Recherche d'une hiérarchie de variables ordinales pour modéliser une variable ordinale

Christian Derquenne

Recherche d'une hiérarchie de variables ordinales pour modéliser une variable ordinale

Christian Derquenne

EDF R&D - 1,avenue du Général de Gaulle - 92141 Clamart Cedex - France

Résumé : Dans de nombreuses applications, notamment en marketing, la modélisation d'un phénomène demande de connaître les variables qui l'influencent, mais aussi celles qui contribuent le plus et notamment leur ordre d'importance. Dans ce papier, nous proposons une méthode pour hiérarchiser des variables ordinales explicatives d'une variable réponse ordinale. Le principal problème est le nombre de variables candidates à l'explication. Une première réponse peut être fournie par des approches de sélection de variables pas à pas, mais cela ne règle pas le problème d'ordre de contribution des variables car il demande d'analyser l'ensemble des permutations des variables retenues. La méthode proposée utilise la statistique d'ordre, le modèle logit ordinal et l'algorithme du recuit simulé.

Abstract : In many applications, in marketing for instance, the modelling of a response variable has not only a goal to know the main explanatory variables, but also the most contributive and the top that the importance order. In this paper, we propose a method to put in order explanatory ordinal variables to modelize a response ordinal variable. The main problem is the number of explanatory variables. A first answer is provided by stepwise approach to select variables, but it does not offer a solution concerning order of explanatory variables. Indeed, it must analyse all the permutations of these last ones. The proposed method uses order statistic, ordinal logit model and simulated annealing algorithm.

Mots-clés : Modèles semi-paramétriques et non paramétriques, marketing.

1 Contexte, motivations et objectif

Dans le cadre de l'étude du comportement de la clientèle et plus précisément dans les enquêtes de satisfaction de nombreux types d'études sont réalisés. En particulier, il peut être intéressant pour l'expert en marketing de connaître les principaux facteurs et leurs poids qui influencent le marché. En effet, cela peut l'aider à construire sa stratégie en fonction de cette hiérarchie. Une solution raisonnable, en terme d'aide à la décision, est de rechercher les variables explicatives les plus contributives qui jouent un rôle sur un facteur global et de construire des poids normés associés, par exemple la satisfaction globale à l'égard d'un produit ou d'un service. Dans ce cas, les facteurs explicatifs correspondront à des variables de satisfaction partielle. Dans ce papier, nous restreindrons notre champ aux variables ordinales car cette échelle est généralement la plus utilisée dans les applications.

Le principal problème pour hiérarchiser la contribution de variables expliquant un facteur global est leur nombre. En effet, si les techniques de régression pas à pas offrent une première réponse pour sélectionner les facteurs les plus explicatifs, elles ne fournissent pas une permutation optimale de ces variables. Nous proposons dans ce papier, une méthode qui fournit une hiérarchie optimale des variables explicatives ordinales pour expliquer la variable de réponse ordinale. Cette méthode est fondée sur la statistique d'ordre, le modèle logit ordinal, ou éventuellement sur le modèle linéaire gaussien du type analyse de la variance à un facteur, et l'algorithme du recuit simulé.

2 Méthode proposée

Soit Y le facteur global (la variable réponse) distribué sur une échelle ordinale à R réponses et soient $X_1, ..., X_p$, les facteurs partiels (les variables candidates à l'explication) également ordinaux avec $m_1, ..., m_p$ modalités.

Une première solution (bivariée) est de rechercher tout d'abord une corrélation entre Y et chacune des p variables. Le Tau-b de Kendall ou le Tau-c de Stuart (1990) peuvent être utilisés car ils possèdent de bonnes propriétés statistiques pour traiter les modalités ex-aequo existant naturellement dans les tableaux de contingence ordonnés. Afin de juger de la qualité de la force de corrélation entre les deux variables ordinales, il est nécessaire de diviser le coefficient de corrélation par son écart-type asymptotique. Plus ce rapport est élevé en valeur absolue, plus la liaison est forte. La hiérarchie peut paraître évidente dans certains cas, mais il n'en est pas toujours de même. De plus l'apport individuel de chaque variable en tenant compte des autres n'est pas pris en compte, ce qui sera d'autant plus pénalisant si le nombre de variables est élevé. Pour pallier ce problème, il est possible d'utiliser des méthodes de sélection de variables pas à pas dans un modèle de régression (ici un modèle logit ordinal, par exemple). Un nombre restreint de variables explicatives peut alors être obtenu grâce à différents critères d'arrêt, tel que le chi-2 d'adéquation entre les probabilités cumulées estimées et observées, ou encore les tests marginaux du rapport de vraisemblances pour chaque variable candidate à l'explication. Nous pouvons obtenir une hiérarchie des variables sélectionnées en ordonnant de façon croissante les p-valeurs associées à ce test. Cependant, cette solution n'est pas satisfaisante si le praticien désire garder l'ensemble des variables candidates à l'explication pour mieux juger de leur apport mutuel. Il peut alors être intéressant de construire une démarche statistique pour "estimer" l'ordre d'importance de toutes les variables $X's$. Un moyen raisonnable est de calculer une statistique pour chacune des permutations des p variables permettant d'évaluer la force de l'ordre des variables par rapport à Y et donc fournir une nouvelle hiérarchie des variables.

La démarche proposée est la suivante : Pour un ordre de variables $X's$ donné, dont les modalités sont triées de façon croissante, le rang 1 est attribué au premier p-uplet, puis ce rang est incrémenté de 1 dés que le p-uplet change. Ce procédé est réalisé pour

l'ensemble des n individus et fournit donc une variable de rang résumant une permutation donnée des p variables. Cette démarche repose sur les quatre propriétés suivantes.

Propriété 1 : Le nombre minimum de rangs différents est égal au nombre de modalités de la variable X_j qui possède le plus de modalités, alors que le nombre maximum de rangs est égal au produit des $m_1, ..., m_p$ modalités, indépendamment de la distribution de Y.

Propriété 2 : Le nombre de rangs différents sera le même quelle que soit la permutation, independamment de la distribtion de Y.

Ces deux propriétés permettent de raisonner sur le même support de référence en terme de distribution de la variable de rang quelle que soit la permutation, ce qui offre la possibilité de comparer les différentes valeurs des statistiques obtenues pour chaque permutation.

Propriété 3 : Les distributions des rangs pour chacune des modalités de la variable Y sont généralement différentes d'une permutation à l'autre.

Cette propriété permet de proposer la règle suivante : *pour chaque modalité ordonnée de la variable Y, moins les distributions des rangs seront dispersées et plus les centres seront différents, alors plus la hiérarchie des variables $X's$ sera meilleure en terme de corrélation avec Y*. Le modèle linéaire gaussien d'analyse de la variance à un facteur, représenté ici par Y et dans lequel le rang sera la variable réponse, semble le plus adapté pour répondre à la règle émise précédemment. La statistique de Fisher est alors choisie pour calculer la force de liaison entre Y et la variable de rang (permutation des $X's$).

Propriété 4 : La distribution de Y pour chaque rang est généralement différente d'une permutation à l'autre.

Cette propriété permet de proposer la règle suivante : *moins la distribution de Y par rang ordonné sera dispersée et plus les fonctions de répartition seront différentes alors plus la hiérarchie des variables $X's$ sera meilleure en terme de corrélation avec Y*. Le modèle logit ordinal parait le plus raisonnable pour répondre à la règle précédente. Dans ce modèle la variable réponse est cette fois-ci Y, alors que la seule variable explicative est le rang. La statistique utilisée est celle du rapport des vraisemblances estimées avec et sans la variable de rang.

Si ces deux statistiques paraissent adaptées pour faire apparaître la qualité d'une hiérarchie, il n'en reste pas moins que le fléau de la dimension liée au nombre de hiérarchies possibles subsiste. En effet, ce nombre est tout d'abord composé des $p!$ permutations des p variables entrant dans la hiérarchie, mais aussi de l'ordre croissant ou décroissant de chacune des p variables ordinales avec 2^p possibilités. Cependant, cet ordre est symétrique. En effet, si les p variables sont toutes triées par ordre croissant, alors les résultats seront identiques à un tri par ordre décroissant des p variables. Cela permet de réduire de moitié le nombre de possibilités, c'est-à-dire que le nombre total de hiérarchies est finalement de $2^{p-1}p!$.

Les solutions proposées dans la littérature sont nombreuses et tirent leurs premières applications du trajet du voyageur de commerce. En effet, s'il doit passer par n villes, quel est le

plus court chemin ? Il y a alors $n!$ chemins possibles et le critère à optimiser (à minimiser) est la distance parcourue. Dans notre cas, le critère à optimiser est soit le F de Fisher, soit le rapport de vraisemblances. Ces deux statistiques seront alors à maximiser. Pour cela, la solution généralement proposée est de raisonner de façon stochastique, c'est-à-dire de calculer une probabilité de continuer ou de remettre en cause la solution proposée. Nous avons choisi une méthode classique qui a fait ses preuves : le recuit simulé (Simulated Annealing) introduit par Kirkpatrik (1983).

3 Application à une enquête de satisfaction

Une enquête de satisfaction à l'égard de différents produits et services a été réalisée auprès de 923 personnes. Nous avons sélectionné une partie du questionnaire relatif à un produit A. Celle-ci comporte une question de satisfaction globale et 16 questions de satisfaction partielle à l'égard de ce produit. Nous nommerons Y la variable réponse à la satisfaction globale et $X_1, ..., X_j, ..., X_{16}$, les variables candidates à l'explication. Comme nous pouvons le constater le nombre de possibilités est très élevé : $2^{15}16!$=6,86E+17 hiérarchies possibles dont on comprend aisément qu'il serait déraisonnable de calculer toutes les solutions.

Nous avons appliqué les deux méthodes de permutations proposées avec la statistique de Fisher, nommée "perm-anova", et avec la statistique du rapport de vraisemblances, nommée "perm-logit", que nous allons comparer à d'autres méthodes décrites ci-après.

(1) Un modèle logit ordinal classique où Y est la réponse et les p variables candidates à l'explication sont toutes incluses dans le modèle.

(2) La régression séquentielle pas à pas développée par Bachelet (1996). Cette méthode consiste à réaliser une régression pas à pas descendante jusqu'à épuisement des p variables (les variables non significatives sont également retenues à la fin), une hiérachie est alors obtenue et des poids normés associés à chaque variable sont calculés par rapport à la variable la plus significative.

(3) Un réseau Bayesien, voir Pearl (1985), est un graphe orienté sans circuit, associé à une loi de probabilité dans lequel les noeuds représentent des variables probabilisées, et les liens, des relations de dépendances probabilistes. Dans notre cas, Y est le noeud cible et les $X's$ sont les noeuds enfants.

(4) La régression PML (Partial Maximum Likelihood), développée par Derquenne (2004) pour des variables catégorielles, est une extension de la régression PLS (Partial Least Squares) introduite par Wold et al. (1983). Ce type de régression permet de s'affranchir du problème de multicolinéarité qui peut exister entre les variables candidates à l'explication. Ce qui est très intéressant dans notre cas, car les variables de satisfaction partielle sont généralement liées entre elles. Les résultats obtenus fournissent pour chaque variable un poids normé représentant sa contribution à la modélisation de Y.

(5) Classification de variables et approche PML. Comme nous l'avons indiqué pour la régression PML, les variables candidates à l'explication dans le cas de réponses subjectives, comme la satisfaction, peuvent être très liées entre elles et même représenter une seule et même dimension. Cependant s'il y a multidimensionnalité, la qualité de la modélisation et de la prévision de Y risque d'être moins bonne. Dans ce cas, il peut être judicieux de procéder au préalable à une classification de variables (cf. Derquenne, (1998)) pour obtenir un certain nombre de nouvelles dimensions contenant les variables initiales. Puis, sur la structure de classes obtenue, nous avons appliqué l'approche PML développée par Derquenne (2005) pour des variables de tout nature (notamment ordinale), qui est une extention de l'approche PLS introduite par Wold (1982). Dans ce cas, les p variables de satisfaction partielle et la satisfaction globale correspondent aux variables manifestes et leurs relations avec leurs variables latentes (les dimensions construites) constituent le modèle de mesure. Ici le modèle de structure met en relation chaque variable latente relative aux groupes de variables de satisfaction partielle avec la variable latente de la satisfaction globale. La satisfaction globale représente à elle seule une variable latente.

(6) Classification de variables et approche RFGPC (Regression on the First Generalized Principal Components). La démarche est la même que la précédente, à part que nous utilisons l'approche RFGPC développée par Derquenne (2005) pour des variables catégorielles qui est une extension de l'approche RFPC introduite par Derquenne et al. (2002). Cette méthode permet de s'affranchir du problème de tests biaisés lors de l'estimation finale du modèle de structure qui est présent dans l'approche PLS.

Les ordres des facteurs partiels de satisfaction obtenus par perm-anvova, perm-logit et la régression PML sont en total accord sur les trois premières variables : X_5, X_{14}, X_{12}. La régression logistique ordinale fournit quasiment le même ordre, en placant la variable X_{12} en quatrième position. L'ordre obtenu avec perm-logit est $X_5, X_{14}, X_{12}, X_9, X_3, X_{11}, X_6, X_2, X_{15}, X_7, X_{16}, X_1, X_8, X_{10}, X_{13}, X_4$. Celui-ci est relativement proche de celui fourni par le modèle logit ordinal. Cela peut éventuellement s'expliquer par le fait que dans les deux cas le critère à optimiser est identique, à savoir le rapport de vraisemblances, même si d'une part, la seule variable explicative est le rang pour notre approche et d'autre part l'ensemble des $X's$ pour la régression logistique classique. Enfin, la variable X_5 est systématiquement placée en première ou en deuxième position quelle que soit la méthode.

4 Apports, limites, applications et voies futures

La méthode proposée de hiérarchisation par permutation des variables de satisfaction partielle pour "expliquer" la satisfaction globale est très prometteuse, car elle correspond à une approche semi-paramétrique. En effet, l'utilisation des rangs possède le coté "non paramétrique", alors que le modèle d'analyse de la variance ou le modèle logit représentent le coté "paramétrique" puisque des coefficients sont estimés et reposent d'une part sur le

modèle linéaire gaussien usuel, et d'autre part sur le modèle linéaire généralisé avec toutes leurs contraintes. De plus, le coté multivarié des régressions logistique, séquentielles et PML, des approches PML et RFGPC est également présent dans les deux méthodes proposées puisque les rangs tiennent compte de l'ensemble des variables $X's$. Par ailleurs, le nombre très élevé de permutations possibles pour tenter de trouver la "meilleure" pourrait être un gros inconvénient de cette méthode. Mais l'algorithme du recuit simulé proposé est une solution raisonnable pour pallier ce problème. Enfin, l'inconvénient qui subsiste pour cette méthode est pour l'instant l'impossibilité de proposer une solution satisfaisante pour calculer un poids de contribution de chaque variable de satisfaction partielle à la satisfaction globale. Des recherches sont actuellement en cours pour estimer ces poids.

Bibliographie

[1] Bachelet D., (1996), La mesure de la satisfaction du consommateur ou la chaîne, l'arbre et la cascade, *ESOMAR, Customer Satisfaction*, 199-228, Madrid, Espagne.

[2] Derquenne Ch., (1998), Clustering of Mixture of Variables to Research Significant Ideas in Electric Heating Survey, IFCS, Roma, Italy.

[3] Derquenne Ch., Hallais C., (2002), Une méthode alternative à l'approche PLS : Comparaison et application aux modèles conceptuels marketing, *XXXVièmes Journées de Statistique*, Bruxelles, Belgique.

[4] Derquenne Ch., (2003), A Multivariate Modeling Method based on the Partial Maximum Likelihood, *Third International Symposium on PLS and Related Methods*, Lisbon, Portugal.

[5] Derquenne Ch., (2005), Generalized Path Modelling based on the Partial Maximum Likelihood Approach, *PLS'05, Fourth International Symposium on PLS and Related Methods*, Barcelona, Spain.

[6] Kendall M., Gibbons J. D., (1990), *Rank Correlation Methods*, Fifth Edition, Edwards Arnold, London, Melbourne, Auckland.

[7] Kirkepatrick S., Gelatt C., Vecchi M., (1983), Optimization by Simulated Annealing, *Science*, 20, pp. 671-680.

[8] Pearl J., (1980), Bayesian Networks: A Model of Self-activated Memory for Evidential Reasoning, *In proceedings of the 7th Conference of the Cognitive Science Society*, University of California, Irvine, CA, pp. 329-334.

[9] Wold H., (1982), Soft modeling: the basic design and some extensions, In: K.G. Jöreskog & H. Wold (eds.), *Systems under Indirect Observations: Causality, Structure, Prediction*, Vol 2, 1-54, Amsterdam, North-Holland.

[10] Wold S., Albano C., Dunn III W.J., Esbensen K., Hellberg S., Johansson E., Sjöström H., (1983), Patter Recognition: Finding and Using Regularities in Multivariate Data in *Proc. IUFOST Conf. "Food Research and Data Analysis"*, Martens J. (Ed.), Applied Sciences publications.