Méthodes pour la construction du modèle conceptuel en vue de l'application de l'approche PLS

Emmanuel Jakobowicz

Laboratoire CEDRIC, CNAM, 292 rue Saint Martin, 75141 Paris Cedex 03
EDF R&D, 1 avenue du Général de Gaulle, 92141 Clamart Cedex
emmanuel.jakobowicz@free.fr

Résumé : Les modèles d'équations structurelles à variables latentes sont de plus en plus utilisés dans le cadre de l'analyse de la satisfaction du consommateur. Dans ce domaine, c'est l'approche PLS qui domine les pratiques. Cette approche est strictement confirmative et nécessite la connaissance préalable d'un modèle conceptuel. Nous présentons dans cette communication une étude comparative des méthodes utilisées afin d'aider l'expert à mettre en place le modèle conceptuel sus-mentionné. Nous nous appliquerons à étudier chaque méthode en estimant son efficacité dans le cadre de l'application de l'approche PLS. Ces approches se sépareront en deux types, celles concernant la mise en place du modèle de mesure (relations entre les variables observées et les variables latentes) et celles concernant la mise en place du modèle structurel (relations entre variables latentes). L'ensemble de nos recherches se basent sur des données réelles issues de questionnaires de satisfaction.
Mots clés : *PLS*, approche PLS, construction de modèle, marketing.

Abstract : PLS path modeling has found increase interests since being used in the context of customer satisfaction studies. PLS path modeling is a confirmatory method and needs an initial conceptual model. In this talk, we present approaches used to build the conceptual model, compare them and try to see which one is adapted to an application of PLS path modeling. There is two types of approaches, those to build the measurement model (relationships between manifest and latent variables) and those to build the structural model (relationships between latent variables). Our research is based on datasets from Customer Satisfaction Surveys in the French power supply sector.
Keywords : PLS path modeling, model building strategy, marketing.

1 Introduction

L'approche PLS est une méthode confirmative, le modèle conceptuel doit donc être préalablement établi. Dans la majeure partie des cas, c'est à l'aide d'experts que ce modèle est mis en place. L'intérêt des méthodes présentées est de fournir une aide afin de mieux cerner les relations mises en jeu à partir de la structure des données. Tout au long de cette communication, nous nous concentrerons sur le cas de relations réflectives entre

variables manifestes et variables latentes (les variables manifestes sont un reflet du concept latent). Les approches existantes sont pour l'instant centrées sur un sous-modèle. Nous étudierons dans un premier temps les approches pour la mise en place du modèle de mesure (relations entre variables observées et variables latentes), suivra une présentation de celles permettant la mise en place du modèle structurel (relations entre variables latentes). Nous présenterons ensuite une comparaison des différentes approches basée sur des données issues de questionnaires de satisfaction des clients EDF et nous terminerons par quelques remarques sur les résultats ainsi que des ouvertures sur de nouvelles approches possibles.

2 Construction du modèle de mesure

Nous considérons que nous n'avons que les données brutes sans aucune information sur les relations entre variables manifestes et désirons obtenir des blocs unidimensionnels.

2.1 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.

2.1.1 Classification basée sur un algorithme divisif

La principale méthode de ce type est celle développée dans la procédure VARCLUS de SAS/STAT. Elle recherche 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.

2.1.2 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 $1 - |cor(x_i, x_j)|$. 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'unidimensionalité des blocs.

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

$$T = n \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{jk} cov^2(\mathbf{x}_j, \mathbf{c}_k) \text{ avec la contrainte } \mathbf{c}_k' \mathbf{c}_k = 1 \quad (1)$$

où K est le nombre de blocs, $\mathbf{c}_k$ la composante latente du bloc k. $\delta_{jk} = 1$ si x_j est dans le bloc k, 0 sinon. Cette approche est spécialement adaptée pour les modèles structurels à variables latentes.

2.2 Les réseaux bayésiens

Il pourrait sembler étrange d'utiliser des réseaux bayésiens dans notre cas. Les données étant souvent continues et la structure probabiliste n'étant pas vraiment adaptée aux modèles structurels à variables latentes. Cependant, de par leur puissance et leur lisibilité les réseaux bayésiens peuvent apporter une bonne aide à la construction du modèle. Nous avons mis en place une méthodologie en partant des données brutes. Tout d'abord, les données sont discrétisées. Le réseau est créé ensuite par le biais d'un algorithme d'apprentissage de la structure[1]. Une fois le réseau en place, l'utilisation de la distance de Kullback-Liebler nous permettra d'effectuer des associations au sein des variables manifestes déjà reliées. Cette approche n'est malheureusement pas très robuste à cause des étapes de discrétisation et d'apprentissage.

3 Construction du modèle structurel

On considère que les blocs de variables manifestes existent et donc que les variables latentes sont déjà définies. On va chercher les relations entre variables latentes.

[1]Utilisation d'un algorithme heuristique basé sur une notion de score de réseau bayésien

3.1 Les algorithmes de construction pas à pas

La méthode présentée dans Hackl (2003) est basée sur un algorithme dans lequel l'arc à ajouter est choisi de façon à maximiser le $|R|$ entre chaque paire de variables latentes. Les variables latentes sont créées comme la première composante principale du bloc de variables manifestes leur correspondant. A chaque itération, un test t de Student est effectué afin d'éliminer les relations non significatives. L'algorithme est poursuivi jusqu'à l'obtention d'un modèle stable. Cette approche, développée par Schenk, est issue d'une méthode développée par Hui (1982) avec des modifications afin d'éviter les cycles.

3.2 La méthode d'Amato (2003)

L'approche PLS ayant une vocation prédictive, il serait logique de maximiser des critères associés à celle-ci. C'est pour cette raison qu'un algorithme basé sur la redondance (F^2) et la communauté (H^2)(Tenenhaus, Esposito Vinzi, Chatelin et Lauro (2005)) a été mis au point par Amato (2003). A chaque itération de l'algorithme, on maximisera un critère dépendant de ces deux notions 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 comme le GoF ou la moyenne des redondances. L'auteur a utilisé plusieurs critères à optimiser, le principal étant :

$$B^2 = \sum \frac{B_j^2}{J}\,; \quad B_j^2 = \begin{cases} H_j^2 & \text{si } \xi_j \text{ est exogène} \\ F_j^2 & \text{sinon} \end{cases} \tag{2}$$

3.3 Les modèles libres

Cette approche développée par Derquenne et Hallais (2003) est basée sur les corrélations partielles entre les premières composantes principales de chaque bloc. Les modèles obtenus sont validés par des tests statistiques effectués sur des échantillons bootstrap.

4 La comparaison

Nous avons donc implémenté et comparé les méthodes présentées en utilisant des données issues de questionnaires de satisfaction. Le modèle conceptuel utilisé par les experts est le modèle ECSI (European Customer Satisfaction Index) sans les plaintes avec 3 variables manifestes par bloc. (Fig.1)

Dans le tableau 1, nous avons rassemblé le nombre de variables latentes et les associations pour chaque méthode dans le cas de la création du modèle de mesure. Pour chacun des modèles obtenus par les approches de construction du modèle structurel, nous avons rassemblé dans le tableau 2 le nombre de liens ainsi que 3 indices de qualité : la

redondance moyenne, la moyenne des R^2 et le GoF (Tenenhaus, Esposito Vinzi, Chatelin et Lauro (2005)).

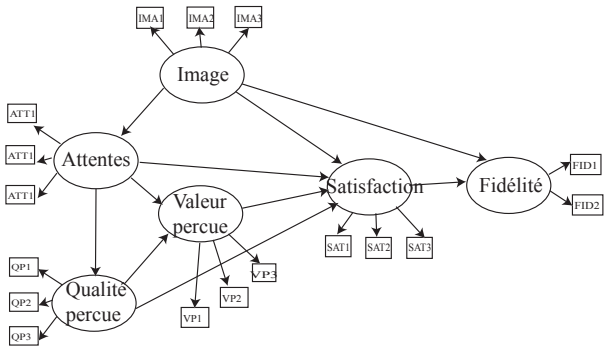

FIG. 1 – Modèle ECSI (mis au point par des experts)

Méthode	VL	Bloc1	Bloc2	Bloc3	Bloc4	Bloc5	Bloc6
ECSI	6	IMA1-3	ATT1-3	QP1-3	VP1-3	SAT1-3	FID1-2
VARCLUS	4	IMA1-3,SAT1,FID1-2	ATT1-3	QP1-3,SAT2-3	VP1-3	-	-
Vigneau et Qannari (2003)	5	IMA1-3,SAT1	ATT1-3	QP1-3,SAT2-3	VP1-3	-	FID1-2
Stan et Saporta (2005)	5	IMA1-3,SAT1	ATT1-3	QP1-3,SAT2-3	VP1-3	-	FID1-2
Réseaux Bayésiens	5	IMA1-3,SAT1	ATT1-3	QP1-3,SAT2-3	VP1-3	-	FID1-2

TAB. 1 – Récapitulatif des groupes obtenus pour former les blocs de variables manifestes

Méthode	Liens	Redondance moy.	R^2 moy.	GoF
ECSI	10	$0,303$	$0,365$	$0,498$
Hackl (2003)	9	$0,466$	$0,384$	$0,563$
Amato (2003)	7	$\mathbf{0,567}$	$\mathbf{0,458}$	$\mathbf{0,619}$
Modèles libres	10	$0,433$	$0,357$	$0,542$

TAB. 2 – Indices associés aux modèles structurels obtenus

5 Conclusion et discussions

Pour le modèle de mesure, il apparaît qu'il est extrêmement difficile de trouver un critère global afin de différencier les modèles obtenus, le modèle structurel n'étant pas encore connu, il est impossible d'obtenir des indices globaux associés à l'approche PLS. Chaque méthode présentée optimise différents critères. Pour se rapprocher de l'approche PLS, on aura tendance à privilégier l'approche de classification de variables autour de composantes latentes (Vigneau et Qannari (2003)). Rien ne nous permet cependant de considérer que le modèle obtenu est le meilleur. Les réseaux bayésiens offrent une autre

solution qui permettra de mieux visualiser les associations avec les désavantages cités précédemment. Il ressort que les modèles obtenus lors de l'application sont très proches du modèle expert et ne se différencient que très peu d'une approche à l'autre. Cette application nous montre que les notions d'image, de satisfaction et de fidélité ont tendance à se confondre.
Pour le modèle structurel, les approches maximisant les $|R|$ sont avantageuses d'un point de vue calculatoire. Elles ne sont, par contre, pas aussi satisfaisantes que la méthode d'Amato (2003), qui explore plus profondément la structure du modèle, et que les modèles libres qui prennent en compte les corrélations partielles. Il faut être prudent avec l'interprétation des très bons résultats de la méthode d'Amato (2003) car cette approche optimise directement les critères de comparaison. Les modèles obtenus sont assez proches avec une tendance à la simplicité par rapport au modèle expert (les indices de qualité sont meilleurs que dans le cas du modèle ECSI).

Cette rapide comparaison nous montre que ces méthodes obtiennent des résultats proches et que c'est le choix du critère de qualité qui entraînera la supériorité d'une approche. Il apparaît que chaque partie du modèle relève de méthodes très différentes n'optimisant pas les mêmes critères. L'approche PLS, du fait de son approche différenciée entre chaque sous-modèle, nécessite cette séparation. Des méthodes combinant les deux types d'approches sont toutefois envisageables.

Bibliographie

[1] Amato, S. (2003) A model building strategy for PLS path modeling. *Actes du Symposium International PLS'03*, Lisbonne, 135–141.
[2] Derquenne, C. et 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.
[3] Hackl, P. (2003) Specification Analysis of Structural Equation Models. *Actes du Symposium International PLS'03*, Lisbonne, 127–134.
[4] Hui, B.S. (1982) On building partial least squares models with interdependent inner relations. *Systems under indirect observation*, North-Holland, vol.2, 249–271.
[5] Stan, V. et Saporta, G. (2005) Conjoint use of variables clustering and PLS structural equation modelling. *Actes du Symposium International PLS'05*, Barcelone, 133–140.
[6] Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M. et Lauro, C. (2005) PLS path modeling. *Computational Statistics and Data Analysis*, 48(1), 159–205.
[7] Vigneau, E. et Qannari, E.M. (2003) Clustering of variables around latent components. *Communications in Statistics (Simulation and Computation)*, 32(4), 1131–1150.