Contributions aux modèles d'équations structurelles à variables latentes

Emmanuel Jakobowicz

CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS
PARIS

Thèse

Pour obtenir le grade de

DOCTEUR DU CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS

Spécialité : INFORMATIQUE

Emmanuel Jakobowicz

CONTRIBUTIONS AUX MODÈLES D'ÉQUATIONS STRUCTURELLES À VARIABLES LATENTES

Soutenue publiquement le 22 octobre 2007

Jury :

M. Gilbert Saporta	Professeur, CNAM	Directeur de thèse
M. Pierre Cazes	Professeur, Université Paris-Dauphine	Rapporteur
M. Vincenzo Esposito Vinzi	Professeur, Université Federico II, Naples, Italie	Rapporteur
M. Gérard d'Aubigny	Professeur, Université de Grenoble	Examinateur
M. Christian Derquenne	Docteur, EDF Recherche et Développement	Examinateur
M. Pierre-Louis Gonzalez	Maître de Conférences, CNAM	Examinateur
M. El Mostafa Qannari	Professeur, INRA - ENITIAA, Nantes	Examinateur
M. Michel Tenenhaus	Professeur, Groupe HEC	Examinateur

Contributions aux modèles d'équations structurelles à variables latentes

Résumé

Les modèles d'équations structurelles à variables latentes constituent des modèles statistiques complexes permettant de mettre en relation des concepts non observables. Les deux méthodes d'estimation de ces modèles, que sont, d'une part, l'analyse de la structure de covariance (méthode LISREL) et, d'autre part, l'approche PLS, offrent des solutions à la fois concurrentes et complémentaires. Dans ce travail, un certain nombre de questionnements liés à ce type de modèles et de méthodes sont abordés aussi bien d'un point de vue théorique qu'empirique. Nous étudions la construction du modèle initial par le biais d'algorithmes itératifs, au niveau des relations du modèle de mesure et du modèle structurel. Nous présentons des tests basés sur des permutations afin de comparer des échantillons non appariés sur un modèle donné. Des transformations dites optimales des variables sont utilisées afin de rechercher des relations non linéaires. Finalement, des méthodes permettant le traitement de données manquantes spécifiques induites par des filtres sont décrites. Pour chaque cas, la théorie existante est présentée et des approches novatrices sont introduites. Nous appliquons l'ensemble de ces méthodes dans le cadre de l'analyse de la satisfaction et de la fidélité des clients sur des données provenant d'Electricité de France.

Mots clés : *Analyse de données, statistiques, modèles d'équations structurelles, variables latentes, approche PLS, LISREL, analyse de la structure de covariance, satisfaction, fidélité.*

Contributions to path modeling with latent variables

Abstract

Path models with latent variables are complex statistical models able to interpret interactions between unobservable concepts. The two leading estimation methods, covariance structure analysis (LISREL) and partial least squares path modeling (PLS-PM), appear complementary rather than competitive. In this work, we first introduce models and methods related to structural equations modeling and, then, we answer several questions on path models with latent variables. We present algorithms to construct the initial measurement and structural models. We introduce methods based on permutation tests for multi-group comparison. Approaches based on variables transformation to find nonlinear relationships are exposed. Finally, methods to handle specific cases of missing data patterns are presented. We first focus on existing theories and, then, introduce new approaches. Finally, the new methods are applied within the framework of customer satisfaction and loyalty to the French electricity sector.

Keywords : *Data analysis, statistics, path analysis, latent variables, structural equations models, partial least squares path modeling (PLS-PM), LISREL, covariance structure analysis, satisfaction, loyalty.*

Equipe méthodes statistiques de *data-mining* et apprentissage - Centre d'études et de recherches en informatique de CNAM
Conservatoire National des Arts et Métiers - 292 rue Saint Martin 75141 Paris Cedex 03

Remerciements

Je tiens à remercier tout spécialement mon directeur de thèse, Gilbert Saporta, pour sa confiance, ses conseils avisés et le temps qu'il m'a accordé. Mes remerciements vont aussi à Christian Derquenne qui m'a accueilli chaleureusement au sein d'EDF et m'a permis d'effectuer cette thèse dans des conditions idéales. Je tiens, par ailleurs, à remercier les deux autres membres de mon comité de thèse, Pierre-Louis Gonzalez et Silvia Squillacciotti pour leurs conseils.

Je tiens à exprimer ma reconnaissance à l'ensemble des membres de mon jury : Pierre Cazes et Vincenzo Esposito Vinzi pour avoir accepté d'être rapporteurs, Michel Tenenhaus qui, par ses relectures attentives, m'a permis d'éviter de tomber dans les pièges de l'approche PLS, Gérard d'Aubigny, Pierre-Louis Gonzalez et El Mostafa Qannari qui ont consenti à être membres de mon jury.

Je remercie l'ensemble du groupe SOAD d'EDF recherche et développement pour l'ambiance sympathique et chaleureuse ainsi que l'équipe de la chaire de statistique appliquée du CNAM et tout spécialement Ndeye Niang pour son soutien.

Un grand merci au professeur Morgenthaler de l'Ecole Polytechnique Fédérale de Lausanne sans qui je n'aurais pas pu rencontrer le professeur Saporta.

Finalement, mes remerciements les plus sincères vont à celle sans qui cette thèse n'aurait même pas pu avoir lieu et qui m'a épaulé chaque jour de ces trois années : ma femme adorée.

Table des matières

Introduction générale **13**

1 Les modèles d'équations structurelles à variables latentes **17**
1.1 La conceptualisation . . . 17
1.1.1 Préliminaires . . . 17
1.1.2 Formalisme . . . 19
1.2 L'approche PLS . . . 20
1.2.1 Le modèle externe . . . 20
1.2.2 Le modèle interne . . . 21
1.2.3 La représentation graphique . . . 21
1.2.4 L'algorithme PLS . . . 22
1.2.5 Critères de validation du modèle . . . 24
1.2.6 L'adéquation aux données particulières . . . 25
1.2.7 Propriétés supplémentaires . . . 26
1.3 Méthode par analyse de la structure de covariance (LISREL) . . . 28
1.3.1 Le cas général . . . 28
1.3.2 La représentation graphique . . . 30
1.3.3 Qualité d'ajustement du modèle aux données . . . 31
1.3.4 L'adéquation aux données particulières . . . 35
1.3.5 Propriétés supplémentaires . . . 37
1.4 Conclusion . . . 38

2 Quelques développements récents **39**
2.1 PLS et LISREL : deux méthodes d'estimation complémentaires . . . 39
2.1.1 Deux approches complémentaires . . . 39
2.1.2 Le formalisme adopté . . . 42
2.2 De LISREL à PLS : un pont possible . . . 42
2.3 Une analyse de la sensibilité des estimations de l'approche PLS . . . 46
2.3.1 Taille de l'échantillon et nombre de variables manifestes par bloc . . . 46
2.3.2 Distribution inégale du nombre de variables manifestes . . . 48
2.3.3 Conclusion . . . 49
2.4 Autres méthodes . . . 49
2.4.1 L'approche Partial Maximum Likelihood (PML) . . . 49
2.4.2 La méthode Generalized Structured Component Analysis (GSCA) . . . 53
2.5 Conclusion . . . 55

3 La construction du modèle conceptuel **57**
3.1 Introduction . . . 57
3.2 Du sens des relations dans le modèle de mesure . . . 58
3.2.1 Modèle réflectif contre modèle formatif . . . 58

3.2.2 Le test des tétrades . . . 59
3.2.3 Une heuristique pour la découverte de construits réflectifs . . . 63
3.3 La construction du modèle de mesure . . . 66
3.3.1 L'unidimensionnalité dans le cadre des modèles d'équations structurelles . . . 66
3.3.2 La classification de variables . . . 67
3.3.3 Les modèles graphiques et les réseaux bayésiens . . . 68
3.3.4 Un algorithme basé sur les modèles graphiques probabilistes . . . 71
3.4 La construction du modèle structurel . . . 74
3.4.1 Les heuristiques d'apprentissage du modèle interne . . . 74
3.4.2 La méthode d'Amato . . . 75
3.4.3 Les modèles libres . . . 76
3.4.4 Une adaptation des modèles libres dans le cadre de l'approche PLS . . . 77

4 La comparaison de groupes d'observations 81
4.1 Introduction . . . 81
4.2 Les niveaux de comparaison . . . 81
4.3 Les tests de permutations . . . 82
4.4 La qualité globale . . . 83
4.4.1 La différence entre les méthodes LISREL et PLS . . . 83
4.4.2 Comparaison de la structure des données . . . 83
4.4.3 Comparaison de la qualité d'ajustement par des tests de permutations . . . 84
4.4.4 La reconstruction du modèle . . . 85
4.5 Les variables latentes . . . 86
4.6 Les coefficients structurels . . . 87
4.6.1 Comparaison paramétrique classique . . . 87
4.6.2 Comparaison basée sur les effets modérateurs . . . 87
4.6.3 Comparaison basée sur un test non paramétrique . . . 89
4.6.4 Comparaison basée sur des tests de permutations . . . 89
4.7 Elaboration d'un processus de comparaison : des données aux coefficients structurels . . . 90
4.7.1 La structure du processus . . . 90
4.7.2 Le processus . . . 91
4.8 Simulations . . . 92
4.9 Conclusion . . . 95

5 La non linéarité dans les modèles d'équations structurelles 97
5.1 Introduction et motivations . . . 97
5.2 La non linéarité dans les modèles d'équations structurelles . . . 97
5.2.1 L'approche LISREL . . . 98
5.2.2 Le cas PLS . . . 99
5.3 Méthodes basées sur des transformations des variables . . . 101
5.3.1 Le choix d'une famille de transformations . . . 101
5.3.2 Les transformations par B-splines monotones . . . 105
5.4 Transformations et approche PLS . . . 106
5.4.1 Au niveau du modèle de mesure . . . 106
5.4.2 Au niveau du modèle interne . . . 109
5.5 Conclusion . . . 113

6 Modèles structurels et données manquantes : le traitement d'un cas spécifique 115
6.1 Introduction . . . 115
6.2 Les données MCAR et MAR . . . 116
6.2.1 Les méthodes classiques . . . 116
6.2.2 La méthode *Full Information Maximum Likelihood* (FIML) et l'approche LISREL 117

6.2.3 L'algorithme *Non-linear Iterative PArtial Least Squares* (NIPALS) et l'approche PLS . . . 117
6.2.4 Exemple sur des données simulées . . . 118
6.3 Les données MNAR . . . 121
6.3.1 Les modèles économétriques . . . 121
6.3.2 Données MNAR et modèles d'équations structurelles . . . 123
6.4 Un cas spécifique : les questions filtrées . . . 124
6.4.1 Un problème théorique et "philosophique" . . . 124
6.4.2 Les méthodes classiques . . . 125
6.4.3 Méthodes alternatives . . . 125
6.5 Conclusion . . . 126

7 Applications à l'analyse de la satisfaction et de la fidélité 129
7.1 Satisfaction et fidélité : concepts et problématiques . . . 129
7.1.1 La notion de satisfaction . . . 130
7.1.2 La fidélité (*loyalty*) et ses problématiques . . . 131
7.1.3 La fidélité : problématiques . . . 132
7.1.4 Les pratiques en matière d'analyse de la satisfaction du client . . . 137
7.2 Les problématiques et les données . . . 140
7.3 Choix de la méthode de traitement : comparaison des approches PLS, LISREL et GSCA 142
7.4 Les modèles de satisfaction : du dire d'expert à la construction du modèle . . . 146
7.4.1 Réflectif vs. formatif . . . 146
7.4.2 Le modèle externe : une construction complexe . . . 148
7.4.3 Le modèle interne : la satisfaction, concept central . . . 151
7.4.4 Conclusion . . . 153
7.5 La relation face à l'ouverture du marché : comparaison de groupes dans le cadre de l'approche PLS . . . 153
7.5.1 Comparaison globale . . . 155
7.5.2 Reconstruction du modèle . . . 155
7.5.3 Comparaison des variables latentes . . . 155
7.5.4 Comparaison des coefficients structurels . . . 156
7.5.5 Conclusion . . . 156
7.6 L'asymétrie des facteurs de la satisfaction et la relation satisfaction - fidélité : la non linéarité dans les modèles d'équations structurelles . . . 157
7.6.1 Les théories de la non linéarité en marketing . . . 157
7.6.2 Le traitement du modèle externe : la mesure de l'asymétrie des impacts des attributs sur la satisfaction . . . 158
7.6.3 Le modèle interne : la non linéarité dans la relation entre satisfaction et fidélité . 162
7.6.4 Conclusion . . . 164
7.7 Le cas des réclamations : illustration du traitement de questions filtrées . . . 165
7.7.1 Segmentation et comparaison . . . 165
7.7.2 Ajout d'une variable supplémentaire . . . 167
7.7.3 Ajout d'une modalité . . . 168
7.7.4 Conclusion . . . 169
7.8 Conclusion des applications . . . 169

Conclusion générale 171

A Formulation alternative de la méthode Generalized Structured Component Analysis 175

B Principaux résultats supplémentaires **179**
B.1 Satisfaction et fidélité des clients 179
B.2 La construction du modèle 179
B.3 La non linéarité 182
B.4 Le traitement des questions filtrées 185
B.5 Programmes informatiques 186

Bibliographie **188**

Table des figures

1.1 Graphe associé à un modèle d'équations structurelles à variables latentes 18
1.2 Types de relations au niveau du modèle externe 20
1.3 Modèle d'équations structurelles pour l'approche PLS 22
1.4 Approche PLS avec le mode A et le schéma centroïde 24
1.5 Modèle d'équations structurelles dans le cas de LISREL 30

2.1 Modèle d'équations structurelles 44
2.2 Comparaison des scores obtenus par l'approche PLS mode A et l'approche ULS de McDonald (1996) 45
2.3 Evolution de la moyenne d'un coefficient structurel du modèle (avec l'écart type) en fonction de (i) la taille de l'échantillon et (ii) du nombre de variables manifestes par bloc 47
2.4 Evolution de la moyenne d'un *loading* du modèle (avec l'écart type) en fonction de (i) la taille de l'échantillon et (ii) du nombre de variables manifestes par bloc 47
2.5 Modèle d'équations structurelles pour l'estimation par l'approche GSCA 54

3.1 Modèles simulés pour la validation de l'algorithme de recherche d'un construit réflectif . 65
3.2 Arbre de classification pour la construction du modèle externe associé à un réseau bayésien initial 73

4.1 Illustration de l'effet modérateur lié à une variable 88
4.2 Illustration de l'effet modérateur lié à une variable dans le cadre de deux variables latentes 88
4.3 Modèle structurel simulé 93
4.4 Courbe de puissance empirique pour le test basé sur les communautés moyennes ($\bar{H}^2$) . 94
4.5 Courbe de puissance empirique pour le test basé sur les redondances moyennes ($\bar{F}^2$) . . 94
4.6 Courbe de puissance empirique pour le test basé sur les coefficients structurels 95
4.7 Processus de comparaison de groupes d'individus dans le cadre de l'approche PLS . . . 96

5.1 B-splines de degré 0 à 2 nœuds internes 102
5.2 Base des B-splines de degré 2 à 2 nœuds internes 103
5.3 Exemple de B-spline de degré 2 à 2 nœuds internes 103
5.4 Exemple de B-spline monotone de degré 2 à 2 nœuds internes 104
5.5 Comparaison des scores PLS et de la première composante de l'ACP pour des données simulées sur une variable endogène 107
5.6 Comparaison des scores PLS et de la première composante de l'ACP pour des données réelles sur une variable (a) endogène et (b) exogène 107
5.7 Processus pour l'application de l'approche PLS avec transformation non linéaire du modèle externe 108
5.8 Transformation simulée et estimée de la variable latente satisfaction 112
5.9 Comparaison des scores PLS avec ou sans la relation entre satisfaction et fidélité pour les variables satisfaction et fidélité 113

6.1 Modèle structurel simulé . . . 118

7.1 Modèle conceptuel marketing dans le cas des services afin de définir les leviers de l'engagement (Morgan et Sonquist, 1963) . . . 134
7.2 Généralisation du modèle de la figure 7.1 . . . 134
7.3 Modèle général développé par Frisou (1998) . . . 135
7.4 Modèle pour le SCSB . . . 138
7.5 Modèle pour l'ACSI . . . 138
7.6 Modèle pour l'ECSI . . . 139
7.7 Modèle structurel associé aux données EDF-1 . . . 141
7.8 Comparaison des scores de la variable latente satisfaction . . . 145
7.9 Comparaison des scores de la variable latente fidélité . . . 145
7.10 Réseau bayésien obtenu avec les groupes de variables associées . . . 149
7.11 Arbre de classification complet pour la méthode basée sur les réseaux bayésiens . . . 149
7.12 Modèles internes obtenus pour chaque méthode de construction . . . 152
7.13 Modèle interne simplifié associé aux données EDF-1 . . . 154
7.14 Modèle interne avec les différences entre hommes et femmes . . . 154
7.15 Modèle interne avec les différences entre favorable et défavorable à l'ouverture du marché de l'électricité . . . 154
7.16 Moyennes et écarts types des variables latentes satisfaction et fidélité pour chacun des groupes d'observations . . . 156
7.17 La chaîne allant des attributs de la satisfaction au profit de l'entreprise par Anderson et Gerbing (1982) . . . 158
7.18 Asymétrie des facteurs basée sur le modèle de Kano . . . 158
7.19 Répartition des variables manifestes par l'approche de Llosa . . . 159
7.20 Modèle structurel pour les approches PLS non linéaire basée sur le modèle externe et PLS classique . . . 161
7.21 Les transformations des variables manifestes associées à la satisfaction . . . 161
7.22 Relation satisfaction - fidélité . . . 162
7.23 Modèle structurel pour les approches PLS non linéaire basée sur le modèle interne et PLS classique . . . 163
7.24 Les transformations des variables latentes expliquant la satisfaction . . . 164
7.25 Modèle interne pour la sous-population ayant réclamé . . . 166
7.26 Modèle interne pour la sous-population ayant eu un contact . . . 166
7.27 Modèle interne pour la sous-population ayant réclamé avec une modalité "non applicable" 168
7.28 Modèle interne pour la sous-population ayant eu un contact avec une modalité "non applicable" . . . 169

A.1 Modèle d'équations structurelles pour l'estimation par l'approche GSCA (seconde formulation) . . . 176

B.1 Arbre de classification pour la méthode de Stan et Saporta (2005) . . . 182
B.2 Méthode de Llosa . . . 183

Liste des tableaux

1.1 Principales notations utilisées 19
1.2 Récapitulatif des propriétés de chaque indice (O : oui, N : non) 34

2.1 Comparaison théorique entre les approches PLS mode A et LISREL-ML 40
2.2 Comparaison pratique entre les approches PLS mode A et LISREL-ML 41
2.3 Problématiques et modèles d'équations structurelles 42
2.4 Biais moyens des estimations d'un coefficient structurel en fonction de la taille de l'échantillon et du nombre de variables manifestes par bloc 48
2.5 Biais moyens (avec intervalles de confiance) des estimations des coefficients structurels en fonction de la distribution du nombre de variables manifestes dans chaque bloc (VM : variables manifestes) 48

3.1 Résultats de simulations sur l'orientation du modèle externe 65

5.1 Comparaison entre la méthode PLS non linéaire et la méthode classique sur des données simulées 112
5.2 Comparaison des indices associés à la méthode PLS lorsque la relation satisfaction - fidélité est présente ou absente 113

6.1 Résultats des simulations sur les données manquantes (avec 5% de données manquantes) 119
6.2 Résultats des simulations sur les données manquantes (avec 10% de données manquantes) 120
6.3 Points forts et points faibles des méthodes de traitement des questions filtrées 127

7.1 Définitions conceptuelles de la satisfaction 130
7.2 Différents profils de clients en fonction des degrés de satisfaction et de fidélité 135
7.3 Définition des blocs de variables pour les données EDF-1 141
7.4 Définition des jeux de données utilisés 141
7.5 *Loadings* et coefficients structurels obtenus par les approches PLS, LISREL et GSCA sur les données EDF-1 143
7.6 Comparaison des R^2 associés aux variables latentes endogènes pour les approches PLS, LISREL et GSCA sur les données EDF-1 144
7.7 Principaux indices associés à l'approche PLS pour les données EDF-1 144
7.8 Processus de validation de l'algorithme 148
7.9 Construction du modèle de mesure 150
7.10 Etapes de la construction du modèle interne par les modèles libres itératifs 151
7.11 Construction du modèle interne 152
7.12 Résultats de la comparaison globale 155
7.13 Résultats des tests au niveau des variables latentes 156
7.14 Résultats des comparaisons des coefficients structurels 157
7.15 Classement des variables par la méthode de Brandt 160

7.16 Indices et coefficients estimés pour le modèle externe transformé, comparés à l'approche classique . . . 160
7.17 Indices pour le modèle interne transformé, comparés à l'approche classique . . . 163
7.18 Résultats des comparaisons pour les réclamations et les contacts . . . 167
7.19 Poids externes pour le cas de l'ajout d'une variable binaire "filtre" . . . 167

B.1 Quelques exemples de questions afin de définir les facettes de la fidélité . . . 179
B.2 Variables manifestes associées aux données GS . . . 180
B.3 Variables manifestes associées aux données BB et BF . . . 180
B.4 Blocs de variables obtenus pour l'ensemble des approches de construction . . . 181
B.5 Classification par l'approche de Brandt . . . 183
B.6 Discrétisation en 3 modalités . . . 184
B.7 Estimation du nombre de degrés et de nœuds des B-splines monotones pour le modèle externe en utilisant les communautés de la satisfaction . . . 184
B.8 Estimation du nombre de degrés et de nœuds des B-splines monotones pour le modèle interne en utilisant le R^2 de la fidélité . . . 184
B.9 Modèle externe dans le cadre de la segmentation des données sur les variables latentes filtrées . . . 185
B.10 Modèle externe dans le cadre de l'ajout d'une modalité aux variables filtrées sur les réclamations . . . 185
B.11 Modèle externe dans le cadre de l'ajout d'une modalité aux variables filtrées sur les contacts . . . 186
B.12 Programmes informatiques créés dans le cadre de ce travail (Macros SAS (SAS Institute Inc., 2004a)) . . . 187
B.13 Logiciels utilisés dans le cadre de ce travail . . . 187

Introduction générale

Les modèles d'équations structurelles à variables latentes constituent une méthode de modélisation de phénomènes apte à bien définir des systèmes complexes en interaction. Ils trouvent leur place dans la statistique et dans l'analyse de données et sont considérés comme des généralisations de nombreux modèles classiques (l'analyse en composantes principales, l'analyse factorielle, l'analyse canonique...). Ils se trouvent à la croisée de plusieurs domaines de recherche extrêmement riches : d'une part, le traitement de variables non observables dites latentes et, d'autre part, l'introduction de la notion de causalité dans les modèles statistiques. Des recherches sont menées à ce sujet depuis le début du XX$^{\text{ème}}$ siècle. Ces deux concepts complexes en font des modèles difficiles à traiter et avec lesquels il faut être très prudent. En ce début de XXI$^{\text{ème}}$ siècle, ils continuent à susciter maintes réactions et polémiques.

De plus, les modèles d'équations structurelles à variables latentes se sont développés en étroite collaboration entre différents domaines d'applications, tout d'abord dans le monde de la sociologie et de la psychologie quantitative grâce à la possibilité de tester des modèles complexes en utilisant des concepts qui ne peuvent pas s'exprimer directement. Puis, dans les sciences de gestion et de management où ils permettent l'évaluation de processus comme la satisfaction des consommateurs.

Deux courants de pensées existent dans l'estimation de ces modèles. Le premier, fondé sur une estimation des covariances en utilisant généralement l'estimateur du maximum de vraisemblance, est issu des recherches de Karl Jöreskog et est très appliqué dans la sociologie et la psychologie. Le second, initié par Herman Wold (qui était d'ailleurs professeur de Jöreskog), est basé sur une estimation par moindres carrés dits partiels. Ces deux courants s'opposent depuis de nombreuses années avec un léger avantage pour le premier grâce à une théorie plus approfondie et à des logiciels plus développés.

Cette thèse répond à une réelle demande de rationalisation et d'évolution venant d'un secteur particulier et dans lequel les modèles d'équations structurelles connaissent un essor singulier : l'analyse de la satisfaction et de la fidélité des clients. Nous contribuons ici à l'étude des modèles structurels, d'une part, par la formulation et la présentation de ces modèles et de leurs techniques d'estimation dans des termes simples et, d'autre part, par des apports théoriques sur des points clés dans l'utilisation de ces modèles. La construction de modèles, la comparaison de groupes d'observations, l'intégration de non linéarité et le traitement des données manquantes avec, en particulier, les questions dites filtrées, sont les domaines dans lesquels cette thèse apporte des innovations.

Cette étude, effectuée au Conservatoire National des Arts et Métiers avec le support du département Recherche et Développement de l'entreprise Electricité de France, aborde à la fois des aspects théoriques par la présentation de nouvelles méthodes mais aussi des points méthodologiques par l'étude des méthodes sur des données réelles.

L'analyse des modèles d'équations structurelles à variables latentes n'est pas récente dans l'industrie. Cependant, un certain nombre de points importants restaient en suspens dans leur traitement. Cette thèse a donc été initiée dans le but de répondre de manière rigoureuse à ces questions afin d'apporter des contributions, aussi bien pour l'application industrielle, que pour la communauté scientifique.

Nous sommes donc partis des données (questionnaires de satisfaction des clients) afin de cerner les problèmes et de les résoudre. Ces recherches se sont basées sur de nombreux résultats empiriques obtenus par la mise en place de multiples programmes informatiques.

La thèse est séparée en sept chapitres. Tout d'abord, deux chapitres permettent d'introduire les modèles d'équations structurelles et leurs méthodes d'estimation. Ainsi, nous commençons par une mise au point sur les modèles, leurs principes et leurs fondations. Nous présentons ensuite les méthodes d'estimation de ces modèles, que sont la méthode LISREL introduite par Jöreskog (1970) et l'approche PLS proposée par Wold (1982). Nous décrivons leurs principales propriétés, ainsi que leurs adaptations au cas de données spécifiques. Dans le chapitre 2, les deux méthodes d'estimation sont comparées et les différences de notations induites par chacune justifiées. Nous mettons ainsi en valeur la complémentarité de ces deux approches qui sont basées sur des principes très différents. Une analyse de sensibilité des estimations de l'approche PLS est présentée. Nous introduisons des méthodes développées récemment et qui nous paraissent intéressantes de par leurs propriétés. Ainsi, la méthode issue de LISREL basée sur la fonction ULS (*Unweighted Least Squares*) et qui suppose que les variables latentes sont des combinaisons linéaires de leurs variables manifestes associées, permet d'obtenir des estimations des scores des variables non observables proches de celles l'approche PLS (McDonald, 1996; Tenenhaus, 2007). Par ailleurs, la méthode *Partial Maximum Likelihood* facilite le traitement de données ayant des échelles de mesures différentes (Derquenne, 2005; Jakobowicz et Derquenne, 2007) et l'approche *Generalized Structured Component Analysis* utilise les avantages des méthodes PLS et LISREL tout en évitant leurs inconvénients par le biais d'une nouvelle formulation du modèle et par l'utilisation de l'algorithme des moindres carrés alternés (Hwang et Takane, 2004). Ainsi, les deux premiers chapitres permettent de mettre en place la théorie nécessaire aux autres développements de ce travail.

Les chapitres 3 à 6 sont consacrés à la présentation de contributions novatrices dans des cas particuliers. Dans chacun des chapitres, nous présentons les méthodes existantes afin d'obtenir un panorama complet pour chaque cas. Le chapitre 3 propose des méthodes permettant de construire le modèle conceptuel initial qui devra être estimé par les méthodes PLS ou LISREL. Nous nous attachons à l'orientation des relations du modèle de mesure, à sa construction et à la construction du modèle structurel. Pour le premier cas, nous présentons un algorithme fondé sur le test des tétrades (Bollen, 1990a) et vérifions ses propriétés par une étude de simulation par méthode de Monte-Carlo. Pour le second cas, nous utilisons l'apprentissage des réseaux bayésiens et, pour le troisième, une méthode itérative basée sur les corrélations partielles entre variables latentes. Le chapitre 4 est consacré à la comparaison de groupes d'observations sur un modèle d'équations structurelles avec l'introduction de tests permettant de comparer des groupes à différents niveaux du modèle. Ces méthodes sont introduites dans le cadre de l'approche PLS, qui ne suppose pas d'hypothèses de distribution des données. Les tests présentés sont basés sur des permutations de l'échantillon. Dans le chapitre 5, nous proposons des méthodes permettant de détecter des relations non linéaires dans le cadre de l'application de l'approche PLS. Des méthodes issues de l'*optimal scaling* et une famille de fonctions appelées B-splines monotones sont utilisées. Nous traitons aussi bien le modèle externe par l'emploi de l'analyse en composantes principales non linéaires que le modèle interne par l'utilisation de transformations permettant la linéarisation des relations. Le dernier chapitre de cette partie s'intéresse aux données manquantes dans les modèles structurels et, spécifiquement, aux variables dites "filtrées", c'est-à-dire aux variables non observées pour certaines observations car celles-ci sont dites "non applicables". Nous présentons les méthodes de traitement des données manquantes classiques avec un exemple sur données simulées, et les méthodes de traitement des données qui manquent non aléatoirement. Finalement, trois approches simples sont mises en place afin de travailler sur les variables "filtrées".

Le dernier chapitre est consacré à l'application des méthodes pour l'analyse de la satisfaction et la fidélité des clients. Il est séparé en deux sous-chapitres, le premier présente la théorie marketing sous-

jacente à l'analyse de la satisfaction et de la fidélité et, le second, propose un cas pratique. L'application se fait sur des données réelles issues d'un questionnaire posé aux clients d'EDF. Nous illustrons la mise en oeuvre des méthodes présentées tout au long de la thèse en tentant de répondre à des questions clés issues d'expériences industrielles dans le traitement des modèles d'équations structurelles.

Ce travail se conclut sur une discussion générale rappelant les apports, les limites et les orientations de recherche induites par cette thèse.

Chapitre 1

Les modèles d'équations structurelles à variables latentes

1.1 La conceptualisation

1.1.1 Préliminaires

Les modèles d'équations structurelles à variables latentes sont basés sur un certain nombre de concepts que nous allons définir dans le cadre de ce premier chapitre. Ils sont fondés sur des équations dites structurelles pouvant être retranscrites par des modèles graphiques.

Les méthodes d'estimation de modèles d'équations structurelles sont toutes issues de recherches datant du début du XX$^{\text{ème}}$ siècle. Ces recherches se sont basées sur deux axes : l'analyse de relations structurelles (*path analysis*) introduite par Wright (1918, 1921) et la conceptualisation de la notion de variable latente lancée par les travaux de Spearman (1904). Les méthodes d'estimation ont été mises en place dans les années 1970 par Jöreskog (1970) et Wold (1973). Ces méthodes sont issues aussi bien de recherches en statistique, qu'en psychologie, en sociologie ou encore en économétrie.

Les approches que nous allons présenter sont basées sur un modèle conceptuel préétabli. Certains chercheurs les appellent confirmatoires (*confirmatory*). Elles ont actuellement des applications dans de nombreux domaines dont la psychologie, la sociologie, le marketing...
Avant tout développement des méthodes en elles-mêmes, il est important de définir clairement les deux types de variables présentes.

Définition 1.1. *Une* ***variable manifeste*** *est une variable pour laquelle une mesure peut être directement recueillie (observée, mesurée, etc.).*

Définition 1.2. *Une* ***variable latente*** *correspond à une caractéristique qui n'est pas directement observable et qui ne peut donc pas être mesurée directement.*

Les variables latentes étant inconnues, elles pourront être estimées à partir des variables manifestes. Nous utilisons un terme qui peut porter à confusion pour définir le calcul des scores associés aux variables latentes pour chaque observation. Nous parlerons de l'estimation d'une variable latente lorsqu'on veut évoquer les valeurs associées à cette variable obtenues à partir des données. Nous tenterons néanmoins de minimiser cette utilisation. Ainsi, les variables latentes seront estimées en combinant l'information recueillie à l'aide d'un ensemble de variables manifestes et en isolant leur portion de variance commune. Cette façon de faire permet de contrôler et d'isoler les erreurs de mesure. La théorie

liée aux variables latentes est complexe et a été développée par de nombreux chercheurs comme Bollen (2002) ou Borsboom et al. (2003). Dans le cadre de cette thèse, nous appelons variables latentes aussi bien les variables latentes au sens de facteurs (généralement indéterminés, c'est-à-dire que les facteurs sont des variables latentes théoriques non estimés au niveau de chaque observation comme dans le cas de l'analyse factorielle) que celles au sens de composantes (calculées comme combinaisons linéaires des variables manifestes associées comme dans le cas de l'analyse en composantes principales). Nous différencions ces deux cas lorsque, pour des raisons de compréhension, cela est nécessaire.

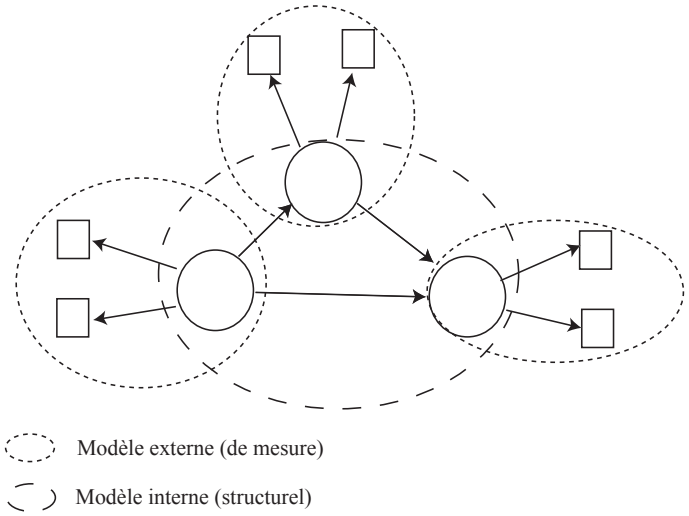

FIG. 1.1 – Graphe associé à un modèle d'équations structurelles à variables latentes

Un modèle d'équations structurelles à variables latentes consiste en un système d'équations dites structurelles pouvant être représentées par un graphe orienté. Les nœuds de ce graphe représentent les variables (sous forme de carrés pour les variables manifestes et de cercles pour les variables latentes) et les arcs modélisent des relations de causalité. Chaque variable manifeste est associée à une seule variable latente et il existe des relations entre les variables latentes (voir la fig. 1.1). Dans le cadre de l'utilisation de variables latentes, on sépare généralement le modèle en deux sous-modèles.

Définition 1.3. *Le* ***modèle de mesure*** *ou* ***modèle externe*** *est une sous-partie du modèle complet incluant les relations entre variables manifestes et latentes (fig. 1.1).*

Définition 1.4. *Le* ***modèle structurel*** *ou* ***modèle interne*** *est une sous-partie du modèle complet incluant les relations entre les variables latentes (fig. 1.1).*

Dans le cadre des méthodes d'estimation du modèle complet, les deux sous-modèles sont estimés soit simultanément (cas de la méthode par analyse de la structure de covariance, généralement appelée LISREL pour *LInear Structural RELationships*), soit alternativement (cas de l'approche PLS (*Partial Least Squares*)).
D'autre part, la littérature sur le sujet étant très étendue, il est important de définir clairement le vocabulaire utilisé. Ainsi, les modèles d'équations structurelles à variables latentes appartiennent à différentes familles de méthodes.
Comme des représentations graphiques des modèles existent, nous considérerons qu'elles appartiennent aux "modèles graphiques" au sens large. Cependant, nous réserverons le terme de modèles graphiques

aux modèles développés par Lauritzen (1996) et Whittaker (1990), ainsi qu'aux réseaux bayésiens développés par Pearl (1988).
Les méthodes d'estimations sont dites confirmatoires, c'est-à-dire qu'elles visent à confirmer un modèle théorique préétabli. Nous verrons par la suite que ce point de vue généralement partagé est à nuancer dans certains cas.

1.1.2 Formalisme

Les modèles d'équations structurelles à variables latentes peuvent être estimés par deux méthodes issues de domaines de recherches différents et formulés sur des principes distincts. Les notations associées à ces deux principes d'estimation sont différentes. Dans un souci de clarté, nous présenterons chacune des approches dans son formalisme habituel : d'une part, pour PLS en utilisant un seul type de variables latentes, d'autre part, dans le cadre de l'approche LISREL, nous différencions les variables latentes qui ne dépendent d'aucune autre, nommées **exogènes** de celles étant expliquées par d'autres variables latentes nommées **endogènes**. Le tableau 1.1 rassemble les notations et indices communs aux deux approches que nous utiliserons dans toute cette thèse.

Concept	Notation
Variable latente	$\boldsymbol{\xi}$, $\boldsymbol{\eta}$
Variable manifeste	$\mathbf{x}$
Termes d'erreurs	$\boldsymbol{\epsilon}$, $\boldsymbol{\delta}$, $\boldsymbol{\zeta}$
Coefficient structurel	β, γ
Loading	π
Indice des variables latentes	k
Indice des variables manifestes du bloc k	kj
Nombre d'observations	N
Nombre de variables latentes (endogènes, exogènes)	K (K_{endo}, K_{exo})
Nombre de variables manifestes dans le bloc k	p_k
Nombre total de variables manifestes	P
$\boldsymbol{\xi}_j$ explique $\boldsymbol{\xi}_k$ dans le modèle	$\boldsymbol{\xi}_j \rightarrow \boldsymbol{\xi}_k$
$\boldsymbol{\xi}_j$ explique $\boldsymbol{\xi}_k$ ou $\boldsymbol{\xi}_k$ explique $\boldsymbol{\xi}_j$ dans le modèle	$\boldsymbol{\xi}_j \leftrightarrow \boldsymbol{\xi}_k$

Tab. 1.1 – Principales notations utilisées

Par ailleurs, dans le cadre du chapitre 2.1 et des comparaisons entre les deux approches d'estimation, nous justifions ces différences de notations et présentons une notation basée sur la formulation appelée *Reticular Action Model* (RAM, McDonald et McArdle (1984); McDonald (1996)) qui représente la forme d'expression la plus simple du modèle d'équations structurelles à variables latentes.

Nous présentons donc chacune des approches indépendamment et effectuerons des recoupements dans le cadre du chapitre 2.

1.2 L'approche PLS

L'approche PLS (*partial least squares*) permet d'estimer un modèle d'équations structurelles, elle a été introduite pour la première fois par Wold (1973, 1980b). Elle est l'aboutissement des recherches de ce dernier qui avaient débuté avec un article sur l'analyse en composantes principales dans lequel l'algorithme NILES (*nonlinear iterative least squares*, Wold (1966)) a été présenté. Celui-ci deviendra NIPALS (*nonlinear iterative partial least squares*, Wold (1973)) pour aboutir à l'approche PLS. Elle est aussi l'application des théories de la méthode du point fixe développées par le même auteur (Wold, 1980a).

L'approche PLS est une méthode très générale qui contient comme cas particulier l'analyse en composantes principales, l'analyse canonique, l'analyse des redondances, la régression PLS, l'analyse canonique généralisée au sens de Horst ou de Carroll, au niveau de la première composante (Tenenhaus, 1999).

L'approche PLS est issue d'une théorie ancienne, celle de l'estimation des moindres carrés et elle se base sur des régressions simples et multiples. En conséquent, elle nécessite peu d'hypothèses et c'est pour cette raison qu'elle est appelée modélisation douce (*soft modeling*, Wold (1982)). Des hypothèses liées à la nature même du modèle sont nécessaires et nous les détaillerons au cours des prochains paragraphes.

1.2.1 Le modèle externe

Il existe plusieurs schémas de modélisation du modèle externe qui modifieront la manière dont les variables latentes seront construites. Il existe trois façons de relier les variables manifestes aux variables latentes (voir fig. 1.2).

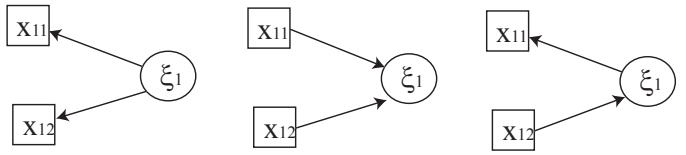

FIG. 1.2 – Types de relations au niveau du modèle externe

Le schéma réflectif

C'est celui adopté dans la plupart des utilisations des modèles d'équations structurelles à variables latentes. Chaque variable manifeste est reliée à sa variable latente par une régression simple.

Définition 1.5. *Les relations du modèle externe sont dites* ***réflectives*** *si, pour chaque variable latente* $\boldsymbol{\xi}_k$*, la relation entre cette variable et l'ensemble des variables manifestes qui lui sont associées s'écrit :*

$$\mathbf{x}_{kj} = \pi_{kj}\boldsymbol{\xi}_k + \boldsymbol{\epsilon}_{kj} \tag{1.1}$$

avec comme contraintes :
- $cor(\boldsymbol{\epsilon}_{ij}, \boldsymbol{\xi}_k) = 0\,,\ \forall i, k = 1, ..., K\,,\ \forall j = 1, ..., p_k$ *;*
- $cor(\boldsymbol{\epsilon}_{ij}, \boldsymbol{\epsilon}_{lm}) = 0\,,\ \forall i, j, l, m = 1, ..., p_k\,,\ (i, j) \neq (l, m)\,,\ \forall k = 1, ..., K.$

où $\mathbf{x}_{kj}$ *est le vecteur associé à la* j^e *variable manifeste de la variable latente* $\boldsymbol{\xi}_k$*,* π_{kj} *est un loading associé à* $\mathbf{x}_{kj}$ *et* $\boldsymbol{\epsilon}_{kj}$ *est un terme d'erreur (erreurs de mesure des variables manifestes).*

Les valeurs prises par les variables manifestes sont des "conséquences" de la variable latente. La relation de causalité va de la variable latente vers les variables manifestes. Nous verrons dans le chapitre 3.1 de cette thèse que ce schéma est parfois utilisé de manière abusive.

Le schéma formatif

Le schéma formatif est moins fréquemment utilisé. On suppose que chaque variable latente est une combinaison linéaire de ses variables manifestes correspondantes.

Définition 1.6. *Les relations du modèle externe sont dites* ***formatives*** *si, pour chaque variable latente* $\boldsymbol{\xi}_k$*, la relation entre cette variable et l'ensemble des variables manifestes qui lui sont associées s'écrit :*

$$\boldsymbol{\xi}_k = \sum_j \omega_{kj} \mathbf{x}_{kj} + \boldsymbol{\delta}_k \tag{1.2}$$

où ω_{kj} *est un poids et* $\boldsymbol{\delta}_k$ *est un vecteur d'erreur qui est supposé de moyenne nulle et non corrélé aux variables manifestes.*

Ce schéma entraîne une modification de la signification de la variable latente et conduit à une nouvelle façon de modéliser le modèle externe. La variable latente est ici un construit, elle est formée à partir des variables manifestes qui lui sont associées, elle ne s'intègre pas dans les définitions classiques de variables latentes (Bollen, 2002).

Le schéma MIMIC

C'est un mélange des deux schémas précédents.

Définition 1.7. *Les relations du modèle externe sont dites* ***MIMIC*** *si, pour une variable latente* $\boldsymbol{\xi}_k$*, les* q_k *premières variables manifestes sont modélisées avec le schéma réflectif et les* $p_k - q_k$ *variables restantes avec le schéma formatif.*

Nous approfondirons ces notions complexes dans le cadre du chapitre 3.

1.2.2 Le modèle interne

Il est défini par des équations linéaires reliant les variables latentes entre elles. Pour toute $\boldsymbol{\xi}_k$ endogène, on a :

$$\boldsymbol{\xi}_k = \sum_{i:\boldsymbol{\xi}_i \to \boldsymbol{\xi}_k} \beta_{ki} \boldsymbol{\xi}_i + \boldsymbol{\zeta}_k \tag{1.3}$$

avec comme contraintes :
- $cor(\boldsymbol{\zeta}_k, \boldsymbol{\epsilon}_{ki}) = 0\,, \quad \forall k = 1, .., K, \forall i = 1, ..., p_k$
- $cor(\boldsymbol{\zeta}_k, \boldsymbol{\xi}_j) = 0\,, \quad \forall j \neq k.$

où β_{ki} représente le coefficient structurel associé à la relation entre les variables $\boldsymbol{\xi}_k$ et $\boldsymbol{\xi}_i$, et $\boldsymbol{\zeta}_k$ est un terme d'erreur associé à la variable latente endogène $\boldsymbol{\xi}_k$.

Ces équations peuvent être exprimées sous forme matricielle, mais nous nous restreindrons à cette expression dans ce chapitre.

1.2.3 La représentation graphique

Les équations structurelles peuvent être représentées par un *path diagram*, c'est-à-dire un graphe orienté. La figure 1.3 représente un modèle à 3 variables latentes associées chacune à trois variables manifestes.

Ce graphe (*path diagram*) représente le modèle d'équations structurelles défini par les équations suivantes :

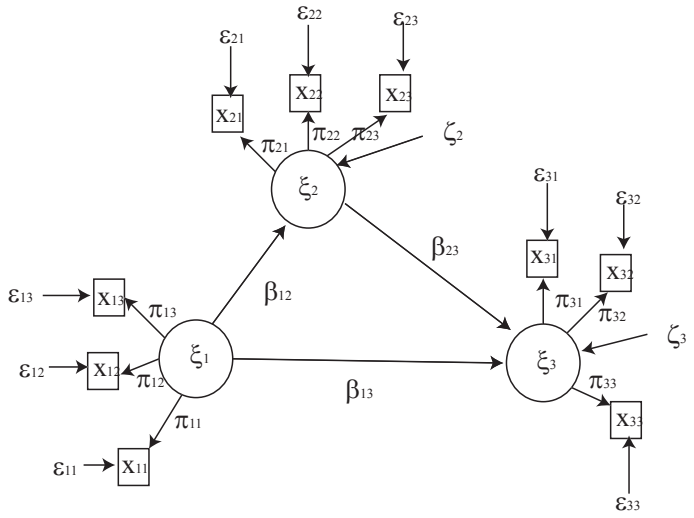

FIG. 1.3 – Modèle d'équations structurelles pour l'approche PLS

– On a trois variables latentes $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$ et $\boldsymbol{\xi}_3$ représentées comme suit :

$$\boldsymbol{\xi}_2 = \beta_{12}\boldsymbol{\xi}_1 + \boldsymbol{\zeta}_2$$

$$\boldsymbol{\xi}_3 = \beta_{13}\boldsymbol{\xi}_1 + \beta_{23}\boldsymbol{\xi}_2 + \boldsymbol{\zeta}_3$$

– Les variables manifestes pour le bloc associé à la variable latente $\boldsymbol{\xi}_1$ sont définies par :

$$\begin{pmatrix} \mathbf{x}_{11} \\ \mathbf{x}_{12} \\ \mathbf{x}_{13} \end{pmatrix} = \begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{13} \end{pmatrix} \boldsymbol{\xi}_1 + \begin{pmatrix} \boldsymbol{\epsilon}_{11} \\ \boldsymbol{\epsilon}_{12} \\ \boldsymbol{\epsilon}_{13} \end{pmatrix}$$

1.2.4 L'algorithme PLS

L'approche PLS est basée sur un algorithme itératif qui alterne une construction des variables latentes en se basant sur le modèle externe avec une autre construction se basant sur le modèle interne. Après convergence, les coefficients du modèle peuvent être estimés par régressions ordinaires simples ou multiples.

– *Estimation externe de $\boldsymbol{\xi}_k$:*
On commence par calculer les variables latentes standardisées par combinaison linéaire de leurs variables manifestes centrées :

$$\mathbf{y_k} \propto \pm \mathbf{X}_k \mathbf{w}_k \tag{1.4}$$

où le symbole $\propto$ signifie que le terme de gauche est égal au terme de droite standardisé et $\pm$ montre l'ambiguïté du signe. On choisit le signe de façon à ce que $\mathbf{y}_k$ soit positivement corrélée au plus de colonnes de $\mathbf{X}_k$ possible. Les éléments de $\mathbf{w}_k$ sont appelés **poids externes**.

– *Estimation interne :*
L'estimation interne des variables latentes standardisées est définie par :

$$\mathbf{z}_k \propto \sum_{j=1}^{K} c_{kj} e_{kj} \mathbf{y}_j \tag{1.5}$$

où $\mathbf{C} = (c_{kj})$ est une matrice de 0 et de 1 telle que $c_{kj} = 1$ si $\boldsymbol{\xi}_k \leftrightarrow \boldsymbol{\xi}_j$ et $\boldsymbol{E} = (e_{kj})$ est la matrice des poids internes définie par :

Schéma centroïde :

$$e_{kj} = sgn(\mathbf{y}_k' \mathbf{y}_j)$$

Schéma factoriel :

$$e_{kj} = \mathbf{y}_k' \mathbf{y}_j$$

Schéma structurel :

$$e_{kj} = \begin{cases} \text{coefficient de régression dans la régression de } \mathbf{y}_k \text{ sur } \mathbf{y}_j \text{ si } \boldsymbol{\xi}_j \rightarrow \boldsymbol{\xi}_k \\ cor(\mathbf{y}_k, \mathbf{y}_j) \text{ si } \boldsymbol{\xi}_k \rightarrow \boldsymbol{\xi}_j \end{cases}$$

A partir de cette étape, il existe deux façons de mettre à jour les poids externes w_{kj} :

Mode A :

$$\mathbf{w}_k = \frac{1}{\mathbf{z}_k' \mathbf{z}_k} \mathbf{X}_k' \mathbf{z}_k$$

Mode B :

$$\mathbf{w}_k = (\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k' \mathbf{z}_k$$

Sous la contrainte $\mathbf{w}_k' \mathbf{X}_k' \mathbf{X}_k \mathbf{w}_k = N$.

– *Algorithme d'estimation des poids :*

(1) Fixer arbitrairement les poids externes w_{kj}

(2) Calculer $\mathbf{y}_k$ avec l'équation 1.4

(3) Calculer $\mathbf{z}_k$ avec l'équation 1.5

(4) Recalculer les poids externes en appliquant le mode A ou le mode B

(5) Si convergence, aller en (7)

(6) Aller en (2)

(7) Calculer les coefficients structurels par régressions ordinaires

Propriété 1.1. *L'algorithme précédent est convergent avec une probabilité* 1 *pour* $K \leq 2$ *(Lyttkens et al., 1975).*

Au-delà de deux blocs, cette convergence n'a été que constatée dans la pratique.

Valeur initiale des poids externes

La première étape de l'algorithme consiste à fixer l'ensemble des poids externes. Ces poids sont ensuite standardisés afin d'obtenir des variables latentes de variance 1.
Un choix usuel pour ces valeurs initiales est de prendre :

$$w_{kj} = sign(cor(\mathbf{x}_{kj}, \mathbf{y}_k))$$

ou plus simplement,

$$w_{kj} = \begin{cases} sign(cor(\mathbf{x}_{kj}, \mathbf{y}_k)) & \text{pour } k = 1, \\ 0 & \text{sinon.} \end{cases}$$

Tenenhaus et al. (2005) conseillent une autre méthode :
Pour chaque bloc, on prend les éléments du premier vecteur propre obtenu lors de l'ACP avec une majorité de signes positifs. En cas d'égalité entre les signes positifs et négatifs, la variable avec la plus grande corrélation en valeur absolue prend le signe positif.

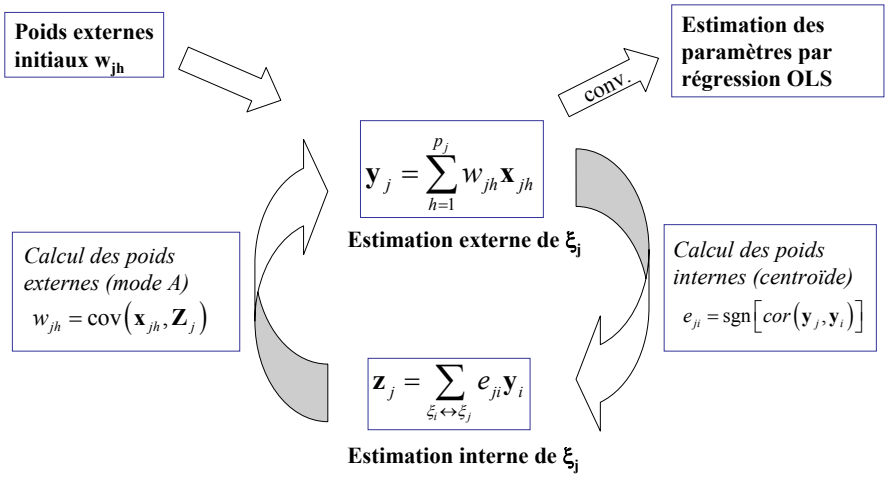

FIG. 1.4 – Approche PLS avec le mode A et le schéma centroïde

Evaluation des équations structurelles

L'équation structurelle 1.3 est estimée à l'aide d'une régression multiple (OLS) dans laquelle les variables latentes sont remplacées par leurs scores $\mathbf{y}_k$. De la même façon, l'équation 1.1 est estimée par une régression simple et, dans le cas formatif, l'équation 1.2 est estimée par une régression multiple en utilisant ces mêmes scores.
L'utilisation de régressions multiples peut entraîner des coefficients négatifs en cas de présence de multicolinéarité entre les variables latentes estimées. Dans ce cas, on utilisera une régression PLS (Tenenhaus, 1998).

1.2.5 Critères de validation du modèle

Dans le cadre de l'application de la méthode PLS sur un modèle, on a trois niveaux de validation du modèle. La qualité du modèle externe, celle du modèle interne et la qualité de chaque équation structurelle de régression.

La communauté, la redondance et un indice absolu

Définition 1.8. *La **communauté** (communality, H^2) évalue la qualité du modèle de mesure pour chaque bloc. Pour un bloc k :*

$$H_k^2 = \frac{1}{p_k}\sum_{i=1}^{p_k} cor^2(\mathbf{x}_{ki}, \mathbf{y}_k) \tag{1.6}$$

Elle représente la proportion de la variance des variables manifestes expliquée par leur variable latente associée.

Définition 1.9. *La **redondance** (redundancy, F^2) évalue la qualité du modèle structurel pour chaque bloc endogène k en prenant en compte le modèle de mesure, on peut l'écrire :*

$$F_k^2 = H_k^2 \times R^2(\mathbf{y}_k, \{les\ \mathbf{y}_i\ qui\ expliquent\ \mathbf{y}_k\}) \tag{1.7}$$

Elle part de la même idée que la communauté mais la variable latente est remplacée par son estimation à partir des variables latentes voisines.

Ces deux indices s'attachent chacun à la qualité de l'un des deux sous-modèles ; Tenenhaus et al. (2004) ont introduit un indice de qualité d'ajustement global.

Définition 1.10. *On définit le* ***GoF*** *(Goodness of Fit) par la moyenne géométrique de la moyenne des communauté sur l'ensemble des variables latentes ($\bar{H}^2$) et de la moyenne des R^2 associés aux variables latentes endogènes ($\bar{R}^2$) :*

$$GoF = \sqrt{\bar{H}^2 \times \bar{R}^2} \tag{1.8}$$

Ces indices sont obtenus directement à partir des estimations sans aucune hypothèse distributionnelle sous-jacente (ce qui empêche toute notion d'intervalle de confiance). Afin de pallier le risque de manque de robustesse et, grâce à l'amélioration de la puissance des outils informatiques, des indices obtenus par rééchantillonnage peuvent être utilisés.

Indices rééchantillonnés

L'utilisation d'indices rééchantillonnés est très bien adaptée au cas de l'approche PLS. En effet, l'absence d'hypothèse de distribution ne nous permet pas d'obtenir d'indice de qualité d'ajustement avec des intervalles de confiance.
Wold (1982), Chin (2005) et Tenenhaus et al. (2005) ont détaillé des procédures de rééchantillonnage d'indice de qualité prédictive.

- Wold (1982) propose d'utiliser le Q^2 de Stone (1975) et Geisser (1974), qui "va comme un gant" à l'approche PLS (Wold, 1982). Pour une variable latente $\boldsymbol{\xi}_k$, celui-ci est calculé par :

$$Q_k^2 = 1 - \frac{\sum_{j=1}^{p_k}\sum_{i=1}^{N}(x_{kji} - \bar{x}_{kj} - \hat{\pi}_{kj} \times Pred(y_{kj}))^2}{\sum_{j=1}^{p_k}\sum_{i=1}^{N}(x_{kji} - \bar{x}_{kj})^2} \tag{1.9}$$

  où x_{kji} est la i^e observation de la j^e variable du bloc k et

$$Pred(\mathbf{y}_k) = \sum_{k' : \boldsymbol{\xi}_{k'} \to \boldsymbol{\xi}_k} \hat{\boldsymbol{\beta}}_{k'} \mathbf{y}_{k'}$$

  est la prédiction de la variable latente $\boldsymbol{\xi}_k$ obtenue à partir des variables latentes qui l'expliquent dans le modèle structurel.

- Tenenhaus et al. (2005) clarifient les notions d'indices validés par validation croisée : la qualité du modèle de mesure pour chaque bloc est mesurée par la cv-communauté. C'est une sorte de R^2 par validation croisée entre les variables manifestes et leurs variables latentes endogènes. La qualité de chaque équation structurelle est calculée par l'indice de cv-redondance, appelé aussi Q^2 de Stone-Geisser. C'est une sorte de R^2 par validation croisée entre les variables manifestes d'une variable latente endogène et toutes les variables manifestes associées aux variables latentes expliquant la variable latente initiale. Les niveaux de significativité peuvent être calculés en utilisant le test t de Student ou en utilisant des méthodes comme le bootstrap ou le jackknife.

Ces indices constituent des indices d'estimation de la qualité prédictive du modèle et ne permettent pas d'étudier la qualité d'ajustement du modèle aux données. L'absence de ce type d'indices dans le cadre de l'approche PLS constitue l'un des défauts de cette méthode.

1.2.6 L'adéquation aux données particulières

Les données mixtes

Cette approche est très bien adaptée dans le cas de variables manifestes continues. Etant basée sur des régressions ordinaires, elle peut traiter des variables binaires. L'utilisation de variables manifestes

catégorielles ordonnées est possible mais le traitement se fera en considérant que celles-ci sont continues. Comme aucune distribution sous-jacente n'est supposée, ceci ne posera pas de gros problèmes. Cependant, dans le cas de variables catégorielles à peu de modalités ou non ordonnées, l'approche ne pourra pas s'appliquer.
Pour pouvoir utiliser des données dont on ne peut pas supposer une distribution continue sous-jacente, plusieurs adaptations ont été exposées. Nous présenterons l'une d'elles dans le second chapitre (*Partial Maximum Likelihood*, Derquenne (2005); Jakobowicz et Derquenne (2007)). Par ailleurs, Betzin et Henseler (2005) proposent l'utilisation des moindres carrés alternés afin de "quantifier" les variables catégorielles dans le cadre de l'approche PLS. Cet angle de recherche se rapproche de nos travaux sur la non linéarité et seront repris dans le chapitre 5.

Les données manquantes

Nous traiterons en détail ce cas dans le cadre du chapitre 6 de cette thèse.

1.2.7 Propriétés supplémentaires

L'approche PLS étant basée sur un algorithme itératif, peu de propriétés théoriques ont été démontrées. Certaines sont communément supposées car elles se vérifient dans la pratique. La multiplicité des modes et des schémas d'estimation rend l'obtention de propriétés générales d'autant plus difficile.

Dijkstra (1981), Mathes (1993) ou Hanafi (2004) font parti des chercheurs à avoir tenter de prouver des propriétés générales associées à l'approche PLS. Nous présentons donc soit des propriétés avec des références pour les preuves, soit des conjectures généralement supposées mais sans preuves mathématiques approfondies.

Conjecture 1.1. *Les estimations des coefficients du modèle sont consistantes au sens large (Wold, 1982). C'est-à-dire qu'elles se rapprochent de la valeur réelle du paramètre lorsque le nombre de variables manifestes par bloc et le nombre d'observations tendent vers l'infini.*

Cette conjecture, émise sans preuve classique, a été étudiée par Dijkstra (1981) qui a montré que les valeurs estimées se rapprochaient bien de la valeur réelle mais que rien ne pouvait prouver qu'elles tendaient vers cette valeur à la limite. Nous étudierons cette propriété dans le cadre de l'analyse de sensibilité des estimations des paramètres de l'approche PLS (chap. 2.3, p. 46).

L'approche PLS de par son processus itératif et l'alternance entre deux optimisations par moindres carrés ne présente pas de fonction globale à optimiser dans le cas général. Néanmoins, dans certains cas particuliers, une fonction globale existe.
Nous ne nous attardons pas ici sur le cas de deux blocs de variables (dans ce cas, on peut voir Tenenhaus (1999)). Pour plus de deux blocs, nous avons :

Conjecture 1.2. *Pour le mode B, en choisissant le schéma centroïde, les variables latentes sont obtenues en maximisant le critère :*

$$\sum_{k,l:\boldsymbol{\xi}_l \leftrightarrow \boldsymbol{\xi}_k} |cor(\mathbf{X}_k\mathbf{w}_k, \mathbf{X}_l\mathbf{w}_l)| \tag{1.10}$$

Conjecture 1.3. *Pour le mode B, en choisissant le schéma factoriel,les variables latentes sont obtenues en maximisant le critère :*

$$\sum_{k,l:\boldsymbol{\xi}_l \leftrightarrow \boldsymbol{\xi}_k} cor^2(\mathbf{X}_k\mathbf{w}_k, \mathbf{X}_l\mathbf{w}_l) \tag{1.11}$$

Les deux critères introduits montrent que la différence entre les variantes centroïde et factorielle réside dans le degré de prise en compte de la liaison entre les variables latentes.

Propriété 1.2. *Au niveau de l'estimation du modèle interne, deux variantes (centroïde et factorielle) traitent l'ensemble des variables latentes (exogènes et endogènes) d'une manière symétrique, contrairement à la variante structurelle qui distingue les variables latentes explicatives des variables latentes expliquées.*

Par ailleurs, nous avons :

Propriété 1.3. *L'estimation du modèle est sensible à la mise à l'échelle uniquement lors de l'utilisation du mode A (Dijkstra, 1981).*

Propriété 1.4. *Les estimations par les modes A et B sont similaires lorsque la matrice traitée est orthogonale.*

Ceci s'explique par le fait que le mode A est basé sur des régressions simples et le mode B sur des régressions multiples. Or, sur une matrice orthogonale, régressions simples et multiples sont équivalentes.

Wold (1982), dès ses premiers travaux, a voulu associer au mode A, le schéma réflectif et au mode B, le schéma formatif. Cette formalisation est rentrée dans les pratiques mais n'est basée sur aucune démonstration de la part de l'auteur. La différence entre les modes n'étant pas associée au schéma, elle est associée à l'objet de l'étude. Ainsi, le choix du mode pourra être motivé par différents facteurs :

- si on veut donner un plus grand poids au modèle externe, on utilisera le mode A,
- si on veut donner un plus grand poids au modèle interne, on utilisera le mode B,
- si on a beaucoup de variables manifestes par bloc, on favorisera le mode A.

Le premier point se justifie par le fait que l'application du mode A sur un seul bloc revient à une analyse en composante principale, sa généralisation à plusieurs blocs favorisera donc le modèle externe (nous étudierons cette propriété dans le chapitre 5). Lorsqu'on a deux blocs, le mode B équivaut à une analyse canonique qui maximise les corrélations entre les deux facteurs. Sa généralisation à plusieurs blocs favorisera donc le modèle interne. Finalement, le troisième point s'explique par le fait que lorsqu'on a beaucoup de variables dans un bloc, les risques de multicolinéarité avec le mode B deviennent plus importants.

Dans le cadre de cette thèse, nous nous focaliserons sur le mode A car, d'une part, il est le plus utilisé dans la littérature et reste associé au schéma réflectif et, d'autre part, car nous pourrons avoir des blocs avec de nombreux indicateurs.

1.3 Méthode par analyse de la structure de covariance (LISREL)

La méthode par analyse de la structure de covariance utilise un système d'équations structurelles basé sur l'estimation de la matrice de covariance. Elle permet de juger la qualité d'ajustement du modèle aux données par ses propriétés distributionnelles. Cette méthode vise donc à établir la qualité d'un modèle préétabli en se basant sur les données.
Cette méthode a été développée par Jöreskog (1970) qui a créé le logiciel de référence LISREL (Jöreskog et Sörbom, 1996). On peut trouver une présentation claire et précise de celle-ci dans Bollen (1989). Elle a plusieurs noms dans la littérature, les principaux étant LISREL, *Structural Equation Modeling (SEM)*, *Covariance Structure Analysis*... Dans un souci de simplicité, nous utiliserons le terme "méthode LISREL" dans le cadre de cette thèse.

Etant initialement basée sur le maximum de vraisemblance, elle est exigeante en terme d'hypothèses probabilistes. Ainsi, l'indépendance des observations et la normalité multivariée des données sont requises préalablement à son application.

1.3.1 Le cas général

Dans la littérature classique sur cette approche, les variables latentes endogènes et exogènes sont généralement représentées séparément. On aura deux types d'équations :

- pour le modèle de mesure :
$$\mathbf{y} = \mathbf{\Lambda_y}\boldsymbol{\eta} + \boldsymbol{\epsilon} \qquad \mathbf{x} = \mathbf{\Lambda_x}\boldsymbol{\xi} + \boldsymbol{\delta} \tag{1.12}$$
- pour le modèle structurel :
$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{1.13}$$

avec

- $\boldsymbol{\eta}$: variables latentes endogènes, c'est-à-dire une variable latente qui est expliquée par une ou plusieurs autres variables latentes,
- $\boldsymbol{\xi}$: variables latentes exogènes, c'est-à-dire une variable latente explicative qui n'est pas endogène,
- $\mathbf{y}$: variables manifestes relatives aux variables latentes endogènes
- $\mathbf{x}$: variables manifestes relatives aux variables latentes exogènes
- $\boldsymbol{\epsilon}$: erreurs de mesure associées aux $\mathbf{y}$
- $\boldsymbol{\delta}$: erreurs de mesure associées aux $\mathbf{x}$
- $\mathbf{\Lambda_y}$: matrice des coefficients reliant $\mathbf{y}$ à $\boldsymbol{\eta}$ appelés *loadings* et notés π,
- $\mathbf{\Lambda_x}$: matrice des coefficients reliant $\mathbf{x}$ à $\boldsymbol{\xi}$ appelés *loadings* et notés π,
- $\mathbf{\Theta}_\epsilon$: matrice de covariance de $\boldsymbol{\epsilon}$,
- $\mathbf{\Theta}_\delta$: matrice de covariance de $\boldsymbol{\delta}$,
- $\boldsymbol{\zeta}$: erreurs de mesure des variables latentes endogènes,
- $\mathbf{B}$: matrice des coefficients structurels des relations entre les variables latentes endogènes,
- Γ : matrice des coefficients structurels des relations entre les variables latentes endogènes et exogènes,
- $\mathbf{\Phi}$: matrice de covariance de $\boldsymbol{\xi}$,
- $\mathbf{\Psi}$: matrice de covariance de $\boldsymbol{\zeta}$,

avec comme contraintes :

- $\boldsymbol{\epsilon}$ et $\boldsymbol{\eta}$ non corrélées,
- $\boldsymbol{\xi}$ et $\boldsymbol{\delta}$ non corrélées,
- $\boldsymbol{\epsilon}$, $\boldsymbol{\delta}$ et $\boldsymbol{\zeta}$ non corrélées,
- $\boldsymbol{\delta}$ et $\boldsymbol{\eta}$ non corrélées,
- $\boldsymbol{\epsilon}$ et $\boldsymbol{\xi}$ non corrélées.

Les données étant supposées normales multivariées, l'ensemble de l'information est contenu dans les moments de degré 1 et 2 (moyenne et covariance). On va donc estimer la matrice de covariance en utilisant les équations structurelles.

Soit $\boldsymbol{\theta}$ le vecteur des paramètres à estimer ($\boldsymbol{\theta}$ contient t paramètres, les coefficients de toutes les matrices). L'hypothèse de base du modèle général d'équations structurelles est :

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

où $\boldsymbol{\Sigma}$ est la matrice de covariance des colonnes de $\mathbf{X}$, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ la matrice de covariance écrite comme fonction des paramètres du modèle.

En utilisant une décomposition de cette matrice et la non corrélation entre les différentes variables, on obtient :

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{yy}} & \boldsymbol{\Sigma}_{\mathbf{yx}} \\ \boldsymbol{\Sigma}_{\mathbf{xy}} & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_{\mathbf{y}}(\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'+\boldsymbol{\Psi})[(\mathbf{I}-\mathbf{B})^{-1}]'\boldsymbol{\Lambda}_{\mathbf{y}}' + \boldsymbol{\Theta}_{\boldsymbol{\epsilon}} & \boldsymbol{\Lambda}_{\mathbf{y}}(\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Lambda}_{\mathbf{x}}' \\ \boldsymbol{\Lambda}_{\mathbf{x}}\boldsymbol{\Phi}\boldsymbol{\Gamma}'[(\mathbf{I}-\mathbf{B})^{-1}]'\boldsymbol{\Lambda}_{\mathbf{y}}' & \boldsymbol{\Lambda}_{\mathbf{x}}\boldsymbol{\Phi}\boldsymbol{\Lambda}_{\mathbf{x}}' + \boldsymbol{\Theta}_{\boldsymbol{\delta}} \end{bmatrix} \tag{1.14}$$

Cette matrice peut être estimée à partir de la matrice de covariance empirique notée $\mathbf{S}$. On doit donc trouver $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$ telle que $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$ et $\mathbf{S}$ soient le plus proche possible au sens d'une fonction à optimiser préalablement définie.

La fonction à optimiser aura les propriétés suivantes :
- $F(\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ est un scalaire
- $F(\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \geq 0$
- $F(\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) = 0 \quad \Longleftrightarrow \mathbf{S} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$
- $F(\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ est continue en $\mathbf{S}$ et en $\boldsymbol{\Sigma}(\boldsymbol{\theta}))$

Le plus fréquemment, on utilise comme fonction une fonction de vraisemblance (qui se rapproche de celle utilisée en analyse en facteurs communs et spécifiques) basée sur le maximum de vraisemblance (ML) :

$$F_{ML}(\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + tr(\mathbf{S} \cdot \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}) - \log|\mathbf{S}| - P \tag{1.15}$$

Il existe d'autres méthodes permettant d'estimer les paramètres du modèle en utilisant les fonctions suivantes :
- Méthode des moindres carrés généralisés (GLS) :

$$F_{GLS} = \frac{1}{2}Tr(\mathbf{S}^{-1}(\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta}))^2) \tag{1.16}$$

- Méthode des moindres carrés non pondérés (ULS) :

$$F_{ULS} = \frac{1}{2}Tr((\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta}))^2) \tag{1.17}$$

- Méthode des moindres carrés pondérés (notée WLS ou ADF) :

$$F_{ADF/WLS} = (\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta}))^T\mathbf{W}^{-1}(\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})) \tag{1.18}$$

  où la matrice $\mathbf{W}$ est une matrice de poids. Elle est composée des covariances des éléments de $\mathbf{S}$.

Le maximum de vraisemblance est généralement préféré pour ses propriétés asymptotiques fortes. Le désavantage vient de l'hypothèse de multinormalité des données, celle-ci doit donc être vérifiée. Nous notons, lorsque c'est nécessaire, LISREL-ML pour la méthode LISREL estimée par le maximum

de vraisemblance (LISREL-GLS, LISREL-ULS ou LISREL-WLS pour les autres méthodes).
La méthode LISREL-WLS serait plus avantageuse si elle ne nécessitait pas de très grands jeux de données.
Nous verrons que LISREL-ULS a de nombreux avantages dans le cadre du chapitre 2.2.
Chacune de ces méthodes possède une littérature développée, pour LISREL-GLS et LISREL-ULS, on peut voir Bollen (1989), le cas de LISREL-WLS est traité en particulier dans le cas de données non normales ou ordinales qui sera développé plus loin.

Avant toute utilisation de la méthode LISREL, il faut vérifier l'unidimensionnalité des blocs de variables (voir chapitre 3.2).

1.3.2 La représentation graphique

Comme pour PLS, les équations structurelles peuvent être représentées par un *path diagram* c'est-à-dire un graphe orienté dans lequel les variables latentes sont représentées par des cercles et les variables manifestes par des carrés. La figure 1.3.2 représente un modèle avec 3 variables latentes et chacune trois variables manifestes.

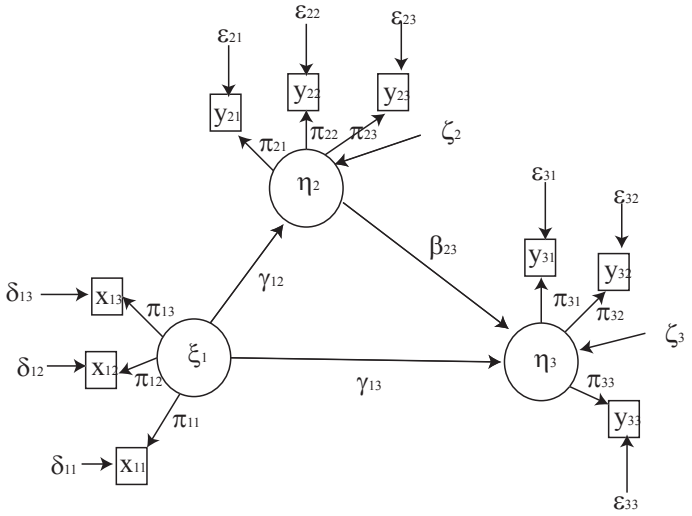

FIG. 1.5 – Modèle d'équations structurelles dans le cas de LISREL

Ce graphe représente donc un modèle structurel défini par les équations suivantes :
– On a deux variables endogènes $\boldsymbol{\eta}_2$ et $\boldsymbol{\eta}_3$, on a donc :

$$\boldsymbol{\eta}_2 = \gamma_{12}\boldsymbol{\xi}_1 + \boldsymbol{\zeta}_2$$

$$\boldsymbol{\eta}_3 = \gamma_{13}\boldsymbol{\xi}_1 + \beta_{23}\boldsymbol{\eta}_2 + \boldsymbol{\zeta}_3$$

– Les variables manifestes de la variable latente exogène ξ_1 sont définies par :

$$\mathbf{x}_{1j} = \pi_{1j}\boldsymbol{\xi}_1 + \boldsymbol{\delta}_{1j}, \qquad j = 1, 2, 3$$

– Les variables manifestes des variables latentes endogènes sont définies par :

$$\mathbf{y}_{kj} = \pi_{kj}\boldsymbol{\eta}_k + \boldsymbol{\epsilon}_{kj}, \qquad k = 2, 3, \qquad j = 1, 2, 3$$

1.3.3 Qualité d'ajustement du modèle aux données

Avant toute interprétation d'indices de qualité d'ajustement, une vérification des résultats est nécessaire pour éviter les solutions erronées. En effet, dans le cas de l'estimation par LISREL, des problèmes d'identification du modèle et de convergence des méthodes d'estimation apparaissent. Certains indices permettent de repérer ces problèmes :

- des variances négatives
- des |corrélations| > 1.00
- des matrices de corrélations ou de covariances singulières non définies positives.

Nous reviendrons plus en détails sur ces problèmes par la suite.

Une fois ces premières vérifications effectuées, on peut alors évaluer la qualité d'ajustement globale du modèle.

La littérature sur les critères de qualité d'ajustement du modèle est extrêmement riche. Au cours de ces 20 dernières années un grand nombre d'indices sont apparus et beaucoup d'articles ont été publiés sur le sujet dans des revues principalement axées sur les sciences sociales comme *Sociological Review*, *Multivariate Behavorial Research*, *Structural Equation Modeling* ou encore *Psychometrika*.
Les quelques indices, dont nous allons parler, sont ceux qui sont les plus robustes et les plus "capables" de juger de la qualité d'un modèle d'équations structurelles. Ce sont ceux qui, à la date d'aujourd'hui, sont les plus utilisés dans la recherche pour les modèles structurels à variables latentes par analyse de la structure de covariance.

Pour une présentation exhaustive des différents indices absolus et relatifs, on peut voir Yoder (1998). Dans ce mémoire, des comparaisons théoriques et empiriques ont permis de sélectionner les critères les plus avantageux.

Tanaka (1993) a mis au point 6 dimensions pour classer les critères :

1. Dépendant ou non de la taille de l'échantillon
2. Favorisant la simplicité ou la complexité
3. Indices basés sur une population ou indices basés sur un échantillon
4. Dépendant ou non de la méthode d'estimation
5. Normé ou non
6. Absolu ou relatif

Nous présentons les indices et les classons dans le sens de Tanaka (1993) dans le tableau 1.2.

- **Le test du χ^2 :** Jöreskog (1967) a présenté dans le cadre des modèles structurels un test simple dans le cas de l'utilisation du maximum de vraisemblances :

$$\begin{aligned} H_0 : &\quad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) \\ H_1 : &\quad \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}(\boldsymbol{\theta}) \end{aligned}$$

  Pour tester cette hypothèse, on utilise un rapport de vraisemblance

$$\Lambda = \frac{\max_{\Omega_0} L(\boldsymbol{\theta})}{\max_{\Omega} L(\boldsymbol{\theta})} \tag{1.19}$$

  où Λ représente la statistique du rapport de vraisemblances, $\boldsymbol{\theta}$ le vecteur des paramètres libres à estimer. Le numérateur est la fonction du maximum de vraisemblance (qui est le produit de la densité normale de probabilité sous l'hypothèse d'indépendance et de distribution normale des variables aléatoires $\mathbf{Z}$) sous l'hypothèse nulle, et le dénominateur représente la fonction du maximum de vraisemblance sans restriction. L'espace des paramètres pour le numérateur

(Ω_0) est de dimension t, et l'espace sans restriction des paramètres (Ω) a une dimension de $s = (1/2)(P_{endo} + P_{exo})(P_{endo} + P_{exo} + 1)$ où P_{endo} (resp. P_{exo}) est le nombre de variables manifestes associées aux variables latentes endogènes (resp. exogènes).

Propriété 1.5. *Sous certaines conditions de régularité et avec un grand échantillon, la statistique suivante :*

$$T \equiv -2ln(\Lambda) = (N-1)(ln|\hat{\mathbf{\Sigma}}(\boldsymbol{\theta})| - ln|\mathbf{S}| + Tr(\mathbf{S}\hat{\mathbf{\Sigma}}(\boldsymbol{\theta})^{-1}) - (P_{endo} + P_{exo})) \tag{1.20}$$

a une distribution χ^2_v où $v = s - t$ est le nombre de degrés de liberté du modèle.

Inconvénients :
De nombreuses hypothèses sont nécessaires et celle sur la normalité tient rarement (Satorra et Bentler, 1999). La méthode d'estimation des paramètres par maximum de vraisemblance n'est que très peu biaisée par la présence de non normalité alors que le calcul de cet indice est très influencé. Il aura tendance à rejeter trop de modèles. (Meijer, 1998; West et al., 1995)
Ceci a poussé les chercheurs à mettre au point d'autres indices, souvent basés sur le χ^2, mais minimisant ses défauts.

– **Le test statistique du χ^2 standardisé**
Pour un modèle M_t, il a la forme :

$$T_t^s = \frac{T_t - v_t}{\sqrt{2v_t}} \tag{1.21}$$

où T_t est la statistique du χ^2 et v_t le nombre de degrés de liberté. Il calcule la déviation entre l'estimation à partir du modèle et la covariance réelle avec une unité de déviation standard (Bollen, 1986). Il se trouve sur l'intervalle $]-\infty, \infty[$.
Il met en balance la qualité du modèle et sa complexité et pénalise la trop grande complexité du modèle.

– **Le test statistique du χ^2 de Satorra-Bentler** (Satorra et Bentler, 1999)
Pour un modèle M_t, il a la forme :

$$T_t^{SB} = c^{-1}T_t \tag{1.22}$$

où c^{-1} représente un facteur de correction qui est l'inverse de l'aplatissement.
Il se trouve sur l'intervalle $[0, \infty[$. Un modèle théorique d'intérêt est acceptable quand il a une p-valeur relativement grande (p $> .05$) une fois comparé à la variable aléatoire (centrale) χ^2 avec v_t degrés de liberté.
Ce test est spécialement bien adapté aux données déviant de la normalité.

– **La qualité d'ajustement (GFI)**
Elle est définie par :

$$GFI = 1 - \frac{Tr(\mathbf{W}^{-1/2}(\mathbf{S} - \hat{\mathbf{\Sigma}})\mathbf{W}^{-1/2})^2}{Tr(\mathbf{W}^{-1/2}\mathbf{S}\mathbf{W}^{-1/2})^2} \tag{1.23}$$

où $\mathbf{W}$ représente la matrice des poids qui dépend de la méthode d'estimation utilisée. Par exemple, pour LISREL-ML, $\mathbf{W} = \hat{\mathbf{\Sigma}}$, pour LISREL-GLS, $\mathbf{W} = \mathbf{S}$.
Le GFI donne la proportion d'information expliquée par la matrice $\mathbf{S}$. Elle se trouve dans $[0, 1]$. Elle n'est pas directement reliée à la taille de l'échantillon.
Empiriquement, le modèle est généralement accepté si $GFI \geq 0.9$.

– **La qualité d'ajustement ajustée (AGFI)**
Elle est définie par :

$$AGFI = 1 - (1 - GFI)\frac{(P_{endo} + P_{exo})(P_{endo} + P_{exo} + 1)}{2v_t} \tag{1.24}$$

L'AGFI mesure la proportion de l'information (ajustée aux degrés de liberté) dans la matrice S de covariance d'échantillon qui est expliquée par le modèle, et se trouve sur l'intervalle $[0, 1]$. Elle est souvent très proche du GFI ; elle défavorise les modèles trop complexes.

– **Racine carrée de la moyenne des erreurs d'approximations (RMSEA)**
Cet indice est calculé à partir de la fonction $\tilde{F}$ qui est la fonction à optimiser pour l'ensemble de la population. L'utilisation de cette fonction le différencie des autres indices basés sur la fonction à optimiser au niveau de l'échantillon. Pour un modèle M_t, le RMSEA a la forme :

$$RMSEA = \sqrt{\frac{\tilde{F}}{v_t}} \tag{1.25}$$

où $\tilde{F} = Tr(\mathbf{\Sigma}\,\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}) - P + ln(det\,\mathbf{\Sigma}(\boldsymbol{\theta})) - ln(det\,\mathbf{\Sigma})$
Cet indice d'ajustement mesure la racine carrée de la déviation moyenne de la statistique du χ^2 de sa valeur prévue par degré de liberté, et il se trouve sur l'intervalle $[0, \infty[$.

Son approximation en utilisant S donne :

$$\widehat{RMSEA} = \sqrt{\frac{\hat{F}}{v_t} - \frac{1}{N-1}} = \sqrt{\frac{\hat{\lambda}}{v_t(N-1)}}$$

où $\hat{F} = Tr(\mathbf{S}\,\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}) - P + ln(det\,\mathbf{\Sigma}(\boldsymbol{\theta})) - ln(det\,\mathbf{S})$.
Empiriquement, on tend à accepter le modèle si $RMSEA \leq 0.05$.

Cet indice est l'un des plus utilisés, il a été développé par Steiger et Lind (1980) et amélioré par Browne et Cudeck (1993) et Steiger (2000).
Il a l'avantage sous certaines conditions de suivre une loi de probabilité du χ^2 non central remis à l'échelle avec v_t taille de l'échantillon et λ paramètre de non centralité. On peut donc obtenir un intervalle de confiance :

$$CI = (\sqrt{\frac{\hat{\lambda}_L}{v_t(N-1)}}; \sqrt{\frac{\hat{\lambda}_U}{v_t(N-1)}}) \tag{1.26}$$

Cependant, cette hypothèse de distribution repose sur un aplatissement multivarié pas trop fort, une bonne taille d'échantillon et une erreur d'approximation qui ne doit pas être trop grande par rapport à l'erreur d'estimation.
Dans Curran et al. (2003), il ressort que le RMSEA est surestimé en cas de petits échantillons ainsi qu'en cas de non normalité des données. Sauf en cas de très mauvaise spécification, les intervalles de confiance restent bons.

– **Indice d'ajustement non normé (NNFI)**
On l'appelle aussi indice de Tucker-Lewis, TLI, NNIFI ... Pour un modèle M_t, il a la forme :

$$NNFI = \frac{\frac{T_0}{v_0} - \frac{T_t}{v_t}}{\frac{T_0}{v_0} - 1} = \frac{\frac{F_0}{v_0} - \frac{F}{v_t}}{\frac{F_0}{v_0} - \frac{1}{n-1}} \tag{1.27}$$

avec T_0 statistique du χ^2 pour le modèle de référence et F_0 fonction à optimiser pour le modèle de référence.
Il mesure l'augmentation de la qualité d'ajustement lorsque l'on passe du modèle de référence (*null model*, M_0, Rust et al. (1995)) au modèle étudié. Il se trouve en général dans l'intervalle $[0, 1]$.
On pose en général que si $NNFI \geq 0.95$, alors le modèle est accepté.

– **Indice d'ajustement normé (NFI)**
Pour un modèle M_t, il a la forme :

$$NFI = \frac{T_0 - T_t}{T_0} \tag{1.28}$$

– **Indice comparatif d'ajustement (CFI)**
Pour un modèle M_t, il a la forme :

$$CFI = 1 - \frac{\max\{T_t - v_t, 0\}}{\max\{T_t - v_t, T_0 - v_0, 0\}} = \frac{[(N-1)F_0 - v_0] - [(N-1)F - v_t]}{[(N-1)F_0 - v_0]} \tag{1.29}$$

où F_0 est le F minimum pour le modèle indépendant et v_0 est le degré de liberté maximal de ce même modèle. Empiriquement, le modèle est accepté si $CFI \geq 0.9$.
Cet indice compare le modèle étudié au modèle d'indépendance complète.

Récapitulatif sur les indices de qualité d'ajustement

Dans le tableau 1.2, nous avons rassemblé les différents critères vus jusqu'ici et nous les avons identifiés avec les 6 dimensions définies par Tanaka (1993). Les deux premières colonnes représentent la dépendance directe ou indirecte à la taille de l'échantillon. Celle-ci a été définie par Bollen (1990b). Les 3 colonnes suivantes déterminent l'influence de la complexité du modèle :
– La colonne 3 (Nb.VL) représente la relation de l'indice au nombre de variables latentes
– La colonne 4 (VM/VL) représente la relation de l'indice au nombre de variables manifestes par bloc
– La colonne 5 (*Load.*) montre s'il y a une relation ou non entre le critère et la taille des *loadings* des facteurs.

Les colonnes 6 (distr.) et 7 (estim.) indique s'il y a une relation ou non entre le critère et, respectivement, la distribution des données et la méthode d'estimation (LISREL-ML ou LISREL-GLS). La colonne 8 (Normé) indique si le critère est normé ou non, c'est-à-dire s'il est défini sur un intervalle $[0, 1]$. La dernière colonne (Abs./Rel.) permet de savoir si le critère est relatif (R) ou absolu (A), c'est-à-dire si le calcul de l'indice dépend d'un modèle dit *null model* ou non.

Indice	Dép. de N		Complexité			Rel. à		Normé	Abs./Rel.
	Direct	Indirect	Nb.VL	VM/VL	*Load.*	distr.	estim.		
T_t	O	–	O	N	O	O	O	N	A
T_t^s	O	–	–	–	–	O	–	N	A
T_t^{SB}	O	–	–	–	–	N	–	N	A
GFI	N	O	O	–	O	O	O	O	A
AGFI	N	O	O	O	N	O	–	O	A
RMSEA	O	–	N	O	–	–	N	N	A
NNFI	O	–	N	O	N	–	N	O	R
NFI	N	O	O	O	O	N	O	O	R
CFI	O	–	N	O	O	N	O	O	R

Tab. 1.2 – Récapitulatif des propriétés de chaque indice (O : oui, N : non)

Par ailleurs, la validation croisée (Browne et Cudeck, 1989) et le bootstrap peuvent être utilisés (Bollen et Stine, 1993). En effet, l'évolution des capacités en matière de calcul a permis de mettre en place des techniques de rééchantillonnage. A partir d'un échantillon, on va recréer un grand nombre d'échantillons composés d'éléments de l'échantillon original et estimer les paramètres du modèle sur ces échantillons "rééchantillonnés". Ceci nous permettra de traiter des données en s'affranchissant des

contraintes de distributions.
D'autre part, la majorité des indices ont une distribution inconnue, on ne peut donc pas obtenir d'intervalle de confiance ou effectuer de tests d'hypothèses. Les techniques de bootstrap permettent de résoudre ces problèmes.

1.3.4 L'adéquation aux données particulières

Dans le cas de données réelles, la plupart des études se fait sur des données mixtes à la fois continues et catégorielles et fréquemment non normales. Lee et al. (1995) ont montré qu'utiliser un traitement classique sur ce type de données comme si elles étaient toutes continues pouvait amener à des conclusions erronées.

Les données mixtes

La méthode la plus connue pour étudier les données catégorielles ordonnées est celle de Jöreskog et al. (2000). Elle est basée sur les corrélations polychoriques.

On considère que les variables catégorielles ont une distribution continue sous-jacente. On aura :

$$x = l \quad \Leftrightarrow \quad \tau_{l-1} < \tilde{x} < \tau_l$$

avec $\tilde{x}$ variable continue sous-jacente et τ_l seuil à calculer. La variable continue $\tilde{x}_i$ est supposée de distribution normale. On note $\phi(u)$ et $\Phi(u)$ respectivement les fonctions de densité et de probabilité de $\tilde{x}$. On a :

$$\pi_l = Pr(x = l) = Pr(\tau_{l-1} < \tilde{x} < \tau_l) = \int_{\tau_{l-1}}^{\tau_l} \phi(u)du = \Phi(\tau_l) - \Phi(\tau_{l-1})$$

$$\Longrightarrow \quad \tau_l = \Phi^{-1} \sum_{i=1}^{l} \pi_i$$

Cette valeur peut être estimée par :

$$\hat{\tau}_l = \Phi^{-1} \sum_{i=1}^{l} p_i$$

avec p_i pourcentage des réponses pour la modalité i.
L'estimation de ces seuils permet de calculer les corrélations polychoriques. La matrice de corrélation polychorique entre deux variables catégorielles se calcule en supposant que les variables continues sous-jacentes suivent une distribution normale bivariée. A partir de cette distribution, les corrélations entre les différentes modalités peuvent être calculées (Jöreskog et al., 2000). Par des calculs, en se basant sur la distribution des variables, on obtient ces corrélations entre chaque paire de variables. A partir des matrices de corrélations polychoriques, on estime la matrice de corrélation asymptotique W en utilisant les moindres carrés pondérés (LISREL-WLS). Cette méthode est efficace à condition d'avoir de "grands" jeux de données et un modèle relativement simple. Pour une explication claire, on peut voir Aris (2001, p. 64).

D'autres méthodes existent et s'inspirent de cette dernière. Ainsi dans Lee et al. (1992) et dans Lee et al. (1995), les auteurs ont développé des méthodes utilisant les corrélations polychoriques et la matrice de covariance asymptotique, mais ils utilisent une méthode d'estimation basée sur le maximum de vraisemblance et une méthode alternative pour l'estimation des seuils.

Les données non normales

Les variations des variables manifestes sont totalement résumées dans la matrice de covariance lorsque les variables sont distribuées selon une loi normale multivariée. En présence de non normalité, on aura besoin d'informations venant de moments d'ordres plus grands.

Nous pourrons donc traiter les données non normales :

1. soit en utilisant les méthodes avec hypothèse de distribution (LISREL-ML ou LISREL-GLS) tout en sachant que les estimations des paramètres restent consistantes et non biaisées mais ne sont plus efficaces. Ceci entraînera donc :
   - Le test du χ^2 rejettera trop de "bons" modèles.
   - Les tests sur toutes les estimations des paramètres seront biaisés, donnant trop de résultats significatifs.

   Dans Lei et Lomax (2005), une étude détaillée de l'effet de la non normalité des données est effectuée. De façon à ne pas tirer de conclusions trop hâtives, les auteurs ont testé des données avec différents degrés d'aplatissement et d'asymétrie, pour des tailles d'échantillons variables, avec les méthodes d'estimation LISREL-ML et LISREL-GLS. Les comparaisons ont été faites sur les biais et écarts types des estimations des paramètres ainsi que sur l'analyse de la variance d'un certain nombre d'indices.
   Ils arrivent aux conclusions suivantes :
   - Les écarts types des estimations des paramètres ne sont que très peu modifiés lorsqu'on fait varier la non normalité des données, quelles que soient la méthode d'estimation et la taille de l'échantillon.
   - La non normalité a un effet sur le biais des estimations des paramètres surtout lorsque l'échantillon est petit (100).
   - La non normalité a un effet significatif sur l'estimation des *loadings* et des coefficients structurels.
   - La non normalité joue un rôle plus important que la taille de l'échantillon et que la méthode d'estimation.
   - Les conditions de non normalité ont un effet significatif sur le test du χ^2.

   Il faut nuancer ces conclusions. L'étude a été effectuée sur des données simulées avec une asymétrie et un aplatissement uniformes, ce qui n'est pas le cas pour des données réelles. Ceci peut donc entraîner des résultats inattendus.

2. soit en utilisant les méthodes ne dépendant pas de la distribution des données. La méthode LISREL-WLS de Browne (1984) est la plus couramment utilisée. On utilise une matrice de poids $\mathbf{W}$ et une estimation du type LISREL-GLS. On remplace dans LISREL-GLS la matrice de poids ($\mathbf{S}^{-1}$) par une matrice $\mathbf{W}$ qui est la covariance des éléments de $\mathbf{S}$ (elle-même matrice de covariance), on obtient donc une combinaison de moments d'ordres deux et quatre. Cette matrice est appelée matrice de covariance asymptotique. Cet estimateur produit une estimation asymptotiquement non biaisée du test du χ^2, des estimations des paramètres et des écarts types. Malheureusement, il a le désavantage d'être très demandeur en temps de calcul et d'exiger un échantillon de très grande taille.
   Par ailleurs, la méthode plus classique qu'est LISREL-ULS constitue une alternative ne dépendant pas de la distribution de l'échantillon. De plus, elle converge beaucoup plus fréquemment que les autres méthodes. Malheureusement, la validation du modèle est basée sur le test classique du χ^2 qui ne permet pas de se détacher de l'hypothèse de normalité multivariée.

3. soit en utilisant la statistique du χ^2 de Satorra-Bentler aussi appelée "χ^2 mis à l'échelle" présenté précédemment (Satorra et Bentler, 1999).

4. soit en appliquant les méthodes d'équations structurelles non linéaires. En transformant les données initiales par des fonctions non linéaires, on peut rendre les données normales. On peut voir la thèse de Meijerink (1995) ou l'article de West et al. (1995).

Par ailleurs, Yuan et al. (2004) montrent dans leur article que l'aplatissement est plus important que l'asymétrie. Pour une étude comparative des estimations par le maximum de vraisemblance et par l'estimateur sans distribution asymptotique, on peut voir Gold et al. (2003).

Le traitement des données manquantes

Nous développons cet aspect dans le cadre du chapitre 6.

Le problème du calcul des scores des variables latentes au niveau de chaque observation

Dans le cadre de l'approche PLS, les scores des variables latentes sont obtenus automatiquement à la convergence de l'algorithme. A l'inverse, dans le cadre de l'analyse de la structure de covariance, les scores des variables latentes ne sont pas directement explicités. Deux méthodes existent afin de les calculer *a posteriori*.

- *Méthode de Jöreskog (2000) :* Cette approche consiste à calculer les scores des variables latentes de façon à ce qu'ils aient la même matrice de covariance que celle des variables latentes théoriques. Dans ce cas, le score de la variable latente pour chaque observation est obtenu à partir de l'ensemble des variables manifestes du modèle.
  Cette méthode est difficile à mettre en oeuvre. De plus, elle n'est pas intuitive car on calcule les scores des variables latentes à partir de l'ensemble des variables manifestes du modèle. On tente de trouver des scores dont les covariances sont similaires à celles calculées à partir du modèle. Pour un développement détaillé, on peut voir Jöreskog (2000).

- *Méthode de Tenenhaus et al. (2005) :* Elle est basée sur le calcul des scores en s'inspirant de l'approche PLS, et en remplaçant les poids externes par les *loadings* obtenus lors de l'estimation du modèle :
$$\hat{\boldsymbol{\xi}}_k = \sum_j \hat{\pi}_{kj}(\mathbf{x}_{kj} - \bar{\mathbf{x}}_{kj}) \tag{1.30}$$

Ce point a fait l'objet d'applications et de simulations qui tendent à montrer que les scores calculés par la seconde approche se rapprochent de ceux estimés par l'approche PLS (cf. chapitre 2).

1.3.5 Propriétés supplémentaires

Dans le cadre de la méthode LISREL, un problème central se pose au praticien : la non identification du modèle. Ce problème est récurrent et trouve son explication dans plusieurs problèmes liés aux matrices à estimer :

- La matrice de covariance des données $\mathbf{S}$ n'est pas définie positive, alors on ne pourra pas l'inverser dans le cadre de l'estimation par moindres carrés généralisés (LISREL-GLS), par ailleurs, dans le cadre du maximum de vraisemblance (LISREL-ML), on inverse $\boldsymbol{\Sigma}$ mais le fait que $\mathbf{S}$ ne soit pas définie positive rend la maximisation de la similarité entre $\boldsymbol{\Sigma}$ et $\mathbf{S}$ plus difficile et le modèle obtenu aura une mauvaise qualité d'ajustement.
- La matrice des poids $\mathbf{W}$ dans le cadre des approches par moindres carrés non pondérés n'est pas définie positive ;
- L'estimation de la matrice de covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ n'est pas définie positive ;
- D'autres matrices de covariances associées au modèle peuvent ne pas être définies positives.

Le fait que des matrices ne soient pas définies positives pose d'autres problèmes, par exemple sur les variances. En effet, le déterminant d'une matrice de covariance est une variance généralisée, or celui-ci est nul ou négatif si la matrice n'est pas définie positive, ce qui pose un problème comme une variance ne peut pas être négative et une variance nulle correspond à un scalaire.

Le fait qu'une matrice ne soit pas définie positive peut être expliqué par une dépendance linéaire entre certaines variables. Ceci peut être expliqué par des fluctuation d'échantillonnage, surtout lorsque l'échantillon est petit, la matrice $\mathbf{S}$ n'est alors pas définie positive (Anderson et Gerbing, 1984).

Cette propriété de la méthode LISREL posera des problèmes lorsque l'échantillon traité est de petite taille. On peut voir Chen et al. (2001) pour une étude plus détaillée des problèmes d'identification du modèle et Wothke (1993) sur les problèmes de matrices non définies positives.

1.4 Conclusion

Cette présentation des modèles d'équations structurelles à variables latentes et de leurs méthodes d'estimation nous a permis de voir à quel point les approches possibles étaient variées. Nous avons pu mettre en valeur la complexité de ces modèles et les solides propriétés théoriques de leurs méthodes d'estimation. Nous étudierons dans le cadre du dernier chapitre de cette thèse l'application de ces méthodes à l'analyse de la satisfaction des clients qui nous permettra de nous rendre compte de leur étendue d'application. Nous introduisons des comparaisons et des méthodes novatrices dans le cadre du chapitre suivant.

Chapitre 2

Quelques développements récents

Depuis le début des années 1980, de nombreux chercheurs ont consacré leurs travaux aux modèles d'équations structurelles à variables latentes. Les avancées ont surtout touché le domaine de l'analyse de la structure de covariance avec Jöreskog, Bollen, Muthén, Arbuckle... Mais elles concernent aussi l'approche PLS avec Wold, Löhmoller, Fornell, Bagozzi, Chin, Tenenhaus...

Nous commençons par comparer les deux approches d'estimation du modèle puis justifions les différences liées aux notations. Nous introduisons par la suite une manière de relier les approches PLS et LISREL, introduite par McDonald (1996). Une analyse de la sensibilité des paramètres estimés par l'approche PLS en fonction de différents facteurs est effectuée. Finalement, nous présentons des méthodes alternatives récentes et innovantes apportant de réelles avancées dans le cadre des modèles d'équations structurelles à variables latentes.

2.1 PLS et LISREL : deux méthodes d'estimation complémentaires

De nombreux travaux depuis celui mené par les créateurs des deux approches (Wold et Jöreskog, 1982) ont tenté de comparer LISREL et PLS. Nous en synthétisons les principaux aspects.

2.1.1 Deux approches complémentaires

Les techniques d'estimation des modèles d'équations structurelles font l'objet de comparaisons depuis le début des années 1980 (Wold et Jöreskog, 1982; Fornell et Bookstein, 1982; Dijkstra, 1983; Chin, 1995; Tenenhaus et Gonzalez, 2001; Kressman et Muller, 2002/2003; Tenenhaus, 2003; Vilares et al., 2005, 2007; Barroso et al., 2005).

Nous constatons par l'étude des deux approches et de l'ensemble des analyses comparatives que les méthodes LISREL et PLS sont plus complémentaires que concurrentes. Leurs principes, leurs objectifs, leurs contraintes et leurs applications sont différentes.
Nous ne désirons pas ici énumérer les points communs et les différences sous la forme d'une liste. Nous ne présentons pas non plus d'analyses basées sur des simulations comme l'ont déjà fait Vilares et al. (2005, 2007). Nous rassemblons dans les tableaux 2.1 et 2.2 les différences et points communs notables, d'une part, d'un point de vue théorique (2.1) et, d'autre part, dans le cadre d'applications empiriques (2.2). Nous en approfondissons certains illustrant particulièrement les différences entre les approches. Les comparaisons se font entre les deux méthodes classiquement utilisées : l'approche LISREL-ML et l'approche PLS mode A. Nous détaillons les propriétés d'autres méthodes d'estimation par la suite qui permettent de réduire les inconvénients de LISREL dans certains cas. La méthode LISREL-ULS

constitue une alternative qu'il faudra étudier.

Caractéristiques	LISREL-ML	PLS mode A
Objectif	Validation d'un modèle	Prévision (calcul des VL)
Mesures	Covariances	Variances
Principe d'optimisation	Global	Partiel
Estimation des paramètres	Optimale, consistante	Consistante au sens large
Validation	Tests statistiques	Indices de qualité prédictive
Sous-modèle favorisé	Modèle interne	Modèle externe
Relations du modèle de mesure	Réflectives	Réflectives et formatives
Variables latentes	Facteurs non estimés	Combinaison linéaire des variables manifestes

Tab. 2.1 – Comparaison théorique entre les approches PLS mode A et LISREL-ML

Comme le développe Chin (1995), PLS est à LISREL ce que l'analyse en composantes principales (ACP) est à l'analyse en facteurs communs et spécifiques. Le choix de l'utilisation de l'une de ces deux approches se fait de la même manière qu'entre une analyse en composantes principales et une analyse en facteurs communs et spécifiques. Les points les plus déterminants dans ce choix sont la relation recherchée entre les données et le modèle, les facteurs influant les estimations des paramètres et l'objectif de l'analyse. Par ailleurs, les deux approches différencient la notion même de variables latentes, la méthode LISREL travaille sur un facteur qui par essence est indéterminé et qui s'inscrit dans l'espace de l'ensemble des variables du modèle. L'approche PLS est basée sur des composantes générées par l'espace de leurs variables manifestes associées. Ceci explique aussi le fait que PLS mode A favorise le modèle externe et LISREL-ML, le modèle interne (Vilares et al., 2005).

La méthode LISREL doit être appliquée lorsque la théorie sous-jacente est riche et que les hypothèses imposées sont vérifiées. Cette méthode est basée sur la validation d'une théorie. Les indices de qualité qui lui sont associés sont basés sur des tests de qualité d'ajustement du modèle aux données. A l'inverse, l'approche PLS, même si elle se base sur un modèle prédéfini, n'utilise pas d'indices de qualité d'ajustement mais des indices de qualité prédictive généralement validés par rééchantillonnage.

La méthode LISREL offre une précision statistique dans le cadre d'hypothèses contraignantes alors que PLS échange l'efficacité des estimations des paramètres contre la simplicité et peu d'hypothèses contraignantes. C'est pour cette raison qu'on les oppose dans, d'une part, le *hard modeling* et, d'autre part, le *soft modeling*.

D'un point de vue pratique, la méthode LISREL possède un handicap majeur, la non détermination du modèle dans de nombreux cas (cf. chap. 1). Elle nécessite de grands échantillons et des modèles simples avec un respect des hypothèses initiales.

Les autres faiblesses, souvent avancées à l'égard de la méthode LISREL-ML, peuvent être contournées grâce aux avancées récentes dans le domaine :

- Le traitement des données non normales : l'utilisation d'une méthode d'estimation du type LISREL-WLS ou LISREL-ULS permet d'éviter certains problèmes d'identification et de biais des estimations.
- Le traitement des données ordinales : l'utilisation des corrélations polychoriques associées à LISREL-WLS permet un traitement, certes complexe, mais efficace.
- Le traitement des données manquantes : l'approche *Full Information Maximum Likelihood* permet de traiter directement le modèle avec les données disponibles (cf. chap. 6).

Caractéristique	LISREL-ML	PLS mode A
Hypothèses	Indépendance et distribution normale multivariée	-
Cas réflectif	Unidimensionnalité	Unidimensionnalité
Convergence et identification	Matrices non définies positives (cf. chap. 1)	Convergence observée (cf. chap. 1)
Taille d'échantillon nécessaire	Grande	Faible
Complexité du modèle possible pour l'identification	Faible	Elevée
Traitement des données ordinales	Corrélations polychoriques (cf. chap. 1) et LISREL-WLS	Direct (ou approche PML, cf. chap. 2.4)
Traitement des données manquantes	Méthode FIML (cf. chap. 6)	Algorithme NIPALS (cf. chap. 6)
Domaines d'application	Sociologie, psychologie, marketing,...	Marketing, analyse sensorielle,...

Tab. 2.2 – Comparaison pratique entre les approches PLS mode A et LISREL-ML

Sur le calcul des scores des variables latentes, malgré la non estimation des facteurs lors de l'application de la méthode LISREL, l'utilisation de méthodes de calcul a posteriori telle que celle de l'équation 1.30 (p. 37) permet d'obtenir des scores proches de ceux obtenus par l'approche PLS mode A (Tenenhaus et al., 2005). Par ailleurs, les approches définies par McDonald (1996) permettent d'obtenir des scores avec des corrélations de l'ordre de 0.99 avec les scores de l'approche PLS mode A (cf. chapitre 2.2, p. 42).

Finalement, Vilares et al. (2005, 2007) ont mené des études par simulation afin de comparer les estimations des coefficients pour des modèles complexes. Il ressort que PLS mode A a tendance à surestimer les *loadings* alors que LISREL-ML a tendance à les sous-estimer et que LISREL-ML surestime les coefficients structurels alors que PLS mode A les sous-estime. Les implications théoriques permettant d'expliquer ces comportements restent à trouver. Par ailleurs, il ressort aussi de cette étude que lorsque les données dévient de la normalité et, lorsque les variables manifestes sont formatives, les biais associés aux estimations par la méthode LISREL-ML augmentent.

Cette étude nous montre que dans le cadre de la validation d'une théorie et lorsque les hypothèses préalables sont vérifiées, la méthode LISREL-ML doit être utilisée. Par contre, dès que les connaissances sont plus faibles, ou que l'on désire étudier les variables latentes sous forme de scores, alors l'approche PLS sera préférée.
On voit donc que, suivant les hypothèses et la problématique, chaque approche s'applique de façon spécifique. Ainsi les méthodes PLS et LISREL s'avèrent plus complémentaires que concurrentes. Nous rassemblons dans le tableau 2.3 les problématiques et les méthodes à préférer (un (+) signifie que cette méthode est généralement plus adaptée dans ce cas). Ces conseils sont à mettre en rapport avec la problématique spécifique liée au cas pratique. De plus, les hypothèses de travail dépendent de la méthode d'estimation choisie.

Problématique	LISREL-ML	PLS mode A
Validation d'une théorie :		
théorie solide basée sur une population bien identifiée	+	-
théorie peu solide et données ne vérifiant pas les hypothèses de travail	-	+
Utilisation d'un modèle afin d'effectuer des prévisions	-	+
Analyse des coefficients structurels :		
sur des données vérifiant les hypothèses de travail	+	-
sur des données ne vérifiant pas les hypothèses de travail	-	+
Type de données et modèles traités :		
traitement d'un modèle complexe	-	+
traitement de jeux de données réduits	-	+
traitement de construits formatifs	-	+

Tab. 2.3 – Problématiques et modèles d'équations structurelles

2.1.2 Le formalisme adopté

Nous avons présenté dans le premier chapitre deux formalismes différents dans le cadre des approches d'estimation PLS et LISREL. Cette différence majeure entre les deux approches rend la communication entre leurs domaines d'applications plus complexe et défavorise les mises en relation.

L'utilisation de ces deux formalismes se justifie par la différenciation nécessaire entre les variables latentes endogènes et exogènes dans le cadre de l'approche LISREL. L'utilisation de deux types de variables latentes dans le cadre de l'approche PLS pose un problème d'estimation des modèles externe et interne. Les étapes de l'algorithme PLS doivent être doublées et la complexité de l'estimation est alors augmentée.

Dans le cadre de la méthode LISREL, le modèle externe est défini par deux types d'équations. L'utilisation d'une seule équation dans le même style que PLS nous amène à un modèle du type *Reticular Action Model* (RAM, McDonald et McArdle (1984); McDonald (1996)). Néanmoins, une différence entre variables latentes endogènes et exogènes est toujours nécessaire afin de ne pas avoir de matrices non inversibles dans le calcul de la matrice de covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ (cf. équation 1.14, p. 29).

2.2 De LISREL à PLS : un pont possible

Très peu d'études tentant de rapprocher l'approche PLS et l'analyse de la structure de covariance ont vu le jour.
L'une des principales différences théoriques entre ces deux approches réside dans la façon dont sont considérées les variables latentes. Dans le cadre de l'approche PLS, les variables latentes sont des combinaisons linéaires des variables manifestes du bloc (*composite variables*), alors que dans le cas de LISREL, ce sont des facteurs non estimés. McDonald (1996) propose différentes méthodes d'estimation du modèle, dans le cadre de l'approche LISREL, en supposant que les variables latentes sont des combinaisons linéaires des variables manifestes.

Tenenhaus (2007) a repris récemment ces travaux. Ainsi, l'estimation par une fonction ULS soumise à quelques contraintes supplémentaires permet d'obtenir des estimations très proches de celles du mode A de l'approche PLS.

McDonald (1996) utilise une formulation simplifiée du modèle structurel appelée *Reticular Action Model* (RAM, McDonald et McArdle (1984)) et suppose que chaque variable latente $\boldsymbol{\xi}_k$ est estimée

comme une somme pondérée par :

$$\mathbf{t}_k = \mathbf{w}_k' \mathbf{x}_k \tag{2.1}$$

où $\mathbf{t}_k$ est l'estimation du construit latent et $\mathbf{w}_k$ est un vecteur de poids.

Nous introduisons d'abord le modèle RAM :

Définition 2.1. *On appelle* ***reticular action model*** *ou* ***modèle RAM*** *la formulation d'un modèle d'équations structurelles la plus simple possible basée sur deux équations :*

$$\mathbf{X} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\epsilon} \tag{2.2}$$

$$\boldsymbol{\xi} = \mathbf{B}\boldsymbol{\xi} + \mathbf{f} \tag{2.3}$$

On aura :
- $\mathbf{X}$ matrice des données observées,
- $\boldsymbol{\Lambda}$ sommes directes des matrices de *loadings* associées à chaque bloc,
- $\boldsymbol{\epsilon}$ terme d'erreur associés au modèle de mesure,
- $\boldsymbol{\xi}$ matrice rassemblant l'ensemble des variables latentes du modèle,
- $\mathbf{B}$ matrice de coefficients structurels (la j$^{\text{ème}}$ ligne de $\mathbf{B}$ est composée de 0 si la j$^{\text{ème}}$ variable latente est exogène),
- $\mathbf{f}$ matrice comportant des variables latentes exogènes et les erreurs associées au modèle interne.

On définit par ailleurs, les matrices de covariances :
- $\boldsymbol{\Psi}$ matrice de covariance associée aux termes de $\mathbf{f}$,
- $\boldsymbol{\Delta}$ matrice de covariance des termes d'erreur du modèle de mesure.

Par des calculs matriciels, on obtient :

$$\boldsymbol{\Sigma} = cov(\mathbf{X}) = \boldsymbol{\Lambda} cov(\boldsymbol{\xi}) \boldsymbol{\Lambda}' + \boldsymbol{\Delta}$$

On a donc :

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\Psi} (\mathbf{I} - \mathbf{B})'^{-1} \boldsymbol{\Lambda}' + \boldsymbol{\Delta} = \boldsymbol{\Lambda}\mathbf{P}(\boldsymbol{\theta})\boldsymbol{\Lambda}' + \boldsymbol{\Delta} \tag{2.4}$$

L'estimation des paramètres $\boldsymbol{\theta}$ se fait à l'aide de la matrice de covariance estimée à partir des données $\mathbf{S}$. On peut rappeler que dans le cadre des moindres carrés non pondérés (ULS), on minimise :

$$\frac{1}{2} Tr((\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta}))^2) = \frac{1}{2} Tr((\mathbf{S} - (\boldsymbol{\Lambda}\mathbf{P}(\boldsymbol{\theta})\boldsymbol{\Lambda}' + \boldsymbol{\Delta}))^2) \tag{2.5}$$

L'ensemble des approches présentées par McDonald (1996) suppose que les variances des erreurs de mesure sont nulles $\boldsymbol{\Delta} = \mathbf{0}$.

Tenenhaus (2007) s'appuie sur l'une d'entre elles afin d'estimer un modèle structurel du type PLS avec une estimation de $\boldsymbol{\theta}$ par ULS. Ceci revient à minimiser :

$$Tr((\mathbf{S} - (\boldsymbol{\Lambda}\mathbf{P}(\boldsymbol{\theta})\boldsymbol{\Lambda}'))^2) = \|S - (\boldsymbol{\Lambda}\mathbf{P}(\boldsymbol{\theta})\boldsymbol{\Lambda}')\|^2$$

Cette approche est une généralisation de l'analyse en composantes principales au cas de plusieurs blocs. Elle maximise la qualité d'ajustement des variables à l'ensemble du modèle (aussi bien interne qu'externe) au sens des moindres carrés. La différence majeure avec l'approche ULS classique réside dans l'utilisation d'une composante à la place d'un facteur indéterminé (en fixant $\boldsymbol{\Delta} = \mathbf{0}$). Ceci revient à remplacer l'équation $\mathbf{x}_{kj} = \pi_{kj}\boldsymbol{\xi}_k + \boldsymbol{\epsilon}_{kj}$ par $\mathbf{x}_{kj} = \pi_{kj}\mathbf{t}_k$. McDonald (1996) utilise comme solution de l'equation 2.1 :

$$\mathbf{w}_k' = (\boldsymbol{\pi}_k'\boldsymbol{\pi}_k)^{-1}\boldsymbol{\pi}_k'$$

Cette approche permet d'utiliser l'ensemble des fonctionnalités issues des logiciels estimant le modèle par l'approche LISREL (contraintes sur les paramètres) tout en obtenant des estimations très proches de celles de l'approche PLS mode A.

La relation entre cette méthode et l'approche PLS mode A n'est pas directe. Cette dernière est basée sur la maximisation des covariances entre les variables latentes connectées alors que la première est basée sur une qualité d'ajustement globale. Néanmoins, les applications effectuées ont montré que ces deux approches conduisent à des résultats très proches.

Nous utilisons des données simulées afin de vérifier l'accord entre ces deux approches. Nous simulons le modèle issu de la figure 2.1 à l'aide de données normales. Nous obtenons donc des scores pour les trois variables latentes. Les graphes issus de cette comparaison se trouvent dans la figure 2.2, on voit que les scores obtenus sont très proches avec des corrélations toujours supérieures à 0.99. En terme de coefficients structurels, les biais remarqués plus tôt dans les comparaisons existent toujours.

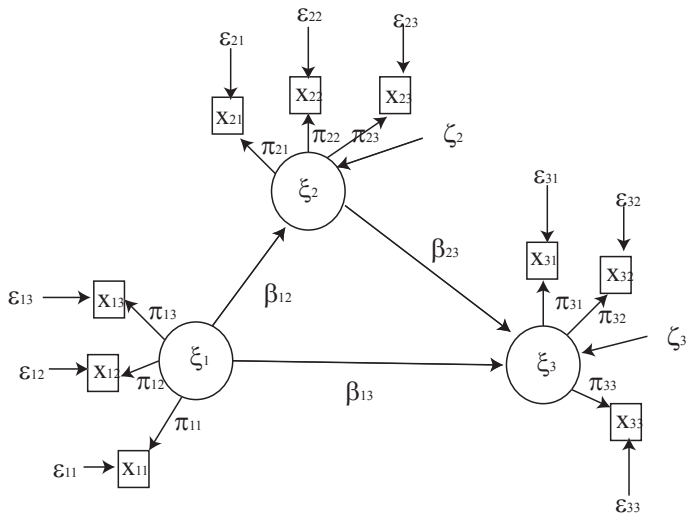

FIG. 2.1 – Modèle d'équations structurelles

Par ailleurs, McDonald (1996) tente d'obtenir une équivalence directe entre le mode A de l'approche PLS et un critère spécifique. Sa méthode est complexe et plusieurs problèmes se posent :

1. il ne prend pas en compte le schéma d'estimation du modèle interne et les problèmes de signe qui en découlent ;
2. il part d'une hypothèse non démontrée introduite par Streissguth et al. (1993) qui dit : *Mode A algorithm must converge to give rank-one least-squares fit to the cross-correlations between the individual indicators of all pairs of blocks that are directly connected by a path.* McDonald (1996) traduit cette hypothèse par la minimisation du critère :

$$\sum_{i,j:\xi_i \leftrightarrow \xi_j} Tr((\mathbf{S}_{ij} - \varpi_i \varpi_j \mathbf{w}_i \mathbf{w}_j)^2)$$

avec ϖ_i et ϖ_j solutions de l'équation :

$$\varpi_i \varpi_j = (\mathbf{w}_i \mathbf{S}_{ij} \mathbf{w}_j)(\mathbf{w}_i' \mathbf{w}_i \mathbf{w}_j' \mathbf{w}_j)^{-1}$$

en supposant $\mathbf{w}_i \mathbf{S}_{ii} \mathbf{w}_i = 1$ et $\mathbf{S}_{ij}$ matrice de covariance entre les éléments associés aux blocs i et j.

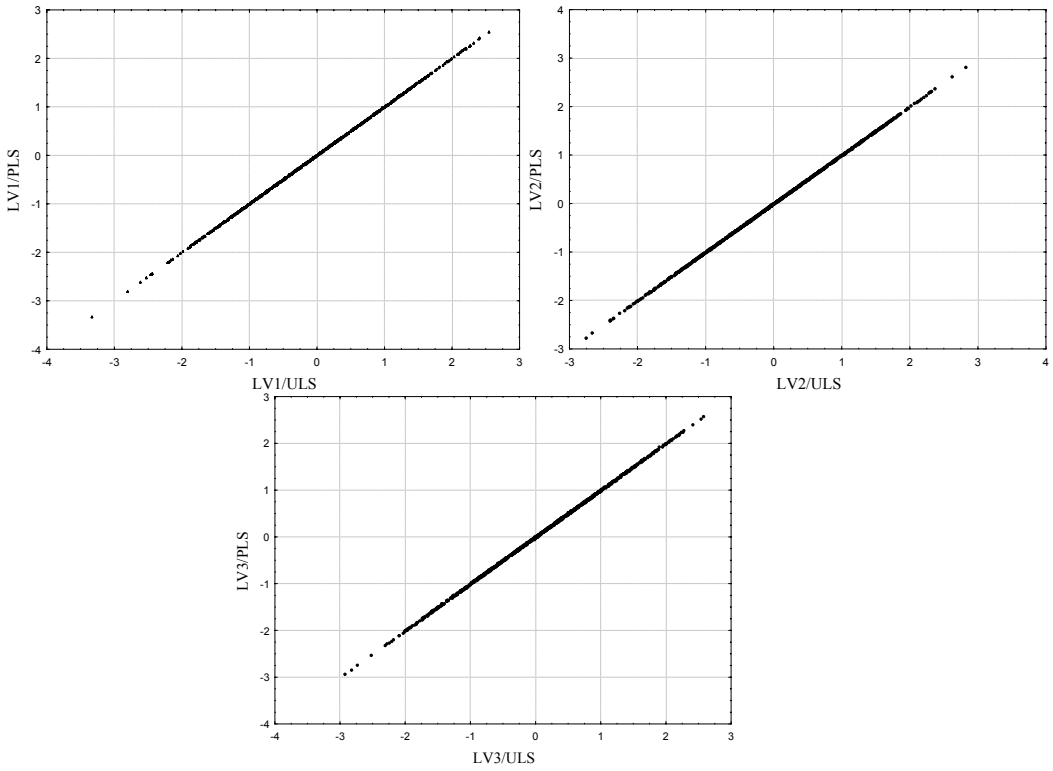

FIG. 2.2 – Comparaison des scores obtenus par l'approche PLS mode A et l'approche ULS de McDonald (1996)

Nous ne développerons pas ici cette approche car sa présentation originale n'est pas rigoureuse et de plus amples recherches sur le sujet seraient nécessaires. Cette voie de recherche reste néanmoins intéressante.

Les recherches de McDonald (1996) permettent donc d'arriver à relier PLS et LISREL et les études empiriques montrent que les scores obtenus sont très proches ainsi que les coefficients structurels. L'équivalence supposée, malgré des problèmes de démonstration et de mise en pratique, ouvre des voies de communication entre les deux méthodes d'estimation.

2.3 Une analyse de la sensibilité des estimations de l'approche PLS

L'aspect complexe de l'algorithme PLS et le manque de fonction globale à optimiser rend une étude analytique des résultats ardue. Dans ce contexte, un certain nombre de chercheurs ont tenté d'étudier dans le cadre empirique les propriétés de l'approche PLS.
Nous menons une étude de sensibilité basée sur des simulations de Monte Carlo afin de confirmer ou d'apporter de nouvelles remarques sur la sensibilité de l'approche PLS à différents facteurs.
Goodhue et al. (2006); Kristensen et al. (2003); Cassel et al. (1999, 2000); Chin et al. (1996) se sont tous attelés à simuler des modèles structurels afin de vérifier un certain nombre de propriétés non vérifiables directement. L'ensemble de ces recherches se sont basées sur le mode A d'estimation des poids externes et le schéma centroïde d'estimation des poids internes.

Jusqu'alors, les chercheurs se sont concentrés sur les effets des différentes paramétrisations de l'approche PLS, notamment sur les scores des variables latentes (Kristensen et al., 2003; Cassel et al., 1999). Cassel et al. (1999) ont étudié la robustesse de PLS à des données asymétriques, à la présence de multicolinéarité entre les variables manifestes et à une mauvaise spécification du modèle. Il ressort que les biais d'estimation des paramètres du modèle ne sont pas modifiés par ces facteurs.

Nous nous intéressons à l'influence de la complexité du modèle sur les estimations des paramètres. Nous utilisons des méthodes de Monte Carlo afin de simuler un modèle issu de l'*European Customer Satisfaction Index* (Tenenhaus et al. (2005), figure 7.4). Nous utilisons le mode A et le schéma centroïde lors de l'estimation des paramètres du modèle.

Nous étudions différents facteurs :
- Taille de l'échantillon N.
- Nombre de variables manifestes par variable latente.
- Distribution inégale du nombre de variables manifestes par bloc.

Le processus de simulation est basé sur le principe classique de la simulation de modèles réflectifs. Ainsi, les variables latentes exogènes sont simulées en utilisant une distribution prédéfinie, classiquement la distribution normale. Par ailleurs, les termes d'erreurs sont simulés. On utilise les équations du modèle de mesure et du modèle structurel afin d'obtenir les variables manifestes (cf. équations 1.1 et 1.3, p. 20 et 21). Les variables manifestes ainsi obtenues sont mises sur des échelles de 1 à 10. Kristensen et al. (2003) ont montré que la distribution choisie n'avait pas d'impact dans le cadre de l'approche PLS.

2.3.1 Taille de l'échantillon et nombre de variables manifestes par bloc

Ces deux paramètres sont reliés par la propriété de consistance au sens large. Comme nous l'avons vu plus tôt, il n'y a pas de démonstration rigoureuse de cette propriété. Nous allons donc tenter de la vérifier par des simulations.

Afin d'évaluer, dans le cadre d'un modèle complexe, les biais induits, nous simulons des données avec $N = 20, ..., 10000$ (9 modèles) et $p_k = 2, ..., 10$ (9 modèles) sur le modèle ECSI. Chaque cas est appliqué 100 fois. On obtient donc $9 \times 9 \times 100 = 8100$ jeux de données.
Nous comparons les coefficients structurels, les *loadings* et les significativités de ces coefficients. Ces dernières sont obtenues par des méthodes de type bootstrap.

Dans la figure 2.3, nous présentons les résultats pour un coefficient structurel de valeur simulée 0.6 en fonction de la taille de l'échantillon (4 variables par bloc fixé) et du nombre de variables par bloc (500 observations fixées).

Il apparaît que pour un échantillon de petite taille, les biais induits par l'estimation sont importants ainsi que les écarts types. Pour plus de 200 cas, ces biais sont clairement négatifs mais se stabilisent.

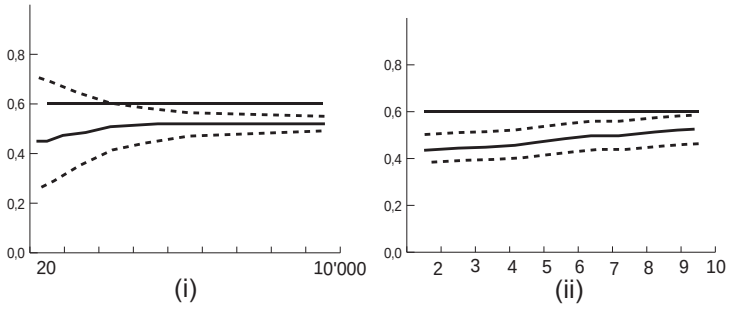

FIG. 2.3 – Evolution de la moyenne d'un coefficient structurel du modèle (avec l'écart type) en fonction de (i) la taille de l'échantillon et (ii) du nombre de variables manifestes par bloc

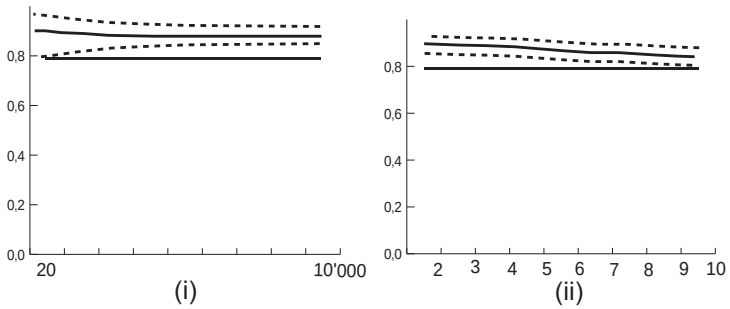

FIG. 2.4 – Evolution de la moyenne d'un *loading* du modèle (avec l'écart type) en fonction de (i) la taille de l'échantillon et (ii) du nombre de variables manifestes par bloc

En ce qui concerne le nombre de variables manifestes dans chaque bloc, les biais toujours négatifs décroissent constamment. Les écarts types baissent à mesure que la taille de l'échantillon augmente mais ne dépendent pas du nombre de variables manifestes.
La significativité dépend de la taille de l'échantillon mais pas du nombre de variables manifestes.

Par ailleurs, les *loadings* ont le même comportement que les coefficients structurels avec un biais positif et le même type de conclusions peuvent être apportées. Nous illustrons, dans la figure 2.4, le cas d'une variable manifeste associée à la variable latente satisfaction avec un *loading* fixé à 0.8.

La principale conclusion vient du fait qu'un échantillon minimal est nécessaire afin d'obtenir des coefficients structurels utilisables. Pour les *loadings*, les résultats sont constants pour toutes les tailles d'échantillon. Par contre, au-delà de ce seuil, la taille de l'échantillon n'a pas d'impact sur les biais d'estimations (ceci concorde avec les résultats de Kristensen et al. (2003)).

Lorsqu'on fait varier ces deux paramètres simultanément, les biais réduisent plus rapidement que lorsqu'on travaille sur l'un des deux facteurs indépendamment. Cependant, même avec un échantillon de taille 10000 et 10 variables manifestes par bloc, les coefficients structurels restent biaisés négativement et les *loadings* positivement. Nous illustrons ce cas pour le coefficient structurel entre satisfaction et fidélité. Dans le tableau 2.4, nous rassemblons les biais moyens obtenus en croisant la taille de l'échantillon et le nombre de variables manifestes par bloc.

L'étude des estimations de ces coefficients dans le cadre de petits échantillons est possible mais le chercheur devra prendre en compte le fait que ceux-ci peuvent être largement biaisés.

Les études effectuées précédemment ont montré que les effets trouvés ici n'apparaissent pas au

	2	3	4	6	8
20	0.125	0.123	0.121	0.086	0.081
50	0.132	0.124	0.101	0.074	0.078
100	0.128	0.118	0.096	0.085	0.071
200	0.139	0.114	0.097	0.086	0.073
500	0.134	0.114	0.095	0.076	0.070
1000	0.139	0.113	0.097	0.075	0.068

TAB. 2.4 – Biais moyens des estimations d'un coefficient structurel en fonction de la taille de l'échantillon et du nombre de variables manifestes par bloc

niveau du calcul des scores des variables latentes (Cassel et al., 1999).

2.3.2 Distribution inégale du nombre de variables manifestes

Nous développons un processus de simulation afin de tester l'effet d'une distribution inégale du nombre de variables manifestes par bloc. Nous utilisons 24 variables manifestes (4 par bloc) et nous modifions le nombre de variables manifestes par bloc (entre 2 et 10). Nous nous focalisons sur les relations structurelles entre les variables latentes exogènes (image) et endogènes (satisfaction et fidélité). Nous obtenons dix jeux de données (répliqués chacun 100 fois). Nous rassemblons les résultats dans le tableau 2.5. Les valeurs entre parenthèses sont les intervalles de confiance et lorsqu'on a 10 variables manifestes pour une variable latente, alors les autres variables latentes du modèle ont 2 variables manifestes. De la même façon, lorsqu'on a deux variables manifestes sur un bloc alors on en aura entre 5 et 6 dans chacun des autres blocs.

Cas	Biais des coef. structurels	
	Image-Fidélité	Satisfaction-Fidélité
Equidistribution	-0.05 (-0.15 ;0.05)	-0.14 (-0.23 ;-0.05)
10 VM sur l'image	-0.05 (-0.16 ;0.03)	-0.21 (-0.30 ;-0.14)
10 VM sur la satisfaction	-0.10 (-0.19 ;-0.02)	-0.24 (-0.35 ;-0.16)
10 MV sur la fidélité	-0.10 (-0.18 ;-0.04)	-0.12 (-0.22 ;-0.07)
2 VM sur l'image	-0.11 (-0.17 ;-0.04)	-0.14 (-0.23 ;-0.05)
2 VM sur la satisfaction	-0.07 (-0.14 ;0.00)	-0.18 (-0.25 ;-0.11)
2 VM sur la fidélité	-0.12 (-0.20 ;-0.03)	-0.13 (-0.22 ;-0.06)

TAB. 2.5 – Biais moyens (avec intervalles de confiance) des estimations des coefficients structurels en fonction de la distribution du nombre de variables manifestes dans chaque bloc (VM : variables manifestes)

Le tableau 2.5 montre que lorsqu'on travaille sur un modèle équidistribué, le biais est faible. Le biais le plus fort apparaît lorsque la satisfaction a 10 indicateurs (alors que les autres blocs en ont 2). Cependant, les différences obtenues ne sont pas significatives lorsqu'on utilise les intervalles de confiances associés (ils se superposent). Les résultats sont similaires pour les autres coefficients structurels.

Il ressort donc que cet aspect n'a pas d'impact sur le biais des estimations. Néanmoins, cette distribution inégale pourra avoir un impact sur l'interprétation du modèle estimé.

2.3.3 Conclusion

Cette courte étude donne des indices afin de bien analyser les estimations obtenues par l'approche PLS. Elle confirme les résultats de Cassel et al. (1999), Chin et al. (1996) et Kristensen et al. (2003) sur les biais relatifs à la taille de l'échantillon et au nombre de variables manifestes par bloc. Elle va dans le même sens que Dijkstra (1981) sur la consistance au sens large. On se rapproche bien de la valeur simulée du paramètre mais même pour des très grands échantillons et un grand nombre de variables manifestes, on a toujours un biais significatif. On ne peut pas conclure que les estimations des paramètres atteignent la valeur réelle du paramètre à l'infini.
D'autre part, cette étude montre que de fortes inégalités dans la distribution du nombre de variables manifestes dans les blocs affectent principalement l'interprétation du modèle et n'ont pas d'effet sur les biais d'estimation.
L'ensemble de ces résultats est difficile à généraliser, mais peut nous amener à quelques conseils. Pour un modèle aussi complexe que celui étudié, afin d'obtenir des coefficients interprétables, un échantillon d'au moins 200 observations est conseillé, les variables latentes doivent au moins avoir 4 indicateurs chacune. Ces valeurs restent faibles en rapport avec celles de l'approche LISREL-ML. Celle-ci nécessite dans le cas de données déviant de la normalité multivariée, au moins 10 observations par paramètre à estimer (ceci nécessite de l'ordre de 1000 observations dans le cadre du modèle ECSI, Bentler et Bonnet (1980)).
Cette recherche confirme les règles générales (*rule of thumbs*) mises en places dans le cadre des nombreuses simulations effectuées sur les biais des estimations des coefficients structurels (Kristensen et al., 2003) :

– La variabilité (σ représente la variance des variables manifestes) :

$$biais(k\sigma, N, P) = k \times biais(\sigma, N, P)$$

– La taille de l'échantillon (N) :

$$biais(\sigma, kN, P) = \frac{1}{\sqrt{k}} \times biais(\sigma, N, P)$$

– Le nombre de variables manifestes (P) :

$$biais(\sigma, N, kP) = \frac{1}{k} \times biais(\sigma, N, P)$$

2.4 Autres méthodes

D'autres méthodes traitant des modèles structurels à variables latentes ont vu le jour récemment, nous nous attardons sur deux d'entre elles très intéressantes. La première est adaptée au cas de données sur des échelles variées et la seconde permet d'avoir une vision différente du modèle.

2.4.1 L'approche Partial Maximum Likelihood (PML)

Cette approche initialement développée par Derquenne (2005) et approfondie dans Derquenne (2006), Jakobowicz et Derquenne (2007) permet de traiter des données ayant tous types d'échelles.

Cette approche s'appuie sur des modèles linéaires généralisés combinés à l'approche PLS classique. Elle est basée sur la même structure que l'approche PLS en se référant au principe du point fixe (Wold, 1980a). Les estimations par régressions linéaires ordinaires sont remplacées en utilisant des estimations

par des méthodes adaptées aux échelles des variables manifestes. Cette méthode utilise, dans l'estimation externe des variables latentes, des modèles linéaires généralisés plutôt que des régressions. Elle permet de traiter des données aussi bien ordinales que nominales qui posent des problèmes à l'approche PLS. Les études effectuées jusqu'alors ont montré que les résultats étaient similaires entre PLS et PML pour des données continues et catégorielles ordonnées mais que l'approche PML dépasse l'approche PLS pour des données binaires et nominales (Jakobowicz et Derquenne, 2007).
Nous présentons son principe et nous l'utiliserons dans le cadre du chapitre 6.4 (p. 124) et dans les applications 7.7 (p. 165).

L'approche Partial Maximum Likelihood (PML) se décompose en trois étapes principales :

1. La quantification des variables latentes qui équivaut à l'estimation externe initiale de l'algorithme PLS.
2. L'algorithme itératif qui équivaut aux itérations de l'algorithme PLS.
3. L'estimation des coefficients du modèle par régressions simples ou multiples.

Tout au long de l'application de l'approche PML, il faut connaître le type d'échelle associé à chaque variable manifeste (continue, ordinale, nominale, binaire).

Etape de quantification

Afin d'obtenir une première estimation des variables latentes à partir des variables manifestes, il faut sélectionner un poids externe initial pour chaque relation. Celui-ci n'est pas choisi aléatoirement comme dans le cadre de l'approche PLS. Nous définissons donc la notion de variable de référence.

Nous savons que $w_{kj} = cov(\mathbf{x}_{kj}, \mathbf{z}_k) = cov(\mathbf{x}_{kj}, \sum_{i:\boldsymbol{\xi}_i \leftrightarrow \boldsymbol{\xi}_k} e_{ki}\mathbf{y}_i)$. Lors de l'étape initiale de l'algorithme, on peut supposer que $\mathbf{y}_i = \mathbf{x}_{i1}$ et dans un souci de simplicité, on pose $w^0_{kj} = cov(\mathbf{x}_{kj}, \mathbf{y}_i) = cov(\mathbf{x}_{kj}, \mathbf{x}_{i1})$. Des simulations effectuées par ailleurs montrent que cette simplification n'a pas d'effet sur les résultats de l'algorithme. De plus, le choix de la variable latente $\boldsymbol{\xi}_i$ connectée à $\boldsymbol{\xi}_k$ et de la variable manifeste de référence $\mathbf{x}_{i1}$ n'a pas d'effet sur les estimations des coefficients.
Le poids externe initial de $\mathbf{x}_{kj}$ du bloc k sera donc :

$$w^0_{kj} = cov(\mathbf{x}_{kj}, \mathbf{x}_{i1}). \tag{2.6}$$

La variable manifeste de référence $\mathbf{x}_{i1}$ peut être choisie dans n'importe quel bloc connecté au bloc k.

Dans le cadre de l'approche PML, la variable de référence $\mathbf{x}_{i1}$ correspond à la variable réponse et la variable $\mathbf{x}_{kj}$ correspond à une variable explicative.

L'estimation externe initiale $\mathbf{y}_k$ de $\boldsymbol{\xi}_k$ est faite en utilisant le mode A (schéma réflectif) et dépendra du type de variables (aussi bien de référence que la variable manifeste étudiée). Nous écrivons "la variable $\mathbf{x}$ est ajustée par la variable $\mathbf{y}$", ceci peut être reformulé par "la variable $\mathbf{x}$ est régressée par la variable $\mathbf{y}$".

1. **La variable de référence est continue et elle est ajustée par une variable continue** : On se retrouve alors dans le cas classique de l'approche PLS.

2. **La variable de référence est continue et elle est ajustée par une variable catégorielle :** Soit $\mathbf{x}_{i1}$ continue et $\mathbf{x}_{kj}$ catégorielle, alors on peut appliquer un modèle d'analyse de la variance à un effet (ANOVA).
   Pour le bloc k connecté au bloc i, nous avons :

$$\mathbf{y}_k^{(t=0)} = \sum_{j=1}^{p_k} \sum_{l=1}^{L_j} w_{kjl}^{(t=0)} \mathbf{x}_{kjl}, \tag{2.7}$$

où $\mathbf{x}_{kjl}$ est un vecteur binaire de dimension égale au nombre de modalités L_j $(l = 1, , L_j)$ dans $\mathbf{x}_{kj}$ et $w_{kjl}^{(t=0)}$ est la moyenne de $\mathbf{x}_{i1}$ sur la modalité l.

3. **La variable de référence est binaire ou ordinale et elle est ajustée par une variable continue :** Soit $\mathbf{x}_{i1}$ catégorielle (binaire ou ordinale) et $\mathbf{x}_{kj}$ continue. Dans ce cas, un modèle logit simple peut être appliqué.
   Pour le bloc k connecté au bloc i, nous avons :

$$\mathbf{y}_k^{(t=0)} = \sum_{j=1}^{p_k} w_{kj}^{(t=0)} \mathbf{x}_{kj}, \tag{2.8}$$

   où $w_{kj}^{(t=0)}$ est le coefficient de régression logistique de $\mathbf{x}_{kj}$ sur $\mathbf{x}_{i1}$ obtenu par l'estimateur du maximum de vraisemblance.

4. **La variable de référence est binaire ou ordinale et elle est ajustée par une variable catégorielle :** Soit $\mathbf{x}_{i1}$ catégorielle (binaire ou ordinale) et $\mathbf{x}_{kj}$ catégorielle. Dans ce cas, un modèle logit à un effet (données groupées) peut être appliqué.
   Pour le bloc k connecté au bloc i, nous avons :

$$\mathbf{y}_k^{(t=0)} = \sum_{j=1}^{p_k} \sum_{l=1}^{L_j} w_{kjl}^{(t=0)} \mathbf{x}_{kjl}, \tag{2.9}$$

   où $\mathbf{x}_{kjl}$ est une variable binaire associée à la modalité l et $w_{kjl}^{(t=0)}$ est le coefficient de régression logistique associé.

5. **La variable de référence est nominale et elle est ajustée par une variable continue :** Soit $\mathbf{x}_{i1}$ nominal avec R modalités et soit $\mathbf{x}_{kj}$ continue, alors on peut appliquer un modèle logit polytomique simple.
   Pour le bloc k connecté au bloc i, nous avons :

$$\mathbf{y}_k^{(t=0)} = \sum_{j=1}^{p_k} w_{kj(r)}^{(t=0)} \mathbf{x}_{kj} \text{ quand } \mathbf{x}_{i1} \text{ prend la valeur } r, \tag{2.10}$$

   où $w_{kj(r)}^{(t=0)}$ est le coefficient de la régression logistique simple généralisée de $\mathbf{x}_{kj}$ sur $\mathbf{x}_{i1}$.

6. **La variable de référence est nominale et elle est ajustée par une variable catégorielle :** Soit $\mathbf{x}_{i1}$ nominal avec R modalités et soit $\mathbf{x}_{kj}$ catégorielle, on peut appliquer un modèle logit généralisé à un effet (avec des données groupées).
   Pour le bloc k connecté au bloc i, nous avons :

$$\mathbf{y}_k^{(t=0)} = \sum_{h=1}^{p_k} \sum_{l=1}^{L_j} w_{kjl(r)}^{(t=0)} \mathbf{x}_{kjl} \text{ quand } \mathbf{x}_{i1} \text{ prend la valeur } r, \tag{2.11}$$

   où $\mathbf{x}_{kjl}$ est une variable binaire pour la modalité l et $w_{kjl(r)}^{(t=0)}$ est le coefficient de la régression logistique associée.

Cette étape peut aussi s'appliquer à des variables de comptage avec un modèle log-linéaire.

Algorithme itératif

Les estimations des variables latentes obtenues à l'étape précédente sont toutes continues, alors que les variables manifestes ont toujours des échelles différentes. En effet, lors de chaque étape de quantification, l'estimation de la variable latente a la forme d'un score qui peut être considéré comme continu. Ce score est en fait la somme des produits entre les $\mathbf{x}_{kj}$ et les coefficients de régression.
Comme dans le cadre de l'approche PLS, l'algorithme se divise en deux estimations séparées :

(1) *Estimation interne :*

$$\mathbf{z}_k^{(t)} = \sum_{i:\xi_i \leftrightarrow \xi_k} e_{ki} \mathbf{y}_i^{(t-1)}$$

(2) *Estimation externe :*

1. $\mathbf{x}_{kj}$ continue :

$$w_{kj}^{(t)} = cor(\mathbf{x}_{kj}, \mathbf{z}_k^{(t)})$$

$$\mathbf{y}_k^{(t)} = \sum_{j=1}^{p_k} w_{kj} \mathbf{x}_{kj}$$

2. $\mathbf{x}_{kj}$ catégorielle :

$$\mathbf{y}_k^{(t)} = \sum_{j=1}^{p_k} \sum_{l=1}^{L_j} w_{kjl}^{(t)} \mathbf{x}_{kjl}$$

avec $\mathbf{x}_{kjl}$ est une variable binaire associée à la modalité l et $w_{kjl}^{(t)}$ est la moyenne de $\mathbf{z}_k$ sur la modalité l.

Répéter jusqu'à convergence.

L'estimation interne est effectuée de la même façon que dans l'approche PLS car les variables latentes sont considérées comme continues. Les poids internes e_{ki} doivent être choisis en fonction du schéma sélectionné (centroïde, factoriel ou structurel).
L'estimation interne dans le cas catégoriel revient à un modèle d'analyse de la variance car $\mathbf{x}_{kj}$ est catégorielle et $\mathbf{y}_k^{(t)}$ est continue.

Lorsqu'une variable manifeste est catégorielle, on obtient un poids associé à chacune des modalités de cette variable. Ces poids correspondent aux coefficients de l'analyse de la variance avec comme variable de réponse l'estimation de la variable latente. Ces poids individuels sont normalisés :

$$\tilde{w}_{khl} = \frac{(\hat{w}_{khl} - \overline{w}_{kh})^2}{\sum_{l=1}^{L_h} (\hat{w}_{khl} - \overline{w}_{kh})^2}$$

où $\hat{w}_{khl}$ est le poids externe individuel associé à la modalité l et $\overline{w}_{kh}$ est leur moyenne pour la variable manifeste $\mathbf{x}_{kh}$.

Finalement, le poids externe global de la variable manifeste $\mathbf{x}_{kh}$ qui a L_h modalités est :

$$\hat{w}_{kh} = \sqrt{\frac{\sum_{l=1}^{L_h} (\hat{w}_{khl} - \overline{w}_{kh})^2}{L_h}}$$

L'estimation des équations structurelles

Les coefficients structurels du modèle interne sont estimés par des régressions linéaires classiques entre les estimations finales des variables latentes.

Conclusion

Cette approche apporte un bon complément à l'approche PLS dans le cas de traitement de données ordinales ou binaires. Elle permet de prendre en compte des échelles de mesure de chacune des variables manifestes du modèle. Par contre, elle ne résout pas le problème de l'absence d'optimisation globale liée à l'approche PLS. Néanmoins, elle permet d'obtenir des informations supplémentaires comme des poids au niveau de chaque modalité des variables catégorielles. De nombreuses recherches supplémentaires sont nécessaires afin de compléter les connaissances sur cette approche. Les notions de convergence et d'optimisation pourront être étudiées en détail dans le cadre de recherches futures.

2.4.2 La méthode Generalized Structured Component Analysis (GSCA)

Introduction et principe

Cette approche développée par Hwang et Takane (2004), et ayant fait l'objet de recherches et adaptations ces dernières années (Hwang et al., 2007), tire son origine dans la recherche d'une méthode à la croisée des approches PLS et LISREL. Elle converge dans tous les cas (à l'inverse de LISREL) et a une fonction de perte globale à optimiser (à l'inverse de PLS). Elle est issue d'une représentation spécifique du modèle structurel appelé *Reticular Action Model* (McDonald et McArdle, 1984), des travaux sur l'*optimal scaling* du début des années 1980 et plus spécifiquement sur l'algorithme des moindres carrés alternés (Young, 1981).

On peut la rapprocher des recherches sur la méthode PATHALS (Coolen et De Leeuw (1987), Coolen et De Leeuw (1988), De Leeuw (1987)) développée pour traiter des modèles d'équations structurelles par les moindres carrés alternés. PATHALS a été créée afin de traiter tous types de données dans la *path analysis*, elle utilise les moindres carrés alternées afin de trouver une transformation des variables et ensuite optimise une fonction globale bien définie. Cette fonction ne prend pas en compte le modèle interne et ne permet donc pas de traiter des modèles d'équations structurelles à variables latentes. L'approche GSCA en étend les capacités et en minimise les défauts en utilisant une formulation spécifique.

Le modèle

Le modèle est basé sur une équation :

$$\mathbf{XV} = \mathbf{XWA} + \mathbf{E} \tag{2.12}$$

qui peut être écrite :

$$\mathbf{\Psi} = \mathbf{\Gamma A} + \mathbf{E} \tag{2.13}$$

avec

- P_{refl} nombre de variables manifestes qui suivent un schéma réflectif,
- $\mathbf{X}$ matrice $N \times P$ des variables observées,
- $\mathbf{V}$ matrice $P \times (K_{endo} + P_{refl})$ des poids (w_{ij}) des variables manifestes associées aux variables latentes endogènes,
- $\mathbf{W}$ matrice $P \times K$ des poids des variables manifestes (w_{ij}),
- $\mathbf{A}$ matrice $K \times (K_{endo} + P_{refl})$ composée de deux sous matrices, la première composée des *loadings* (c_{ij}) associés aux variables latentes ayant un schéma réflectif et la seconde composée des coefficients structurels,

– $\mathbf{E}$ matrice $N \times (K_{endo} + P_{refl})$ rassemblant l'ensemble des résidus,
– $\mathbf{\Psi} = \mathbf{XV}$ et $\mathbf{\Gamma} = \mathbf{XW}$.

Cette formulation permet d'inclure dans un processus d'optimisation global l'ensemble des paramètres du modèle (les *loadings*, les poids externes et les coefficients du modèle interne). Afin de mieux comprendre cette formulation peu intuitive nous en présentons un exemple simple et nous introduisons une formulation plus intuitive en annexe (cf. annexe A, p. 175).

Illustration

Nous illustrons cette notation complexe qui permet de prendre en compte des construits formatifs par un exemple simple à 3 variables latentes et deux variables manifestes par variable latente (cf. fig. 2.5).

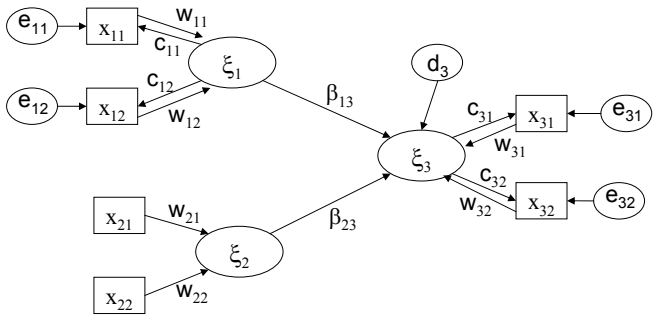

FIG. 2.5 – Modèle d'équations structurelles pour l'estimation par l'approche GSCA

On définit :

$$\mathbf{X} = \left[\begin{array}{cccccc} \mathbf{x}_{11} & \mathbf{x}_{12} & \mathbf{x}_{21} & \mathbf{x}_{22} & \mathbf{x}_{31} & \mathbf{x}_{32} \end{array} \right] \tag{2.14}$$

$$\mathbf{E} = \left[\begin{array}{ccccc} \mathbf{e}_{11} & \mathbf{e}_{12} & \mathbf{e}_{31} & \mathbf{e}_{32} & \mathbf{d}_3 \end{array} \right] \tag{2.15}$$

L'équation du modèle est alors :

$$\mathbf{X} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & w_{31} \\ 0 & 0 & 0 & 1 & w_{32} \end{bmatrix} = \mathbf{X} \begin{bmatrix} w_{11} & 0 & 0 \\ w_{12} & 0 & 0 \\ 0 & w_{21} & 0 \\ 0 & w_{22} & 0 \\ 0 & 0 & w_{31} \\ 0 & 0 & w_{32} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & 0 & 0 & \beta_{13} \\ 0 & 0 & 0 & 0 & \beta_{23} \\ 0 & 0 & c_{31} & c_{32} & 0 \end{bmatrix} + \mathbf{E} \tag{2.16}$$

L'estimation des paramètres

Les paramètres du modèle sont estimés de manière à minimiser la somme des carrés de l'ensemble des résidus du modèle. Ceci revient à minimiser :

$$f = SS(\mathbf{XV} - \mathbf{XWA}) = SS(\mathbf{\Psi} - \mathbf{\Gamma A}) \tag{2.17}$$

L'estimation de l'ensemble n'est pas possible directement. Les auteurs préconisent l'utilisation d'un algorithme basé sur les moindres carrés alternés (ALS, Young (1981)). Dans la première étape, $\mathbf{A}$ est estimé par moindres carrés avec $\mathbf{V}$ et $\mathbf{W}$ fixés. Dans la seconde étape, c'est $\mathbf{A}$ qui est fixé et $\mathbf{V}$ et $\mathbf{W}$

sont estimés par moindres carrés. Ces deux étapes sont répétées jusqu'à convergence.
Cet algorithme converge et on peut ainsi obtenir un indice d'ajustement :

$$FIT = 1 - \frac{SS(\mathbf{\Psi} - \mathbf{\Gamma A})}{SS(\mathbf{\Psi})} \tag{2.18}$$

Cet indice global permet de juger de la qualité d'ajustement du modèle aux données, il permet de juger à quel point les résidus sont minimisés après l'estimation des paramètres du modèle. Il est compris entre 0 et 1 et donne la proportion de variance expliquée par le modèle.
Hwang et al. (2007) introduisent l'indice FIT ajusté, noté AFIT, défini par :

$$AFIT = 1 - (1 - FIT)\frac{N}{N - v}$$

où N est la taille de l'échantillon et v le nombre de paramètres libres. Cet indice prend en compte le nombre de degrés de liberté associés au modèle estimé. Il permet de favoriser les modèles simples.

La formulation introduite dans ce chapitre, bien qu'adaptée, n'est pas intuitive. Nous introduisons en annexe A (p. 175) une formulation alternative plus proche de celles associées aux modèles structurels classiques. Celle-ci est adaptée dans le cas d'un modèle composé uniquement de variables réflectives.

Extensions et ouvertures

Hwang et Takane (2004) et Hwang et al. (2007) ont développé en parallèle un certain nombre d'adaptations de cette approche permettant de faire du *clustering* et de la comparaison inter-groupes (comme dans le cadre du chapitre 4 de cette thèse). Ils sont actuellement en train de mettre en place des simulations afin de comparer les approches PLS, LISREL et GSCA. Finalement, un programme informatique convivial est disponible (Hwang, 2007).

Par ailleurs, l'intégration de relations non linéaire dans le cadre de cette méthode pourra se faire facilement par l'utilisation d'une fonction d'optimisation globale et par l'utilisation de l'algorithme ALS. Hwang et Takane (2002) en présente le principe dans le cas d'une méthode légèrement différente : *extended redundancy analysis*.

Cette méthode récente est très prometteuse, les propriétés asymptotiques des estimations sont encore inconnues et font l'objet de travaux en cours de développement. Nous illustrons son fonctionnement dans le cadre de l'application à l'analyse de la satisfaction des clients dans le chapitre 7.

2.5 Conclusion

Les méthodes d'estimation de modèles d'équations structurelles à variables latentes sont de plus en plus nombreuses avec de multiples adaptations spécifiques. Les deux méthodes présentées nous paraissent intéressantes. La méthode PML doit être utilisée en complément de l'approche PLS. La méthode GSCA pourrait venir en remplacement de l'approche PLS dans certains cas, à condition que ses propriétés asymptotiques s'avèrent bonnes.

Ces deux premiers chapitres de présentation ont tenté de mettre en place une théorie complexe qui permet d'estimer les modèles d'équations structurelles à variables latentes. Les points présentés nous paraissent les plus pertinents dans le cadre de cette thèse consacrée à de nombreux sujets en rapport avec le domaine des modèles structurels à variables latentes.

Ces chapitres introductifs nous permettent donc de mieux aborder les points spécifiques que nous traitons par la suite dans le cadre des chapitres 4 à 6 de cette thèse.

Chapitre 3

La construction du modèle conceptuel

3.1 Introduction

Les modèles d'équations structurelles à variables latentes sont des approches basées sur une théorie posée *a priori* : le modèle de mesure et le modèle structurel. Lorsque les informations liées au modèle sont trop peu nombreuses ou peu fiables, Wold (1980b) conseille l'utilisation de l'approche PLS du fait du peu d'hypothèses nécessaires à son application. Cependant, il arrive que la mise en place du modèle pose un problème de manque d'information sur la nature des relations entre les variables manifestes et les variables latentes ou même sur l'existence de ces dernières. Dans ces cas, des méthodes de construction du modèle à partir des données peuvent être utilisées. Ces méthodes s'appliquent aussi bien dans le cadre de l'approche PLS que dans celui de la méthode LISREL mais la notion de modélisation douce (*soft modeling*) associée à la première aura tendance à la favoriser.

Le modèle conceptuel est représenté par un *path diagram* dans lequel les arcs représentent des relations de causalité. Cette notion est complexe et nécessite quelques éclaircissements. C'est Pearl (2000) qui a le plus clairement rassemblé les notions d'équations structurelles et de causalité. Ce dernier écrit qu'à l'origine, les approches par modélisation d'équations structurelles partaient du principe de causalité entre les variables sans aucune autre hypothèse. Au cours de l'évolution de la pratique de ces méthodes, les chercheurs ont ajouté des conditions pour que cette causalité soit avérée.
Pearl, dans son ouvrage, donne deux raisons principales :

- La notion de causalité a été mise de côté dans un souci de respectabilité. En effet, cette notion n'est pas vérifiable statistiquement, elle est donc "embarrassante".
- Les méthodes ont été traduites dans un langage algébrique dans lequel la causalité ne peut pas apparaître explicitement, c'est pour cela qu'elle a été progressivement "oubliée".

Pearl réussit donc à mettre au point un formalisme permettant d'exprimer avec les modèles graphiques des notions de causalité dans les modèles d'équations structurelles.

Nous n'avons pas la prétention d'introduire des notions de causalité dans les développements statistiques que nous introduisons. Nous n'aborderons donc pas ce thème complexe qui suscitent de nombreux débats. Des recherches ont été entreprises afin de rassembler modèles structurels et causalité. On peut voir, entre autres, Pearl (1998), Bollen (1989) ou Aris (2001, chap. 2) pour la méthode LISREL et les actes de la conférence PLS'07 (*Causalities explored by indirect observation*) dans le cas de PLS.

Dans ce chapitre, nous présentons trois aspects de la construction du modèle conceptuel. Nous étudions tout d'abord le choix de l'orientation des relations entre variables manifestes et latentes. Nous présentons un algorithme de sélection de variables permettant de connaître les variables suivant des relations réflectives au sein du modèle. Dans un second point, nous étudions la construction du modèle de mesure par l'association de variables manifestes dans des blocs unidimensionnels et cohérents. Nous

terminons par des méthodes pour la construction du modèle interne à partir d'un modèle de mesure déjà défini.

3.2 Du sens des relations dans le modèle de mesure

Deux schémas de relations entre les variables manifestes et les variables latentes existent, le schéma réflectif et le schéma formatif. Le premier est le plus utilisé mais il arrive souvent qu'il soit appliqué, alors qu'en réalité le modèle suit un schéma formatif. Le schéma réflectif est généralement préféré car il permet un traitement direct par l'approche LISREL. Bollen (1990a) a introduit le test des tétrades afin de vérifier le type de schéma associé au modèle étudié.

L'existence d'items (variables observées, mesures, variables manifestes, indicateurs), qui ne s'intègrent pas dans leur construit (variable latente) réflectif constitue l'un des problèmes dans la mise en place du modèle de mesure. Cet aspect peut être dû, par exemple, aux erreurs dans la construction de l'échelle (l'existence des items qui ne mesurent pas le construit) ou au fait que certains sont formatifs et dans ce cas on parle d'un modèle hybride MIMIC (*Multiple Indicators MultIple Causes*, cf. chapitre 1).

Après avoir présenté le test des tétrades, nous développons un algorithme itératif permettant d'obtenir un bloc réflectif à partir d'un bloc rejetant l'hypothèse de réflexivité. Il est basé sur la recherche du plus grand sous-groupe de variables réflectives. La détermination du type de modèle se fait par le test des tétrades.

3.2.1 Modèle réflectif contre modèle formatif

Dans la théorie classique des tests, les variables observées sont directement dépendantes d'un construit, de sorte que toute variation dans le construit est reflétée par une variation des valeurs de ses indicateurs. Selon la même théorie, la variation des valeurs des mesures liées à un construit est considérée comme une fonction de sa vraie valeur et d'un terme d'erreur. Selon Jarvis et al. (2003), ce sens supposé de la causalité (de la variable latente à ses mesures) est approprié conceptuellement dans de nombreux cas, mais pas dans tous.

Dans la littérature, on peut identifier deux modèles principaux de mesure : le modèle en facteurs principaux (ou réflectif) et le modèle en composantes principales constitué à partir de plusieurs indicateurs (ou formatif).

De manière formelle, on peut écrire :

– Pour le modèle réflectif (un modèle de régression simple de chaque manifeste sur sa latente) :

$$\mathbf{x}_j = \pi_j \boldsymbol{\xi} + \boldsymbol{\epsilon}_j \tag{3.1}$$

– Pour le modèle formatif (la variable latente est une fonction linéaire des variables manifestes associées) :

$$\boldsymbol{\xi} = \sum_j \omega_j \mathbf{x}_j + \boldsymbol{\zeta} \tag{3.2}$$

Selon Fornell et Bookstein (1982), le modèle formatif indique que les mesures sont des causes du concept qu'elles construisent. Au contraire, dans le modèle réflectif elles sont spécifiées comme le reflet du construit qui rend compte de leurs variances et covariances observées. MacCallum et Browne (1993) considèrent que : "lorsqu'un construit n'a que des indicateurs formatifs, ce construit n'est plus une variable latente dans son sens traditionnel". Dans ce cas, les auteurs préfèrent le terme de variable composite (ou composante).

Pour le modèle de mesure réflectif, les blocs de variables manifestes, outre leur cohérence théorique, doivent satisfaire un certain nombre de propriétés dont :

$$cor(\mathbf{x}_{kj}, \boldsymbol{\xi}_k) > cor(\mathbf{x}_{kj}, \boldsymbol{\xi}_i) \, \forall i \neq k$$

L'hypothèse principale associée à un bloc de variables réflectif est son unidimensionnalité (celle-ci peut être vérifiée par l'importance de la première composante principale par rapport à la seconde lors

de l'analyse en composantes principales sur chaque bloc). Nous ne nous attarderons pas ici sur les notions d'unidimensionnalité et de consistance interne des construits et les traiterons dans la partie sur la construction du modèle externe.

Par contre, le modèle formatif ne présuppose pas que les mesures soient toutes causées par un seul et même construit sous-jacent. Les mesures formatives influencent le construit latent, elles peuvent être corrélées, mais le modèle ne le présuppose pas ou ne le nécessite pas (Bollen et Lennox, 1991; Jarvis et al., 2003). En conséquence, pour un modèle de type formatif, l'évaluation de la consistance interne n'est pas appropriée pour juger de l'adéquation des mesures.

Dans les applications pratiques réalisées, le modèle réflectif est le plus utilisé (Baumgartner et Homburg, 1996), car plus proche de la théorie classique de la mesure. Cette prédominance du modèle réflectif peut avoir plusieurs explications. L'une d'elles est la difficulté d'identification et d'estimation du modèle formatif (Jarvis et al., 2003).

Une autre explication peut être trouvée dans le débat sur le "statut théorique de la variable latente" (Bollen, 2002). Dans la théorie classique de la mesure, il est nécessaire que la variable latente satisfasse le principe d'indépendance locale. Le variable latente est une cause commune (hypothèse d'indépendance locale), donc les indicateurs de mesure sont indépendants conditionnellement à la variable latente. Dans le cadre du modèle formatif, le construit ne peut pas être considéré comme une variable latente, dans le sens de cette définition.

En ce qui concerne le choix entre le modèle de mesure réflectif et formatif, plusieurs auteurs (Bollen et Lennox, 1991; Edwards et Bagozzi, 2000) ont formulé des recommandations. L'application de ces critères est basée sur la subjectivité du chercheur. Leur pertinence est incontestable, cependant ils sont insuffisants.

Même si le chercheur a de l'expérience dans le domaine, il fait seulement des hypothèses qui doivent être testées. C'est pour cette raison que nous considérons que l'algorithme que nous proposons peut être un outil intéressant dans la spécification du modèle de mesure. Comme celui-ci est basé sur le test des tétrades, nous introduisons d'abord ce test. Nous insistons sur le fait que la démarche que nous proposons doit être considérée comme une aide et qu'elle doit compléter l'avis des experts.

3.2.2 Le test des tétrades

En 1904, Spearman (1904) a mis au point les bases de ce qui devait devenir l'analyse factorielle. Dans ce travail et ultérieurement (Spearman, 1927), il a démontré qu'un seul facteur, étant à la base d'au moins quatre variables observées, implique que la différence dans les produits de certaines paires de covariances de ces variables doit être nulle. Ce qui sera nommé tétrades nulles ou évanescentes (*vanishing tetrads*).

Glymour et al. (1987), Bollen et Ting (1993, 2000) ont proposé des tests et des algorithmes afin de construire et valider le modèle externe en se basant sur la notion de tétrades. Les premiers se sont attachés à la construction du modèle de mesure alors que les seconds se sont intéressés à la validation du schéma réflectif ou formatif. Dans le cadre du test des tétrades introduit par Bollen (1990a), on parle d'analyse confirmatoire des tétrades (*confirmatory tetrad analysis*). Celle-ci permet de tester un ou plusieurs modèles spécifiques (le modèle est donc établi à l'avance).

Avant de définir ce test, il est important d'introduire quelques notions :

Définition 3.1. *Une* ***tétrade*** τ_{ijkl} *est une combinaison de covariances associées aux variables* $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_k$ *et* $\mathbf{x}_l$ *définie par :*

$$\tau_{ijkl} = \sigma_{ij}\sigma_{kl} - \sigma_{ik}\sigma_{jl} \tag{3.3}$$

avec $\sigma_{ij} = cov(\mathbf{x}_i, \mathbf{x}_j)$.

Ainsi, soit $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ 4 variables appartenant au même construit (pour la construction des tétrades il est nécessaire d'avoir au moins quatre variables). Une tétrade est donc définie comme la

différence entre deux paires de covariances. Comme quatre indicateurs produisent six covariances, on obtient les trois tétrades :

$$\begin{aligned}\tau_{1234} &= \sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24}\\ \tau_{1342} &= \sigma_{13}\sigma_{42} - \sigma_{14}\sigma_{23}\\ \tau_{1423} &= \sigma_{14}\sigma_{32} - \sigma_{12}\sigma_{34}\end{aligned}$$

La notation utilisée dans la définition précédente a été introduite par Kelley (1928).

Définition 3.2. *Une tétrade est dite* ***évanescente*** *ou* ***nulle*** *si* $\tau_{ijkl} = 0$.

Nous rapprochons maintenant les notions de tétrades de celles de modèles réflectifs ou formatifs.

Le cas du modèle réflectif

Propriété 3.1. *Les tétrades associées aux variables d'un construit réflectif sont nulles.*

Démonstration : Pour chaque variable manifeste du modèle, on a :

$$\mathbf{x}_j = \pi_j \boldsymbol{\xi} + \boldsymbol{\epsilon}_j$$

On a donc si on pose ϕ variance de $\boldsymbol{\xi}$:

$$\begin{aligned}\sigma_{ij} &= cov(\mathbf{x}_i, \mathbf{x}_j) = cov(\pi_i\boldsymbol{\xi} + \boldsymbol{\epsilon}_i, \pi_j\boldsymbol{\xi} + \boldsymbol{\epsilon}_j)\\ &= cov(\pi_i\boldsymbol{\xi}, \pi_j\boldsymbol{\xi}) + cov(\pi_i\boldsymbol{\xi}, \boldsymbol{\epsilon}_j) + cov(\boldsymbol{\epsilon}_i, \pi_j\boldsymbol{\xi}) + cov(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j)\\ &= \pi_i\pi_j cov(\boldsymbol{\xi}, \boldsymbol{\xi}) + \pi_i cov(\boldsymbol{\xi}, \boldsymbol{\epsilon}_j) + \pi_j cov(\boldsymbol{\epsilon}_i, \boldsymbol{\xi}) + cov(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j)\\ &= \pi_i\pi_j\phi\end{aligned}$$

car $cov(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j) = 0$, $cov(\boldsymbol{\xi}, \boldsymbol{\epsilon}_k) = 0$. On obtient pour un modèle réflectif :

$$\begin{aligned}\tau_{ijkl} &= \sigma_{ij}\sigma_{kl} - \sigma_{ik}\sigma_{jl}\\ &= \pi_i\pi_j\pi_k\pi_l\phi^2 - \pi_i\pi_k\pi_j\pi_l\phi^2 = 0\end{aligned}$$

En conséquence, les tétrades d'un modèle réflectif (quelles que soient les valeurs des paramètres du modèle) sont nulles.

□

Afin de vérifier la réflectivité du modèle, on va donc tester le respect des contraintes sur les tétrades reliées au modèle de mesure, c'est-à-dire leur nullité.

Dans la pratique, la matrice de covariance de la population entre les variables du modèle est inconnue ; elle peut être estimée en utilisant des échantillons finis. En conséquence, les tétrades théoriquement nulles sont en valeur absolue positives à cause de l'erreur d'échantillonnage. C'est pour cette raison qu'un test statistique adéquat doit être utilisé afin de vérifier l'hypothèse de nullité des tétrades.

Le cas du modèle formatif

Par contre, les tétrades associées aux variables d'un construit formatif ne doivent pas nécessairement être nulles.

Les seuls cas où les tétrades seront nulles apparaissent lorsque les variables sont indépendantes (ceci est peu plausible si on suppose la présence d'un construit) ou alors lorsque les produits de covariances s'annulent. Ces cas seront rares et on pourra supposer que dans le cas d'un modèle formatif, on a :

$$\tau_{ijkl} \neq 0,\ \forall i, j, k, l,\ i \neq j \neq k \neq l$$

Présentation du test statistique

Bollen (1990a) propose un test statistique permettant de tester la nullité de l'ensemble des tétrades générées par un construit simultanément. Ce test ne demande pas d'hypothèse de distribution sur les données et dans sa version classique nécessite au moins 4 variables manifestes.

Le nombre de tétrades générées par un construit à P variables est de $\frac{P!}{(P-4)!4!}$. Tester l'ensemble de ces tétrades serait trop exigeant. On va donc définir un type de tétrades spécifique :

Définition 3.3. *Une tétrade est dite indépendante si elle est non redondante dans le cas où l'hypothèse de nullité des tétrades est vérifiée.*

Une tétrade indépendante sera obtenue si, par des calculs algébriques, on ne peut pas à partir de plusieurs tétrades obtenir cette tétrade. Les tétrades redondantes sont exclues du test. Bollen et Ting (1993) ont mis au point une règle générale afin de repérer les tétrades indépendantes. A chaque fois que la même paire de covariances apparaît dans deux tétrades, alors elles sont redondantes (on doit alors en supprimer une).

Soit $\boldsymbol{\tau}$ le vecteur rassemblant l'ensemble des tétrades τ_{ijkl} non redondantes issues d'un construit. On veut tester :

$$H_0 : \boldsymbol{\tau} = 0 \tag{3.4}$$

Soit t_{ijkl} la réalisation d'une tétrade pour l'échantillon traité et $\mathbf{t}$ le vecteur rassemblant l'ensemble des réalisations des tétrades associées à $\boldsymbol{\tau}$.

La statistique du test a la forme suivante :

$$T = N\,\mathbf{t}'\boldsymbol{\Sigma}_{tt}^{-1}\mathbf{t} \tag{3.5}$$

avec N taille de l'échantillon et $\boldsymbol{\Sigma}_{tt}$ est la matrice de covariance de $\mathbf{t}$ lorsque N tend vers l'infini.

Propriété 3.2. *La statistique T suit une distribution du χ^2 à v degrés de liberté où v représente le nombre de tétrades non redondantes (Bollen, 1990a).*

Un résultat significatif renvoie à ce que les tétrades sont significativement différentes de 0 et que le modèle réflectif doit être rejeté.

L'obtention des valeurs de cette statistique se fait à partir des moments d'ordre 2 et 4 issus des données. Ainsi $\boldsymbol{\Sigma}_{tt}$ s'obtient en 3 étapes :

1. Construire le vecteur $\boldsymbol{\sigma}$ dans lequel sont rassemblées les covariances non redondantes de $\boldsymbol{\tau}$.
2. Construire une matrice $\boldsymbol{\Sigma}_{ss}$ issue de la distribution limite des covariances observées dans $\boldsymbol{\sigma}$. Les éléments de cette matrice sont :

   $$(\boldsymbol{\Sigma}_{ss})_{ijkl} = \sigma_{ijkl} - \sigma_{ij}\sigma_{kl}$$

   avec σ_{ijkl} moment d'ordre 4 associé aux variables i, j, k, l.
   Cette matrice peut être alors estimée grâce à l'estimateur

   $$S_{ijkl} = \frac{1}{N}\sum(x_i - \bar{x}_i)(x_j - \bar{x}_j)(x_k - \bar{x}_k)(x_l - \bar{x}_l)$$

3. La matrice $\boldsymbol{\Sigma}_{tt}$ est obtenue par :

   $$\boldsymbol{\Sigma}_{tt} = \frac{\partial \boldsymbol{\tau}'}{\partial \boldsymbol{\sigma}}\,\boldsymbol{\Sigma}_{ss}\frac{\partial \boldsymbol{\tau}}{\partial \boldsymbol{\sigma}}$$

Remarques :
- Par ailleurs, il est important de préciser que ce test comporte certaines limites d'application. Il sera sensible à un certain nombre de facteurs, comme l'erreur d'échantillonnage (celle-ci devra être "limitée" de façon à trouver des valeurs proches de zéro dans le cas réflectif (risque de classer un modèle réflectif en formatif) ou encore la taille de l'échantillon et le nombre d'items qui composent le construit (Bollen et Ting, 1998).
- Il faut faire attention à ce que les hypothèses préalables soient vérifiées (relation d'indépendances entre les erreurs de mesure, cohérence du construit)
- Certaines améliorations ont été développées :
  - Bollen et Ting (1998) ont mis au point une méthode afin de traiter moins de 4 indicateurs
  - La statistique du test tend asymptotiquement vers un χ^2, ceci peut poser des problèmes lorsque les échantillons sont petits. Bollen et Ting (1998) proposent d'utiliser un processus issu du bootstrap pour éviter ce problème.

Propriété 3.3. *Lorsque le modèle réflectif n'est pas rejeté, la consistance interne du construit est vérifiée.*

Démonstration : Lorsque le modèle réflectif n'est pas rejeté, les tétrades associées sont nulles. Sans perte de généralité, on peut remplacer les covariances des tétrades par des corrélations. On a alors

$$cor(\mathbf{x}_i, \mathbf{x}_j)cor(\mathbf{x}_k, \mathbf{x}_l) - cor(\mathbf{x}_i, \mathbf{x}_k)cor(\mathbf{x}_j, \mathbf{x}_l) = 0$$

Selon Hart et Spearman (1913), la consistance interne pour un construit avec au moins 4 variables est définie par :

$$\frac{cor(\mathbf{x}_i, \mathbf{x}_j)}{cor(\mathbf{x}_i, \mathbf{x}_k)} = \frac{cor(\mathbf{x}_l, \mathbf{x}_j)}{cor(\mathbf{x}_l, \mathbf{x}_k)}$$

On a donc :

$$\begin{aligned} cor(\mathbf{x}_i, \mathbf{x}_j)cor(\mathbf{x}_k, \mathbf{x}_l) - cor(\mathbf{x}_i, \mathbf{x}_k)cor(\mathbf{x}_j, \mathbf{x}_l) &= 0 \\ cor(\mathbf{x}_i, \mathbf{x}_j)cor(\mathbf{x}_k, \mathbf{x}_l) &= cor(\mathbf{x}_i, \mathbf{x}_k)cor(\mathbf{x}_j, \mathbf{x}_l) \\ \frac{cor(\mathbf{x}_i, \mathbf{x}_j)}{cor(\mathbf{x}_i, \mathbf{x}_k)} &= \frac{cor(\mathbf{x}_j, \mathbf{x}_l)}{cor(\mathbf{x}_k, \mathbf{x}_l)} \end{aligned}$$

Un construit réflectif a donc la propriété de consistance interne.

□

Pour conclure, selon Bollen et Ting (2000), "ce test, comme les autres tests dans le cadre des modèles d'équations structurelles, a les meilleurs performances quand beaucoup de recherches préalables sur la construction du modèle ont été menées". Il faut préciser que ce test n'est intégré dans aucun logiciel de modèles d'équations structurelles. Cependant, une macro SAS est à la disposition des chercheurs (http ://www.cuhk.edu.hk/soc/ting/, tab. B.13, p. 187).

3.2.3 Une heuristique pour la découverte de construits réflectifs

Comme nous l'avons vu précédemment, l'utilisation d'un modèle réflectif est largement dominante dans le cadre de l'application des modèles structurels. Cette utilisation mène souvent à une mauvaise spécification du modèle. La méthode que nous présentons part de ce constat. Dans le cadre de l'utilisation de l'approche LISREL, il sera plus judicieux de supprimer des variables plutôt que d'appliquer la méthode sur un construit non réflectif. De la même façon, pour l'approche PLS, l'utilisation du modèle MIMIC permettra d'éviter les erreurs de spécification.

Le principe

Nous présentons un algorithme itératif basé sur le test des tétrades qui permet d'obtenir, à partir d'un modèle rejetant l'hypothèse réflective, un sous-modèle réflectif.

Nous utilisons le test statistique présenté plus haut qui, malgré sa dépendance à la taille de l'échantillon et au nombre de variables dans le bloc, ne dépend pas de la distribution des données.

Plusieurs hypothèses importantes doivent être vérifiées dont :

$$cov(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j) = 0, \forall i, j, \, i \neq j$$

Le construit initial devra avoir une cohérence théorique et devra refléter un concept commun.

Nous partons d'un modèle comportant l'ensemble des variables à classer (défini *a priori* par les experts) et testons tous les modèles possibles jusqu'à ce que l'hypothèse de réflectivité ne soit plus rejetée. Cet algorithme doit permettre de sélectionner un modèle réflectif de taille maximale.

L'indice à maximiser

Le test des tétrades étant basé sur la statistique du χ^2, nous utilisons celle-ci dans le cadre du critère à maximiser afin de choisir quelle variable doit être écartée. Nous proposons un indice basé à la fois sur la valeur de la statistique (χ^2) et sur le nombre de degrés de liberté (dl). Pour le modèle complet, cet indice sera :

$$\Delta^{(0)} = \frac{\chi^{2(0)}}{dl^{(0)}}$$

A l'étape r, on retire la variable j et on teste le nouveau modèle, on aura alors :

$$\Delta_j^{(r)} = \frac{\chi_j^{2(r)}}{dl_j^{(r)}} - \frac{\chi^{2(r)}}{dl^{(r)}} \tag{3.6}$$

La maximisation de cet indice permet d'obtenir la variable qui, par sa suppression, entraîne la plus grande diminution du χ^2 (la variable sélectionnée est celle qui éloigne le plus le modèle de l'hypothèse de réflectivité, en terme de χ^2 associé au test des tétrades). Le choix d'un indice basé sur le χ^2 et les degrés de liberté est le plus adapté car le test des tétrades se base sur cette statistique afin de choisir le type de modèle à sélectionner. L'utilisation de cette statistique permet de prendre en compte simultanément la nullité de l'ensemble des tétrades du modèle. L'utilisation d'un critère basé sur les tétrades indépendantes n'est pas adaptée car ceux-ci sont différents à chaque application du test des tétrades. Les résultats de la statistique du χ^2 pour un nombre de degrés égal sont comparables et permettent donc de se rapprocher le plus rapidement possible d'un modèle ne rejetant pas le schéma réflectif.

L'algorithme

Soit P le nombre de variables dans le construit initial, on a $P \geq 4$. Soit S un ensemble contenant initialement l'ensemble des variables du construit, on note $\sharp S \leq P$ le nombre d'éléments de S. On

commence par $r = 0$ et $\sharp S = P$.

(1) Appliquer le test des tétrades sur les variables de S. Si l'hypothèse $H_0 : \boldsymbol{\tau} = 0$ est rejetée, aller en (2). Si elle n'est pas rejetée, aller en (6).

(2) Pour chacune des $\sharp S$ variables de S :
- Retirer la variable x_i du modèle
- Appliquer le test des tétrades
- Calculer $\Delta_i^{(r)}$

(3) Sélectionner la variable x_i telle que

$$i = \arg\max_k \Delta_k^{(r)}$$

(4) Retirer la variable x_i du modèle et de S :

$$S = S - \{x_i\}\,,\ \sharp S = \sharp S - 1\,,\ r = r + 1$$

(5) Aller en (1) tant que $\sharp S \geq 4$, sinon aller en (6).

(6) Vérification de l'unidimensionnalité du bloc (voir prochain paragraphe).

(7) Les variables dans S forment un construit réflectif.

Cet algorithme permet donc de mettre en valeur un modèle de mesure réflectif à partir de celui mis au point par les experts. Deux solutions sont alors possibles : les variables ne s'intégrant pas sont éliminées du modèle ou alors elles sont prises en compte en tant qu'indicateurs formatifs dans le cadre d'un modèle MIMIC.

Dans la septième étape, nous demandons la vérification de l'unidimensionnalité du bloc de façon à rassembler l'ensemble des hypothèses nécessaires à l'application d'une méthode d'estimation que ce soit la méthode LISREL ou l'approche PLS avec le mode A.

Cette méthode nous permettra donc, dans le cadre de construits complexes, d'éviter des erreurs dans la spécification de ceux-ci. Nous présentons maintenant des études de simulation afin de valider l'utilisation de cet algorithme.

Simulations

La complexité de cet algorithme nous pousse à étudier ces propriétés à l'aide de simulations. Nous utilisons l'algorithme sur des données simulées issues de deux processus différents. Dans chacun des cas, nous faisons varier la taille de l'échantillon de 200 à 1000 observations (200, 500 et 1000). Le nombre de variables manifestes initial est de 9. Nous supposons que toute les variables suivent une distribution normale (car le test des tétrades n'est pas sensible à la non normalité). Les variables latentes ainsi que les erreurs de mesure sont simulées à partir d'une distribution normale et les variables manifestes sont obtenues à partir des équations du modèle de mesure. Pour chaque processus, l'algorithme est répété 1000 fois. Nous illustrons les deux modèles utilisés pour les processus (P1) et (P2) dans la figure 3.1.

Le premier processus (P1) consiste en la simulation d'un modèle réflectif monofactoriel : en utilisant la loi normale, on simule une variable latente et 9 termes d'erreurs. Ensuite, les variables manifestes sont obtenues à partir des 9 équations du modèle de mesure réflectif. Il ressort que, dans tous les cas, le modèle obtenu est celui qui a été simulé, quel que soit l'échantillon utilisé (200, 500 ou 1000 observations). L'algorithme converge à la première étape avec 9 variables.

Le second processus (P2) consiste en la simulation d'un modèle bi-factoriel avec d'une part une variable latente associée à 5 variables et d'autre part, une autre associée aux 4 restantes. Les résultats

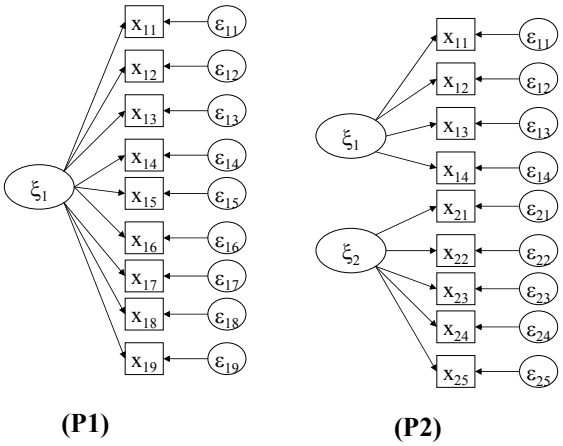

FIG. 3.1 – Modèles simulés pour la validation de l'algorithme de recherche d'un construit réflectif

sont rassemblés dans le tableau 3.1. Dans la troisième colonne apparaît le pourcentage de modèles exacts (c'est-à-dire le même modèle que celui qui est simulé), dans la quatrième le pourcentage de modèles qui diffèrent du modèle simulé à une variable près. Dans la dernière colonne, nous rassemblons le pourcentage de modèles largement différents du modèle simulé (plus d'une erreur).

Avec entre 64 et 68% de modèles exacts (c'est-à-dire que le modèle obtenu par l'algorithme représente l'un des deux blocs réflectifs simulé), on arrive pour une heuristique aussi simple à de bons résultats. On peut noter qu'on obtient entre 26 et 28% de modèles comportant une erreur d'allocation. Ceci s'explique par la structure de l'algorithme. En effet, plus on se rapproche du modèle réflectif, plus le critère Δ tend à être proche d'une variable à une autre. On remarque que c'est généralement lors de la dernière itération de l'algorithme que l'erreur d'allocation a lieu et on peut vérifier qu'à cette étape les différents Δ sont très proches et le choix d'une variable plutôt qu'une autre devient problématique. Ceci nous pousse donc à être vigilant sur cette dernière étape et à consulter l'avis des experts quand le choix se fait entre deux Δ très proches. Le pourcentage de modèles très différents du modèle simulé est de l'ordre de 7%, ce qui apparaît comme faible au vu de la complexité du modèle (9 variables). Il ressort que la taille de l'échantillon n'a pas un effet significatif sur le résultat de l'algorithme.

Processus	Nombre d'observations	Modèle exact	1 erreur d'allocation	Autre
P1	1000	100%	0%	0%
	500	100%	0%	0%
	200	100%	0%	0%
P2	1000	66%	27%	7%
	500	68%	26%	6%
	200	64%	28%	8%

TAB. 3.1 – Résultats de simulations sur l'orientation du modèle externe

Il ressort de cette courte étude par simulation que les modèles obtenus sont stables et que l'algorithme fournit de bons résultats. Nous ne détaillons pas les résultats liés à l'unidimensionnalité mais il s'avère qu'elle est toujours vérifiée dans le modèle réflectif obtenu. Il serait intéressant de faire varier par ailleurs le nombre initial de variables dans le modèle car le test des tétrades pose des problèmes lorsque ce nombre devient très grand afin de valider un modèle réflectif.

L'algorithme que nous avons présenté est donc une aide utile à l'expert afin de vérifier et d'ajuster

un modèle de mesure pour appliquer une méthode d'estimation avec des relations réflectives.
Nous illustrerons dans le chapitre 7.4 (p. 146), l'efficacité de cet algorithme dans le cas de données réelles pour l'analyse de la satisfaction et de la fidélité.

3.3 La construction du modèle de mesure

Dans la suite de ce chapitre, nous supposons que le modèle de mesure est réflectif. Ainsi, la construction du modèle de mesure revient à rassembler des variables manifestes dans des blocs (ou construits) qui représentent un concept latent et qui auront la propriété d'unidimensionnalité.

Les méthodes de construction du modèle de mesure sont en fait des méthodes visant à l'association de variables manifestes dans des blocs cohérents. Ceci nous amène fort logiquement à deux grandes classes de méthodes, tout d'abord la classification de variables avec des critères adaptés et, par ailleurs, les modèles graphiques qui permettent de visualiser les variables et donc de les classer.

Nous commençons par introduire la notion d'unidimensionnalité dans le cadre des modèles d'équations structurelles à variables latentes. Par la suite, nous présentons les méthodes de classification de variables généralement utilisées. Finalement, nous introduisons une approche originale basée sur des modèles graphiques probabilistes qui s'applique très bien dans le cadre de données sur des échelles de type Likert (1932).

3.3.1 L'unidimensionnalité dans le cadre des modèles d'équations structurelles

La notion d'unidimensionnalité est depuis longtemps sujette à de nombreux débats. Elle se différencie de celle de consistance interne généralement mesurée à l'aide de l'α de Cronbach. De nombreux chercheurs se sont penchés sur ces concepts et nous tentons d'en éclaircir les principes.

L'unidimensionnalité d'un groupe de variables peut être définie par l'existence d'un construit latent sous-jacent à l'ensemble des mesures associées à ces variables. Elle est généralement mesurée par le biais de l'analyse en composantes principales. Soit $\mathbf{X}$ la matrice dont les colonnes $\mathbf{x}_j$ représentent les P variables associées au bloc. Les premières composantes principales de l'analyse en composantes principales effectuée sur la matrice $\mathbf{X}$ sont notées $\mathbf{t}_1, \mathbf{t}_2, ...$ avec comme valeurs propres associées $\lambda_1, \lambda_2, ...$

Propriété 3.4. *Un bloc de variable est unidimensionnel si :*

$$\lambda_1 >> 1\,,\ \lambda_2 < 1$$

Par ailleurs, d'autres indices ont été proposés (Tenenhaus et al., 2005; Sahmer et al., 2005) :
– L'α de Cronbach (1951), défini par :

$$\alpha = \frac{\sum_{i=1}^{P}\sum_{j\neq i} cor(\mathbf{x}_i, \mathbf{x}_j)}{P + \sum_{i=1}^{P}\sum_{j\neq i} cor(\mathbf{x}_i, \mathbf{x}_j)} \times \frac{P}{P-1} \tag{3.7}$$

De nombreux articles ont montré que cet indice ne mesurait pas l'unidimensionnalité mais la consistance interne d'un bloc de variables (Green et al., 1977; Ten Berge et Socan, 2004). Il mesure la force des relations entre les variables du bloc et non leur unidimensionnalité. Il ne peut pas être directement lié à l'unidimensionnalité. Il constitue une condition nécessaire mais pas suffisante à l'unidimensionnalité d'un bloc (Anderson et Gerbing, 1982). La consistance interne est vérifiée lorsque $\alpha > 0.7$ (Tenenhaus et al., 2005). Cet indice devra donc être utilisé avec précaution, car un α élevé ne traduira pas forcément une seule dimension associée au construit.

– Le ρ de Dillon et Goldstein (1984), défini par :

$$\rho = \frac{(\sum_{i=1}^{P} cor(\mathbf{x}_i, \mathbf{t}_1))^2}{(\sum_{i=1}^{P} cor(\mathbf{x}_i, \mathbf{t}_1))^2 + \sum_{i=1}^{P}(1 - cor^2(\mathbf{x}_i, \mathbf{t}_1))} \tag{3.8}$$

Celui-ci mesure directement l'unidimensionnalité et peut être plus facilement relié à l'analyse en composantes principales. L'unidimensionnalité est vérifiée lorsque $\rho > 0.7$ (Tenenhaus et al., 2005).

– D'autres méthodes ont été présentées (Sahmer et al., 2005) notamment en se basant sur des tests, mais nous nous restreindrons à celles introduites jusqu'alors.

Il ressort que la vérification de l'unidimensionnalité se fait avant tout à partir d'une analyse en composantes principales. Nous verrons dans le cadre des méthodes de construction si cette validation est bien prise en compte.

3.3.2 La classification de variables

Comme pour la classification d'individus, il existe deux grandes familles de méthodes de classification de variables. D'une part, les méthodes hiérarchiques permettent d'obtenir un arbre de classification ou une succession de partitions emboîtées de l'ensemble des variables en groupes homogènes. Elles sont elles-mêmes divisées en deux groupes : les méthodes ascendantes basées sur un algorithme agglomératif type classification ascendante hiérarchique et les méthodes descendantes reposant sur un algorithme divisif. D'autre part, il existe des méthodes de partitionnement direct.

Classification basée sur un algorithme divisif

La principale méthode de ce type est celle développée dans la procédure VARCLUS de SAS Institute Inc. (2004b). Cette méthode permet de rechercher des classes unidimensionnelles, c'est-à-dire décrites par une seule composante principale. L'algorithme consiste à réaliser une analyse factorielle particulière sur l'ensemble des variables et à retenir les composantes principales correspondant aux deux plus grandes valeurs propres si la seconde est supérieure à 1. Chaque variable est alors affectée à la composante principale dont elle est la plus proche au sens du carré du coefficient de corrélation linéaire, formant ainsi deux groupes de variables. Ceux-ci sont, à leur tour, divisés selon la même méthode. La partition obtenue est telle que les variables d'une même classe sont les plus corrélées possible et deux variables de deux classes différentes sont les moins corrélées possible.

Classification basée sur un algorithme agglomératif

Les techniques de classification ascendante hiérarchique d'un ensemble de variables reposent sur le choix d'un indice de dissimilarité entre variables et d'une stratégie d'agrégation qui permet de construire un système de classes de variables de moins en moins fines par regroupements successifs. Cette méthode a été adaptée par Stan et Saporta (2005) pour les modèles structurels à variables latentes avec comme distance de dissimilarité $1 - |cor(\mathbf{x}_i, \mathbf{x}_j)|$. Il suffit ensuite d'appliquer les mêmes stratégies d'agrégation que pour la classification d'individus : critère de Ward, critère du saut minimal, du diamètre, ou de la moyenne. L'arbre est coupé de manière à maximiser l'unidimensionnalité des blocs. Les auteurs utilisent l'α de Cronbach et sa maximisation afin de choisir les blocs les plus adaptés.

Classification autour de composantes latentes

Cette approche, développée par Vigneau et Qannari (2003), offre un moyen d'organiser des données multivariées dans des structures significatives. La stratégie consiste à faire une classification hiérar-

chique puis à appliquer une méthode de partitionnement. On cherche à maximiser :

$$T = N \sum_{k=1}^{K} \sum_{j=1}^{p_k} \delta_{kj} cov^2(\mathbf{x}_j, \mathbf{c}_k) \text{ avec la contrainte } \mathbf{c}_k' \mathbf{c}_k = 1 \tag{3.9}$$

où K est le nombre de blocs, $\mathbf{c}_k$ la composante latente du bloc k. $\delta_{kj} = 1$ si $\mathbf{x}_j$ est dans le bloc k, 0 sinon. Cette approche est spécialement adaptée pour les modèles structurels à variables latentes. Elle permet d'obtenir des blocs dont les variables sont les plus corrélées possible avec une composante latente représentant la variable latente.

D'un point de vue formel, on cherche des composantes latentes $\mathbf{c}_k = \mathbf{d}_k' \mathbf{X}_k$, combinaisons linéaires des variables manifestes, $\mathbf{X}_k$ contient l'ensemble des variables associées à la composante $\mathbf{c}_k$. Soit $\mathbf{\Sigma}_k$ la matrice de covariance de $\mathbf{X}_k$, on veut donc maximiser :

$$\sum_{k=1}^{K} \sum_{j=1}^{p_k} \delta_{kj} cov^2(\mathbf{x}_j, \mathbf{c}_k) = \sum_{k=1}^{K} \mathbf{d}_k' \mathbf{\Sigma}_k^2 \mathbf{d}_k = \sum_{k=1}^{K} \lambda_k$$

avec la contrainte $\mathbf{d}_k' \mathbf{\Sigma}_k \mathbf{d}_k = 1$, $\forall k = 1, ..., K$. $\mathbf{d}_k$ est le vecteur propre de $\mathbf{\Sigma}_k$ associé à la plus grande valeur propre λ_k.

Le développement de cette méthode a fait l'objet d'une thèse (Sahmer, 2006) dans laquelle des cas particuliers sont développés (données manquantes entre autres).
Le nombre de blocs K est défini à l'avance dans la méthode présentée initialement, Sahmer (2006) a mis au point une méthode basée sur des permutations permettant d'estimer K.

Nous comparerons les résultats de ces différentes méthodes dans le cadre des applications à l'analyse de la satisfaction dans le chapitre 7.4 (p. 146).

3.3.3 Les modèles graphiques et les réseaux bayésiens

Les modèles graphiques constituent des méthodes largement utilisées dans la recherche actuelle, leur relation avec les modèles d'équations structurelles à variables latentes est assez intuitive. En effet, la représentation graphique des modèles structurels les rapproche des modèles graphiques et c'est Pearl (2000) qui le premier les associa.
Par la suite, Spirtes et al. (2000) les ont associés dans le cadre d'un logiciel nommé TETRAD qui permet à la fois de construire des modèles par l'intermédiaire de l'apprentissage de modèles graphiques mais aussi d'estimer les relations de ce modèle par l'estimation de la structure de covariance (Scheines et al., 1998).

Nous commençons par présenter les réseaux bayésiens et le principe d'apprentissage de la structure d'un réseau. Nous introduisons les principes des méthodes développées par Spirtes et al. (2000); Silva (2005) et par la suite nous proposons un algorithme basé sur les réseaux bayésiens afin de construire le modèle de mesure lorsque les données sont sur des échelles de Likert (1932).

Les réseaux bayésiens

Les réseaux bayésiens sont des outils puissants permettant de modéliser des relations entre variables en utilisant des algorithmes d'apprentissage basés sur des probabilités conditionnelles.

Les réseaux bayésiens ont été mis au point par Pearl (1988). On peut voir pour une introduction détaillée le livre de Naïm et al. (2004).

Définition 3.4. *Un **réseau bayésien** est un graphe orienté sans circuit (DAG) tel que :*
- *1 nœud ↔ 1 v.a. discrète ou continue ;*
- *1 nœud sans parent ↔ 1 loi de probabilité ;*
- *ensemble des arcs pointant sur 1 nœud ↔ 1 table de probabilités conditionnelles ;*
- *le réseau bayésien est compatible avec la loi conjointe dès lors que la probabilité qu'un nœud soit dans un état donné, conditionnellement à l'état de ses parents, ne dépend pas des états des nœuds qui ne sont pas ses descendants.*

Dans un langage basé sur la théorie des graphes, on pourra formuler la définition de la manière suivante :

Définition 3.5. *Un **réseau bayésien** est défini par :*
- *un graphe sans circuit orienté $G = (V, E)$, où V est l'ensemble des nœuds de G et E l'ensemble des arcs de G ;*
- *un espace probabilisé (Ω, Z, p) ;*
- *un ensemble de variables aléatoires correspondant aux nœuds du graphe défini sur (Ω, Z, p), tel que :*

$$p(V_1, ..., V_n) = \prod_{i=1}^{n} p(V_i | pV_i) \tag{3.10}$$

où pV_i est l'ensemble des parents de V_i dans le graphe G.

Un réseau bayésien est donc un graphe causal auquel on a associé une représentation probabiliste sous-jacente.
L'utilisation essentielle des réseaux bayésiens réside dans le calcul des probabilités conditionnelles d'événements reliés les uns aux autres par des relations de cause à effet. Cette utilisation s'appelle l'*inférence*.
Le théorème fondamental des réseaux Bayésiens relie les approches graphiques et probabilistes. Nous devons d'abord introduire deux définitions.

Définition 3.6. *Soient (x, y) deux nœuds du graphe G, et soit Z un ensemble de nœuds de ce graphe. Soit S un chemin entre les nœuds x et y. On dit que S est **d-séparé** par Z si :*
- *soit le chemin converge en un nœud w (les arcs sont orientés vers w) tel que $w \notin Z$ et $\forall z \in Z$, $w \notin C(z)$ ($C(Z)$ représente les descendants des nœuds de Z) ;*
- *soit le chemin passe par un nœud $z \in Z$ et il est soit divergent (les arcs ont pour racine z), soit en série en ce nœud (les arcs sont orientés de x vers y ou inversement en passant par z).*

*Soit ϵ une épreuve et (Ω, Z, P) l'espace probabilisé associé. Soient X, Y et Z trois vecteurs de variables aléatoires discrètes associées à ϵ. On dit que X et Y sont **indépendants conditionnellement** à Z, et on note $X \perp Y|Z$ si $P(X|Z,Y) = P(X|Z)$ ou si $P(X,Y|Z) = P(X|Z)P(Y|Z)$.*

La d-séparation revient à dire que si les valeurs des éléments de Z sont connues alors aucune information ne peut circuler entre X et Y. Ou encore Z d-sépare X et Y si et seulement si X et Y sont indépendants conditionnellement à Z.

Théorème 3.1. *Soit $B = (G, P)$ un réseau Bayésien. Soient $X, Y, Z \subset V$ trois sous-ensembles de nœuds. Si X et Y sont d-séparés dans G par Z, alors X et Y sont indépendants conditionnellement à Z, i.e. $X \perp Y|Z$.*

On peut trouver la preuve de ce théorème dans Naïm et al. (2004).

Ce qui nous intéresse dans le cadre de la construction du modèle conceptuel réside dans l'apprentissage de la structure d'un réseau bayésien.

Beaucoup de travaux ont été effectués sur l'apprentissage de la structure des réseaux Bayésiens. Deux types de méthodes existent : tout d'abord la recherche de relations causales (Spirtes et al., 2000) ou recherche sous contraintes et, d'autre part, la recherche de structure par des algorithmes heuristiques basés sur des scores.

Recherche sous contraintes : Cette approche de l'apprentissage de la structure est issue des théories de la causalité développées simultanément par Pearl et Verma et par Spirtes, Scheines et Glymour (Pearl, 1988; Pearl, 2000; Spirtes et al., 2000).
Ces méthodes sont basées sur des tests d'indépendance entre les variables, elles se divisent en trois étapes :

- construction d'un graphe non orienté à partir des tests d'indépendance conditionnelle,
- détection des V-structures, c'est-à-dire des parties du graphe dont l'orientation peut être obtenue par la loi de probabilité sous-jacente (qui sont orientées dans la classe d'équivalence associée à la probabilité jointe),
- propagation des orientations des arcs déjà orientés.

On obtiendra un réseau Bayésien partiellement orienté. Ces méthodes demandent une grande quantité de calcul, notamment lors de la première étape, mais possèdent l'avantage de pouvoir intégrer des variables latentes (dans ce cas on perd la dépendance causale).

L'algorithme le plus connu est l'algorithme PC (Spirtes et al., 2000), introduit en 1993, qui utilise un test statistique pour évaluer s'il y a indépendance conditionnelle entre deux variables. Ce test est un test d'indépendance conditionnelle basé sur l'information mutuelle

$$CE(X,Y|Z) = \sum_z P(z) \sum_{x,y} P(x,y|z) \log(\frac{P(x,y|z)}{P(x|z)P(y|z)})$$

et sur la statistique $G^2 = 2N\,CE(X,Y|A)$ qui suit une loi de χ^2 à $(r_X - 1)(r_Y - 1)\prod_{Z\in A} r_Z$ degrés de liberté où $r_.$ est le nombre de modalités de chaque variable.
Il est alors possible de reconstruire la structure du réseau Bayésien à partir de l'ensemble des relations d'indépendances conditionnelles découvertes. En pratique, un graphe complètement connecté sert de point de départ, et lorsqu'une indépendance conditionnelle est détectée, l'arc correspondant est retiré. D'autres algorithmes existent mais nous ne les développerons pas.

Recherche basée sur un score : Dans le cas de ces méthodes, deux types de recherches existent :

- les travaux sur les scores associés à une structure de réseau et à un jeu de données, ces scores nous donnent une valeur de la qualité d'ajustement de la structure aux données ;
- les travaux sur les algorithmes afin d'obtenir des structures avec de bons scores.
- *Les scores :* Il existe plusieurs types de scores suivant l'interprétation donnée au réseau par ceux-ci.
  Les scores les plus usités considèrent que la structure d'un réseau Bayésien est un ensemble de contraintes d'indépendances sur une distribution de probabilités associée aux données. Le score le plus connu dont nous nous servirons est le score MDL (*minimum description lenght*, Bouckaert (1993)). Soit D les données, B représente le réseau étudié, on aura :

$$S_{MDL}(B,D) = \log L(D|\theta^{MV}, B) - |A_B| \log N + c\dim(B) \tag{3.11}$$

  où θ^{MV} est l'ensemble des paramètres du réseau estimé par maximum de vraisemblance, $|A_B|$ est le nombre d'arcs du réseau, N le nombre d'observations et c est le nombre de bits utilisés pour stocker chaque paramètre numérique et $\dim(B)$ le nombre de paramètres pour décrire le réseau. Ce score revient dans le même temps à coller le mieux possible aux données (par la vraisemblance des données connaissant le réseau, $L(D|\theta^{MV}, B)$) et à privilégier la simplicité du réseau (par le terme $dim(B)$). D'autres scores peuvent être utilisés comme l'AIC, le BIC, l'entropie ou les scores BD (Bayesian Dirichlet).

– *Les algorithmes de maximisation des scores :* Les algorithmes utilisés pour obtenir des structures avec des scores les meilleurs possibles sont nombreux (François et Leray, 2003). Robinson (1977) a montré que le nombre de structures différentes pour un réseau Bayésien à n nœuds est de :

$$NS(n) = \begin{cases} 1 & , \quad n = 0 \text{ ou } 1 \\ \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-1)} NS(n-i) & , \quad n > 1 \end{cases}$$

La plupart des algorithmes d'apprentissage sont donc des heuristiques qui travaillent dans l'ensemble des graphes orientés sans circuit (DAG). Nous n'utiliserons que des heuristiques classiques de recherche issues de la théorie des graphes tel que la recherche gloutonne, l'arbre de poids maximal (Heckerman et al., 1994), la recherche Taboo, les colonies de fourmis ou encore les algorithmes génétiques.

Cette introduction nous permet de travailler sur des réseaux bayésiens afin de construire le modèle de mesure.

Les recherches de causalité avec TETRAD

Spirtes et al. (2000) ont développé une approche sur laquelle il est intéressant de se pencher. En effet, elle est basée sur des notions de causalité et intègre la notion de variables latentes.
Les théories de Spirtes et al. (2000) et Pearl (2000) sur la recherche de causalité ont abouti à de nombreux algorithmes permettant de construire le modèle à variables latentes. Dans leurs approches, les chercheurs ont mis au point des algorithmes de recherche orientés vers la recherche d'un modèle à variable latente. Ils utilisent l'apprentissage sous contraintes présenté plus haut et la notion de d-séparation afin d'obtenir un modèle de mesure "pur", i.e. une variable manifeste associée à un concept sera totalement indépendante des autres concepts du modèle. Les auteurs adaptent des algorithmes bien connus en réseaux bayésiens comme PC pour le cas où on aurait des variables latentes. Cette étape a un désavantage majeur, on perd la notion de suffisance causale.

Ces méthodes apportent beaucoup de flexibilité. Ainsi lorsqu'on recherche le modèle externe, il reste toujours des variables manifestes non assignées qui sont difficiles à classifier. C'est-à-dire des variables manifestes qui ne sont pas intégrées dans un bloc et reste exclues des blocs. Comme on ne peut pas créer de variables latentes avec un seul indicateur, elles sont exclues de la construction du modèle.
On peut retrouver l'ensemble de ces algorithmes dans Spirtes et al. (2000) ou dans Silva (2005).

3.3.4 Un algorithme basé sur les modèles graphiques probabilistes

Nous présentons un algorithme permettant, à partir de variables sur des échelles de Likert (1932), d'obtenir des blocs dont les variables ont des relations de probabilité inférentielle fortes (Jakobowicz et Derquenne, 2007). De plus, la puissance et la lisibilité des réseaux bayésiens apportent une aide supplémentaire à la construction du modèle.

Le principe

Dans le cadre de la construction du modèle de mesure, lorsque les variables manifestes sont catégorielles ordonnées, les approches précédentes par classification de variables peuvent être appliquées mais ne sont pas adaptées.
Nous proposons donc une approche basée sur les réseaux bayésiens afin d'associer des variables ayant des relations fortes. Nous partons d'un postulat fréquemment évoqué dans les recherches sur les modèles graphiques.

Polstulat 3.1. *Si des relations de probabilités fortes existent entre un ensemble $G = (\mathbf{x}_1, ..., \mathbf{x}_P)$ de variables observées, alors on peut supposer qu'il existe une variable non observée $\boldsymbol{\xi}$ tel que $\boldsymbol{\xi}$ d-sépare chaque paire de variable manifeste, ainsi :*

$$\mathbf{x}_i \perp \mathbf{x}_j | \boldsymbol{\xi}, \quad \forall i, j = 1, ..., P, \, i \neq j$$

En partant de cette hypothèse, nous construisons le modèle de mesure par le biais d'un algorithme associant deux à deux les variables observées ayant des relations fortes (on suppose alors qu'il existe ξ qui d-sépare ces deux variables).

On utilise les variables manifestes du modèle afin d'effectuer un apprentissage de la structure du réseau bayésien par le biais d'une heuristique basée sur un score et sur une méthode de recherche. Nous utilisons le score MDL (*Minimum Description Length*) qui maximise une vraisemblance du modèle par rapport aux données tout en favorisant la simplicité de ce modèle. Les relations qui augmentent le score sont ajoutées au modèle en utilisant un algorithme de recherche du type Taboo (Jouffe et Munteanu, 2001), celui-ci permet d'éviter de tomber dans un optimum local comme c'est souvent le cas avec la recherche gloutonne.

Le réseau obtenu consiste en un réseau le plus proche possible des données en terme de vraisemblance tout en privilégiant la simplicité (sinon l'ensemble des relations serait sélectionné). Il permet donc de visualiser des relations de probabilité entre des variables connectées. Nous introduisons une mesure de dissimilarité entre deux variables reliées applicable aux réseaux bayésiens :

Définition 3.7. *La divergence de Kullback et Leibler (1951) est définie pour deux variables* $\mathbf{x}_{jh}$ *et* $\mathbf{x}_{lm}$ *par :*

$$D_{KL}(p(\mathbf{x}_{jh}|\mathbf{x}_{lm})||p(\mathbf{x}_{jh})) = \sum_{r=1}^{L_m} p(\mathbf{x}_{jh}|\mathbf{x}_{lm}=r) \log \frac{p(\mathbf{x}_{jh}|\mathbf{x}_{lm}=r)}{p(\mathbf{x}_{jh})}, \tag{3.12}$$

avec L_m *nombre de modalités de* $\mathbf{x}_{lm}$ *et* D_{KL} *estime la force de la relation entre* $\mathbf{x}_{jh}$ *et* $\mathbf{x}_{lm}$*.*

Cette divergence n'est pas symétrique. Elle représente la quantité d'information, ou le gain d'information, sur $\mathbf{x}_{jh}$ qui est apportée en découvrant la valeur de $\mathbf{x}_{lm}$. Pour des raisons de simplicité de notation, la divergence définie dans l'équation 3.12 sera notée $D_{KL}(\mathbf{x}_{jh}, \mathbf{x}_{lm})$.

En utilisant l'apprentissage de la structure et la divergence de Kullback et Leibler (1951), nous développons un algorithme qui permet d'obtenir des groupes associés à un concept latent.

L'algorithme

L'algorithme part des données brutes rassemblées dans un tableau $\mathbf{X}$ à P colonnes et N lignes. On cherche à former des groupes de variables G_i. Initialement, aucun groupe n'est créé.

(1) Apprentissage de la structure du réseau bayésien à partir de $\mathbf{X}$

(2) Soit $S = \{(\mathbf{x}_i, \mathbf{x}_j) | \mathbf{x}_i \rightarrow \mathbf{x}_j$ dans le réseau bayésien, $\forall i, j = 1, ..., P\}$.
Soit $G = \{$ens. des groupes de variables créés$\}$, $G = \emptyset$.

(3) $\forall (\mathbf{x}_i, \mathbf{x}_j) \in S$, calculer $D_{KL}(\mathbf{x}_i, \mathbf{x}_j)$

(4) Sélectionner $(\mathbf{x}_i, \mathbf{x}_j)$ tel que :

$$(\mathbf{x}_i, \mathbf{x}_j) = \arg \max_{(\mathbf{x}_r, \mathbf{x}_s) \in S} D_{KL}(\mathbf{x}_r, \mathbf{x}_s)$$

- Si $\mathbf{x}_i, \mathbf{x}_j \notin G_k$, $\forall k \Rightarrow G_l = \{\mathbf{x}_i, \mathbf{x}_j\}$ et $G = G \cup G_l$
- Si $\mathbf{x}_i \in G_m$ et $\mathbf{x}_j \notin G_k$, $\forall k \Rightarrow G_k = G_k \cup \{\mathbf{x}_j\}$. Mise à jour des éléments de G.
- Si $\mathbf{x}_i \in G_m$ et $\mathbf{x}_j \in G_n \Rightarrow G_m = G_m \cup G_n$. Mise à jour des éléments de G.

(5) $S = S - (\mathbf{x}_i, \mathbf{x}_j)$

(6) Répéter (4) et (5) jusqu'à atteindre le critère d'arrêt.

On obtient donc un ensemble $G = \{G_1, ..., G_K\}$ composé de groupes de variables qui forment le modèle externe.

Le critère d'arrêt

Le critère d'arrêt de cet algorithme revêt une place très importante. Celui-ci définit la forme du modèle de mesure final. Lorsque $S = \emptyset$, G est alors composé d'un seul groupe. On ne peut donc pas se contenter d'attendre la fin de l'algorithme.
Plusieurs approches sont possibles :

- Utilisation d'un seuil sur D_{KL}. Si on fixe un seuil s tel que $\max D_{KL} \leq s$, alors l'algorithme s'arrête directement. Malheureusement ce seuil est difficile à fixer (il ne représente pas une grandeur interprétable). Les simulations effectuées montrent qu'en général un seuil $s \simeq 0.3$ est acceptable. Il est plus judicieux de le choisir à partir des valeurs des D_{KL} pour le jeu de données étudié. Par exemple, si on recherche peu de groupes, on pourra prendre comme seuil :

$$s = \frac{1}{\sharp S} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in S} D_{KL}(\mathbf{x}_i, \mathbf{x}_j)$$

  Ce seuil entraînera que certaines variables ne seront pas intégrées au modèle de mesure. Par ailleurs, d'autres méthodes de validation pourraient être utilisées.

- Utilisation d'un nombre fixe de groupes pour le modèle final (K). Dès que ce nombre est atteint, l'algorithme s'arrête.

- Utilisation d'une représentation du type arbre de classification. Les feuilles représentent les variables et la racine représente le cas $S = \emptyset$ (un seul groupe). Le figure 3.2 illustre cet arbre pour 6 variables. La découpe de l'arbre peut alors se faire en utilisant soit des dires d'experts, soit un indice associé aux propriétés recherchées dans le modèle de mesure (par exemple, l'unidimensionnalité).

Ces méthodes offrent des alternatives intéressantes, la dernière de par la visualisation des associations semble plus pratique d'usage et plus simple à interpréter.

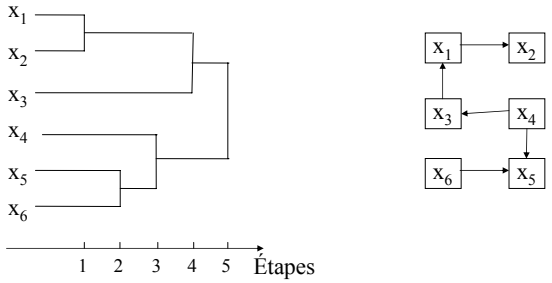

FIG. 3.2 – Arbre de classification pour la construction du modèle externe associé à un réseau bayésien initial

Les contraintes

L'un des avantages de cette approche réside dans la possibilité d'imposer un grand nombre de contraintes de manière aisée grâce à la flexibilité des réseaux bayésiens. Ainsi nous détaillons certaines d'entre elles rapidement :

- Si certaines associations de variables sont indiscutables, il suffit initialement de modifier l'ensemble G, on posera $G \neq \emptyset$ et les variables à associer seront dans des groupes prédéfinis auxquels on ajoutera les variables rentrant dans les groupes au cours de l'algorithme.

– Si certaines associations entraînent des non sens dans les blocs, il suffit d'interdire la mise en relation de ces variables dans l'algorithme par des contraintes simples.
– En ce qui concerne la taille des blocs et le nombre de blocs, il est aussi aisé de le prédéfinir et, de plus, ceci facilitera l'obtention d'un critère d'arrêt.

Conclusion

Cette méthode possède donc l'avantage de prendre en compte les échelles des variables issues généralement de questionnaires. Elle permet de construire un modèle de mesure robuste (cf. chapitre 7.4) sur lequel un modèle structurel peut être adapté. Le fait de pouvoir y intégrer des contraintes constitue un gros avantage supplémentaire dans des cas pratiques. Les nombreuses visualisations possibles associées, d'une part, aux réseaux bayésiens eux-mêmes et, d'autre part, à l'algorithme et à l'arbre de classification apportent une convivialité d'utilisation et une clarté de mise en place peu visible dans les autres méthodes.
Nous présentons dans le cadre de l'analyse de la satisfaction, des simulations et des applications de cet algorithme dans le chapitre 7.4 (p. 146).

3.4 La construction du modèle structurel

Une fois le modèle de mesure estimé, les relations structurelles entre les variables latentes peuvent être recherchées.
On considère que les blocs de variables manifestes existent et donc que les variables latentes sont déjà définies. On va chercher à construire les relations entre ces variables latentes. Les méthodes que nous présentons permettent, d'une part, de construire le modèle structurel et, d'autre part, d'obtenir des relations orientées entre les arcs.
Nous présentons des algorithmes mis en pace dans le cadre de l'approche PLS. Nous ne présentons pas ici les algorithmes développés dans le même cadre que ceux associés au logiciel TETRAD pour la construction du modèle de mesure, on peut voir Silva (2005); Spirtes et al. (2000).

3.4.1 Les heuristiques d'apprentissage du modèle interne

Trois méthodes construites à partir d'algorithmes itératifs et développées en relation avec l'approche PLS existent. Chacune vise à la maximisation d'un critère et fonctionne de manière itérative.

L'approche de Hui

Hui (1982) fut le premier à s'intéresser à la construction du modèle interne dans le cadre de l'approche PLS. Il présente une méthode simple et intuitive afin de créer le modèle structurel. Afin d'éviter les problèmes liés à la causalité, Hui (1982) demande que les variables latentes endogènes soient connues à l'avance. Ceci permet de connaître le nombre d'équations du type $\boldsymbol{\xi}_k = \sum_i \beta_{ki}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_k$ dans le modèle.

L'algorithme a la forme suivante :

(1) Effectuer une ACP sur chaque bloc et obtenir les premières composantes principales associées à chaque bloc $\mathbf{t}_1^1, ..., \mathbf{t}_1^K$.

(2) Calculer la matrice de corrélation : $\mathbf{R}_{ACP} = (cor(\mathbf{t}_1^i, \mathbf{t}_1^j))_{ij}$

(3) Pour chaque variable endogène, la variable avec laquelle elle a la plus forte corrélation entre dans l'équation du modèle.

(4) Estimation du modèle par l'approche PLS.

(5) Vérification des t-valeurs associées aux équations, on conserve la variable dans l'équation si $|t| > 1.65$.

(6) Calculer à partir des scores PLS obtenus en (4) : $\mathbf{R}_{PLS} = (cor(\mathbf{y}_i, \mathbf{y}_j))_{ij}$

(7) Aller en (3) jusqu'à convergence (obtention de la stabilité des équations structurelles).

Cette méthode permet d'obtenir un modèle structurel orienté dont les relations sont significatives en terme de régressions ordinaires. Néanmoins, elle pose un problème central, les modèles obtenus peuvent contenir des cycles et on peut avoir des relations à double sens. L'algorithme suivant permet de contourner ces deux défauts.

La méthode de Schenk et Hackl

La méthode présentée dans Hackl (2003) est inspirée de celle de Hui (1982). Elle permet d'éviter les cycles. Pour cela, elle force la matrice des coefficients structurels **B** à être triangulaire inférieure. Cette méthode demande un ordre préalable entre les variables latentes (quelle variable précède quelle autre dans une relation de cause à effet).

L'algorithme a la forme suivante :

(1) Effectuer une ACP sur chaque bloc afin d'obtenir $\mathbf{t}_1^1, ..., \mathbf{t}_1^K$.

(2) Calculer à partir des premières composantes principales : $\mathbf{R}_{ACP} = (cor(\mathbf{t}_1^i, \mathbf{t}_1^j))_{ij}$

(3) Les deux variables ayant la plus forte corrélation sont reliées et forment le modèle initial.

(4) Estimation du modèle par l'approche PLS.

(5) Vérification des t-valeurs associées à la relation ajoutée, on conserve la variable dans l'équation si $|t| > 1.65$.

(6) Calculer à partir des scores PLS obtenus en (4) : $\mathbf{R}_{PLS} = (cor(\mathbf{y}_i, \mathbf{y}_j))_{ij}$

(7) La variable ayant la plus forte corrélation à l'une des variables du modèle est intégrée ou si cette corrélation se trouve entre deux variables du modèle, la relation est ajoutée au modèle.

(8) Aller en (4) jusqu'à convergence (obtention de la stabilité du modèle).

Cet algorithme par l'ajout de l'étape (7) permet d'obtenir une matrice **B** triangulaire inférieure et ainsi d'avoir un modèle sans cycle (Hackl, 2003).

3.4.2 La méthode d'Amato

Amato (2003) propose une méthode qui prend en compte le fait que PLS recherche un compromis entre les estimations internes et externes des variables latentes. Il propose un algorithme qui n'est pas basé sur les composantes principales et sur de simples indices de corrélations mais sur des indices associés à l'approche PLS. Il utilise donc soit la redondance (F^2), soit la communauté (H^2)(Tenenhaus

et al., 2005), soit un indice hybride ayant la formule :

$$B^2 = \sum_k \frac{B_k^2}{K}\,; \quad B_k^2 = \begin{cases} H_k^2 & \text{si } \boldsymbol{\xi}_k \text{ est exogène,} \\ F_k^2 & \text{sinon.} \end{cases} \tag{3.13}$$

A chaque itération de l'algorithme, on maximisera l'un de ces critères afin d'ajouter un arc au modèle jusqu'à obtenir le modèle saturé. Le meilleur modèle sera choisi en effectuant des tests statistiques sur des échantillons bootstrap ou en utilisant un indice de qualité du modèle interne.

Soit $G_{ij}^{2(l)}$ le critère d'optimisation calculé pour le modèle à l'itération l avec l'ajout d'un arc entre $\boldsymbol{\xi}_i$ et $\boldsymbol{\xi}_j$. On part de $l = 0$.
L'algorithme a la forme suivante :

(1) Début de l'étape l
(2) Ajouter un arc entre $\boldsymbol{\xi}_i$ et $\boldsymbol{\xi}_j$
(3) Appliquer l'approche PLS
(4) Calculer $G_{ij}^{2(l)}$
(5) Retirer l'arc entre $\boldsymbol{\xi}_i$ et $\boldsymbol{\xi}_j$
(6) Répéter (2) à (5) pour toutes les relations n'existant pas dans le modèle
(7) Sélectionner l'arc entre $\boldsymbol{\xi}_k$ et $\boldsymbol{\xi}_l$ tel que :

$$G_{kl}^{2(l)} = \max_{i,j} G_{ij}^{2(l)}$$

(8) Ajouter l'arc au modèle et enregistrer le modèle l et la valeur $g^{2(l)} = G_{kl}^{2(l)}$
(9) $l = l + 1$
(10) Si $l = \frac{K(K-1)}{2} - 1$, l'algorithme s'arrête
(11) Sinon aller en (1)

Une fois le modèle saturé obtenu, le modèle m choisi est tel que

$$g^{2(m)} = \max_{l=1,\ldots,\frac{K(K-1)}{2}-1} g^{2(l)}$$

Cette méthode est intéressante mais pose un problème de quantité de calcul avec l'estimation de $\frac{K(K-1)}{2}(\frac{K(K-1)}{2}+1)$ modèles PLS. En fonction du type d'indices choisi, certains problèmes de cycles pourront apparaître. Le choix des orientations se fait uniquement en fonction de ces indices, il faut dont être prudent afin de ne pas intégrer des contresens dans la structure du modèle (des relations de cause à effet inversées). On aura tendance à choisir le F^2 afin d'éviter les cycles au niveau local et le GoF au niveau global. La validation globale pourra aussi se faire par des méthodes de rééchantillonnage.

3.4.3 Les modèles libres

Soit $\mathbf{t}_1^1, ..., \mathbf{t}_1^K$, les premières composantes principales issues de l'analyse en composantes principales effectuée sur les K blocs de variables manifestes. Si $\mathbf{t}_1^j$ représente une variable latente, toutes les autres variables latentes du modèle expliquent potentiellement $\mathbf{t}_1^j$. L'utilisation de corrélations classiques comme dans les algorithme de Hui (1982) ou de Hackl (2003) est possible. Néanmoins, des relations de dépendances complexes peuvent exister entre ces concepts et parfois une forte corrélation entre deux variables peut être expliquée par une troisième. Derquenne et Hallais (2004) ont donc décidé d'utiliser des corrélations partielles :

$$r_{\mathbf{t}_1^i,\mathbf{t}_1^j|\mathbf{Z}_{ij}} = \frac{r_{\mathbf{t}_1^i,\mathbf{t}_1^j} - r_{\mathbf{t}_1^i,\mathbf{Z}_{ij}} r_{\mathbf{t}_1^j,\mathbf{Z}_{ij}}}{\sqrt{(1 - r^2_{\mathbf{t}_1^i,\mathbf{Z}_{ij}})(1 - r^2_{\mathbf{t}_1^j,\mathbf{Z}_{ij}})}}, \tag{3.14}$$

avec $\mathbf{Z}_{ij}$ ensemble des premières composantes principales excepté $\mathbf{t}_1^i$ et $\mathbf{t}_1^j$.

Les associations entre variable latentes sont choisies en utilisant un seuil obtenu par un test statistique classique sur les corrélations partielles.

En terme de complexité, il y a $\frac{K(K-1)}{2}$ coefficients à calculer mais ceci ne permet d'obtenir que des relations non orientées.

A partir de ce point, Derquenne et Hallais (2004) utilisent un test basé sur le bootstrap afin de vérifier que le modèle de référence n'est pas apparu au hasard. Si l'hypothèse de ce test n'est pas rejetée, on peut alors orienter les arêtes du modèle.

Pour l'orientation des arcs, Hui (1982); Hackl (2003) ont proposé différentes alternatives basées sur un *a priori* des experts. Dans Jakobowicz et Derquenne (2007), nous tentons de développer une méthode statistique, qui ne prendra pas en compte la causalité, mais qui permet de maximiser un indice, dans notre cas le R^2. Des modèles ayant des qualités prédictives grandes seront obtenus mais l'expert devra toujours donner son avis afin de valider l'interprétation pratique des relations.

Soit s le nombre de liens obtenus par les tests effectués sur les premières composantes principales, on aura ainsi 2^s orientations possibles. Pour le modèle m, nous définissons :

$$\bar{R}_m^2 = \frac{1}{v_m} \sum_{r=1}^{v_m} R_m^2(\mathbf{y}_r; \mathbf{Z}_r), \tag{3.15}$$

avec $\mathbf{Z}_r$ ensemble des scores des variables latentes expliquant $\boldsymbol{\xi}_r$ et v_m nombre de variables endogènes dans le modèle m.

Une fois l'ensemble des modèles testés, le modèle M sélectionné est tel que :

$$M = \arg \max_{m=1,\ldots,s} \bar{R}_m^2. \tag{3.16}$$

Cette méthode apporte une nouvelle vision sur la construction du modèle, il pose toujours le problème des cycles dans les modèles et nécessite des précautions permanentes lors de son application. De plus, dans le cadre de PLS, comme les variables latentes n'évoluent pas (ce sont des composantes principales), il ne permet pas de visualiser l'évolution des construits et de prendre en compte les interactions au niveau du modèle interne.

La notion de corrélation partielle ne peut cependant pas être remplacée par la corrélation classique qui fera perdre tout l'avantage de cette méthode.

3.4.4 Une adaptation des modèles libres dans le cadre de l'approche PLS

Comme nous l'avons vu, les modèles libres proposent une alternative intéressante aux algorithmes classiques de construction du modèle interne par l'utilisation de corrélations partielles. Nous introduisons une approche issue de cette dernière mais qui devra éviter ses défauts.

Le principe

L'utilisation des premières composantes principales tout au long de la mise en place des modèles libres pose un problème de prise en compte des interactions au sein du modèle interne dans la construction de celui-ci. Malgré le faible impact de ce sous-modèle sur les estimations des variables latentes dans l'approche PLS, il nous a semblé important de le prendre en compte. Par ailleurs, on voudra éviter les cycles dans le processus d'orientation des arêtes.

Notre méthode est basée sur un algorithme itératif. A chaque étape, le lien sélectionné est celui qui maximise les corrélations partielles. De plus, son orientation se fait de manière à maximiser le R^2 moyen du modèle complet tout en évitant des cycles.

L'algorithme que nous présentons est complexe mais trouve une implémentation simple dans le cadre de l'approche PLS et donne des résultats satisfaisants comme nous le verrons dans le cadre du chapitre 7.4 de cette thèse.

L'algorithme

Soit $\mathbf{t}_1^1, ..., \mathbf{t}_1^K$ les premières composantes principales associées à chacun des blocs de variables latentes. Soient $\mathbf{y}_1, ..., \mathbf{y}_K$ les scores des variables latentes obtenus par l'approche PLS. Soit (i, j), l'arête non orientée entre les variables latentes $\boldsymbol{\xi}_i$ et $\boldsymbol{\xi}_j$. Soit $S = (i, j)_{\forall i,j,\, i \neq j}$ un ensemble regroupant initialement l'ensemble des arêtes possibles associées aux relations entre les variables latentes, S possède $\frac{K(K-1)}{2}$ éléments.

(1) Calculer $\mathbf{t}_1^1, ..., \mathbf{t}_1^K$

(2) Calculer la matrice de corrélations partielles $K \times K$ associées aux composantes principales :

$$R_{part} = (r_{\mathbf{t}_1^i, \mathbf{t}_1^j | T_{ij}})_{ij}$$

avec T_{ij} ensemble des premières composantes principales excepté $\mathbf{t}_1^i$ et $\mathbf{t}_1^j$

(3) Sélectionner l'arête (k, l), telle que :

$$(k, l) = \arg \max_{(i,j) \in S} (R_{part})_{ij}$$

(4) Ajouter le lien (k, l) au modèle

(5) Appliquer l'approche PLS avec l'arc $\boldsymbol{\xi}_k \rightarrow \boldsymbol{\xi}_l$ puis avec l'arc $\boldsymbol{\xi}_l \rightarrow \boldsymbol{\xi}_k$

(6) Sélectionner l'arc qui maximise :

$$\bar{R}_m^2 = \frac{1}{v_m} \sum_{r=1}^{v_m} R_m^2(\mathbf{y}_r; \mathbf{Z}_r)$$

avec $\mathbf{Z}_r$ ensemble des scores des variables latentes expliquant $\boldsymbol{\xi}_r$ et v_m nombre de variables endogènes dans le modèle m.

(7) Vérification de l'absence de cycle et des t-valeurs associées à la relation ajoutée, on conserve la variable dans l'équation si $|t| > 1.65$.

(8) On pose $S = S - \{(k, l)\}$

(9) Calculer la matrice de corrélations partielles issue des scores PLS :

$$R_{part} = (r_{\mathbf{y}_i, \mathbf{y}_j | Z_{ij}})_{ij}$$

avec Z_{ij} ensemble des scores PLS associés aux variables latentes excepté $\mathbf{y}_i$ et $\mathbf{y}_j$

(10) Aller en (3) et continuer jusqu'à obtenir un modèle stable.

Propriétés

La stabilité est atteinte comme dans le modèle de Hackl (2003), cette méthode est une combinaison des modèles libres (pour l'utilisation des corrélations partielles et la méthode d'orientation) et des approches itératives classiques (pour la convergence et la maximisation d'un indice).

L'utilisation d'une convergence basée sur la stabilité du modèle peut entraîner l'arrêt de l'algorithme au niveau d'un optimum local. Néanmoins, la simplification en terme de complexité du modèle est importante. En effet, si on se place dans le cadre de l'obtention du modèle saturé, on aura $\frac{K(K-1)}{2}$ matrices de corrélations partielles à calculer et le même nombre de modèles à estimer. Dans la pratique, le fait d'arrêter l'algorithme lorsque la stabilité est atteinte, entraîne généralement un nombre d'itération faible et l'obtention d'un modèle simple. Afin de minimiser le risque d'optimum local associé à

la statistique t, plusieurs itérations de l'algorithme pourront être nécessaires afin de valider le modèle obtenu.

Les résultats pratiques obtenus sont convaincants mais, comme dans les deux dernières approches, l'orientation des arcs doit être soumise à l'avis d'expert dans le domaine d'application car nous travaillons sur la maximisation d'un critère et non sur de vraies relations de cause à effet.

Conclusion

Cette approche offre une option intéressante car elle permet de combiner les avantages des approches de Hui (1982); Hackl (2003) et ceux des modèles libres. D'une part, on empêche les cycles, les arcs sont orientés et, d'autre part, on se base sur les scores PLS. Cette approche nous apparaît comme une alternative efficace à l'approche de Amato (2003).

Nous pourrions modifier le critère d'arrêt en se basant sur un autre indice mais ceci n'a pas été mené dans le cadre de cette thèse.

Nous comparons l'ensemble de ces méthodes sur des données réelles dans le cadre du chapitre 7.4 (p. 146).

Chapitre 4

La comparaison de groupes d'observations

4.1 Introduction

La comparaison de groupes d'observations sur un modèle d'équations structurelles à variables latentes constitue un domaine de recherche très dynamique. Le besoin de comparer des populations est fréquent et, dans le cadre de modèles complexes, nécessite des précautions. En ce qui concerne l'approche LISREL, de nombreuses méthodes existent, la majorité d'entre elles étant paramétriques. La comparaison de critères de qualité d'ajustements du modèle aux données y est fréquente (pour une présentation des indices, voir Bollen (2002) et pour une méthodologie complète, on peut voir Liao (2002)). Par ailleurs, l'utilisation de contraintes sur les valeurs des *loadings* permet de comparer des coefficients structurels.

Dans le cadre de l'approche PLS, il n'existe pas de fonction à optimiser basée sur une loi de probabilité comme celle du χ^2 et il faudra utiliser d'autres méthodes sans hypothèse de distribution. De nombreuses techniques ont été présentées. Celles-ci visent à comparer des coefficients structurels directement entre des groupes d'individus. Ces approches, bien que très adaptées, ne touchent qu'à un seul point de comparaison issu du modèle. Le nombre de paramètres associés à ce type de problèmes étant grand et les interactions entre ceux-ci nombreuses, les comparaisons devront se faire à plusieurs niveaux. Nous commençons par introduire les différents niveaux de comparaisons possibles, avec la présentation de méthodes classiques et d'autres alternatives. Nous présentons ensuite un processus de comparaison de modèles structurels permettant de mieux connaître les différences entre les deux groupes d'observations testés.

4.2 Les niveaux de comparaison

Un modèle d'équations structurelles à variables latentes est un modèle complexe dans lequel chaque relation est dépendante de nombreux paramètres. Lorsqu'on possède plusieurs groupes d'observations, l'application de chacun à un modèle n'aura pas les mêmes résultantes à différents niveaux.

Dans son ouvrage, Liao (2002) souligne que pour les modèles d'équations structurelles, on ne peut pas se limiter à un seul niveau de comparaison. L'auteur en rassemble plusieurs dans le cadre de l'approche LISREL :

- *Comparaison de distribution :* des méthodes non paramétriques existent afin de discerner les différences entre distributions.
- *Comparaison de la structure des données :* par le biais par exemple des corrélations.

– *Comparaison de la structure du modèle :* les relations entre variables peuvent différer d'un échantillon à un autre.
– *Comparaison des paramètres du modèle.*

Chacun de ces points a une importance capitale dans la compréhension des différences entre deux groupes d'observations par rapport à un modèle d'équations structurelles à variables latentes.

Dans le cadre de l'approche PLS, les recherches ont surtout concerné la comparaison de coefficients structurels. Or, on peut dénombrer trois niveaux de comparaisons différents et importants dans l'analyse des deux groupes d'observations :
– Comparaison de la structure des données : par le biais d'indices de qualité locaux ou globaux.
– Comparaison des scores des variables latentes.
– Comparaison des coefficients structurels.

Nous approfondissons chacun de ces niveaux par la suite.

4.3 Les tests de permutations

Nous introduisons un type de test statistique pour la comparaison d'échantillons non appariés qui sera spécialement bien adapté dans le cas de l'approche PLS. En effet, ce test est non paramétrique et constitue une alternative à l'utilisation du bootstrap dans le cas de deux échantillons.
Les tests basés sur des permutations des observations liées aux 2 groupes sont des méthodes efficaces et largement étudiées (Edgington, 1987).

Définition 4.1. *Soient deux échantillons G_1 et G_2, soit s une statistique permettant de tester une hypothèse H_0 associée aux deux échantillons. Un* ***test de permutations*** *est un test non paramétrique basé sur la méthodologie suivante :*

(1) *Calculer s pour les échantillons originaux en utilisant G_1 et $G_2 \rightarrow s_{original}$.*

(2) *Rassembler les deux échantillons dans un seul jeu de données $G_1 \cup G_2$.*

(3) *Permuter aléatoirement l'échantillon complet.*

(4) *Diviser l'échantillon permuté en deux groupes de la taille de G_1 et G_2.*

(5) *Calcul de s pour les échantillons permutés $\rightarrow s_{permut}$.*

(6) *Répéter (2) à (5) un grand nombre de fois N_{permut}.*

(7) *Obtention de la distribution de s_{permut} en supposant H_0 vraie.*

(8) *Si $s_{original}$ est une valeur extrême de la distribution de s_{permut} alors l'hypothèse nulle est rejetée.*

La p-valeur est obtenue à partir de la distribution de probabilité de s_{permut}.

Par ailleurs, on définit la probabilité :

$$P(s_{original} < s_{permut}) = \frac{1}{N_{permut}+1}\Big(\sum_{i=1}^{N_{permut}} I(s_{original} < s_{permut_i}) + 1\Big) \tag{4.1}$$

avec

$$I(s_{original} < s_{permut_i}) = \begin{cases} 1 \text{ si } s_{original} < s_{permut_i} \\ 0 \text{ sinon.} \end{cases}$$

L'hypothèse H_0 est rejetée lorsque cette probabilité est en dessous d'un certain seuil ($P < \alpha$).

Nous utiliserons ce test non paramétrique aux différents niveaux de comparaison dans le modèle d'équations structurelles.

4.4 La qualité globale

Dans le cadre des méthodes pour l'estimation des modèles structurels, les indices de qualité globale sont différents suivant la méthode.

4.4.1 La différence entre les méthodes LISREL et PLS

Dans le cadre de l'approche LISREL, les indices de qualité sont des indices de qualité d'ajustement du modèle aux données, ceux-ci sont majoritairement paramétriques et permettent des comparaisons directes. Ainsi, certains indices présentés dans le chapitre 1.3 sont tout à fait comparables d'un groupe à un autre.

Des tests sur les coefficients du modèle sont possibles, ainsi tous les paramètres peuvent être comparés : $\mathbf{\Lambda}_x$, $\mathbf{\Lambda}_y$, $\mathbf{\Gamma}$, $\mathbf{B}$, ... Si on veut comparer les coefficients structurels pour deux groupes d'individus, on testera :

$$H_0 \, : \, (\mathbf{B}_{G_1} = \mathbf{B}_{G_2}) \wedge (\mathbf{\Gamma}_{G_1} = \mathbf{\Gamma}_{G_2})$$

avec $\wedge$ signifie "et". Ces hypothèses sont généralement validées avec un test basé sur les multiplicateurs de Lagrange.

D'autre part, comme indice de qualité globale, on pourra comparer l'ensemble des paramètres du modèle associé à chacun des groupes : $\boldsymbol{\theta}_{G_1}$ et $\boldsymbol{\theta}_{G_1}$. On vérifiera donc l'hypothèse nulle :

$$H_0 \, : \, \boldsymbol{\theta}_{G_1} = \boldsymbol{\theta}_{G_2}$$

Pour cela on utilise un test du rapport de vraisemblance. Pour deux groupes, on a :

$$LRT = -2(L_R - L_U) = -2(L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_{G_1}) - L(\boldsymbol{\theta}_{G_2})) \sim \chi^2 \tag{4.2}$$

avec $L(\boldsymbol{\theta})$ log-vraisemblance de $\boldsymbol{\theta}$ sur $G_1 \cup G_2$ et $L(\boldsymbol{\theta}_{G_i})$ log-vraisemblance de $\boldsymbol{\theta}$ sur G_i. Des paramètres plus précis peuvent être testés en posant des contraintes d'égalité sur les paramètres restants.
Le rejet de H_0 nous amènera à considérer que chaque groupe représente une population différente en rapport avec ce modèle. Il faut donc remettre en question le modèle lui-même. Des adaptations non paramétriques sont possibles et pourront améliorer la qualité des tests (en effet, la relation au χ^2 est asymptotique et certaines perturbations peuvent advenir, voir Liao (2002)).

Par ailleurs, les indices comparatifs basés sur un modèle de référence (voir chapitre 1.3 et Bentler (1990)) peuvent être comparés pour deux groupes d'observations différents (NFI, $NNFI$ et les indices basés sur les multiplicateurs de Wald et Lagrange, Bentler (1990)).

Dans le cadre de l'approche PLS, l'aspect paramétrique lié à la méthode LISREL doit être abandonné. Par ailleurs, les indices ne sont plus ici des indices de qualité d'ajustement mais des indices de qualité prédictive. L'estimation de la qualité du modèle se fait indépendamment sur chacune des sous-parties du modèle (modèle de mesure et modèle structurel). Nous utilisons donc un éventail d'indices plus important. Nous nous attardons en 4.4.3 sur 3 d'entre eux.

4.4.2 Comparaison de la structure des données

Une comparaison des données sans prendre en compte le modèle est possible, on travaille alors sur la structure des données.
Afin de tester les différences au niveau de la structure des données, nous suivons les indications de Liao (2002). Comme nous évoluons dans le cadre de l'approche PLS, aucune hypothèse de distribution n'est nécessaire. Nous devons donc utiliser des tests non paramétriques.

La comparaison la plus "globale" consiste en la comparaison statistique des matrices de covariances des données (des variables manifestes), avec le test :

$$H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$$

Or, une différence entre deux matrices de covariances peut être expliquée par une différence entre les variances des variables des deux échantillons ($\boldsymbol{\sigma}_g^2$), cette différence n'entraîne pas forcément une structure différente. Même si l'égalité entre les variances nous intéresse par la suite. Nous préférons donc tester l'égalité entre les matrices de corrélations par le test :

$$H_0 : \boldsymbol{C}_1 = \boldsymbol{C}_2$$

avec $\boldsymbol{\Sigma}_g = \boldsymbol{\sigma}_2^T \boldsymbol{C}_g \boldsymbol{\sigma}_2$. Ce test ne peut pas se faire avec un test type Box M, une contrepartie non paramétrique doit être utilisée.

Cette procédure permet en cas de non rejet de supposer que les structures internes aux données sont proches et que, par conséquent, les données peuvent être comparées dans le cadre d'une méthode particulière telle que l'approche PLS.
D'autres tests sont possibles, mais celui-ci paraît le plus adapté. Néanmoins, étant très général, nous ne l'approfondirons pas davantage.

4.4.3 Comparaison de la qualité d'ajustement par des tests de permutations

Chin (2003) a appliqué les tests de permutations, présentés en 4.3, afin de comparer des coefficients structurels. Nous présentons l'adaptation de ces tests sur les indices de qualité prédictive dans le cadre de l'approche PLS.

Nous partons donc de l'hypothèse nulle :

$$H_0 : GF_{G_1} = GF_{G_2} \tag{4.3}$$

avec GF_{G_i} indice global associé au groupe G_i. Comme un test paramétrique n'est pas adapté, nous adaptons les tests de permutations à ce cas spécifique.

Ce type de tests est non paramétrique et consiste en la permutation des individus associés à deux échantillons distincts. Nous choisissons la statistique s pour les échantillons G_1 et G_2. Soit s la statistique du test, on prendra :

$$s = |GF_{G_1} - GF_{G_2}| \tag{4.4}$$

Le test présenté en 4.3 est adapté. A chaque étape, le calcul de GF se fait par l'application de l'approche PLS mode A sur chacun des échantillons permutés.

La probabilité P obtenue représente la probabilité que les deux modèles soient de qualité égale avec un degré de confiance donné par le seuil α, on prend généralement $\alpha = 0.05$ pour une confiance à 95%. Celui-ci doit être évalué en fonction du contexte et de la problématique.

Ce test permet donc de comparer des qualités prédictives globales d'un échantillon à un autre et peut amener à un certain nombre de remarques sur les différents modèles.

Les indices GF testés seront de différents types :

– Pour le modèle externe, la communauté moyenne. Elle représente la part de variance expliquée des variables latentes par le modèle externe :

$$\bar{H}^2 = \frac{1}{K}\sum_{j=1}^{K} H_j^2 = \frac{1}{K}\sum_{j=1}^{K}\frac{1}{p_j}\sum_{i=1}^{p_j} cor^2(\mathbf{x}_{ji}, \mathbf{y}_j)$$

où $\mathbf{y}_j$ est le score de la variable latente $\boldsymbol{\xi}_j$.
– Pour le modèle interne, la redondance moyenne :

$$\bar{F}^2 = \frac{1}{K_{endo}} \sum_{j:\boldsymbol{\xi}_j \text{ endogène}} F_j^2$$

– Pour le modèle global, le GoF (Tenenhaus et al., 2004) qui est une combinaison d'indicateurs de la validité du modèle interne et d'indicateurs de la validité du modèle externe

$$GoF = \sqrt{\bar{H}^2 \times \bar{R}^2}$$

Nous étudions les propriétés de ces tests par des simulations en 4.8. Une fois les tests appliqués, deux cas se profilent et des actions différentes sont nécessaires :

– L'hypothèse H_0 est rejetée ⇒ L'un des groupes a une qualité prédictive significativement meilleure que l'autre. On a alors deux possibilités, soit on respécifie le modèle conceptuel afin de vérifier que cette différence ne vient pas de la modélisation des équations structurelles, soit on passe à des comparaisons plus poussées sur d'autres paramètres. Dans le second cas, les résultats des tests globaux permettront de connaître la partie du modèle à étudier.

– H_0 n'est pas rejetée ⇒ Le modèle s'adapte aussi bien aux deux groupes. La structure des deux groupes par rapport au modèle est équivalente, on peut alors effectuer d'autres comparaisons à des niveaux plus détaillés.

4.4.4 La reconstruction du modèle

Dans le cas où les deux modèles ne sont pas comparables (les hypothèses associées aux tests précédents sont rejetées), il est possible de construire un nouveau modèle mieux adapté. Ce nouveau modèle global doit être mis en place grâce à une stratégie de construction du modèle.

Le type de méthode de construction varie en fonction des résultats des tests globaux.

On différencie trois cas :
– L'hypothèse $H_0 : \bar{H}^2_{G_1} = \bar{H}^2_{G_2}$ est rejetée et l'hypothèse $H_0 : \bar{F}^2_{G_1} = \bar{F}^2_{G_2}$ n'est pas rejetée. Alors on applique une méthode de classification de variables basées sur les composantes latentes en laissant le nombre de variables latentes fixe sur l'échantillon $G_1 \cup G_2$. On maximise :

$$T = N\sum_{k=1}^{K}\sum_{j=1}^{p_k} cov^2(\mathbf{x}_j, \mathbf{c}_k) \text{ avec la contrainte } \mathbf{c}_k'\mathbf{c}_k = 1$$

On peut voir le chapitre 3.3.2 (p. 67).
– L'hypothèse $H_0 : \bar{H}^2_{G_1} = \bar{H}^2_{G_2}$ n'est pas rejetée et l'hypothèse $H_0 : \bar{F}^2_{G_1} = \bar{F}^2_{G_2}$ est rejetée. Alors on applique la méthode de construction du modèle interne basée sur les corrélations partielles en conservant les blocs initiaux et les relations de causalité initiales sur l'échantillon $G_1 \cup G_2$ (voir chapitre 3.4.4, p. 77).
– Les hypothèses $H_0 : \bar{H}^2_{G_1} = \bar{H}^2_{G_2}$ et $H_0 : \bar{F}^2_{G_1} = \bar{F}^2_{G_2}$ sont rejetées. Alors on applique sur l'échantillon $G_1 \cup G_2$ une méthode de construction du modèle externe avec un nombre de variables

latentes non fixé (classification autour de composantes latentes associée à des permutations), puis une méthode de construction du modèle interne basée sur les corrélations partielles.

Une fois le nouveau modèle obtenu, les tests globaux sont appliqués sur le nouveau modèle. Si les hypothèses associées ne sont pas rejetées alors le nouveau modèle s'adapte aussi bien aux deux groupes et les comparaisons suivantes peuvent être effectuées. Si l'hypothèse du test est rejetée, il faudra considérer que les deux groupes doivent être modélisés par des équations structurelles différentes. La comparaison, comme nous la traitons, ne sera plus possible.

4.5 Les variables latentes

Les variables latentes constituent des indicateurs intéressants dans le cadre de l'approche PLS. L'un des principaux atouts de cette approche réside dans le calcul direct des variables latentes à partir des variables manifestes associées. Leur étude est fréquente comme, par exemple, dans l'analyse des indices de satisfaction des clients (Fornell, 1992).
Il est important dans cette optique de comparer les scores que l'on notera $\hat{\boldsymbol{\xi}}_k$. Nous utilisons le formalisme généralement utilisé pour les variables latentes, sur une échelle de 0 à 100 :

$$\hat{\boldsymbol{\xi}}_k = \frac{\sum w_{kj}\mathbf{x}_{kj}}{\sum w_{kj}}$$

$$\hat{\boldsymbol{\xi}}_k^* = 100 \times \frac{\hat{\boldsymbol{\xi}}_k - \mathbf{x}_{min}}{\mathbf{x}_{max} - \mathbf{x}_{min}}$$

en supposant l'ensemble des variables manifestes sur la même échelle et $\mathbf{x}_{min}$ (resp. $\mathbf{x}_{max}$) est le minimum (resp. maximum) de cette échelle.

Leur étude peut se faire de deux manières pour deux groupes G_1 et G_2 sur un modèle M :

– En comparant les moyennes $\mu_{\boldsymbol{\xi}_k^{*(G_1)}}$ et $\mu_{\boldsymbol{\xi}_k^{*(G_2)}}$. L'hypothèse nulle est :

$$H_0 : \mu_{\boldsymbol{\xi}_k^{*(G_1)}} = \mu_{\boldsymbol{\xi}_k^{*(G_2)}} \tag{4.5}$$

On utilise $\hat{\mu}_{\boldsymbol{\xi}_k^{*(G_1)}} = E(\hat{\boldsymbol{\xi}}_k^{*(G_1)})$ et $\hat{\mu}_{\boldsymbol{\xi}_k^{*(G_2)}} = E(\hat{\boldsymbol{\xi}}_k^{*(G_2)})$ afin de valider cette hypothèse ainsi qu'une procédure non paramétrique basée sur un test de permutation. La statistique s utilisée sera :

$$s = |E(\hat{\boldsymbol{\xi}}_k^{*(G_2)}) - E(\hat{\boldsymbol{\xi}}_k^{*(G_1)})|$$

En cas de rejet de l'hypothèse nulle, les différences permettent d'évaluer si le degré de satisfaction de chacune des populations est différent.

– En comparant les variances $\sigma^2_{\boldsymbol{\xi}_k^{*(G_1)}}$ et $\sigma^2_{\boldsymbol{\xi}_k^{*(G_2)}}$. L'hypothèse nulle est alors :

$$H_0 : \sigma^2_{\boldsymbol{\xi}_k^{*(G_1)}} = \sigma^2_{\boldsymbol{\xi}_k^{*(G_2)}} \tag{4.6}$$

On utilise $\hat{\sigma}^2_{\boldsymbol{\xi}_k^{*(G_1)}} = S_{\hat{\boldsymbol{\xi}}_k^{*(G_1)}}$ et $\hat{\sigma}^2_{\boldsymbol{\xi}_k^{*(G_2)}} = S_{\hat{\boldsymbol{\xi}}_k^{*(G_2)}}$ associé à une procédure non paramétrique. Nous utilisons de la même façon que précédemment un test de permutation basé sur une statistique s définie par :

$$s = |S_{\hat{\boldsymbol{\xi}}_k^{*(G_2)}} - S_{\hat{\boldsymbol{\xi}}_k^{*(G_1)}}|$$

Les variances ont un rapport direct avec le modèle, en effet, elles sont directement reliées à la méthode d'estimation et à l'obtention des coefficients structurels. Ce test nous permettra de comparer la dispersion de chacun des groupes en se basant sur le concept associé à la variable latente.

La comparaison des variables latentes permet d'apporter des éléments très importants dans la compréhension et l'interprétation des différences entre deux groupes d'observations sur un modèle structurel.

4.6 Les coefficients structurels

Le dernier niveau de comparaison se fait entre les coefficients structurels estimés. De nombreux chercheurs ont abordé cette problématique, en général, dans le cadre du *multigroup comparison*. Plusieurs approches de comparaison ont été développées pour l'approche PLS, elles sont toutes basées sur des principes de rééchantillonnage.

Quatre approches ont été introduites :
- L'approche paramétrique classique.
- L'approche par effet de modération.
- L'approche par test non paramétrique.
- L'approche basée sur les permutations.

4.6.1 Comparaison paramétrique classique

Elle est basée sur le bootstrap classique, on procède de la manière suivante :
- Pour chacun des groupe G_1 et G_2, estimer les paramètres du modèle par l'approche PLS $\hat{\beta}_{ij}^{G_k}$.
- Estimation des écarts types $SE^2_{\hat{\beta}_{ij}^{G_k}}$ par bootstrap.
- Calcul de la statistique t définie par :

$$t = \frac{\hat{\beta}_{ij}^{G_1} - \hat{\beta}_{ij}^{G_2}}{[\sqrt{\frac{(N_1-1)^2}{N_1+N_2-2}SE^2_{G_1} + \frac{(N_2-1)^2}{N_1+N_2-2}SE^2_{G_2}}][\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}]} \tag{4.7}$$

- Lorsqu'un certain nombre d'hypothèses sont vérifiées, cette statistique suit asymptotiquement une loi t de Student à $N_1 + N_2 - 2$ degrés de liberté.

Les hypothèses sont la normalité des résidus et des variances proches entre les coefficients pour les deux groupes.

4.6.2 Comparaison basée sur les effets modérateurs

Les variables modératrices sont des variables catégorielles qui influencent la force d'une relation entre deux variables : l'une exogène, l'autre endogène.

On peut ramener l'effet de ce type de variables à celui de plusieurs groupes d'observations. Baron et Kenny (1986) les utilisent sans variables latentes. Ainsi, la figure 4.1 illustre la recherche d'un effet modérateur. Si le coefficient c est significatif, alors il existe un effet modérateur induit par la variable modératrice.

Chin et al. (1996) en ont illustré l'application dans le cadre de variables latentes, au lieu d'utiliser des variables observées, ils utilisent des variables latentes. Supposons que la variable latente endogène soit constituée de p_1 variables manifestes et que la variable latente modératrice de p_2 variables manifestes. Alors, la variable produit sera constituée de $p_1 \times p_2$ indicateurs.

On peut appliquer cette étude à l'effet modérateur de deux groupes d'observations au cas de l'approche PLS.

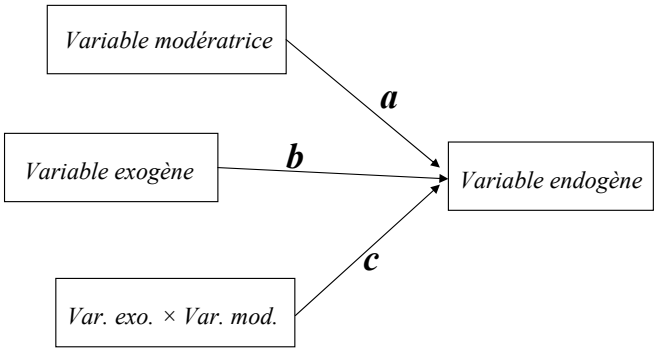

FIG. 4.1 – Illustration de l'effet modérateur lié à une variable

Ainsi, si on cherche à modéliser l'impact de groupes sur un coefficient, on peut utiliser cette notion de variable modératrice. Pour deux groupes, on aura une variable modératrice x^{mod}. Elle est définie par :

$$x_i^{mod} = \begin{cases} 1 & \text{si l'observation } i \text{ appartient au groupe } G_1 \\ 0 & \text{si l'observation } i \text{ appartient au groupe } G_2 \end{cases}$$

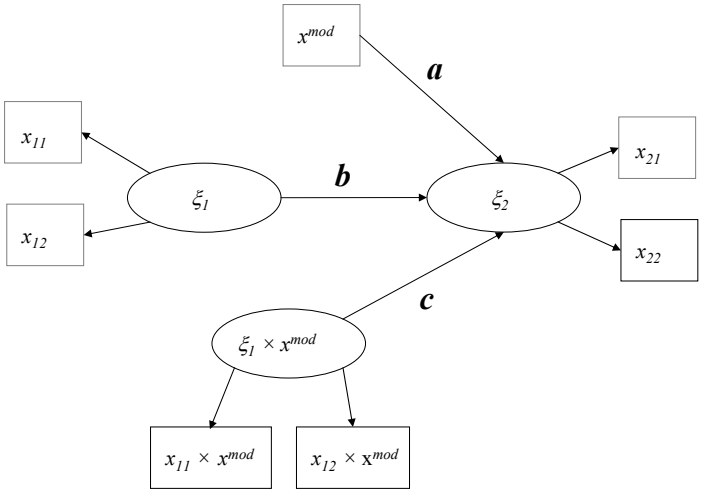

FIG. 4.2 – Illustration de l'effet modérateur lié à une variable dans le cadre de deux variables latentes

Le modèle aura alors la forme de celui de la figure 4.2 pour deux variables latentes ayant deux indicatrices chacune. L'estimation de la significativité du paramètre c se fait par le biais du bootstrap.

Cette approche est intéressante car elle ne possède pas d'hypothèses de distribution. Cependant, elle pose plusieurs problèmes, le fait que l'effet obtenu ne pourra pas être clairement défini. L'effet modérateur est difficile à interpréter. D'autre part, l'hypothèse d'erreurs de mesure non corrélées sera difficilement tenable dans ce nouveau modèle. Chin et al. (1996) ont vérifié que la violation de cette hypothèse n'entraînait pas de graves problèmes dans le cadre de simulations de Monte-Carlo.

4.6.3 Comparaison basée sur un test non paramétrique

Henseler (2007) a présenté un test simple afin de comparer des coefficients. Il tente de valider l'hypothèse :

$$H_0 : \beta_{ij}^{G_1} > \beta_{ij}^{G_2}$$

Pour cela, il met en place un processus facilement applicable avec des logiciels classiques de traitement de l'approche PLS. La méthode est la suivante :

(1) Pour chacun des groupes, estimer le paramètre d'intérêt en utilisant l'approche PLS.

(2) Pour chacun des groupes, appliquer du bootstrap sur ce même paramètre. On obtient J échantillons bootstrap.

(3) Construire l'ensemble des combinaisons possibles des paramètres bootstrap entre chacun des groupes (J^2 comparaisons).

(4) Compter le nombre de fois que l'un des coefficients est plus grand que l'autre. La fréquence obtenue est une probabilité :

$$P(\hat{\beta}_{ij}^{G_1} > \hat{\beta}_{ij}^{G_2}) = \sum_{l=1}^{J} \sum_{m=1}^{J} I(\hat{\beta}_{ij}^{G_1(l)} > \hat{\beta}_{ij}^{G_2(m)})$$

avec $I(\cdot)$ fonction indicatrice et l'exposant (l) représente l'échantillon bootstrap l sur les J échantillons.

Cette méthode est intéressante car elle se détache de toute notion de distribution de probabilité. Elle se rapproche des notions de tests de rangs ou du type Mann-Whitney.

4.6.4 Comparaison basée sur des tests de permutations

Comme nous l'avons vu précédemment, les tests de permutations constituent de bonnes techniques afin de valider des hypothèses (Edgington, 1987). Chin (2003) propose une approche non paramétrique basée sur ce type de tests afin de valider la significativité des différences entre les paramètres.

L'application de ce test se fait de la même façon que celle présentée en 4.3. L'hypothèse nulle testée est :

$$H_0 : \beta_{ij}^{G_1} = \beta_{ij}^{G_2}$$

Chin (2003) ne précise pas quel type de statistique s il utilise dans son article, on peut prendre :

$$s = |\hat{\beta}_{ij}^{G_1} - \hat{\beta}_{ij}^{G_2}| \qquad ou \qquad s = (\hat{\beta}_{ij}^{G_1} - \hat{\beta}_{ij}^{G_2})^2$$

On obtient ainsi une probabilité P de rejet de l'hypothèse qui permet de savoir si l'égalité entre les coefficients doit être rejetée avec un seuil α fixé.

Cette méthode combine les avantages des autres approches et permet d'obtenir des informations fiables sur la significativité de la différence entre les coefficients structurels. Nous étudions la puissance de celle-ci dans le cas de données simulées en 4.8.

4.7 Elaboration d'un processus de comparaison : des données aux coefficients structurels

La comparaison de groupes dans le cadre des modèles d'équations structurelles à variables latentes impose, du fait de la complexité des relations, un certain nombre de précautions d'utilisation. Le nombre de paramètres associés à ce type de problématiques est grand et les interactions entre ceux-ci sont nombreuses. On ne pourra pas simplement comparer deux coefficients en utilisant la statistique t indépendamment du reste des paramètres comme cela se fait largement dans la pratique et même dans certains articles de recherche (Thompson et al., 1994).
La plupart des applications dans ce domaine ne s'intéressent qu'aux coefficients structurels (Eberl, 2007; Sanchez-Franco, 2006; Keil et al., 2006). Or, afin d'analyser deux groupes sur des modèles d'équations structurelles, chaque niveau de comparaison devra être étudié.

Les méthodes classiques de comparaison des coefficients ne prennent pas en compte la structure des données. La procédure habituelle consiste à comparer des coefficients du modèle obtenu sur chacun des échantillons. La validation se fait généralement par des méthodes du type bootstrap et un test de Student permet d'estimer la significativité des différences. Comme le constate Chin (2003), cette procédure basique pose un problème car les tests effectués supposent que la structure de chacun des échantillons soit similaire, qu'ils aient des tailles proches et que les résidus soient distribués normalement.

Nous présentons donc un processus afin de comparer deux groupes d'observations sur un modèle prédéfini. Cette approche vise à qualifier globalement les différences entre les groupes au niveau de ce modèle. Il est particulièrement bien adapté au cas de l'approche PLS car il se base sur des tests non paramétriques.

Dans le cadre de ce processus, nous prenons en compte les différents niveaux évoqués plus haut. Nous nous inspirons, entre autres, des recherches de Amato et Balzano (2003) sur le développement de processus afin de comparer des échantillons. Ils utilisent pour cela des méthodes de construction du modèle et des indices globaux. Une fois le modèle construit pour chaque échantillon, des indices globaux sont calculés et validés par bootstrap non paramétrique. Ainsi un "overall fit" peut être obtenu et les qualités d'ajustement des modèles peuvent être comparés.
Par ailleurs, nous nous différencions ici des travaux sur la discrimination se basant sur le modèle dans le cadre de l'approche PLS. Ces travaux visent à obtenir des groupes d'individus à partir d'un modèle et d'une distance entre les individus par rapport au modèle. De nombreuses approches sont en concurrence, FIMIX-PLS (Ringle et al., 2007), PLS Typological path modeling (Squillacciotti, 2007), REBUS-PLS (Trinchera et al., 2007), ou PATHMOX (Sanchez et Aluja, 2006), mais ne s'intègrent pas dans ce travail.

4.7.1 La structure du processus

Le but de cette recherche est de mettre en place un processus de comparaison de deux groupes d'observations issus ou non de la même population. Ces groupes sont associés à un même modèle structurel à variables latentes. La comparaison se fait en tenant compte de la complexité du modèle.
Nous nous inspirons des différents points de comparaison de Liao (2002). De plus, nous devrons prendre en compte le fait que les données pourront suivre des distributions différentes, avoir des structures différentes et des tailles d'échantillons variables.

Le modèle sous-jacent doit donc être similaire entre les deux échantillons. Ce point est primordial dans un traitement comparatif de deux échantillons. Si la structure des données n'est pas la même que celle du modèle ou est différente d'un échantillon à un autre, alors toute conclusion sur les variations des coefficients structurels, même validée par bootstrap n'a pas de valeur. C'est pour cette raison que,

préalablement à l'application du bootstrap ou d'un test de comparaison non paramétrique sur les coefficients structurels, des tests, d'une part, sur l'adéquation des données au modèle et, d'autre part, des tests afin d'évaluer les différences entre les structures des échantillons doivent être menés.

Avant toute comparaison, les conditions d'application de l'approche PLS doivent être vérifiées pour chacun des échantillons (consistance interne pour le cas réflectif).

Supposons que l'on dispose de deux échantillons issus de deux populations différentes sur les mêmes variables. La première question qui se pose dans le cadre des modèles d'équations structurelles à variables latentes est la comparabilité des échantillons. Comme aucune hypothèse de distribution n'est nécessaire dans le cadre de l'approche PLS, nous ne testons pas la différence de distribution. Nous allons donc nous intéresser aux structures des relations entre les variables de chacun des échantillons.

4.7.2 Le processus

Première étape : Comparaison des qualités globales

Nous utilisons le test présenté en 4.4.3 (p. 84) sur deux indices différents : $\bar{H}^2$ et $\bar{F}^2$.
Soit G_1 et G_2 deux échantillons, nous utilisons le test de permutation basé sur les hypothèses nulles :

$$H_0^{H^2} : \bar{H}^2_{G_1} = \bar{H}^2_{G_2}, \quad H_0^{F^2} : \bar{F}^2_{G_1} = \bar{F}^2_{G_2}$$

La probabilité P obtenue pour chacun des tests permet de connaître le degré de significativité de la différence. Dans la pratique, on suppose que si $P < 0.05$, l'hypothèse associée au test est rejetée et les indices sont significativement différents.

- Aucune hypothèse n'est rejetée : les indices de qualité prédictive sont proches et les données s'adaptent de la même façon au modèle.
- L'une des hypothèse est rejetée : le traitement dépendra de la problématique voulue.
  - Si on désire chercher un modèle qui s'adapte aux deux échantillons, on reconstruit le modèle $\longrightarrow$ Seconde étape.
  - Si le modèle est fixé et ne peut pas être modifié, alors on effectue des comparaisons plus détaillées tout en sachant que les groupes ont un fort risque d'avoir des différences significatives au niveau des autres paramètres du modèle.

Seconde étape : Reconstruction du modèle

On va chercher un modèle qui s'adapte bien aux deux échantillons. On utilise les méthodes de construction présentées dans le chapitre 3 et en 4.4.4 (p. 85). On reconstruit le modèle conceptuel en se basant sur l'échantillon $G_1 \cup G_2$ en fonction des hypothèses rejetées sur la comparaison globale de la première étape.
Si les hypothèses des tests sur le nouveau modèle ne sont pas rejetées, on pourra passer à la troisième étape.
Si les hypothèses des tests sur le nouveau modèle sont rejetées, il faudra soit utiliser le modèle initial afin de comparer les autres paramètres du modèle, soit construire deux modèles différents et les étudier indépendamment.

Troisième étape : La comparaison des variables latentes

On suit les indications données en 4.5 (p. 86). Ces comparaisons nous amènent à de nouvelles interprétations sur les différences entre les groupes.

Quatrième étape : La comparaison des coefficients

Si les modèles obtenus sont comparables en terme de qualité globale, nous pouvons alors nous attacher à la comparaison des coefficients. Quelques tests simples doivent être préalablement

appliqués : on doit vérifier l'égalité des variances des coefficients, la normalité des résidus et l'équivalence des tailles d'échantillons. En fonction des résultats, différentes approches pourront être suivies. Ces tests sont basés sur des méthodes de rééchantillonnage.

Pour chaque comparaison, quelques tests simples devront préalablement être appliqués :
– Test sur l'égalité des variances (T1)
– Test sur la normalité des résidus (T2)
En fonction des résultats, différentes approches pourront être suivies.

1. *Variances proches, tailles équivalentes et ne déviant pas trop de la normalité.* Dans ce cas, la validation se fait par des tests de Student classiques avec :

$$t = \frac{\hat{\beta}_{ij}^{G_1} - \hat{\beta}_{ij}^{G_2}}{[\sqrt{\frac{(N_1-1)^2}{N_1+N_2-2}SE_{G_1}^2 + \frac{(N_2-1)^2}{N_1+N_2-2}SE_{G_2}^2}][\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}]} \tag{4.8}$$

   où N_1 et N_2 sont les tailles des échantillons G_1 et G_2, et SE^2 représente la variance de chaque estimation de coefficient par bootstrap. Ce t suit une t-distribution à $N_1 + N_2 - 2$ degrés de liberté.
2. *Variances différentes, tailles équivalentes et ne déviant pas trop de la normalité.* On utilise alors un test de Smith-Satterthwait :

$$t = \frac{\hat{\beta}_{ij}^{G_1} - \hat{\beta}_{ij}^{G_2}}{\sqrt{SE_{G_1}^2 + SE_{G_2}^2}};\ DF = \frac{(SE_{G_1}^2 + SE_{G_2}^2)^2}{\frac{SE_{G_1}^2}{N_1+1} + \frac{SE_{G_2}^2}{N_2+1}} - 2 \tag{4.9}$$

   avec un nombre de degrés de liberté égal à l'entier le plus proche de DF. Cependant, cette approche est aussi basée sur un test paramétrique et ne pourra pas s'appliquer dans le cas où les résidus ne sont pas normaux.
3. *Autres cas.* Nous préférons l'approche de Chin (2003) basée sur les permutations car elle s'intègre bien dans ce processus et elle permet de tester l'hypothèse H_0 : $\beta_{ij}^{G_1} = \beta_{ij}^{G_2}$ en obtenant une probabilité facilement interprétable.

Le processus obtenu permet donc de tester les principaux niveaux de comparaison nécessaires à l'étude poussée de deux groupes d'observations sur un modèle structurel à variables latentes dans le cadre de l'approche PLS. La figure 4.7 (p. 96) constitue un résumé de ce processus complexe.

4.8 Simulations

Nous effectuons des simulations afin de connaître les propriétés des tests présentés. Nous commençons par introduire les notions qui nous intéressent afin de valider les tests :

Définition 4.2. *La **probabilité d'erreur de première espèce** est définie par la probabilité :*

$$P(H_0 \text{ rejeté } | H_0 \text{ vraie}) = \alpha$$

Définition 4.3. *La **probabilité d'erreur de seconde espèce** est définie par la probabilité :*

$$P(H_0 \text{ non rejetée } | H_0 \text{ fausse}) = \beta$$

La probabilité $1 - \beta$ représente la puissance du test.

Nous allons donc essayer d'estimer la sensibilité de la puissance des tests $(1-\beta)$ en fonction des différences entre les groupes pour un α donné.

Nous simulons donc des données par méthode de Monte Carlo en utilisant un modèle très simple dans lequel nous faisons varier différents facteurs. On obtient 3 processus, chacun basé sur la variation des paramètres du modèle d'un groupe à l'autre.

(P1) Tous les coefficients du modèle externe varient. On utilise l'indice et l'hypothèse suivants :

$$\Delta_{\bar{H}^2} = |cor(\mathbf{x}_{kj}^{G_1}, \boldsymbol{\xi}_k^{G_1}) - cor(\mathbf{x}_{kj}^{G_2}, \boldsymbol{\xi}_k^{G_2})| \quad k = 1,2,3,\ j = 1,2$$
$$H_0^{H^2} : \bar{H}_{G_1}^2 = \bar{H}_{G_2}^2$$

(P2) Tous les coefficients du modèle interne varient. On utilise l'indice et l'hypothèse suivants :

$$\Delta_{\bar{F}^2} = |\hat{\beta}_{ij}^{G_1} - \hat{\beta}_{ij}^{G_2}| \quad (i,j) = (1,2),(1,3)$$
$$H_0^{F^2} : \bar{F}_{G_1}^2 = \bar{F}_{G_2}^2$$

(P3) Un coefficient structurel particulier varie. On utilise l'indice et l'hypothèse suivants :

$$\Delta_\beta = |\hat{\beta}_{13}^{G_1} - \hat{\beta}_{13}^{G_2}|$$
$$H_0 : \beta_{13}^{G_1} = \beta_{13}^{G_2}$$

Dans le cadre de ces simulations, nous fixons $\alpha = 0.05$ et analysons β en fonction des variations de Δ.

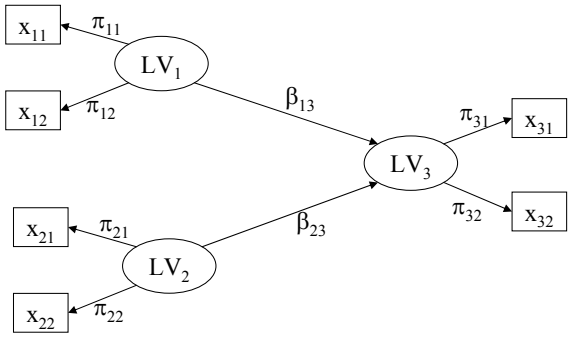

FIG. 4.3 – Modèle structurel simulé

Le modèle utilisé est celui de la figure 4.3 et l'obtention du modèle est basée sur l'utilisation de variables simulées pour les variables latentes exogènes et pour les termes d'erreur. Le modèle comporte 3 variables latentes et 2 variables manifestes par variable latente. Chaque processus de simulation et de test est répété 1000 fois.
Pour chaque processus et pour différentes tailles d'échantillons, nous obtenons la courbe de puissance empirique. Celle-ci est obtenue en faisant varier Δ. Nous obtenons, pour 3 tailles d'échantillons $(N = 100, 500, 1000)$ avec les groupes équidistribués (50, 250 et 500 observations par groupe), les courbes des figures 4.4 pour la communauté, 4.5 pour la redondance et 4.6 pour le coefficient structurel.

Notons que l'échelle des Δ est différente pour les communautés. Il s'avère que le test sur les communautés est plus sensible que les autres tests aux modifications du modèle. Ainsi, pour une différence de 0.3, on a de l'ordre de 99% de rejet pour toutes les tailles d'échantillon. Les courbes sur les redondances

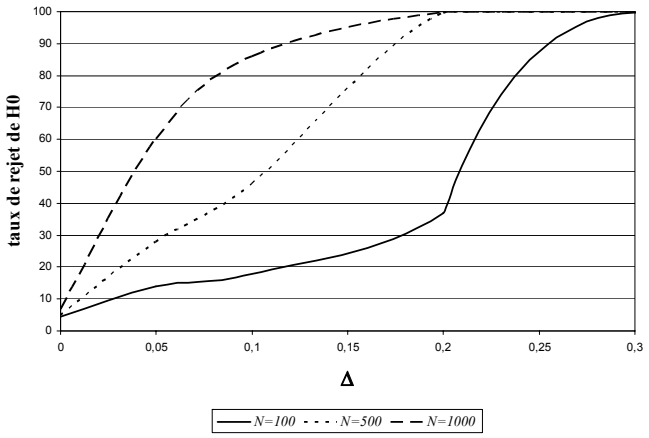

FIG. 4.4 – Courbe de puissance empirique pour le test basé sur les communautés moyennes ($\bar{H}^2$)

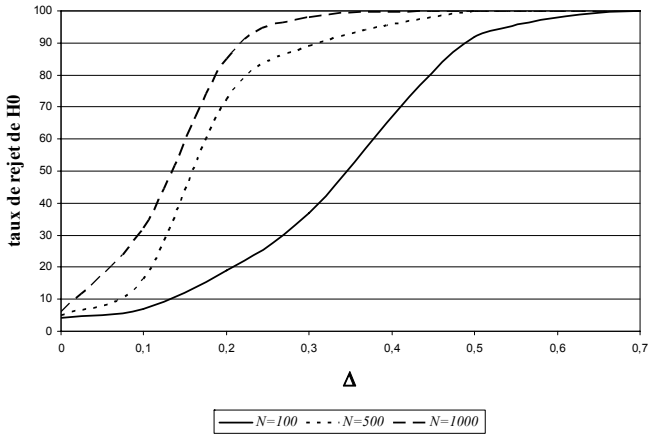

FIG. 4.5 – Courbe de puissance empirique pour le test basé sur les redondances moyennes ($\bar{F}^2$)

et les coefficients structurels sont très proches. Ces deux tests ont le même type de comportement en fonction de leur Δ. Par contre, si on étudie la courbe de puissance empirique du test sur les coefficients structurels en faisant varier $\Delta_{\bar{F}^2}$, on obtient de mauvais résultats. On aura, même pour une grande différence, un taux de rejet faible.
Si on étudie la taille des groupes, on voit que pour des groupes de taille 50, les courbes sont aplaties. Par contre, pour des groupes de 500 observations, les courbes donnent de bons résultats. Une variation de l'ordre de 0.2 est donc nécessaire afin d'obtenir un effet significatif pour des échantillons de plus de 200 observations.

Des simulations supplémentaires seraient utiles afin de tester l'influence d'autres paramètres. Par exemple, on pourrait utiliser deux groupes de tailles différentes ou étudier les courbes de puissance empirique associées aux autres tests présentés dans ce chapitre. Néanmoins, cette étude nous a permis de voir que les tests introduits avaient des puissances élevées tant que la différences entre les coefficients est "assez" grande (de l'ordre de 0.2) et que les groupes sont suffisamment grands (au moins 200 observations).

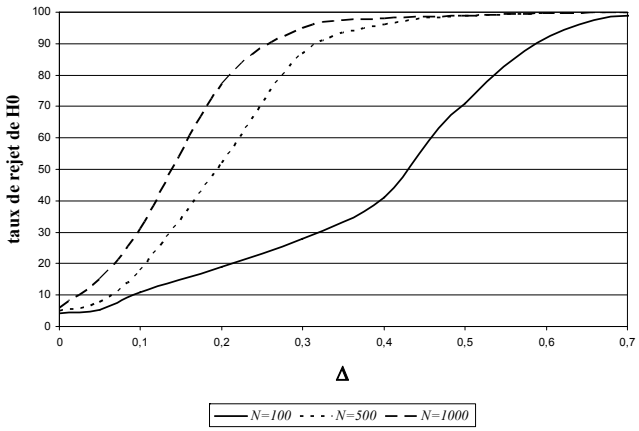

FIG. 4.6 – Courbe de puissance empirique pour le test basé sur les coefficients structurels

4.9 Conclusion

Nous avons présenté les techniques de comparaison de groupes d'observations en se basant sur un modèle d'équations structurelles à variables latentes. Il ressort de cette étude que l'analyse de deux échantillons ne se limite pas à l'impact de chaque groupe sur un coefficient structurel et que cette facette largement étudiée ne constitue qu'une partie de la comparaison. Le processus proposé permet d'explorer, niveau par niveau, les différentes facettes de comparaisons entre deux groupes d'observations (cf. fig. 4.7). Les simulations présentées nous ont montré que les tests basés sur des permutations obtenaient de bon résultats.
Nous illustrons le processus de comparaison dans le cadre des applications à l'analyse de la satisfaction des clients dans le chapitre 7.

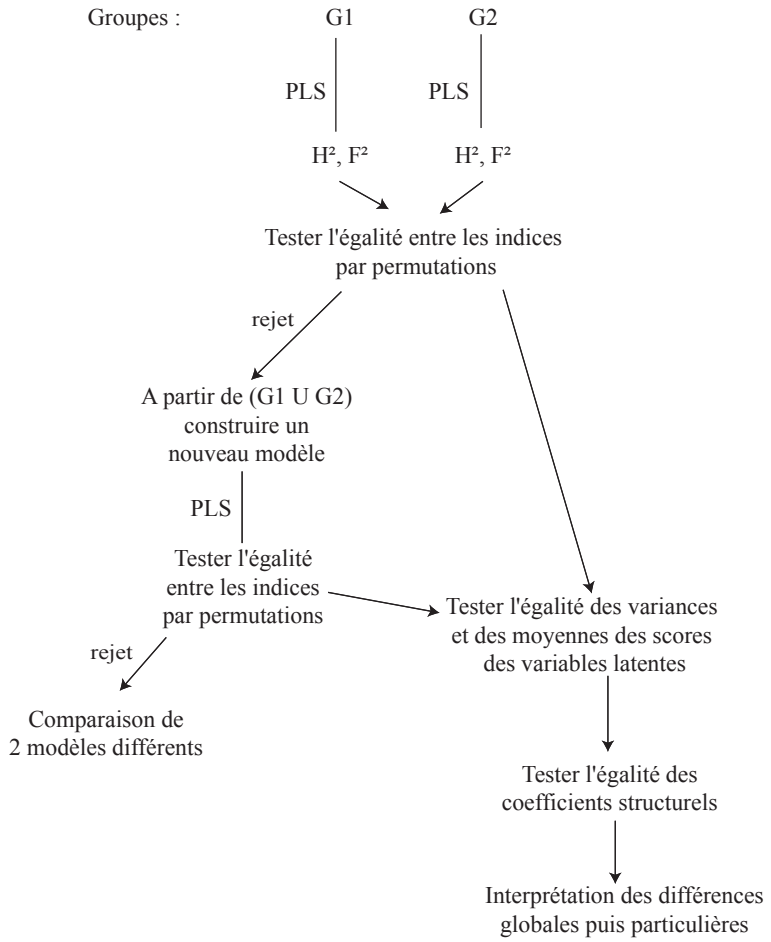

FIG. 4.7 – Processus de comparaison de groupes d'individus dans le cadre de l'approche PLS

Chapitre 5

La non linéarité dans les modèles d'équations structurelles

5.1 Introduction et motivations

Les modèles d'équations structurelles à variables latentes permettent de modéliser des relations complexes en se basant sur des relations linéaires (toutes les équations du modèle sont linéaires). L'introduction de transformations non linéaires pourrait apporter des informations supplémentaires sur les relations existant entre les variables (qu'elles soient manifestes ou latentes).

Dans le cadre de modèles aussi complexes que ceux développés dans le cadre des modèles structurels, on peut supposer que certaines relations du modèle ne sont pas linéaires. De plus, les recherches dans les domaines d'application montrent parfois des relations non linéaires entre certains concepts. Par exemple, pour l'analyse de la satisfaction et de la fidélité des clients, la littérature marketing montre que la relation entre satisfaction et fidélité s'avère souvent non linéaire, ainsi que celle entre la satisfaction et ses facteurs (les variables manifestes associées à la variable latente satisfaction). Nous développons ce cas spécifique dans le cadre des applications de cette thèse (cf. chapitre 7.7, p. 153). Nous nous intéressons donc à la découverte de relations non linéaires dans des modèles complexes.

Nous séparons nos travaux en deux parties. Dans un premier temps, nous présentons les grandes lignes des méthodes déjà mises en place, d'une part, dans le cadre de la méthode LISREL et, d'autre part, dans celui de l'approche et de la régression PLS. Dans un second temps, nous proposons des approches afin de traiter des modèles d'équations structurelles en utilisant les théories basées sur la transformation optimale des variables. Ces méthodes s'appliquent dans le cadre de l'approche PLS.

5.2 La non linéarité dans les modèles d'équations structurelles

Du fait de la complexité des modèles structurels, on ne pourra pas étudier la non linéarité de toutes les relations simultanément. Il faudra d'abord choisir à quel niveau on désire travailler. Ces niveaux sont les suivants :

- les relations du modèle externe, modélisées par des équations de régressions linéaires classiques entre variables latentes et manifestes (cf. équation 1.12 (p. 28) pour la méthode LISREL et équation 1.1 (p. 20) pour l'approche PLS) ;
- les relations du modèle interne modélisées par des équations de régressions entre les variables latentes (cf. équation 1.13 (p. 28) pour la méthode LISREL et équation 1.3 (p. 21) pour l'approche PLS).

Nous séparons le cas de chacune des méthodes d'estimation du modèle. Leurs procédés étant "par essence" différents, on ne pourra pas mettre en place un processus adaptable aux deux approches (cf. chapitre 2, p. 39).

Par ailleurs, une méthode pour traiter des modèles d'équations structurelles non linéaires devra vérifier trois hypothèses :

1. La simplicité de mise en place.
2. L'interprétabilité des résultats.
3. Le respect des "vraies" relations entre les variables (si la relation est linéaire, alors on obtiendra bien une relation linéaire).

5.2.1 L'approche LISREL

Les principaux travaux sur la non linéarité dans les modèles d'équations structurelles se sont focalisés sur l'approche LISREL dont l'application est aujourd'hui plus développée que celle de l'approche PLS. De plus, le fait d'utiliser une estimation par maximum de vraisemblance donne un réel avantage sur l'approche PLS qui nécessite une méthode d'estimation basée sur les moindres carrés comme l'algorithme PLS lui-même.

Jusqu'à présent, les méthodes d'équations structurelles à variables latentes sont basées sur des équations linéaires. Néanmoins, il est admis que prendre en compte des relations non linéaires dans le modèle permet de mieux représenter la réalité. On peut voir entre autres Ajzen et Fishbein (1980) et les articles dans Marcoulides et Schumacker (1998). L'analyse factorielle non linéaire a été étudiée par McDonald (1962, 1967), puis Etezadi-Amoli et McDonald (1983), et Zhu et Lee (1999).

Depuis le début des années 1980, les chercheurs ont présenté des approches afin d'intégrer de la non linéarité dans les modèles structurels. Busemeyer et Jones (1983) sont les premiers à introduire des produits de variables latentes dans les modèles, mais ils le font pour des variables latentes à un seul indicateur ce qui limite grandement l'utilisation de cette méthode. Kenny et Judd (1984) utilisent une méthode plus générale afin d'intégrer des produits ou des carrés de variables latentes dans le modèle. Pour cela, ils ajoutent des équations qui reflètent par le biais de produits des variables manifestes, les variables latentes produits. Le modèle ainsi transformé est estimé par une fonction de moindres carrés généralisés (GLS) qui permet une estimation consistante des paramètres. Néanmoins, cette approche n'est définie que lorsqu'on a une seule équation structurelle, ce qui limite aussi grandement les possibilités d'application. Cette approche a été améliorée par Jaccard et Wan (1995), Jöreskog et Yang (1996), et Ping (1996).

Bollen (1995) introduit une méthode plus générale s'appliquant au modèle LISREL classique, les équations du modèle structurel deviennent :

$$\boldsymbol{\xi} = \mathbf{B}_1\boldsymbol{\xi} + \mathbf{B}_2 f(\boldsymbol{\xi}) + \boldsymbol{\zeta} \tag{5.1}$$

avec $\mathbf{B}_1$ matrice de coefficients traduisant les effets des variables latentes endogènes sur les autres variables latentes, $f(.)$ fonction non linéaire et $\mathbf{B}_2$ matrice de coefficients traduisant les effets de $f(\boldsymbol{\xi})$ sur $\boldsymbol{\xi}$.
Pour le modèle de mesure, on a :

$$\boldsymbol{x} = \boldsymbol{\Lambda}_1\boldsymbol{\xi} + \boldsymbol{\Lambda}_2 f(\boldsymbol{\xi}) + \boldsymbol{\epsilon} \tag{5.2}$$

avec $\boldsymbol{\Lambda}_1$ et $\boldsymbol{\Lambda}_2$ matrices de coefficients associées à $\boldsymbol{\xi}$ et $f(\boldsymbol{\xi})$.
Ce modèle est une généralisation du modèle LISREL classique avec la possibilité d'introduire de la non linéarité. Bollen (1995) utilise une méthode 2SLS (*two-stage least squares*) afin d'estimer ce modèle en fixant le *loading* associé à la première variable de chaque bloc à 1. Nous ne développons pas ici la méthode d'estimation, on peut voir Bollen (1995). Cette méthode, bien que prometteuse, n'a pas fait

l'objet de nombreuses recherches et n'est plus utilisée.

Deux approches dominent actuellement, celle de Lee et Zhu (2002) basée sur le maximum de vraisemblance et celle de Arminger et Muthén (1998) dite bayésienne. Ces deux méthodes sont basées sur des données normales multivariées.
Dans les travaux de Lee et Zhu (2002), ce sont les relations entre les variables latentes qui sont traitées. La non linéarité y est prise en compte mais la fonction $f(.)$ associée à chaque variable latente doit être connue. Le modèle externe reste le même mais les équations du modèle interne diffèrent :

$$\boldsymbol{\eta}_i = \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i \tag{5.3}$$

où $\mathbf{F}(\boldsymbol{\xi}_i) = (f_1(\boldsymbol{\xi}_i), ..., f_r(\boldsymbol{\xi}_i))$ est un vecteur de fonctions différentiables indépendantes. L'estimation par maximum de vraisemblance se fait classiquement avec des adaptations pour la nouvelle équation interne.

Arminger et Muthén (1998) se réfèrent au même modèle que Lee et Zhu (2002) mais changent la méthode d'estimation, ils n'utilisent plus du maximum de vraisemblance, mais des statistiques bayésiennes. Ainsi, ils utilisent des méthodes de MCMC (*Markov Chain Monte Carlo*). Parmi celles-ci, le Gibbs sampler et l'algorithme de Metropolis-Hasting servent à estimer la distribution des paramètres du modèle connaissant les données. En fixant des distributions associées à l'ensemble des paramètres du modèle, les auteurs arrivent à obtenir une distribution pour chacune des matrices à estimer, ainsi la matrice de covariance du modèle connaissant la matrice de covariance empirique suit une loi inverse de Wishart. Les paramètres de ces distributions sont estimés par des MCMC. Cette méthode est développée dans Arminger et Muthén (1998) associée à quelques simulations afin de vérifier son efficacité. Néanmoins, elle repose sur des hypothèses de distributions fortes qui ne sont pas forcément remplies dans des applications réelles. De plus, les fonctions (produits de variables ou polynômes) doivent être prédéterminées.

Sur les méthodes basées sur le maximum de vraisemblance, dans sa thèse, Meijerink (1995) applique des transformations non linéaires aux variables manifestes avant de les intégrer dans le modèle. Ainsi, au lieu d'avoir simplement des variables manifestes classiques, celles-ci seront transformées afin d'optimiser leur pouvoir explicatif. Les estimations des paramètres des fonctions non linéaires et de ceux du modèle structurel pourront se faire, soit pas à pas, soit simultanément par le maximum de vraisemblance. Dans son ouvrage, il utilise les fonctions de splines afin de transformer des données non normales en données normales multivariées. Les théories de l'*optimal scaling* ont inspiré nos recherches et nous utilisons par la suite certains principes mis en place par Meijerink (1995).

Ces travaux sont intéressants mais n'ont pas vu beaucoup d'applications dans le domaine des modèles d'équations structurelles à variables latentes. Ceci peut être expliqué par le fait que, parmi les logiciels commerciaux, seul M-Plus (Muthén et Muthén, 1998) intègre la prise en compte de non linéarités dans le modèle. D'autre part, l'utilisation de ces méthodes pose un problème d'interprétation des variables ou des relations transformées.

5.2.2 Le cas PLS

Dans le cadre des méthodes PLS, les travaux sur la non linéarité ont surtout concerné la régression PLS. Les travaux sur l'approche PLS sont encore balbutiants. Wold (1982), dès ses premiers articles, insiste sur le fait que les variables manifestes peuvent être transformées sans modifier l'algorithme PLS. Il propose des transformations du type : $\mathbf{x}^2, \log \mathbf{x}, \mathbf{x} \cdot \mathbf{y}, \mathbf{x}/\mathbf{y}$. Ces transformations pourront être utiles mais l'existence de non linéarité sera difficile à déceler.

Krämer (2005) a présenté un algorithme pour inclure de la non linéarité entre les variables manifestes et latentes. L'auteur se sert de l'astuce du noyau (*kernel trick*) et du mode B d'estimation du modèle. L'approche PLS dans le cadre de l'application du mode B peut être modélisée par un problème

d'optimisation sous contraintes, on aura alors :

$$\begin{array}{l} \max \sum_{k,l=1}^{K} c_{kl} cov^2(\mathbf{X}_k \mathbf{w}_k, \mathbf{X}_l \mathbf{w}_l) \\ s.c. \parallel \mathbf{X}_k \mathbf{w}_k \parallel = 1 \, \forall k \end{array} \tag{5.4}$$

On cherche donc les $\mathbf{w}_k$ tel que l'estimation du score de la variable latente $\boldsymbol{\xi}_k$ s'écrive :

$$\mathbf{y}_k = \mathbf{X}_k \mathbf{w}_k \tag{5.5}$$

On devra alors trouver la solution optimale par le biais du Lagrangien. Afin de passer au cas non linéaire, on va transformer les données en utilisant une fonction noyau. Pour cela, on suppose que n'importe quelle solution du problème d'optimisation précédent peut s'écrire :

$$\mathbf{w}_k = \mathbf{X}_k^T \boldsymbol{\alpha}_k, \boldsymbol{\alpha}_k \in R^n$$

L'algorithme PLS peut alors être réécrit en fonction du vecteur $\boldsymbol{\alpha}_k$ et de la matrice $\mathbf{X}_k \mathbf{X}_k^T$. Cette matrice est appelée matrice de Gram et représente les produits internes entre les observations. L'idée de la transformation sera de remplacer les matrices $\mathbf{X}_k \mathbf{X}_k^T$ dans l'algorithme par une fonction noyau :

$$k(\mathbf{X}, \mathbf{Z}) = < \Phi(\mathbf{X}), \Phi(\mathbf{Z}) >$$

Nous savons que ces fonctions sont efficaces et permettent de ne pas se limiter au cas linéaire. Les $\boldsymbol{\alpha}_k$ estimés seront des coefficients associés à la matrice produit transposée dans un espace de très grande dimension (voire infini). Cette approche paraît intéressante, les paramètres du noyau sont estimés par validation croisée. Cependant, elle comporte un défaut quasi-insurmontable, l'interprétation des résultats. En considérant qu'on arrive à revenir dans le cadre du poids externe $\mathbf{w}_k$ alors celui-ci explique le bloc $\Phi(\mathbf{X}_k)$ où Φ est une fonction inconnue. Cette approche posera des problèmes trop complexes pour être utilisée "telle quelle". Dans la pratique, malgré l'apparition de relations fortes entre les variables latentes, nous aurons besoin de connaître précisément la façon dont les variables latentes sont construites. C'est pour cette raison que cette approche devra être abandonnée ou du moins simplifiée.

Betzin et Henseler (2005) proposent de traiter des données catégorielles en utilisant des transformations de variables issues des théories de l'*optimal scaling*. Ils utilisent une quantification des variables manifestes en appliquant l'algorithme des moindres carrés alternés. De la même façon que Krämer (2005), ils travaillent avec le mode B afin de pouvoir utiliser la fonction globale à optimiser. Il est simple de remplacer la fonction de quantification par une fonction non linéaire. Les méthodes que nous développons sont basées sur ce principe sans se limiter au mode B. Ceci posera le même problème d'optimisation que dans le cadre du mode A.

Par ailleurs, la régression PLS a été adaptée au cas non linéaire par de nombreux chercheurs (Wold et al., 1989; Wold, 1992). On peut voir la thèse de Vivien (2002, partie II, chap. 1) rassemblant l'ensemble de ces approches. Nous ne développons pas ces méthodes ici mais elles figurent comme des voies de recherches intéressantes car la régression PLS est un cas particulier de l'approche PLS. La recherche de généralisation de ces approches constitue une ouverture importante pour l'étude de non linéarités.

Comme on peut le voir, dans le cadre de l'approche PLS, la non linéarité n'est que peu prise en compte et il serait intéressant d'utiliser des méthodes statistiques afin de l'inclure dans le modèle.

5.3 Méthodes basées sur des transformations des variables

Nos travaux sont issus d'un principe simple : plutôt que de modifier les relations au sein du modèle, nous modifions les variables en les transformant. Ainsi les relations entre les variables transformées seront linéaires et les estimations pourront se faire classiquement.

Une fois ce principe admis, reste à le mettre en pratique. Dans la littérature, ce type de transformations est rassemblé sous le titre d'*optimal scaling*. Il faudra toutefois être plus prudent car les modèles structurels sont des modèles complexes dans lesquels de nombreuses interactions existent et dont l'interprétation n'est pas aisée. Il faut garder à l'esprit les trois conditions nécessaires à l'introduction de non linéarité données en 5.2. Il est donc important de conserver la possibilité d'interprétation des résultats, une relation non linéaire complexe ne pourra pas être transcrite dans une interprétation globale du modèle. Ces raisons nous poussent à choisir une famille de fonctions à la fois flexible et s'appliquant bien aux variables traitées. Ces fonctions devront être monotones afin de mettre en avant des tendances dans les évolutions des relations sans remettre en cause l'interprétation du modèle.

A partir de cette hypothèse de travail, nous partons de deux méthodes :

- Celle consistant à transformer les variables manifestes. Nous utilisons des fonctions monotones afin de rendre les variables manifestes le plus corrélées les unes aux autres. Ceci nous permettra de comprendre les mécanismes de rapprochement afin d'obtenir des construits plus cohérents. Ces méthodes sont basées sur les recherches de Kruskal et Shepard (1974), Young et al. (1978), Young (1981) et Winsberg et Ramsay (1983). L'étude des différences entre les modèles transformés et les modèles non transformés permet de mettre en valeur les avantages de chaque transformation.

- Celle consistant à transformer les scores des variables latentes. On maximise ainsi la qualité prédictive de la transformation d'une variable latente sur une variable cible. Nous travaillons avec des transformations optimales afin d'ajuster des régressions simples ou multiples (Young et al. (1976), Winsberg et Ramsay (1980) et Breiman et Friedman (1985)) au moyen de fonctions monotones. Une fois la transformation obtenue, elle devra être réinjectée dans l'estimation PLS des variables latentes.

Ces deux approches sont approfondies distinctement par la suite et donnent lieu à des applications dans le cadre du chapitre 7.7 (p. 165).

5.3.1 Le choix d'une famille de transformations

Les transformations non linéaires des données sont nombreuses et nous devons nous restreindre à certaines d'entre elles du fait de contraintes inhérentes aux modèles et aux données.

- Les modèles d'équations structurelles, avant d'être des modèles prédictifs (dans le cadre de PLS), sont des modèles confirmatoires. L'interprétation des paramètres estimés est donc primordiale. Afin de conserver la possibilité d'interprétation, nous utilisons des fonctions continues monotones sur un intervalle donné. On cherche donc $f(x, \theta)$ tel que f soit différentiable et :

$$\frac{df(x,\theta)}{dx} \geq 0$$

- Les données issues de questionnaires sont ordinales avec un grand nombre de modalités, elles sont donc basées sur un intervalle borné $[a, b]$. Il faut donc travailler avec des fonctions définies sur $[a, b]$.

Nous recherchons donc des fonctions flexibles. Les polynômes constituent des fonctions trop déterminées (leur comportement dans une région détermine leur comportement dans une autre). Ces

contraintes nous poussent à choisir une famille de fonctions largement développées et qui satisfait l'ensemble de ces conditions : **les B-splines monotones**. Afin de mieux comprendre ces fonctions, il est nécessaire de détailler leur mise en place.

Une spline est l'équivalent d'un polynôme par morceaux. De plus, une spline d'un degré donné et avec une séquence de nœuds connue, forme un espace linéaire. On peut voir sur le sujet De Boor (1978) et Schumacker (1981).
Une B-spline est une forme particulière de spline. Elle possède l'avantage d'être dans un espace simple dans lequel des bases de fonctions peuvent être trouvées facilement. Ainsi on définit :

Définition 5.1. *Pour un intervalle $[a, b]$, une suite de $m + 1$ points $a = z_0 < ... < z_m = b$ dans $[a, b]$. On appelle B-spline d'ordre $(k+1)$ ayant pour nœuds simples les points $z_1, ..., z_m$ toute fonction $B : [a, b] \longrightarrow \mathbb{R}$ telle que :*
- *B est continue sur $[a, b]$;*
- *B est non nulle sur un certain nombre $k < m$ de sous-intervalles de $[a, b]$. Ce nombre est égal au degré $(k + 1)$ de cette B-spline.*

Propriété 5.1. *Toute B-spline de degré donné et ayant une séquence de nœuds connue peut être réécrite comme une combinaison linéaire d'éléments de la base associée à cette B-spline.*

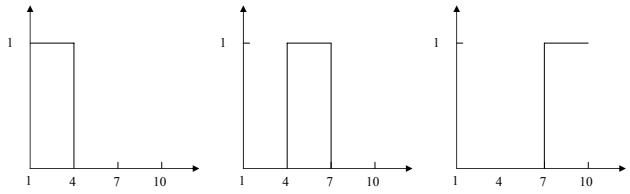

FIG. 5.1 – B-splines de degré 0 à 2 nœuds internes

La figure 5.1 illustre 3 exemples de B-splines de degré 0 à 2 nœuds intérieurs. Toute spline de l'espace des splines de degré 0 à 2 nœuds peut être écrite comme une combinaison linéaire de ces 3 splines qui forment la base des B-splines de degré 0 sur cette séquence de nœuds. Il est rare que l'on utilise des B-splines de degré 0 car elles ne sont pas différentiables. Ainsi, pour des B-splines de degrés plus grands, on pourra toujours former une base complète et indépendante dont les éléments sont notés B_l et sont au nombre de $m + k$. Ainsi, pour une spline de degré k définie sur $m - 1$ nœuds, on aura :

$$y = \sum_{l=1}^{m+k} \gamma_l B_l(x) \tag{5.6}$$

avec y variable transformée et $B_l(.)$ transformations de la base. De manière matricielle, on peut écrire $y = \mathbf{q}'\boldsymbol{\gamma}$ avec $\boldsymbol{\gamma}$ vecteur des paramètres et $\mathbf{q}$ vecteur composé des $q_l = B_l(x)$.
Pour un degré égal à 2 et 2 nœuds internes, on aura 5 éléments dans la base (voir figure 5.2) et dans le cas de la B-spline de la figure 5.3, l'équation sera donnée par :

$$y = \sum_{l=1}^{5} \gamma_l B_l(x) = 0.1\, B_1(x) + 10\, B_2(x) + 0.0\, B_3(x) + 1.0\, B_4(x) + 0.1\, B_5(x)$$

Cette B-spline n'est pas monotone, nous devrons donc ajouter des contraintes sur celle-ci.

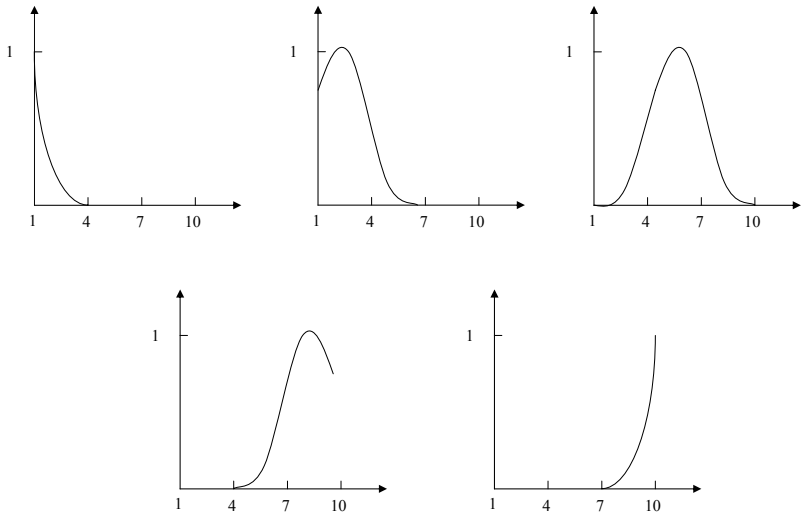

FIG. 5.2 – Base des B-splines de degré 2 à 2 nœuds internes

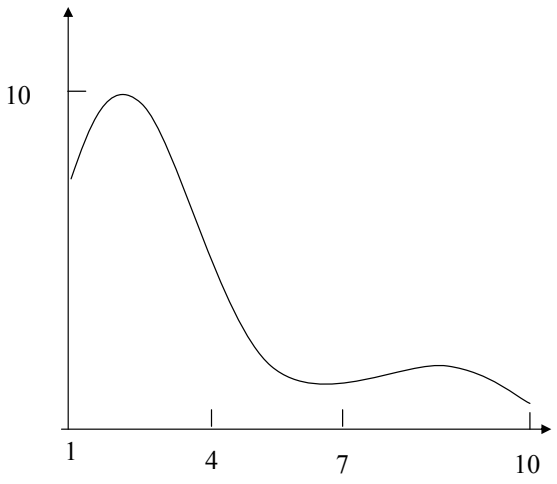

FIG. 5.3 – Exemple de B-spline de degré 2 à 2 nœuds internes

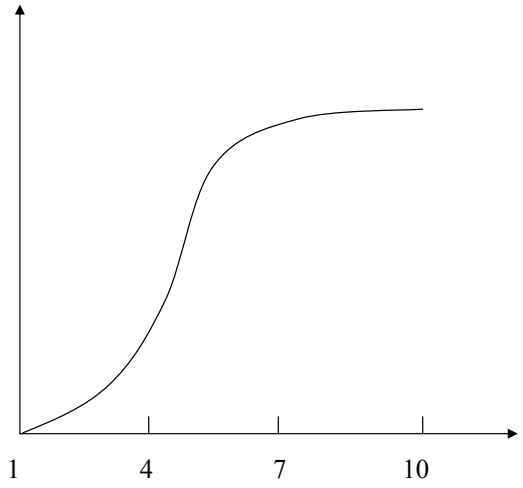

FIG. 5.4 – Exemple de B-spline monotone de degré 2 à 2 nœuds internes

Afin d'obtenir des B-splines monotones, il existe plusieurs techniques. La première consiste en l'utilisation d'intégrales des fonctions de la base des B-splines.

Définition 5.2. *On appelle* ***I-spline*** *de degré d à k nœuds, la fonction obtenue comme l'intégrale d'une B-spline de degré d à k nœuds. On a :*

$$I_{d,k}(x) = \int_a^x B_{d,k}(u)du$$

Propriété 5.2. *Les I-splines sont des fonctions différentiables monotones.*

En tant qu'intégrales de fonctions différentiables positives, les I-splines sont donc différentiables et monotones.
La transformation est alors basée sur une combinaison linéaire de la base des I-splines :

$$y = \sum \gamma_l I_l(x) \quad (\gamma_l > 0)$$

Ceci permet d'obtenir une fonction monotone très flexible (Winsberg et Ramsay, 1981; Ramsay, 1988).

Une autre manière d'obtenir des B-splines monotones est d'effectuer des combinaisons linéaires de fonctions issues de la base des B-splines de sorte que les paramètres γ_l soient dans un ordre croissant $\gamma_1 < \gamma_2 < ... < \gamma_{m+k}$. Cette condition est une condition nécessaire à l'obtention de B-splines monotones.
La figure 5.4 illustre le cas de l'équation :

$$y = 0.0\, B_1(x) + 0.1\, B_2(x) + 9.7\, B_3(x) + 9.8\, B_4(x) + 9.9\, B_5(x)$$

Nous pourrons aussi bien utiliser des I-splines que des B-splines monotones dans les applications. D'un point de vue intuitif, la seconde version permet une meilleure visualisation. Pour des raisons pratiques, nous continuons à utiliser le terme de B-spline monotone. Le type de construction ne change pas le principe des méthodes que nous présentons.

5.3.2 Les transformations par B-splines monotones

Dans l'ensemble de cette partie, nous utilisons des méthodes qui étendent les moindres carrés ordinaires en utilisant des transformations optimales des variables par le biais des moindres carrés alternés (Young, 1981).
Nous nous focalisons donc sur les transformations des variables du modèle par des B-splines monotones. Ces transformations nécessitent l'estimation de paramètres supplémentaires, on les notera $\boldsymbol{\gamma}$. Ainsi, dans le cadre des modèles d'équations structurelles, on aura deux transformations possibles :

- transformation du modèle externe, l'équation 1.1 (p. 20) s'écrit alors :

$$B(\mathbf{x}_{kj}, \boldsymbol{\gamma}) = \pi_{kj}\boldsymbol{\xi}_k + \boldsymbol{\epsilon}_{kj} \tag{5.7}$$

  où les inconnues sont π_{kj} et $\boldsymbol{\gamma}$. On veut que la spline B rende la relation entre $B(\mathbf{x}_{kj}, \boldsymbol{\gamma})$ et $\boldsymbol{\xi}_k$ linéaire ;
- transformation du modèle interne, l'équation 1.3 (p. 21) s'écrit alors :

$$\boldsymbol{\xi}_i = \sum_{k:\boldsymbol{\xi}_k \to \boldsymbol{\xi}_i} \beta_{ik} B^k(\boldsymbol{\xi}_k, \boldsymbol{\gamma}_k) + \boldsymbol{\zeta}_i \tag{5.8}$$

  où les inconnues sont les β_{ik} et les paramètres $\boldsymbol{\gamma}_k$ associés à la spline de la variable $\boldsymbol{\xi}_k$. De la même façon que pour la première transformation, on voudra obtenir des transformations qui rendent la relation entre les variables latentes linéaire.

L'estimation de l'ensemble des paramètres π, β et $\boldsymbol{\gamma}$ ne peut pas se faire directement par les moindres carrés classiques en minimisant :

$$\|B(\mathbf{x}_{kj}, \boldsymbol{\gamma}) - \pi_{kj}\boldsymbol{\xi}_k\|^2 \quad ou \quad \|\boldsymbol{\xi}_i - \sum_{k:\boldsymbol{\xi}_k \to \boldsymbol{\xi}_i} \beta_{ik} B^k(\boldsymbol{\xi}_k, \boldsymbol{\gamma}_k)\|^2 \tag{5.9}$$

Le nombre de paramètres étant trop grand, on se base donc sur un algorithme itératif. Il se divise en deux étapes :

- recherche des estimations par moindres carrés des paramètres du modèle (π ou β) avec les valeurs associées aux paramètres des transformations fixées ($\boldsymbol{\gamma}$)
- recherche des estimations par moindres carrés des paramètres liés aux transformations des données ($\boldsymbol{\gamma}$) en laissant les paramètres du modèle fixes (π ou β).

Cet algorithme, appelé moindres carrés alternés, converge (Young, 1981) et il est appliqué dans le cadre de la linéarisation d'une régression (Young et al., 1976; Breiman et Friedman, 1985; De Leeuw, 1988) ou dans celui de l'ACP non linéaire (Young et al., 1978). Il est fréquemment utilisé afin de traiter des données catégorielles (Young et al., 1976; Tenenhaus, 1977, 1979).

Nous supposons dans l'ensemble de ce chapitre que le degré et les nœuds associés aux B-splines monotones sont connus. Il est possible d'estimer ces paramètres supplémentaires par le biais de méthodes de classification hiérarchique en fonction d'un critère global. On part d'un nombre de nœuds et de degrés faibles et en augmentant ces paramètres, on vérifie l'augmentation de la qualité d'explication associée aux données transformées. On utilisera souvent les théories sous-jacentes au domaine d'application. Celles-ci nous aiguilleront vers un choix adapté ou vers un ensemble de valeurs plus réduit.

5.4 Transformations et approche PLS

L'application des transformations dans le cadre de modèles de régression classique est fréquente, nous tentons de les intégrer afin d'augmenter le pouvoir explicatif des modèles d'équations structurelles à variables latentes. Nous nous basons sur le cas de l'approche PLS et spécifiquement sur le mode A qui est aujourd'hui le plus utilisé. Ces recherches sont motivées par la proximité entre les recherches sur l'approche par moindres carrés partiels (PLS) et celles sur la méthode par moindres carrés alternés (ALS). Nous n'avons pas ici la prétention de résoudre les problèmes liés à l'approche PLS (absence de fonction globale à optimiser) mais nous tentons d'introduire dans les équations structurelles des transformations non linéaires. Des chercheurs ont tenté de minimiser les désavantages de l'approche PLS en utilisant l'algorithme ALS, par exemple, la méthode PATHALS (Coolen et De Leeuw (1987), Coolen et De Leeuw (1988), De Leeuw (1987)) ou l'approche GSCA de Hwang et Takane (2004) (voir chap. 2.4.2, p. 53).
Nous introduisons deux types de transformations : l'une au niveau des variables manifestes dans le modèle externe par une méthode à deux pas, l'autre au niveau du calcul des scores des variables latentes dans le modèle structurel. Pour chacune, nous présentons, dans le chapitre 7.5 (p. 153), des applications dans le cadre de l'analyse des non linéarités dans les relations associées à la satisfaction.

5.4.1 Au niveau du modèle de mesure

L'utilisation d'une transformation des variables manifestes dans le cadre de l'approche PLS est problématique. On ne peut pas utiliser l'équation 5.7 car le score de la variable latente n'est pas connu initialement et il est obtenu à partir d'une combinaison linéaire des colonnes de $\mathbf{X}_k$. Cette transformation sera donc inefficace du fait de la linéarité de la relation (par construction). Nous nous focalisons donc sur une transformation de l'ensemble des variables manifestes associées à une variable latente simultanément (en optimisant une seule fonction globale), on aura $\mathbf{B}_k(\mathbf{X}_k, \boldsymbol{\gamma}_k) = \{B_{ki}(\mathbf{x}_{ki}, \boldsymbol{\gamma}_{ki}) | i = 1, ..., p_k\}$. Dans le cadre de l'algorithme PLS, ceci revient à utiliser lors de l'étape d'estimation externe des scores des variables latentes, l'estimation :

$$\mathbf{y}_k = \mathbf{B}_k(\mathbf{X}_k, \boldsymbol{\gamma}_k)\mathbf{w}_k$$

Le problème qui se pose réside dans l'estimation des paramètres $\boldsymbol{\gamma}_k$ de $\mathbf{B}_k(\cdot, \cdot)$ qui ne peuvent pas être estimés directement par moindres carrés car $\mathbf{w}_k$ et $\mathbf{y}_k$ sont fixés en fonction de $\mathbf{B}_k(\mathbf{X}_k, \boldsymbol{\gamma}_k)$.

Nous supposons lors de cette étape que $\mathbf{y}_k$ est la première composante principale issue du bloc des variables transformées. On applique l'approche PLS mode A à laquelle on ajoute une étape préalable de transformation des données basée sur un critère choisi.

Ce type d'approche se justifie dans le cadre de l'approche PLS par la proximité entre les scores finaux des variables latentes et les premières composantes principales associées à chaque bloc de variables latentes. Dans le cadre du mode A, lorsqu'on a un seul bloc, l'approche PLS revient à une ACP. De plus, Dijkstra (1981) montre que le mode A est une extension de l'ACP et que les scores des variables latentes sont des combinaisons linéaires des indicatrices avec des poids proportionnels à la covariance entre les indicatrices et la variable latente.

Les nombreuses applications effectuées tendent à valider cette proximité, nous présentons des simulations et des comparaisons empiriques afin de vérifier cette hypothèse. Nous simulons un modèle structurel en utilisant le procédé du chapitre 2.3 (p. 46 et utilisons un jeu de données réelles issu de la littérature basé sur une analyse de la satisfaction de clients d'opérateurs de téléphones mobiles (Tenenhaus et al., 2005)). Ces applications s'appuient sur le modèle ECSI (cf. fig. 7.4, p. 138). Nous illustrons la comparaison des scores PLS et des premières composantes principales dans le cadre d'une variable exogène (l'image) et d'une variable endogène (la satisfaction). Les résultats sont illustrés dans les figures 5.5 (cas simulé) et 5.6 (cas réel). Dans la figure 5.6, les corrélations obtenues apparaissent.

Nous voyons que les résultats sont très proches, tout spécialement pour le cas exogène et pour le modèle simulé.

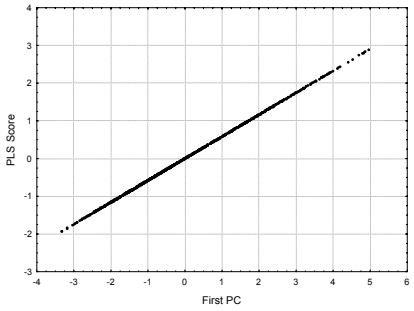

FIG. 5.5 – Comparaison des scores PLS et de la première composante de l'ACP pour des données simulées sur une variable endogène

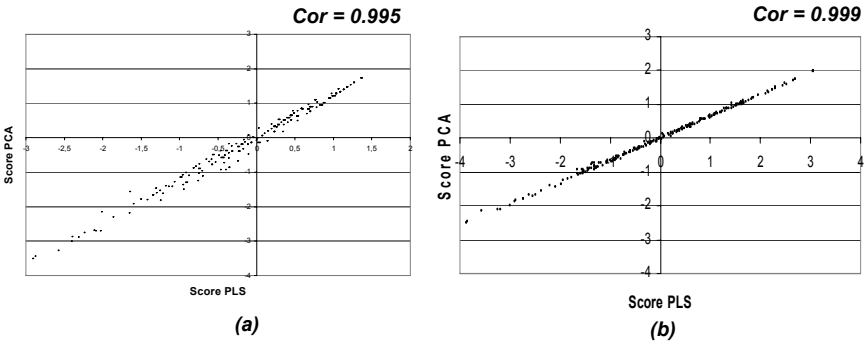

FIG. 5.6 – Comparaison des scores PLS et de la première composante de l'ACP pour des données réelles sur une variable (a) endogène et (b) exogène

Cette petite étude nous montre que le modèle interne a peu d'impact sur le calcul du score des variables latentes dans un cas classique. Des études par simulations plus poussées seraient nécessaires afin de comprendre dans quels cas cette hypothèse s'avère fausse.

Nous cherchons donc à obtenir des transformations qui maximisent le pourcentage de variance expliquée par la première composante principale associée au bloc étudié en utilisant des B-splines monotones.

L'ACP est une méthode linéaire au sens où les composantes principales sont des combinaisons linéaires des variables observées et parce qu'elle est basée sur les coefficients de corrélation linéaire. Si les relations ne sont pas linéaires, on peut transformer les variables de façon à se rapprocher de la linéarité. Or, lorsqu'on tente de transformer des données afin de rendre leur relation linéaire, aucune transformation ne pourra améliorer la relation entre ces variables si celles-ci sont normales. Le théorème suivant peut s'appliquer (Saporta, 2006) :

Théorème 5.1. *Si* $(\mathbf{X}, \mathbf{Y})$ *est un couple de variables aléatoires gaussiennes, on ne peut pas trouver de transformations* $\varphi(\mathbf{X})$, $\phi(\mathbf{Y})$ *augmentant en valeur absolue le coefficient de corrélation :*

$$cor^2(\varphi(\mathbf{X}), \phi(\mathbf{Y})) \leq cor^2(\mathbf{X}, \mathbf{Y})$$

Pour obtenir les corrélations maximales, on cherchera donc à rendre normales les variables observées. Dans le cadre de l'ACP non linéaire, on va chercher à transformer l'ensemble des variables d'un bloc afin que le pourcentage d'inertie expliqué par le premier axe de l'ACP soit maximal. Ceci revient à trouver la transformation $B(\cdot, \boldsymbol{\gamma})$ qui maximise :

$$\max_{B(\cdot,\boldsymbol{\gamma})} \lambda_1$$

avec λ_1 plus grande valeur propre de la matrice $B(\mathbf{X})'B(\mathbf{X})$. Ceci revient à maximiser la variance expliquée par la première composante principale.

Cette méthode est basée sur un algorithme itératif avec estimation des paramètres par les moindres carrés alternés. A chaque itération, l'algorithme alterne entre une ACP classique (Hotelling, 1933) et de l'*optimal scaling* (Young, 1981). Quand toutes les variables sont ordinales, ceci correspond à MDPREF de Carroll (1972).

D'un point de vue pratique, on effectue donc une ACP classique sur le tableau des données augmenté (les données et les transformations dans la base des B-splines monotones). On aura N lignes et $P(\sharp(\text{degrés}) + \sharp(\text{noeuds}) + 1)$ colonnes.

Une fois les transformations estimées, l'approche PLS mode A classique peut être appliquée. Le processus complet est décrit dans la figure 5.7.

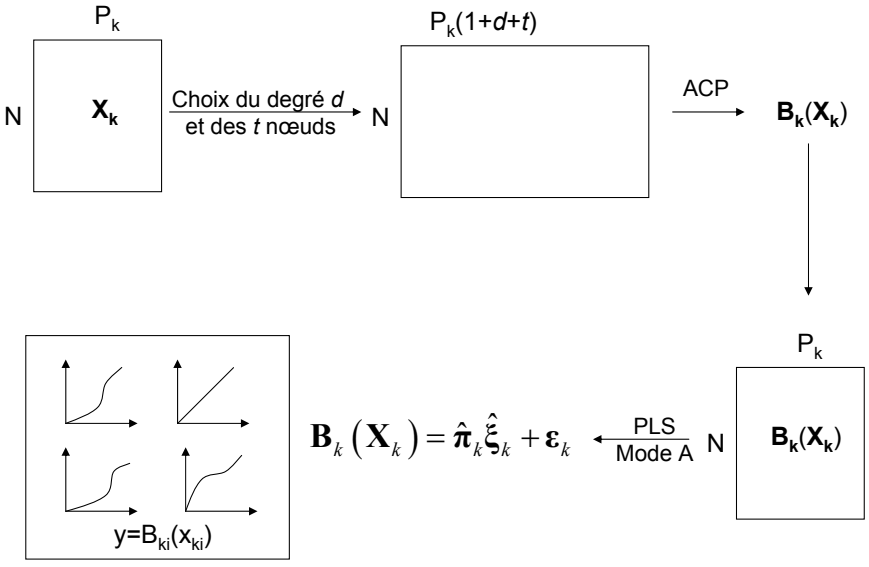

FIG. 5.7 – Processus pour l'application de l'approche PLS avec transformation non linéaire du modèle externe

Cette méthode permet donc de maximiser la qualité d'explication du modèle externe. La communauté de chaque variable latente en sera améliorée. Nous illustrons cette approche par une application sur des données avec visualisation de l'ensemble des transformations dans le cadre du chapitre 7.6 (p. 157).

5.4.2 Au niveau du modèle interne

L'approche que nous avons présentée dans la partie précédente sert à augmenter la qualité prédictive du modèle externe. Les problèmes y étant associés ne se posent pas dans le cadre du modèle interne, les variables latentes sont estimées indépendamment et on pourra donc travailler directement sur l'équation 5.8 (p. 105).

Nous introduisons donc une modification de l'approche PLS mode A afin de prendre en compte des relations non linéaires dans le modèle structurel. Dans cette optique, nous utilisons des transformations optimales des scores des variables latentes par le biais de B-splines monotones.

Les transformations de régressions

L'équation 5.8 (p. 105) est une équation de régression multiple entre une variable latente expliquée et des variables latentes explicatives transformées, à condition de remplacer les variables latentes par leurs scores, on aura :

$$\hat{\boldsymbol{\xi}}_i = \sum_{k:\boldsymbol{\xi}_k \to \boldsymbol{\xi}_i} \beta_{ik} B_k(\hat{\boldsymbol{\xi}}_k, \boldsymbol{\gamma}_k) + \boldsymbol{\zeta}_i \tag{5.10}$$

On va tenter pour une variable $\boldsymbol{\xi}_i$ donnée d'obtenir les estimations des $\boldsymbol{\gamma}_k$ tel que le critère de moindres carrés suivant soit minimisé :

$$\|\hat{\boldsymbol{\xi}}_i - \sum_{k:\boldsymbol{\xi}_k \to \boldsymbol{\xi}_i} \beta_{ik} B_k(\hat{\boldsymbol{\xi}}_k, \boldsymbol{\gamma}_k)\|^2 \tag{5.11}$$

L'estimation de l'ensemble des paramètres de cette équation se fait par alternance : on estime donc β_{ik} en fixant $\boldsymbol{\gamma}_k$ puis $\boldsymbol{\gamma}_k$ en fixant β_{ik}. On utilise donc les moindres carrés alternés (Young, 1981). Après convergence, on obtient une estimation des β_{ik} et des $\boldsymbol{\gamma}_k$. Les paramètres associés aux B-splines monotones (nombre de degrés et séquence de nœuds) doivent être définis à l'avance.
Les variables transformées ainsi obtenues ont des relations linéaires avec leurs variables expliquées. Dans le cadre de cette approche, nous utilisons donc un algorithme basé sur la maximisation du R^2 d'une variable dite cible (Young et al., 1976).

Transformation du modèle interne

Un certain nombre d'hypothèses sont nécessaires afin d'appliquer ce type de méthodes aux modèles d'équations structurelles à variables latentes et spécifiquement à l'approche PLS. Une variable latente endogène $\boldsymbol{\xi}_K$ dite cible doit être choisie, on prendra généralement la variable ayant le plus de variables explicatives $\boldsymbol{\xi}_j$. Dans le cadre de l'analyse de la fidélité, on prendra ce concept. Pour la satisfaction, on prendra la variable latente satisfaction. Les variables explicatives devront être transformées. D'autre part, une nouvelle étape dans la construction des variables latentes doit être ajoutée pour l'approche PLS mode A.

L'intégration de transformations de variables par le biais des théories de l'*optimal scaling* dans l'approche PLS est motivée par la proximité existant entre ces deux concepts. L'estimation des transformations est basée sur un algorithme de moindres carrés alternés alors que l'approche PLS est basée sur l'alternance de deux critères de moindres carrés. Il est donc cohérent d'intégrer dans l'estimation interne un critère de moindres carrés.

Le modèle structurel complet obtenu est le suivant :

$$\begin{aligned} \mathbf{x}_{lm} &= \pi_{lm}\tilde{\boldsymbol{\xi}}_l + \boldsymbol{\epsilon}_{lm} \\ \tilde{\boldsymbol{\xi}}_l &= \textstyle\sum_i \delta_{li}\beta_{li}\tilde{\boldsymbol{\xi}}_i + \boldsymbol{\zeta}_l \end{aligned} \tag{5.12}$$

où $\mathbf{x}_{lm}$ est une variable manifeste associée à la variable latente $\boldsymbol{\xi}_l$.

$$\delta_{ij} = \begin{cases} 1 & \text{si } \boldsymbol{\xi}_j \rightarrow \boldsymbol{\xi}_i \\ 0 & \text{sinon} \end{cases}$$

$$\tilde{\boldsymbol{\xi}_j} = \begin{cases} \phi_j(\boldsymbol{\xi}_j) & \text{si } \boldsymbol{\xi}_j \rightarrow \boldsymbol{\xi}_K \\ \boldsymbol{\xi}_j & \text{sinon} \end{cases}$$

avec $\phi_j(\boldsymbol{\xi}_j)$ transformation "optimale" de $\boldsymbol{\xi}_j$ en fonction de la variable cible $\boldsymbol{\xi}_K$.

Notre but est de transformer l'ensemble des variables latentes expliquant la variable latente cible $\boldsymbol{\xi}_K$ de façon à rendre les relations entre $\boldsymbol{\xi}_j$ transformées et $\boldsymbol{\xi}_K$ linéaires. Les questions sur les interactions, la convergence et l'effet sur les scores des autres variables latentes sont traitées par la suite.

Soit $\mathbf{y}_K = \mathbf{X}_K \mathbf{w}_K$ l'estimation externe du score de la variable latente cible $\boldsymbol{\xi}_K$. Soient $\mathbf{y}_j = \mathbf{X}_j \mathbf{w}_j$ les estimations des scores des variables latentes $\boldsymbol{\xi}_j$ expliquant la variable $\boldsymbol{\xi}_K$ dans le modèle structurel, on définit :

$$\mathbf{Y}_J = \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & ... & \mathbf{y}_R \end{pmatrix} \tag{5.13}$$

avec $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, ..., \boldsymbol{\xi}_R$ sont des antécédents directs de $\boldsymbol{\xi}_K$ dans le modèle structurel. On recherche une transformation spline monotone $B_j(.)$ de $\mathbf{y}_j$ tel que le carré des corrélations multiples de $\mathbf{y}_K$ soit maximisé. Soit $\tilde{\mathbf{Y}}_J = B_J(\mathbf{Y}_J)$, on cherche à maximiser :

$$\max_{B_j} R^2 = \max_{B_j} \mathbf{y}'_K \tilde{\mathbf{Y}}_J (\tilde{\mathbf{Y}}'_J \tilde{\mathbf{Y}}_J)^{-1} \tilde{\mathbf{Y}}'_J \mathbf{y}_K \tag{5.14}$$

Ce qui revient à minimiser :

$$\lambda^2 = (\tilde{\mathbf{Y}}_J \boldsymbol{\alpha} - \mathbf{y}_K)'(\tilde{\mathbf{Y}}_J \boldsymbol{\alpha} - \mathbf{y}_K) \tag{5.15}$$

avec $\boldsymbol{\alpha}$ est le vecteur des coefficients de régression multiple entre les $\mathbf{y}_j$ et $\mathbf{y}_K$. Nous supposons que toutes les estimations $\mathbf{y}_i$ sont standardisées à chaque étape.

L'estimation des paramètres $\boldsymbol{\gamma}_j$ de la transformation et du vecteur $\boldsymbol{\alpha}$ se fait par moindres carrés alternés. Nous présentons les étapes de l'algorithme PLS ainsi obtenu :

(1) Initialisation des poids externes $\mathbf{w}_i$, choix de la variable latente cible $\boldsymbol{\xi}_K$;

(2) Estimation externe de $\boldsymbol{\xi}_i$: $\mathbf{y}_i = \mathbf{X}_i \mathbf{w}_i$.

(3) Transformation optimale des variables latentes par des B-splines monotones :
- Pour chaque variable latente $\boldsymbol{\xi}_j$ expliquant la variable latente endogène cible $\boldsymbol{\xi}_K$:
  1. Estimation des paramètres γ de la transformation par B-spline monotone des variables latentes expliquant $\mathbf{y}_K$ par moindres carrés alternés en maximisant le $R^2_{\xi_K}$.
  2. Calcul de : $\tilde{\mathbf{y}}_j = B_j(\mathbf{y}_j)$.
- Pour les variables latentes $\boldsymbol{\xi}_l$ non connectées à $\boldsymbol{\xi}_K$ et pour $\boldsymbol{\xi}_K$, $\tilde{\mathbf{y}}_l = \mathbf{y}_l$.

(4) Estimation des poids internes : $e_{im} = sgn(\tilde{\mathbf{y}}'_i \tilde{\mathbf{y}}_m)$.

(5) Estimation interne : $\mathbf{z}_i = \sum_{m:\boldsymbol{\xi}_m \leftrightarrow \boldsymbol{\xi}_i} e_{im} \tilde{\mathbf{y}}_m$.

(6) Mise à jour des poids externes par le mode A : $\mathbf{w}_i = \frac{1}{\mathbf{z}'_i \mathbf{z}_i} \mathbf{X}'_i \mathbf{z}_i$.

(7) Répéter (2) à (6) jusqu'à convergence.

(8) Estimation des coefficients du modèle externe par régression OLS entre $\mathbf{x}_{lm}$ et $\tilde{\mathbf{y}}_l$, estimation des coefficients structurels par régressions multiples entre les $\tilde{\mathbf{y}}_l$.

Quelques observations peuvent être ajoutées. Pour comprendre la signification des coefficients estimés, il faut qu'ils soient analysés en présence des transformations obtenues. La visualisation de celles-ci offre un outil complémentaire important dans l'analyse d'un modèle structurel. Cependant, tant qu'on utilise des transformations monotones, les paramètres peuvent être interprétés sans prendre en compte la transformation.

La convergence de cet algorithme constitue un point important. Il s'intègre dans les problèmes de convergence liés au mode A de l'approche PLS. L'algorithme ALS intégré converge et les applications que nous avons effectuées tendent à valider cette hypothèse. La réponse à ce problème reste en suspens tant que les problèmes de convergence de l'approche PLS ne sont pas résolus.

Cette approche permet de renforcer l'importance du modèle interne dans l'estimation du modèle par l'approche PLS mode A car les scores des variables latentes sont transformés en fonction du modèle interne.

L'étude des interactions reste à approfondir. L'optimalité des transformations est obtenue en fonction de la variable cible, ceci pénalise la qualité prédictive associée aux autres relations du modèle et entraîne des diminutions sur certains indices globaux. Des études empiriques poussées doivent être effectuées afin de mieux comprendre ces aspects importants.

A chaque étape de l'approche PLS, nous estimons donc une nouvelle transformation qui a un impact sur l'estimation interne des scores des variables latentes et donc sur la mise à jour des poids externes. Il s'avère que cette transformation ne varie que très peu d'une itération à l'autre de l'approche PLS. Ceci s'explique par la rapide convergence de la méthode et par le fait que les poids externes arrivent très rapidement à une valeur constante (souvent dès la seconde itération). La transformation ne semble pas entraîner de modifications importantes dans les estimations.

Nous présentons une étude de simulation de cette approche afin de vérifier certaines propriétés car une étude analytique serait trop complexe.

Données simulées

Nous simulons donc des données en utilisant le modèle ECSI (fig. 7.4, p. 138). La relation entre image et fidélité est non linéaire et suit une fonction quadratique par morceaux avec deux nœuds internes.
Nous utilisons donc notre algorithme en fixant le degré égal à 2 et les nœuds prédéfinis. Nous illustrons la recherche du degré adéquat et du nombre de nœuds adapté dans le cadre des applications du chapitre 7.6. Nous rassemblons les résultats dans la figure 5.8 et le tableau 5.1. La figure 5.8 montre une bonne identification de la transformation par notre approche. Les résultats associés au R^2 de la fidélité et au F^2 sont légèrement améliorés et le GoF reste constant. Ce dernier résultat peut être expliqué par le fait que cette méthode favorise la qualité prédictive de la variable cible au détriment des autres variables latentes endogènes du modèle.
Les estimations des coefficients structurels nous montrent que le biais obtenu est plus faible pour l'approche non linéaire. La transformation réduit le biais en rendant la relation linéaire.

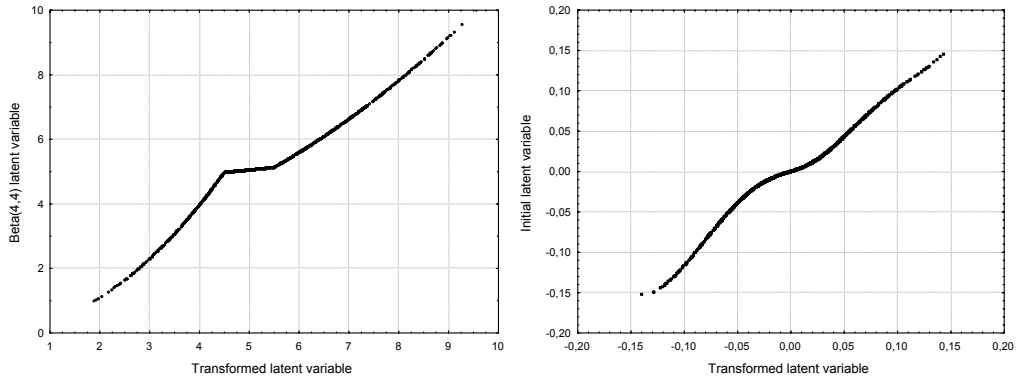

FIG. 5.8 – Transformation simulée et estimée de la variable latente satisfaction

Méthode	PLS non linéaire	PLS
R^2_{fid}	0.816	0.810
H^2_{fid}	0.869	0.869
F^2_{fid}	0.715	0.704
Coef. struct. (Estimé/Simulé)	0.704/0.8	0.676/0.8
GoF	0.868	0.865

TAB. 5.1 – Comparaison entre la méthode PLS non linéaire et la méthode classique sur des données simulées

Ce résultat révèle que les transformations sont bien estimées mais que les indices de qualité prédictive sont peu améliorés. Ceci peut être expliqué par deux causes :

- L'approche PLS mode A favorise le modèle externe, comme nous l'avons vu précédemment par la proximité entre scores PLS et ACP. Pour voir en quoi les estimations changent, nous comparons sur des données réelles et simulées les principaux indices ainsi que les scores des variables latentes lorsque la relation satisfaction - fidélité existe et lorsqu'elle n'existe pas. Les résultats sont présentés dans le tableau 5.2 et la figure 5.9. La proximité est ici marquante.
- Le fait d'utiliser une transformation proche du cas linéaire dans le processus de simulation entraîne une faible modification des indices et des coefficients. L'utilisation d'une transformation monotone plus marquée pourrait être intéressante. Cependant, le chercheur devra toujours utiliser des fonctions monotones afin de garder la possibilité d'interprétation des coefficients du modèle estimé.

Les résultats empiriques dans le cadre du mode A montrent une amélioration sensible de la qualité globale et soulignent surtout le faible impact du modèle interne dans le calcul des scores associés aux variables latentes.

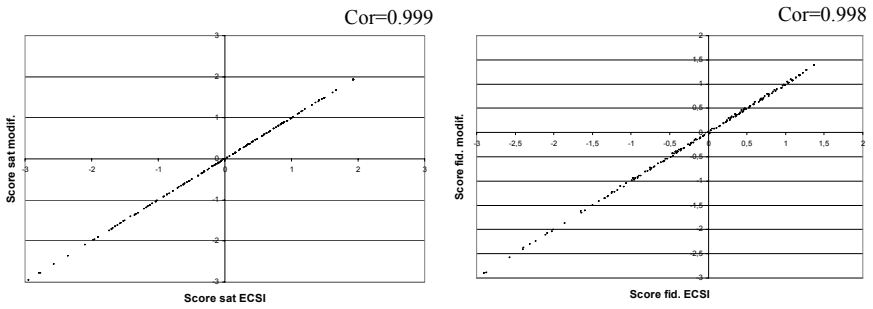

FIG. 5.9 – Comparaison des scores PLS avec ou sans la relation entre satisfaction et fidélité pour les variables satisfaction et fidélité

Méthode	avec relation	sans relation
R^2_{fid}	0.957	0.947
H^2_{fid}	0.973	0.973
F^2_{fid}	0.931	0.914
GoF	0.942	0.940

TAB. 5.2 – Comparaison des indices associés à la méthode PLS lorsque la relation satisfaction - fidélité est présente ou absente

5.5 Conclusion

Jusqu'ici, tous les travaux sur la non linéarité dans le cadre de l'approche PLS ont traité de la combinaison de PLS avec une autre méthode sans automatisation, on peut voir Pagès et Tenenhaus (2001) pour l'utilisation d'une AFM préalablement à l'approche PLS, mais dans une approche non intégrée.
Néanmoins, dans le cadre de la méthode LISREL, une méthode à deux pas proche de celle que nous présentons a été utilisée fréquemment et défendue par De Leeuw (1988).

Les deux méthodes introduites permettent d'estimer la présence de non linéarité dans le cas de l'approche PLS avec des indicateurs réflectifs (mode A) et, en cas de relations non linéaires, d'estimer les paramètres de fonctions B-splines monotones adaptées. Les relations du modèle peuvent être interprétées ainsi que les transformations obtenues.

Des travaux supplémentaires sont nécessaires et ces approches constituent des voies de recherche afin de traiter des non linéarités dans le cadre de l'approche PLS. Des études sur les propriétés des transformations pourraient être menées. Une analyse complète basée sur des simulations par méthodes de Monte Carlo est possible afin de valider les résultats.

De nombreux points restent en suspens et pourront être traités par la suite. Par ailleurs, d'autres méthodes pourraient être développées :

- L'approche GSCA de Hwang et Takane (2004) utilise les moindres carrés alternés pour l'estimation des coefficients du modèle. On pourra facilement intégrer des transformations, soit pour quantifier des variables non continues (Hwang et Takane, 2002), soit comme dans le cadre de ce chapitre afin d'estimer des non linéarités.

- La généralisation de la méthode PLSS développée par Durand (2001). Celle-ci a été mise en place dans le cadre de la régression PLS qui apparaît comme un cas particulier de l'approche PLS.

Cette méthode permet une transformation par B-splines des variables prédictives.

– L'utilisation du mode B de l'approche PLS : celui-ci peut plus facilement être traité du fait de l'existence d'une fonction globale à optimiser qui permet d'utiliser un algorithme basé sur l'estimation de cette fonction. Ainsi, pour le modèle externe, ceci revient à trouver une transformation Ψ telle que :

$$\begin{array}{l} \max \sum_{k,l=1}^{K} c_{kl} cov^2(\Psi_k(\mathbf{X}_k)\mathbf{w}_k, \Psi_l(\mathbf{X}_l)\mathbf{w}_l) \\ s.c. \parallel \Psi_k(\mathbf{X}_k)\mathbf{w}_k \parallel = 1 \, \forall k \end{array} \tag{5.16}$$

et pour le modèle interne, ceci revient à trouver une transformation Υ telle que :

$$\begin{array}{l} \max \sum_{k,l=1}^{K} c_{kl} cov^2(\Upsilon_k(\mathbf{X}_k\mathbf{w}_k), \Upsilon_l(\mathbf{X}_l\mathbf{w}_l)) \\ s.c. \parallel \Upsilon_k(\mathbf{X}_k\mathbf{w}_k) \parallel = 1 \, \forall k \end{array} \tag{5.17}$$

L'utilisation de B-splines monotones pourra de la même façon constituer une option intéressante.

Ces ouvertures montrent des voies de recherche afin de travailler sur la non linéarité dans les modèles d'équations structurelles. Nous nous sommes focalisés, dans le cadre de cette thèse, sur le cas réflectif et donc sur le mode A d'estimation des poids externes. Nous présenterons dans le cadre du chapitre 7.7 (p. 165) des applications sur des données de satisfaction des clients.

Chapitre 6

Modèles structurels et données manquantes : le traitement d'un cas spécifique

6.1 Introduction

Tout statisticien lorsqu'il travaille sur des données réelles doit prendre en compte la présence de données manquantes. Le cas des modèles d'équations structurelles à variables latentes ne fait pas exception, les données manquantes dans les questionnaires y sont fréquentes et proviennent souvent de différents processus d'absence. Leur traitement se fait par de nombreuses méthodes développées dans la littérature (voir Allison (2001), Little et Rubin (2002) ou Schafer et Graham (2002)). Néanmoins, le choix d'une méthode de traitement dépendra fortement du processus sous-jacent à l'absence des données. Il existe trois types de données manquantes :

- Les données qui manquent totalement aléatoirement (MCAR, *missing completely at random*). La probabilité qu'une donnée associée à la variable $\mathbf{y}$ soit manquante (on notera $\mathbf{y}_{manquant}$) ne dépend ni de la valeur de $\mathbf{y}$, ni de celle d'autres variables. Cette hypothèse est forte et on lui préfère souvent la suivante.
- Les données qui manquent aléatoirement (MAR, *missing at random*). La probabilité qu'une valeur de la variable $\mathbf{y}$ soit manquante ne dépend pas de la variable $\mathbf{y}$. Elle peut dépendre d'autres variables. On aura :

$$P(\mathbf{y}_{manquant}|\mathbf{y},\mathbf{x}) = P(\mathbf{y}_{manquant}|\mathbf{x}), \quad \forall\, \mathbf{x}$$

- Les données qui manquent non aléatoirement (MNAR, *missing not at random*). Dans ce cas, les données manquantes suivent un processus qui peut être défini par une fonction, ainsi la probabilité d'avoir une donnée manquante dépend à la fois d'autres variables et de la valeur de la variable $\mathbf{y}$. Ce cas doit être traité lorsque le processus est bien connu. On aura :

$$P(\mathbf{y}_{manquant}|\mathbf{y},\mathbf{x}) = P(\mathbf{y}_{manquant}|\mathbf{y},\mathbf{x}), \quad \forall\, \mathbf{x}$$

  C'est clairement le cas de données censurées, tronquées ou sélectionnées.

Pour plus de détails sur ces notions dans le cadre des modèles structurels, on peut voir Gold et Bentler (2000) pour MAR et MCAR, et Muthén et al. (1987) pour MNAR.

Nous commençons par introduire rapidement les méthodes de traitement des données manquantes dans le cadre des modèles structurels à variables latentes. Par la suite, nous nous focalisons sur le

cas de données qui manquent non aléatoirement. Finalement, nous présentons des ouvertures pour le traitement d'un type de données spécifiques, les données dites "non applicables" ou filtrées.

6.2 Les données MCAR et MAR

6.2.1 Les méthodes classiques

Nous rassemblons les méthodes usuelles en détaillant brièvement leur principe.

- **Méthode par délétion :** chaque observation ayant sur l'une de ses variables une donnée manquante est supprimée. Dans la pratique, il est fréquent d'avoir des données manquantes sur plusieurs observations différentes. L'utilisation d'une méthode de délétion entraînera donc une forte perte de données et des biais importants.

- **Remplacement par la moyenne :** méthode fréquemment utilisée mais produisant des biais importants sur les variances et covariances. La variabilité des variables en est réduite.

- **L'imputation Hot-Deck :** Pour chaque observation ayant une valeur manquante, on recherche une observation similaire sur l'ensemble des autres variables et on impute en utilisant cette observation. Cette méthode est très utilisée dans la théorie des enquêtes (Ford, 1983).

- **La méthode d'imputation simple par algorithme EM (*Expectation Maximisation*) :** On estime les valeurs manquantes en utilisant un algorithme de maximisation de l'attente (EM). Il est basé sur deux étapes :

  Etape (E) Celle-ci consiste à remplacer les données manquantes par la moyenne conditionnelle étant donné les données disponibles. Celle-ci est obtenue grâce au maximum de vraisemblance.

  Etape (M) Celle-ci consiste en l'obtention d'une mise à jour de la matrice de covariance.

  On répète ces deux pas jusqu'à convergence (matrice de covariance identique entre deux pas). Cette méthode permet d'obtenir une imputation basée sur le maximum de vraisemblance. Elle est souvent utilisée, l'une des critiques qui peut lui être faite concerne les variances qui ne sont pas valides (Dempster et al., 1977).

- **Les méthodes d'imputation multiples :** proposées par Rubin (1987), elles utilisent les méthodes d'imputation classiques basées soit sur le maximum de vraisemblance, soit sur l'algorithme EM. En répétant ces procédures d'imputation simple, on obtient différentes imputations. La variabilité des estimations permet d'obtenir des variances plus grandes que dans le cas de l'imputation simple et donc plus proches de leur valeur réelle.

De nombreux auteurs ont comparé les différentes méthodes de traitement des données manquantes. Les plus adaptées dans le cas général sont celles utilisant l'imputation multiple et l'algorithme EM. Pour des comparaisons et des présentations de ces différentes méthodes, voir Brown (1994), Gold et Bentler (2000) ou Olinsky et al. (2003).

Ces méthodes permettent donc d'obtenir un jeu de données complet afin d'appliquer les méthodes d'estimation des modèles d'équations structurelles à variables latentes. D'autres méthodes existent mais nous allons nous focaliser sur celles spécialement adaptées aux modèles structurels.

6.2.2 La méthode *Full Information Maximum Likelihood* (FIML) et l'approche LISREL

A la différence des approches classiques, plutôt que de compléter les données avant d'estimer le modèle, on va utiliser les données disponibles afin d'estimer les paramètres (Arbuckle, 1996).

Soit $\mathbf{x}$ une variable à N observations avec l données manquantes. On pourra donc créer une variable $\mathbf{x}^\star$ à $N-l$ observations telle que $\mathbf{x}^\star = \mathbf{V}\,\mathbf{x}$ avec $\mathbf{V}$ matrice de sélection (cette matrice est une matrice identité à laquelle on a retiré les lignes associées aux observations manquantes). On suppose alors que les variables ainsi transformées suivent une distribution normale de moyenne $\mathbf{V}\boldsymbol{\mu}$ et de matrice de covariance $\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}'$. A partir de ces données, il est simple d'estimer les paramètres du modèle d'équations structurelles en se basant sur le maximum de vraisemblance.

Ainsi pour chaque observation i de l'échantillon, on estime la fonction de log-vraisemblance associée à l'observation :

$$\log L_i = K_i - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{y}_i^\star - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i^\star - \boldsymbol{\mu}_i)$$

avec $\mathbf{y}_i^\star$ vecteur associé à toutes les variables complètes de l'observation i, $\boldsymbol{\mu}_i$ vecteur des moyennes de l'ensemble des variables non manquantes pour le cas i, K_i constante associée au nombre de données complétées pour le cas i. $\boldsymbol{\Sigma}_i$ est la matrice de covariance des variables disponibles pour le cas i. On a :

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} \log L_i$$

L'avantage de cette approche réside dans le fait que les variances et covariances sont obtenues directement sans biais associé à la taille de l'échantillon. Enders et Bandalos (2001) montrent par de nombreuses simulations que, dans le cadre de l'estimation par maximum de vraisemblance de l'approche LISREL, cette approche est meilleure que les précédentes en terme de biais d'estimation.

6.2.3 L'algorithme *Non-linear Iterative PArtial Least Squares* (NIPALS) et l'approche PLS

En cas de présence de données manquantes, l'approche PLS peut travailler sur les données disponibles (Wold, 1973) par le biais de l'algorithme NIPALS. Celui-ci est antérieur à l'approche PLS, il permet de faire une ACP d'un bloc de variables en utilisant uniquement les données disponibles par le biais d'une succession de régressions.
Il s'intègre directement dans la première partie de l'algorithme PLS.

Pour une présentation de l'algorithme, on peut voir Tenenhaus (1998, p.61). Son intégration dans l'approche PLS se fait par deux étapes :

Etape 1 : $\mathbf{t}_1 \propto \mathbf{X}\mathbf{w}_1$

Etape 2 : $\mathbf{w}_1 = \frac{\mathbf{X}'\mathbf{t}_1}{N}$

Si on a des données manquantes, les deux étapes sont modifiées. Pour la première, on aura :

$$t_j = \frac{\sum_{i:x_{ji},w_i \text{ présentes}} x_{ji} w_i}{\sum_{i:x_{ji},w_i \text{ présentes}} w_i^2}$$

Pour la seconde étape, on aura :

$$w_i = \frac{\sum_{j:x_{ji} \text{ présentes}} x_{ji} t_j}{\sum_{j:x_{ji} \text{ présentes}} t_j^2}$$

Ces deux étapes sont répétées jusqu'à convergence, on estime donc la variable latente en se basant sur le modèle externe et en prenant uniquement en compte les données présentes.

Cette méthode permet de garder l'aspect sans distribution de l'approche PLS et sera donc recommandée dans ce cas spécifique.

6.2.4 Exemple sur des données simulées

Nous illustrons le traitement des données manquantes sur des données simulées. Nous simulons des données en nous basant sur un modèle simple à 2 variables manifestes par variable latente et trois variables latentes (cf. fig. 6.1). Nous retirons aléatoirement, dans un premier cas, 5% des données et, dans un second cas, 10% des données. Sur les jeux de données obtenus, nous appliquons les méthodes d'imputation suivies des approches LISREL et PLS. Par ailleurs, nous appliquons NIPALS et FIML sur les données non complétées. Les méthodes de délétion sont laissées de côté car le nombre de cas conservés est trop faible.
Nous rassemblons dans les tableaux 6.1 et 6.2 les résultats associés aux coefficients structurels et aux indices de qualité d'ajustement pour, respectivement, 5 et 10% de données manquantes. Les méthodes de complétion utilisées sont l'imputation par la moyenne (IM), l'imputation multiple par algorithme EM (MI) et les méthodes FIML et NIPALS. Les comparaisons se font aussi par rapport aux données sans données manquantes (NM). Les termes entre parenthèses représentent les écarts types obtenus par bootstrap lors de l'application de LISREL et de PLS.

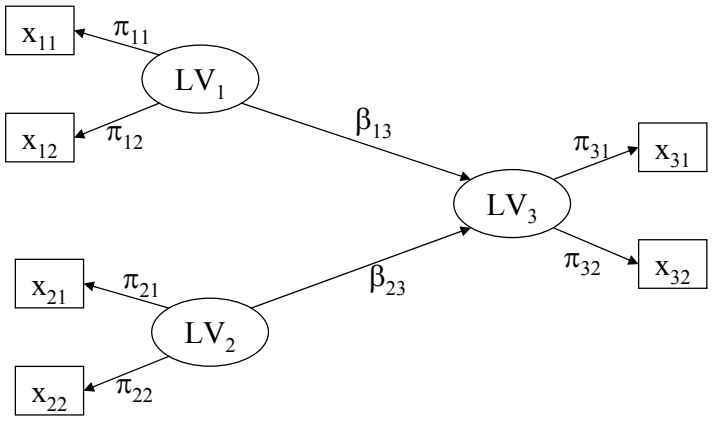

FIG. 6.1 – Modèle structurel simulé

Cet exemple illustre bien les biais induits par les méthodes d'estimation du modèle structurel. Les biais des estimations des coefficients sont plus importants pour l'approche PLS, ce qui se justifie par l'adaptation du modèle aux conditions associées à la méthode LISREL (normalité des données et modèle bien spécifié) mais aussi par la notion de consistance au sens large (le nombre d'observations est grand, mais on a uniquement deux variables manifestes par construit).
Au niveau des coefficients structurels, les résultats ne varient pas en fonction du nombre de données manquantes. Les écarts types sont plus élevés. L'algorithme NIPALS obtient néanmoins des écarts types plus faibles.
Au niveau des *loadings*, les résultats sont très proches. Les données étant simulées à partir d'une distribution normale, les méthodes de complétion basées sur le maximum de vraisemblance possèdent un avantage sur les autres. Elles permettent de bien reconstituer les données supprimées aléatoirement, on voit que l'imputation multiple (MI) et la méthode FIML permettent d'obtenir des coefficients très proches du cas complet.

La méthode FIML obtient de très bons résultats et sera généralement préférée car elle est spécialement adaptée aux modèles structurels, c'est aussi la conclusion des travaux sur le sujet de Olinsky

Méthode	Simulé	LISREL				PLS (mode A)			
Compl.	-	NM	IM	MI	FIML	NM	IM	MI	NIPALS
β_{13}	0.6	0.52(0.039)	0.53(0.031)	0.52(0.037)	0.52(0.039)	0.42(0.071)	0.40(0.069)	0.43(0.071)	0.41(0.068)
β_{23}	0.7	0.62(0.040)	0.62(0.041)	0.61(0.039)	0.61(0.041)	0.47(0.069)	0.45(0.067)	0.48(0.063)	0.46(0.057)
π_{11}	0.9	0.86(0.044)	0.76(0.043)	0.87(0.042)	0.87(0.046)	0.89(0.053)	0.91(0.045)	0.90(0.051)	0.91(0.047)
π_{12}	0.8	0.80(0.051)	0.72(0.052)	0.77(0.049)	0.77(0.053)	0.79(0.080)	0.77(0.060)	0.80(0.069)	0.79(0.061)
π_{21}	0.9	0.97(0.051)	0.91(0.051)	0.97(0.048)	0.97(0.053)	0.90(0.042)	0.90(0.038)	0.90(0.038)	0.90(0.033)
π_{22}	0.7	0.70(0.049)	0.69(0.049)	0.70(0.047)	0.71(0.051)	0.75(0.093)	0.74(0.081)	0.75(0.087)	0.76(0.081)
π_{31}	0.8	0.88(0.025)	0.84(0.029)	0.88(0.036)	0.88(0.032)	0.93(0.012)	0.93(0.008)	0.94(0.014)	0.94(0.012)
π_{32}	0.9	1.01(0.034)	0.96(0.037)	1.02(0.037)	1.01(0.038)	0.95(0.008)	0.94(0.013)	0.95(0.008)	0.95(0.011)
χ^2	-	5.4 (6dl)	10.53 (6dl)	5.92 (6dl)	9.14 (6dl)	-	-	-	-
GoF	-	-	-	-	-	0.550	0.531	0.533	0.530

TAB. 6.1 – Résultats des simulations sur les données manquantes (avec 5% de données manquantes)

Méthode	Simulé	LISREL				PLS (mode A)			
Compl.	-	NM	IM	MI	FIML	NM	IM	MI	NIPALS
β_{13}	0.6	0.52(0.039)	0.50(0.042)	0.52(0.035)	0.52(0.041)	0.42(0.071)	0.37(0.069)	0.43(0.069)	0.39(0.060)
β_{23}	0.7	0.62(0.040)	0.61(0.045)	0.61(0.040)	0.61(0.044)	0.47(0.069)	0.44(0.063)	0.48(0.067)	0.44(0.047)
π_{11}	0.9	0.86(0.044)	0.80(0.037)	0.86(0.040)	0.86(0.048)	0.89(0.053)	0.91(0.037)	0.91(0.045)	0.91(0.041)
π_{12}	0.8	0.80(0.051)	0.69(0.049)	0.79(0.048)	0.79(0.055)	0.79(0.080)	0.75(0.060)	0.80(0.070)	0.79(0.051)
π_{21}	0.9	0.97(0.051)	0.88(0.043)	0.95(0.047)	0.95(0.057)	0.90(0.042)	0.91(0.037)	0.91(0.036)	0.91(0.033)
π_{22}	0.7	0.70(0.049)	0.62(0.038)	0.69(0.046)	0.69(0.053)	0.75(0.093)	0.70(0.080)	0.75(0.079)	0.75(0.066)
π_{31}	0.8	0.88(0.025)	0.81(0.041)	0.88(0.035)	0.88(0.038)	0.93(0.012)	0.93(0.015)	0.95(0.013)	0.94(0.013)
π_{32}	0.9	1.01(0.034)	0.93(0.030)	1.00(0.030)	1.01(0.038)	0.95(0.008)	0.94(0.014)	0.96(0.007)	0.95(0.013)
χ^2	-	5.4 $(6dl)$	8.9 $(6dl)$	5.9 $(6dl)$	8.8 $(6dl)$	-	-	-	-
GoF	-	-	-	-	-	0.550	0.510	0.551	0.503

TAB. 6.2 – Résultats des simulations sur les données manquantes (avec 10% de données manquantes)

et al. (2003) et de Verleye et al. (1999). La méthode par imputation par la moyenne (IM) obtient des biais plus importants et minimise l'écart type car elle baisse la variabilité des variables manifestes.

Pour l'approche PLS, le fait d'utiliser des données normales favorise l'imputation multiple qui permet de mieux reconstruire les données. L'algorithme NIPALS obtient de bons résultats avec une tendance à rendre les écarts types trop petits. Dans le cas de données réelles non normales, il sera préféré afin de conserver l'aspect sans distribution de l'approche PLS.

Cet exemple illustre les différences entre les approches. Une étude basée sur des simulations par méthodes de Monte Carlo serait nécessaire afin de valider ces commentaires. Sur l'approche LISREL, les études effectuées jusqu'alors offrent des conclusions proches de celles données dans le cadre de cet exemple. Dans le cadre de l'approche PLS, aucune étude n'a été menée.

6.3 Les données MNAR

Les données qui manquent non aléatoirement suscitent de plus en plus de recherches et tout spécialement dans le domaine de l'économétrie. Ces méthodes sont paramétriques et nécessitent une très bonne connaissance du processus d'absence des données. Il existe deux types de méthodes : les *sample selection model* et les modèles de mélanges. Nous nous intéressons aux premiers. Soit X une variable d'intérêt et Y une variable binaire qui vaut 1 si X est observée et 0 sinon. Soit $f(X,Y)$ la densité de probabilité jointe. Choisir un modèle associé aux données manquantes revient à choisir $f(X,Y)$. Pour les *sample selection model*, on utilisera :

$$f(X,Y) = Pr(Y|X)f(X)$$

On modélise X comme si aucune donnée ne manquait, par exemple, on peut supposer que X est normale de moyenne μ et de variance σ^2 et que

$$Pr(Y=1|X) = \begin{cases} p_1 & \text{si } X > 0 \\ p_2 & \text{si } X \leq 0 \end{cases}$$

Le modèle peut alors être estimé par le maximum de vraisemblance.

6.3.1 Les modèles économétriques

L'utilisation des principes du modèle Tobit (Tobin, 1958) apparaît comme une solution. Muthén (1989) a mis au point une analyse factorielle Tobit qui pourrait être intégrée dans les modèles d'équations structurelles. Cependant, quelques points posent problème. Le modèle Tobit est basé sur la notion de variable censurée, ce qui limite le traitement des données qui manquent non aléatoirement. Dans un cadre plus large, le modèle Heckit (Heckman, 1979), qui est basé sur les *sample selection model*, sera plus adapté.

Le modèle Tobit

Le modèle Tobit est appliqué en économétrie sur la consommation de certains biens durables par les ménages en fonction de leurs revenus. Cette consommation présente la particularité suivante : la somme consacrée pendant une certaine période à l'achat d'un bien durable donné peut prendre toute valeur positive, mais est nulle pour beaucoup de ménages (Gourieroux, 1989). Le modèle est alors construit en deux phases car on suppose qu'il s'agit d'un comportement séquentiel :

Première phase : l'individu va décider s'il achète ou non le bien ; cette première décision peut être décrite par un modèle catégoriel dichotomique fondé sur un critère :

$$\begin{cases} \text{si } y_{2i}^* > 0, \text{ l'individu } i \text{ achète du bien} \\ \text{si } y_{2i}^* \leq 0, \text{ l'individu } i \text{ n'en achète pas} \end{cases} \tag{6.1}$$

Seconde phase : l'individu fixe la somme y_{1i}^* qu'il va consacrer à l'achat de bien. La variable observée y_i est alors :

$$y_i = \begin{cases} y_{1i}^* & \text{si } y_{2i}^* > 0 \\ 0 & \text{sinon} \end{cases} \tag{6.2}$$

Cette modélisation permet notamment de faire apparaître la plus ou moins grande corrélation existant entre les deux décisions.

Plaçons-nous dans le cadre de variables manifestes quantitatives, pour la variable x_i pour laquelle sur les N observations de l'échantillon, uniquement l ont répondu.
On aura y_{2i}^* est positif pour ces l observations et négatif pour les autres, la variable y_i sera définie par :

$$y_i = \begin{cases} y_{1i}^* & \text{si } y_{2i}^* > 0 \\ 0 & \text{sinon} \end{cases} \tag{6.3}$$

avec $y_{1i}^* = x_{1i}b_1 + u_{1i}$ et $y_{2i}^* = x_{2i}b_2 + u_{2i}$. On a que y_{1i} n'est observé que lorsque y_{2i} est positif.

On définit deux ensembles :
– $\Im_0 : y_i = 0$, avec $\#(\Im_0) = N - l$
– $\Im_1 : y_i \neq 0$, avec $\#(\Im_1) = l$
On utilise alors une estimation en deux étapes afin d'estimer les paramètres de y_i, cette méthode n'est pas efficace (dans le sens anglophone de "efficient") mais a l'avantage d'être rapide.
On définit :

$$z_i = \begin{cases} 1 & \text{si } y_{2i}^* > 0 \\ 0 & \text{sinon} \end{cases} \tag{6.4}$$

On a donc $P(z_i = 1) = P(y_{2i}^* \geq 0) = \Phi(x_{2i}\frac{b_2}{\sigma_2})$, qui nous permet de trouver une estimation $\hat{c}_2$ de $c_2 = \frac{b_2}{\sigma_2}$. Jusqu'ici on a utilisé la partie qualitative du modèle, on utilise maintenant la partie quantitative, c'est-à-dire les indices $i \in \Im_1$, on aura :

$$E(y_i | i \in \Im_1) = x_{1i}b_1 + \rho\sigma_1 \frac{\varphi(\frac{x_{2i}b_2}{\sigma_2})}{\Phi(\frac{x_{2i}b_2}{\sigma_2})} \tag{6.5}$$

On obtient donc des estimateurs sans biais de b_1 et de $\rho\sigma_1$ par régression et le résidu de cette régression va nous permettre d'estimer σ_1 :

$$\hat{v}_{1i} = y_i - x_{1i}\hat{b}_1 - (\hat{\rho\sigma_1})\frac{\varphi(x_{2i}\hat{c}_2)}{\Phi(x_{2i}\hat{c}_2)} \tag{6.6}$$

et donc

$$\hat{\sigma}_1^2 = \frac{1}{n_1}\sum_{i\in\Im_1}\hat{v}_{1i} + \frac{(\hat{\rho\sigma_1})^2}{n_1}\sum_{i\in\Im_1}(x_{2i}\hat{c}_2\frac{\varphi(x_{2i}\hat{c}_2)}{\Phi(x_{2i}\hat{c}_2)} + (\frac{\varphi(x_{2i}\hat{c}_2)}{\Phi(x_{2i}\hat{c}_2)})^2) \tag{6.7}$$

Cette procédure donne des estimateurs convergents, asymptotiquement normaux.

Malheureusement, le modèle Tobit se limite au cas où les données sont censurées ou tronquées à gauche ou à droite.

Modèle de Heckman (1979)

Les modèles du type *sample selection model* développés par Heckman, que nous nommons modèles Heckit, sont plus adaptés aux modèles structurels (Heckman, 1979; Sigelman et Zeng, 1999).

Dans ce cas, on aura deux équations, la première définissant la variable filtre et la seconde définissant la variable filtrée elle-même. L'estimation des paramètres peut alors se faire par maximum de vraisemblance.

$$\begin{cases} z_i^* = w_i\gamma + \mu_i \\ y_i = x_i\beta + \epsilon_i, \text{ observée si } z_i^* > 0 \end{cases} \tag{6.8}$$

Les termes d'erreur sont supposés suivre une distribution normale bivariée de moyenne nulle. Comme dans le cas du modèle Tobit, l'estimation par régressions ordinaires donne des estimations biaisées qui ne prennent pas en compte le processus de sélection. En utilisant la distribution conjointe, on peut obtenir le fonction de vraisemblance de y_i et, ainsi, estimer les paramètres en utilisant le maximum de vraisemblance.
Les modèles présentés supposent de fortes hypothèses de distribution. Ceci rend leur application difficile dans le cadre de l'approche PLS. Par contre, combinées à des modèles estimés par le maximum de vraisemblance, ces méthodes trouvent une application directe.

6.3.2 Données MNAR et modèles d'équations structurelles

Dans la littérature, le traitement des données qui manquent non aléatoirement se limite au cas de LISREL. Dans le cas de PLS, des problèmes dus à la modélisation douce apparaîtront.

Muthén et al. (1987), Tang et Lee (1998) et Lee et Tang (2006) ont traité ce type de problèmes dans le cadre des modèles d'équations structurelles à variables latentes. Les méthodes précédemment présentées sont utilisées quand le processus d'absence des données est supposé connu.

Muthén et al. (1987) présentent une méthode dans laquelle le processus d'absence des données peut être modélisé soit par les variables manifestes du modèle, soit par les variables latentes. Ainsi, ils supposent que $\mathbf{x}_{kl}^\star$ représente la variable sous-jacente à la variable observée $\mathbf{x}_{kl}$ ne comportant aucune donnée manquante. $\mathbf{s}_{kl}^\star$ représente une variable de sélection tel que si $s_{kli}^\star \geq \tau_{kli}$, alors x_{kli} est observée et si $s_{kli}^\star < \tau_{kli}$ alors x_{kli} est manquante. $\mathbf{x}_{kl}^\star$ et $\mathbf{s}_{kl}^\star$ sont supposées normales multivariées. On aura donc comme équation :

$$\begin{cases} \mathbf{x}_{kl}^\star = \boldsymbol{\pi}_{kl}\boldsymbol{\xi}_k + \boldsymbol{\epsilon}_{kl} \\ \boldsymbol{\xi} = \mathbf{B}\boldsymbol{\xi} + \boldsymbol{\zeta} \\ \mathbf{s}_{kl}^\star = \boldsymbol{\Gamma}_{\boldsymbol{\xi}}\boldsymbol{\xi}_k + \boldsymbol{\Gamma}_{\mathbf{x}_{kl}}\mathbf{x}_{kl}^\star + \boldsymbol{\delta}_{kl} \end{cases} \tag{6.9}$$

On veut donc estimer ce modèle, ce que l'on fait par maximum de vraisemblance. Soit $\boldsymbol{\theta}$ un vecteur rassemblant l'ensemble des paramètres associés au modèle, soit $\boldsymbol{\psi}$ un vecteur rassemblant les paramètres associés au processus d'absence des données. Dans leurs travaux, les auteurs considèrent que le processus d'absence est connu et estime ainsi les paramètres du modèle en supposant les $\boldsymbol{\psi}$ connus. L'utilisation d'une fonction de vraisemblance adaptée est alors nécessaire. Nous ne la détaillerons pas ici, on peut voir Muthén et al. (1987). Par l'utilisation de données simulées, les auteurs montrent l'efficacité de leur technique tout en soulignant la nécessaire connaissance du processus d'absence des données.

Par ailleurs, Tang et Lee (1998) présentent une méthode basée sur l'algorithme EM afin de traiter des données MNAR censurées ou tronquées. Ils reformulent le processus d'absence des données en estimant sa distribution. En utilisant cette distribution, ils présentent un algorithme EM permettant d'estimer les paramètres $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}(\theta)$ associés aux données $\mathbf{X}$.

Finalement, Lee et Tang (2006) utilisent des méthodes bayésiennes afin d'analyser le modèle avec absence de données. Ils supposent que cette absence suit un processus modélisé par un modèle de régression logistique.

Ces méthodes, pourtant bien développées dans la littérature, ont trouvé peu d'applications dans la pratique. Ceci peut être expliqué par les faibles relations entre les modèles structurels et les modèles économétriques, mais aussi par le fait que le processus d'absence des données est difficile à modéliser. D'autre part, le fait qu'un seul logiciel commercial (Muthén et Muthén, 1998) propose ce type de méthodes explique aussi cette désaffection.

6.4 Un cas spécifique : les questions filtrées

6.4.1 Un problème théorique et "philosophique"

Dans leur application originale, les modèles d'équations structurelles à variables latentes ont été créés pour traiter des données complètes ou des données contenant des valeurs qui manquent aléatoirement (MAR ou MCAR). Comme nous l'avons vu, le concept de données MNAR peut être traité dans le cadre de la méthode LISREL.

Les variables latentes sont des concepts définis pour l'ensemble de la population. Cependant, dans des cas pratiques, il peut arriver que certaines données soient absentes du fait de la construction même des enquêtes utilisées. C'est le cas des questions dites filtrées. Ce type de variables ne concerne qu'une partie de l'échantillon traité, en général, on utilise le terme "non applicable" afin de définir les données manquantes. Néanmoins, l'utilisation des données disponibles sur ces variables peut faire apparaître de nouvelles facettes de certains concepts et elle est donc importante. Par exemple, prenons le cas des réclamations. Seule une partie de la population est concernée et on voudrait étendre les conclusions à l'ensemble du modèle. Un processus déterministe sous-tend la présence ou l'absence d'une réponse. Ces variables ne peuvent pas être considérées comme des variables avec données manquantes, les non réponses sont en fait des "non applicables", on les nomme aussi *skip pattern*. Elles ont été étudiées dans le cadre de la théorie des enquêtes et servent à réduire la taille des questionnaires ou à aborder des sujets spécifiques à certains clients sans risquer d'erreur de cible. Il n'y a pas encore eu de recherches dans le cadre des modèles d'équations structurelles sur ce type de données. On utilise généralement des modèles séquentiels ou des graphes (Fagan et Greenberg, 1988) afin de les étudier. Il est généralement admis qu'on ne peut pas les traiter comme des données manquantes (Acock, 2005). Le processus d'absence ne dépend pas des valeurs prises par les données, il dépend de la structure de l'enquête. On peut modéliser l'absence ou la présence de données par une variable binaire. On ne peut pas supposer que, pour chaque variable exprimée par un sous-échantillon, il existe une variable complète sous-jacente ayant une distribution spécifique (pour tout l'échantillon). L'absence ou la présence peuvent être prédites mais la valeur à imputer ne peut pas être estimée.

Cette analyse nous amène à des questions sur la possibilité de traitement de ce type de variables dans les modèles structurels. D'un point de vue plus global, une question préliminaire s'impose : *Peut-on tirer des conclusions sur un modèle global alors que certaines variables ne sont pas complétées par tous les individus ?* Cette question devra rester à l'esprit des chercheurs. Les modèles d'équations structurelles à variables latentes tentent de donner une vision globale du modèle induit par l'échantillon et les influences dans le sous-modèle incomplet pourront-elles être intégrées dans le modèle global ?

Toutes ces questions devront être prises en compte dans les méthodes proposées et serviront à les critiquer, les mettre en valeur ou les disqualifier.

6.4.2 Les méthodes classiques

Comme nous l'avons vu plus haut, on ne peut pas dans ce cas supposer que les données manquent aléatoirement. Les méthodes de complétion entraîneront de graves erreurs d'interprétation. Les méthodes d'imputation par la moyenne et Hot-Deck posent, en plus des problèmes classiques, un problème de généralisation des valeurs associées aux répondants au cas des non répondants. Ceci ne peut pas être supposé. L'utilisation du maximum de vraisemblance est aussi problématique. On ne peut pas supposer qu'il existe une distribution sous-jacente commune aux données présentes et absentes. Les méthodes FIML et NIPALS, en plus de problèmes soulevés plus haut, feront face à des problèmes d'identification car la part de données manquantes est très importante. La seule approche acceptable est la délétion par liste qui amène au traitement du sous-échantillon ayant répondu aux variables filtrées.

On ne peut pas appliquer les méthodes présentées dans le cadre des données MNAR car la valeur manquante ne dépend pas des valeurs présentes. En effet, la donnée "non applicable" ne dépend pas des valeurs existantes, elle n'est donc pas MNAR par définition de ce concept. De plus, on ne peut pas trouver de processus permettant d'estimer la valeur à imputer, on ne pourra que prédire la notion absence/présence.

Il faut donc mettre en place d'autres techniques.

6.4.3 Méthodes alternatives

Nous proposons trois méthodes afin d'intégrer une partie ou toutes ces données dans un modèle d'équations structurelles en perdant le moins d'information possible.

La segmentation des données associée à la comparaison de modèle

Lorsque les filtres sont peu nombreux et si on a suffisamment de réponses aux questions filtrées, la meilleure solution est de traiter deux modèles indépendamment, l'un sans questions filtrées sur toute la population et l'autre avec les questions filtrées sur la part de la population y ayant répondu.

Ainsi, on suppose que le filtre est modélisé par une variable latente $\boldsymbol{\xi}_f$ avec $\mathbf{X}_f$ matrice des variables manifestes associées à cette variable latente. Ces variables sont définies pour un échantillon de taille $N - l$ où l est le nombre d'individus "non applicables". Nous illustrerons ce cas dans le cadre des applications. A partir de ce sous-modèle, les conclusions habituelles associées aux modèles structurels peuvent être effectuées en se restreignant au sous-échantillon traité.
Un problème important se pose. Les questions filtrées peuvent être assez précises, il y aura donc un nombre d'observations faible dans les sous-modèles obtenus. Il faudra donc être prudent avec ce type de questions lors de la mise en place du questionnaire.

Par ailleurs, il est possible de tester les différences entre deux groupes d'observations. Nous pouvons séparer l'échantillon en deux groupes : les individus ayant répondu au filtre et ceux n'y ayant pas répondu. Une fois cette séparation effectuée, l'utilisation de méthodes de comparaison de groupes d'observations est possible. La forte inégalité entre les tailles d'échantillon nous pousse à utiliser des méthodes avec rééchantillonnage du type test de permutation. L'ensemble des approches présentées dans le chapitre 4 de cette thèse sont ici applicables. Ceci permettra de connaître l'influence du filtre sur les différentes parties du modèle.

Cette méthode reste la plus adaptée afin de garder une théorie statistique sous-jacente robuste. Malheureusement, cela ne donne pas une vue globale de la problématique et les résultats pour des sous-échantillons petits sont souvent instables (cf. chap. 2.3).

L'ajout d'une variable supplémentaire

Afin de prendre en compte le fait de répondre au filtre, et l'impact de ce choix sur un concept spécifique, il est possible d'ajouter une variable binaire de la forme :

$$x_{iINDk} = \begin{cases} 1 & \text{si l'individu a répondu} \\ 0 & \text{sinon} \end{cases} \tag{6.10}$$

x_{iINDk} est définie pour le filtre i pour l'individu k. L'ajout de cette variable permet de faire apparaître dans le modèle la notion de corrélation entre la non réponse et le reste du modèle qui n'était pas possible plus tôt. L'utilisation d'une variable binaire avec l'approche PLS peut poser des problèmes, on pourra donc utiliser l'approche PML (chap. 2.4 et Jakobowicz et Derquenne (2007)).

Cette approche possède le désavantage de ne pas traiter l'ensemble des questions associées au filtre, de plus, l'ajout de cette nouvelle variable est parfois difficile à intégrer dans le modèle structurel.

L'utilisation d'une modalité supplémentaire

L'ajout d'une modalité à chaque variable afin de modéliser la notion de "non applicable" est ce qui est généralement fait. Mais, dans le cadre des modèles structurels, les données sont supposées ordinales et le choix de la modalité aura un effet sur les résultats. Nous proposons donc une méthode de substitution :

(1) Construire le modèle structurel avec l'ensemble des variables.

(2) Pour les variables ordonnées filtrées :
- Discrétisation de la variable en l modalités.
- Ajout de la modalité $l+1$ associée à la notion de "non applicable".

(3) Application de l'approche PML sur le modèle global en supposant que :
- Les variables non filtrées sont continues.
- Les variables filtrées sont nominales à $l+1$ modalités.

Ceci revient à appliquer du PLS classique pour les variables originales et du PML sur données nominales pour les variables filtrées. On obtient des estimations des apports de chaque modalité dans le cadre des questions filtrées. On peut supposer que la modalité "non applicable" aura un fort impact sur le modèle mais ceci permet d'avoir une vision complète du modèle.
La perte d'information générée par la discrétisation (que l'on considérera experte) et le traitement comme données nominales devra être étudié.

6.5 Conclusion

Le traitement des données manquantes constitue un terrain de recherche très important de la statistique.
Nous avons donc pu constater que les méthodes de traitement basées sur l'utilisation des données disponibles obtiennent de très bons résultats (méthode FIML et algorithme NIPALS). Pour ce qui est du traitement des questions filtrées, les trois approches présentées sont complémentaires, les deux premières sont très proches et la segmentation doit rester une priorité dans ce type de recherches. La troisième approche constitue une voie intéressante et elle reste à approfondir afin de connaître sa robustesse. Elle est bien adaptée dans le cadre de données sur des échelles différentes. Nous rassemblons dans le tableau 6.3 les points forts et les points faibles de chacune.

Une remarque très importante concerne la construction même de l'enquête. Certains filtres pourront dépendre d'événements aléatoires dans la population. Par exemple, le cas d'un contact aléatoire

venant d'une entreprise vers ses clients. Dans ce cas, le traitement par des méthodes classiques de données manquantes est possible. On peut finalement conclure que l'utilisation de questions filtrées dans les enquêtes doit se faire avec le plus grand soin et être évitée lorsqu'on travaille sur des concepts primordiaux pour l'analyse statistique des résultats.

La validation de l'efficacité de ces méthodes est complexe. En effet, l'utilisation de simulations afin de tester ces méthodes pose un problème majeur. La simulation de données manquantes peut se faire, soit aléatoirement (comme dans l'exemple de la page 118), soit en suivant un processus non aléatoire. Dans le premier cas, les méthodes présentées en 6.2 sont adaptées alors que dans le second, ce sont celles introduites en 6.3 qui peuvent être appliquées. La simulation de variables dites filtrées n'est pas possible, car dans le cas de données réelles, le filtre est induit par un processus différent pour chaque observation qui ne peut pas être considéré comme aléatoire. Nous préférons présenter des applications sur des données réelles afin d'illustrer le fonctionnement des approches introduite en 6.4. Ces applications sont rassemblées dans le chapitre 7.7 (p. 165) et traitent de l'analyse des réclamations des clients.

Méthode	Points forts	Points faibles
Segmentation et comparaison	- Traitement classique par l'approche PLS.	- Résultats locaux.
	- Utilisation des comparaisons de modèles (*multigroup comparison*).	- Problème de taille d'échantillon.
Ajout d'une variable	- Prise en compte du filtre.	- Les réponses aux variables filtrées ne sont pas prises en compte.
	- Traitement classique par PLS ou PML.	- Difficulté d'association de la nouvelle variable.
Ajout d'une modalité	- Traitement global du modèle.	- Problème de stabilité de la discrétisation. - Traitement comme nominal de données ordinales.

Tab. 6.3 – Points forts et points faibles des méthodes de traitement des questions filtrées

Chapitre 7

Applications à l'analyse de la satisfaction et de la fidélité

Les recherches intégrées dans cette thèse s'appuient sur un besoin réel : celui de l'analyse de la satisfaction et de la fidélité des clients. Nous présentons donc dans ce chapitre des applications illustrant l'utilité de ces approches dans le cadre de données, soit issues d'Electricité de France (EDF), soit issues de questionnaires de satisfaction venant d'autres domaines d'application. Nous présentons de manière simple l'application de l'ensemble des méthodes présentées dans cette thèse. Nos travaux sont illustrés en nous plaçant d'un point de vue industriel avec un questionnaire de satisfaction et des données.

Nous nous posons donc une question générale : que peut-on dire de la satisfaction et de la fidélité des clients d'une entreprise ? Quels en sont les antécédents, les facettes et les relations ? Nous suivons ce fil conducteur à travers chacun des sous-chapitres.
Nous partons de questionnaires de satisfaction dans le cadre d'une étude complète, nous allons répondre à 5 questions fréquentes dans la pratique :

1. Quelle méthode d'estimation du modèle choisir ?
2. Quel modèle conceptuel utiliser ?
3. Que peut-on dire du rapport des clients face à l'ouverture du marché de l'électricité ?
4. Peut-on supposer des relations linéaires dans les modèles structurels ?
5. Quel est l'impact des réclamations et des contacts sur la satisfaction et la fidélité des clients ?

Ces cinq questions sont récurrentes dans le traitement des modèles structurels à variables latentes. Nous allons illustrer la réponse à chacune d'entre elles par des applications sur des données réelles. Nous commençons par introduire les notions marketing clés afin de comprendre la construction de concepts aussi complexes que la satisfaction ou la fidélité.

7.1 Satisfaction et fidélité : concepts et problématiques

Dans le cadre de l'économie de marché, la réaction des consommateurs évolue très vite. D'où l'idée d'étudier les notions de satisfaction et de fidélité des clients qui en découlent. En marketing, le développement d'outils permettant de mettre à jour les leviers et les composantes de la satisfaction des clients, et par là même de la fidélisation, tient une place très importante avec la croissance de l'utilisation de stratégies du type CRM (*Customer Relationship Management*).

Avant de rechercher les leviers de la satisfaction et de la fidélité par des méthodes statistiques, il nous semble important de clarifier ces notions qui ont été étudiées dans des champs très divers (de la philosophie au marketing en passant par la psychologie) et qui ne sont pas clairement définies.

7.1.1 La notion de satisfaction

Définition du dictionnaire :
1. Etat d'esprit de quelqu'un dont les besoins, les envies, les souhaits sont satisfaits : contentement, plaisir. - LOC. "donner satisfaction à", "être un sujet de contentement de". - 2. Action par laquelle quelqu'un obtient réparation d'une offense qui lui a été faite. - 3. Fait d'accorder à quelqu'un ce qu'il demande.
Dictionnaire HACHETTE

Historiquement : Emprunté au latin *satisfactio* "excuse, justification, amende honorable", puis "réparation" et "action d'acquitter une dette ". Dérive de *satisfactum* supin de *satisfacere*. Désigne d'abord l'acte par lequel on obtient réparation d'une offense. Il s'est employé (1280) pour l'acte par lequel on paye ce que l'on doit. Dans le langage courant, (1611) désigne le plaisir qui résulte de l'accomplissement de ce que l'on désire. Le nom se dit ensuite (1836) de l'action de satisfaire un besoin, un désir.
Dictionnaire Historique de la Langue Française (LE ROBERT)

En marketing :
Un aperçu de la littérature montre qu'il existe beaucoup de définitions de la satisfaction. Dans Giese et Cote (2000), les auteurs tentent de rassembler ces définitions pour donner un cadre plus clair à la notion de satisfaction du client.
Ces définitions ont trois composantes communes :
- La satisfaction est une réponse
- Cette réponse a lieu sur un sujet particulier (attente, expérience...)
- Cette réponse a lieu à un moment particulier (après consommation, après un certain nombre d'expériences)

Dans le tableau 7.1, nous avons rassemblé les différentes définitions conceptuelles de la satisfaction.

Source	**Définition conceptuelle**
Howard et Sheth (1969, p.145)	Le sentiment de l'acheteur d'être suffisamment ou insuffisamment récompensé au vu de ce qu'il a subi.
Westbrook (1980, p.49)	Se rapporte à l'évaluation subjective des nombreuses expériences associées à l'utilisation d'un produit.
Oliver (1980, p.27)	Évaluation de la surprise associée à une acquisition ou à une expérience de consommation.
Day (1984, p.496)	L'évaluation de la réponse à un acte de consommation. L'évaluation en comparaison avec les attentes d'une expérience de consommation.
Woodruff et al. (1983, p.305)	Conceptualisation d'un sentiment développé lors de l'évaluation d'une expérience.
Fornell (1992, p.11)	Une évaluation générale suivant un achat.
Halstead et al. (1994, p.122)	Une réponse affective spécifique à une transaction résultant de la comparaison entre la performance perçue et la performance attendue.
Oliver (1997, p.13)	Le contentement du consommateur. Jugement sur le degré de plaisir qu'offre la consommation d'un produit.

Tab. 7.1 – Définitions conceptuelles de la satisfaction

Elle peut être définie comme :
***Une réponse affective d'intensité variable**. Le type exact de cette réponse peut varier en fonction du domaine de recherche. C'est au chercheur de la définir en fonction du contexte. Sa détermination se fait à **un moment spécifique et dure peu de temps**. C'est au chercheur de trouver le moment et la durée les plus adaptés. Elle est reliée à **l'acquisition ou à la consommation d'un produit**. Le chercheur devra identifier les différents acteurs de cette action et prendre en compte les plus importants (Giese et Cote, 2000).*

Remarque : Le paradigme de la "disconfirmation" des attentes est une autre façon de définir la satisfaction élaborée à la fin des années 1970 notamment par Oliver (1980), puis approfondie au cours des années 1980. Ce modèle décrit la formation de la satisfaction comme un processus comparatif comportant quatre construits principaux :

- Le jugement porté sur la performance du produit ou service (ou qualité perçue) au cours de l'expérience de consommation.
- Les attentes (ou *expectations*) formées par le consommateur préalablement à l'achat et à la consommation du produit ou service concerné.
- La "disconfirmation", qui correspond à la comparaison entre la performance et les attentes. Elle est positive lorsque les performances sont supérieures aux attentes ; neutre en cas d'égalité entre attentes et performances ; négative dans le cas où les performances sont inférieures au standard de référence des consommateurs.
- La "disconfirmation" va générer l'évaluation globale de l'expérience de consommation, c'est-à-dire la satisfaction.

Dans le cas particulier d'EDF :
La satisfaction peut être vue comme la réponse à une attente qui serait, dans le cadre du marché de l'électricité, l'approvisionnement en électricité et le tarif. Un client est "satisfait" si la réalité a au moins égalé cette attente. Dans le cadre d'EDF, le moment dans le temps est continu. Il faut toutefois faire attention à certains biais lors de l'estimation de la satisfaction. En effet, un client aura tendance à porter son attention plus particulièrement sur le tarif si l'enquête est effectuée durant une période de facturation, ou alors sur la qualité du service (approvisionnement) si l'on se trouve entre deux prélèvements.

7.1.2 La fidélité (*loyalty*) et ses problématiques

La notion de fidélité

Définition du dictionnaire :
1. Qualité d'une personne fidèle à ses engagements - 2. Attachement constant (à quelqu'un, à quelque chose) - 3. Respect de la réalité (fidélité d'un récit) - PHYS : qualité d'un appareil de mesure fidèle.
DICTIONNAIRE HACHETTE

Historiquement : A suivi l'évolution sémantique de fidèle : "qualité d'une personne fidèle" (1155), "constance dans les affections" (1670), honnêteté (1691), sorti d'usage aujourd'hui et par ailleurs "justesse, vérité" (1690) et "exactitude" en parlant d'un instrument de mesure (début du XX^{ème} siècle). A donné haute fidélité (Hi-Fi).
DICTIONNAIRE HISTORIQUE DE LA LANGUE FRANÇAISE (LE ROBERT)

En marketing :
Une définition classique (Jacoby et Kyner, 1973) :
La fidélité est définie comme une réponse comportementale (un achat), biaisée (non aléatoire), exprimée dans le temps, par une unité de décision, au regard d'une ou plusieurs marques alternatives, issues d'un ensemble, résultant de processus psychologiques (prise de décision, évaluation).

Ou encore, nous appelons fidélité une tendance latente du client au contrôle de ses actions, s'exprimant sur le long terme (plusieurs années), ayant pour effet d'augmenter ses achats en faveur de la marque, en dépit de l'influence des situations qui pourraient l'en détourner.

Dans l'entreprise, la fidélité est vue comme un phénomène sur le long terme, impliquant de la part des clients des comportements d'achat durables qui se traduisent par une croissance du chiffre d'affaires dégagé individuellement.

Il ressort de ces observations que les causes induisant la fidélité du client sont difficiles à définir, cependant un certain nombre de leviers peuvent être cherchés dans quatre directions :
- La fidélité attitudinale à travers ses composantes stables (attitude, valeurs) ou moins stable (confiance).
- L'apprentissage des processus de fidélisation, ou de l'accumulation des points.
- L'engagement comportemental.
- Les contributions économiques (*sunk costs*, *pledges*) comme les abonnements, les forfaits, les achats prépayés.

Remarques :
- Certains auteurs font une différence entre fidélité et rétention de clients. La fidélisation est expliquée par la satisfaction et explique la rétention du client. On peut voir par exemple Gerpott et al. (2001) dans le cas des télécommunications en Allemagne avec la création d'un nouveau modèle conceptuel marketing basé sur la satisfaction, la fidélité et la rétention de clients.
- On peut aussi voir la fidélité comme un concept multidimensionnel. En effet, Lam et al. (2002) divisent la fidélité en deux dimensions dans le modèle structurel. On aura une fidélité ne concernant que l'attachement à l'entreprise, c'est-à-dire le réachat (*patronage*), et une fidélité qui entraîne la recommandation de l'entreprise auprès d'autres potentiels clients (*recommend*). Ces deux facettes sont associées à des questions de l'enquête de satisfaction bien définies.

Dans le cadre de cette thèse, nous considérons que la fidélité est un concept unidimensionnel et que rétention et fidélité forment un seul concept dans nos modèles.

Dans le cas particulier d'EDF :
Le cas d'Electricité de France est spécifique à deux titres. Tout d'abord, on se trouve dans un cas de fidélisation d'un client fidèle a priori, du fait de la position historique de monopole de l'entreprise. On va donc essayer de garder des clients qui ne sont jamais partis et qui n'ont jamais eu de choix différents. La seconde spécificité découle directement de la première : comme EDF sort d'une position de monopole, il n'existe aucun historique de la fidélité et ceci pourra entraîner des difficultés dans la définition des leviers.

7.1.3 La fidélité : problématiques

L'expérience a montré depuis quelques années qu'acquérir un nouveau client demande des moyens très importants sans pour autant pouvoir être assuré que ce client rapportera à l'entreprise. Sachant ceci, les entreprises ont décidé d'axer leurs efforts sur la rétention de clients qui est bien moins onéreuse et qui assure une pérennité de l'entreprise. En effet, des campagnes de fidélisation concerneront les clients les plus potentiellement rentables et laisseront de côté les clients n'apportant aucun bénéfice à l'entreprise. Ainsi, cette fidélisation du client passe par un certain nombre d'étapes. La première est la compréhension des processus qui poussent un client à être fidèle. Il faudra donc définir les facettes de cette fidélité et découvrir ses leviers.

La littérature sur la fidélité en marketing est extrêmement conséquente. Cette notion a tout d'abord été vue comme un comportement unidimensionnel (Cunningham, 1956). Toutefois, elle est rapidement

devenue comportementale en plus d'être attitudinale (Jacoby et Chestnut, 1978). Par la suite, d'autres facettes ont été ajoutées. Rundle-Thiele (2005) en a identifié 5 qui devraient être intégrées aux questionnaires afin d'assurer une prise de décision cohérente. Elles sont :

- **La fidélité attitudinale :** Elle a été définie par Jacoby et Chestnut (1978) comme la prédisposition d'un client pour une marque. Elle est fonction de processus psychologiques. Pour la déceler, on utilise différents leviers : la préférence, l'intention de réachat, l'attachement. Souvent le bouche à oreille est inclus dans cette facette de la fidélité. C'est la partie la plus simple à mesurer, elle est en général intégrée dans les questionnaires de satisfaction.
- **L'attitude en cas de contact :** Elle concerne aussi bien des réclamations que des compliments. Les chercheurs sur le sujet sont opposés entre l'insertion de cette facette dans la fidélité et la considération de celle-ci comme une conséquence ou une cause de la fidélité (Singh, 1989). Nous développons ce concept dans le cadre des applications du chapitre 7.7 (p. 165).
- **La tendance à être fidèle :** Certains chercheurs voient la fidélité comme un trait de caractère. Ceci pourra être pris en compte. En effet, certaines personnes auront tendance à ne pas changer alors que d'autres seront toujours tentées par le changement (Raju, 1980). Dans de nombreux secteurs, on considère que l'âge est un levier important de cette facette.
- **Le comportement de résistance à la concurrence :** La relation entre cette résistance et la fidélité n'est pas claire. De plus, les chercheurs la perçoivent soit comme une facette, soit comme une cause, soit comme une conséquence de la fidélité ce qui posera des problèmes dans son intégration (Ganesh et al., 2000). On peut y associer les coûts de changement (en téléphonie par exemple).
- **La fidélité situationnelle :** Elle a été définie par Farley (1964). Elle suppose que la fidélité est expliquée uniquement par des situations rencontrées par le client. Le démarchage à domicile pourrait être intégré dans cette facette. Par contre, l'analyse de réponses associées à ce sujet entraînera la nécessité de traiter des questions filtrées.

Ces facettes de la fidélité devront être présentes dans les questionnaires de satisfaction orientés sur l'analyse de la fidélité. Nous présentons, en annexe dans le tableau B.1 (p. 179), un certain nombre de questions qu'il faudrait poser pour mieux identifier les différentes facettes de la fidélité dans le cadre d'une entreprise X.
Il s'avère que les questionnaires ne traitent en général que la fidélité attitudinale et parfois les réclamations.

Autres approches

D'autres points de vue existent, Bowen et Shoemaker (1998) voient la fidélité comme l'engagement ("commitment") qui est expliqué par la confiance. Pour Morgan et Sonquist (1963), la confiance entraîne, en général, la fidélité des clients. La figure 7.1 est un graphe représentant la théorie mise au point par Morgan et Sonquist (1963) pour le cas des services. Ce graphe de causalité est assez complexe, il fait ressortir les deux concepts cités précédemment. L'application se fait dans le secteur de l'hôtellerie de luxe. Bowen et Shoemaker (1998) concluent que la fidélité est basée sur les bénéfices en cas de fidélité, le coût de changement et la valeur perçue. D'autre part, l'engagement entraîne la recommandation et plus de dépenses de la part du client. Le fait d'intégrer la confiance comme levier de la fidélité est aussi appuyé par Harris et Goode (2004) qui démontrent son importance dans le cas des achats en ligne à l'aide de la méthode LISREL.

Les avantages représentent ceux liés à la fidélité à cette entreprise mais aussi à d'autres entreprises ayant des accords de programmes de fidélité avec elle (compagnie aérienne et chaînes d'hôtels, trains et locations de voitures...). Les actions et réactions opportunistes sont le fait de démarches des concurrents qui, de par leur importance, feront partir un client.
Nous voyons donc que les principaux concepts liés à la fidélité sont ici : l'engagement, la confiance, la communication, la valeur perçue et le prix. Nous avons rassemblé dans la figure 7.2 ces concepts pour obtenir un modèle simplifié, adapté à un cas plus général.

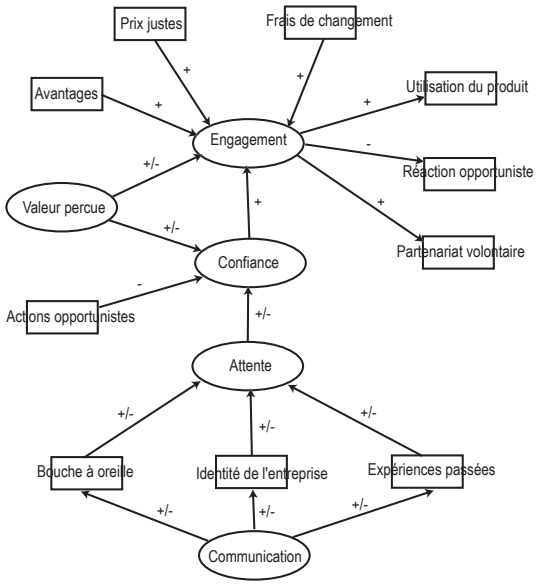

FIG. 7.1 – Modèle conceptuel marketing dans le cas des services afin de définir les leviers de l'engagement (Morgan et Sonquist, 1963)

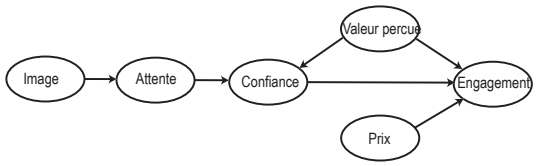

FIG. 7.2 – Généralisation du modèle de la figure 7.1

Dans ce modèle, l'image intègre la communication, le bouche à oreille, les expériences passées et l'identité de l'entreprise. Le prix prend en compte les avantages, la justesse du prix et les coûts de changement. Nous n'avons pas intégré les actions opportunistes car elles sont difficiles à définir.

Frisou (1998) a mis au point un modèle rassemblant toutes les approches marketing de la fidélité. C'est celui illustré dans la figure 7.3. On voit que l'auteur fait une synthèse entre le modèle présenté précédemment et le modèle classique satisfaction-fidélité. Le niveau cognitif est inspiré des théories de Morgan et Sonquist (1963) et les relations en pointillés sont des relations qui ont été avérées par Kotler et Dubois (1993) et qui montrent une grande perspicacité de la part du client. L'auteur a validé ce modèle sur des données concernant le marché de la téléphonie et il s'avère avec la méthode LISREL que les relations sont significatives. Les indices de qualité d'ajustement du modèle aux données sont bons. Il ressort que le client a un comportement qui prend en compte toutes les facettes de la fidélité.

De nombreux chercheurs se sont éloignés de la notion de fidélité en elle-même, ils se sont intéressés au lien entre satisfaction et fidélité et aux effets d'autres variables sur la force de ce lien (*compatibility effects*). Ainsi, Auh et Johnson (2005) montrent que les deux leviers de la satisfaction, la qualité et le prix, ont un effet différent sur la satisfaction et sur la fidélité. Une baisse du prix n'aura pas le même type d'effet sur la satisfaction que sur la fidélité. Ceci revient à démontrer que la relation satisfaction-fidélité n'est pas forcément linéaire. Ce sont Bloemer et Kasper (1995) qui ont posé cette

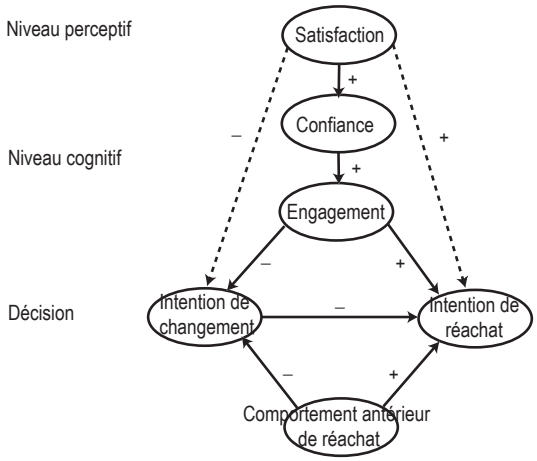

FIG. 7.3 – Modèle général développé par Frisou (1998)

		Fidélité	
		Faible	Forte
Satisfaction	Faible	*Déserteurs*	*Otages*
	Forte	*Mercenaires*	*Fidèles*

TAB. 7.2 – Différents profils de clients en fonction des degrés de satisfaction et de fidélité

hypothèse les premiers. Cette non linéarité pourra être contournée en ajoutant des nouveaux liens dans le modèle conceptuel marketing incluant satisfaction, prix, valeur et fidélité. Une autre approche consiste à définir plusieurs dimensions de satisfaction. Ainsi, Bloemer et Kasper (1995) utilisent deux types de satisfaction : la satisfaction manifeste et la satisfaction latente. La première est celle ressentie par le client après une expérience alors que la seconde est celle perçue par le client à tout moment. Nous nous intéressons à cette relation dans le cadre des applications sur la non linéarité dans le chapitre 7.6 (p. 157). Par ailleurs, ce lien entre satisfaction et fidélité peut nous permettre d'obtenir des profils de clients. Il en existe 4 principaux qui sont rassemblés dans le tableau 7.3. Ainsi, la répartition dans chacune des catégories sera très différente en fonction du secteur d'activité. Par exemple, lorsque les coûts de changement seront importants, le nombre "d'otages" sera grand. Les mercenaires sont clairement les clients dont la tendance à ne pas être fidèles est grande. Ce tableau est basé sur les recherches de Jones et Sasser (1995).

Fidélité et marché de l'électricité

Ces quinze dernières années, de nombreux pays ont décidé de libéraliser leur marché de l'électricité. La France en faisant partie, EDF est très intéressé par les expériences à l'étranger. Même si chaque pays, par son contexte socio-économique, par son histoire, par sa culture, réagit différemment à ce type de changement, il semble primordial de connaître les erreurs faites ailleurs et ainsi de les éviter.
Le cas le plus proche de la France est celui de la Grande Bretagne. La marché de l'électricité y est ouvert depuis 1991 et il ressort que 20% des clients ont changé de fournisseur après 5 ans. Les différentes études effectuées montrent que la fidélité a pour leviers principaux le prix, l'habitude et la commodité. D'autre part, l'image de marque est très importante. La relation entre fidélité et satisfaction est là aussi assez peu accentuée.
La commodité apparaît comme deux fois plus importante que le prix. Ceci peut vouloir dire que la fidélité attitudinale est primordiale dans ce secteur. Il s'avère que le changement de fournisseur est la plupart du temps le fait d'un démarchage à domicile. D'autre part, Percebois et Wright (2001) ont

fait une étude relative au prix de l'électricité entre la France et la Grande Bretagne dans les années 90. Il s'avère que pour les particuliers, en 1990, les prix étaient plus bas en France qu'en Grande Bretagne et qu'avec l'ouverture du marché britannique, en 2000, les prix français étaient toujours les plus bas. La baisse des prix étant du même ordre dans les deux pays. En Grande Bretagne, les prix au niveau de chaque région ont varié assez différemment. Cette étude fait ressortir que pour les particuliers, l'ouverture du marché ne correspond pas forcément à une baisse des prix. Il s'avère que pour les grandes entreprises, l'ouverture du marché britannique a apporté une forte baisse du prix de l'électricité, plus forte qu'en France où EDF continuait d'exercer sa position de monopole. A l'inverse de la Grande-Bretagne, où les prix au niveau de chaque région se sont mis à varier, les pays nordiques dans lesquels les prix étaient très variables en fonction de la région, l'ouverture de marché a permis une uniformisation. Johnsen (2003) présente le cas de la Norvège où la dérégulation s'est faite progressivement de 1991 à 1997. En 2002, seulement 20% des foyers norvégiens ont changé de fournisseur et cette déréglementation a entraîné une uniformisation des prix. Avec ces deux exemples, nous avons pu voir que la déréglementation du marché de l'électricité n'entraîne pas les mêmes conséquences en fonction du pays et du type d'énergie. En Grande Bretagne, c'est l'énergie fossile qui est préférée alors que l'hydroélectricité est omniprésente en Norvège. En France, c'est le nucléaire qui domine et il semble donc difficile de faire la moindre comparaison avec les autres pays. Cependant, nous devons les prendre en compte dans la mise en place des leviers de la fidélité. La majorité des chercheurs sur le sujet sont d'accord pour dire que le principal facteur d'infidélité, dans le marché de l'électricité, est le démarchage à domicile d'un concurrent. Cette facette devra être intégrée aux questionnaires.

La problématique générale : satisfaction client - fidélité

Le cas général : Toute entreprise veut fidéliser sa clientèle. Pour se faire, il faut identifier les composantes qui auront le plus d'importance dans cette fidélisation. Le lien entre satisfaction et fidélité est un lien central dans l'étude de ces deux concepts, clairement la satisfaction est l'une des causes de la fidélité, d'autres existent et seront spécifiques au cas traité. Par exemple, dans une étude sur le secteur de la téléphonie mobile (Lam et al., 2002), les leviers de la fidélité sont la satisfaction qui ressort comme le plus important, mais aussi la différence entre valeur attendue et valeur perçue, et le coût de changement d'opérateur (*switching cost*). D'autre part, cette relation n'est pas forcément linéaire. Heskett et al. (1997) ont montré que la fidélité augmente très vite après le passage d'un seuil de la satisfaction. Une littérature très importante sur le sujet existe et nous ne pourrions pas développer tous les points de vue ici (Bhote, 1996; Gitomer, 1998; Gronholdt et al., 2000). Il faut être très attentif à différencier la fidélité et l'intention de fidélité. La première ne peut être mesurée qu'a posteriori alors que la seconde est intégrée dans les questionnaires de satisfaction et n'a pas obligatoirement comme conséquence la fidélité du client.

L'industrie énergétique : Le secteur énergétique est à bien des égards un secteur particulier. En effet, il s'agit de fournir un bien capital auquel peu de valeur peut être ajoutée (électricité, gaz ...). De plus, c'est un secteur où la confiance est très importante. Un client acceptera sans trop de problèmes d'avoir une coupure de téléphone pendant quelques heures alors qu'une coupure d'électricité entraînerait un grand mécontentement.
C'est la raison pour laquelle il faut considérer le cas du secteur énergétique comme un cas totalement différent du cas général.

Le cas particulier d'Electricité de France : En plus d'être dans le secteur énergétique, EDF quitte une situation de monopole. Jusqu'alors, la satisfaction du client était très importante mais la partie fidélisation n'avait pas lieu d'être. Il va donc falloir associer à l'étude de la satisfaction client la notion de fidélisation de la clientèle face aux concurrents du secteur. Il faut donc s'appuyer sur l'expérience passée mais surtout sur celle de fournisseurs d'autres pays (dans lesquels le marché de l'électricité est déjà ouvert). Par exemple, nous pourrons nous appuyer sur

EDF Energy en Grande Bretagne qui a une longue expérience sur le comportement de fidélité des clients.

Une fois les relations obtenues, les méthodes d'estimation des modèles peuvent être utilisées par les différents acteurs de l'économie. La question qui se pose est de savoir quelle est la méthode la plus adaptée. Existe-t-il des spécificités liées à chaque méthode qui les rendent adaptées à certains secteurs d'activité ?

7.1.4 Les pratiques en matière d'analyse de la satisfaction du client

La plupart des grandes entreprises ont des stratégies de fidélisation de leur clientèle. C'est en partant de cette idée que nous avons décidé d'étudier le comportement d'autres entreprises face à ces problèmes.
Un certain nombre de méthodes ont été développées dans une optique d'amélioration de la satisfaction et de fidélisation. Nous allons donc les présenter avec, de plus, la présentation des stratégies d'un certain nombre d'acteurs de l'économie.

Les instruments de mesure

La qualité des données est capitale dans le cadre de l'estimation de la satisfaction, tout le processus d'obtention et de traitement des données doit être minutieusement effectué. Plusieurs grandes étapes apparaissent :

- La problématique de l'enquête.
- Le type d'enquête choisi.
- Le choix des personnes à interroger.
- La mise au point du questionnaire.
- La vérification et la transformation des données brutes.

Pour avoir une bonne synthèse de ces méthodes, on peut voir Ray (2001).

Les indices

Les indices de satisfaction donnent des évaluations de la satisfaction globale en utilisant des modèles d'équations structurelles à partir d'enquêtes de satisfaction auprès des clients. La majorité de ces indices ont été créés au niveau national afin d'obtenir un nouvel indicateur de la santé des entreprises globalement ou par secteur d'activité. C'est Claes Fornell qui fut le premier à développer la notion d'indice de satisfaction du client. Il a commencé par mettre au point l'indice suédois, le SCSB (*Swedish Customer Satisfaction Barometer*, Fornell (1992), Fig. 7.4). Celui-ci est constitué de 5 concepts, la satisfaction est expliquée par l'attente du client et la valeur perçue. Elle explique la fidélité et les réclamations. Il est basé sur 130 entreprises de 32 secteurs d'activité.
Le même auteur a ensuite mis au point l'ACSI (*American Customer Satisfaction Index*) aux Etats-Unis (Fornell et al., 1996). Pour ce dernier, le *path diagram* est celui de la figure 7.5, on voit qu'a été ajoutée la notion de valeur perçue. Il est basé sur 200 entreprises de 34 secteurs d'activité.

L'indice de satisfaction le plus récent est l'ECSI (*European Customer Satisfaction Index*) mis au point dans le cadre du projet ESIS (*European Satisfaction Index System*) soutenu par l'Union Européenne. Le modèle est celui de la figue 7.6 et la notion d'image de l'entreprise a été ajoutée.
Pour un récapitulatif des différents indices existants et la méthodologie de création d'un nouvel indice, on peut voir Johnson et al. (2001).

Les modèles développés dans le cadre des entreprises sont souvent issus des modèles ACSI et ECSI avec des adaptations pour les spécificités de chaque secteur. Dans le cas d'EDF, le modèle actuellement utilisé est confidentiel.

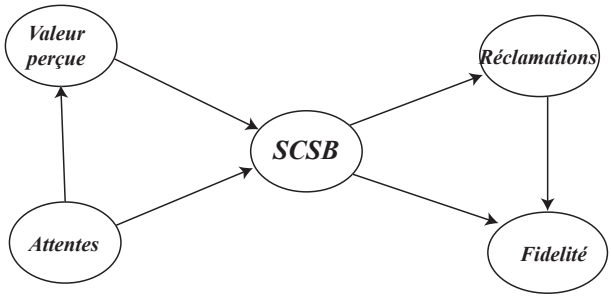

FIG. 7.4 – Modèle pour le SCSB

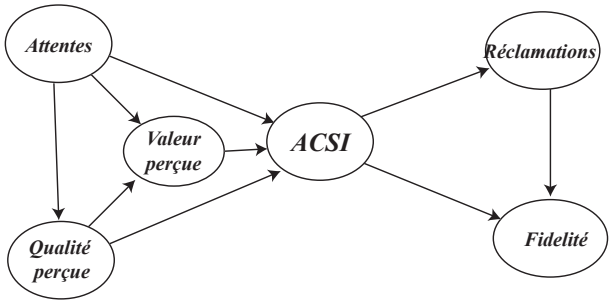

FIG. 7.5 – Modèle pour l'ACSI

Ces modèles ont été créés dans le cadre de l'application de l'approche PLS et peuvent poser des problèmes de non détermination avec la méthode LISREL.

Les usages

Il est intéressant de rechercher les pratiques en matière d'analyse de la satisfaction et de la fidélité des clients dans les principaux secteurs de l'économie.

Chez EDF : Dans le cadre de l'étude de la satisfaction des clients, EDF utilise un processus bien développé, mis en place par la R&D. Les questionnaires de satisfaction ont été mis en place depuis plusieurs années directement au sein d'EDF. Les réponses se font majoritairement sur des échelles de 1 à 10 et se séparent en plusieurs blocs représentant chacun une variable latente.
Les résultats sont traités par des méthodes d'équations structurelles à variables latentes : en premier lieu, l'approche PLS ; en second lieu, une méthode alternative issue de PLS et développée au sein d'EDF par Derquenne et Hallais (2004) appelée RFPC (*Regression on First Principal Components*). Dans le cas de présence de données manquantes, l'algorithme NIPALS (*NonLinear Iterative Partial Least Squares*) est utilisé. Par ailleurs, certaines méthodes développées dans cette thèse sont en cours de test.
EDF apparaît donc comme un précurseur dans l'utilisation des modèles structurels en France dans le milieu industriel.

Dans le secteur énergétique : Le groupe EDF international

- *EDF Energy* en Grande-Bretagne place EDF dans une autre optique qu'en France. En effet, il cherche à gagner des parts de marché. Des questionnaires de satisfaction y ont été développés et les résultats sont traités de la même manière que chez EDF France. C'est-à-dire avec l'approche PLS et l'algorithme NIPALS.
- *Yello* en Allemagne n'applique encore aucune méthode basée sur les modèles d'équations structurelles à variables latentes. Cependant, des modèles probabilistes de "churn" avec mise en place d'une typologie de "churneurs" sont utilisés.

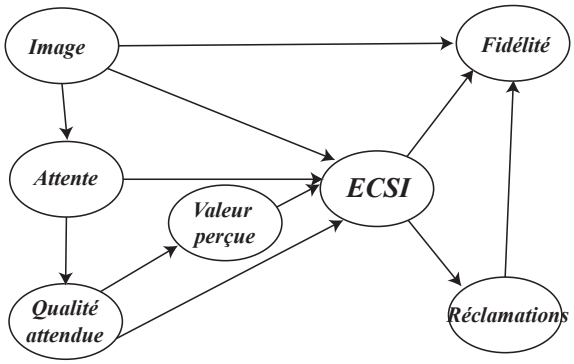

FIG. 7.6 – Modèle pour l'ECSI

Sur les autres acteurs du marché international, peu d'informations sont disponibles.

Hors secteur énergétique : Nous avons donc tenté d'obtenir un panorama des usages en la matière en consultant quelques acteurs de l'économie.

Tout d'abord, au niveau national, nous avons pu voir que la prise en compte de la satisfaction des usagers ou clients commence à apparaître comme un sujet important. En effet, en 2001, le gouvernement a demandé à la Cour des Comptes un rapport sur les méthodes d'évaluation de la satisfaction des usagers, dans le but d'obtenir des recommandations en vue de l'amélioration de la satisfaction dans les administrations publiques (Lorit et al., 2002). Ce rapport n'a pas d'aspect méthodologique ou statistique. Il rassemble quelques expériences et des recommandations. Les auteurs arrivent à la conclusion qu'aujourd'hui, les méthodes utilisées sont très variées et souvent peu efficaces. Ils préconisent la mise en place de processus d'estimation de la satisfaction des usagers mieux définis et de méthodologies plus robustes et plus uniformes au sein des services publics français. Ce document nous montre à quel point l'étude de la satisfaction des consommateurs occupe une place de premier plan dans de nombreuses entreprises.

D'autre part, l'utilisation de plus en plus fréquente de certification de qualité dans les entreprises comme ISO 9000, qui reconnaît l'efficacité de l'organisation et permet de garantir la confiance et la satisfaction des clients, montre un grand intérêt pour la satisfaction des consommateurs. Il nous a donc semblé judicieux d'aller directement poser des questions à quelques entreprises.

Bouygues Telecom est la première d'entre elles. C'est une entreprise assez jeune qui dédie beaucoup de temps et de personnel à l'étude de la satisfaction et de la fidélité. L'analyse des enquêtes de satisfaction se cantonne pour l'instant à des études statistiques de base. Les modèles d'équations structurelles ne sont pas encore envisagés. En fait, Bouygues Telecom, plutôt que de chercher les leviers de la satisfaction et de la fidélité, s'intéresse à la probabilité de départ. Ils interviennent au dernier moment lorsqu'un client est sur le point de partir (analyse du *churn*).

D'autre part, il nous a semblé intéressant de consulter les grands instituts de sondages qui proposent une démarche CRM (*Customer Relationship Management*). L'un d'entre eux a répondu à notre demande. Celui-ci prend en charge des études de satisfaction pour des clients de petite taille. L'institut met en avant son savoir-faire, c'est-à-dire l'enquête en elle-même. Il n'utilise aucune des approches présentées dans cette thèse. L'approche PLS n'est pas connue, LISREL est mal vue (à cause de ses problèmes de convergence) et les réseaux bayésiens n'en sont qu'à leurs débuts.

Cet institut ne travaille que sur les blocs de variables indépendamment, ils ne considèrent pas d'équations structurelles entre variables latentes. Ils utilisent simplement des régressions PLS et des régressions sur les axes principaux pour estimer les relations.

Les seuls points traités de façon originale sont, d'une part, la partition de la fidélité en fidélité attitudinale et fidélité comportementale et, d'autre part, la segmentation des bases de données sur des questions spécifiques afin de faire ressortir des profils d'individus.

Ces deux exemples de pratiques en matière d'utilisation des modèles d'équations structurelles nous confortent dans le fait qu'EDF est à la pointe de l'utilisation de ces méthodes. La plupart des entreprises consultées montrent un grand intérêt pour ces approches mais admettent ne pas assez les maîtriser pour les appliquer. Cette étude fait donc ressortir l'intérêt de cette thèse qui permet de mettre l'accent sur les points forts de ces méthodes.
Au niveau international, la réussite du groupe CFI (groupe créé par C. Fornell) montre l'intérêt des grandes entreprises pour l'analyse de la satisfaction de leurs clients par l'approche PLS. La méthode LISREL, du fait de son association aux domaines de la sociologie et de la psychologie, reste en retrait dans les applications industrielles.

7.2 Les problématiques et les données

La première partie de ce chapitre nous a permis d'introduire les notions marketing importantes. Nous allons maintenant vérifier la validité de ces théories et répondre aux cinq questions posées en début de chapitre par des applications. Nous utilisons tout au long de ces applications des programmes informatiques créé dans la cadre de ce travail. Pour plus de détails, on peut voir le tableau B.12 (p. 187) en annexe.

Dans la pratique, le choix de la méthode d'estimation du modèle n'est pas simple (voir chap. 2.1, spécialement le tableau 2.3, p. 42). Nous illustrons ce choix par une comparaison des méthodes d'estimations sur des données issues d'un questionnaire de satisfaction des clients d'EDF. Une fois la méthode choisie, nous nous proposons d'analyser en profondeur un jeu de données. Nous commençons par une application lorsque les connaissances associées au domaine d'application sont très faibles. Nous appliquons les méthodes de construction présentées dans le chapitre 3 (p. 57). Une fois le modèle validé, nous comparons des groupes d'observations ayant des comportements différents. Nous introduisons des relations non linéaires dans les modèles d'équations structurelles à variables latentes par le biais de l'analyse de la relation satisfaction - fidélité et de celle entre la satisfaction et ses facteurs. Nous terminons ce cas pratique par l'introduction de variables dites "filtrées" au sujet des réclamations des clients.

Dans le cadre de l'analyse de la satisfaction des clients, l'analyse de modèles d'équations structurelles à variables latentes se base sur des questionnaires de satisfaction. A l'intérieur de ces questionnaires, plusieurs types de données peuvent être trouvés :

- Des questions précises concernant des facettes de certains concepts (la satisfaction, la fidélité, l'image...). Elles sont généralement modélisées par des variables catégorielles ordonnées avec entre 5 et 10 modalités (échelles de Likert (1932)).
- Des questions d'ordre informationnelle sur les clients (variables socio-culturelles). Elles sont généralement modélisées soit par des variables binaires, soit par des variables nominales. Parfois on peut trouver des variables continues (revenu, consommation électrique...).
- Des questions dites "filtrées". Ces questions ne concernent qu'une sous-partie de l'échantillon, généralement de taille modérée, qui est concernée par un sujet spécifique.

Les jeux de données que nous utilisons contiennent surtout des variables du premier type.

Nous ne pourrons pas ici utiliser un seul jeu de données. En effet, les problématiques associées à chaque chapitre de la thèse étant très différentes, nous utilisons plusieurs jeux de données. Néanmoins, nous aurons un jeu de données principal. Il est composé de 29 variables manifestes, nous ne les détaillons pas ici pour des raisons de confidentialité. Ces variables sont réparties en six groupes unidimensionnels. Les références des variables manifestes, ainsi que leur bloc d'appartenance, sont détaillés dans le tableau 7.3. Ce jeu de données est basé sur 1988 observations et sera noté EDF-1.

Les variables latentes sont supposées réflectives et la consistance interne de chaque bloc a été vérifiée $\alpha > 0.7$.

Variable latente	Libellé des VM	Echelles
Valeur perçue	VP1, VP2	Likert 10 modalités
Satisfaction	SAT1, SAT2, SAT3, sat4, SAT5, SAT6	Likert 10 modalités
Comp. futur	CF1, CF2	Likert 10 modalités
Fidélité	Fid1, Fid2, Fid3, Fid4, Fid5	Likert 10 modalités
Réputation	Rep1, Rep2, Rep3	Likert 10 modalités
Image	I1, I2, I3, I4, I5, I6, I7, I8, I9, I10, I11	Likert 10 modalités

Tab. 7.3 – Définition des blocs de variables pour les données EDF-1

Le modèle structurel initial est décrit dans la figure 7.7. Pour des raisons de confidentialité, ce modèle n'est pas le modèle utilisé chez EDF, nous avons mis au point ce modèle en nous basant sur les théories marketing associées. L'image, la réputation et le comportement futur sont des variables exogènes, la satisfaction est expliquée par l'image et la valeur perçue, et cette dernière est elle-même expliquée par l'image. La fidélité a comme antécédent l'ensemble des variables du modèle exceptée la valeur perçue.

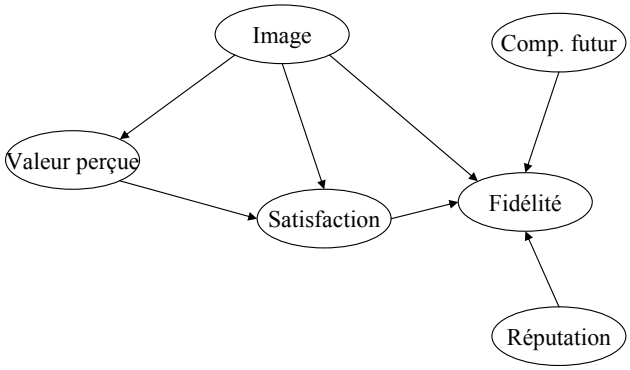

Fig. 7.7 – Modèle structurel associé aux données EDF-1

Nous rassemblons l'ensemble des jeux de données utilisés dans ces applications dans le tableau 7.4. Nous détaillons le nombre de variables manifestes (VM), le nombre d'observations (N), le type de données (LIK = échelles de Likert (1932)), l'origine des données et les domaines d'application dans la thèse. Les modèles conceptuels associés sont détaillés dans le courant des applications.

Données	VM	N	Echelles	Origine	Applications
EDF-1	29	1988	LIK	EDF	Jeu de données principal
BF	5	1510	LIK	Banques françaises	7.4.1 (p. 146)
BB	13	1256	LIK	Banques belges	7.4.1 (p. 146)
GS	7	1000	LIK	Grandes surfaces US	7.4.1 (p. 146)

Tab. 7.4 – Définition des jeux de données utilisés

7.3 Choix de la méthode de traitement : comparaison des approches PLS, LISREL et GSCA

Le choix d'une méthode d'estimation du modèle pourra se faire différemment en fonction du but recherché comme nous l'avons vu dans le chapitre 2. Nous nous plaçons dans un contexte général afin de commencer et détaillerons ensuite les problématiques possibles.

Nous utilisons le jeu de données EDF-1 pour les méthodes PLS, LISREL et GSCA. Une fois le jeu de données sélectionné, la problématique recherchée devra nous orienter vers l'une des méthode présentées plus tôt (chapitres 1 et 2). Ainsi, nous appliquons 3 méthodes : l'approche PLS classique avec le mode A et le schéma centroïde, la méthode LISREL-ML par estimation de la matrice de covariance en se basant sur le maximum de vraisemblance, et l'approche GSCA avec le cas réflectif. D'autres méthodes sont possibles comme l'estimation par LISREL et la fonction ULS mais nous nous limitons d'abord à ces trois approches.
Nous nous servons de trois programmes distincts :

- Pour l'approche PLS : une macro SAS (SAS Institute Inc., 2004a) mise au point dans le cadre de cette thèse (tab. B.12, p. 187) et le logiciel SmartPLS (Ringle et al. (2005), tab. B.13, p. 187).
- Pour la méthode LISREL : les logiciels LISREL 8.57 (Jöreskog et Sörbom, 1996) et AMOS 5 (Arbuckle, 2005).
- Pour l'approche GSCA : le logiciel VisualGSCA 1.0 récemment développé par Hwang (2007).

Nous utilisons le modèle associé aux données EDF-1 (cf. fig. 7.7).
Dans le tableau 7.3, nous avons rassemblé les *loadings* et les coefficients structurels obtenus pour les 3 approches. Nous voyons que les estimations des paramètres sont très proches pour les méthodes PLS et GSCA. Ceci s'explique par la proximité de principe entre les deux approches, les coefficients sont estimés par moindres carrés en supposant que les variables latentes sont des combinaisons linéaires des variables manifestes qui leur sont associées. Par ailleurs, les *loadings* associés à l'approche LISREL sont plus faibles que ceux associés à PLS et GSCA. Ceci s'explique par le fait qu'ils sont calculés comme des régressions simples pour PLS et que PLS se focalise sur le modèle externe. Inversement, les coefficients structurels sont plus élevés pour le cas de LISREL qui a tendance à favoriser le modèle interne. Néanmoins, une cohérence générale entre ces approches semble exister pour les estimations des paramètres.
Les indices de qualité du modèle étant différents d'une approche à l'autre, nous ne pouvons pas rassembler les différents indices sous la forme d'un tableau. Néanmoins, nous regroupons dans le tableau 7.6, les R^2 associés à chaque variable latente endogène pour les 3 approches. Il apparaît que ceux associés à l'approche LISREL sont les plus élevés. Ceci vient du fait que, dans le cadre de l'approche LISREL, les variables latentes utilisées pour le calcul des R^2 sont les variables latentes théoriques. Les R^2 sont donc plus élevés que dans le cas de PLS ou de GSCA où les variables latentes sont calculées à partir des variables manifestes qui leur sont associées. De même les coefficients structurels obtenus par la méthode LISREL sont obtenus à partir des variables latentes théoriques (non calculées au niveau de chaque observation).

Ces 3 approches étant très différentes, nous détaillons les résultats pour chacune des méthodes avec les critères spécifiques.

Pour PLS, les principaux résultats sont rassemblés dans le tableau 7.7 où AVE (*Average Variance Extracted*) est la moyenne de la variance expliquée.
Pour GSCA, on obtient $FIT = 0.568$, $AFIT = 0.572$, $GFI = 0.991$ (*Goodness of Fit* défini p. 32 avec comme matrice $\mathbf{W}$ la matrice de covariance estimée à partir du modèle) et $SRMR = 0.107$ (*Standardized Root Mean Square Residual*, cet indice est défini par la différence standardisée entre les covariances observées et les covariances prédites par le modèle, il est souvent associé à la méthode LISREL). Il ressort que la méthode permet d'extraire plus de 55% de la variance et obtient un GFI au-dessus du seuil généralement utilisé. Ces remarques tendent à montrer que la qualité d'ajustement est bonne dans le cas de la méthode GSCA.

loading	LISREL	PLS	GSCA
VP1	0.779	0.893	0.892
VP2	0.763	0.893	0.892
SAT1	0.427	0.530	0.566
SAT2	0.479	0.587	0.617
SAT3	0.658	0.729	0.727
SAT4	0.714	0.784	0.780
SAT5	0.766	0.820	0.809
SAT6	0.819	0.834	0.814
CF1	0.651	0.808	0.865
CF2	0.761	0.912	0.865
Fid1	0.779	0.829	0.802
Fid2	0.582	0.698	0.735
Fid3	0.767	0.844	0.839
Fid4	0.588	0.708	0.726
Fid5	0.648	0.752	0.744
Rep1	0.607	0.720	0.775
Rep2	0.683	0.826	0.784
Rep3	0.448	0.677	0.671
I1	0.802	0.820	0.822
I2	0.776	0.798	0.790
I3	0.788	0.812	0.810
I4	0.514	0.568	0.570
I5	0.687	0.727	0.734
I6	0.597	0.649	0.658
I7	0.709	0.743	0.746
I8	0.700	0.733	0.740
I9	0.592	0.640	0.639
I10	0.789	0.808	0.809
I11	0.843	0.846	0.839
Coefficients structurels	LISREL	PLS	GSCA
Image → Satisfaction	0.42	0.474	0.469
Image → Valeur perçue	0.64	0.503	0.525
Image → Fidélité	0.37	0.397	0.402
Valeur perçue → Satisfaction	0.63	0.454	0.449
Satisfaction → Fidélité	0.25	0.187	0.173
Réputation → Fidélité	0.09	0.076	0.067
Comp. futur → Fidélité	-0.35	-0.294	-0.291

TAB. 7.5 – *Loadings* et coefficients structurels obtenus par les approches PLS, LISREL et GSCA sur les données EDF-1

R^2	LISREL	PLS	GSCA
Fidélité	0.617	0.486	0.388
Valeur perçue	0.408	0.281	0.262
Satisfaction	0.906	0.659	0.632

TAB. 7.6 – Comparaison des R^2 associés aux variables latentes endogènes pour les approches PLS, LISREL et GSCA sur les données EDF-1

	AVE	Composite Reliability	R^2	H^2	F^2
Comp. futur	0.743	0.852	-	0.663	-
Fidélité	0.591	0.877	0.486	0.827	0.085
Image	0.556	0.931	-	0.918	-
Réputation	0.553	0.786	-	0.598	-
VP	0.797	0.887	0.281	0.746	0.224
Satisfaction	0.523	0.865	0.659	0.814	0.236

TAB. 7.7 – Principaux indices associés à l'approche PLS pour les données EDF-1

Pour l'approche LISREL, nous avons commencé par une analyse basée sur le maximum de vraisemblance. Dans ce cas, le modèle est rejeté par le test du χ^2 ($\chi^2/dl = 4.7$ et $p - valeur = 0.00$) même si les indices relatifs sont très bons $NFI = 0.98$, $CFI = 0.98$, $RMSEA = 0.05$ et $GFI = 0.92$. Ceci peut s'expliquer par la déviation de la distribution normale multivariée. Nous utilisons donc la méthode d'estimation ULS. Cette méthode ne dépend pas de la distribution des données, mais les indices de qualité obtenus en dépendent. Les résultats sont donc similaires au cas du maximum de vraisemblance. L'utilisation de la méthode développée dans le chapitre 2 par McDonald (1996) et Tenenhaus (2007) constituerait une alternative intéressante mais ne sera pas intégrée ici.

Si nous nous intéressons aux variables latentes, nous avons rassemblé dans les figures 7.8 et 7.9 les relations entre les scores des variables latentes pour les 3 approches sur la satisfaction et la fidélité. Nous utilisons les scores classiques pour PLS, les scores issus de l'équation 1.30 (p. 37) pour l'approche LISREL (méthode de Tenenhaus et al. (2005)) et les scores issus de l'équation A.3 (p. 175) pour l'approche GSCA. Les scores obtenus sont très proches et montrent une cohérence entre les méthodes d'estimation.

Malgré les mauvaises qualités d'ajustement (χ^2) obtenues par la méthode LISREL, qui peuvent être expliquées par les propriétés spécifiques associées aux données (non normalité), et au vu des autres indices de qualité, on peut donc considérer que le modèle s'adapte bien aux données et l'interprétation est alors possible. Ainsi la fidélité est assez mal expliquée par les variables du modèle interne ($R^2 = 0.48$). Néanmoins, c'est l'image qui a le plus fort impact sur l'intention de fidélité. Par ailleurs, l'apport de la réputation sur la fidélité n'est pas significatif et le lien entre satisfaction et fidélité est assez faible (cf. tab. 7.3). Nous étudierons ce lien plus en détails dans le cadre des études sur la non linéarité.

L'utilisation d'une méthode plutôt qu'une autre se justifiera donc plus par des principes liés à la problématique que par des problèmes liés aux estimations. Ainsi, la méthode LISREL devra être utilisée lorsque les connaissances sur la population sont fortes et que l'on désire des estimations des paramètres non biaisés. De plus, elle permet d'obtenir des indices de qualité d'ajustement extrêmement utiles dans un processus de validation.
Si on travaille avec peu d'hypothèses sur des modèles complexes, que l'on veut étudier les variables latentes et des indices de reconstruction, on préférera l'approche PLS.
La méthode GSCA offre une alternative intéressante à l'approche PLS avec une fonction globale à optimiser.

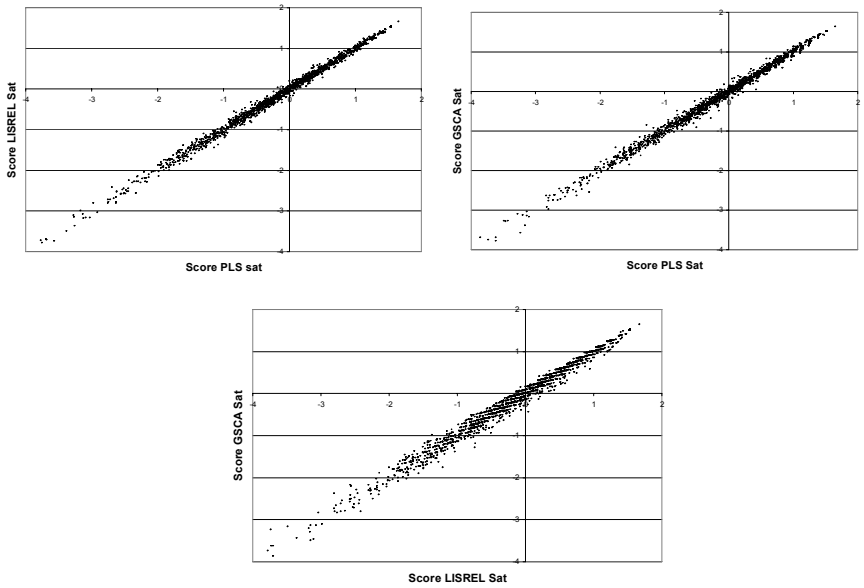

FIG. 7.8 – Comparaison des scores de la variable latente satisfaction

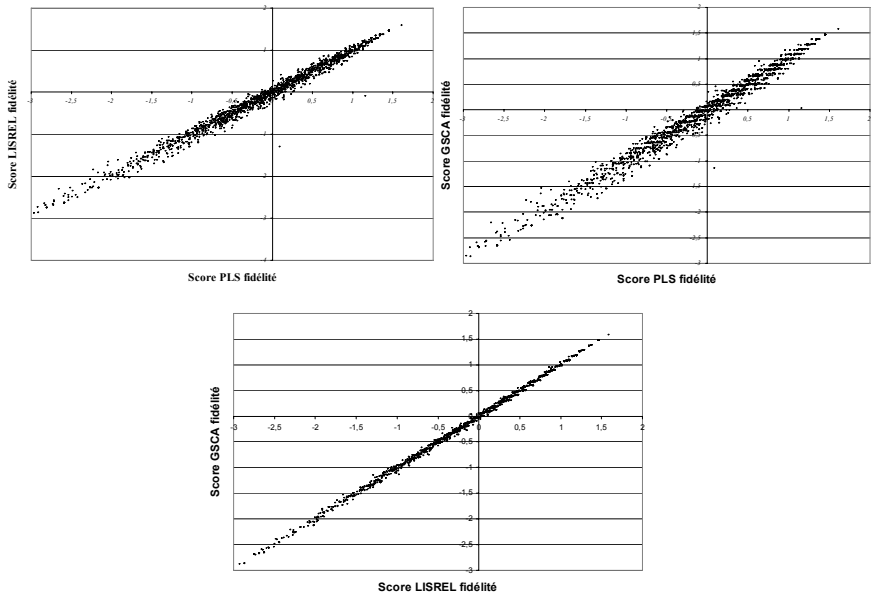

FIG. 7.9 – Comparaison des scores de la variable latente fidélité

Dans le cadre de cette application, d'un point de vue empirique, c'est l'approche PLS qui semble la plus adaptée. D'autres méthodes pourraient être appliquées telle que, par exemple, l'estimation ULS avec des variables composites (voir chap. 2.2, p. 42, McDonald (1996)).

7.4 Les modèles de satisfaction : du dire d'expert à la construction du modèle

Il est reconnu que les modèles structurels à variables latentes sont des méthodes servant à confirmer une théorie existante. Néanmoins, nous avons vu que quand le contexte est complexe, la connaissance des experts pour la construction du modèle n'est pas suffisante. Il faut alors s'aider des données afin de définir le modèle. Dans le cadre d'EDF, l'ouverture du marché de l'électricité pose des questions sur la construction et les modifications du modèle conceptuel marketing. De plus, les variables spécifiques au système français et les attitudes particulières des français dans leur consommation font que l'utilisation de modèles venant de pays ayant un marché ouvert est difficile (d'autant que le Royaume-Uni ou les pays nordiques ont eux-mêmes des évolutions différentes, cf. chap. 7.1, p. 129). Nous traitons ici les 3 points développés dans le chapitre 3 (p. 57). Nous commençons par étudier des modèles mal définis en terme de relations réflectives et formatives, pour ceci nous utiliserons 3 jeux de données issus, pour le premier, d'une étude sur les grandes surfaces aux Etats-Unis et, pour les deux suivants, de questionnaires de satisfaction sur les banques, respectivement en France et en Belgique.

Nous revenons ensuite à des données EDF, afin d'appliquer la méthode de construction du modèle externe basée sur les réseaux bayésiens et comparons les résultats obtenus aux autres approches.

Finalement, nous utilisons les données EDF-1 pour construire le modèle interne en maximisant les corrélations partielles. De la même façon, nous comparons les résultats aux méthodes classiques.

7.4.1 Réflectif vs. formatif

Dans le cadre des modèles d'équations structurelles, le choix entre les schémas réflectif et formatif est difficile et une erreur dans cette décision entraîne bien souvent des biais d'estimation importants. Ainsi, le modèle réflectif est généralement choisi et pourtant il est rarement adapté. Nous utilisons des données de questionnaires de satisfaction afin de vérifier cette hypothèse et appliquons notre algorithme (cf. chap. 3.1, p. 57) afin d'obtenir un modèle réflectif. Nous traitons chaque jeu de données séparément.

Données GS

Nos données sont issues d'un questionnaire de satisfaction sur les grandes surfaces américaines. Ce jeu de données comporte 1000 individus. Nous nous intéressons à la variable latente satisfaction des clients. Les questions sur ce construit sont mesurées sur une échelle de 1 à 10 (1 signifiant "pas du tout d'accord" à 10 signifiant "tout à fait d'accord"). Les experts ont considéré que le modèle de mesure pour cette variable est de type réflectif. Nous utilisons d'abord le test des tétrades pour voir si nous obtenons le même résultat. Les variables associées à ce concept sont rassemblées en annexe (tableau B.2, p. 180). Si on utilise le test des tétrades sur le bloc complet, le modèle réflectif est rejeté, ceci nous amène donc à appliquer notre algorithme pour voir quels sont les items qui posent des problèmes.

Modèle obtenu à partir des variables concernant la satisfaction : sur les 7 variables du bloc, l'algorithme en conserve seulement 4, afin d'obtenir un modèle réflectif (V7 : the customer service you receive ; V9 : knowledgeable employees ; V10 : availability of employees to assist you et V11 : convenient locations). L'algorithme a détecté trois variables qui ne s'intègrent pas dans le construit réflectif satisfaction : V1 (overall satisfaction), V6 (the range of products available) et V8 (the physical store

environment). Le modèle avec les 4 variables obtient pour 2 degrés de liberté, un χ^2 de 1.568 et une p-valeur de 0.45. La consistance interne de ce bloc est vérifiée, l'α de Cronbach est de 0.84.

Comment peut-on expliquer le fait que les trois items ne s'intègrent pas dans le construit réflectif satisfaction ?
La satisfaction globale est un item couramment utilisé dans les échelles de mesures de la satisfaction. Cependant, dans notre cas, "overall satisfaction" pose problème et est une cause du rejet de l'hypothèse associée au test des tétrades. On ne peut pas considérer cet item comme formatif, car cette affirmation va à l'encontre de la théorie marketing. Mais d'autre part, comme on peut le constater, celui-ci est un item global. Selon Westbrook et Oliver (1982), *this measure is a highly generalized satisfaction indicator*. Le résultat de l'algorithme pose donc la question de la pertinence de l'utilisation d'une question aussi globale dans un construit complexe comme la satisfaction.
Quant aux deux autres items, ce sont des items qui caractérisent plutôt l'image du point de vente (Lindquist, 1974).

Données BF et BB

Nous disposons de deux jeux de données concernant la satisfaction dans le secteur bancaire. Le premier concerne des banques françaises (1510 observations) et le second des banques belges (1256 observations). Nous allons étudier le construit image (pour les deux jeux de données) et la qualité perçue (seulement pour les banques belges). Chaque construit image est composé de 5 variables manifestes et la qualité perçue de 8 (voir annexe, tableau B.2, p. 180). Les items sont mesurés sur une échelle de 1 à 10 (1 signifiant "pas du tout d'accord" et 10 "tout à fait d'accord").

Les experts ont considéré que les 3 construits suivent un schéma réflectif. Comme pour l'exemple antérieur, nous utilisons d'abord le test des tétrades pour voir si nous obtenons ainsi le même résultat. Si l'hypothèse de réflectivité est rejetée, alors nous appliquons notre algorithme afin de détecter les items qui ne s'intègrent pas dans leur construit réflectif.

- **Le construit image :**
  - *Modèle obtenu sur les données BF :* sur les données complètes, le modèle à 5 variables est réflectif. L'algorithme converge donc à la première étape avec 5 degrés de liberté (dl), un χ^2 de 9.153, une p-valeur de 0.103 et un α de 0.79.
  - *Modèle obtenu sur les données BB :* le modèle à 5 variables n'est pas réflectif ($\chi^2 = 16,097$, $df = 5$, $p - val = 0.006$). L'algorithme a détecté que la mesure qui ne s'intègre pas dans ce construit est "la banque apporte une contribution sociale à la société".
    Pour le modèle composé par les 4 variables, on a alors $dl = 2$, $\chi^2 = 1.087$, une $p - val = 0.580$ et un α de 0.84.
- **Le construit qualité perçue :**
  - *Modèle obtenu sur les données BB :* Le test des tétrades permet de rejeter l'hypothèse de réflectivité sur le modèle complet. L'algorithme trouve que la mesure qui ne s'intègre pas dans le construit analysé est "comment jugeriez-vous X par rapport au service aux clients et aux conseils personnels donnés". Le modèle composé par les 7 autres variables est donc réflectif avec 9 degrés de liberté, $\chi^2 = 12.7$, une $p - val = 0.177$ et un α de 0.88.

Dans un second temps, nous mettons en place un processus de validation de l'algorithme proposé sur les données BB et BF. Dans ce but, les données de chaque construit analysé (image et qualité perçue) sont divisées en 3 parties de tailles similaires. Sur les deux premiers sous-échantillons, l'algorithme est appliqué et sur le troisième sous-échantillon, les α de Cronbach des blocs définis lors des deux premières applications de l'algorithme sont estimés. Les résultats sont rassemblés dans le tableau 7.8.

Sur les trois sous-échantillons, il ressort que, pour chacun des construits analysés, le modèle réflectif est toujours sélectionné. Pour chacun des concepts, le modèle obtenu avec les deux sous-échantillons testés est le même que celui obtenu sur l'échantillon complet. Comme on peut le constater, les alphas de

Modèle - Echantillon	Nb. VM	χ^2	p-valeur	α
Image BF				
Echantillon complet	5	9.153	0.103	0.79
Echantillon 1	5	6.04	0.302	0.83
Echantillon 2	5	4.74	0.449	0.81
Echantillon 3	5	-	-	0.79
Image BB				
Echantillon complet	4	1.087	0.580	0.84
Echantillon 1	4	0.196	0.923	0.83
Echantillon 2	4	0.214	0.898	0.84
Echantillon 3	4	-	-	0.82
Qualité perçue BB				
Echantillon complet	7	12.70	0.177	0.88
Echantillon 1	7	6.83	0.654	0.87
Echantillon 2	7	7.16	0.620	0.89
Echantillon 3	7	-	-	0.88

TAB. 7.8 – Processus de validation de l'algorithme

Cronbach sont comparables entre les 3 sous-échantillons. Les χ^2 très faibles pour les sous-échantillons 1 et 2 s'expliquent par le nombre d'observations qui change de l'échantillon original aux sous-échantillons (le résultat du test du χ^2 est sensible à la taille de l'échantillon, mais celui du test des tétrades l'est beaucoup moins). On peut donc affirmer, au moins sur nos exemples, que l'algorithme proposé est "stable".

Ces applications nous dévoilent à quel point il est difficile de définir le modèle externe, nous avons montré que notre algorithme obtient des résultats cohérents dans des cas réels et que les résultats obtenus sont interprétables. Cette étape de la construction du modèle est importante lorsque les blocs sont connus et afin de bien appliquer les méthodes d'estimation du modèle. Nous allons maintenant aborder le cas où le modèle externe est totalement inconnu.

7.4.2 Le modèle externe : une construction complexe

Comme nous l'avons vu dans le chapitre 3.2 (p. 58), la construction du modèle de mesure est primordiale dans le traitement des modèles d'équations structurelles à variables latentes. Dans le cadre de l'approche PLS, c'est celui-ci qui revêt la plus grande importance dans l'estimation du modèle. Il est donc nécessaire, lorsque les informations sont peu nombreuses d'utiliser une stratégie de construction efficace et compréhensible. Il s'agit ici d'associer des variables observées de façon à ce qu'elles forment des groupes unidimensionnels et cohérents.

La construction à l'aide des réseaux bayésiens

Nous commençons par appliquer l'algorithme basé sur les réseaux bayésiens introduit dans le chapitre 3.3.4 (p. 71). Nous utilisons le jeu de données EDF-1, les variables associées à ces données sont ordinales à 10 modalités. Nous détaillons l'application de l'algorithme basé sur les réseaux bayésiens. Dans la figure 7.10, le réseau bayésien obtenu à partir du logiciel BayesiaLab (2005) est présenté. La figure 7.11 illustre les étapes de l'algorithme par le biais d'un arbre de classification. En maximisant l'unidimensionnalité des blocs, on obtient une découpe de l'arbre équivalent à un seuil sur la distance de Kullback et Leibler (1951) de 0.3.
Trois différences notables existent avec le modèle expert :

– On obtient 5 blocs au lieu de 6.

– Les blocs satisfaction et valeur perçue sont fusionnés. Ceci se justifie par la proximité entre ces deux concepts et par la formulation des questions dans le questionnaire.
– La variable manifeste I4 est exclue du modèle. Aucune information ne peut a priori expliquer cette séparation.

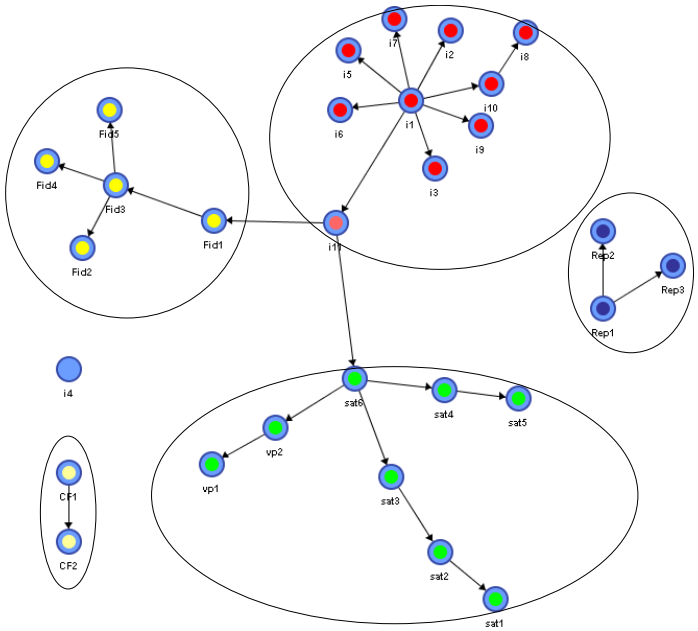

FIG. 7.10 – Réseau bayésien obtenu avec les groupes de variables associées

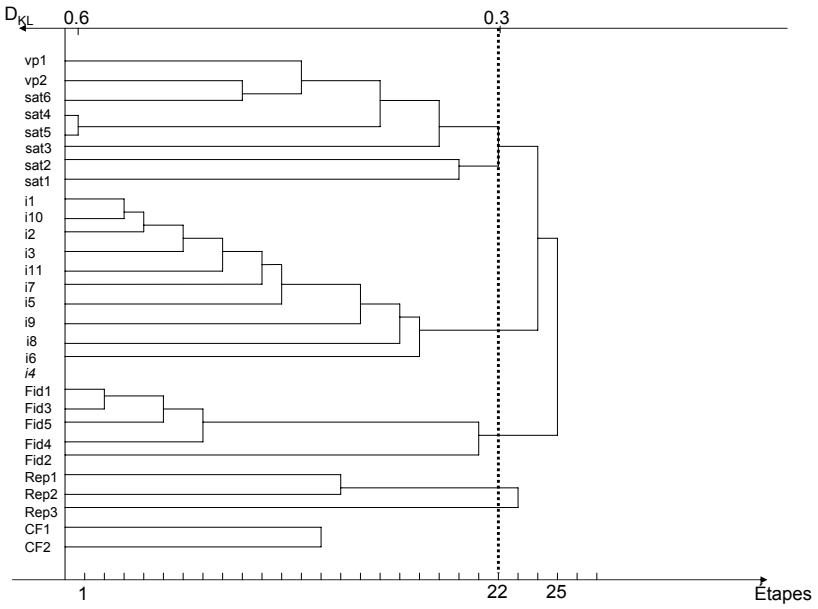

FIG. 7.11 – Arbre de classification complet pour la méthode basée sur les réseaux bayésiens

Comparaisons

Nous appliquons sur les données EDF-1 l'ensemble des méthodes de construction et rassemblons le nombre de variables manifestes prises en compte, le nombre de blocs obtenus et l'α de Cronbach moyen dans le tableau 7.9. Un indice de dissimilarité simple entre la méthode étudiée et le modèle expert est ajouté. Celui-ci est défini par :

$$D_{exp} = \frac{1}{P} \sum_{i=1}^{P_r} F_{x_i} \tag{7.1}$$

avec P nombre total de variables, P_r nombre de variables sélectionnées par la méthode de construction et $F(x_i) = 1$ si la variable x_i se trouve dans le même groupe dans les deux partitions (experte et obtenue par la méthode).

Les méthodes sont donc :
- la procédure VARCLUS de SAS Institute Inc. (2004b),
- l'algorithme de classification autour de composantes latentes (CCL) (nous utilisons des macros SAS développées par Sahmer (2006) ainsi qu'une méthode de permutation afin de connaître le nombre de blocs),
- la méthode de Stan et Saporta (2005) (on peut trouver l'arbre de classification en annexe dans la figure B.1 (p. 182)),
- la méthode basée sur le logiciel TETRAD,
- l'algorithme basé sur l'apprentissage de réseaux bayésiens (RBappr).

Méthode	Nb. de VM	Nb. de blocs	$\bar{\alpha}$	D_{exp}
Expert	29	6	0.800	1
VARCLUS	29	5	0.820	0.931
CCL	29	5	0.820	0.931
Stan et Saporta (2005)	29	5	0.786	0.931
TETRAD	20	4	0.766	0.690
RBappr	28	5	0.821	0.897

TAB. 7.9 – Construction du modèle de mesure

Nous avons rassemblé l'ensemble des partitions en annexe (tab. B.4, p. 181).

La méthode basée sur les réseaux bayésiens apparaît comme une alternative intéressante. Nous n'avons pas utilisé la possibilité d'ajouter des contraintes mais celles-ci peuvent être mises en place simplement. De plus, lorsque les variables sont nominales ou binaires, cette approche a d'autant plus de sens. Pour un exemple dans ce cadre, on peut voir Jakobowicz et Derquenne (2007). Par ailleurs, la stabilité de l'apprentissage du réseau et de l'algorithme a été vérifiée par des simulations dans Jakobowicz (2006a).

Du point de vue de l'application, toutes les méthodes arrivent à des résultats proches de ceux des experts, ce qui tend à montrer que la connaissance de ceux-ci et l'information fournie par les données concordent. Cependant, quelques différences existent. La valeur perçue est souvent intégrée dans le construit satisfaction. Le nombre de blocs sélectionnés est toujours plus faible que dans le cas expert. Les modèles obtenus à partir des 3 premières méthodes sont exactement similaires, l'algorithme que nous présentons obtient l'α moyen le plus élevé. La méthode basée sur l'algorithme TETRAD produit des résultats largement différents. Ceci s'explique par le plus grand nombre d'hypothèses initiales qui sont souvent rarement vérifiées dans les cas pratiques (normalité des données, corrélation nulle entre les variables manifestes associées à des blocs différents...).

L'approche que nous proposons (construction basée sur l'apprentissage des réseaux bayésiens) apparaît donc comme une alternative intéressante aux méthodes classiques avec des résultats similaires et des possibilités de visualisation et d'ajout de contraintes qui la rendent plus compréhensible et interactive.

7.4.3 Le modèle interne : la satisfaction, concept central

Une fois le modèle externe défini, il est important de bien définir le modèle interne, celui-ci contient les relations de causalité entre les concepts et est donc très important. Nous travaillons sur les données EDF-1.

Nous avons présenté une approche basée sur les corrélations partielles tout en nous appuyant sur le processus PLS. Nous nommons cette méthode "modèles libres itératifs". Elle permet de prendre en compte le modèle interne dans l'estimation des relations à intégrer.
Nous détaillons donc le processus de construction du modèle. Dans le tableau 7.10, nous rassemblons les corrélations partielles maximales à chaque étape de l'algorithme et la relation associée jusqu'à obtenir un modèle stable (en vérifiant la significativité des relations).

Etapes	Relation	Corr. part.
1	Valeur perçue - Satisfaction	0.453
2	Image - Satisfaction	0.315
3	Image - Fidélité	0.359
4	Comp. Futur - Fidélité	0.292
5	Réputation - Image	0.195
6	Satisfaction - Fidélité	0.101

TAB. 7.10 – Etapes de la construction du modèle interne par les modèles libres itératifs

Le modèle final peut être observé dans la figure 7.12 (modèle libre itératif). La relation image - réputation est ajoutée, ce qui d'un point de vue marketing semble logique et la relation réputation - fidélité est exclue (au profit de la précédente). De plus, la relation image - valeur perçue n'existe pas. Dans les étapes de l'algorithme, la relation satisfaction - fidélité n'est ajoutée qu'à la dernière étape (avec une significativité $t = 3.01$) avant la relation réputation - image qui est non significative et qui entraîne l'arrêt de l'algorithme.

Les orientations sont proches de celles du modèle expert. La méthode d'orientation des arcs pousse à favoriser l'obtention de lien allant vers des variables latentes endogènes. Ceci permet d'augmenter le R^2 de ces variables et le R^2 global, alors que de transformer une variable exogène en variable endogène au cours de l'algorithme entraîne généralement une baisse du R^2 moyen. L'orientation obtenue doit être validée à partir des connaissances dans le domaine d'application.

Comparaisons

Nous comparons les approches présentées dans le chapitre 3.3 (p. 66). Ainsi nous appliquons :
- la méthode de Hui (1982),
- l'algorithme de Hackl (2003),
- la méthode d'Amato en se basant sur l'indice F^2 (Amato, 2003),
- les modèles libres avec une probabilité de 5% (Derquenne et Hallais, 2004),
- les modèles libres itératifs.

Quelques hypothèses supplémentaires sont nécessaires. Nous utilisons les blocs définis dans le tableau 7.3 (p. 141) et définissons pour les 2 premières approches les relations de causalité entre les

variables latentes.
Pour la méthode de Hui (1982), les variables latentes endogènes seront : la satisfaction, la fidélité et la valeur perçue (en se basant sur l'avis des experts). Pour la méthode de Hackl (2003), le choix est plus difficile car il demande une orientation générale de la causalité. On prend : Image → Valeur perçue → Réputation → Satisfaction → Comportement futur → Fidélité. On se rend compte que cette orientation a un impact important sur la modélisation du modèle interne, et une forte connaissance du domaine d'application est nécessaire. Dans le cadre des 3 autres méthodes, l'orientation se fait automatiquement, mais sans hypothèses causales. La figure 7.12 rassemble les modèles obtenus pour les 5 approches et le modèle expert. Dans le tableau 7.11, les indices moyens pour chaque modèle sont regroupés.

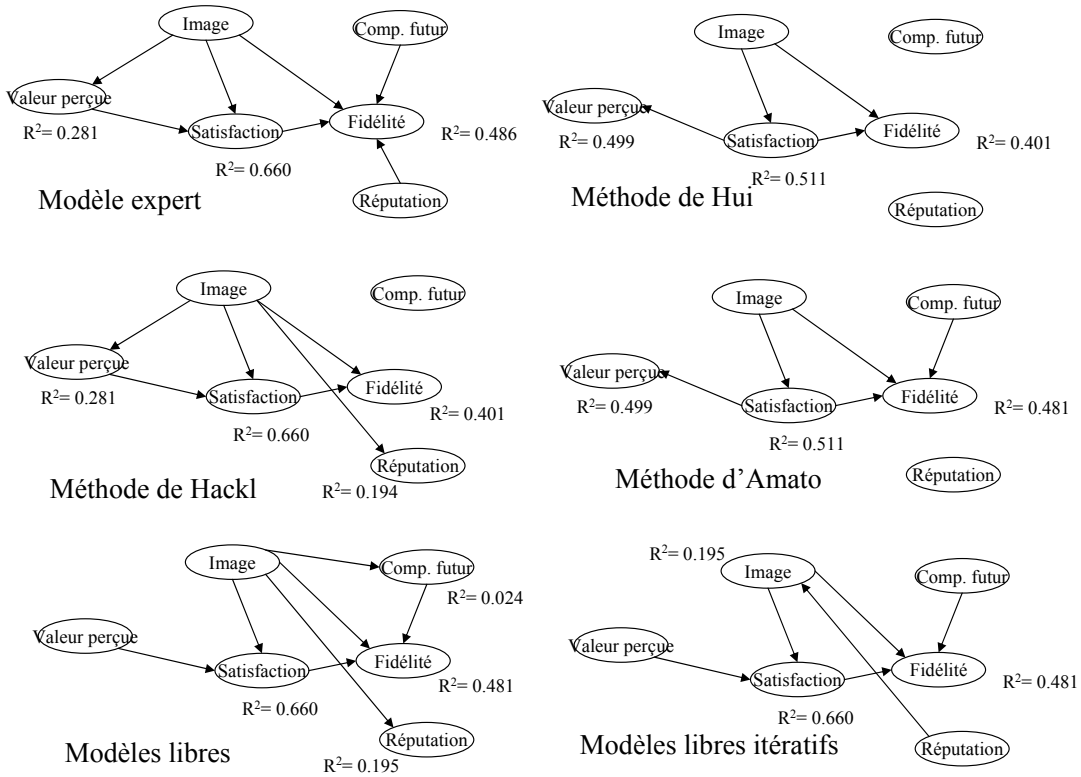

FIG. 7.12 – Modèles internes obtenus pour chaque méthode de construction

Méthode	$\bar{R}^2$	$\bar{F}^2$	GoF
Experts	0.475	0.181	0.546
Méthode de Hui	0.471	0.169	0.538
Méthode de Hackl	0.383	0.191	0.555
Méthode d'Amato	0.497	0.285	0.565
Modèles libres	0.340	0.137	0.461
Mod. libres itératifs	0.448	0.143	0.528

TAB. 7.11 – Construction du modèle interne

L'application de la méthode de Hui a nécessité l'interdiction des cycles. Par exemple, on obtenait un cycle entre image - satisfaction - valeur perçue qui a dû être modifié afin de pouvoir appliquer l'approche PLS.

Si on étudie les différences entre les modèles, on se rend compte qu'un certain nombre de relations non orientées sont communes à tous : image - satisfaction, valeur perçue - satisfaction, image - fidélité et satisfaction - fidélité.
La méthode d'Amato donne les meilleurs résultats en terme d'indices de qualité prédictive. Ceci semble cohérent car la construction même du modèle se base sur ces indices. Les modèles libres posent un problème de relations à inclure. Ainsi, ceux-ci incluent une relation non significative dans le modèle. La probabilité fixée lors de l'application du test n'a pas de relation directe avec le modèle en lui-même. Au niveau des orientations des relations, on voit que les 3 approches basées sur des méthodes automatiques d'orientation donnent des résultats souvent différents de ceux des experts. Les résultats obtenus n'ont pas de valeur de causalité et lorsqu'une information sur celle-ci existe, il faut l'utiliser en priorité. D'un point de vue marketing, la satisfaction apparaît comme un concept central et la fidélité comme une variable "cible" endogène.

L'approche que nous proposons (modèles libres itératifs) donne de meilleurs résultats sur les indices globaux que les modèles libres. De plus, elle utilise un critère d'arrêt associé au modèle qui donne une cohérence à l'approche dans le cadre de la méthode PLS. Cette approche permet de travailler dans un contexte plus large que les méthodes de Hui et de Hackl en utilisant les corrélations partielles tout en étant plus liée aux principes de l'approche PLS que les modèles libres (par l'utilisation des scores PLS et d'un critère d'arrêt basé sur la significativité des liens). Par ailleurs, son utilisation est simple (grâce à une macro SAS mise en place, tab ; B.12, p. 187).

7.4.4 Conclusion

Les méthodes présentées dans le chapitre 3 donnent donc de bons résultats, que ce soit pour le choix d'une orientation du modèle externe, pour la construction du modèle externe, ou pour la création du modèle interne. Les comparaisons effectuées nous ont montré que l'ensemble des résultats était cohérent et que, dans le cadre des données EDF, l'information issue des données coïncide avec le modèle provenant des connaissances marketing sur le sujet.

7.5 La relation face à l'ouverture du marché : comparaison de groupes dans le cadre de l'approche PLS

L'application des comparaisons présentées dans le chapitre 4 (p. 81) se fait, dans le cadre d'EDF, en supplément des méthodes de discrimination de classes d'individus (Squillacciotti, 2007). La nécessité de comparer des groupes d'observations est évidente dans le cadre de l'analyse de la satisfaction. Dans une optique marketing, le but de l'entreprise est d'augmenter la satisfaction de ses clients mais surtout d'augmenter leur fidélité. Ainsi, l'utilisation de campagnes ciblées sur des sous-populations s'avère nécessaire (on peut rarement supposer qu'un effet s'applique à la population complète, il sera plus modéré dans certains groupes que dans d'autres).
Nous présentons une application basée sur les données EDF-1 et le modèle simplifié de la figure 7.13. Elle se base sur deux variables de différenciation des groupes : l'une classique qui a rarement un impact, le sexe, l'autre plus spécifique à l'analyse de la satisfaction chez EDF, le rapport à l'ouverture du marché.

Dans le cadre de l'analyse de la satisfaction et de la fidélité des clients, la connaissance de l'impact de la classe d'appartenance d'un individu est primordiale. Celui-ci doit être mesuré à plusieurs niveaux. Nous nous attachons à trois niveaux :

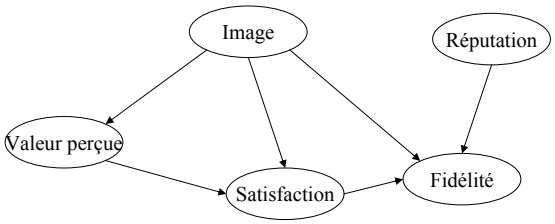

FIG. 7.13 – Modèle interne simplifié associé aux données EDF-1

– le modèle global,
– les variables latentes satisfaction et fidélité,
– la relation entre satisfaction et fidélité.

Nous illustrons ces différents points dans les figures 7.14 et 7.15. A première vue, les différences sont plus marquées entre les individus favorables ou défavorables à l'ouverture du marché qu'entre les hommes et les femmes. Néanmoins, il est nécessaire de valider la significativité de ces différences.

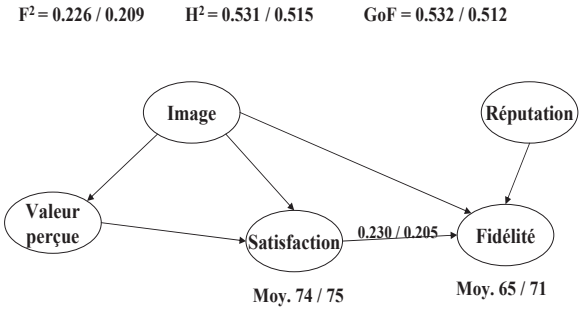

FIG. 7.14 – Modèle interne avec les différences entre hommes et femmes

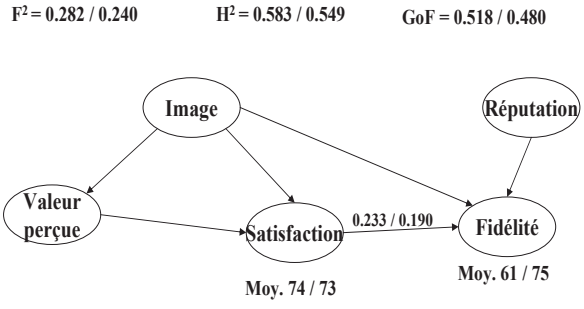

FIG. 7.15 – Modèle interne avec les différences entre favorable et défavorable à l'ouverture du marché de l'électricité

7.5.1 Comparaison globale

Nous rassemblons les résultats des tests sur 3 indices globaux ($\bar{H}^2$, $\bar{F}^2$ et GoF) dans le tableau 7.12. On utilise la statistique $s = |GF_{G_1} - GF_{G_2}|$ et un nombre de permutations $N_{perm} = 1000$.

Groupes	H/F	Ouverture
Le modèle externe		
Différence des $\bar{H}^2$	Non significative ($P = 0.191$)	Significative ($P = 0.019$)
Le modèle interne		
Différence des $\bar{F}^2$	Non significative ($P = 0.226$)	Non significative ($P = 0.075$)
La qualité globale		
Différence des GoF	Non significative ($P = 0.479$)	Non significative ($P = 0.062$)

Tab. 7.12 – Résultats de la comparaison globale

On voit qu'il n'apparaît pas de différences significatives entre les hommes et les femmes, on peut donc s'attendre à ce que cette variable ne soit pas discriminante. Nous étudions par la suite les autres niveaux de comparaison.

En ce qui concerne le sentiment par rapport à l'ouverture, le degré de significativité de la différence interprété par la probabilité est significatif pour la communauté (< 0.05) et proche de la significativité pour les deux autres indices. Dans ce modèle, on peut s'attendre à des différences aux autres niveaux de la comparaison. Sur cette variable, nous testons donc l'effet de la reconstruction du modèle.

7.5.2 Reconstruction du modèle

Dans le cadre de l'analyse du comportement face à l'ouverture du marché, nous avons vu que la qualité globale du modèle externe varie en fonction du groupe choisi. Nous appliquons donc une méthode de reconstruction du modèle externe sur l'échantillon complet en fixant le nombre de variables latentes à 6 (méthode de classification autour de composantes latentes, voir p. 67, Vigneau et Qannari (2003)). Le modèle obtenu est exactement similaire au modèle expert, il semble donc difficile de trouver un modèle mieux adapté. Nous appliquons donc le modèle expert, tout en sachant que des différences risquent d'apparaître entre les deux groupes d'individus au niveau du modèle externe.

7.5.3 Comparaison des variables latentes

Les scores des variables latentes obtenus par l'approche PLS sur chacun des groupes sont utilisés. Nous nous intéressons dans les deux cas aux variables latentes satisfaction et fidélité. Les moyennes des variables latentes sont comparées en les transformant de façon à ce qu'elles représentent un score sur 100. Les tests appliqués sont basés sur des permutations. Soit $\bar{\xi}_k^{100(G_1)}$ la moyenne du score de la variable $\boldsymbol{\xi}_k$ pour les individus du groupe 1 remis sur une échelle sur 100. On va donc employer la statistique s définie par :

$$s = |\bar{\xi}_k^{100(G_2)} - \bar{\xi}_k^{100(G_1)}|$$

Les résultats sont rassemblés dans le tableau 7.13. Nous illustrons les moyennes obtenues dans la figure 7.16 avec des écarts types calculés par bootstrap.

Il ressort que la différence observée sur les moyennes entre les hommes et les femmes au niveau de la fidélité est significative. De plus, comme on pouvait s'y attendre, l'attitude face à l'ouverture du marché a un effet significatif sur la moyenne de la fidélité.

Groupes	H/F	Concurrence
La satisfaction		
Différence des moyennes	Non significative ($p = 0.752$)	Non significative ($p = 0.212$)
La fidélité		
Différence des moyennes	Significative ($p = 0.021$)	Significative ($p = 0.001$)

TAB. 7.13 – Résultats des tests au niveau des variables latentes

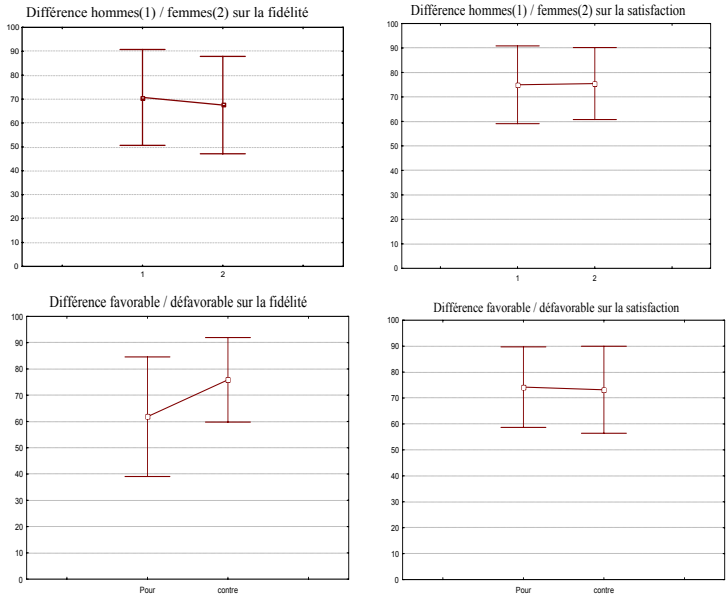

FIG. 7.16 – Moyennes et écarts types des variables latentes satisfaction et fidélité pour chacun des groupes d'observations

7.5.4 Comparaison des coefficients structurels

Nous comparons donc la relation la plus étudiée dans le cadre de l'analyse de la satisfaction, celle entre satisfaction et fidélité. Les variances sont proches et les résidus dévient légèrement de la normalité (par le test de Kolmogorov-Smirnoff sur les résidus issus des coefficients obtenus sur les échantillons bootstrap). Nous appliquons donc le test de Student et le test basé sur les permutations, pour celui-ci nous employons la statistique $s = |\hat{\beta}^{G_2}_{sat-fid} - \hat{\beta}^{G_1}_{sat-fid}|$.

Nous rassemblons dans le tableau 7.14 les résultats des différents tests.

L'hypothèse nulle n'est pas rejetée dans le cas du test de Student, on pourra considérer que les coefficients sont égaux.

Les résultats issus des deux types de tests sont similaires. Il ressort, comme nous l'avions prévu, que le sexe n'a pas d'impact significatif sur ce coefficient. Par contre, le résultat associé à l'attitude par rapport à l'ouverture du marché est plus surprenant. Cette attitude n'a donc pas d'impact sur la relation satisfaction - fidélité.

7.5.5 Conclusion

Ce cas spécifique à EDF nous a montré que le sexe de l'interviewé n'avait d'impact qu'au niveau de la moyenne de la variable latente fidélité. L'attitude face à la concurrence semble entretenir une

Groupes	Hommes/Femmes	Concurrence
Tests préalables		
Normalité des résidus	Non normaux	Non normaux
Différence entre les var. des coef.	Non significative	Non significative
Taille d'échantillon	751/1237	877/1111
Test t de Student		
Test d'égalité	H_0 non rejeté	H_0 non rejeté
Test de permutation		
Différence	Non significative ($P = 0.542$)	Non significative ($P = 0.296$)

Tab. 7.14 – Résultats des comparaisons des coefficients structurels

relation différente par rapport au modèle. Sur la relation d'intérêt spécifique satisfaction - fidélité, elle n'a pas d'impact. Par contre, elle en a au niveau du modèle global et de la moyenne de la variable latente fidélité. Il serait intéressant de tester les différences aux niveaux des poids externes afin de comprendre la différences entre les H^2. D'autres tests sont nécessaires afin de mieux comprendre les processus mais nous nous cantonnons à cet exemple.

7.6 L'asymétrie des facteurs de la satisfaction et la relation satisfaction - fidélité : la non linéarité dans les modèles d'équations structurelles

Dans le cadre de cette partie, nous utilisons les données EDF-1 et le modèle de la figure 7.13 (p. 154). Nous recherchons les non linéarités à deux niveaux :

- entre la satisfaction et ses variables manifestes associées (Sat1 à Sat6),
- entre la satisfaction et l'intention de fidélité.

Après avoir étudié le contexte marketing pour chaque cas, nous présentons les résultats des approches du chapitre 5 (p. 97).

7.6.1 Les théories de la non linéarité en marketing

Dans la littérature marketing, l'hypothèse de relations linéaires dans les modèles incluant la satisfaction et la fidélité a été bien souvent mise à mal. Dans leur article, Anderson et Gerbing (1982) étudient l'ensemble des relations allant des facteurs de la satisfaction aux profits de l'entreprise (voir fig. 7.17). Ils arrivent, par l'étude de recherches antérieures, à la conclusion que toutes ces relations sont non linéaires. Ils envisagent les relations entre la satisfaction et ses attributs et entre la satisfaction et la fidélité comme des relations asymétriques.
La théorie de l'asymétrie des variables est très importante dans le cadre de l'analyse de la satisfaction et a été déjà largement étudiée. Il serait intéressant de développer celle-ci en rapport avec d'autres concepts tels que la fidélité (ceci relève plutôt de la recherche en marketing) ou de l'adapter aux modèles d'équations structurelles à variables latentes. Récemment, Ray (2006) en a fait une synthèse bien structurée en allant, de la même façon que Anderson et Gerbing (1982), des attributs de la satisfaction au profit.

L'asymétrie induit une modélisation non linéaire particulière. L'asymétrie désigne "l'absence de correspondance exacte en forme, taille et position de deux parties opposées" (Petit Robert). Afin de la modéliser, plutôt que d'employer une fonction linéaire classique, on utilisera une fonction linéaire par morceaux.
Les premiers travaux en rapport avec l'asymétrie et la satisfaction concernaient l'asymétrie des impacts

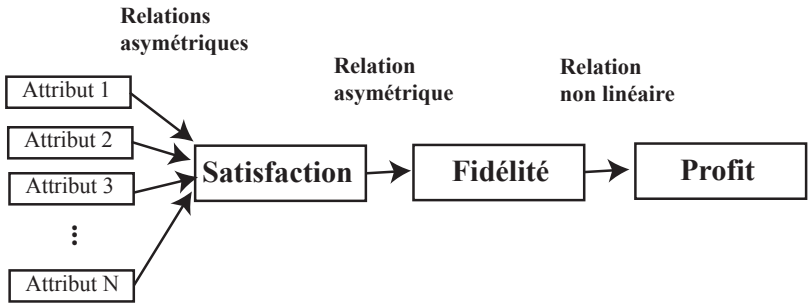

FIG. 7.17 – La chaîne allant des attributs de la satisfaction au profit de l'entreprise par Anderson et Gerbing (1982)

des attributs sur la satisfaction. Ils sont dus à Herzberg et al. (1959) puis aux qualiticiens japonais (Kano et al., 1984).

7.6.2 Le traitement du modèle externe : la mesure de l'asymétrie des impacts des attributs sur la satisfaction

Les méthodes classiques

Herzberg et al. (1959) contestent la théorie d'un continuum entre satisfaction et insatisfaction en mettant en avant des facteurs qui agissent uniquement sur la satisfaction et d'autres sur l'insatisfaction. Kano et al. (1984) développe cette théorie bi-factorielle en définissant trois "familles" de facteurs : les basiques (contribuant uniquement à l'insatisfaction), les unidimensionnels (contribuant aussi bien à la satisfaction qu'à l'insatisfaction) et les attractifs (bonus, contribuant uniquement à la satisfaction). Ils sont représentés dans la figure 7.18.

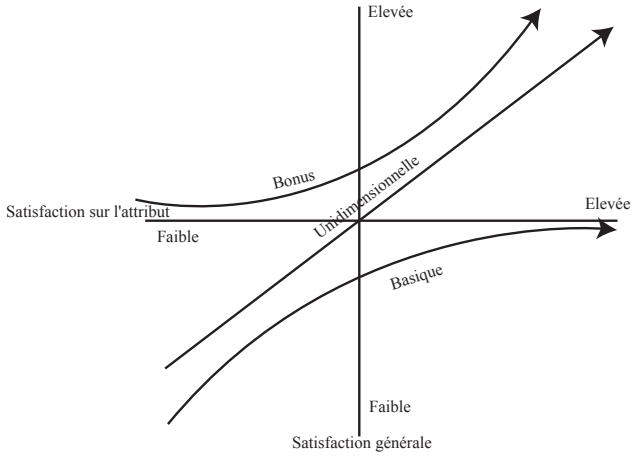

FIG. 7.18 – Asymétrie des facteurs basée sur le modèle de Kano

Deux types de méthodes peuvent être séparées : les méthodes dites d'estimations directes où une question associée au type de variable est intégrée (*Dual importance Mapping*, Martilla et James (1977) et *Simulation Method*, Kano et al. (1984)) et les méthodes dites d'estimation indirecte des poids, que nous développons car elles s'appliquent au cadre des données classiques de questionnaires.

Rivière et al. (2005) ont appliqué des méthodes directes à l'analyse sensorielle sur la satisfaction des testeurs avec une application dans le cadre de la régression PLS. Malheureusement, les données de satisfaction, nous forceront à travailler avec des méthodes indirectes. Il existe 4 principales alternatives permettant de définir le type d'impact des attributs de la satisfaction (Ray, 2001) :

1. Méthode basée sur un recueil très spécifique : la méthode des incidents critiques de Herzberg et al. (1959) et celle du questionnaire dual de Kano et al. (1984). On utilise alors une mesure de la perception sur une échelle spécifique en cinq modalités, d'une offre répondant aux attentes des clients et d'une offre n'y répondant pas. Cette approche est difficile à mettre en oeuvre à cause du recueil spécifique des données.
2. Méthode croisant l'importance déclarée avec l'importance calculée d'un facteur (Venkitaraman et Jaworski, 1993). Il faut disposer de variables associées à l'importance de chaque attribut.
3. Méthode regroupant les modalités des items de satisfaction en deux classes (modalité positive opposée à modalité négative), afin de distinguer les contributions respectivement positives et négatives de chaque facteur. Un poids de la contribution sur la satisfaction de chacun de ces items peut alors être calculé (Llosa, 1997; Brandt et Scharioth, 1998; Mittal et al., 1998). Dans ce cas, c'est le seuil de séparation qui est difficilement estimable. On peut voir en annexe le détail de l'application des méthodes de Llosa et de Brandt, ainsi que les méthodes d'estimation des seuils (p. 182).
4. Méthode créant pour chaque facteur trois catégories de consommateurs selon leur niveau de satisfaction sur le facteur afin de déduire le mode de contribution de chaque facteur (Brandt, 1988; Vanhoof et Swinnen, 1998). De la même façon, le seuil est difficile à estimer.
5. Méthode basée sur un "continuum" (Ray, 2006) en utilisant la régression logistique.

Sur les données EDF-1, les résultats obtenus par les approches de Llosa (1997) et de Brandt et Scharioth (1998) sont rassemblés dans la figure 7.19 et dans le tableau 7.15 pour les cinq premières variables associées à la satisfaction (la sixième a été laissée de côté pour des questions techniques).

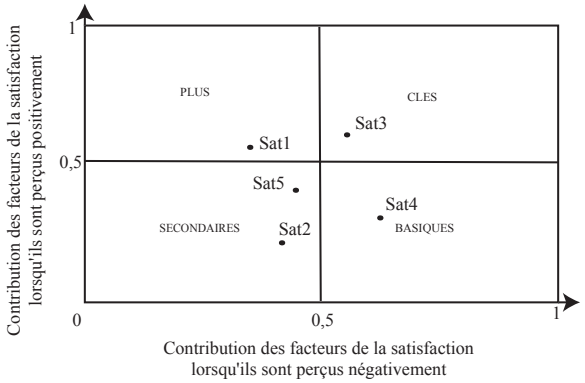

FIG. 7.19 – Répartition des variables manifestes par l'approche de Llosa

On a donc une variable (Sat1) "plus" (ou bonus) contribuant principalement à la satisfaction, une variable basique (Sat4, qui contribue principalement à l'insatisfaction), deux variables secondaires et une variable clé, ces dernières seront donc unidimensionnelles. L'approche de Brandt donne des tables de contingence qui permettent de classer les variables manifestes (tab. 7.15, x est le pourcentage des insatisfaits sur l'attribut qui sont globalement satisfaits, y est le pourcentage des neutres sur l'attribut qui sont satisfaits globalement et z représente le pourcentage des satisfaits sur l'attribut qui sont satisfaits au global).

Cette approche est basée uniquement sur le pourcentage de satisfaits au global et elle est extrêmement sensible aux seuils choisis.

Variable	x	y	z	Type de variable
Sat1	4.9	5.2	32.7	Bonus
Sat2	6.1	9.4	32.8	Performance
Sat3	2.5	10.6	58.3	Performance
Sat4	4.5	5.3	40.6	Bonus
Sat5	5.1	12.7	55.6	Performance

TAB. 7.15 – Classement des variables par la méthode de Brandt

Méthode	PLS non linéaire	PLS classique (mode A)
H^2_{sat}	0.565	0.524
F^2_{sat}	0.336	0.336
R^2_{sat}	0.595	0.656
F^2_{fid}	0.237	0.240
R^2_{fid}	0.402	0.401
GoF	0.497	0.501
w_{sat1}	0.253	0.246
w_{sat2}	0.287	0.286
w_{sat3}	0.402	0.411
w_{sat4}	0.442	0.434
w_{sat5}	0.479	0.475
w_{sat6}	0.516	0.524

TAB. 7.16 – Indices et coefficients estimés pour le modèle externe transformé, comparés à l'approche classique

L'ensemble de ces approches permet de traiter une seule relation du modèle. Nous nous intéressons par la suite à l'ensemble des facteurs de la satisfaction simultanément.

Transformation par B-splines monotones

La transformation des variables manifestes Sat1 à Sat6 par la méthode que nous avons présentée ne peut se faire qu'après avoir choisi le degré et la séquence de nœuds associés. Plusieurs possibilités s'offrent à nous :

- L'utilisation des théories marketing présentées en annexe afin d'estimer le seuil. Celles-ci nous invitent à poser un seuil de l'ordre de 7, ou 2 seuils autour de 4 et 8 pour une note sur 10. Dans le cadre de l'asymétrie classique, on pourra poser un degré de 1 mais l'utilisation de degrés supérieurs apportera des informations supplémentaires.
- L'utilisation d'une méthode pas à pas comme celle introduite par Durand (2001). A la différence de ce dernier, nous nous baserons sur l'augmentation de la communauté du groupe de variables étudié. On augmente le nombre de nœuds et de degrés et on obtient des B-splines monotones à 2 nœuds et 2 degrés (voir tableau B.7 en annexe, p. 184).

Le choix final est d'utiliser des B-splines de degré 2 à 2 nœuds internes. Nous rassemblons les indices généraux et les coefficients du modèle externe (pour la satisfaction) dans le tableau 7.16. Le modèle structurel est représenté dans la figure 7.20. Les 6 transformations sont illustrées dans la figure 7.21.

Il ressort qu'au niveau du modèle externe la communauté de la satisfaction est améliorée. Cependant, les autres indices sont plutôt moins bons pour le modèle transformé. Au niveau des poids externes, les poids sont légèrement modifiés (augmentation pour Sat1 et Sat4 et diminution pour Sat3 et Sat6). Les formes des transformations montrent pour les deux premières variables une asymétrie

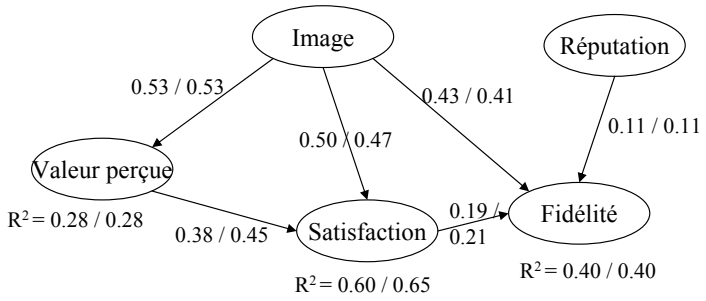

FIG. 7.20 – Modèle structurel pour les approches PLS non linéaire basée sur le modèle externe et PLS classique

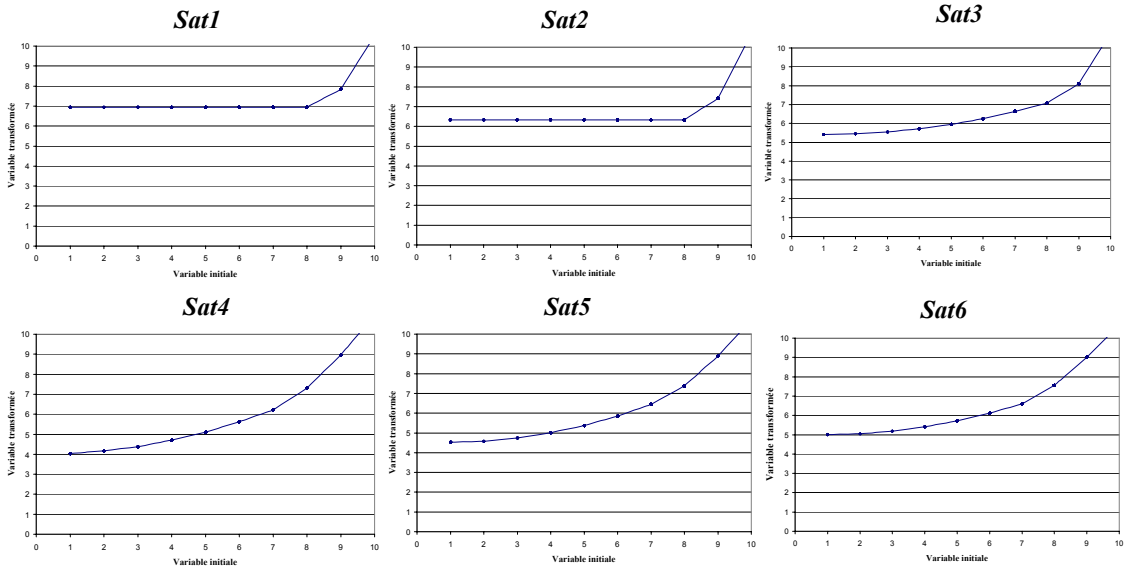

FIG. 7.21 – Les transformations des variables manifestes associées à la satisfaction

marquée, elles entrent dans la classification de Kano et al. (1984) comme des variables bonus. Les autres transformations, malgré un aspect quadratique, ne s'éloignent pas fortement de la linéarité et peuvent donc être classées comme unidimensionnelles.

Cette application montre que notre approche a un effet positif au niveau local, sur le modèle externe. La faiblesse du modèle interne associé au mode A de l'approche PLS semble accrue. Les transformations obtenues permettent des analyses supplémentaires. Ainsi, pour Sat1 et Sat2, il existe un seuil à partir duquel la variable manifeste a un effet sur la variable latente satisfaction et celui-ci se trouve au niveau de la note 8/10. Pour la variable Sat1 (bonus), Sat3 et Sat5 (unidimensionnelles), les résultats coïncident avec les approches marketing classiques de Llosa (1997) et de Brandt et Scharioth (1998).

7.6.3 Le modèle interne : la non linéarité dans la relation entre satisfaction et fidélité

Satisfaction - fidélité : une relation complexe

Le lien entre satisfaction et intention de fidélité peut être appréhendé aussi bien en terme d'asymétrie que de non linéarité (Mittal et al., 1998). Ce lien a donné lieu à de nombreuses controverses dans la recherche en marketing. Ainsi, Steukens et De Ruyter (2004) considèrent que la non linéarité entre satisfaction et fidélité n'améliore pas significativement les R^2 (du moins dans les domaines étudiés par les deux auteurs). De plus, les notions de fidélité et d'intention de fidélité sont loin d'être communes et les différentes facettes de chacun de ces concepts, qui seront prises en compte dans l'analyse, donneront différents types de relations. Nous étudions dans cette application l'impact de la satisfaction sur l'intention de fidélité.

Les recherches de Ray (2006) montrent qu'il existe trois types de notes pour la satisfaction qui permettent de mieux étudier la satisfaction et l'insatisfaction. Tout d'abord, on trouve les clients largement insatisfaits, puis des clients dont l'avis n'est pas déterminé, et finalement des clients extrêmement satisfaits. Les clients se trouvant au centre n'auront pas une grande valeur pour le décideur mais devront être pris en compte. De la même façon, pour la relation entre satisfaction et intention de fidélité, trois types de relations en fonction de deux seuils apparaissent (Ngobo, 1999) comme on peut le voir dans la figure 7.22. Les recherches sur ce sujet sont, néanmoins, controversées et les avis divergent d'une application à une autre.

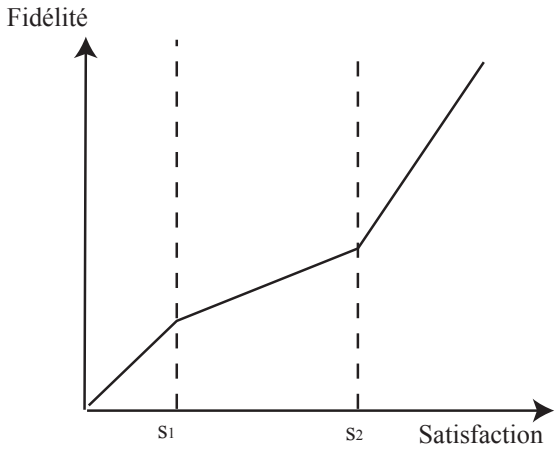

FIG. 7.22 – Relation satisfaction - fidélité

Dans le cadre de nos travaux, nous allons mettre en avant ces différentes non linéarités et étudier les modèles obtenus dans le cadre des équations structurelles à variables latentes.

Transformation par B-splines monotones

Nous utilisons toujours les données EDF-1 pour travailler sur la relation entre la fidélité et ses antécédents en se focalisant sur la relation entre la satisfaction et la fidélité. Nous appliquons donc la méthode basée sur la transformation du modèle interne en se servant des B-splines monotones. Afin de fixer le nombre de nœuds et le degré des transformations, nous employons le même procédé que pour le modèle externe. L'indice à maximiser dans ce cas là est le R^2 associé à la variable latente fidélité. Le tableau B.8 (p. 184) indique donc 2 nœuds internes et 3 degrés.

Le choix final est donc d'utiliser des B-splines de degré 3 à 2 nœuds internes. Nous rassemblons les indices généraux dans le tableau 7.17. Le modèle structurel est représenté dans la figure 7.23. Finalement, les 3 transformations sont rassemblées dans la figure 7.24.

Méthode	PLS non linéaire	PLS classique (mode A)
F^2_{fid}	0.243	0.240
R^2_{fid}	0.412	0.401
GoF	0.503	0.501

Tab. 7.17 – Indices pour le modèle interne transformé, comparés à l'approche classique

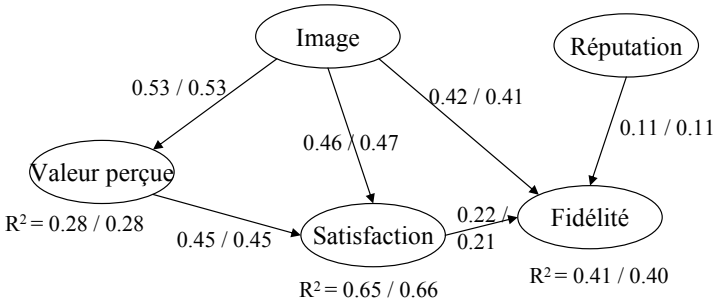

Fig. 7.23 – Modèle structurel pour les approches PLS non linéaire basée sur le modèle interne et PLS classique

On remarque une amélioration des indices de qualité prédictive de la fidélité, mais dans une faible mesure. Ceci s'explique, d'une part, par le faible impact de la satisfaction sur la fidélité (de l'ordre de 0.2) et, d'autre part, par l'allure des transformations estimées. En effet, les trois transformations de la figure 7.24 sont très proches du cas linéaire. Pour la satisfaction, il y a une tendance à un aplatissement de la courbe pour les valeurs proches de 0 et de 10. Cette analyse nous pousse à supposer que la relation entre la satisfaction et la fidélité est linéaire dans le cadre des données d'EDF. Un point important n'est pas pris en compte, nous travaillons sur l'intention de fidélité et non sur la fidélité réelle et ceci aura un impact sur les relations. De plus, EDF se trouvait en situation de monopole lors de la mise en place de cette enquête, ce qui peut expliquer la linéarité entre satisfaction et intention de fidélité.

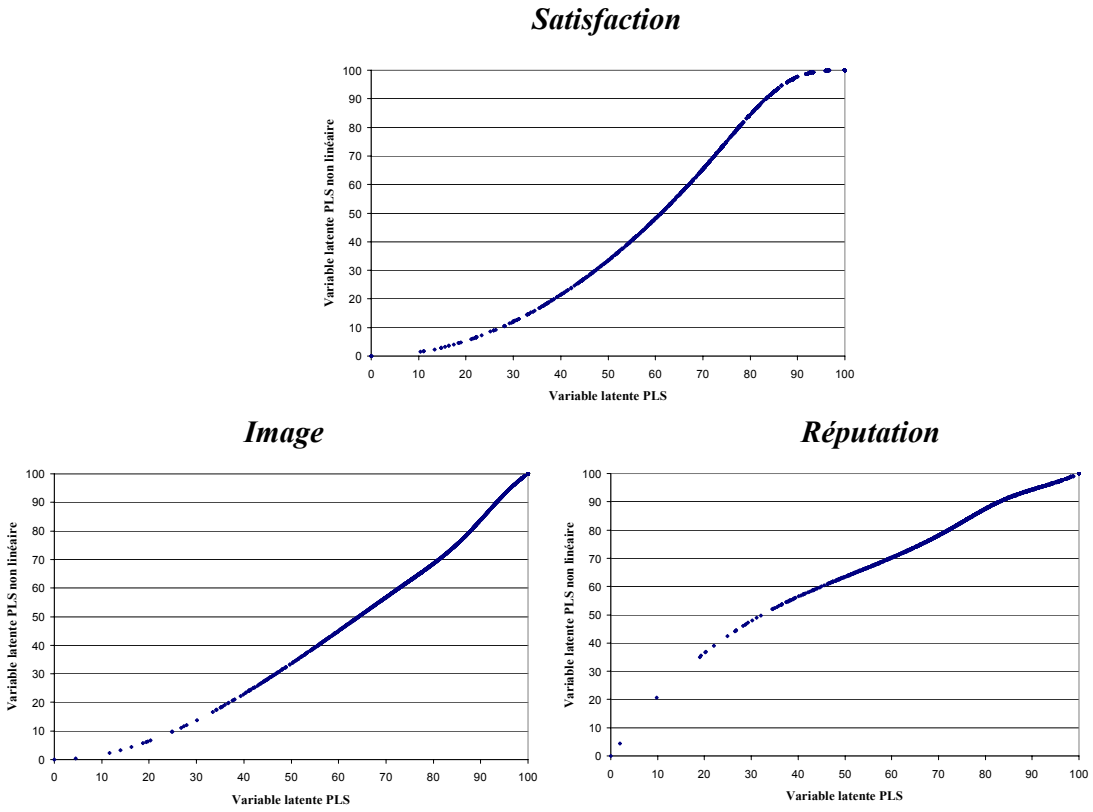

FIG. 7.24 – Les transformations des variables latentes expliquant la satisfaction

7.6.4 Conclusion

L'application à la recherche de non linéarité dans le cadre des données d'EDF nous a montré que les méthodes présentées dans le chapitre 5 obtenaient des résultats cohérents et que les transformations étaient interprétables. Cette étude valide aussi le faible impact du modèle interne dans l'estimation des paramètres du modèle dans le cadre de l'approche PLS mode A. Ainsi, on peut conclure que les relations entre la satisfaction et ses indicateurs sont asymétriques, alors que celle entre satisfaction et fidélité est linéaire dans le cas des données EDF-1.

7.7 Le cas des réclamations : illustration du traitement de questions filtrées

Nous illustrons un cas particulier de l'analyse de la satisfaction et de la fidélité : le traitement des réclamations des clients.

Les réclamations en elles-mêmes ont 2 facettes. Tout d'abord, l'action par le client de réclamer aura une conséquence négative sur la fidélité. Cet effet peut être inversé et peut même augmenter la fidélité si la réponse à la réclamation satisfait le client. Nous considérons l'attitude en cas de réclamations plutôt comme un levier que comme une facette de la fidélité. Sur l'attitude en cas de réclamations, Karatepe (2006) construit un modèle permettant de relier dans le cadre du tourisme les réactions en cas de réclamation et la fidélité des clients. En effet, erreurs et réclamations sont inévitables dans les services (Hart et Spearman, 1913). Comme on ne peut pas toutes les éviter, il faut apprendre à y répondre et à en tirer profit. Il a été montré qu'une bonne réponse aux réclamations entraîne souvent la transformation d'un client mécontent en un client fidèle (Gilly et Hansen, 1992).

Cependant, les recherches sur le sujet en sont à leurs balbutiement (Johnston et Mehra, 2002; Tax et al., 1998). Quelques études théoriques existent (Smith et Bolton, 1998; McCollough et al., 2000) et peu de recherches empiriques ont été effectuées (Davidow, 2000, 2003). La prise en compte des réclamations dans les modèles d'équations structurelles est ancienne et, dans le cadre du calcul d'indices comme le SCSB, l'ACSI ou l'ECSI (cf. chap. 7.1), une variable latente réclamations est incluse comme antécédent de la fidélité et conséquence de la satisfaction.

Lorsqu'on souhaite traiter des réclamations dans un modèle, on doit faire face au fait qu'uniquement une partie de la population a fait une réclamation. L'utilisation de méthodes de traitement de données manquantes n'est ici pas adaptée. En effet, on ne peut pas supposer que le fait de réclamer suit un processus aléatoire. Karatepe (2006) montre que, suivant le type de réclamation, tout le processus est différent. De plus, il n'est pas possible de connaître le processus modélisant l'absence ou la présence de données, celui-ci est très complexe et pourra dépendre de variables différentes pour chaque individu. Les réclamations constituent donc des questions filtrées que nous traitons en utilisant les méthodes présentées dans le chapitre 6.

Nous nous servons toujours des données EDF-1 avec 4 variables supplémentaires mesurant la satisfaction par rapport à une réclamation (Recl1-Recl4). Le filtre est modélisé par la variable binaire Recl_filtre. On obtient 80 réclamants sur 1988 individus. Ceci constituera un problème dans le traitement de cette variable.

Afin d'illustrer un cas avec plus d'observations, nous employons un autre filtre qui risque de donner des résultats plus significatifs : les contacts. Ils consistent aussi bien en contacts entrants que sortants et sont modélisés par 8 variables complétées par 384 individus sur 1988.

Nous appliquons donc les 3 méthodes d'analyse : tout d'abord, la segmentation et la comparaison de groupes, ensuite, l'utilisation d'une variable filtre supplémentaire, et finalement, l'utilisation d'une modalité "non applicable" pour les variables filtrées.

7.7.1 Segmentation et comparaison

La segmentation des données nous amène à traiter deux jeux de données, l'un de 80 observations avec une variable latente réclamations qui explique la satisfaction et la fidélité, et un second, avec 384 observations avec une variable latente contact qui explique la satisfaction et la fidélité. Les modèles internes obtenus après application de l'approche PLS apparaissent dans les figures 7.25 et 7.26. Nous détaillons les poids externes et les *loadings* en annexe (tab. B.9, p. 185).

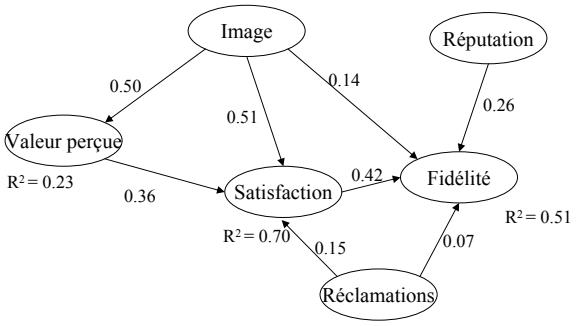

FIG. 7.25 – Modèle interne pour la sous-population ayant réclamé

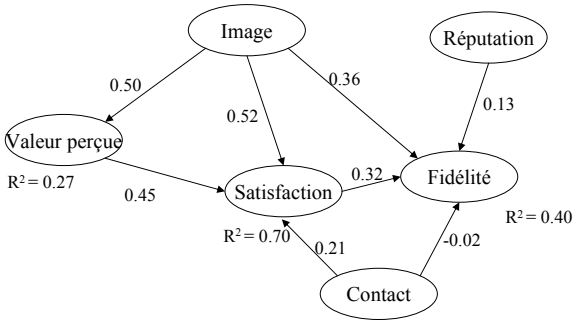

FIG. 7.26 – Modèle interne pour la sous-population ayant eu un contact

Les tailles d'échantillons étant faibles, surtout pour le premier cas, les résultats peuvent être instables. On voit que les réclamations ont un effet significatif sur la satisfaction, même s'il est beaucoup plus faible que celui de l'image ou de la valeur perçue. Par contre, sur la fidélité, il n'y a pas d'effet significatif. En ce qui concerne les contacts, le lien contact - satisfaction est fort, alors que celui entre contact et fidélité n'est pas significatif. Il serait intéressant maintenant de comparer les modèles associés, d'une part, aux clients ayant réclamé et, d'autre part, à ceux qui n'ont pas réclamé.

Comme les tailles d'échantillons sont extrêmement différentes, les tests basés sur des permutations sont adaptés. Nous appliquons les mêmes tests que dans le chapitre 7.5 (p. 153). Les résultats sont rassemblés dans le tableau 7.18.

Au niveau global, les redondances sur les réclamations sont les seuls indices significativement différents d'une classe à l'autre. Les résultats sur les coefficients structurels entre satisfaction et fidélité sont surprenants car pour les réclamations on a une différence importante (0.205 pour ceux qui ont réclamé contre 0.400 pour les autres), mais le fait d'avoir des tailles d'échantillons si différentes rend les estimations instables (voir l'étude de sensibilité de l'approche PLS dans le chapitre 1, p. 46). Ainsi, la différence n'est pas significative. Au niveau des variables latentes, les différences sont significatives, les clients ayant fait une réclamation sont moins satisfaits et moins fidèles et ceux ayant eu un contact sont moins satisfaits. Par contre, le contact n'a pas d'impact sur la fidélité.

Cette première technique permet d'étudier les influences dans les sous-populations par rapport au modèle structurel et d'expliquer les différences de structure existant entre les classes d'individus.

Groupes	Réclamations	Contacts
Le modèle externe		
Différence des $\bar{H}^2$	Non significative ($P = 0.227$)	Non significative ($P = 0.260$)
Le modèle interne		
Différence des $\bar{F}^2$	Significative ($P = 0.029$)	Non significative ($P = 0.115$)
La qualité globale		
Différence des GoF	Non significative ($P = 0.098$)	Non significative ($P = 0.163$)
La satisfaction		
Différence des moyennes	Significative ($p = 0.002$)	Significative ($P = 0.045$)
La fidélité		
Différence des moyennes	Significative ($P = 0.001$)	Non significative ($P = 0.407$)
Le coefficient structurel satisfaction - fidélité		
Différence	Non significative ($P = 0.159$)	Non significative ($P = 0.125$)

Tab. 7.18 – Résultats des comparaisons pour les réclamations et les contacts

7.7.2 Ajout d'une variable supplémentaire

Nous ajoutons donc une variable binaire supplémentaire qui modélise le fait que l'individu soit concerné ou non par le filtre. L'intégration de cette variable se fera à deux niveaux pour chaque cas traité, sur la variable latente satisfaction et sur la variable latente fidélité. Nous utilisons deux méthodes, l'approche PLS pour laquelle on admet généralement l'utilisation de variables binaires et l'approche PML avec une adaptation spécifique pour la variable filtre. Nous testons donc quatre hypothèses :

H_1 **:** le fait de réclamer a un effet significatif sur la satisfaction.

H_2 **:** le fait de réclamer a un effet significatif sur l'intention de fidélité.

H_3 **:** le fait de contacter EDF a un effet significatif sur la satisfaction.

H_4 **:** le fait de contacter EDF a un effet significatif sur l'intention de fidélité.

Dans le tableau 7.19, nous donnons les poids externes obtenus entre les variables manifestes et la variable latente pour les deux construits et pour les deux méthodes d'estimation.

	Réclamations		Contacts	
	PLS	PML	PLS	PML
Satisfaction				
filtre	-0.037	0.131	-0.0192	0.034
sat1	0.136	0.098	0.138	0.228
sat2	0.1601	0.093	0.1583	0.212
sat3	0.2317	0.148	0.2329	0.339
sat4	0.2429	0.137	0.2451	0.31
sat5	0.2665	0.168	0.2683	0.382
sat6	0.2959	0.222	0.2969	0.502
Fidélité				
filtre	-0.0405	0.127	-0.0196	0.047
Fid1	0.3459	0.268	0.3483	0.318
Fid2	0.1881	0.099	0.1882	0.118
Fid3	0.2849	0.159	0.2842	0.189
Fid4	0.2161	0.125	0.2184	0.148
Fid5	0.2357	0.137	0.2374	0.181

Tab. 7.19 – Poids externes pour le cas de l'ajout d'une variable binaire "filtre"

Le tableau fait apparaître une différence entre les estimations PLS et PML. L'approche PLS considère que la variable binaire ajoutée est continue, le fait qu'elle soit en fait binaire et donc nominale a tendance à minimiser son effet sur le modèle. Ainsi, on voit que quel que soit le cas, le poids externe associé à cette variable supplémentaire est non significatif. Dans le cas de l'approche PML, les ordres de grandeur pour les variables classiques du modèle sont proches du cas PLS. Par contre, pour la variable filtre, son apport est significatif pour les réclamations, que ce soit sur la satisfaction ou sur la fidélité. Cependant, il est non significatif pour les contacts.
Il est donc difficile de donner une réponse définitive aux hypothèses posées. Ainsi, les hypothèses H_1 et H_2 semblent validées même si l'effet du filtre est faible. Par contre, les hypothèses H_3 et H_4 sont rejetées.

Cette méthode permet de mettre en valeur l'effet du filtre. Les estimations par PLS et PML n'amènent pas aux mêmes conclusions, nous préférerons PML car l'échelle est prise en compte, il est difficile de supposer l'existence d'une variable continue sous-jacente à une variable binaire.

7.7.3 Ajout d'une modalité

La dernière méthode suppose que le fait de ne pas répondre peut être modélisé par une modalité appelée généralement "non applicable". On va donc ajouter une nouvelle modalité aux variables testées et vérifier son apport sur le modèle structurel complet. Comme les variables associées aux données EDF-1 sont sur des échelles de Likert, on ne pourra pas ajouter une modalité sans modifier le traitement du modèle. Nous utilisons une discrétisation et un traitement nominal des variables discrétisées. Pour les variables de satisfaction sur les réclamations et sur les contacts, les experts conseillent une discrétisation en 4 classes (1-4, 5-6, 7-8 et 9-10). On aura donc 5 modalités par variable. Nous utilisons l'approche PML et supposons que les variables filtrées sont nominales. Nous employons les mêmes modèles que dans la première application et obtenons les modèles des figures 7.27 pour les réclamations et 7.28 pour les contacts. Au sein du modèle externe, nous rassemblons les résultats en annexe, afin de connaître l'impact des modalités de chaque variable filtrée sur la variable latente filtrée (voir tab. B.10, p. 185 et tab. B.11, p. 186).

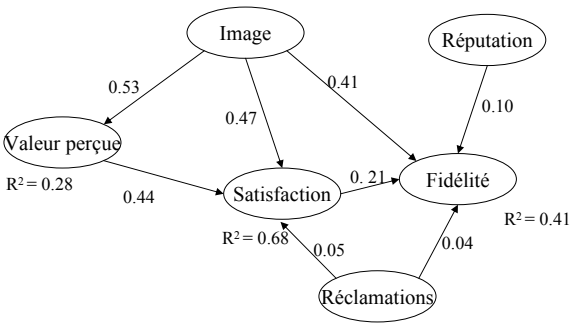

FIG. 7.27 – Modèle interne pour la sous-population ayant réclamé avec une modalité "non applicable"

Nous voyons que les nouvelles variables ainsi traitées n'ont pas d'impact significatif sur la satisfaction et la fidélité. Ceci peut s'expliquer par la présence d'une modalité très fortement représentée (90% des réponses pour les réclamations et près de 80% pour les contacts). En regardant les sous-modèles des figures 7.25 et 7.26, on voit que les contributions sont soit non significatives, soit très faibles. L'utilisation de l'ensemble des données les affaiblit davantage. De plus, le fait de traiter ces variables comme nominales réduit l'information apportée par celles-ci. Si on regarde les poids externes, la modalité "non applicable" a un poids fort mais celui-ci ne domine pas forcément les autres. Ainsi, pour les variables manifestes associées aux réclamations, au niveau de la variable Recl3, c'est la modalité 5-6 qui a le plus

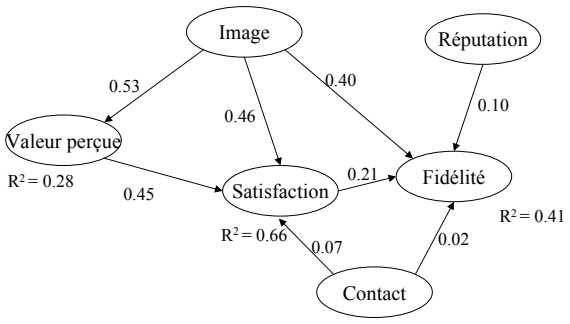

FIG. 7.28 – Modèle interne pour la sous-population ayant eu un contact avec une modalité "non applicable"

fort impact. Pour les contacts, mis à part le cas de Cont1, la modalité "non applicable" a toujours un impact très faible et ce sont les modalités 1-4 et 9-10 qui en ont un fort. Finalement, si on compare les poids externes globaux associés à chaque variable et ceux obtenus sur les modèles segmentés (tab. B.9, p. 185), on voit que les apports de chacune des variables sur les concepts sont comparables.

Cette méthode, appliquée sur des concept très discriminants au niveau du modèle global et avec plus de réponses, pourra donner des résultats intéressants, mais, dans notre cas, n'apporte pas d'informations supplémentaires et tend à confirmer le faible impact sur le modèle de ces deux concepts que sont les réponses aux réclamations et aux contacts.

7.7.4 Conclusion

Les méthodes introduites constituent des solutions possibles afin de traiter des questions filtrées dans le cadre des modèles structurels. Elles n'ont pas les mêmes principes et ne pourront pas être interprétées de la même façon. La méthode de segmentation et l'utilisation d'une modalité supplémentaire permettent de faire apparaître l'impact sur le modèle de la satisfaction par rapport à une réclamation ou par rapport à un contact. La comparaison de modèles et l'utilisation d'une variable binaire permettent de connaître l'impact d'une réclamation ou d'un contact avec EDF. L'ensemble de ces applications nous pousse à être prudent sur les conclusions. Même si les résultats sont cohérents, on aura tendance à conclure que ces concepts sont complexes et ne peuvent pas être traités globalement. Ainsi, le fait d'avoir fait une réclamation a tendance à faire baisser la satisfaction et le contact n'a pas d'effet significatif.

7.8 Conclusion des applications

Les applications que nous avons présentées permettent de répondre aux questions posées en début de chapitre et nous donnent des informations sur l'utilisation des méthodes introduites dans cette thèse.

1. La méthode à utiliser pour l'estimation d'un modèle d'équations structurelles à variables latentes dépendra de la problématique associée aux travaux. Ainsi, sur les données EDF, nous supposons que l'on ne possède pas une connaissance poussée des mécanismes en jeu. Ceci s'explique par le fait que le marché de l'électricité est actuellement en mutation et la réaction des clients est difficile à prédire. Nous nous plaçons dans un cadre où l'approche PLS constitue une bonne alternative afin d'estimer le modèle. Néanmoins, les deux autres approches, que sont LISREL et GSCA, ont leurs avantages qui ne les excluent pas de cette analyse.

2. Il s'avère que le modèle conceptuel présenté par les experts est un bon modèle car, à partir des données, on retrouve le même type de modèle. Il serait possible de rassembler les construits satisfaction et valeur perçue. La méthode basée sur les réseaux bayésiens obtient des résultats qui concordent avec ceux des autres approches. Pour le modèle interne, les modèles libres itératifs donnent de meilleurs résultats que les modèles libres et des résultats proches des heuristiques classiques.

3. Le fait d'être favorable à l'ouverture du marché de l'électricité a un impact sur la qualité du modèle interne et sur la moyenne du score de la fidélité. Etonnamment, cela n'a pas d'impact sur la relation entre satisfaction et fidélité. L'utilisation des tests présentés permet de mieux connaître les différences entre les classes d'individus par rapport à un modèle.

4. La relation entre satisfaction et fidélité, dans le cadre des données EDF traitées, est linéaire. Ceci confirme la complexité du lien entre satisfaction et fidélité qui, même dans la littérature marketing, n'est pas bien défini. Les relations entre la satisfaction et ses variables manifestes sont non linéaires, et l'application valide le fait que certaines variables ont une relation asymétrique avec la satisfaction. Les méthodes introduites permettent de bien visualiser les non linéarités et aident ainsi à l'interprétation du modèle.

5. Les réclamations et les contacts ont donc peu d'effet sur le modèle. Leur impact se trouve au niveau de la satisfaction et c'est le fait de réclamer qui a la plus forte influence. La satisfaction suite au traitement de la réclamation n'a pas d'impact significatif. Les méthodes présentées permettent d'introduire la notion de question filtrée dans un modèle structurel mais il faudra que l'impact de ce filtre soit très fort afin de pouvoir arriver à des conclusions significatives.

Ces applications permettent de mettre en valeur les apports de cette thèse dans un cadre industriel et valident les méthodes introduites sur la large gamme de sujets traités.

Conclusion générale

L'objectif de cette thèse était de mieux appréhender les modèles d'équations structurelles à variables latentes et de comprendre leur utilisation, mais aussi d'étudier un certain nombre de points précis nécessaires à leur traitement. Ce travail de recherche nous a permis d'aborder un grand nombre de sujets dans le cadre de l'analyse de données et de la statistique. Dans cette conclusion, nous rassemblons l'ensemble des apports, nous décrivons quelques limites de ce travail et proposons des perspectives de recherches.

Apports

Tout au long de ce travail, nous avons proposé des méthodes originales comme alternatives aux méthodes existantes. La validation de celles-ci s'est faite par le biais d'études sur données simulées et par une étude de cas sur des données réelles issues d'un questionnaire de satisfaction auprès des clients d'EDF.

Nous avons commencé par mettre en valeur les différences entre les méthodes d'estimation du modèle structurel et nous avons rassemblé des techniques alternatives, issues de la littérature, qui nous paraissaient novatrices (LISREL-ULS, GSCA, PML).

Des méthodes originales pour la construction du modèle conceptuel ont été introduites. Pour le modèle externe, l'algorithme fondé sur le test des tétrades permet d'éviter une erreur largement répandue : celle de traiter un construit formatif comme s'il était réflectif. Il nous paraît plus adapté de ne pas prendre en compte certaines variables et, ainsi, de vérifier les hypothèses initiales pour l'application des méthodes d'estimation plutôt que de violer ces hypothèses en utilisant l'ensemble des variables. Les tests et applications effectués nous ont permis de vérifier le bon fonctionnement de cette technique. La méthode de construction du modèle externe basée sur les réseaux bayésiens permet d'obtenir des résultats proches des méthodes de classification de variables utilisées classiquement. Elle propose, en plus, l'intégration simple de contraintes sur le modèle et offre une visualisation fort utile lors de l'application de ce type de méthodes.
Pour la construction du modèle interne, la méthode présentée permet d'utiliser les avantages des modèles libres (utilisation des corrélations partielles) et ceux des méthodes pas à pas (utilisation des scores estimés des variables latentes) sans leurs défauts (problème de choix du critère d'arrêt pour les modèles libres et utilisation d'indices ne prenant pas en compte les interactions pour les méthodes pas à pas).

Les tests développés dans le cadre de la comparaison de groupes d'observations permettent une compréhension globale des différences entre groupes au niveau d'un modèle, et une meilleure interprétation des dissemblances aux multiples niveaux présentés. L'utilisation de tests non paramétriques se justifie par l'aspect *soft modeling* de l'approche PLS. Les tests basés sur des permutations constituent une autre solution face au bootstrap, considéré comme plus classique. Ceux-ci sont particulièrement bien adaptés dans le cadre des modèles structurels et de l'approche PLS avec deux échantillons non appariés. Les tests présentés permettent d'obtenir une vision plus complète de la différence entre deux

groupes d'observations par rapport à un modèle. L'étude de simulations présentée a montré leur efficacité.

Supposer que toutes les relations sont linéaires peut poser des problèmes lorsque les théories sous-jacentes affirment que les relations ne le sont pas. Ainsi, les deux méthodes introduites permettent de mettre en valeur les non linéarités par l'utilisation des moindres carrés alternés. Ceux-ci constituent un complément adapté aux moindres carrés partiels de l'approche PLS. Les transformations par B-splines monotones permettent de conserver les possibilités d'interprétation des coefficients du modèle. Les applications ont montré l'efficacité de ces approches dans le cadre du mode A de l'approche PLS.

Finalement, le traitement des données manquantes a permis d'introduire l'ensemble des approches adaptées et fait ressortir un avantage pour les méthodes directes que sont FIML dans le cas de LISREL et NIPALS dans le cas de l'approche PLS. Les méthodes présentées, afin de prendre en compte les questions filtrées, constituent des solutions pour ce type de variables difficilement intégrables dans un modèle. Cependant, les résultats obtenus dans les applications restent complexes à interpréter.

Les travaux exposés dans cette thèse ont fait l'objet de plusieurs publications. Ainsi, un article est paru sur l'approche PML et les méthodes de construction du modèle externe et interne (Jakobowicz et Derquenne, 2007), des communications ont été présentées sur l'approche PML (Jakobowicz et al., 2005), l'analyse de sensibilité des estimations des paramètres de l'approche PLS (Jakobowicz, 2006c), les méthodes de construction du modèle (Jakobowicz, 2006b; Stan et al., 2007), le choix de l'orientation du modèle externe (Stan et al., 2007), les tests pour la comparaison de groupes (Jakobowicz, 2007a) et les approches non linéaires (Jakobowicz, 2007b; Jakobowicz et Saporta, 2007).

Limites

La principale limite de cette thèse est liée à l'étendue de son sujet. Nous avons traité des thèmes extrêmement variés touchant à beaucoup de domaines de la statistique et de l'analyse de données. Ainsi, en plus de la théorie associée aux modèles d'équations structurelles à variables latentes, nous avons abordé des sujets comme les modèles graphiques probabilistes avec les réseaux bayésiens (chap. 3.3), la théorie des tests statistiques (chap. 4), les principes de l'*optimal scaling* (chap. 5), les modèles linéaires généralisés (chap. 2.4), les méthodes de traitement des données manquantes ou encore les modèles économétriques du type Tobit (chap. 6). Ainsi, certains points auraient pu à eux seuls constituer le sujet d'une thèse. Cette étude propose donc des solutions et ouvre de nouvelles voies de recherche afin d'améliorer les résultats obtenus.

L'une des limites de ce travail réside dans le fait que nous nous sommes focalisés sur l'approche PLS et le mode A d'estimation du modèle externe. Ce choix délibéré est contestable. En effet, rien ne justifie le principe mis en place par H. Wold et repris dans toute la littérature qui veut que des construits réflectifs soient modélisés par le mode A. De plus, ce mode favorise le modèle externe, ce qui rend les scores des variables latentes très proches des premières composantes principales. Il serait donc intéressant de transposer les méthodes présentées dans le cas du mode B d'estimation du modèle externe. D'autre part, l'ensemble des travaux présentés aurait pu s'appliquer, avec certaines modifications, aux approches LISREL-ML, LISREL-ULS, PML ou GSCA.

Les méthodes exposées se basent sur des algorithmes complexes difficiles à étudier analytiquement. Nous avons présenté quelques études de simulations (cf. chapitre 2.3, p. 46, chapitre 3.2, p. 92 et chapitre 4.8, p. 58) et des exemples sur données simulées (cf. chapitre 5.4, p. 106 et chapitre 6.2, p. 116). Mais, de plus amples analyses sur chacune des méthodes seraient nécessaires afin de compléter l'étude des propriétés des approches et de mieux comparer les nouvelles méthodes aux anciennes.

De ces travaux ressortent aussi de nombreuses voies de recherche, qui pourront faire l'objet d'études ultérieures.

Ouvertures et perspectives

La première perspective est directement issue d'une limite. En effet, la majorité des travaux effectués peut être adaptée au cas du mode B d'estimation du modèle dans le cadre de l'approche PLS.

Sur la construction du modèle, les principes développés par Pearl (2000) sur les notions de causalité pourraient être intégrés avec plus de souplesse que dans le cas du logiciel TETRAD (Spirtes et al., 1996). Au niveau du choix entre schéma réflectif ou formatif, Bollen et Ting (2000) ont développé des alternatives afin de traiter moins de quatre variables ou de s'affranchir de la distribution du χ^2 (Bollen et Ting, 1998) qui peut poser problème dans le cas de petits échantillons. Pour cela, les auteurs utilisent du bootstrap. Il serait intéressant d'intégrer ces fonctionnalités dans l'algorithme introduit. La seconde alternative permettra de stabiliser les estimations du χ^2 lorsque les tailles des échantillons changent. Les réseaux bayésiens appliqués pour la mise en place du modèle externe peuvent être utilisés afin de construire le modèle interne. On se base alors sur des variables latentes issues de profils associés aux réponses données sur les variables manifestes de chaque bloc.

Les tests introduits, au niveau de la comparaison de groupes d'observations, peuvent compléter une segmentation des données par les méthodes récemment développées par Ringle et al. (2007), Sanchez et Aluja (2006), Squillacciotti (2007) ou encore Trinchera et al. (2007). L'étude de simulation devra être complétée afin de mieux comprendre le comportement de ces tests en fonction de différents facteurs.

Sur la non linéarité, l'utilisation de méthodes d'estimation du modèle ayant une fonction à optimiser tel que le mode B de l'approche PLS ou la méthode GSCA permettrait de se baser sur une fonction globale prenant en compte tout le modèle. Hwang et Takane (2002) se rapprochent de ce principe en utilisant une méthode globale avec estimation par moindres carrés alternés associée à un traitement des données catégorielles. On peut remplacer la technique de transformation des données catégorielles par une transformation non linéaire basée sur les B-splines. Ceci permettra d'optimiser directement l'estimation des paramètres du modèle et ceux des transformations sur une même fonction globale par moindres carrés alternés. Les méthodes présentées peuvent aussi s'adapter au cas LISREL-ULS qui, lui-même, consiste en une estimation par moindres carrés. Par ailleurs, l'adaptation des théories sur la régression PLS non linéaire (Durand, 2001) constitue un domaine de recherche porteur (l'approche PLS étant une généralisation de la régression PLS).

Sur les données manquantes, une analyse complète et approfondie des différentes méthodes sur des données simulées pour l'approche PLS est nécessaire. La seule étude a été menée par Kristensen et Eskildsen (2005). De plus, les méthodes de traitement des questions filtrées doivent être plus approfondies afin de mieux connaître leurs principes et leurs propriétés de façon à mieux interpréter les résultats.

Finalement, après avoir développé théoriquement ces approches, il serait intéressant de les rendre applicables dans un cadre industriel. D'une part, les programmes informatiques (macros SAS) développés dans le courant de la thèse doivent être complétés et commentés (cf. tab. B.12, p. 187). D'autre part, les méthodes pourraient être appliquées dans le cadre de problématiques actuelles et pour lesquelles elles pourront aider à prendre des décisions pour l'amélioration de la satisfaction et de la fidélité des clients.

Annexe A

Formulation alternative de la méthode Generalized Structured Component Analysis

Le modèle présenté dans le cadre du chapitre 2.4.2 (p. 53) peut être reformulé afin d'être plus intuitif, nous présentons ici cette formulation adaptée au cas réflectif ainsi qu'une illustration et la fonction globale à optimiser.

Le modèle

Le modèle initial développé par Hwang et Takane (2004) peut être reformulé dans le cas réflectif par 3 équations : la première est l'équation classique du modèle de mesure, la seconde, celle du modèle structurel, et la troisième relie les variables latentes aux variables manifestes par le biais des poids externes :

$$\mathbf{X} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\epsilon} \tag{A.1}$$

$$\boldsymbol{\xi} = \mathbf{B}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{A.2}$$

$$\boldsymbol{\xi} = \mathbf{W}\mathbf{X} \tag{A.3}$$

avec
- $\mathbf{X}$ matrice des variables observées,
- $\boldsymbol{\Lambda}$ matrice comportant les *loadings* du modèle de mesure,
- $\boldsymbol{\xi}$ matrice des variables latentes,
- $\boldsymbol{\epsilon}$ termes d'erreur du modèle de mesure,
- $\mathbf{B}$ matrice rassemblant les coefficients structurels du modèle interne,
- $\boldsymbol{\zeta}$ terme rassemblant les erreurs associées aux variables latentes endogènes,
- $\mathbf{W}$ matrice des poids associés aux variables manifestes.

Ces trois équations rappellent les équations classiques des modèles structurels à variables latentes. Elles peuvent être rassemblées dans l'équation matricielle suivante :

$$\mathbf{u} = \mathbf{A}\mathbf{u} + \mathbf{e} \tag{A.4}$$

avec $\mathbf{u} = \begin{pmatrix} \mathbf{I} \\ \mathbf{W} \end{pmatrix} \mathbf{X}$, $\mathbf{A} = \begin{pmatrix} \mathbf{0} & \boldsymbol{\Lambda} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$ et $\mathbf{e} = \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\zeta} \end{pmatrix}$. Dans cette équation, $\mathbf{u}$ représente l'ensemble des variables du modèle avec les variables manifestes exprimées par le produit $\mathbf{I} \cdot \mathbf{X}$ et les variables latentes

sous forme de combinaisons linéaires des variables manifestes en utilisant l'équation A.3.

Cette formulation se rapproche du modèle RAM (*Reticular Action Model*) introduit précédemment mais, à la différence du cas de LISREL, les variables latentes sont des composantes issues d'une combinaison linéaire des poids externes et des variables manifestes et il n'y a pas de différenciation entre variables latentes exogènes et endogènes.

Illustration

Nous présentons un exemple simple basé sur le modèle de la figure A.1.

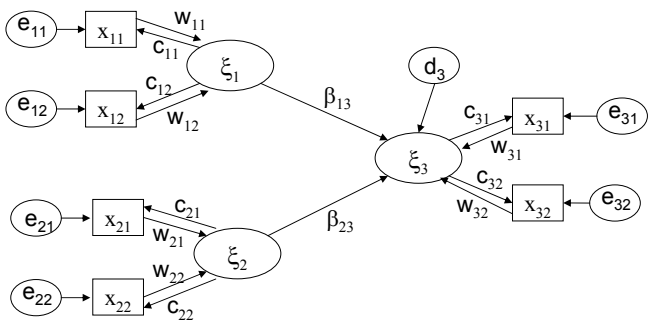

FIG. A.1 – Modèle d'équations structurelles pour l'estimation par l'approche GSCA (seconde formulation)

On écrit :

$$\mathbf{X} = \left[\begin{array}{cccccc} \mathbf{x}_{11} & \mathbf{x}_{12} & \mathbf{x}_{21} & \mathbf{x}_{22} & \mathbf{x}_{31} & \mathbf{x}_{32} \end{array} \right]'$$

$$\mathbf{e} = \left[\begin{array}{cccccccccc} \mathbf{e}_{11} & \mathbf{e}_{12} & \mathbf{e}_{21} & \mathbf{e}_{22} & \mathbf{e}_{31} & \mathbf{e}_{32} & 0 & 0 & \mathbf{d}_3 \end{array} \right]'$$

L'équation du modèle est alors :

$$\begin{aligned} & \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ w_{11} & w_{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & w_{21} & w_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{31} & w_{32} \end{bmatrix} \cdot \mathbf{X} \\ = & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & c_{11} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & c_{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & c_{21} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & c_{22} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & c_{31} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & c_{32} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \beta_{13} & \beta_{23} & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ w_{11} & w_{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & w_{21} & w_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{31} & w_{32} \end{bmatrix} \cdot \mathbf{X} + \mathbf{e} \end{aligned} \tag{A.5}$$

Estimation

Afin d'estimer les paramètres du modèle, on minimise simultanément les termes de mesure rassemblés dans la matrice **e**. Ils sont estimés de façon à ce que la somme des carrés de l'ensemble des résidus **e** soit aussi petite que possible. Ceci revient à minimiser :

$$\phi = \sum(\mathbf{u} - \mathbf{Au})'(\mathbf{u} - \mathbf{Au}) \tag{A.6}$$

sous la contrainte $\sum_{i=1}^{N} \boldsymbol{\xi}_i \boldsymbol{\xi}_i = 1$.

Les paramètres inconnus de cette équation sont donc **W**, $\boldsymbol{\Lambda}$ et **B**, qui sont modélisés dans l'équation A.4 par **W** et **A**. Ceux-ci sont estimés par le biais d'un algorithme de moindres carrés alternés.

Cette seconde formulation permet donc de rapprocher les méthode GSCA, PLS et LISREL dans leur formulation la plus compacte (modèle RAM).

Annexe B

Principaux résultats supplémentaires

Nous rassemblons dans ces annexes un certain nombre de résultats associés au chapitre 7 dont l'intégration dans le cours du chapitre n'était pas nécessaire, mais qui peuvent apporter des éclairages intéressants.

B.1 Satisfaction et fidélité des clients

	Attitudinale
1	Avez-vous l'intention de rester chez X ?
2	Avez-vous l'intention de n'utiliser que X ?
3	Avez-vous l'intention de recommander X à vos amis ?
	Réponse aux réclamations
1	Avez-vous l'intention de faire des remarques négatives sur X à vos amis ?
2	Avez-vous l'intention de décourager vos amis de venir chez X ?
3	Si vous n'obtenez pas de réponse à votre réclamation, avez-vous l'intention d'abîmer la réputation de X ?
	Tendance à la fidélité
1	J'ai tendance à présenter de nouvelles marques à mes amis
2	Je n'achète pas de marques que je ne connais pas
3	J'attends que d'autres testent une marque avant d'essayer
	Résistance à la concurrence
1	Etes-vous prêt à payer 5% de plus chez X ?
2	Etes-vous prêt à rester chez X même si la presse le décrie ?
3	Etes-vous prêt à rester chez X sans regarder le prix ?
4	Etes-vous prêt à rester chez X si le service des concurrents est meilleur ?
	Fidélité situationnelle
1	Si vous déménagez, resteriez-vous chez X ?

Tab. B.1 – Quelques exemples de questions afin de définir les facettes de la fidélité

B.2 La construction du modèle

Satisfaction des clients :
- Overall satisfaction (V1) ;
- Your expectations (V2) ;
- The range of products available (V6) ;
- The customer service you receive (V7) ;
- The physical store environment (V8) ;
- Knowledgeable employees (V9) ;
- Availability of employees to assist you (V10) ;
- Convenient locations (V11).

Tab. B.2 – Variables manifestes associées aux données GS

Image banques belges :
- On peut avoir confiance en ce que ma banque dit et fait ;
- Ma banque est une banque stable, bien établie ;
- Ma banque apporte une contribution sociale à la société ;
- Ma banque se soucie de ses clients ;
- Ma banque est innovatrice et est tournée vers l'avenir.

Image banques françaises :
- Very professional ;
- Close to their clients ;
- For all kinds of people ;
- Innovative ;
- Reliable.

Qualité perçue banques françaises :
- Qualité générale ;
- Comment jugeriez-vous X par rapport à la qualité des produits et services de la banque ;
- Comment jugeriez-vous X par rapport au service client et aux conseils personnels donnés ;
- Comment jugeriez-vous X par rapport à la "joignabilité" à l'aide de nouvelles technologies (banque par Internet, par téléphone,...) ;
- Comment jugeriez-vous X par rapport à la fiabilité et la précision des produits et services offerts ;
- Comment jugeriez-vous X par rapport à la gamme de produits et services ;
- Comment jugeriez-vous X par rapport à la clarté et la transparence des informations données ;
- Comment jugeriez-vous X par rapport à la disponibilité et la qualité des agences.

Tab. B.3 – Variables manifestes associées aux données BB et BF

Blocs	Expert	VARCLUS	CCL	Stan et Saporta (2005)	Rbappr	TETRAD
Groupe 1	i1	i1	i1	i1	i1	i1
	i2	i2	i2	i2	i2	i2
	i3	i3	i3	i3	i3	i3
	i4	i4	i4	i4	-	i4
	i5	i5	i5	i5	i5	i5
	i6	i6	i6	i6	i6	i6
	i7	i7	i7	i7	i7	-
	i8	i8	i8	i8	i8	i8
	i9	i9	i9	i9	i9	i9
	i10	i10	i10	i10	i10	i10
	i11	i11	i11	i11	i11	i11
α	0.918	0.918	0.918	0.918	0.919	0.910
Groupe 2	vp1	-	vp1	-	-	-
	vp2	-	vp2	-	-	-
α	0.746	-	0.746	-	-	-
Groupe 3	sat1	sat1	sat1	sat1	sat1	sat1
	sat2	sat2	sat2	sat2	sat2	sat2
	sat3	sat3	sat3	sat3	sat3	sat3
	sat4	sat4	sat4	sat4	sat4	sat4
	sat5	sat5	sat5	sat5	sat5	sat5
	sat6	sat6	sat6	sat6	sat6	sat6
	-	vp1	vp1	vp1	vp1	-
	-	vp2	vp2	vp2	vp2	-
α	0.814	0.860	0.860	0.860	0.860	0.814
Groupe 4	Fid1	Fid1	Fid1	Fid1	Fid1	-
	Fid2	Fid2	Fid2	Fid2	Fid2	Fid2
	Fid3	Fid3	Fid3	Fid3	Fid3	-
	Fid4	Fid4	Fid4	Fid4	Fid4	Fid4
	Fid5	Fid5	Fid5	Fid5	Fid5	Fid5
α	0.827	0.827	0.827	0.827	0.827	0.687
Groupe 5	Rep1	Rep1	Rep1	Rep1	Rep1	Rep1
	Rep2	Rep2	Rep2	Rep2	Rep2	Rep2
	Rep3	Rep3	Rep3	Rep3	Rep3	-
α	0.723	0.723	0.723	0.723	0.723	0.654
Groupe 6	CF1	CF1	CF1	CF1	CF1	-
	CF2	CF2	CF2	CF2	CF2	-
α	0.773	0.773	0.773	0.773	0.773	-
Non classés	-	-	-	-	i4	i7, vp1, vp2, fid1, fid3, Rep3, CF1,CF2
α moyen	0.800	0.820	0.820	0.820	0.821	0.766

Tab. B.4 – Blocs de variables obtenus pour l'ensemble des approches de construction

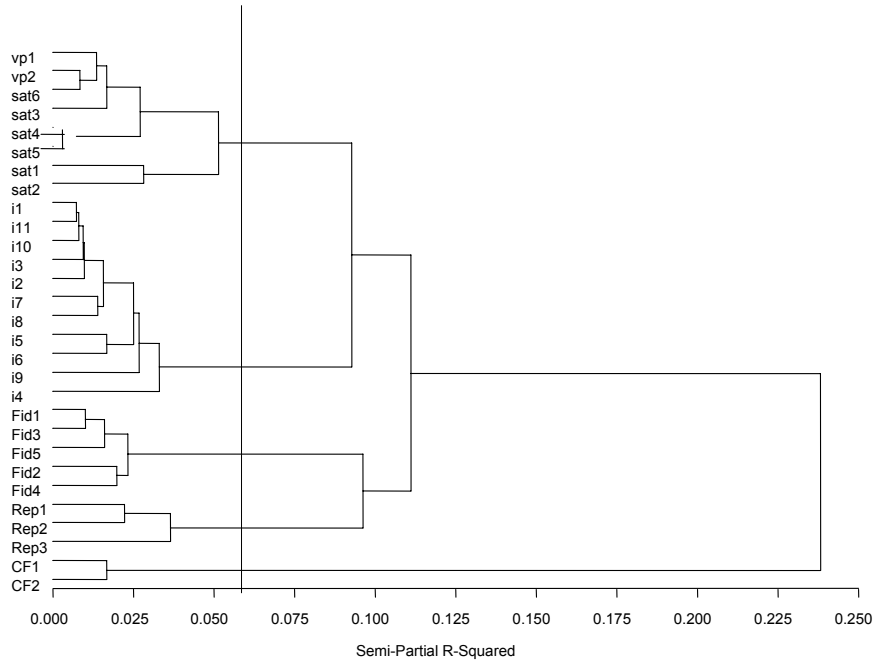

FIG. B.1 – Arbre de classification pour la méthode de Stan et Saporta (2005)

B.3 La non linéarité

L'approche de Llosa

Parmi les méthodes du type, la plus ingénieuse et la plus puissante est celle de Llosa (1997). Elle est basée sur l'utilisation de l'analyse factorielle des correspondances afin de distinguer les contributions négatives et positives sur la satisfaction. Les étapes sont les suivantes :

- recueil des données en note sur 5 ;
- la satisfaction globale est notée sur plusieurs échelles, leur somme constitue la satisfaction générale ;
- les variables à traiter sont dichotomisées ;
- l'AFC est faite sur le tableau de contingence avec en colonne les modalités de la satisfaction globale dichotomisées et en ligne les modalités des attributs dichotomisés. Le premier axe de l'AFC représente 100% de l'inertie du nuage de points, les coordonnées des attributs sur cet axe correspondent aux contributions de chacun sur la partie négative et positive. On pivote l'axe et on obtient un graphe en 2 dimensions qui permet une interprétation aisée.

Chaque partie du graphe permettra de classer une variable en tant que basique (standard), clé (unidimensionnelle), plus (bonus), secondaire (sans intérêt). Voir figure B.2.

La méthode de Brandt et ses améliorations

Cette approche (Brandt, 1988) possède un avantage sur la précédente, elle suppose l'existence de trois modalités ce qui a été fréquemment validé dans les analyses marketing sur le sujet. C'est une méthode simple basée sur une comparaison de pourcentages dans des tableaux de contingences.
Brandt (1988) propose une démarche en deux étapes :

- Rassemblement des modalités de départ en 3 modalités (insatisfaits, neutres et satisfaits).

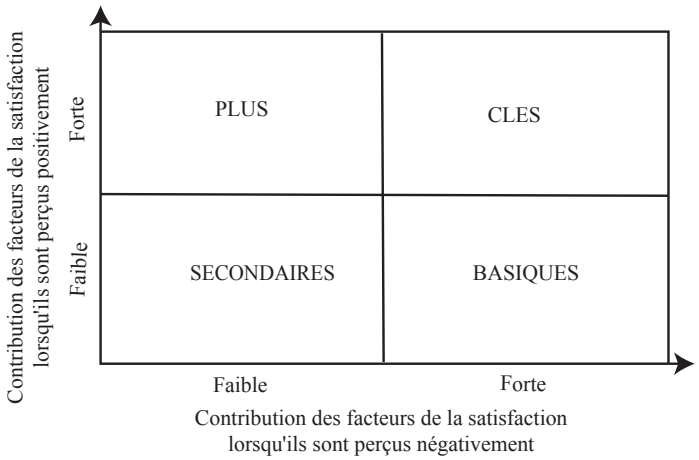

FIG. B.2 – Méthode de Llosa

Classe de variable	Propriétés
Basique	$x < y$ et $y \simeq z$
Unidimensionnelle	$x \neq y \neq z$
Bonus	$z > x$ et $x \simeq y$

TAB. B.5 – Classification par l'approche de Brandt

– Pour chacun des sous-groupes définis plus haut, la part de satisfait (en terme de satisfaction globale) est calculée. Pour chaque attribut, on obtient 3 pourcentages (x est le pourcentage des insatisfaits sur l'attribut qui sont globalement satisfaits, y est le pourcentage des neutres sur l'attribut qui sont satisfaits globalement et z représente le pourcentage des satisfaits sur l'attribut qui sont satisfaits au global). La comparaison des pourcentages permet de classer les variables dans l'un des 3 types caractérisés. Nous rassemblons dans le tableau B.5 la règle utilisée.

Vanhoof et Swinnen (1998) ont amélioré cette approche en prenant en compte l'impact sur l'insatisfaction générale et en utilisant une règle de classement basée sur les V de Cramer calculés.

Recherche des seuils afin de séparer clients satisfaits et clients insatisfaits

Nous nous baserons sur des notes de 1 à 10. Dans le cadre de nos recherches, il est difficile de différencier ces deux états (client satisfait/insatisfait). De plus, les clients sont généralement plutôt satisfaits (avec la moyenne des notes à 7/10). Cependant, on peut imaginer qu'un item comme la qualité du courant a une caractéristique de standard et qu'au-delà d'une certaine limite les résultats sont équivalents. Dans ce cas, le seuil doit être choisi suffisamment haut pour prendre en compte les petites variations.

Plusieurs approches sont possibles :

– Prendre la valeur de 5 comme seuil. Malheureusement, comme les questionnaires sont téléphoniques, le client n'aura pas forcément comme note médiane "imaginée" la valeur de 5.
– Prendre la moyenne ou la médiane sur la variable étudiée comme seuil. Ceci suppose le fait que le nombre de clients satisfaits est proche du nombre de clients insatisfaits.
– Utiliser un seuil expert. Généralement, on prend 7 comme seuil mais l'appartenance de la valeur 7 reste indécise.
– Effectuer une classification des individus et regarder si deux groupes de notes se séparent. On choisit comme seuil la moyenne inter-groupes.

Type de clients	Modalités initiales
Insatisfait	$1-6$
Neutre	$7-8$
Satisfait (*delighted*)	$9-10$

TAB. B.6 – Discrétisation en 3 modalités

		Nb. de nœuds			
		0	1	2	3
Nb.	1	0.524	0.551	0.563	0.563
de	2	0.549	0.559	**0.565**	0.564
degrés	3	0.560	0.562	0.564	0.563
	4	0.563	0.563	0.564	0.563

TAB. B.7 – Estimation du nombre de degrés et de nœuds des B-splines monotones pour le modèle externe en utilisant les communautés de la satisfaction

- Estimer le seuil à l'aide d'une question supplémentaire du type : "à partir de quelle note vous considérez-vous comme satisfait ou pas satisfait ?"
- Utiliser une ACM sur les variables à séparer de façon à voir si certaines modalités se séparent sur le premier axe de celle-ci.
- Appliquer la méthode de Llosa, basée sur la segmentation de l'indice de satisfaction par la méthode AID (Morgan et Sonquist, 1963) en deux groupes dont la variance inter-classe est maximisée, et la variance intra-classe minimisée. Cette méthode permet d'expliquer une variable quantitative à l'aide d'un ensemble de variables qualitatives. La variable de satisfaction est utilisée comme variable quantitative et cette même variable, considérée comme qualitative, est employée comme variable explicative.

Une méthode classiquement appliquée est le croisement des approches afin d'estimer le seuil. Nous testerons les approches applicables.

D'autre part, dans le cadre de l'approche de Brandt (1988), deux seuils sont nécessaires. L'utilisation des méthodes citées précédemment pour le cas binaire est possible mais généralement une discrétisation experte est choisie, elle est détaillée dans le tableau B.6.

		Nb. de nœuds			
		0	1	2	3
Nb.	1	0.408	0.410	0.411	0.412
de	2	0.410	0.411	0.411	0.412
degrés	3	0.411	0.411	**0.412**	0.412
	4	0.411	0.412	0.412	0.413

TAB. B.8 – Estimation du nombre de degrés et de nœuds des B-splines monotones pour le modèle interne en utilisant le R^2 de la fidélité

B.4 Le traitement des questions filtrées

Variable	Poids	*Loading*	Variable	Poids	*Loading*
Réclamations					
recl1	0.1786	0.7153	recl3	0.3233	0.9069
recl2	0.2936	0.8858	recl4	0.3132	0.9294
Contacts					
cont1	0.1371	0.5859	cont5	0.1608	0.8811
cont2	0.1410	0.7308	cont6	0.1503	0.8819
cont3	0.1331	0.8886	cont7	0.1574	0.8659
cont4	0.1487	0.8511	cont8	0.1805	0.8934

Tab. B.9 – Modèle externe dans le cadre de la segmentation des données sur les variables latentes filtrées

Modalité	Poids	Modalité	Poids
Recl1	0.1727	Recl3	0.2953
1-4	0.1872	1-4	0.0687
5-6	0.0711	5-6	0.3093
7-8	0.1503	7-8	0.0590
9-10	0.2796	9-10	0.2815
N/A	0.3117	N/A	0.2815
Recl2	0.2503	Recl4	0.2817
1-4	0.2668	1-4	0.2318
5-6	0.1926	5-6	0.2343
7-8	0.0061	7-8	0.0041
9-10	0.2672	9-10	0.2649
N/A	0.2672	N/A	0.2649

Tab. B.10 – Modèle externe dans le cadre de l'ajout d'une modalité aux variables filtrées sur les réclamations

Modalité	Poids	Modalité	Poids
Cont1	0.2197	Cont5	0.1075
1-4	0.2178	1-4	0.1954
5-6	0.0167	5-6	0.1037
7-8	0.0085	7-8	0.0001
9-10	0.0023	9-10	0.2399
N/A	0.7294	N/A	0.1323
Cont2	0.0852	Cont6	0.1247
1-4	0.3690	1-4	0.3712
5-6	0.0303	5-6	0.1710
7-8	0.0339	7-8	0.0469
9-10	0.4696	9-10	0.0283
N/A	0.0486	N/A	0.1275
Cont3	0.1564	Cont7	0.0838
1-4	0.0173	1-4	0.3179
5-6	0.0001	5-6	0.1979
7-8	0.0347	7-8	0.0088
9-10	0.1651	9-10	0.4377
N/A	0.0351	N/A	0.0010
Cont4	0.1093	Cont8	0.1135
1-4	0.1518	1-4	0.0637
5-6	0.1796	5-6	0.0477
7-8	0.0241	7-8	0.0933
9-10	0.3010	9-10	0.5283
N/A	0.0753	N/A	0.0005

Tab. B.11 – Modèle externe dans le cadre de l'ajout d'une modalité aux variables filtrées sur les contacts

B.5 Programmes informatiques

Dans le cadre de cette thèse de nombreux programmes ont été créé et d'autres uniquement utilisés. Dans le tableau B.12, nous rassemblons les programmes créé, ceux-ci sont basé sur le logiciel SAS (SAS Institute Inc., 2004b) et permettent d'appliquer les méthodes présentées. Dans le tableau B.13, nous détaillons les logiciels utilisés avec les versions et l'utilisation qui en a été faite.

Nom de la macro	Type d'applications
Méthode générale	
%approche_PLS	L'approche PLS (mode A/B, schéma centroïde, factoriel ou structurel)
Construction du modèle (chap. 3, p. 57)	
%iter_tetrad	Algorithme de sélection d'un construit réflectif
%hackl	Implémentation de la méthode de Hackl (2003)
%hui	Implémentation de la méthode de Hui (1982)
%amato	Implémentation de la méthode de Amato (2003)
%mod_lib_iter	Implémentation des modèles libres itératifs
Comparaison de groupes d'observations (chap. 4, p. 81)	
%test_perm_H2	Test de comparaison des $\bar{H}^2$ par permutation
%test_perm_F2	Test de comparaison des $\bar{F}^2$ par permutation
%test_perm_coef	Test de comparaison des coef. struct. par permutation
%test_perm_moy	Test de comparaison des moyennes des scores des var. latentes par permutation
Transformations non linéaires (chap. 5, p. 97)	
%trans_non_lin_ext	Transformation optimale du modèle externe combinée à l'approche PLS
%trans_non_lin_int	Transformation optimale du modèle interne combinée à l'approche PLS

Tab. B.12 – Programmes informatiques créés dans le cadre de ce travail (Macros SAS (SAS Institute Inc., 2004a))

Nom du programme	Version	Type d'applications
Logiciels généralistes		
SAS (SAS Institute Inc., 2004b)	9.1	Traitements statistiques généraux et approche LISREL
Statistica	6.0	Traitements statistiques généraux
R	2.5.1	Traitements statistiques généraux
Logiciels pour l'approche PLS		
SmartPLS (Ringle et al., 2005)	2.0	Approche PLS
PLS-Graph (Chin, 2001)	3.0	Approche PLS
Logiciels pour la méthode LISREL		
LISREL (Jöreskog et Sörbom, 1996)	8.57	Approche LISREL
AMOS (Arbuckle, 2005)	5.0	Approche LISREL
Autres logiciels		
TETRAD (Spirtes et al., 1996)	4.3.8	Méthodes de construction du modèle basées sur la théorie de la causalité
VisualGSCA (Hwang, 2007)	1.0	Méthode GSCA
%tetrad_test (Bollen et Ting, 2000)		Macro SAS pour le test des tétrades

Tab. B.13 – Logiciels utilisés dans le cadre de ce travail

Bibliographie

Acock, A.C., 2005. Working with missing values. *Journal of Marriage and Family* 67, 1012–1028. 124

Ajzen, I., Fishbein, M., 1980. *Understanding Attitudes and Predicting Social Behavior*. Prentice-Hall, Englewood Cliffs, NJ. 98

Allison, P.D., 2001. *Missing data*. Sage, Thousand Oaks, CA. 115

Amato, S., 2003. A model building strategy for PLS path modeling. In : Vilares, M., Tenenhaus, M., Coelho, P., Esposito Vinzi, V., Morineau, A. (Eds.), *PLS and Related Methods - Proceedings of the PLS'03 International Symposium*. Decisia, pp. 135–141. 75, 79, 151, 187

Amato, S., Balzano, S., 2003. Exploratory approaches to group comparison in PLS path models. In : Vilares, M., Tenenhaus, M., Coelho, P., Esposito Vinzi, V., Morineau, A. (Eds.), *PLS and Related Methods - Proceedings of the International Symposium PLS'03*. Decisia, pp. 443–451. 90

Anderson, J.C., Gerbing, D.W., 1982. Some methods for respecifying measurement models to obtain unidimensional construct measurement. *Journal of Marketing Research* 19, 453–460. 10, 66, 157, 158

Anderson, J.C., 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. 38

Arbuckle, J.L, 1996. Full information estimation in the presence of incomplete data. In : Marcoulides, G.A., Schumacker, R.E. (Eds.), *Advanced Structural Equation Modeling : Issues and Techniques*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, pp. 243–277. 117

Arbuckle, J.L., 2005. *AMOS 6.0*. AMOS Development Corporation, Spring House, PA. 142, 187

Aris, E., 2001. *Statistical Causal Models for Categorical Data*. Dutch University Press. 35, 57

Arminger, G., Muthén, B., 1998. A bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika* 63(3), 271–300. 99

Auh, S., Johnson, M.D., 2005. Compatibility effects in evaluations of satisfaction and loyalty. *Journal of Economic Psychology* 26, 35–57. 134

Baron, R.M., Kenny, D.A., 1986. The moderator-mediator variable distinction in social psychological research : Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51(6), 1173–1182. 87

Barroso, C., Cepeda, G., Roldán, J., 2005. Applying maximum likelihood and PLS on different sample sizes : Studies on SERVQUAL model and employee behaviour model. In : Aluja, T., Casanova, J., Esposito Vinzi, V., Morineau, A., Tenenhaus, M. (Eds.), *PLS and Related Methods, proceedings of the PLS'05 International Symposium, Barcelona*. SPAD, pp. 95–102. 39

Baumgartner, H., Homburg, C., 1996. Applications of structural equation modeling in marketing and consumer research : a review. *International Journal of Research in Marketing* 13(2), 139–161. 59

BayesiaLab, 2005. *BayesiaLab 3.2.a*. Bayesia SA. 148

Bentler, P.M., 1990. Comparative fit indexes in structural models. *Psychological Bulletin* 107(2), 238–246. 83

Bentler, P.M., Bonnet, D.G., 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* 88(3), 588–606. 49

Betzin, J., Henseler, J., 2005. Looking at the antecedents of perceived switching costs. A PLS path modeling approach with categorical indicators. In : Aluja, T., Casanova, J., Esposito Vinzi, V., Morineau, A., Tenenhaus, M. (Eds.), *PLS and Related Methods, proceedings of the PLS'05 International Symposium, Barcelona*. 26, 100

Bhote, K.R., 1996. *Beyond customer satisfaction to customer loyalty : the key to greater profitability*. American Management Association, New York, USA. 136

Bloemer, M.M., Kasper, H.D., 1995. The complex relationship between consumer satisfaction and brand loyalty. *Journal of Economic Psychology* 16, 311–329. 134, 135

Bollen, K., 1986. Sample size and Bentler and Bonett's nonnormed fit index. *Psychometrika* 51(3), 375–377. 32

Bollen, K., 1989. *Structural equations with latent variables*. Wiley-Interscience. 28, 30, 57

Bollen, K., 1990a. Outlier screening and distribution-free test for vanishing tetrad. *Sociological Methods and Research* 19, 80–92. 14, 58, 59, 61

Bollen, K., 1990b. Overall fit in covariance structure models : Two types of sample size effects. *Psychological Bulletin* 107(2), 256–259. 34

Bollen, K., 1995. Structural equation models that are nonlinear in latent variables : a least-squares estimator. *Sociological Methodology* 25, 223–251. 98

Bollen, K., 2002. Latent variables in psychology and the social sciences. *Annual Review of Psychology* 53, 605–634. 18, 21, 59, 81

Bollen, K., Lennox, R., 1991. Conventional wisdom on measurement : a structural equation perspective. *Psychological Bulletin* 110(2), 305–314. 59

Bollen, K., Stine, R.A., 1993. Bootstrapping goodness-of-fit measures in structural equation models. In : Bollen, K., Long, J.S. (Eds.), *Testing Structural Equation models*. Sage, Newbury Park, CA, pp. 111–135. 34

Bollen, K., Ting, K.F., 1993. Confirmatory tetrad analysis. *Sociological Methodology* 23, 147–176. 59, 61

Bollen, K., Ting, K.F., 1998. Bootstrapping a test statistic for vanishing tetrad. *Sociological Methods and Research* 27, 77–102. 62, 173

Bollen, K., Ting, K.F., 2000. A tetrad test for causal indicators. *Psychological Methods* 5(1), 3–22. 59, 62, 173, 187

Borsboom, D., Mellenbergh, G.J., Van Heerden, J., 2003. The theoretical status of latent variables. *Psychological Review* 110(2), 203–219. 18

Bouckaert, R.R., 1993. Probabilistic network construction using the minimum description length principle. *Lecture Notes in Computer Science* 747, 41–48. 70

Bowen, J.T., Shoemaker, S., 1998. The antecedents and consequences of customer loyalty. *Cornell Hotel and Restaurant Administration Quarterly* 39(1), 12–25. 133

Brandt, D.R., 1988. How service marketers can identify value-enhancing service elements. *The Journal of Services Marketing* 2(3), 35–41. 159, 182, 184

Brandt, D.R., Scharioth, J., 1998. Attribute life cycle analysis. Alternatives to the kano method. In : *Proceedings of the 51st ESOMAR-Congress*. pp. 413–429. 159, 162

Breiman, L., Friedman, J.H., 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* 80, 580–598. 101, 105

Brown, R.L., 1994. Efficacity of the indirect approach for estimating structural equation models with missing data. *Structural Equation Modeling* 1(4), 287–316. 116

Browne, M.W., 1984. Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* 37, 1–21. 36

Browne, M.W., Cudeck, R., 1989. Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research* 24(4), 445–455. 34

Browne, M.W., Cudeck, R., 1993. Alternative ways of assessing model fit. In : Bollen, K., Long, J.S. (Eds.), *Testing Structural Equation Models*. Sage, Newbury Park, CA, pp. 136–162. 33

Busemeyer, J.R., Jones, L.E., 1983. Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin* 93(3), 549–562. 98

Carroll, J.D., 1972. Individual differences and multidimensional scaling. In : Shepard, R.N., Romney, A.K., Nerlove, S.B. (Eds.), *Multidimensional Scaling : Theory and Applications in Behavioral Sciences, vol. I*. pp. 115–155. 108

Cassel, C., Hackl, P., Westlund, A.H., 1999. Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of Applied Statistics* 26, 435–446. 46, 48, 49

Cassel, C., Hackl, P., Westlund, A.H., 2000. On measurement of intangible assets : A study of robustness of partial least squares. *Total Quality Management* 11, 897–907. 46

Chen, F., Bollen, K., Paxton, P., Curran, P.J., Kirby, J.B., 2001. Improper solutions in structural equation models. *Sociological Methods & Research* 29(4), 468–508. 38

Chin, W., 1995. Partial least squares is to LISREL as principal component analysis is to common factor analysis. *Technology Studies* 2, 315–319. 39, 40

Chin, W.W., 2001. *PLS-Graph User's Guide*. C.T. Bauer College of Business, University of Houston, USA. 187

Chin, W., 2003. A permutation procedure for multi-group comparison of PLS models. In : Vilares, M., Tenenhaus, M., Coelho, P., Esposito Vinzi, V., Morineau, A. (Eds.), *PLS and Related Methods - Proceedings of the International Symposium PLS'03*. Decisia, pp. 33–43. 84, 89, 90, 92

Chin, W., 2005. Bootstrap cross-validation indices for PLS path model assessment. In : Aluja, T., Casanova, J., Esposito Vinzi, V., Morineau, A., Tenenhaus, M. (Eds.), *PLS and Related Methods, proceedings of the PLS'05 International Symposium, Barcelona*. SPAD, pp. 43–55. 25

Chin, W., Marcolin, B.L., Newsted, P.R., 1996. A partial least squares latent variables modeling approach for measuring interaction effects : Results from a monte carlo simulation study and voice mail emotion/adoption study. In : *Proceedings of the 17th Int. Conf. on Information Systems*. pp. 21–41. 46, 49, 87, 88

Coolen, H., De Leeuw, J., 1987. Least squares path analysis with optimal scaling. Rapport technique, no. RR-87-03, Department of Data Theory, University of Leiden. 53, 106

Coolen, H., De Leeuw, J., 1988. Least squares path analysis with optimal scaling. In : Diday, E. (Ed.), *Data Analysis and Informatics, V*. North-Holland, pp. 71–78. 53, 106

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16(3), 297–333. 66

Cunningham, R., 1956. Brand loyalty : what, where, how much ? *Harvard Business Review* Jan.-Feb., 116–128. 132

Curran, P.J., Bollen, K., Chen, F., Paxton, P., Kirby, J.B., 2003. Finite sampling properties of point estimates and confidence intervals of the RMSEA. *Sociological Methods & Research* 32(2), 208–252. 33

Davidow, M., 2000. The bottom line impact of organizational responses to customer complaints. *Journal of Hospitality and Tourism Research* 24(4), 473–490. 165

Davidow, M., 2003. Organizational responses to customer complaints : what works and what doesn't. *Journal of Service Research* 5(3), 225–250. 165

Day, R.L., 1984. Modeling choices among alternative responses to dissatisfaction. *Advances in consumer Research* 11, 496–499. 130

De Boor, C., 1978. *A practical Guide to Splines*. Springer. 102

De Leeuw, J., 1987. Path analysis with optimal scaling. In : Legendre, P., Legendre, L. (Eds.), *Developments in Numerical Ecology*. Springer-Verlag, pp. 381–404. 53, 106

De Leeuw, J., 1988. Multivariate analysis with linearizable regressions. *Psychometrika* 53(4), 437–454. 105, 113

Dempster, A., Laird, N., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B : Methodological* 39, 1–38. 116

Derquenne, C., 2005. Generalized path modeling based on the partial maximum likelihood approach. In : Aluja, T., Casanova, J., Esposito Vinzi, V., Morineau, A., Tenenhaus, M. (Eds.), *PLS and Related Methods, proceedings of the PLS'05 International Symposium, Barcelona*. Decisia, pp. 159–166. 14, 26, 49

Derquenne, C., 2006. Traitements statistiques de données catégorielles : Recherche exploratoire de structures et modélisation de phénomènes. Thèse de doctorat, Université Paris Dauphine, France. 49

Derquenne, C., Hallais, C., 2004. Une méthode alternative à l'approche PLS : comparaison et application aux modèles conceptuels marketing. *Revue de Statistique Appliquée* LII(3), 37–72. 76, 77, 138, 151

Dijkstra, T., 1981. Latent variables in linear stochastic models : Reflections on "maximum likelihood" and "partial least squares" methods. Thèse de doctorat, Univ. Groningen. 26, 27, 49, 106

Dijkstra, T., 1983. Some comments on maximum likelihood and partial least squares methods. *Journal of Econometrics* 22, 67–90. 39

Dillon, W.R., Goldstein, M., 1984. *Multivariate analysis (Methods and applications)*. John Wiley and Sons Inc., USA. 67

Durand, J.F., 2001. Local polynomial additive regression through PLS and splines : PLSS. *Chemometrics and Intelligent Laboratory Systems* 58, 235–246. 113, 160, 173

Eberl, M., 2007. An application of PLS in multi-group analysis : The need for differentiated corporate-level marketing in the mobile communications industry. In : Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H. (Eds.), *Handbook of Partial Least Squares : Concepts, Methods and Applications in Marketing and related fields*. Springer. 90

Edgington, E.S., 1987. *Randomization tests*. Marcel Dekker, Inc. 82, 89

Edwards, J.R., Bagozzi, R.P., 2000. On the nature and direction of relationships between constructs and measures. *Psychological Methods* 5(2), 155–174. 59

Enders, C.K., Bandalos, D.L., 2001. The relative performance of the Full Information Maximum Likelihood estimaton for structural equation modeling with missing data. *Structural Equation Modeling* 8(3), 430–457. 117

Etezadi-Amoli, P., McDonald, R.P., 1983. A second generation nonlinear factor analysis. *Psychometrika* 48(3), 315–342. 98

Fagan, J., Greenberg, B.V., 1988. Using graph theory to analyse skip patterns in questionnaires. Rapport technique, Bureau of the Census, Statistical Research Division, Washington D.C. 124

Farley, J.U., 1964. "Brand loyalty" and the economics of information. *Journal of Business* 37, 370–381. 133

Ford, B.L., 1983. An overview of hot-deck procedures. In : Madow, W.G., Olkin, I., Rubin, D.B. (Eds.), *Incomplete Data in Sample Surveys, vol. II*. Academic Press, New York, pp. 85–207. 116

Fornell, C., 1992. A national customer satisfaction barometer : The swedish experience. *Journal of Marketing* 56, 6–21. 86, 130, 137

Fornell, C., Bookstein, F.L., 1982. Two structural equation models : LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research* 19, 440–452. 39, 58

Fornell, C., Johnson, M.D., Anderson, E.W., Cha, J., Bryant, B.E., 1996. The american customer satisfaction index : Nature, purpose and findings. *Journal of Marketing* 60, 7–18. 137

François, O., Leray, P., 2003. Etude comparative d'algorithmes d'apprentissage de structure dans les réseaux bayésiens. In : De Saint-Cyr, F.D. (Ed.), *6èmes Rencontres Nationales des Jeunes Chercheurs en Intelligence Artificielle*. Presses Universitaires de Grenoble, pp. 167–180. 71

Frisou, J., 1998. Premiers jalons pour une théorie éclectique de la fidélité des clients : Un essai de validation empirique sur le marché des services de télécommunication. In : *Congrès International de l'Association Française du Marketing, Bordeaux, France*. 10, 134, 135

Ganesh, J., Arnold, M.J., Reynolds, K.E., 2000. Understanding the customer base of service providers an examination of the differences between switchers and stayers. *Journal of Marketing* 64, 65–87. 133

Geisser, S., 1974. A predictive approach to the random effect model. *Biometrika* 61(1), 101–107. 25

Gerpott, T.J., Rams, W., Schindler, A., 2001. Customer retention, loyalty, and satisfaction in the german mobile cellular telecommunications market. *Telecommunications Policy* 25, 249–269. 132

Giese, J.L., Cote, J.A., 2000. Defining consumer satisfaction. *Academy of Marketing Science Review* (online). 130, 131

Gilly, M.C., Hansen, R.W., 1992. Consumer complaint handling as a strategic marketing tool. *The Journal of Product and Brand Management* 1(3), 5–16. 165

Gitomer, J., 1998. *Customer Satisfaction is Worthless, Customer Loyalty is Priceless*. Bard Press, Texas, USA. 136

Glymour, C., Spirtes, P., Scheines, R., Kelly, K., 1987. *Discovering causal structure*. Academic Press, Orlando. 59

Gold, M.S., 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. 115, 116

Gold, M.S., Bentler, P.M., Kim, K.H., 2003. A comparison of maximum-likelihood and asymptotical distribution free methods of treating incomplete nonnormal data. *Structural Equation Modeling* 10(1), 47–79. 37

Goodhue, D., Lewis, W., Thompson, R., 2006. PLS, small sample, and statistical power in MIS research. In : *39th Hawaii International Conference on System Sciences*. 46

Gourieroux, J., 1989. *Econométrie des Variables Qualitatives*, 2nde Edition. Economica. 121

Green, S.B., Lissitz, R.W., Mulaik, S.A., 1977. Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement* 37, 827–838. 66

Gronholdt, L., Martensen, A., Kristensen, K., 2000. The relationship between customer satisfaction and loyalty : cross-industry differences. *Total Quality Management* 11(4-6), 509–514. 136

Hackl, P., 2003. Specification analysis of structural equation models. In : Vilares, M., Tenenhaus, M., Coelho, P., Esposito Vinzi, V., Morineau, A. (Eds.), *PLS and Related Methods - Proceedings of the International Symposium PLS'03*. Decisia, pp. 127–134. 75, 76, 77, 78, 79, 151, 152, 187

Halstead, D., Hartman, D., Schmidt, S.L., 1994. Multisource effects on the satisfaction formation process. *Journal of the Academy of Marketing Science* 22, 114–129. 130

Hanafi, M., 2004. Approche PLS : une hiérarchie des stratégies pour la détermination des variables latentes. In : *Actes des 36èmes journées de statistique de la SFDS - Montpellier*. 26

Harris, L.C., Goode, M.H., 2004. The four levels of loyalty and the pivotal role of trust : a study of online service dynamics. *Journal of Retailing* 80, 139–158. 133

Hart, B., Spearman, C., 1913. General ability, its existence and nature. *British Journal of Psychology* 5, 51–84. 62, 165

Heckerman, D., Geiger, D., Chickering, M., 1994. Learning bayesian networks : The combination of knowledge and statistical data. In : De Mantaras, R.L., Poole, D. (Eds.), *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, pp. 293–301. 71

Heckman, J., 1979. Sample selection bias analyzing censored and sample-selected data with tobit and heckit models. *Econometrica* 47, 153–161. 121, 123

Henseler, J., 2007. A new and simple approach to multi-group analysis in partial least squares path modeling. In : *PLS'07 - 5th International symposium, Oslo*. 89

Herzberg, F., Mausner, B., Snyderman, B.B., 1959. *The motivation to work*, 2nde Edition. John Wiley & Sons, New York. 158, 159

Heskett, J.L., Sasser, E.W., Schlesinger, L.A., 1997. *The Service Profit Chain*. The Free Press, New York. 136

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, 417–441,498–520. 108

Howard, J.A., Sheth, J.N., 1969. *The Theory of Buyer Behavior*. John Wiley and Sons, New York. 130

Hui, B.S., 1982. On building partial least squares models with interdependent inner relations. In : Jöreskog, K.G., Wold, H. (Eds.), *Systems under indirect observation*. Vol. 2. North-Holland, Amsterdam, pp. 249–271. 74, 75, 76, 77, 79, 151, 152, 187

Hwang, H., 2007. *VisualGSCA 1.0*. Department of Psychology, McGill University, Montreal, Canada. 55, 142, 187

Hwang, H., DeSarbo, W.S., Takane, Y., 2007. Fuzzy clusterwise generalized structured component analysis. *Psychometrika* 72(2), 181–198. 53, 55

Hwang, H., Takane, Y., 2002. Structural equation modeling by extended redundancy analysis. In : Nishisato, S., Baba, Y., Bodzogan, H., Kanefuji, K. (Eds.), *Measurement and Multivariate Analysis*. Springer, pp. 115–124. 55, 113, 173

Hwang, H., Takane, Y., 2004. Generalized structured component analysis. *Psychometrika* 69(1), 81–99. 14, 53, 55, 106, 113, 175

Jaccard, J., 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), 348–357. 98

Jacoby, J., Chestnut, R., 1978. *Brand Loyalty Measurement and Management*. Wiley, New York. 133

Jacoby, M.D., Kyner, A., 1973. Brand loyalty vs. repeat purchasing behavior. *Journal of Marketing Research* 10, 1–9. 131

Jakobowicz, E., 2006a. Les réseaux bayésiens et l'analyse de la satisfaction et de la fidélité des clients d'EDF. Rapport technique, EDF R&D, Clamart. 150

Jakobowicz, E., 2006b. Méthodes pour la construction du modèle conceptuel en vue de l'application de l'approche PLS. In : *Journées de Statistique de la SFdS, Clamart*. 172

Jakobowicz, E., 2006c. Understanding PLS path modeling parameters estimates : a study based on monte carlo simulation and customer satisfaction surveys. In : *COMPSTAT 2006, 17th Symposium on Computational Statistics, Rome, Italie*. Physica-Verlag, pp. 721–728. 172

Jakobowicz, E., 2007a. Comparaison de groupes d'observations dans le cadre de l'approche PLS. In : *Journées de Statistique de la SFdS, Angers, juin*. 172

Jakobowicz, E., 2007b. Latent variable transformation using monotonic B-splines in PLS path modeling. In : *IASC 07 - Statistics for Data Mining, Learning and Knowledge Extraction, août-septembre*. 172

Jakobowicz, E., Derquenne, C., 2007. A modified PLS path modeling algorithm handling reflective categorical variables and a new model building strategy. *Computational Statistics and Data Analysis* 51(7), 3666–3678. 14, 26, 49, 50, 71, 77, 126, 150, 172

Jakobowicz, E., Derquenne, C., Casacci, V., 2005. Methods for the analysis of customer satisfaction and loyalty : the experience of Electricité de France. In : *3rd world conference on Computational Statistics & Data Analysis, Limassol, Chypre*. 172

Jakobowicz, E., Saporta, G., 2007. A nonlinear PLS path modeling based on monotonic B-spline transformations. In : *Causalities explored by indirect observations - PLS'07, 5th International symposium on PLS and related methods, Oslo, septembre*. 172

Jarvis, C.B., MacKenzie, S.B, Podsakoff, P.M., 2003. A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research* 30, 199–218. 58, 59

Johnsen, A., 2003. Residential customers and competitive electricity markets : The case of Norway. *The Electricity Journal* 16(1), 74–79. 136

Johnson, M.D., Gustafsson, A., Andreassen, T.W., Lervik, L., Cha, J., 2001. The evolution and future of national customer satisfaction index models. *Journal of Economic Psychology* 22, 217–245. 137

Johnston, R., Mehra, S., 2002. Best-practice complaint management. *Academy of Management Executive* 16(4), 145–154. 165

Jones, T., Sasser, W., 1995. Why satisfied customers defect. *Harvard Business Review* Nov.-Dec., 88–99. 135

Jöreskog, K.G., Yang, F., 1996. Nonlinear structural equation models : The Kenny-Judd model with interaction effects. In : Marcoulides, G.A., Schumacker, R.E. (Eds.), *Advanced structural equation modeling : Issues and Techniques*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, pp. 57–88. 98

Jouffe, L., Munteanu, P., 2001. New search strategies for learning bayesian networks. In : *Proceedings of the 10th International Symposium on Applied Stochastic Models and Data Analysis, Compiègne, France*. pp. 591–596. 72

Jöreskog, K.G., 1967. Some contributions to maximum likelihood factor analysis. *Psychometrika* 32(4), 443–482. 31

Jöreskog, K.G., 1970. A general method for analysis of covariance structures. *Biometrika* 57(2), 239–251. 14, 17, 28

Jöreskog, K.G., 2000. *Latent Variable Scores and their uses*. Scientific Software International Inc. 37

Jöreskog, K.G., Sörbom, D., 1996. *LISREL 8 : User's Reference Guide*. Scientific Software International Inc. 28, 142, 187

Jöreskog, K.G., Sörbom, D., Du Toit, S., Du Toit, M., 2000. *LISREL 8 : New Statistical Features*. Scientific Software International Inc. 35

Kano, N., Seraku, N., Takahashi, F., Tsuji, S., 1984. Attractive quality and must-be quality. *Hinshitsu : The Journal of the Japanese Society for Quality Control* 14, 39–48. 158, 159, 162

Karatepe, O.M., 2006. Customer complaints and organizational responses : the effects of complainants' perceptions of justice on satisfaction and loyalty. *Hospitality Management* 25, 69–90. 165

Keil, M., Tan, B.C., Wei, K.K., Saarinen, T., Tuunainen, V., Wassenaar, A., 2006. A cross-cultural study on escalation of commitment behavior in software projects. *Behaviour & Information Technology* 25(1), 19–36. 90

Kelley, T.L., 1928. *Crossroads in the Mind of Mind*. Stanford University Press, Stanford. 60

Kenny, D.A., Judd, C.M., 1984. Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin* 96(1), 201–210. 98

Kotler, P., Dubois, B., 1993. Satisfaire la clientèle à travers la qualité, le service et la valeur. *Revue Française du Marketing* 144/145(4/5), 35–52. 134

Krämer, N., 2005. Nonlinear partial least squares path models. In : *3rd World conference of the IASC, Cyprus*. 99, 100

Kressman, F., Muller, R.H., 2002/2003. Comparing linear structural relationship modeling and partial least squares for business excellence. Rapport technique, Forschungsmethodik I of Prof. Boutellier, Prof. Fahrni and Prof. Westlund. 39

Kristensen, K., Eskildsen, J.K., 2005. Missing values, partial least squares and the estimation of customer satisfaction. In : Aluja, T., Casanova, J., Esposito Vinzi, V., Morineau, A., Tenenhaus, M. (Eds.), *PLS and Related Methods, proceedings of the PLS'05 International Symposium, Barcelona*. SPAD, pp. 33–42. 173

Kristensen, K., Eskildsen, J.K., Juhl, H.J., Ostergaard, P., 2003. PLS structural equation modeling for customer satisfaction : Methodological and application issues. In : *13th Int. and 68th Annual American Meeting of the Psychometric Society, Cagliari, Sardinia*. 46, 47, 49

Kruskal, J.B., Shepard, R.N., 1974. A nonmetric variety of linear factor analysis. *Psychometrika* 39(2), 123–157. 101

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22(1), 79–86. 72, 148

Lam, S.Y., Shankar, V., Erramilli, M.K., Murthy, B., 2002. Investigating the interrelationships among customer value, customer satisfaction, switching costs and customer loyalty. Rapport technique, Nanyang Business School, Singapore. 132, 136

Lauritzen, S.L., 1996. *Graphical Models*. Oxford University Press. 19

Lee, S.Y., Poon, W.Y., Bentler, P.M., 1992. Structural equation models with continuous and polytomous variables. *Psychometrika* 57(1), 89–105. 35

Lee, S.-Y., Poon, W.-Y., Bentler, P., 1995. A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology* 48, 339–358. 35

Lee, S.-Y., Tang, M.L., 2006. Bayesian analysis of nonlinear structural equation models with nonignorable missing data. *Psychometrika* 71(3), 541–564. 123, 124

Lee, S.-Y., Zhu, H.-T., 2002. Maximum likelihood estimation of nonlinear structural equation models. *Psychometrika* 67(2), 189–210. 99

Lei, M., Lomax, R.G., 2005. The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling* 12(1), 1–27. 36

Liao, T.F., 2002. *Statistical Group Comparison*. Wiley. 81, 83, 90

Likert, R., 1932. *A Technique for the Measurement of Attitudes*. Vol. 140. Archives of Psychology. 66, 68, 71, 140, 141

Lindquist, J.D., 1974. Meaning of image : survey of empirical and hypothetical evidence. *Journal of Retailing* 50, 29–38. 147

Little, R.J., Rubin, D.B., 2002. *Statistical Analysis with missing data*, 2nde Edition. Wiley. 115

Llosa, S., 1997. L'analyse de la contribution des éléments du service à la satisfaction : Un modèle "tétraclasse". *Décisions Marketing* 10, 81–88. 159, 162, 182

Lorit, J.F., Barichard, S., Brunetiere, X., Pavé, F., Thierrée, J., 2002. Les méthodes d'évaluation de la satisfaction des usagers. Rapport technique, Cour des Comptes, Paris. 139

Lyttkens, E., Areskoug, B., 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

MacCallum, R.C., Browne, M.W., 1993. The use of causal indicators in covariance structure models : some practical issues. *Psychological Bulletin* 114(3), 533–541. 58

Marcoulides, G.-A., Schumacker, R.-E., 1998. *Interaction and Non-Linear Effects in Structural Equation*. Lawrence Erlbaum Associates. 98

Martilla, J.A., James, J.C., 1977. Importance performance analysis. *Journal of Marketing* 41, 77–79. 158

Mathes, H., 1993. Global optimization criteria of the PLS-algorithm in recursive path models with latent variables. In : Haagen, K., Bartholomew, D.J., Deistler, M. (Eds.), *Statistical Modelling and Latent Variables*. Elsevier Science Publishers, pp. 229–248. 26

McCollough, M.A., Berry, L.L., Yadav, M.S., 2000. An empirical investigation of customer satisfaction after service failure and recovery. *Journal of Service Research* 3(2), 121–137. 165

McDonald, R.P., 1962. A general approach to nonlinear factor analysis. *Psychometrika* 27(4), 397–415. 98

McDonald, R.P., 1967. Factor interaction in nonlinear factor analysis. *British Journal of Mathematical and Statistical Psychology* 20, 205–215. 98

McDonald, R.P., 1996. Path analysis with composite variables. *Multivariate Behavioral Research* 31(2), 239–270. 9, 14, 19, 39, 41, 42, 43, 44, 45, 144, 146

McDonald, R.P., McArdle, J.J., 1984. Some algebraic properties of the reticular action model. *British Journal of Mathematical and Statistical Psychology* 37, 234–251. 19, 42, 53

Meijer, E., 1998. *Structural Equation Models for Nonnormal Data*. DSWO Press, Leiden. 32

Meijerink, F., 1995. *A Nonlinear Structural Relations Model*. DSWO Press, Leiden. 37, 99

Mittal, V., Ross, W.T., Baldasare, P.M., 1998. The asymmetric impact of negative and positive attribute-level performance on overall satisfaction and repurchase intentions. *Journal of Marketing* 62, 33–47. 159, 162

Morgan, J.J., Sonquist, J.A., 1963. Problems in the analysis of survey data and proposal. *Journal of the American Statistical Association* 58, 415–435. 10, 133, 134, 184

Muthén, B., 1989. Tobit factor analysis. *British journal of mathematical and statistical psychology* 42, 241–250. 121

Muthén, B., Kaplan, D., Hollis, M., 1987. On structural equation modeling with data that are not missing completely at random. *Psychometrika* 5(3), 431–462. 115, 123

Muthén, B., Muthén, L.K., 1998. *Mplus : User's Guide*. Muthén and Muthén, Los Angeles, CA. 99, 124

Naïm, P., Wuillemin, P.-H., Leray, P., Becker, A., Pourret, O., 2004. *Réseaux bayésiens*. Eyrolles. 68, 69

Ngobo, P.-V., 1999. Decreasing returns in customer loyalty : does it really matter to delight the customers ? *Advances in Customer Research* 26, 469–476. 162

Olinsky, A., Chen, S., Harlow, L., 2003. The comparative efficacity of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research* 151, 53–79. 116, 118

Oliver, R., 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research* 17, 460–469. 130, 131

Oliver, R., 1997. *Satisfaction : A Behaviour Perspective on the Consumer*. The McGraw Hill Company, Inc., New York. 130

Pagès, J., Tenenhaus, M., 2001. Multiple factor analysis combined with PLS path modelling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgements. *Chemometrics and Intelligent Laboratory Systems* 58, 261–273. 113

Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann. 19, 68, 70

Pearl, J., 1998. Graphs, causality, and structural equation models. *Sociological Methods & Research* 27, 226–284. 57

Pearl, J., 2000. *Causality : Models, Reasoning, and Inference*. Cambridge University Press. 57, 68, 70, 71, 173

Percebois, J., Wright, P., 2001. Electricity consumers under the state and the private sector : comparing the price performance of the French and UK electricity industries 1990-2000. *Utilities Policy* 10, 167–179. 135

Ping, R.A., 1996. Latent variable regression : A technique for estimating interaction and quadratic coefficients. *Multivariate Behavioral Research* 31(1), 95–120. 98

Raju, P., 1980. Optimum stimulation level its relationship to personality, demographics and exploratory behaviour. *Journal of Consumer Research* 7, 272–282. 133

Ramsay, J.O., 1988. Monotone regression splines in action. *Statistical Science* 3(4), 425–461. 104

Ray, D., 2001. *Mesurer et développer la satisfaction clients*. Editions d'Organisation. 137, 159

Ray, D., 2006. L'asymétrie dans la chaîne attributs-satisfaction-fidélité : aspects théoriques et méthodologiques. Thèse de doctorat, Université Paris I Panthéon-Sorbonne, Paris, France. 157, 159, 162

Ringle, C.M., Wende, S., Will, A., 2005. *SmartPLS 2.0*. University of Hamburg, Germany, www.smartpls.de. 142, 187

Ringle, C.M., Wende, S., Will, A., 2007. Finite mixture partial least squares analysis : Methodology and numerical examples. In : Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H. (Eds.), *Handbook of Partial Least Squares : Concepts, Methods and Applications in Marketing and related fields*. Springer. 90, 173

Rivière, P., Saporta, G., Pagès, J., Monrozier, R., 2005. Kano's satisfaction model applied to external preference mapping : a new way to handle non linear relationships between hedonic evaluations and product characteristics. In : Aluja, T., Casanova, J., Esposito Vinzi, V., Morineau, A., Tenenhaus, M. (Eds.), *PLS05, 4th International Symposium on PLS and related methods, Barcelona*. 158

Robinson, R.W., 1977. Counting unlabeled acyclic digraphs. In : Little, C.H.C. (Ed.), *Combinatorial Mathematics V, volume 622 of Lecture Notes in Mathematics*. Springer, Berlin, pp. 28–43. 71

Rubin, D.B., 1987. *Multiple imputation for nonresponse in surveys*. Wiley, New York. 116

Rundle-Thiele, S., 2005. Elaborating customer loyalty : exploring loyalty to wine retailers. *Journal of Retailing and Consumer Services* 12, 333–344. 133

Rust, R.T., Lee, C., Valente Jr., E., 1995. Comparing covariance structure models : a general methodology. *International Journal of Research in Marketing* 12, 279–291. 33

Sahmer, K., 2006. Propriétés et extensions de la classification de variables autour de composantes latentes. Application en évaluation sensorielle. Thèse de doctorat, Université Rennes II - Universität Dortmund. 68, 150

Sahmer, K., Hanafi, M., Qannari, E.M., 2005. Assessing unidimensionality within PLS path modeling framework. In : *From Data and Information Analysis to Knowledge Engineering - Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation, University of Magdeburg*. Springer, pp. 222–229. 66, 67

Sanchez, S., Aluja, T., 2006. A PLS-PM segmentation algorithm. In : *Proceedings of KNEMO 2006, September 4-6, 2006, Capri, Italy*. 90, 173

Sanchez-Franco, M.J., 2006. Exploring the influence of gender on the web usage via partial least squares. *Behaviour & Information Technology* 25(1), 19–36. 90

Saporta, G., 2006. *Probabilités, analyse des données et statistique*, 2nde Edition. Editions Technip, Paris. 107

SAS Institute Inc., 2004a. *SAS 9.1 Macro Language : Reference*. SAS Institute Inc., Cary, NC. 12, 142, 187

SAS Institute Inc., 2004b. *SAS/STAT 9.1 User's Guide*. SAS Institute Inc., Cary, NC. 67, 150, 186, 187

Satorra, A., Bentler, P.M., 1999. Corrections to test statistics and standard errors in covariance structure analysis. In : Von Eye, A., Clogg, C. (Eds.), *Latent Variables Analysis, Applications to Developmental Research*. Thousand Oaks, Ca, Sage, pp. 399–419. 32, 36

Schafer, J.L., Graham, J.W., 2002. Missing data : our view of the state of the art. *Psychological Methods* 7(2), 147–177. 115

Scheines, R., Spirtes, P., Glymour, C., Meek, C., Richardson, T., 1998. The TETRAD project : Constraint based aids to causal model specification. *Multivariate Behavioral Research* 33(1), 65–117. 68

Schumacker, R.-E., 1981. *Spline functions, basic theory*. Wiley, New York. 102

Sigelman, L., Zeng, L., 1999. Analyzing censored and sample-selected data with tobit and heckit models. *Political Analysis* 8(2), 167–182. 123

Silva, R., 2005. Automatic discovery of latent variable models. Thèse de doctorat, School of Computer Science - Carnegie Mellon University - Pittsburgh. 68, 71, 74

Singh, J., 1989. Determinants of consumer's decisions to seek third party redress an empirical study of dissatisfied patients. *The Journal of Consumer Affairs* 23, 329–363. 133

Smith, M.A., Bolton, L.L., 1998. An experimental investigation of customer reactions to service failure and recovery encounters : paradox or peril ? *Journal of Service Research* 1(1), 65–81. 165

Spearman, C., 1904. General intelligence, objectively determined and measured. *American Journal of Psychology* 15, 201–293. 17, 59

Spearman, C., 1927. *The abilities of man*. Macmillan, New York. 59

Spirtes, P., Glymour, C., Scheines, R., 2000. *Causation, prediction, and search*. MIT Press. 68, 70, 71, 74

Spirtes, P., Scheines, R., Meek, C., Richardsonand C. Glymour, T., Hoijtink, H., Boomsma, A., 1996. *TETRAD 3 :Tools for Causal Modeling, Users Manual*. 173, 187

Squillacciotti, S., 2007. Prediction oriented classification in PLS path modelling. In : Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H. (Eds.), *Handbook of Partial Least Squares : Concepts, Methods and Applications in Marketing and related fields*. Springer. 90, 153, 173

Stan, V., Jakobowicz, E., Calciu, M., 2007. Aide à la spécification du modèle de mesure pour les modèles d'équations structurelles utilisés en marketing. In : *1ères journées de la satisfaction et de la fidélité, Grenoble, janvier*. 172

Stan, V., Saporta, G., 2005. Conjoint use of variables clustering and PLS structural equation modelling. In : Aluja, T., Casanova, J., Esposito Vinzi, V., Morineau, A., Tenenhaus, M. (Eds.), *PLS and Related Methods, proceedings of the PLS'05 International Symposium, Barcelona*. SPAD, pp. 133–140. 10, 67, 150, 181, 182

Steiger, J.H., 2000. Point estimation hypothesis testing and interval estimation using the RMSEA : Some comments and reply to Hayduk and Glaser. *Structural Equation Modeling* 7(2), 149–162. 33

Steiger, J.H., Lind, J.C., 1980. Statistically based tests for the number of common factors. In : *Annual meeting of the Psychometric Society*. 33

Steukens, S., De Ruyter, K., 2004. Reconsidering nonlinearity and asymmetry in customer satisfaction and loyalty models : An empirical study in three retail service settings. *Marketing Letters* 15(2-3), 99–111. 162

Stone, M., 1975. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B* 36(2), 111–133. 25

Streissguth, C., Bookstein, F.L., Sampson, L., Barr, H.M., 1993. *The enduring effects of prenatal alcohol exposure on child development*. University of Michigan Press, Ann Arbor. 44

Tanaka, J.S., 1993. Multifactor conceptions of fit in structural equation models. In : Bollen, K., Long, J.S. (Eds.), *Testing Structural Equation Models*. Sage, Newbury Park, CA, pp. 10–39. 31, 34

Tang, M.L., Lee, S.-Y., 1998. Analysis of structural equation models with censored or truncated data via EM algorithm. *Computational Statistics and Data Analysis* 27(1), 33–46. 123

Tax, S.S., Brown, S.W., Chandrashekaran, M., 1998. Customer evaluations of service complaint experiences : implications for relationship marketing. *Journal of Marketing* 62, 60–67. 165

Ten Berge, J.M., Socan, G., 2004. The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika* 69(4), 613–625. 66

Tenenhaus, M., 1977. Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de Statistique Appliquée* 25(2), 39–56. 105

Tenenhaus, M., 1979. La régression qualitative. *Revue de Statistique Appliquée* 27(2), 5–21. 105

Tenenhaus, M., 1998. *La Régression PLS*. Editions Technip, Paris. 24, 117

Tenenhaus, M., 1999. L'approche PLS. *Revue de Statistique Appliquée* 47(2), 5–40. 20, 26

Tenenhaus, M., 2003. Comparison between PLS and LISREL approaches for structural equation modeling : Application to the measure of customer satisfaction. In : Vilares, M., Tenenhaus, M., Coelho, P., Esposito Vinzi, V., Morineau, A. (Eds.), *PLS and Related Methods - Proceedings of the International Symposium PLS'03*. Decisia, pp. 111–126. 39

Tenenhaus, M., 2007. A bridge between PLS path modelling and ULS-SEM. In : *Proceedings of the International Symposium PLS'07, Aas, Norvège*. 14, 42, 43, 144

Tenenhaus, M., Esposito Vinzi, V., Amato, S., 2004. A global goodness-of-fit index for PLS structural equation modelling. In : *Atti de la reunion Scientifica della SIS, Barri*. pp. 739–742. 25, 85

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M., Lauro, C., 2005. PLS path modeling. *Computational Statistics and Data Analysis* 48(1), 159–205. 23, 25, 37, 41, 46, 66, 67, 75, 106, 144

Tenenhaus, M., Gonzalez, P.L., 2001. Comparaison entre les approches PLS et LISREL en modélisation d'équations structurelles : Application à la mesure de la satisfaction clientèle. In : *Compte rendu du Club SAS*. 39

Thompson, R.L., Higgins, C.A., Howell, J.M., 1994. Influence of experience on personal computer utilization : testing a conceptual model. *Journal of Management Information Systems* 11(1), 167–187. 90

Tobin, J., 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26, 24–36. 121

Trinchera, L., Squillacciotti, S., Esposito Vinzi, V., Tenenhaus, M., 2007. PLS path modeling in presence of a group structure : REBUS-PLS, a new response-based approach. In : *PLS'07 - 5th International symposium, Oslo*. 90, 173

Vanhoof, K., Swinnen, G., 1998. Attribute importance - assessing nonlinear patterns of factors contributing to customer satisfaction. *ESOMAR Publication Series* 204, 160–171. 159, 183

Venkitaraman, R.K., Jaworski, C., 1993. Restructuring customer satisfaction measurement for better resource allocation decisions : an integrated approach. In : *Fourth Annual Advanced Research Techniques Forum of the American Marketing Association*. 159

Verleye, G., Pepermans, R., Despontin, M., 1999. Missing at random data problems and maximum likelihood structural equation modeling. *Kwantitatieve Methoden* 62, 95–110. 121

Vigneau, E., Qannari, E.M., 2003. Clustering of variables around latent components. *Communications in Statistics (Simulation and Computation)* 32(4), 1131–1150. 67, 155

Vilares, M., Almeida, M., Coelho, P., 2005. Comparison of likelihood and PLS estimators for structural equation modeling. a simulation with customer satisfaction data. In : Aluja, T., Casanova, J., Esposito Vinzi, V., Morineau, A., Tenenhaus, M. (Eds.), *PLS and Related Methods, proceedings of the PLS'05 International Symposium, Barcelona*. SPAD. 39, 40, 41

Vilares, M., Almeida, M., Coelho, P., 2007. Sample size effect on the comparison of likelihood and PLS estimators for structural equation modeling : a simulation with customer satisfaction data. In : *PLS'07 International Symposium, Aas, Norway*. 39, 41

Vivien, M., 2002. Approches PLS linéaires et non-linéaires pour la modélisation de multi-tableaux : théorie et applications. Thèse de doctorat, Université Montpellier I. 100

West, S.G., Finch, J.F., Curran, P.J., 1995. Structural equation models with nonnormal variables : Problems and remedies. In : Hoyle, R.H. (Ed.), *Structural Equation Modeling : Concepts, Issues and Applications*. Sage, Thousand Oaks, CA, pp. 56–75. 32, 37

Westbrook, R.A., 1980. Intrapersonal affective influences on consumer satisfaction with products. *Journal of Consumer Research* 7, 49–54. 130

Westbrook, R.A., Oliver, R.L., 1982. Developing better measures of consumer satisfaction : some preliminary results. *Advances in Consumer Research* 8(1), 94–99. 147

Whittaker, J., 1990. *Graphical Models in Applied Multivariate Statistics*. Wiley. 19

Winsberg, S, Ramsay, J.O., 1980. Monotonic transformations to additivity using splines. *Biometrika* 67(3), 669–674. 101

Winsberg, S., Ramsay, J.O., 1981. Analysis of pairwise preference data using integrated B-splines. *Psychometrika* 46(2), 171–186. 104

Winsberg, S., Ramsay, J.O., 1983. Monotone spline transformations for dimension reduction. *Psychometrika* 48(4), 575–599. 101

Wold, H., 1966. Estimation of principal components and related models by iterative least squares. In : Krishnaiah, P.R. (Ed.), *Multivariate Analysis*. Academic Press, New York, pp. 391–420. 20

Wold, H., 1973. Non-linear iterative partial least squares (NIPALS) modelling. Some current developments. In : Krishnaiah, P.R. (Ed.), *Multivariate Analysis, Vol. III*. Academic Press, New York, pp. 383–407. 17, 20, 117

Wold, H., 1980a. *The fix-point approach to interdependent systems*. North Holland, Amsterdam. 20, 49

Wold, H., 1980b. Model construction and evaluation when theoretical knowledge is scarce. In : Kmenta, J., Ramsey, J.B. (Eds.), *Evaluation of econometric models*. Academic Press, pp. 47–74. 20, 57

Wold, H., 1982. Soft modeling : the basic design and some extensions. In : Jöreskog, K.G., Wold, H. (Eds.), *Systems under Indirect Observation*. Vol. 2. North-Holland, Amsterdam, pp. 1–54. 14, 20, 25, 26, 27, 99

Wold, H., Jöreskog, K.G., 1982. The ML and PLS techniques for modeling with latent variables : Historical and comparative aspects. In : Jöreskog, K.G, Wold, H. (Eds.), *Systems under indirect observation*. Vol. 1. North-Holland, Amsterdam, pp. 263–270. 39

Wold, S., 1992. Nonlinear partial least squares modeling II, spline inner relation. *Chemolab* 14, 71–84. 100

Wold, S., Kettaneh-Wold, N., Skagerberg, B., 1989. Nonlinear PLS modelling. *Chemometrics and Intelligent Laboratory Systems* 7, 53–65. 100

Woodruff, R.B., Cadotte, E.R., Jenkins, R.L., 1983. Modelling consumer satisfaction processes using experience-based norms. *Journal of Marketing Research* 20, 296–304. 130

Wothke, W., 1993. Nonpositive definite matrices in structural modeling. In : Bollen, K., Long, J.S. (Eds.), *Testing Structural Equation models*. Sage, Newbury Park, CA, pp. 256–293. 38

Wright, S., 1918. On the nature of size factors. *Genetics* 3, 367–374. 17

Wright, S., 1921. Correlation and causation. *Journal of Agricultural Research* 20, 557–585. 17

Yoder, K.A., 1998. Alternatives indices for testing goodness-of-fit in structural equation modeling. Mémoire de Master, Iowa State University. 31

Young, F.W., 1981. Quantitative analysis of qualitative data. *Psychometrika* 46(4), 357–388. 53, 54, 101, 105, 108, 109

Young, F.W., De Leeuw, J., Takane, Y., 1976. Regression with qualitative and quantitative variables : An alternating least squares method with optimal scaling features. *Psychometrika* 41(4), 505–529. 101, 105, 109

Young, F.W., De Leeuw, J., Takane, Y., 1978. The principal components of mixed measurement level multivariate data : An alternating least squares method with optimal scaling features. *Psychometrika* 43(2), 279–281. 101, 105

Yuan, K.H., Lambert, P.L., Fouladi, R.T., 2004. Mardia's multivariate kurtosis with missing data. *Multivariate Behavioral Research* 39(3), 413–437. 37

Zhu, H.-T., Lee, S.-Y., 1999. Statistical analysis of nonlinear factor analysis models. *British Journal of Mathematical and Statistical Psychology* 52, 225–242. 98