall m \in [\![1,2]\!]$.

As to the factors, we assume that :

— $\forall m \in [\![1,2]\!] : f^m \sim \mathcal{N}(0, \mathrm{I}_n)$ with f^1, f^2 independent ;

— In each block $\forall m \in [\![1,2]\!]$ the observed variables X^m (e.g. $x_j^m, j \in [\![1, q_p]\!]$) depend linearly on the block's factor (f^m) and a block of extra-covariates (e.g. T^m), conditional on which they are independent. The covariates are useful to test a treatment, typically ;

— y depend linearly on the factors f^1 and f^2 and an extra-covariate vector T ;

— ε^y, ε^m, f^m, $\forall m \in [\![1,2]\!]$ are mutually independent.

2.2 Estimation using the EM estimation method

The EM estimation method has been proposed in chapter 3 to estimate structural models. It is based on the EM likelihood-maximization algorithm[27], which allows to estimate not only the coefficients of the model, but also the values of its factors at unit-level. The idea of this approach is to consider factors as missing data, which EM allows to estimate.

The complete log-likelihood function

Let $z = (y, \mathrm{X}^1, \mathrm{X}^2)$ be the observed variables and $h = (f^1, f^2)$ the factors. The EM algorithm is based on the log-likelihood associated with the complete data (z, h). Let $p(z, h; \theta)$ denote the probability density of the complete data.

$$\begin{aligned}\mathscr{L}(\theta; z, h) = -\frac{1}{2}\sum_{i=1}^{n}\{ & ln(\sigma_y^2) + q_1 ln(\sigma_1^2) + q_2 ln(\sigma_2^2) \\ & + \sigma_y^{-2}(y_i - t_i d - f_i^1 c^1 - f_i^2 c^2)^2 \\ & + \sigma_1^{-2}(x_i^1 - \mathrm{T}_i^1 \mathrm{D}^1 - f_i^1 a^1)'(x_i^1 - \mathrm{T}_i^1 \mathrm{D}^1 - f_i^1 a^1) \\ & + \sigma_2^{-2}(x_i^2 - \mathrm{T}_i^2 \mathrm{D}^2 - f_i^2 a^2)'(x_i^2 - \mathrm{T}_i^2 \mathrm{D}^2 - f_i^2 a^2) \\ & + (f_i^1)^2 + (f_i^2)^2\} + \lambda\end{aligned}$$

Where λ is a constant and θ is the following K-dimensional set of parameters :

$$\theta = \{d, \mathrm{D}^1, \mathrm{D}^2, a^1, a^2, c^1, c^2, \sigma_y^2, \sigma_1^2, \sigma_2^2\}$$

Thus, we have :

$$\mathrm{K} = 5 + r_{\mathrm{T}} + \sum_{m=1}^{2} q_m (r_m + 1)$$

Estimation of the structural equation model

To maximize this function, in the framework of the EM algorithm, we have to solve :

$$\mathbb{E}_z^h[\frac{\partial}{\partial \theta}\mathscr{L}(\theta; z, h)] = 0 \tag{3}$$

To make it, we need the derivatives of the log-likelihood function and the distribution $p_{z_i}^{h_i}$ of h_i conditional on z_i for each observation $i \in [\![1, n]\!]$. Classically, the use of numeric methods are needed, but we deal with Gaussian distributions in our specific case and this causes $p_{z_i}^{h_i}$ to be an explicit conditional Gaussian distribution. Let us introduce the following notations :

$$p_{z_i}^{h_i} = \mathscr{N}(\mathrm{M}_i = \begin{pmatrix}\mathrm{M}_{1i} \\ \mathrm{M}_{2i}\end{pmatrix}, \Sigma = \begin{pmatrix}\sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22}\end{pmatrix})$$

$$\begin{array}{ll}\widetilde{f_i^1} = \mathbb{E}_{z_i}^{h_i}[f_i^1] = \mathrm{M}_{1i}\,; & \widetilde{\phi_i^1} = \mathbb{E}_{z_i}^{h_i}[(f_i^1)^2] = (\mathbb{E}_{z_i}^{h_i}[f_i^1])^2 + \mathbb{V}_{z_i}^{h_i}[f_i^1] = \mathrm{M}_{1i}^2 + \sigma_{11} \\ \widetilde{f_i^2} = \mathbb{E}_{z_i}^{h_i}[f_i^2] = \mathrm{M}_{2i}\,; & \widetilde{\phi_i^2} = \mathbb{E}_{z_i}^{h_i}[(f_i^2)^2] = (\mathbb{E}_{z_i}^{h_i}[f_i^2])^2 + \mathbb{V}_{z_i}^{h_i}[f_i^2] = \mathrm{M}_{2i}^2 + \sigma_{22}\end{array}$$

Thus, when we follow the same procedure described in Bry et al[25]. We obtain $\forall m \in \{1, 2\}$ the following system characterizing the solution of (3) :

$$\widehat{a^m} = [\overline{\widetilde{f^m} x^m} - \overline{x^m t^{m\prime}}(\overline{t^m t^{m\prime}})^{-1}\overline{\widetilde{f^m} t^m}][\overline{\widetilde{\phi^m}} - \overline{\widetilde{f^m} t^{m\prime}}(\overline{t^m t^{m\prime}})^{-1}\overline{\widetilde{f^m} t^m}]^{-1}$$

$$\widehat{\mathrm{D}^{m\prime}} = [\overline{x^m t^{m\prime}} - \widehat{a^m}(\overline{\widetilde{f^m} t^{m\prime}})][\overline{t^m t^{m\prime}}]^{-1}$$

$$\widehat{c^1} = \frac{[\overline{yt'}(\overline{tt'})^{-1}\overline{\widetilde{f^1} t} - \overline{\widetilde{f^1} y}][\overline{\widetilde{f^2} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^2} t} - \overline{\widetilde{\phi^2}}] - [\overline{\widetilde{f^2} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^1} t} - (\sigma_{12} + \overline{\widetilde{f^1}\widetilde{f^2}})][\overline{yt'}(\overline{tt'})^{-1}\overline{\widetilde{f^2} t} - \overline{\widetilde{f^2} y}]}{[\overline{\widetilde{f^1} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^1} t} - \overline{\widetilde{\phi^1}}][\overline{\widetilde{f^2} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^2} t} - \overline{\widetilde{\phi^2}}] - [\overline{\widetilde{f^2} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^1} t} - (\sigma_{12} + \overline{\widetilde{f^1}\widetilde{f^2}})][\overline{\widetilde{f^1} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^2} t} - (\sigma_{21} + \overline{\widetilde{f^1}\widetilde{f^2}})]}$$

$$\widehat{c^2} = \frac{[\overline{yt'}(\overline{tt'})^{-1}\overline{\widetilde{f^2} t} - \overline{\widetilde{f^2} y}][\overline{\widetilde{f^1} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^1} t} - \overline{\widetilde{\phi^1}}] - [\overline{\widetilde{f^1} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^2} t} - (\sigma_{21} + \overline{\widetilde{f^1}\widetilde{f^2}})][\overline{yt'}(\overline{tt'})^{-1}\overline{\widetilde{f^1} t} - \overline{\widetilde{f^1} y}]}{[\overline{\widetilde{f^1} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^1} t} - \overline{\widetilde{\phi^1}}][\overline{\widetilde{f^2} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^2} t} - \overline{\widetilde{\phi^2}}] - [\overline{\widetilde{f^2} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^1} t} - (\sigma_{12} + \overline{\widetilde{f^1}\widetilde{f^2}})][\overline{\widetilde{f^1} t'}(\overline{tt'})^{-1}\overline{\widetilde{f^2} t} - (\sigma_{21} + \overline{\widetilde{f^1}\widetilde{f^2}})]}$$

$$\widehat{d'} = [\overline{yt'} - \widehat{c^1}\overline{\widetilde{f^1} t'} - \widehat{c^2}\overline{\widetilde{f^2} t'}][\overline{tt'}]^{-1}$$

$$\widehat{\sigma_y^2} = \frac{1}{n}\sum_{i=1}^{n}\{(y_i - \widehat{d'} t_i)^2 - 2(y_i - \widehat{d'} t_i)(\widehat{c^1}\widetilde{f_i^1} + \widehat{c^2}\widetilde{f_i^2}) + (\widehat{c^1})^2\widetilde{\phi^1} + (\widehat{c^2})^2\widetilde{\phi^2} + 2\widehat{c^1}\widehat{c^2}(\sigma_{12} + \widetilde{f_i^1}\widetilde{f_i^2})\}$$

$$\widehat{\sigma_m^2} = \frac{1}{nq_m}\sum_{i=1}^{n}\{||x_i^m - \widehat{\mathrm{D}^{m\prime}} t_i^m||^2 + ||\widehat{a^m}||^2\widetilde{\phi_i^m} - 2(x_i^m - \widehat{\mathrm{D}^{m\prime}} t_i^m)'\widehat{a^m}\widetilde{f_i^m}\}$$

Algorithm

We index with $[s]$ any value currently obtained on the s^{ieth} iteration of the algorithm.

1. Initialization [2] = choice of the initial parameter values $\theta^{[0]}$.

2. Current iteration $s \geq 1$, until stopping condition is met :
   (a) **E-step** : with $\theta^{[s-1]}$,
      i. Calculate explicitly distribution $p_{z_i}^{h_i}$ for each $i \in [\![1, n]\!]$.
      ii. Estimate the factor-values $\widetilde{f^m}^{[s]}$, $m \in \{1,2\}$.
      iii. Calculate $\widetilde{\phi^m}^{[s]}$, $m \in \{1,2\}$.
   (b) **M-step** :
      i. Update θ to $\theta^{[s]}$ by injecting $\widetilde{f^m}^{[s]}$ and $\widetilde{\phi^m}^{[s]}$, $m \in \{1,2\}$ into the formulas of (3).

3. We used the following stopping condition with the smallest possible ϵ :

$$\sum_{k=1}^{\mathrm{K}} \frac{|\theta^{*[s+1]}[k] - \theta^{*[s]}[k]|}{\theta^{*[s+1]}[k]} < \epsilon \tag{4}$$

where θ^* is the K - dimensional vector containing the values of all scalar parameters in θ.

2. In the initialization step, $\forall m \in \{1,2\}$ we propose to obtain $\mathrm{D}^{m[0]}$ by multiple linear regression of X^m on T^m. Then, to initialize the others parameter values, we compute each approximated factor $\widetilde{f^m}^{[0]}$ as the first principal component of $\mathrm{X}^m - \mathrm{T}^m\mathrm{D}^{m[0]}$. Then, we initialize a^m, σ_m^2 (resp. d, c^m, σ_y^2) through a multiple linear regression of $\mathrm{X}^m - \mathrm{T}^m\mathrm{D}^{m[0]}$ on $\widetilde{f^m}^{[0]}$ (resp. of y on T, $\widetilde{f^1}^{[0]}$ and $\widetilde{f^2}^{[0]}$). In practice we use functions *lm()* of R and *PCA()* of the package *FactoMineR*[28] R-package .

2.3 Tests and performances of the algorithm

In the case of a more general structural model, numerical results on simulated data have been presented[25]. A data generation procedure was described and performed one hundred times, each time yielding a set of simulated data matrices (Y, X^1, X^2) [3]. Then for each simulated data, an estimation routine with a threshold value $\epsilon = 10^{-2}$ was run. The results showed that :

— Convergence was observed in almost all cases in less than five iterations.

— The estimations of the parameters θ and the factors were very close to their simulated values.

The quality of estimations was measured through the absolute relative deviation between each simulated scalar parameter in θ^* and its estimation, averaged over the hundred simulations, and through the averaged square correlations between simulated factors and their estimations.
Then a sensitivity analysis was performed to investigate how the quality of estimations could be affected by the number n of subjects and the number q of observed variables in each block. The results were that :

— The sample size n proved to have more impact on the quality of parameter estimation and factor reconstruction than the number of observed variables.

— The quality of factor reconstruction remained high for rather small values of n or q. We advise to use a minimal sample size of $n = 100$ to obtain really stable structural coefficients.

3 A linear mixed model for longitudinal analysis

As in the classical approach, the longitudinal step is performed via a LMM. The aim of the second step is to assess the HRQoL evolution over time, where the GSH is explained by the covariates and the estimated factors obtained in the first step. The response variable consists of repeated measurements on the same subjects over time. For this kind of analysis, the LMM methodology is classically used to take into account the dependence of data from the same subject and to estimate the inter-subject variability. The global HRQoL longitudinal measures y_{iv} at visit v for subject i are described by a standard linear mixed model[4;29] :

$$y_{iv} = \alpha + \boldsymbol{x}'_{iv}\boldsymbol{\beta} + \boldsymbol{u}'_i\boldsymbol{\xi}_i + \varepsilon_{iv}, \tag{5}$$

where

— α is the model intercept ;

— β is the vector of fixed effects and $\boldsymbol{x}_{iv}$ the known design vector containing $\widetilde{f}_{iv} = \left(\widetilde{f}^1_{iv}, \widetilde{f}^2_{iv}\right)$ and others covariates ;

— ξ_i is the vector of subject-specific random effects such that $\xi_i \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$ where Σ is the covariance matrix, and $\boldsymbol{u}_i$ is the known design vector ;

— $\varepsilon_{iv} \sim \mathcal{N}\left(0, \sigma^2_\varepsilon\right)$ is the error term.

3. Where Y is a $(n \times q_Y)$ matrix generalizing the n-length vector in the case of $q_Y \geq 1$ observed variable explained by the factors.

In the HRQoL assessment, two individual random effects are considered : the intercept and the slope, as is classically done in longitudinal analyses on repeated measures by subject[29;30]. In the longitudinal analysis, the random intercept takes into account the variability of the HRQoL perception specific to each patient, while the random slope is associated with the variability between patients concerning the HRQoL evolution. For the fixed part of the model, the factors $\widetilde{f}_{iv} = (\widetilde{f}^1_{iv}, \widetilde{f}^2_{iv})$ estimated at each visiting time are considered to take into account the global symptomatic status and the global functional status of patients. It is thus possible to separate the effects of the two factors in explaining the global evolution of the HRQoL. Moreover, covariates are also considered so as to study their influence on this evolution. The explanatory variables can be the same as those used on the first step, but the interpretation differ between the two steps. In the first one, the covariate's influence at a given time is assessed, while in the second its influence on the evolution of the dependent variable is assessed.

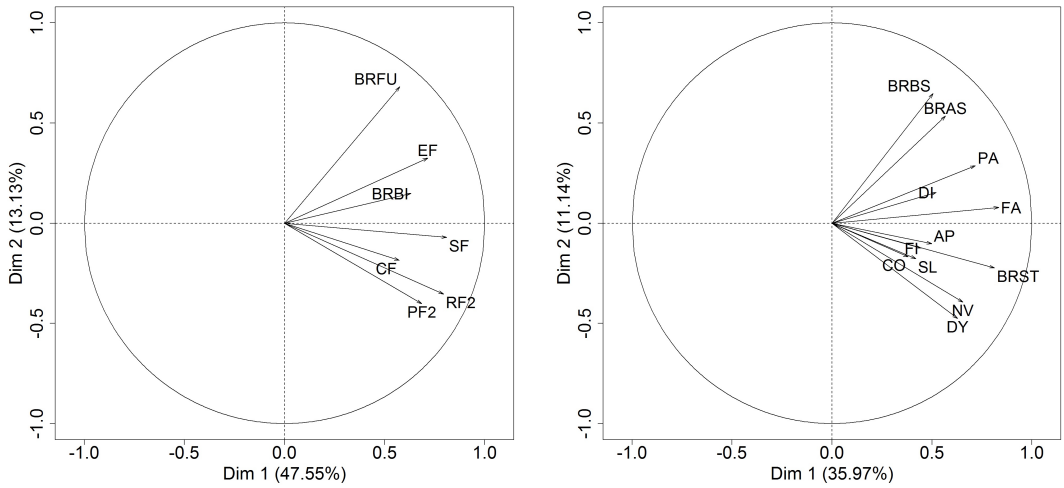
4 Application

The implementation of the approach is performed on HRQoL data from CO-HO-RT clinical trial[31]. This is a randomized phase II trial with one hundred and forty nine breast-cancer patients. The clinical trial aim was to compare two therapy programs with letrozole associated with a radiotherapy. After surgery and chemotherapy, patients from the first group (concomitant arm, $n = 74$) received some letrozole for two years and three weeks after treatment began they started radiotherapy for five weeks. Patients from the second group (sequential arm, $n = 75$) received the same treatments but the radiotherapy was administered before the treatment by letrozole. For both groups, the measurement times were every three months for two years.

At each time, the patients filled out the EORTC QLQ-C30 and the EORTC QLQ-BR23 questionnaires[26]. The latter questionnaire is specific to the breast-cancer and completes the generic EORTC QLQ-C30, and the acronym of its specific dimensions begin by "BR" for "breast". This complementary module adds eight specific HRQoL dimensions to the breast-cancer study : four dimensions among the functional dimensions and four dimensions among the symptomatic dimensions. The scoring procedure proposed by EORTC[26] was used in this analysis. For each HRQoL dimension, a score was calculated as the average of item responses which compose the dimension, and expressed on scale ranging from 0 to 100. A score was considered missing data if more than half of the item responses was missing. In this application, three dimensions (sexual functioning, sexual enjoyment, upset by hair-loss) were not considered because of too much missing data (over fifty to the whole sample).

4.1 Transversal analysis, reconstruction of factors

In this section, we aim at modeling the Global Health Status from the others variables on each visit, while reducing the dimension of the data. In this transversal step, only patients who have a measured score for each HRQoL dimension are considered in the structural equation model (1). The dependent variable y is the Global Health Status variable. Our model is built on a conceptual grouping of functional dimensions and symptomatic dimensions, suggested by the EORTC QLQ-C30 construction. For each visit, a Principal Component Analysis (PCA) of each dimension family illustrated the bundle organization and thus the grouping (Figure 2 at the baseline). On all visits, 121 different patients were



(a) Functional block (b) Symptomatic block

FIGURE 2 – Correlation-scatterplots yielded by the PCAs of the X^1 and X^2 at the baseline.

considered and the number of patient taken into account at each visit was between 84 and 113 (Table 1). The number of iterations needed for the algorithm to converge did not exceed 64 for every measurement time. As for the convergence time, one second was always enough to estimate the model.

TABLEAU 1 – Details associated with the structural model estimation for each measurement time (visit) : sample size, number of iterations and time to convergence. A total of 121 patients have completed the questionnaires to have at least one score for each considered HRQoL dimension.

	Visit v (in month)							
	0	3	6	12	15	18	21	24
Sample size (n_v)	113	106	102	100	102	84	91	90
Number of iterations	34	50	58	45	56	59	64	62
Time (s)	0.42	0.79	0.83	0.71	0.58	0.80	0.50	0.49

The main aim of the first step is the reconstruction of the functional and symptomatic factors for each subject from their different scores. In the factor estimation, the group effect can be also taken into account in the structural model to explain the observed scores, so we took the group as covariate in equations (1a), (1b) and (1c). The group-effect quantifies the score difference between the two groups for each dimension at the considered measurement times. The group covariate is defined so that value "0" corresponds to the group receiving the concurrent radiotherapy and "1" to that receiving the sequential radiotherapy. The first step is also very useful in the descriptive analysis of HRQoL scores. Indeed, each measurement equation allows to divide every HRQoL dimension into a group-effect and a group-deflated factor. The estimations of their parameters for each visit makes it possible, which can be well appreciated by clinicians. Indeed, the estimations give a short overview of covariates' effects or mean scores, and of the HRQoL evolution for each visit and specific HRQoL dimension. The transversal results may be a real benefit for clinicians to study the consistency of estimated effects with the clinical trail protocol. In our application for example, we may expect to a difference between the two groups at the visits 1 month and 3 months. This makes possible to retrieve a quick overview of studied data for a previous descriptive analysis.

For every measurement time, part of the results are showed in Tables 2, 3 and 4. They give respectively the estimations of vector d and matrices D^1 and D^2 related respectively to equations (1c), (1a) and (1b). Two estimations are distinguished. Let g denote the group-index, $(\mu_g)_{g=0,1}$ be the average scores of the two groups and $\delta = \mu_1 - \mu_0$. Regarding Table 2, the difference between the average scores of the two groups increased across time. Indeed, the GHS score of the concurrent group is constant around 70 while the GHS score of the sequential group ($\mu_1 = \mu_0 + \delta$) began to 70 and decreased from the visit at three months.

TABLEAU 2 – For each measurement time, estimations of parameters p associated with vector d from equation (1c) : intercept and treatment-group taken into account in each equation of model (1).

	Visit v (in month)							
p	0	3	6	12	15	18	21	24
μ_0	71.49	69.03	68.24	70.83	70.00	68.60	72.57	70.39
δ	0.98	-0.47	-1.74	-2.72	-1.41	-4.78	-5.90	-2.17

The results presented in Table 3 exhibit three general patterns across the visits. The first one is that the average score for functional dimension is over 60 and for several dimensions, around 80. On the scale ranging from 0 to 100, this suggests the patients had a relatively high perception of their functioning capacity. Secondly, if on the whole visits did not present a difference between the two groups, for some visits the difference could be clinically significant. Indeed, the visits which showed the greatest difference between groups as to the functional dimensions are those at 12 and 21 months. For both, the HRQoL level of patients receiving the sequential radiotherapy was uniformly lower than that of patients receiving the concurrent radiotherapy. For example, there is a difference of 5.48 points between the scores of the two groups for the physical functioning at 12 months, in favor of the concurrent group. When comparing the functional dimensions, the best perception appears to be that of the social functioning and the worst that associated with future perspective. Finally, the average scores did not appear to change the measurement times.

As for the symptomatic block, the results (Table 4) suggested similar conclusions as for the functional block. The observed symptomatic scores are low (between 0 and 45 on the 0-100 scale) that suggest a good level of HRQoL with few symptoms. The dimensions with the worst symptomatic level are insomnia, fatigue and pain. Globally, the scores seemed similar across groups whatever the visit, even if the visits at 3 months, 12 months and 21 months respectively presented scores relatively greater for patient receiving sequential radiotherapy. On the whole, there was no difference between groups which could be explained by the administration method. A difference could be expected at three months after starting (end of radiotherapy), but the difference was not conclusive.

As previously mentioned, the main aim is the estimation of the two unknown factors reflected by symptomatic and functional dimensions, once the effect of covariates is removed. The question arises thus of the influence of covariates on the factor building. Figure 3 shows the correlation between the factors estimated with the group covariates $(\widetilde{f}_{\mathrm{T}}^{m})_{m=1,2}$ and without $(\widetilde{f}^{m})_{m=1,2}$. According to the graphics, there is a low group effect for a few visits. Concerning the functional factor (Figure 3a), the visits at 12 months and 21 months exhibit a clear difference between groups. Indeed, $\widetilde{f}_{\mathrm{T}}^{1} < \widetilde{f}^{1}$ for the concurrent administration group (red), and $\widetilde{f}_{\mathrm{T}}^{1} > \widetilde{f}^{1}$ for the sequential administration group (blue).

TABLEAU 3 – For each measurement time, estimations of parameters p associated with matrix D^1 from equation (1a) : intercept and treatment-group taken into account in each equation of model (1). X^1 corresponds to the functional dimension scores : physical (PF2), role (RF2), emotional (EF), cognitive (CF), social (SF), body image (BRBI) and future perspective (BRFU).

		Functioning dimensions						
Visits	p	PF2	RF2	EF	CF	SF	BRBI	BRFU
Baseline	μ_0	82.81	85.38	75.05	75.73	91.52	84.75	60.23
	δ	1.84	1.52	-1.49	2.84	0.44	2.31	1.08
3 months	μ_0	82.08	85.85	75.42	77.99	91.19	87.89	71.70
	δ	-1.29	-3.77	-1.10	1.57	0.00	-1.89	-5.66
6 months	μ_0	81.70	83.65	74.00	77.36	87.42	88.89	64.15
	δ	-0.88	1.05	0.37	1.55	-0.35	-2.32	3.20
12 months	μ_0	80.83	81.94	73.21	79.86	90.62	86.46	70.14
	δ	-5.74	-6.62	-3.65	2.19	-2.16	-3.12	-7.96
15 months	μ_0	82.80	81.00	74.56	79.67	90.00	89.50	68.00
	δ	0.95	-1.51	-4.52	0.14	-2.18	-6.65	-3.90
18 months	μ_0	84.19	83.33	70.54	78.68	91.86	89.34	70.54
	δ	1.18	-1.22	-0.83	3.03	-0.80	-6.41	-2.25
21 months	μ_0	85.83	82.29	78.30	80.90	90.62	88.19	65.97
	δ	-5.48	0.27	-9.31	-5.32	-3.03	-6.28	-3.18
24 months	μ_0	82.55	81.56	74.53	78.37	89.01	86.17	68.09
	δ	-0.46	2.55	-0.88	0.31	0.14	-2.84	-2.19

TABLEAU 4 – For each measurement time, estimations of parameters p associated with matrix D^2 from equation (1b) : intercept and treatment-group taken into account in each equation of model (1). X^2 corresponds to the symptomatic dimension scores : fatigue (FA), nausea and vomiting (NV), pain (PA), dyspnea (DY), insomnia (SL), appetite loss (AP), constipation (CO), diarrhoea (DI), financial difficulty (FI), systemic therapy side effect (BRST), breast symptom (BRBS) and arm symptoms (BRAS).

		Symptomatic dimensions											
Visits	p	FA	NV	PA	DY	SL	AP	CO	DI	FI	BRST	BRBS	BRAS
Baseline	μ_0	24.95	3.80	17.84	14.62	31.58	8.77	17.54	8.19	4.09	19.26	15.84	16.96
	δ	0.64	0.66	1.51	-1.52	-1.82	-1.63	-1.47	-4.02	4.84	-1.38	-0.81	-2.77
3 months	μ_0	24.11	3.14	22.64	22.01	34.59	5.66	18.87	6.92	5.03	16.71	19.34	17.19
	δ	6.81	1.89	5.35	-1.26	4.40	0.63	-3.77	-1.26	3.77	2.61	2.15	4.82
6 months	μ_0	26.94	3.14	25.16	18.24	38.36	9.43	13.21	8.80	5.03	19.06	18.92	21.91
	δ	-0.52	1.28	2.05	2.17	-5.71	-3.31	5.16	-3.36	0.41	-2.58	-2.37	-6.49
12 months	μ_0	28.01	4.51	27.43	18.75	40.28	6.25	14.58	5.56	6.94	18.14	17.88	21.99
	δ	5.00	0.61	3.66	6.89	-1.18	5.29	4.65	3.42	-1.82	1.63	-0.20	-2.87
15 months	μ_0	26.78	4.67	26.67	18.00	38.67	8.67	14.00	8.67	3.33	18.45	13.72	16.33
	δ	3.56	-0.82	-2.95	2.51	1.72	-2.90	-3.10	-2.90	1.15	1.95	1.02	-0.31
18 months	μ_0	31.01	3.88	20.93	15.50	42.64	10.85	12.40	7.75	6.20	18.11	13.18	14.99
	δ	-2.69	-1.44	6.71	-0.06	-8.49	-6.79	3.04	-0.43	-4.58	-0.96	4.44	3.58
21 months	μ_0	21.18	3.13	25.35	15.28	38.19	4.86	12.50	8.33	4.86	17.87	14.29	16.20
	δ	8.92	2.69	-2.87	7.20	-0.21	2.89	2.23	-4.46	2.89	3.20	1.79	1.50
24 months	μ_0	22.70	3.55	23.76	16.31	32.62	7.09	12.06	7.80	9.22	18.22	13.18	14.78
	δ	1.85	-1.22	1.43	0.74	-0.07	-1.67	7.32	-2.38	-2.24	-0.30	-0.45	3.05

This suggests for these visits, the sequential administration group had a functional health status lower than the concurrent administration group. This confirms the estimation gi-

ven in Table 3. However, no reason from the protocol allowed to explain the small difference for these specific visits. Concerning the measurement model associated with the symptomatic dimensions (Figure 3b), the estimation of the symptomatic factor is hardly affected by the group covariate, with the exception of the visit at 3, 12 and 18 months. Indeed, $\widetilde{f}^2_{\mathrm{T}} > \widetilde{f}^2$ for the concurrent administration group (red) for these measurement times, while for the sequential administration group $\widetilde{f}^2_{\mathrm{T}} < \widetilde{f}^2$. The reverse situation for the symptomatic health status is clinically consistent with the functional factor observations. The symptomatic status of the sequential administration group is higher than that of the concurrent administration group. The latter group seemed to have a low HRQoL given the symptomatic health status for the three visits.

The estimation of factors is impacted by the taking into account of covariates when modeling the dependent variable. When covariates are omitted, their effect on the HRQoL score is conveyed by the factors, the other variable(s) in T and the error term of the measurement models. According to Figure 3a, a lower functional capacity for the sequential administration group caused a global decreasing of the functional factor in this group when "group" covariate was not taken into account. By contrast, the functional factor for the concurrent administration group increased. When the covariate effect is not considered in the model, the estimated factors will be biased with respect to the true hidden factor.

The transversal analysis brought some punctual information on the difference or similarity between the HRQoL at every measurement time across groups, information which was not visible in the classical longitudinal studies by dimension. The sequential administration group presented a better functional health status at the visits at 12 and 21 months after the inclusion, and a lower symptomatic health status three months after the inclusion than the concurrent administration group. Another strong advantage, the structural equation model approach via the algorithm EM is to summarize the symptomatic and functioning scores in two factors which will now be considered in the longitudinal step.

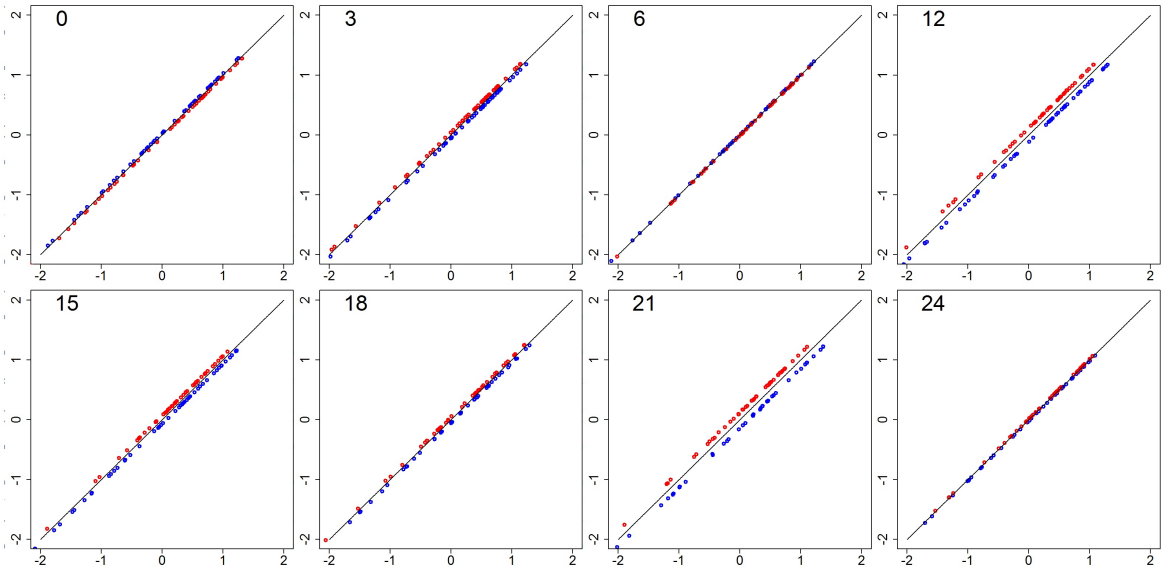
4.2 Longitudinal analysis

After considering the data in the transversely, we are interested in studying the global perception of HRQoL over time. To achieve that, the LMM is used to link the measurements of the same subject. Considering the GSH as the response variable y, the complete LMM ($\mathscr{M}_c$) considered for this specific application is :

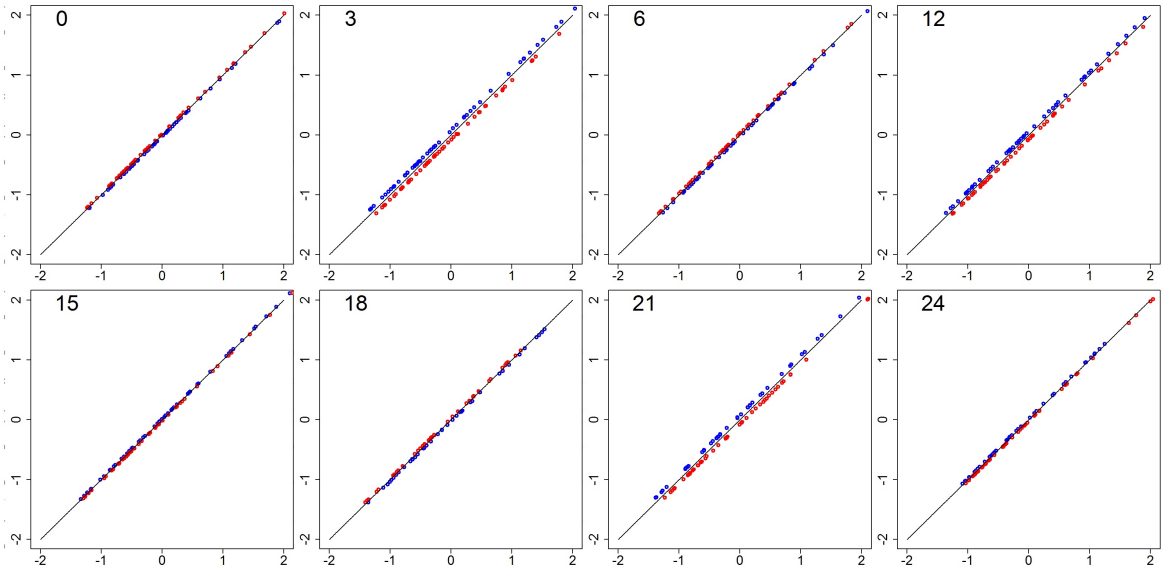
$$\begin{aligned} \mathscr{M}_c: \quad y_{iv} &= \beta_0 + \beta_1 \widetilde{f}^1_{iv} + \beta_2 \widetilde{f}^2_{iv} + \beta_3 x_i + \beta_4 t_v \\ &\quad + \beta_5 x_i t_v + \beta_6 \widetilde{f}^1_{iv} x_i + \beta_7 \widetilde{f}^2_{iv} x_i + \xi_{i0} + \xi_{i1} t_v + \varepsilon_{iv}, \end{aligned} \tag{6}$$

where

- $(\beta_j)_{j=0,\dots,7}$ are the fixed effects ;
- x_i is the affiliation-indicator of the group : $x_i = 0$ if the subject i received the concurrent treatment, and $x_i = 1$ else (sequential administration) ;
- $\widetilde{f}^1_{iv}$ and $\widetilde{f}^2_{iv}$ are respectively the estimated functional and symptomatic factors for subject i ;
- t_v is the time (in month) elapsed since the inclusion (baseline, $t_0 = 0$) and the visit v.



(a) Functional factors ($\widetilde{f}^1$)



(b) Symptomatic factors ($\widetilde{f}^2$)

FIGURE 3 – Scatterplot of estimated factors with (for x-axis) and without (for y-axis) taking into account the treatment group covariate in the both measurement blocks and the dependent variable y. The associated visit-date is mentioned on the top left corner of every plot. The concurrent treatment administration is plotted in red and the sequential one in blue.

TABLEAU 5 – Details of the LMM estimation. The fixed effects, the variance of random intercept and the BIC associated with every model are given. γ is the difference between BIC$(\mathscr{M}_\ell)$ and BIC$(\mathscr{M}_j)_{j=1,\dots,4}$.

Models	$\hat{\beta_0}$	$\hat{\beta_1}$	$\hat{\beta_2}$	$\hat{\sigma}^2_{\xi_0}$	$\hat{\sigma}^2_{\varepsilon}$	BIC	γ
$\mathscr{M}_\ell$	69.11	10.14	-7.61	78.77	160.40	6440.4	0
$\mathscr{M}_1$	68.97	15.89	$\ltimes$	92.91	168.14	6481.6	41.2
$\mathscr{M}_2$	69.07	$\ltimes$	-15.17	77.89	180.71	6516.2	75.8
$\mathscr{M}_3$	69.12	8.65	-8.38	$\ltimes$	236.83	6568.2	127.8
$\mathscr{M}_4$	68.14	$\ltimes$	$\ltimes$	225.64	245.65	6823.0	382.7

No interaction between the time and factors was taken into account, since factors were estimated at every visit, and thus contain the visit-time information. To fit the model, the *lmer* function of the R package *lme4* was used[32]. The Bayesian information criterion was used to select the model the most likely. The factors built while taking into account of the group variable were not considered. In the longitudinal analysis, they were not used because they seemed less relevant than the factors estimated without the knowledge of group. From empirical data, the model $(\mathscr{M}_\ell)$ found the most likely to explain the GSH was :

$$\mathscr{M}_\ell : \quad y_{iv} = \beta_0 + \beta_1 \widetilde{f}^1_{iv} + \beta_2 \widetilde{f}^2_{iv} + \xi_{i0} + \varepsilon_{iv} \tag{7}$$

As anticipated from previous transversal results, no evolution of the GHS over time was, found, nor any difference between groups. The time information included in the factor, if any, is sufficient. There is no difference between groups at inclusion, as expected, because the patients were randomized. Table 5 gives the estimation details of a few models (7). The functional factor seems to give twice as much information as the symptomatic factor to explain the GHS. The individual part conveyed by the random intercept is the most informative component. This suggests that the HRQoL depends strongly of the subject, and confirms the subjective specificity of the HRQoL endpoint. Moreover, if the two factors are not taken into account to explain the GHS, there is a consequent loss of information. Indeed, the two factors are rather strongly correlated (Figure 4, $\rho = -0.81$). Thus, the absence of one is to some extent compensated by the other. Figure 4 shows that each variable family is positively correlated with the corresponding factor.

5 Discussion

The new approach to study the HRQoL in oncology is original and suggests a new course of action. The proposed method is presented to study the HRQoL from a global point of view. The first strong point is the use of a structural equation model estimation at every measurement time. This allows the estimation of each parameter and subject-specific factor. A descriptive analysis can first be done on the estimated parameters at each time. This can give hints as to the influence of observable variables included in the model and allows to compare HRQoL dimension families or specific HRQoL dimensions. From these results, some clinical interpretations can emerge and provide information on how to perform longitudinal analysis. Moreover, the interpretation of coefficients in such a model is the same as in the classical regression equation for each equation regression of the model. One of the interests of the transversal step is to see the potentially different changes across the visits. The strongest asset of our approach is the estimation of latent

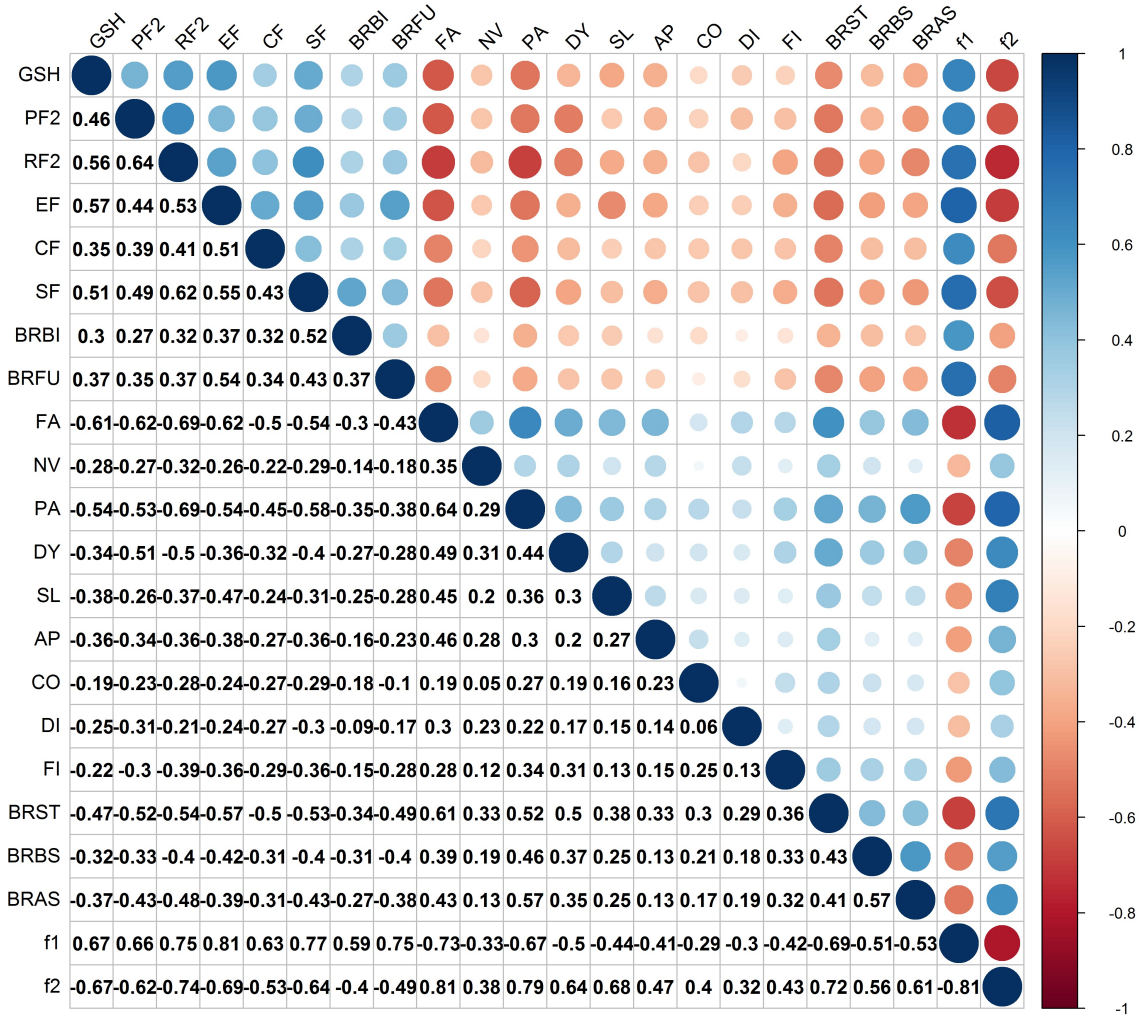


FIGURE 4 – Correlation matrix between each HRQoL dimension score and the estimated factor concerning the CO-HO-RT clinical trial data.

factors. This allows to summarize the information of different variables into a single one for every group of variables measuring a given concept. In the proposed structural equation model, two factors summarizing the functional health status and the symptomatic health status are considered. After the first step, these factors are available for longitudinal analysis. The second and final step performs longitudinal analysis of the global HRQoL through the GHS study. In this article, the use of a classical LMM is proposed to assess the global HRQoL over time as a function of covariates and factors. The two-step approach performs a complete analysis of the HRQoL scores for clinical trial in oncology.

Longitudinal analyses of the functional and symptomatic parts may also be performed. Moreover, the influence of each dimension on the reflected factor was not presented in the results. To compare their influence, it is necessary to have homogeneous variances across variables of the same family (assumption of the structural equation model proposed). Then, the vector of coefficients $(a^m)_{m=1,2}$ represents the intensities of links between the dimensions of the family m and the associated factor $\left(f^m\right)_{m=1,2}$. Finally, we have considered but a single factor reflected by all dimensions in each family. Indeed, this seemed most appropriate for our data. But, for other data, we might imagine that symptomatic or functional dimensions reflect several uncorrelated factors.

To consider that the HRQoL score is normally distributed is questionable as it always is in clinical analyses of the HRQoL. Another consideration is questionable according to the literature about our approach : the use of the GSH to represent the global HRQoL[33]. The GHS is controversial and has to be used with caution because this HRQoL dimension is likely to be affected by the *Response Shift* phenomena[34;35]. To remedy this situation, at least two substitution global scores are proposed in literature, built from all measurements of HRQoL dimensions. The first one is the difference between, on the one hand, the sum of functional dimension scores and the GSH score, and, on the other hand, the sum of symptomatic dimension scores except the financial difficulties[36]. This score belongs to the interval $[-800;600]$ and is considered missing if any one of all HRQoL scores is missing. More recently, a global summary score expressed on a scale from 0 to 100 is proposed as the average of the twenty item responses of the questionnaire EORTC QLQ-C30, given the reverse permutation of response categories of functional item[37]. This one seems more relevant because such as calculation allows to take into account some weighting of the dimensions according to their number of items. Indeed, the more items a dimension contains, the stronger its contribution. In our structural model, we chose to use the global health status score because its calculation is not derived from the other (symptomatic and functional) scores. Using a score which depends on the other scores in our model is not suitable because it would clearly ruin the exogeneity of the dependent score with respect to the explanatory ones.

Références

1. Fiteni F, Westeel V, Pivot X et al. Endpoints in cancer clinical trials. *Journal of Visceral Surgery* 2014 ; 151(1) : 17–22. 102
2. Aaronson NK, Ahmedzai S, Bergman B et al. The European Organization for Research and Treatment of Cancer QLQ-C30 : a quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute* 1993 ; 85(5) : 365–376. 102
3. Blanchin M, Hardouin JB, Neel TL et al. Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes. *Statistics in Medicine* 2011 ; 30(8) : 825–838. 102
4. Anota A, Barbieri A, Savina M et al. Comparison of three longitudinal analysis models for the health-related quality of life in oncology : a simulation study. *Health and Quality of Life Outcomes* 2014 ; 12. 102, 108
5. Gorter R, Fox JP and Twisk JW. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Medical Research Methodology* 2015 ; 15(1) : 55. 102
6. Boeck Pd and Wilson M. *Explanatory Item Response Models : A Generalized Linear and Nonlinear Approach.* 2004 edition ed. New York : Springer, 2004. 102
7. De Ayala RJ. *The theory and practice of item response theory.* New York : Guilford Press, 2009. 102
8. Rasch G. On General Laws and the Meaning of Measurement in Psychology. The Regents of the University of California. 102
9. Samejima F. Estimation of Latent Ability Using a Response Pattern of Graded Scores1. *ETS Research Bulletin Series* 1968 ; 1968(1) : i–169.
10. Masters G. A rasch model for partial credit scoring. *Psychometrika* 1982 ; 42(2) : 149–174.
11. Tutz G. Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology* 1990 ; 43(1) : 39–55. 102
12. Douglas JA. Item response models for longitudinal quality of life data in clinical trials. *Statistics in Medicine* 1999 ; 18(21) : 2917–2931. 102
13. Edelen MO and Reeve BB. Applying item response theory (irt) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 2007 ; 16 Suppl 1 : 5–18. 102
14. Grilli L and Rampichini C. Multilevel Models for Ordinal Data. In Kenett RS and Salini S (eds.) *Modern Analysis of Customer Surveys.* John Wiley & Sons, Ltd, 2011. pp. 391–411. 102
15. Bentler P and Stein J. Structural equation models in medical research. *Statistical Methods in Medical Research* 1992 ; 1(2) : 159–181. 103
16. Rabe-Hesketh S and Skrondal A. Classical latent variable models for medical research. *Statistical Methods in Medical Research* 2008 ; 17(1) : 5–32. 103
17. Lei PW and Wu Q. Introduction to structural equation modeling : Issues and practical considerations. *Educational Measurement : Issues and Practice* 2007 ; 26(3) : 33–43. 103
18. Titman AC, Lancaster GA and Colver AF. Item response theory and structural equation modelling for ordinal data : Describing the relationship between KIDSCREEN and Life-H. *Statistical Methods in Medical Research* 2013 ; . 103
19. Huang CC, Lien HH, Tu SH et al. Quality of life in taiwanese breast cancer survivors

with breast-conserving therapy. *Journal of the Formosan Medical Association = Taiwan Yi Zhi* 2010 ; 109(7) : 493–502. 103

20. King-Kallimanis BL, Oort FJ, Nolte S et al. Using structural equation modeling to detect response shift in performance and health-related quality of life scores of multiple sclerosis patients. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2011 ; 20(10) : 1527–1540. 103
21. Joreskog KG. A General Method for Analysis of Covariance Structures. *Biometrika* 1970 ; 57(2) : 239–251. 103
22. Wold H. Partial Least Squares. In *Encyclopedia of Statistical Sciences*, volume 6. New York : John Wiley & Sons. ISBN 978-0-471-66719-3, 1985. pp. 581–591. 103
23. Wold H. Estimation of Principal Components and Related Models by Iterative Least squares. In *Multivariate Analysis*. Academic Press, 1966. pp. 391–420.
24. Noonan R and Wold H. NIPALS Path Modelling with Latent Variables. *Scandinavian Journal of Educational Research* 1977 ; 21(1) : 33–61. 103
25. Bry X, Lavergne C and Tami M. EM estimation of a structural equation model. *in review* 2016 ; . 103, 106, 108
26. Fayers PM, Aaronson NK, Bjordal K et al. *EORTC QLQ-C30 Scoring Manual (3rd edition)*, volume Brussels : EORTC 2001. 2001. 103, 104, 109
27. Dempster AP, Laird NM and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1977 ; 39(1) : 1–38. 103, 105
28. Husson F, Josse J, Le S et al. *FactoMineR : Multivariate Exploratory Data Analysis and Data Mining*, 2015. R package version 1.29. 107
29. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data : With Applications in R*. CRC Press, 2012. 108, 109
30. Agresti A. *Analysis of Ordinal Categorical Data*. John Wiley & Sons, 2010. 109
31. Azria D, Belkacemi Y, Romieu G et al. Concurrent or sequential adjuvant letrozole and radiotherapy after conservative surgery for early-stage breast cancer (CO-HO-RT) : a phase 2 randomised trial. *The Lancet Oncology* 2010 ; 11(3) : 258–265. 109
32. Bates D, Mächler M, Bolker B et al. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 2015 ; 67(1) : 1–48. 115
33. Phillips R, Gandhi M, Cheung YB et al. Summary scores captured changes in subjects' QoL as measured by the multiple scales of the EORTC QLQ-C30. *Journal of Clinical Epidemiology* 2015 ; . 117
34. Sprangers MA and Schwartz CE. Integrating response shift into health-related quality of life research : a theoretical model. *Social science & medicine (1982)* 1999 ; 48(11) : 1507–1515. 117
35. Schwartz CE and Sprangers MA. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social science & medicine* 1999 ; 48(11) : 1531–1548. 117
36. Nordin K, Steel J, Hoffman K et al. Alternative methods of interpreting quality of life data in advanced gastrointestinal cancer patients. *British Journal of Cancer* 2001 ; 85(9) : 1265–1272. 117
37. Hinz A, Einenkel J, Briest S et al. Is it useful to calculate sum scores of the quality of life questionnaire EORTC QLQ-C30 ? *European Journal of Cancer Care* 2012 ; 21(5) : 677–683. 117

4.5 Discussion, commentaires et perspectives

Alors qu'il pouvait y avoir autant d'analyses et de tests que de dimensions, l'emploi de l'approche EM comme étape préliminaire à l'analyse longitudinale du critère de qualité de vie réduit le nombre d'analyses à une seule par visite. L'objectif de simplification de l'analyse longitudinale de la HRQoL est donc bien satisfait. Pour y parvenir les estimations des deux facteurs à chaque visite jouent le rôle de résumés du statut fonctionnel et du statut symptomatique. Alors, à la seconde étape une seule modélisation linéaire mixte est réalisée et constitue l'analyse longitudinale de la HRQoL où le *statut global de santé* est choisi comme variable réponse. En effet, il est conjecturé comme représentatif de la HRQoL globale, ce qui est discutable. D'ailleurs c'est une dimension controversée. Dans la littérature, il est conseillé de l'utiliser avec prudence car :

— de manière longitudinale, elle serait susceptible d'être en désaccord avec les autres dimensions ;

— elle serait sujette au phénomène de "Response Shift" (ref. Schwartz and Sprangers, 1999) ;

— elle est positionnée à la fin du questionnaire, ce qui influerait sur la qualité des réponses.

La nature gaussienne des données est elle aussi discutable. Bien qu'elles aient été pré-traitées par la procédure de scoring, cela ne les rend pas gaussiennes pour autant. D'ailleurs la procédure de scoring limite les valeurs prises par les différentes variables par les bornes 0 et 100. De plus, les deux facteurs ont été établis de part la construction du questionnaire QLQ-C30 comme résumant les statuts fonctionnel et symptomatique, or ces statuts pourraient être envisagés comme composés de plusieurs facteurs et donc multidimensionnels. Étendre l'approche EM à plusieurs facteurs par famille ou groupe de dimension est une perspective qui s'inscrit dans les perspectives du chapitre 3. Mais dans le cadre de l'application de ce chapitre, cette extension paraît apporter peu d'information par rapport à la modélisation à facteurs unidimensionnels. En effet, envisager plusieurs facteurs par groupe dans le cas de familles comportant beaucoup plus de dimensions pourrait être intéressant mais ce n'est pas le cas ici. En revanche, puisque le questionnaire QLQ-C30 est complété par le QLQ-BR23, comparer la modélisation faite dans ce chapitre avec une modélisation où les deux facteurs seraient bi-dimensionnels semble cohérent. Chaque facteur commun serait composé de deux sous-facteurs résumant les dimensions issues du QLQ-C30 d'une part et celles issues du BR23 d'autre part. Concernant les deux facteurs du modèle linéaire de la seconde étape, leurs estimations sont obtenues avec une forte corrélation de -0.8. Cela pousse à considérer une relation de cause à effet supplémentaire entre les deux facteurs. Il serait intéressant d'estimer un tel modèle et de procéder à une comparaison avec les résultats de l'application. Les facteurs ne seraient alors plus indépendant et cela semble avoir du sens car il est fort probable qu'un patient subissant de forts symptômes verra son statut fonctionnel impacté. Par exemple, une forte fatigue aura des conséquences sur les capacités physiques, tout comme des difficultés financières peuvent avoir des conséquences sur le statut social.

Conclusion

Ce travail a porté sur l'estimation des modèles à équations structurelles et de leurs variables latentes. Après avoir passé en revue les deux principaux paradigmes actuels d'estimation que sont les méthodes de moindres carrés partiels sur composantes et LISREL, nous avons proposé une méthode d'estimation fondée sur la maximisation, via EM, de la vraisemblance globale du modèle.
Nous récapitulons ci-après l'ensemble des apports de ce travail, en soulignons les limites, et esquissons quelques perspectives de recherches.

Apports

L'apport principal de ce travail de thèse est le développement de la méthode d'estimation par algorithme EM du maximum de vraisemblance et la reconstruction des facteurs latents d'un modèle à une équation structurelle. Nous avons procédé à l'extension de ce développement à un modèle multi-blocs avec adjonction de covariables additionnelles à chaque bloc. Cette méthode a été programmée sous R. Une analyse de sensibilité a été conduite afin d'évaluer les performances de la méthode, qui se sont révélées satisfaisantes. Une étude sur des données simulées a montré son efficacité et sa vitesse de convergence.
À la faveur d'une application sur des données réelles environnementales, nous avons montré comment construire pratiquement un modèle, et en évaluer la qualité.
Lors de l'application à l'étude de la qualité de vie dans le domaine de la cancérologie, nous avons montré comment la méthode pouvait être utilisée à l'intérieur d'une modélisation plus complexe pour l'étude de données longitudinales. Nous avons ainsi montré que grâce à une réduction efficace de la dimension des données, elle simplifiait l'analyse longitudinale de la qualité de vie en évitant les tests multiples et pouvait ainsi aider les praticiens à évaluer plus facilement le bénéfice clinique d'un traitement.

Limites et perspectives

L'application à la qualité de vie a toutefois révélé une partie des limites de ce travail. La forte corrélation entre les facteurs estimés lors de l'application présentée au chapitre 4 nous pousse à envisager des relations structurelles supplémentaires. En effet, le modèle pour lequel l'approche est développée est trop simple, malgré le début de généralisation proposé au chapitre 3. L'approche est encore embryonnaire et il faut l'étendre de sorte à pouvoir traiter des modélisations structurelles plus riches. En effet, les relations structurelles devraient pouvoir en pratique impliquer des variables catégorielles comme des liaisons non-linéaires.
Une première perspective pourrait être d'étendre l'approche aux modèles linéaires généralisés. Les variables pourraient alors être discrètes : catégorielles ou de comptage. Par ailleurs, nous pouvons sans grande difficulté envisager la modélisation des interactions entre les variables

à la fois dans les modèles de mesure et dans le modèle structurel. Parmi les généralisations à envisager, il y a aussi la prise en compte de plusieurs facteurs par blocs, ces derniers étant souvent en pratique fondamentalement multi-dimensionnels. La difficulté de cette dernière extension est essentiellement technique, étant liée aux contraintes d'identifiabilité supplémentaires. Enfin, l'illustration de l'utilisation de l'approche au chapitre 3 n'est pas suffisante et mérite d'être enrichie par un indice de validation de modèle. Ainsi, en tenant compte de toutes ces perspectives, la méthode numérique devra être complexifiée de manière à devenir plus générique.
Parmi les perspectives plus proches, nous envisageons la création d'un package R facile d'utilisation avec des pages d'aide et contenant toutes les fonctionnalités proposées en perspective afin que la procédure soit plus complète. À l'aide de ce package, nous conduirons la réalisation d'une étude comparative complète entre l'approche EM et celles de PLS et LISREL.

ANNEXE A

Démonstrations aboutissant aux formules des estimateurs des paramètres du modèle de LISREL

Démonstration 1. *Soit la fonction,*

$$F(B,\Psi) = log|\Sigma^Y| + tr\left(S{\Sigma^Y}^{-1}\right) - log|S| - q$$

que nous cherchons à minimiser. Pour des raisons de simplification de calculs, nous choisissons $\Sigma^Y = BB' + \Psi$ *avec* B *de dimension* $(q \times K)$. *Ce qui est équivalent à la définition de* Σ^Y *faite aux chapitres précédents où* $\Sigma^Y = B'B + \Psi$ *avec* B *de dimension* $(K \times q)$. K *correspond toujours au nombre de facteurs et* Σ^Y *reste de dimension* $(q \times q)$. *Nous nous plaçons sous l'hypothèse que les facteurs sont orthogonaux et non corrélés avec* $\Psi = diag(\sigma_j^2)_{j\in[\![1,q]\!]} = diag(\sigma_i^2)_{i\in[\![1,q]\!]}$ *inconnue. En effet, lorsque* $K > 1$ *il y a trop de paramètres à spécifier de manière unique. Rappelons que* $S = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)'(y_i - \mu)$ *et introduisons les notations suivantes :*

$$S = [s_{ij}]_{\{i,j\}\in[\![1,q]\!]^2},$$

la matrice de variance-covariance observée des y_i *et les deux matrices inconnues :*

$$\Sigma^Y = [\sigma_{Y;ij}]_{\{i,j\}\in[\![1,q]\!]^2}$$

et

$$B = [b_{ik}]_{i\in[\![1,q]\!];k\in[\![1,K]\!]}.$$

Ainsi, à partir de $\Sigma^Y = BB' + \Psi$ *on obtient pour* $(i \neq j)$ *:*

$$\begin{cases} \sigma_{Y;ii} &= \sum_{k=1}^{K} b_{ik}^2 + \sigma_i^2 \\ \sigma_{Y;ij} &= \sum_{k=1}^{K} b_{ik}b_{jk} \end{cases}$$

Afin d'obtenir la dérivée partielle (1.25) *pour* Ψ *fixé, nous allons expliciter les deux dérivées partielles* $\frac{\partial}{\partial B}log|\Sigma^Y|$ *et* $\frac{\partial}{\partial B}tr\left(S{\Sigma^Y}^{-1}\right)$, *toutes deux pour* Ψ *fixé.*

1. *Explicitons* $\frac{\partial}{\partial B}tr\left(S{\Sigma^Y}^{-1}\right)$ *pour* Ψ *fixé :*
   Pour cela nous introduisons quelques notations et résultats que nous justifierons au fur et à mesure. Le premier résultat est le suivant :

$$tr\left(S{\Sigma^Y}^{-1}\right) = \sum_{ij} s_{ij}\sigma_Y^{ij} = \sum_{ij} s_{ij}\frac{{\Sigma^Y}_{ji}}{|\Sigma^Y|} \tag{A.1}$$

   où,

— σ_Y^{ij} *est l'élément de la* i*-ème ligne et de la* j*-ème colonne de* ${\Sigma^Y}^{-1}$*, c'est à dire* ${\Sigma^Y}^{-1} = [\sigma_Y^{ij}]_{\{i,j\}\in[\![1,q]\!]^2}$ *;*

— ${\Sigma^Y}_{ij}$ *est le cofacteur de* $\sigma_{Y;ij}$ *dans* $|\Sigma^Y|$ *qui correspond au produit de* $(-1)^{i+j}$ *par le déterminant de la matrice* Σ^Y *privée de la ligne* i *et de la colonne* j*.*

La partie droite de l'égalité (A.1) *est justifiée par le fait qu'il peut être démontré que :*

$$|\Sigma^Y| = \sum_j \sigma_{Y;ij}\Sigma^Y_{ij} = \sum_j \sigma_{Y;ji}\Sigma^Y_{ji} \tag{A.2}$$

et que pour $i \neq l$*,*

$$|\Sigma^Y| = \sum_j \sigma_{Y;ij}\Sigma^Y_{lj} = \sum_j \sigma_{Y;jl}\Sigma^Y_{jl} = 0$$

Ainsi, si Σ^Y *est non singulière, on peut définir* ${\Sigma^Y}^{-1} = [\sigma_Y^{ij}]$ *par :*

$$\sigma_Y^{ij} = \frac{\Sigma^Y_{ji}}{|\Sigma^Y|} \tag{A.3}$$

d'où le résultat (A.1)*. De plus, puisque* Σ^Y *est symétrique, nous avons* $\Sigma^Y_{ji} = \Sigma^Y_{ij}$*. À partir de ces résultats que nous venons de justifier, nous allons utiliser la partie de droite de l'égalité* (A.1) *qui est de la forme suivante,*

$$\sum_{uw} s_{uw}\frac{{\Sigma^Y}_{uw}}{|\Sigma^Y|} \tag{A.4}$$

pour réécrire $tr\left(S{\Sigma^Y}^{-1}\right)$ *et en expliciter la dérivée partielle par rapport à* B*. Nous introduisons les indices* u *et* w *pour faciliter les calculs des dérivées et nous distinguons* u *et* w *respectivement de* i *et* j*. Le terme* s_{uw} *ne dépendant de* B*, le calcul de la dérivée partielle souhaitée revient au calcul de la dérivée du produit* ${\Sigma^Y}_{uw}.\dfrac{1}{|\Sigma^Y|}$*.*
Or, la dérivée du cofacteur ${\Sigma^Y}_{uw}$ *par rapport à* b_{ik} *est,*

$$2b_{jk}\Sigma^Y_{uw,ij}$$

où $i \neq u$ *et* $j \neq w$*.*
En effet, chercher la dérivée du cofacteur ${\Sigma^Y}_{uw}$ *revient à chercher la dérivée du déterminant de la matrice* Σ^Y *privée de* u *et* w *que nous allons noter* ${\Sigma^Y}_{\boldsymbol{uw}}$*.*
Pour $u \neq i$ *et* $w \neq j$*, la dérivée du déterminant peut s'écrire comme suit,*

$$\begin{aligned}\frac{\partial|{\Sigma^Y}_{\boldsymbol{uw}}|}{\partial b_{ik}} &= \sum_{i,j}\frac{\partial{\Sigma^Y}_{\boldsymbol{uw}}}{\partial\sigma_{Y;uw,ij}}\frac{\partial\sigma_{Y;uw,ij}}{\partial b_{ik}}\\ &= \sum_{i,j}\frac{\partial{\Sigma^Y}_{\boldsymbol{uw}}}{\partial\sigma_{Y;uw,ij}}b_{jk}\\ &= \sum_j 2\Sigma^Y_{\boldsymbol{uw},ij}b_{jk}\end{aligned} \tag{A.5}$$

où $\Sigma^Y_{\boldsymbol{uw},ij}$ *est le cofacteur de* $\sigma_{Y;ij}$ *dans* $\Sigma^Y_{\boldsymbol{uw}}$*. Ce résultat est obtenu car d'après* (A.2) *la dérivée du determinant correspond au cofacteur. En effet, d'après* (A.2)*,*

$$|\Sigma^Y_{\boldsymbol{uw}}| = \sum_j \sigma_{Y;uw,ij}\Sigma^Y_{\boldsymbol{uw},ij}$$

Et puisque $\Sigma^Y_{\boldsymbol{uw}}$ *est symétrique,*

$$\Sigma^Y_{\boldsymbol{uw},ij} = \Sigma^Y_{\boldsymbol{uw},ji}$$

ainsi, pour $i \neq j$,

$$\frac{\partial|\Sigma^Y{}_{\boldsymbol{uw}}|}{\partial\sigma_{Y;uw,ij}} = 2\Sigma^Y{}_{\boldsymbol{uw},ij}.$$

Explicitons la dérivée partielle de $\frac{1}{|\Sigma^Y|}$ *par rapport à* b_{ik} *:*

$$\begin{aligned}\frac{\partial}{\partial b_{ik}}\left(\frac{1}{|\Sigma^Y|}\right) &= -\frac{\partial|\Sigma^Y|}{\partial b_{ik}}\frac{1}{|\Sigma^Y|^2}\\ &= -2\sum_j b_{jk}\frac{\Sigma^Y_{ij}}{|\Sigma^Y|^2}.\end{aligned}$$

En effet, $|\Sigma^Y| = \sum_j \sigma_{Y;ij}\Sigma^Y_{ij}$ *et comme fait en* (A.5),

$$\begin{aligned}\frac{\partial|\Sigma^Y|}{\partial b_{ik}} &= \sum_{ij}\frac{\partial|\Sigma^Y|}{\partial\sigma_{Y;ij}}\frac{\partial\sigma_{Y;ij}}{\partial b_{ik}}\\ &= \sum_j 2\Sigma^Y_{ij}b_{jk}\end{aligned}$$

En utilisant la formule de la dérivée du produit de deux fonctions, pour $u \neq i$ *et* $w \neq j$, *on en déduit le résultat,*

$$\frac{\partial tr\left(S{\Sigma^Y}^{-1}\right)}{\partial B} = 2\sum_{uwj} s_{uw}b_{jk}\Sigma^Y_{\boldsymbol{uw},ij}\frac{1}{|\Sigma^Y|} - \sum_{uw} s_{uw}\Sigma^Y_{uw}2\sum_j b_{jk}\frac{\Sigma^Y_{ij}}{|\Sigma^Y|^2}. \tag{A.6}$$

Pour obtenir une écriture plus élégante de ce résultat, nous utilisons le théorème de Jacobi d'après lequel,

$$\Sigma^Y_{\boldsymbol{uw},ij}|\Sigma^Y| = \Sigma^Y_{uw}\Sigma^Y_{ij} - \Sigma^Y_{uj}\Sigma^Y_{iw}.$$

Si Σ^Y *est non singulière, en multipliant ce résultat par* $|\Sigma^Y|^{-1}$, *et en l'injectant dans* (A.6), *il vient après simplification,*

$$\begin{aligned}\frac{\partial tr\left(S{\Sigma^Y}^{-1}\right)}{\partial b_{ik}} &= -2\sum_{juw} s_{uw}b_{jk}\Sigma^Y_{uj}\Sigma^Y_{iw}\frac{1}{|\Sigma^Y|^2}\\ &= -2\sum_{juw} s_{uw}b_{jk}\sigma^Y_{ju}\sigma^Y_{wi}\end{aligned} \tag{A.7}$$

en utilisant (A.3).

2. *Explicitons* $\frac{\partial}{\partial B}log|\Sigma^Y|$ *pour* Ψ *fixé :*

$$\begin{aligned}\frac{\partial log|\Sigma^Y|}{\partial b_{ik}} &= \frac{\partial|\Sigma^Y|}{\partial b_{ik}}\frac{1}{|\Sigma^Y|}\\ &= 2\sum_j \Sigma^Y_{ij}b_{jk}\frac{1}{|\Sigma^Y|}\\ &= 2\sum_j b_{jk}\sigma^Y_{ji}.\end{aligned} \tag{A.8}$$

Ainsi, à partir des deux dérivées (A.7) *et* (A.8)*, on en déduit,*

$$\frac{\partial F}{\partial b_{ik}} = 2\sum_j b_{jk}\sigma_Y^{ji} - 2\sum_{iuw} b_{jk}\sigma_Y^{ju}s_{uw}\sigma_Y^{wi}.$$

D'où l'écriture matricielle,

$$\begin{aligned}\frac{\partial F}{\partial b_{ik}} &= 2B'\Sigma^{Y^{-1}} - 2B'\Sigma^{Y^{-1}}S\Sigma^{Y^{-1}} \\ &= 2B'\left(I - \Sigma^{Y^{-1}}S\right)\Sigma^{Y^{-1}} \\ &= 2B'\left(\Sigma^{Y^{-1}}\Sigma^Y - \Sigma^{Y^{-1}}S\right)\Sigma^{Y^{-1}} \\ &= 2B'\Sigma^{Y^{-1}}\left(\Sigma^Y - S\right)\Sigma^{Y^{-1}} \\ &= 2\Sigma^{Y^{-1}}\left(\Sigma^Y - S\right)\Sigma^{Y^{-1}}B\end{aligned} \tag{A.9}$$

car la matrice $\Sigma^{Y^{-1}}\left(\Sigma^Y - S\right)\Sigma^{Y^{-1}}$ *est symétrique.* ■

Démonstration 2 (Démonstration du résultat (1.26))*. Nous gardons les notations introduites dans la démonstration précédente.*

1. *Explicitons* $\frac{\partial}{\partial\Psi}tr\left(S\Sigma^{Y^{-1}}\right)$ *pour B fixé :*

   Pour cela nous devons calculer la dérivée partielle $\frac{\partial}{\partial\sigma_i^2}\left(\Sigma_{ji}^Y\frac{1}{|\Sigma^Y|}\right)$*. Or, les termes* σ_i^2 *sont présents uniquement sur la diagonale de* $\Sigma^Y = BB' + \Psi$ *et sont de la forme* $\sigma_{Y;ii}^2 = b_{ik}^2 + \sigma_i^2$*. Ainsi, pour chacune des dérivée à expliciter* $\frac{\partial\Sigma_{ji}^Y}{\partial\sigma_i^2}$ *et* $\frac{\partial}{\partial\sigma_i^2}\left(\frac{1}{|\Sigma^Y|}\right)$*, il suffit de reprendre pour* $i = j$ *les calculs de la démonstration précédente concernant la dérivée de* $tr\left(S\Sigma^{Y^{-1}}\right)$ *par rapport à* b_{ik}*. On en déduit alors aisément le résultat,*

$$\frac{\partial}{\partial\sigma_i^2}tr\left(S\Sigma^{Y^{-1}}\right) = -\sum_{iuw} s_{uw}\sigma_Y^{iu}\sigma_Y^{wi}. \tag{A.10}$$

2. *Explicitons* $\frac{\partial}{\partial\Psi}log|\Sigma^Y|$ *pour B fixé :*
   De même, il suffit de reprendre les calculs de la démonstration précédente pour $i = j$ *et* $\sigma_{Y;ii}^2 = b_{ik}^2 + \sigma_i^2$*. On obtient alors,*

$$\begin{aligned}\frac{\partial}{\partial\sigma_i^2}log|\Sigma^Y| &= \sum_i \Sigma_{ij}^Y\frac{1}{|\Sigma^Y|} \\ &= \sum_i \sigma_Y^{ii}\end{aligned} \tag{A.11}$$

   et d'après (A.10) *et* (A.11)*, on en déduit,*

$$\frac{\partial F}{\partial\sigma_i^2} = \sum_i \sigma_Y^{ii} - \sum_{iuw}\sigma_Y^{iu}s_{uw}\sigma_Y^{wi}.$$

   D'où l'écriture matricielle,

$$\frac{\partial F}{\partial\Psi} = diag\left(\Sigma_Y^{-1} - \Sigma_Y^{-1}S\Sigma_Y^{-1}\right). \tag{A.12}$$

■

Démonstration 3 (Démonstration du résultat (1.28)). *Rappelons, la fonction à minimiser,*

$$F(B,\Psi) = log|\Sigma^Y| + tr\left(S{\Sigma^Y}^{-1}\right) - log|S| - q$$

Pour obtenir les formules (1.28) *des estimateurs* $\hat{B}$ *et* $\hat{\Psi}$ *(*$\hat{\mu}$ *s'obtient de manière triaviale, cf.* (1.20) *et* (1.23)*), nous allons résoudre* $\frac{\partial F}{\partial(B,\Psi)} = 0$. *Nous noterons,* $\hat{\Sigma^Y} = \hat{B}'\hat{B} + \hat{\Psi}$.

$$\left\{\begin{array}{rcl}\frac{\partial F}{\partial B} &=& 0\\ \frac{\partial F}{\partial \Psi} &=& 0\end{array}\right. \Leftrightarrow \left\{\begin{array}{rcl}\hat{B}'{\hat{\Sigma^Y}}^{-1} - \hat{B}'{\hat{\Sigma^Y}}^{-1}S{\hat{\Sigma^Y}}^{-1} &=& 0\\ diag\left({\hat{\Sigma^Y}}^{-1} - {\hat{\Sigma^Y}}^{-1}S{\hat{\Sigma^Y}}^{-1}\right) &=& 0\end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rcl}\hat{B}' - \hat{B}'{\hat{\Sigma^Y}}^{-1}S &=& 0\\ diag\left(\left(\hat{\Sigma^Y} - \hat{B}\hat{B}'\right)\left({\hat{\Sigma^Y}}^{-1} - {\hat{\Sigma^Y}}^{-1}S{\hat{\Sigma^Y}}^{-1}\right)\right) &=& 0\end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rcl}\hat{B}' - \hat{B}'{\hat{\Sigma^Y}}^{-1}S &=& 0\\ diag\left(I - S{\hat{\Sigma^Y}}^{-1} - \hat{B}\hat{B}'{\hat{\Sigma^Y}}^{-1} + \hat{B}\hat{B}'{\hat{\Sigma^Y}}^{-1}S{\hat{\Sigma^Y}}^{-1}\right) &=& 0\end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rcl}\hat{B}' - \hat{B}'{\hat{\Sigma^Y}}^{-1}S &=& 0\\ diag\left(I - S{\hat{\Sigma^Y}}^{-1} - \hat{B}\left[\hat{B}' - \hat{B}'{\hat{\Sigma^Y}}^{-1}S\right]{\hat{\Sigma^Y}}^{-1}\right) &=& 0\end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rcl}\hat{B}' - \hat{B}'{\hat{\Sigma^Y}}^{-1}S &=& 0\\ diag\left(I - S{\hat{\Sigma^Y}}^{-1}\right) &=& 0\end{array}\right. \Leftrightarrow \left\{\begin{array}{rcl}\hat{B}' - \hat{B}'{\hat{\Sigma^Y}}^{-1}S &=& 0\\ diag\left({\hat{\Sigma^Y}}^{-1} - S\right) &=& 0\end{array}\right.$$

Dans un objectif de simplification pour l'implémentation informatique des algorithmes itératifs utilisés en pratique, les identités de Lawley et Maxwell ont été utilisées (cf. (1.27)*). En voici une réécriture selon les notations choisies pour les démonstrations où en particulier* B *de dimension* $(q \times K)$ *et non* $(K \times q)$ *et où nous introduisons* $J = B'\Psi^{-1}B$.

$${\Sigma^Y}^{-1} = \Psi^{-1} - \Psi^{-1}B(I+J)^{-1}B'\Psi^{-1} \tag{A.13}$$

$$B'{\Sigma^Y}^{-1} = (I+J)^{-1}B'\Psi^{-1}. \tag{A.14}$$

En injectant (A.14) *dans la première équation du système, on obtient,*

$$\begin{array}{l}B' = (I+J)^{-1}B'\Psi^{-1}S\\ \Leftrightarrow (I+J)B' = B'\Psi^{-1}S\\ \Leftrightarrow B' + JB' = B'\Psi^{-1}S\\ \Leftrightarrow JB' = B'\Psi^{-1}S - B'\\ \Leftrightarrow B' = J^{-1}\left(B'\Psi^{-1}S - B'\right)\\ \Leftrightarrow B' = J^{-1}\left(B'\Psi^{-1}S - B'\Psi^{-1}\Psi\right)\\ \Leftrightarrow B' = J^{-1}B'\Psi^{-1}\left(S - \Psi\right)\end{array} \tag{A.15}$$

d'où,

$$\hat{B}' = J^{-1}B'\Psi^{-1}\left(S - \Psi\right). \tag{A.16}$$

Concernant la seconde équation du système,

$$\begin{aligned} & diag\left(\Sigma^Y - S\right) = 0 \\ \Leftrightarrow \quad & \forall i, \quad \sigma_{Y;ii} = s_{ii} \\ \Leftrightarrow \quad & \forall i, \quad \sigma_i^2 + \sum_{k=1}^{K} b_{ik}^2 = s_{ii} \\ \Leftrightarrow \quad & \forall i, \quad \sigma_i^2 = s_{ii} - \sum_{k=1}^{K} b_{ik}^2 \end{aligned} \tag{A.17}$$

ce qui est équivalent à l'écriture matricielle,

$$\hat{\Psi} = diag\left(S - \hat{B}\hat{B}'\right). \tag{A.18}$$

■

ANNEXE B

Compléments à la deuxième étape de l'algorithme de Jöreskog

Pour compléter l'algorithme proposé par Jöreskog, cette annexe détaille les étapes 2.(a) et 2.(b) c'est à dire, le calcul du gradient e et les formules utilisées pour la matrice information de Fisher $E\left[\left(\frac{\partial F}{\partial \Psi_{ii}}\frac{\partial F}{\partial \Psi_{jj}}\right)_{\Psi=\Psi^{[t]}}\right]$. *Décrivons d'abord la procédure de calcul itérative du gradient e de l'étape 2.(b). La procédure qui va suivre est une version simplifiée de celle présentée par Jöreskog (1969). En effet, dans cette dernière un paramètre* Φ *supplémentaire est présent dans l'ensemble des paramètres* θ. *J'ai fait le choix de ne pas le présenter car souvent dans le cadre des hypothèses il est supposé contraint à la matrice identité. Et c'est sous cette hypothèse la que je présente la procédure itérative :*
Tout d'abord, lors d'une étape préliminaire, la valeur de $log|S|+q$ *est calculée et stockée pour être utilisée à chacune des itérations où le calcul de la valeur de* F *est fait.*

1. *Calcul de :*
   - $J = B'\Psi^{-1}B$;
   - $I_K + J$;
   - $\boldsymbol{A} = (I_K + J)^{-1}$ *qui est facile à calculer car* $I_K + J$ *est diagonale.*
2. *Calcul de :*
$$det\left(\Sigma^Y\right) = \left(\prod_{i=1}^{q}\Psi_{ii}\right)\left(det\left(I+J\right)\right); \tag{B.1}$$
3. *Calcul de* $\boldsymbol{B} = \Psi^{-1} - \Psi^{-1}B\boldsymbol{A}B'\Psi^{-1} = {\Sigma^Y}^{-1}$;
4. *Calcul de* $\boldsymbol{C} = S\boldsymbol{B} = S{\Sigma^Y}^{-1}$ *et* $tr\left(\boldsymbol{C}\right) = tr\left(S\left({\Sigma^Y}^{-1}\right)\right)$;
5. *Calcul de la valeur courante de la fonction* F *en utilisant les étapes 2., 4. et l'étape préliminaire.*
6. *Calcul de* $\boldsymbol{D} = \boldsymbol{B} - \boldsymbol{B}\boldsymbol{C} = {\Sigma^Y}^{-1}\left(\Sigma^Y - S\right){\Sigma^Y}^{-1}$;
7. *Calcul de* $\boldsymbol{E} = \boldsymbol{D}B$;
8. *Calcul du gradient :*
$$\begin{aligned} e = \frac{\partial F}{\partial \theta} &= \left(\frac{\partial F}{\partial B}, \frac{\partial F}{\partial \Psi}\right) \\ &= \left(2\boldsymbol{E},\ \boldsymbol{diag(E)}\right). \end{aligned}$$

L'équation (B.1) *se retrouve facilement avec la formule suivante des identités de Maxwell* (1.27) *:*

$$\begin{aligned} {\Sigma^Y}^{-1} B' &= \Psi^{-1} B' (I + B\Psi^{-1} B')^{-1} \\ \Rightarrow det\left({\Sigma^Y}^{-1}\right) det\left(B'\right) &= det\left(\Psi^{-1}\right) det\left(B'\right) det\left(\boldsymbol{A}\right) \\ \Leftrightarrow det\left(\Sigma^Y\right) &= det\left(\Psi\right) det\left(\boldsymbol{A}^{-1}\right) \\ \Leftrightarrow det\left(\Sigma^Y\right) &= tr\left(\Psi\right) det\left(I_K + J\right). \end{aligned}$$

Pour finir, décrivons les formules utilisées lors de l'étape 2.(a) décrivant la procédure de calcul itérative de la matrice information de Fisher $E\left[\left(\frac{\partial F}{\partial \Psi_{ii}} \frac{\partial F}{\partial \Psi_{jj}}\right)_{\Psi=\Psi^{[t]}}\right]$*. Ces formules ont été calculées par Lawley (1967) en ne tenant compte ni de la structure de covariance, ni de la loi de Wishart que suit la matrice* S*. De manière indépendante, Lockhart (1967) a lui aussi calculé ces formules mais nous ne présenterons que celles de Lawley que nous limiterons aux différents paramètres de* θ *privé de* Φ *pour les raisons évoquées plus haut :*

1. $E\left[\frac{\partial^2 F}{\partial b_{ir} \partial b_{js}}\right] = 2\left(\sigma^{ij}\gamma_{rs} + \eta_{is}\eta_{jr}\right)$;
2. $E\left[\frac{\partial^2 F}{\partial b_{ir} \partial \Psi_{js}}\right] = 2\sigma^{ij}\eta_{jr}$;
3. $E\left[\frac{\partial^2 F}{\partial \Psi_{ii} \partial \Psi_{jj}}\right] = \left(\sigma^{ij}\right)^2$.

Où $\eta = {\Sigma^Y}^{-1} B$ *de dimension* $(q \times K)$ *et* $\gamma = B' {\Sigma^Y}^{-1} B$ *de dimension* $(K \times K)$.

ANNEXE C

Questionnaire EORTC QLQ-C30 spécifique au cancer

QUESTIONNAIRE SUR LA QUALITE DE VIE
EORTC QLQ-C30 version 3

Nous nous intéressons à vous et à votre santé. Répondez vous-même à toutes les questions en entourant le chiffre qui correspond le mieux à votre situation. Il n'y a pas de "bonne" ou de "mauvaise" réponse. Ces informations sont strictement confidentielles.

Vos initiales : .....

Date de naissance : .....

La date d'aujourd'hui : .....

Au cours de la semaine passée	**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
1. Avez-vous des difficultés à faire certains efforts physiques pénibles comme porter un sac à provision chargé ou une valise ?	1	2	3	4
2. Avez-vous des difficultés à faire une LONGUE promenade ?	1	2	3	4
3. Avez-vous des difficultés à faire un PETIT tour dehors ?	1	2	3	4
4. Etes-vous obligée de rester au lit ou dans un fauteuil la majeure partie de la journée ?	1	2	3	4
5. Avez-vous besoin d'aide pour manger, vous habiller, faire votre toilette ou aller aux W.C. ?	1	2	3	4
6. Etes-vous limitée d'une manière ou d'une autre pour accomplir, soit votre travail, soit vos tâches habituelles chez vous ?	1	2	3	4
7. Etes-vous totalement incapable de travailler ou d'accomplir des tâches habituelles chez vous ?	1	2	3	4

Au cours de la semaine passée	**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
8. Avez-vous eu le souffle court ?	1	2	3	4
9. Avez-vous eu mal ?	1	2	3	4
10. Avez-vous eu besoin de repos ?	1	2	3	4
11. Avez-vous eu des difficultés pour dormir ?	1	2	3	4
12. Vous êtes-vous sentie faible ?	1	2	3	4
13. Avez-vous manqué d'appétit ?	1	2	3	4

14. Avez-vous eu des nausées (mal au cœur) ?	1	2	3	4
15. Avez-vous vomi ?	1	2	3	4
16. Avez-vous été constipée ?	1	2	3	4
Au cours de la semaine passée	**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
17. Avez-vous eu de la diarrhée ?	1	2	3	4
18. Etiez-vous fatiguée ?	1	2	3	4
19. Des douleurs ont-elles perturbé vos activités quotidiennes ?	1	2	3	4
20. Avez-vous eu des difficultés à vous concentrer sur certaines choses par exemple pour lire le journal ou regarder la télévision ?	1	2	3	4
21. Vous êtes-vous sentie tendue ?	1	2	3	4
22. Vous êtes-vous fait du souci ?	1	2	3	4
23. Vous êtes vous sentie irritable ?	1	2	3	4
24. Vous êtes vous sentie déprimée ?	1	2	3	4
25. Avez-vous eu des difficultés pour vous souvenir de certaines choses ?	1	2	3	4
26. Votre état physique ou votre traitement médical vous ont-ils gênée dans votre vie FAMILIALE ?	1	2	3	4
27. Votre état physique ou votre traitement médical vous ont-ils gênée dans vos activités SOCIALES (par exemple sortir avec des amis, aller au cinéma…)	1	2	3	4
28. Votre état physique ou votre traitement médical vous ont-ils causé des problèmes financiers ?	1	2	3	4

POUR LES QUESTIONS SUIVANTES, VEUILLEZ REPONDRE EN ENTOURANT LE CHIFFRE ENTRE 1 ET 7 QUI S'APPLIQUE LE MIEUX A VOTRE SITUATION.

29. Comment évalueriez-vous l'ensemble de votre ETAT PHYSIQUE au cours de la semaine passée ?

1	2	3	4	5	6	7
Très mauvais						Excellent

30. Comment évalueriez-vous l'ensemble de votre QUALITE DE VIE au cours de la semaine passée ?

ANNEXE D

Questionnaire EORTC QLQ-BR23 spécifique au cancer du sein

FRENCH

EORTC QLQ - BR23

Les patientes rapportent parfois les symptômes ou problèmes suivants. Pourriez-vous indiquer, s'il vous plaît, si, durant la semaine passée, vous avez été affectée par l'un de ces symptômes ou problèmes. Entourez, s'il vous plaît, le chiffre qui correspond le mieux à votre situation.

Au cours de la semaine passée:	**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
31. Avez-vous eu la bouche sèche?	1	2	3	4
32. La nourriture et la boisson avaient-elles un goût inhabituel?	1	2	3	4
33. Est-ce que vos yeux étaient irrités, larmoyants ou douloureux?	1	2	3	4
34. Avez-vous perdu des cheveux?	1	2	3	4
35. Répondez à cette question uniquement si vous avez perdu des cheveux : La perte de vos cheveux vous a-t-elle contrariée?	1	2	3	4
36. Vous êtes-vous sentie malade ou souffrante?	1	2	3	4
37. Avez-vous eu des bouffées de chaleur?	1	2	3	4
38. Avez-vous eu mal à la tête?	1	2	3	4
39. Vous êtes-vous sentie moins attirante du fait de votre maladie ou de votre traitement?	1	2	3	4
40. Vous êtes vous sentie moins féminine du fait de votre maladie ou de votre traitement?	1	2	3	4
41. Avez-vous trouvé difficile de vous regarder nue?	1	2	3	4
42. Votre corps vous a-t-il déplu?	1	2	3	4
43. Vous êtes vous inquiétée de votre santé pour l'avenir?	1	2	3	4
Au cours des quatre dernières semaines:	**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
44. Dans quelle mesure vous êtes-vous intéressée à la sexualité?	1	2	3	4
45. Avez-vous eu une activité sexuelle quelconque (avec ou sans rapport)?	1	2	3	4
46. Répondez à cette question uniquement si vous avez eu une activité sexuelle: Dans quelle mesure l'activité sexuelle vous a-t-elle procuré du plaisir?	1	2	3	4

Passez à la page suivante S.V.P.

Au cours de la semaine passée:	Pas du tout	Un peu	Assez	Beaucoup
47. Avez-vous eu mal au bras ou à l'épaule?	1	2	3	4
48. Avez-vous eu la main ou le bras enflé?	1	2	3	4
49. Avez-vous eu du mal à lever le bras devant vous ou sur le côté?	1	2	3	4
50. Avez-vous ressenti des douleurs dans la région du sein traité?	1	2	3	4
51. La région de votre sein traité était-elle enflée?	1	2	3	4
52. La région de votre sein traité était-elle particulièrement sensible?	1	2	3	4
53. Avez-vous eu des problèmes de peau dans la région de votre sein traité (démangeaisons, peau qui pèle, peau sèche)?	1	2	3	4

Bibliographie

Anderson, J. C. and Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, 49(2) :155–173. 29

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, page 1. 38

Andrade, D. t. F. and Helms, R. W. (1984). Maximum likelihood estimates in the multivariate normal with patterned mean and covariance via the em algorithm. *Communications in Statistics - Theory and Methods*, 13(18) :2239–2251. iv

Arbuckle, J. L., Marcoulides, G. A., and Schumacker, R. E. (1996). Full information estimation in the presence of incomplete data. *Advanced structural equation modeling : Issues and techniques*, 243 :277. iv, 32, 35

Azria, D., Belkacemi, Y., Romieu, G., Gourgou, S., Gutowski, M., Zaman, K., Moscardo, C. L., Lemanski, C., Coelho, M., Rosenstein, B., et al. (2010). Concurrent or sequential adjuvant letrozole and radiotherapy after conservative surgery for early-stage breast cancer (co-ho-rt) : a phase 2 randomised trial. *The lancet oncology*, 11(3) :258–265. 89, 99

Bacher, F. (1987). LES MODÈLES STRUCTURAUX EN PSYCHOLOGIE PRÉSENTATION D'UN MODÈLE : LISREL Première partie. *Le Travail Humain*, 50(4) :347–370. iii

Bacher, F. (1988). LES MODÈLES STRUCTURAUX EN PSYCHOLOGIE PRÉSENTATION D'UN MODÈLE : LISREL. *Le Travail Humain*, 51(4) :273–288.

Barbieri, A., Tami, M., Bry, X., Azria, D., Gourgou, S., Bascoul-Mollevi, C., and Lavergne, C. (2016). EM algorithm estimation of a structural equation model for the longitudinal study of the quality of life. *soumis pour publication.*

Bargmann, R. (1957). Study of independence and dependence in multivariate normal analysis. 19, 20

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology. General Section*, 28(1) :97–104. 32

Bartlett, M. S. (1938). Methods of estimating mental factors. *Nature*, 141(3570) :609–610. 32

Beitz, J., Gnecco, C., and Justice, R. (1995). Quality-of-life end points in cancer clinical trials : the us food and drug administration perspective. *Journal of the National Cancer Institute. Monographs*, (20) :7–9. 92

Betzin, J. and Henseler, J. (2005). Looking at the antecedents of perceived switching costs. a pls path modeling approach with categorical indicators. 18

Bollen, K. A. (1995). Structural equation models that are nonlinear in latent variables : A least-squares estimator. *Sociological methodology*, 25 :223–252. 32, 39

Bollen, K. A. (1996). An alternative two stage least squares (2sls) estimator for latent variable equations. *Psychometrika*, 61(1) :109–121. 32, 39

Bollen, K. A. (2014). *Structural Equations with Latent Variables*. John Wiley & Sons. 4, 14, 28

Brown, R. L. (1994). Efficacy of the indirect approach for estimating structural equation models with missing data : A comparison of five methods. *Structural Equation Modeling : A Multidisciplinary Journal*, 1(4) :287–316. 35

Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika*, 33(3) :267–334. 20

Bry, X. (2003). Une méthode d'estimation empirique d'un modèle à variables latentes : l'analyse en composantes thématiques. *Revue de statistique appliquée*, 51(2) :5–45. 18

Bry, X., Lavergne, C., and Tami, M. (2016). EM estimation of a structural equation model. *soumis pour publication*.

Bry, X., Redont, P., Verron, T., and Cazes, P. (2012). THEME-SEER : a multidimensional exploratory technique to analyze a structural model using an extended covariance criterion. *J. Chemometrics*, 26(5) :158–169. :$_2$012

Bry, X., Trottier, C., Verron, T., and Mortier, F. (2013). Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119 :47–60.

Bry, X. and Verron, T. (2015). THEME : THEmatic Model Exploration through Multiple Co-Structure maximization. *Journal of Chemometrics*, 29(12) :637–647. :$_2$015

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., and Kirby, J. B. (2001). Improper solutions in structural equation models causes, consequences, and strategies. *Sociological Methods & Research*, 29(4) :468–508. 29

Chin, W. W., Marcolin, B. L., and Newsted, P. R. (2003). A partial least squares latent variable modeling approach for measuring interaction effects : Results from a monte carlo simulation study and an electronic-mail emotion/adoption study. *Information systems research*, 14(2) :189–217. 39

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38. iii, 34

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in Covariance Components Models. *Journal of the American Statistical Association*, 76(374) :341–353. iv

Derquenne, C. (2005). Generalized path modeling based on the partial maximum likelihood approach. In *PLS and Related Methods, proceedings of the PLS'05 International Symposium, Barcelona. Decisia*, pages 159–166. 18

Dijkstra, T. K. (1981). *Latent variables in linear stochastic models.* PhD thesis, University of Groningen. 16, 17

Emmett, W. (1949). Factor analysis by lawley's method of maximum likelihood. *British Journal of Statistical Psychology*, 2(2) :90–97. 20, 21

Esposito Vinzi, V. and Trinchera, L. (2014). Modèles à équations structurelles, approches basées sur les composantes.

Fayers, P., Aaronson, N., Bjordal, K., Groenveld, M., Curran, D., and Bottomley, A. (2001). The eortc qlq-c30 scoring manual published by : European organisation for research and treatment of cancer. *Brussels, Belgium.* 98

Fayolle, A., Engelbrecht, B., Freycon, V., Mortier, F., Swaine, M., Réjou-Méchain, M., Doucet, J.-L., Fauvet, N., Cornu, G., and Gourlet-Fleury, S. (2012). Geological Substrates Shape Tree Species and Trait Distributions in African Moist Forests. *PLoS ONE*, 7(8).

Fiteni, F., Pam, A., Anota, A., Vernerey, D., Paget-Bailly, S., Westeel, V., and Bonnetain, F. (2015). Health-related quality-of-life as co-primary endpoint in randomized clinical trials in oncology. *Expert review of anticancer therapy*, 15(8) :885–891. 92

Fletcher, R. and Powell, M. J. (1963). A rapidly convergent descent method for minimization. *The Computer Journal*, 6(2) :163–168. 23, 24

Ford, B. L. (1983). An overview of hot-deck procedures. *Incomplete data in sample surveys*, 2(Part IV) :185–207. 34

Foulley, J.-L. (2002). Algorithme EM : Théorie et application au modèle mixte. *Journal de la Société française de statistique*, 143(3-4) :57–109.

Gold, M. S. and Bentler, P. M. (2000). Treatments of missing data : A monte carlo comparison of rbhdi, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, 7(3) :319–355. 33, 35

Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable. part ia modified latent structure approach. *American Journal of Sociology*, pages 1179–1259. 32

Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2) :215–231. 32

Hanafi, M. (2004). Approche pls : une hiérarchie des stratégies pour la détermination des variables latentes. *Actes des 36èmes journées de statistique de la SFDS-Montpellier*, 26. 17

Hayduk, L. A. (1988). *Structural equation modeling with LISREL : Essentials and advances.* Jhu Press. 39

Henseler, J. (2010). On the convergence of the partial least squares path modeling algorithm. *Computational statistics*, 25(1) :107–120. 17

Howe, W. G. (1955). Some contributions to factor analysis. Technical report, Oak Ridge National Lab., Tenn. 19, 20, 22, 23

Husson, F., Josse, J., and Lê, S. (2008). FactoMineR : An R Package for Multivariate Analysis. *Journal of Statistical Software*, page 18. :$_2$008

Hwang, H. and Takane, Y. (2004). Generalized structured component analysis. *Psychometrika*, 69(1) :81–99.

Jaccard, J. and Turrisi, R. (2003). *Interaction effects in multiple regression*. Number 72. Sage.

Jaccard, J. and Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression : Multiple indicator and structural equation approaches. *Psychological bulletin*, 117(2) :348. 39

Jaccard, J. and Wan, C. K. (1996). *LISREL approaches to interaction effects in multiple regression*. Number 114. Sage. 39

Jaccard, J., Wan, C. K., and Turrisi, R. (1990). The detection and interpretation of interaction effects between continuous variables in multiple regression. *Multivariate Behavioral Research*, 25(4) :467–478. 39

Jakobowicz, E. (2007). *Contributions aux modèles d'équations structurelles à variables latentes*. phdthesis, Conservatoire national des arts et metiers - CNAM. iii, 16, 17, 18, 29

Jakobowicz, E. and Derquenne, C. (2007). A modified pls path modeling algorithm handling reflective categorical variables and a new model building strategy. *Computational Statistics & Data Analysis*, 51(8) :3666–3678. 18

Jakobowicz, E. and Saporta, G. (2007). A nonlinear pls path modeling based on monotonic b-spline transformations. In *Causalities explored by indirect observations-PLSt'07, 5th International symposium on PLS and related methods, Oslo, septembre*, volume 172. 32

Jordan, M. I. (2003). *An introduction to probabilistic graphical models*. preparation.

Jöreskog, K. and Yang, F. (1997). Estimation of interaction models using the augmented moment matrix : Comparison of asymptotic standard errors. *SoftStat*, 97 :467–478. 32, 39

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika*, 31(2) :165–178. 23

Joreskog, K. G. and Sorbom, D. (1996). Lisrel 8 : User's reference guide. chicago : Scientific software international. *Inc Chicago : Scientific Software International*, 4 :43. 19, 32

Jöreskog, K. G., Yang, F., Marcoulides, G., and Schumacker, R. (1996). Nonlinear structural equation models : The kenny-judd model with interaction effects. *Advanced structural equation modeling : Issues and techniques*, pages 57–88. 39

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4) :443–482. 13, 20, 21, 23, 24

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2) :183–202. 13, 20, 129

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, 57(2) :239–251. iii, 24

Jöreskog, K. G. (2000). Latent Variable Scores and their uses. Scientific Software International. iii, 6, 29, 32, 37, 38, 50, 54

Jöreskog, K. G. and Sörbom, D. (1982). Recent Developments in Structural Equation Modeling. *Journal of Marketing Research*, 19(4) :404–416. iii

Kaplan, D. (2008). *Structural Equation Modeling : Foundations and Extensions : Foundations and Extensions*, volume 10. Sage Publications. 4

Karl G. Jöreskog, H. W. (1982). The ML and PLS techniques for modeling with latent variables. Historical and comparative aspects. 1 :263–270. iii, 29

Kelley, T. L. (1928). *Crossroads in the Mind of Man*. Stanford University Press. 2

Kenny, D. A. and Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological bulletin*, 96(1) :201. 39

Klein, A. and Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the lms method. *Psychometrika*, 65(4) :457–474. 32, 39

Knott, M. and Bartholomew, D. J. (1999). *Latent variable models and factor analysis*. Number 7. Edward Arnold. 4

Krämer, N. (2005). Nonlinear partial least squares path models. In *3rd World conference of the IASC, Cyprus*, volume 99, page 100. 17

Lafaye de Micheaux, P., Drouilhet, R., and Liquet, B. (2011). *Le logiciel R : Maîtriser le langage - Effectuer des analyses statistiques*. Springer Science & Business Media.

Lawley, D. (1942). Further investigations in factor estimation. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences*, 61(02) :176–185. 19, 20

Lawley, D. (1943). The application of the maximum likelihood method to factor analysis. *British Journal of Psychology. General Section*, 33(3) :172–175. 19, 20

Lawley, D. (1967). Some new results in maximum likelihood factor analysis. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences*, 67(04) :256–264. 19, 130

Lawley, D. and Maxwell, A. (1963). *Factor analysis as a statistical method*. London : Butterworths. 9, 20, 21, 37

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, 60(01) :64–82. 19

Lazarsfeld, P. (1950). The logical and mathematical foundation of latent structure analysis. *In Studies in social psychology in World War II*, IV :362–412. 32

Lazarsfeld, P. (1954). A conceptual introduction to latent structure analysis. in mathematical thinking in the social sciences. *In Mathematical thinking in the social sciences*, page 349–387. 32

Lee, S.-Y. and Tang, N.-S. (2006). Bayesian Analysis of Nonlinear Structural Equation Models with Nonignorable Missing Data. *Psychometrika*, 71(3) :541–564. iv, 36

Lockhart, R. (1967). Asymptotic sampling variances for factor analytic models identified by specified zero parameters. *Psychometrika*, 32(3) :265–277. 130

Lohmöller, J.-B. (2013). *Latent Variable Path Modeling with Partial Least Squares*. Springer Science & Business Media. 13, 15, 18

Lyttkens, E., Areskoug, B., and Wold, H. (1975). The convergence of nipals estimation procedures for six path models with one or two latent variables. *Rapport technique, University of Göteborg*, 23. 16

Marcoulides, G. A. and Moustaki, I. (2014). *Latent Variable and Latent Structure Models*. Psychology Press.

McDonald, R. P. (1996). Path Analysis with Composite Variables. *Multivariate Behavioral Research*, 31(2) :239–270. iii, 28, 32, 37

Moosbrugger, H., Klein, A., Frank, D., and Schermelleh-Engel, K. (1993). On estimating parameters of latent moderator effects in structural equation models. *Arbeiten aus dem Institut f ur Psychologie der JW Goethe-Universit at*, (11). 39

Moosbrugger, H., Klein, A., Frank, D., and Schermelleh-Engel, K. (1996). Zum problem der sch atzung von latenten moderatore ekten (on the problem of estimating latent moderator effects). *Methodische Grundlagen und Anwendungen von Strukturgleichungsmodellen*, 2 :5–35. 39

Moosbrugger, H., Schermelleh-Engel, K., and Klein, A. (1997). Methodological problems of estimating latent interaction effects. *Methods of Psychological Research Online*, 2(2) :95–111. iv, 32, 39

Mortier, F., Trottier, C., Cornu, G., and Bry, X. (2014). SCGLR-An R Package for Supervised Component Generalized Linear Regression. *Journal of Statistical Software*. 57

Muthén, B. O. (1989). Tobit factor analysis†. *British journal of mathematical and statistical psychology*, 42(2) :241–250. 32, 36

Muthén, B., Kaplan, D., and Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52(3) :431–462. iv, 32, 33, 36

Muthén, L. and Muthén, B. (1998). MPlus User's Guide. iv

Noonan, R. and Wold, H. (1977). NIPALS Path Modelling with Latent Variables. *Scandinavian Journal of Educational Research*, 21(1) :33–61.

Olinsky, A., Chen, S., and Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research*, 151(1) :53–79. 35, 36

Ping Jr, R. A. (1996a). Latent variable interaction and quadratic effect estimation : A two-step technique using structural equation analysis. *Psychological Bulletin*, 119(1) :166. 39

Ping Jr, R. A. (1996b). Latent variable regression : A technique for estimating interaction and quadratic coefficients. *Multivariate Behavioral Research*, 31(1) :95–120. 39

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika*, 20(2) :93–111. 19, 20

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. 34

Saidane, M. (2006). *Modèles à Facteurs Conditionnellement Hétéroscédastiques et à Structure Markovienne Cachée pour les Séries Financières*. phdthesis, Université Montpellier II - Sciences et Techniques du Languedoc. 52, 87

Saltelli, A., Chan, K., Scott, E. M., et al. (2000). *Sensitivity analysis*, volume 1. Wiley New York. 57

Saporta, G. (2006). Probabilités, analyse des données et statistique.

Schermelleh-Engel, K., Klein, A., and Moosbrugger, H. (1998). Estimating nonlinear effects using a latent moderated structural equations approach. 39

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2) :201–292. 2

Stan, V. and Saporta, G. (2006). Une comparaison expérimentale entre les approches PLS et LISREL. In *38 èmes Journées de Statistique, Clamart, France*, X, France. iii, 29

Tami, M. (2016). A comparison of pls, lisrel and em approach. *en préparation*.

Tami, M., Baribieri, A., Bry, X., Bascoul-Mollevi, C., and Lavergne, C. (2014a). Analyse longitudinale de la qualité de vie relative à la santé en cancérologie par équation structurelle et modèles mixtes. In *JDS 2016 Montpellier*, Montpellier, France.

Tami, M., Bry, X., and Lavergne, C. (2014b). EM estimation of a structural equation model. In *CASI 2016* , Limerick, Irlande.

Tami, M., Bry, X., and Lavergne, C. (2014c). Estimation of structural equation models with factors by EM algorithm. In *JDS 2014 Renes*, Rennes, France.

Tang, M.-L. and Lee, S.-Y. (1998). Analysis of structural equation models with censored or truncated data via EM algorithm. *Computational statistics & data analysis*, 27(1) :33–46. iv, 36

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2). 18

Tenenhaus, M. (1999). L'approche pls. *Revue de statistique appliquée*, 47(2) :5–40. 17

Tenenhaus, M. (2007). A bridge between pls path modelling and uls-sem. In *Proceedings of the International Symposium PLS'07, Aas, Norvège*, volume 14, pages 42–43. iii, 32

Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M., and Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, 48(1) :159–205. 15, 29

Thomson, G. H. (1948). *The factorial analysis of human ability*. University of London Press. 32

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, 38(5) :406. 2

Verleye, G. (1997). Missing at random data problems and maximum likelihood structural equation modelling. 36

Vinzi, V. E., Chin, W. W., Henseler, J., and Wang, H. (2010). *Handbook of Partial Least Squares : Concepts, Methods and Applications*. Springer Science & Business Media.

W. W. Chin, P. R. N. (1999). Structural equation modeling analysis with small samples using partial least squares.

Wangen, L. E. and Kowalski, B. R. (1989). A multiblock partial least squares algorithm for investigating complex chemical systems. *J. Chemometrics*, 3(1) :3–20.

WHO (1948). Who constitution. *Geneva : WHO*. 92

Wold, H. (1966). Estimation of Principal Components and Related Models by Iterative Least squares. In *Multivariate Analysis.*, pages 391–420. Academic Press. iii, 13

Wold, H. (1973). Nonlinear iterative partial least squares (nipals) modelling : Some current developments. In *Multivariate Analysis*, volume III, pages 383–407. Academic Press. 13

Wold, H. (1980). Model construction and evaluation when theoretical knowledge is scarce. In *Evaluation of econometric models*, pages 47–74. Academic Press.

Wold, H. (1982). Soft modelling : the basic design and some extensions. *Systems under indirect observation, Part II*, pages 36–37. 13, 15, 16, 17

Wold, H. (1985). Partial Least Squares. In *Encyclopedia of Statistical Sciences*, volume 6, pages 581–591. John Wiley & Sons, New York. iii, 13, 18

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. *SAGE FOCUS EDITIONS*, 154 :256–256. 29

Yang Jonsson, F. (1997). Non-linear structural equation models : Simulation studies of the kenny-judd model. 39

Approche EM pour modèles multi-blocs à facteurs à une équation structurelle

Myriam Tami

Résumé : Les modèles d'équations structurelles à variables latentes permettent de modéliser des relations entre des variables observables et non observables. Les deux paradigmes actuels d'estimation de ces modèles sont les méthodes de moindres carrés partiels sur composantes et l'analyse de la structure de covariance. Dans ce travail, après avoir décrit les deux principales méthodes d'estimation que sont PLS-PM et LISREL, nous proposons une approche d'estimation fondée sur la maximisation par algorithme EM de la vraisemblance globale d'un modèle à facteurs latents et à une équation structurelle. Nous en étudions les performances sur des données simulées et nous montrons, via une application sur des données réelles environnementales, comment construire pratiquement un modèle et en évaluer la qualité. Enfin, nous appliquons l'approche développée dans le contexte d'un essai clinique en cancérologie pour l'étude de données longitudinales de qualité de vie. Nous montrons que par la réduction efficace de la dimension des données, l'approche EM simplifie l'analyse longitudinale de la qualité de vie en évitant les tests multiples. Ainsi, elle contribue à faciliter l'évaluation du bénéfice clinique d'un traitement.

Mots-clés : Analyse de données, méthodes d'estimation, modèles à équations structurelles, modèles à facteurs, variables latentes, algorithme EM.

EM estimation of a structural equation model

Myriam Tami

Abstract: Structural equation models enable the modeling of interactions between observed variables and latent ones. The two leading estimation methods are partial least squares on components and covariance-structure analysis. In this work, we first describe the PLS-PM and LISREL methods and, then, we propose an estimation method using the EM algorithm in order to maximize the likelihood of a structural equation model with latent factors. Through a simulation study, we investigate how fast and accurate the method is, and thanks to an application to real environmental data, we show how one can handly construct a model or evaluate its quality. Finally, in the context of oncology, we apply the EM approach on health-related quality-of-life data. We show that it simplifies the longitudinal analysis of quality-of-life and helps evaluating the clinical benefit of a treatment.

Key words: Data analysis, estimation methods, structural equation models, factors models, latent variables, EM algorithm.