Comparaison de groupes d'observations dans le cadre de l'approche PLS

Emmanuel Jakobowicz

CEDRIC, CNAM, 292 rue Saint Martin, 75141 Paris Cedex 03, France et EDF R&D, 1 avenue du Général de Gaulle, 92121 Clamart Cedex, France

Résumé

Les modèles d'équations structurelles à variables latentes sont de plus en plus utilisés dans de nombreux domaines. L'approche PLS permet d'estimer les paramètres de ce type de modèles. A l'inverse de la méthode par analyse de la structure de covariance (aussi appelée LISREL), on ne peut pas se rapporter à un test de qualité d'ajustement paramétrique comme celui basé sur la distribution du χ^2. Il est donc difficile de comparer des sous populations en se basant sur un modèle dans le cadre de l'approche PLS. Nous présentons un processus afin de comparer deux échantillons en se basant sur un modèle d'équations structurelles. Dans ce but, nous partons d'une comparaison globale pour aller jusqu'à la comparaison des coefficients structurels du modèle. Puis, nous présentons des applications dans le cadre de l'analyse de la satisfaction des clients.

Mots-clés : Analyse de données - Marketing.

Abstract

Structural equation models are widely used in various fields. These models can be estimated using PLS path modeling in which, unlike covariance structure analysis, model fit cannot be estimated using a test based on distributional assumptions. It is indeed difficult to make multi-group comparison in that framework. We present a two-sample comparison process adapted to PLS path modeling. It begins with a global models quality comparison and follows until structural coefficients comparison. Applications on customer satisfaction data are then presented.

Key-words : Data analysis - Marketing.

1 Introduction

La comparaison d'échantillons 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 problématique est grand et les interactions entre ceux-ci sont nombreuses. On ne pourra pas simplement comparer deux coefficients indépendamment du reste des paramètres comme cela se fait largement dans la pratique et même dans certains articles de recherche (Thompson, Higgins et Howell (1994)).
Lors de l'application d'estimation par la structure de covariance (LISREL), l'utilisation d'indices basés sur la distribution du χ^2 permet de comparer des modèles sur des données différentes, ce qui est impossible avec l'approche PLS (pour une présentation des indices, voir Bollen et Long (1993), pour une méthodologie complète, voir Liao (2002) et sur l'approche PLS, voir Tenenhaus, Esposito Vinzi, Chatelain et Lauro (2005)).
Dans le cadre de cette communication, nous nous limitons au cas de données indépendantes non appariées.

Nous présentons un processus de comparaison de deux échantillons basés sur le même modèle. Quelques applications sur des données réelles sont ensuite présentées puis nous terminons par des conclusions et des ouvertures.

2 Vers un processus de comparaison

La comparaison d'échantillons dans le cadre de l'approche PLS doit s'affranchir des notions paramétriques largement associées à la méthode LISREL. C'est pour cette raison que nous basons la majorité des tests inclus dans le processus sur des méthodes de rééchantillonnage. Suivant les tests, nous utilisons soit des tests de permutation (Edgington (1987)), soit du bootstrap (Efron et Tibshirani (1993)).

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 d'échantillons proches et que les résidus soient distribués normalement.

2.1 Préalables

Nous nous inspirons des différents points de comparaison développés dans Liao (2002). Lorsque l'on compare des coefficients structurels, si la structure des données n'est pas la même que celle du modèle ou est très différente d'un échantillon à un autre, alors toute conclusion sur les variations de ces coefficients, même validée par rééchantillonnage n'a pas de valeur.
Préalablement à l'application de ce type de procédures, des tests sur l'adéquation des données au modèle et sur les différences au niveau des structures des échantillons doivent

être menés. En fonction des résultats de chacune des séries de tests, les tests suivants pourront être envisagés.

Nous commencerons par introduire quelques notations. Dans le cadre de modèles d'équations structurelles à variables latentes, le modèle peut être représenté par deux équations :

$$\mathbf{X} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\epsilon} \tag{1}$$

$$\boldsymbol{\xi} = \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{2}$$

où $\mathbf{X}$ rassemble l'ensemble des variables manifestes, $\mathbf{\Lambda}$ est une matrice rassemblant les "loadings", $\boldsymbol{\xi}$ est une matrice rassemblant les variables latentes et $\mathbf{\Gamma}$ est une matrice de coefficients structurels. Les coefficients structurels que nous étudions par la suite sont notés γ_{ij}.

2.2 Première étape : La structure des données

Avant toute comparaison, les conditions d'application de l'approche PLS doivent être vérifiées pour chacun des échantillons (indépendance des observations et consistance interne pour le cas réflectif).
Afin d'évaluer les différences au niveau du modèle conceptuel, nous utilisons des indices globaux de qualité prédictive. Par le biais d'un test non paramétrique d'égalité de ces indices, nous pouvons vérifier que les données s'adaptent de façon similaire au modèle. Les deux indices que nous utilisons sont :

- La communauté qui représente la part de variance expliquée des variables latentes par le modèle externe

$$H_j^2 = \frac{1}{p_j}\sum_{i=1}^{p_j} cor^2(\mathbf{x}_{ji}, \mathbf{y}_j), \tag{3}$$

  où $\mathbf{y}_j$ est le score de la variable latente $\boldsymbol{\xi}_j$.
- Le GoF (Tenenhaus, M., Esposito Vinzi, V. et Amato, S. (2003)) qui est une combinaison d'indicateurs de la validité du modèle interne et d'indicateurs de la validité du modèle externe

$$GoF = \sqrt{\bar{communaute} \times \bar{R}^2}. \tag{4}$$

Soit G_1 et G_2 deux échantillons, nous utilisons un test de permutation basé sur l'hypothèse nulle :

$$H_0 : \bar{H}^2_{G_1} = \bar{H}^2_{G_2}, GoF_{G_1} = GoF_{G_2}$$

Comme ces indices mesurent la qualité prédictive du modèle, si on peut considérer que ces indices sont proches pour deux échantillons, alors une étude plus poussée des coefficients du modèle peut être menée. Dans le cas contraire, on peut créer un modèle conceptuel alternatif s'adaptant bien aux deux échantillons comme le font Amato et Balzano (2003). Ceci se fait en utilisant des méthodes de construction de modèles.

2.3 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 se basent sur des méthodes de rééchantillonnage.

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. On devra utiliser une formule du type :

$$t = \frac{\gamma_{ij}^{G_1} - \gamma_{ij}^{G_2}}{[\sqrt{\frac{(n_1-1)^2}{n_1+n_2-2} SE_{G_1}^2 + \frac{(n_2)^2}{n_1+n_2-2} SE_{G_2}^2}][\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}]} \tag{5}$$

   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 :

$$t = \frac{\gamma_{ij}^{G_1} - \gamma_{ij}^{G_2}}{\sqrt{SE_{G_1}^2 + SE_{G_2}^2}};\ DF = \frac{(SE_{G_1}^2 + SE_{G_2}^2)^2}{\frac{SE_{G_1}^2}{n_1+1} + \frac{SE_{G_2}^2}{n_2+1}} - 2 \tag{6}$$

   avec un nombre de degré 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.* Chin (2003) propose une approche non paramétrique basée sur un test de permutation afin de valider la significativité des résultats. Pour deux échantillons G_1 et G_2, l'égalité du coefficient structurel étudié γ_{ij} est testé, l'hypothèse nulle est $H_0 : \gamma_{ij}^{G_1} = \gamma_{ij}^{G_2}$. Ce test est basé sur une permutation de l'ensemble des données suivi de l'application de l'approche PLS avant de comparer les coefficients obtenus. Pour plus de détails sur les tests de permutation, on peut voir Edgington (1987).

3 Application

3.1 Les données

Nous utilisons un questionnaire de satisfaction des clients d'EDF. Nous effectuons deux comparaisons sur des clients EDF ayant des caractéristiques différentes. La première est

basée sur le sexe de l'interviewé et la seconde sur le sentiment par rapport à l'ouverture du marché (favorable/défavorable). Le modèle utilisé est un modèle expert mis en place dans le cadre de cette communication. Il est composé de 5 variables latentes, possédant chacune entre 2 et 10 variables manifestes (soit 27 variables manifestes). La taille de l'échantillon global est de 1988 observations. Nous utilisons le mode A (cas réflectif) pour l'estimation basée sur le modèle externe, et le schéma centroïde pour celle basée sur le modèle interne (figure 1). L'ensemble des tests est effectué à partir de macros SAS développées par l'auteur.

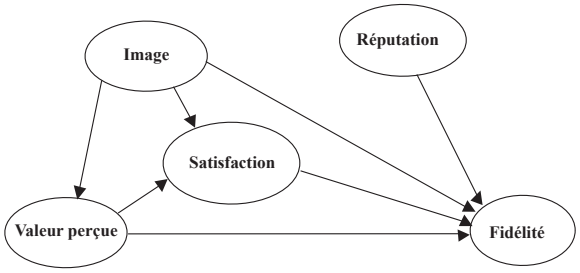

FIG. 1 – Modèle interne conceptuel

3.2 Résultats

Nous rassemblons dans le tableau 1 les résultats des étapes de la comparaison en nous focalisant sur le coefficient structurel entre la satisfaction et la fidélité, qui sont deux concepts clés. Nous utilisons 1000 itérations pour les permutations.

Echantillons	Hommes/Femmes	Concurrence
Test sur les H^2	H_0 acceptée (0.191)	H_0 rejetée (0.019)
Test sur les GoF	H_0 acceptée (0.479)	H_0 rejetée (0.062)
Test sur les variances	Accepté	-
Normalité des résidus	Non normaux	-
Taille d'échantillon	751/1237	877/1111
Choix du test	Test de permutation	Test de permutation
Résultat de la comparaison	H_0 acceptée (0.542)	H_0 acceptée (0.344)

TAB. 1 – Résultat des processus de comparaison (entre parenthèses les p-valeurs)

La première comparaison montre que les indices globaux ne sont pas significativement différents entre les hommes et les femmes alors qu'ils le sont entre les clients pour et ceux contre l'ouverture. Les résidus déviant fortement de la normalité dans les données initiales (aplatissement élevé), nous avons décidé d'appliquer le test de permutation sur le lien satisfaction - fidélité. Il ressort que le coefficient structurel entre satisfaction et fidélité n'est pas significativement différent entre les hommes et les femmes. Nous présentons, à titre

d'exemple, le résultat en rapport avec la concurrence. Ce test appliqué sur des échantillons largement différents au niveau de la qualité globale donne un coefficient structurel égal suivant le type de client. Ce résultat afin d'être validé nécéssiterait la mise en place d'un nouveau modèle mieux adapté aux deux groupes d'observations.

4 Conclusions

Dans cette communication, nous introduisons un processus de comparaison simple dans le cadre de l'application de l'approche PLS. La comparaison des coefficients est souvent effectuée hâtivement sans aucune recherche préalable, nous conseillons donc aux utilisateurs de vérifier si leurs données sont réellement comparables et, dans ce cas, par quelle méthode.
Nous n'approfondissons pas ici le cas de modèles différents qui pourra donner lieu à de plus amples recherches, notamment dans le cadre de la classification de modèles. Il serait intéressant d'autre part de se pencher sur des données réparties dans le temps, nous n'avons pas pu approfondir ce point en raison du manque de données de ce type.

Bibliographie

[1] Amato, S. et Balzano, S. (2003) Exploratory approaches to group comparison in PLS Path Models, actes du Symposium International PLS'03, 443–451.
[2] Bollen, K.A. et Long, J.S. (1993) *Testing Structural Equation Models*, Sage.
[3] Chin, W.W. (2003) A permutation procedure for multi-group comparison of PLS models, actes du Symposium International PLS'03, 33–43.
[4] Edgington, E.S. (1987) *Randomization tests*, Second Edition, Marcel Dekker, Inc.
[5] Efron, B. et Tibshirani, R.J. (1993) *An introduction to the bootstrap*, Chapman and Hall.
[6] Liao, T.F. (2002) *Statistical Group Comparison*, Wiley.
[7] SAS Institute Inc. (2004) *What's New in SAS 9.0, 9.1, 9.1.2, 9.1.3*, Online Documentation.
[8] Tenenhaus, M., Esposito Vinzi, V. et Amato, S. (2005) A global goodness-of-fit index for PLS structural equation modelling, atti de la reunion Scientifica della SIS, Barri, 739–742.
[9] Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M. et Lauro, C. (2005) PLS path modeling, Computational Statistics and Data Analysis, 48(1), 159–205.
[10] Thompson, R.L., Higgins, C.A. et Howell, J.M. (1994) Influence of experience on personal computer utilization : testing a conceptual model, Journal of Management Information Systems, 11(1), 167–187.
[11] Wold, H. (1982) Soft modeling : the basic design and some extensions, in System under indirect observation, vol. 2, North-Holland, Amsterdam, 1–54.