REVUE DE STATISTIQUE APPLIQUÉE

S. BENAMMOU
S. HARBI
G. SAPORTA

Sur l'utilisation de l'analyse conjointe en cas de réponses incomplètes ou de non-réponses

Revue de statistique appliquée, tome 51, n° 4 (2003), p. 31-55.
<http://www.numdam.org/item?id=RSA_2003__51_4_31_0>

Rev. Statistique Appliquée, 2003, **LI** (4), 31-55

SUR L'UTILISATION DE L'ANALYSE CONJOINTE EN CAS DE RÉPONSES INCOMPLÈTES OU DE NON-RÉPONSES

S. BENAMMOU*, S. HARBI *, G. SAPORTA**

** Département de Méthodes quantitatives,*
Faculté de Droit des Sciences Économiques et Politiques de Sousse (Tunisie)
saloua.benammou@fdseps.rnu.tn isnihr@voila.fr

*** Chaire de Statistique appliquée et CEDRIC,*
Conservatoire National des Arts et Métiers, Paris
saporta@cnam.fr

RÉSUMÉ

L'analyse conjointe cherche à expliquer une variable qualitative ordinale en fonction de plusieurs variables qualitatives moyennant une régression multiple par interviewé. Elle se trouve handicapée lorsque certains classements sont manquants. Nous nous intéressons donc dans cet article aux cas où la variable à expliquer présente un ordre partiel uniquement. Nous proposons ici une procédure d'estimation des valeurs manquantes, et montrons sur un exemple que si le nombre de ces dernières n'est pas trop élevé les données restent utilisables, et les résultats proches de ceux de l'ordre complet. Nous confirmons ces résultats par simulation.

Mots-clés : *Analyse conjointe, valeurs manquantes, ex aequo, trade-off*

ABSTRACT

Conjoint analysis tries to explain an ordinal variable according to several categorical variables using a multiple regression. It is handicapped when some of classification-ranks are missing. We are interested in this paper, in the case where the response presents a partial order only and propose a procedure of estimation of the missing values. We show on an example that if the number of missing values is not too large, data remain useful, and results are close to those of the complete order. We confirm these results by simulations.

Keywords : *Conjoint analysis, trade-off, missing values, ex æquo*

1.Introduction

L'analyse conjointe est tout d'abord une méthode d'analyse des données destinée à relier une variable à expliquer ordinale à plusieurs variables explicatives qualitatives à modalités ordonnées ou non. En ce sens l'analyse conjointe est une forme particulière du modèle linéaire général. Dans une optique plus générale l'analyse conjointe est une méthodologie d'étude complète comportant une phase de recueil des observations basée sur un choix de combinaisons de modalités de variables déterminé par application de la technique des plans d'expériences, une phase de traitement

destinée à estimer les paramètres et une phase de simulation. Le domaine privilégié d'application se situe en marketing et on parle alors de méthodes de «Trade Off». Il s'agit d'expliquer les préférences de consommateurs pour des produits définis par des combinaisons d'attributs. Chaque consommateur enquêté note ou classe un ensemble de produits réels ou hypothétiques. L'analyse conjointe décompose alors les préférences selon un modèle d'utilité additive propre à chaque interviewé.

Le terme de «Trade Off» évoque un modèle compensatoire où une bonne note sur un attribut peut être compensée par une note moins bonne sur un autre attribut : le consommateur effectue son choix de produits en mettant en balance avantages et inconvénients.

2. Rappel sur l'analyse conjointe

Soient p variables qualitatives X_1, X_2, ..., X_p à m_1, m_2,...,m_p modalités respectivement. On notera $k = \sum_{i=1}^{p} m_i$ le nombre total de modalités.
On définit un scénario par une combinaison des modalités de chaque variable X_i.

Notons X la matrice (ou plan) d'expérience à q lignes et k colonnes associées aux données, sous forme disjonctive complète $X = (X_1|X_2|\ldots X_i\ldots|X_p)$.

La matrice d'expérience X est donc de la forme :

$$X_{(q,k)} = \left(\begin{array}{ccccc|ccccc|c|ccccc} 1 & 0 & \ldots & 0 & 0 & 1 & 0 & \ldots & 0 & 0 & \ldots & 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 & 1 & \ldots & 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ & & \vdots & & & & & \vdots & & & \vdots & & & \vdots & & \\ 0 & 0 & \ldots & 0 & 1 & 0 & 0 & \ldots & 0 & 1 & \ldots & 0 & 0 & \ldots & 0 & 1 \end{array}\right)$$

Cette matrice comporte autant de lignes que de produits, ou scénarios, et possède un nombre de colonnes égal à la somme des nombres de modalités des variables explicatives (encore appelées facteurs). Les colonnes de X indiquent pour chaque produit les modalités prises sous forme logique (variables indicatrices).

On soumet les produits à un échantillon de n individus (ou consommateurs), chaque individu doit alors classer par ordre de préférence les q scénarios, soit Y_q^j le vecteur (de taille q) des rangs donnés par le $j^{\text{ème}}$ consommateur $(j = 1,,n)$ pour les q scénarios proposés : le scénario préféré reçoit la valeur 1, le dernier reçoit la valeur q.

Nous supposerons dans une première étape que le consommateur donne un ordre complet sur les q scénarios plutôt que de les «noter» pour éviter les ex æquo et pour forcer les enquêtés à comparer réellement les produits

Dans une deuxième étape nous étudierons le cas où le consommateur classe uniquement les q_1 premiers scénarios définissant ainsi un ordre partiel.

2.1. Estimation des paramètres du modèle

2.1.1. Le modèle

On fait l'hypothèse que le classement du consommateur résulte d'une addition de points, « les utilités », associés aux différentes modalités présentes dans les produits (ce qui revient à considérer qu'il arbitre d'une manière additive entre les qualités constituant les produits).

Dans le cas général le modèle s'écrit $f(y) = Xb+e$ où f est une transformation monotone (préservant l'ordre). Il s'agit d'un modèle additif où seuls les effets principaux sont présents. La prise en compte d'interactions est en général impossible du fait du faible nombre de scénarios que l'on peut classer.

Il y a donc $\sum_{i=1}^{p} m_i - p$ coefficients (utilités) estimables.

Ceci signifie que seules les différences entre coefficients ont une signification. En pratique on imposera des contraintes aux coefficients; les plus courantes étant soit d'avoir des coefficients de somme nulle, soit d'annuler un des coefficients, en général le premier ou le dernier ce qui revient à éliminer des colonnes de $\mathbf{X}$ [2].

Remarquons que l'écriture précédente du modèle conduit à des utilités qui sont en fait des anti-utilités car les meilleurs classements sont associés à des valeurs faibles de Y et donc à des coefficients faibles aussi. On procédera si l'on veut de vraies utilités à une transformation des rangs en anti-rangs :

$$y_i \rightarrow q - y_i + 1 \quad \forall i = 1, \ldots, q \; .$$

Il est plus courant en pratique de demander de classer les produits selon un ordre complet plutôt que de les noter (sur 20 par exemple), ceci afin d'éviter les ex æquo et pour forcer les enquêtés à comparer réellement les produits.

Pour estimer les utilités on utilise soit la méthode des moindres carrés ordinaires, soit une méthode de régression monotone (*cf.* M. Tenenhaus et A. Agha) [7], [1].

Considérer le rang de classement comme une variable numérique est discutable : il s'agit uniquement d'une information ordinale et tout système de notation qui conduit au même ordre est acceptable. Comme il s'agit de reconstituer l'ordre de préférences on peut faire subir aux rangs de classement une transformation monotone croissante. Le principe de l'analyse de variance monotone (programme MONANOVA de J.B. Kruskal) consiste à chercher la transformation monotone $f(\mathbf{Y})$ qui conduit à la meilleure régression multiple avec $\mathbf{X}$ (ou $\mathbf{X}_0$, voir 2.1.2) comme matrice de variables explicatives [1] [5] [7].

Quelle méthode d'estimation choisir, moindres carrés ordinaires ou régression monotone? Il est clair que les transformations monotones sont mieux adaptées à la nature ordinale des données recueillies et au modèle additif postulé, mais peuvent conduire à des ajustements artificiels si le nombre de scénarios est insuffisant (risque d'« over fitting »). Il est donc recommandé d'examiner en premier lieu la solution des

moindres carrés ordinaires avant de passer à la régression monotone : les résultats des MCO sont en général robustes.

2.1.2. Estimation des utilités individuelles par les moindres carrés ordinaires

Le modèle précédent est appliqué à chaque enquêté et fournira donc des utilités propres à chacun. Il ne s'agit pas d'un modèle collectif de préférences mais bien d'un modèle tentant d'analyser les préférences de chaque consommateur.

Techniquement il faudra donc réaliser autant de régressions, ordinaires ou monotones, qu'il y a d'enquêtés.

On part ici de $\mathbf{Y} = \mathbf{Xb} + \mathbf{e}$ où $\mathbf{Y}$ est le vecteur des rangs de classements des scénarios et on estime $\mathbf{b}$ par $\hat{b}$ de telle sorte que : $\|Y - Xb\|^2$ soit minimal.

Pour les raisons déjà évoquées plus haut, ces estimations ne sont pas uniques, et on obtient une solution simple en éliminant des colonnes de $\mathbf{X}$. On prendra par exemple $\mathbf{X}_0$ la matrice obtenue en supprimant la première colonne de chaque sous-matrice X_i, et en rajoutant une colonne de 1 au début de la matrice, pour tenir compte de la constante du modèle. X_0 serait alors sous la forme $X_0 = (1|X_{01}|X_{02}|\ldots X_{0i}\ldots|X_{0p})$,

2.1.3. Hiérarchie des facteurs

Pour chacun des enquêtés, on peut ordonner les facteurs ou variables explicatives. Puisqu'il s'agit en fait d'un modèle d'analyse de la variance à effets additifs on pense immédiatement à utiliser les sommes de carrés expliqués par chaque facteur. Il faut cependant être prudent car d'une part les sommes de carrés ont tendance à augmenter avec le nombre de modalités de chaque facteur, et d'autre part ces sommes ne sont additives que dans le cas particulier des plans orthogonaux. Il faut alors utiliser des sommes de carrés de type II avec les tests F associés.

Une solution plus courante pour déterminer les importances relatives des différents facteurs consiste à utiliser les différences entre les utilités.

En effet, lorsque les $\prod_{i=1}^{p} m_i$ scénarios sont réalisables, la différence entre l'utilité maximale et l'utilité minimale pour tous les scénarios (*i.e.* l'étendue totale) est égale à la somme des étendues des utilités partielles de chaque facteur. L'importance d'un facteur est alors égale à $\frac{W_j}{\sum W_j}$, où W_j est l'étendue des utilités du facteur j.

Le passage à une hiérarchie sur l'ensemble des enquêtés suppose l'homogénéité de ces derniers. Il est alors conseillé avant de calculer des importances et des utilités moyennes de procéder à une typologie des interviewés sur la base de leurs utilités estimées.

3. Cas d'ordre partiel

Lorsque l'on demande à l'enquêté de classer un certain nombre de produits ou scénarios plus ou moins complexes, trois cas sont envisageables :

- l'enquêté classe tous les scénarios selon un ordre total : c'est le cas classique de l'Analyse Conjointe, les logiciels usuels de statistique permettent alors d'estimer aisément tous les paramètres.
- l'enquêté classe tous les scénarios mais peut donner le même rang à deux ou plusieurs produits différents : c'est le cas d'ex æquo.
- l'enquêté ne classe pas tous les scénarios : il classe uniquement les q_1 premiers scénarios et omet volontairement de classer les derniers : l'ordre ici est partiel.

Le troisième cas se rencontre lorsque le nombre de produits est important. Il pose un problème d'estimation des valeurs manquantes, car il n'est pas question de supprimer les individus ayant répondu de cette manière.

3.1. Procédé d'estimation

Nous proposons ici quelques procédés simples d'estimation des rangs manquants, dans le cas où le consommateur donne un ordre partiel sur les scénarios.

Nous supposons que le consommateur n'a pas de préférences pour les scénarios qu'il a omis de classer et donc que les rangs de tous ces produits seraient tous égaux.

Trois méthodes différentes d'estimation des valeurs manquantes sont alors envisageables :

a) Les produits non classés reçoivent tous le rang du dernier produit classé augmenté de 1 soit $q_1 + 1$, ce qui revient à estimer les rangs manquants par le rang qui suit celui du dernier classé.

b) Les produits non classés reçoivent tous la moyenne (rendue entière par excès) des rangs manquants.

c) Les produits non classés reçoivent tous le rang maximum, soit q.

Ces trois façons d'estimer les rangs manquants nous permettent de ramener le cas de valeurs manquantes à celui d'ex æquo.

Nous essayons dans la suite, et sur un exemple de déterminer un seuil à partir duquel le classement partiel de scénarios n'est plus exploitable.

3.2. Exemple

3.2.1. Les données

Nous partons ici d'un exemple qui propose à 263 consommateurs de classer des scénarios d'abonnements offerts en téléphone mobile (les données réelles sont légèrement différentes pour des raisons de confidentialité).

Les variables sont :

X_1	prix de l'équipement (Equip)	0 €, 100 €	: (equi1, equi2)
X_2	Frais de mise en service (FMS)	0 €, 30 €	: (fms1, fms2)
X_3	Plage horaire (Plage)	p1, p2	: (plage1, plage2)
X_4	Abonnement mensuel (abonnement)	0 €, 4.5 €,9 €	: (abt1, abt2, abt3)
X_5	Durée de l'abonnement (durée)	6m, 24m	: (dure1, dure2)
X_6	Prix/minute heure pleine (h-pleine)	0.5 €, 0.7 €, 0.9 €	: (hp1, hp2, hp3)
X_7	Prix/minute heure creuse (h-creuse)	0.08 €, 0.1 €, 0.15 €	: (hc1, hc2, hc3)

Les scénarios sont donc obtenus par combinaison de sept variables, dont quatre à deux modalités et trois à trois modalités, le nombre de paramètres linéairement indépendants à estimer étant égal à $\sum_{i=1}^{p} m_i - p$ est ici de 10.

Le plan factoriel complet associé, comprend 432 scénarios. Un plan orthogonal au sens strict doit avoir un nombre de scénarios égal au PPCM $(2 \times 3, 3 \times 3, 2 \times 2)$ soit donc 36 scénarios, ce qui est trop élevé en pratique.

Dans l'étude réelle on a utilisé un plan à 12 scénarios, obtenu par un programme de recherche de plans D-optimaux.

Ainsi dans notre exemple le plan P D-optimal à 12 scénarios les matrices X et X_0 associées à ce plan sont :

$$P = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \end{matrix} \\ \begin{pmatrix} 2 & 2 & 2 & 2 & 3 & 1 & 1 \\ 2 & 2 & 1 & 1 & 3 & 3 & 2 \\ 2 & 2 & 1 & 1 & 1 & 2 & 3 \\ 2 & 1 & 2 & 2 & 1 & 2 & 2 \\ 2 & 1 & 2 & 1 & 2 & 1 & 3 \\ 2 & 1 & 1 & 2 & 2 & 3 & 1 \\ 1 & 2 & 2 & 2 & 1 & 3 & 3 \\ 1 & 2 & 2 & 1 & 2 & 2 & 1 \\ 1 & 2 & 1 & 2 & 2 & 1 & 2 \\ 1 & 1 & 2 & 1 & 3 & 3 & 2 \\ 1 & 1 & 1 & 2 & 3 & 2 & 3 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array} \quad X = \begin{pmatrix} 01 & 01 & 01 & 01 & 001 & 100 & 100 \\ 01 & 01 & 10 & 10 & 001 & 001 & 010 \\ 01 & 01 & 10 & 10 & 100 & 010 & 001 \\ 01 & 10 & 01 & 01 & 100 & 010 & 010 \\ 01 & 10 & 01 & 10 & 010 & 100 & 001 \\ 01 & 10 & 10 & 01 & 010 & 001 & 100 \\ 10 & 01 & 01 & 01 & 100 & 001 & 001 \\ 10 & 01 & 01 & 10 & 010 & 010 & 100 \\ 10 & 01 & 10 & 01 & 010 & 100 & 010 \\ 10 & 10 & 01 & 10 & 001 & 001 & 010 \\ 10 & 10 & 10 & 01 & 001 & 010 & 001 \\ 10 & 10 & 10 & 10 & 100 & 100 & 100 \end{pmatrix}$$

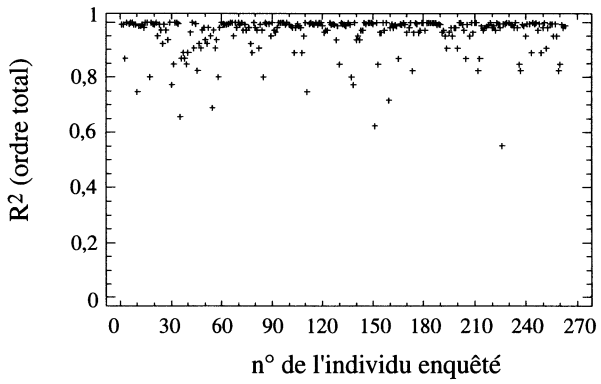
$$X_0 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 01 & 00 & 00 \\ 1 & 1 & 1 & 0 & 0 & 01 & 01 & 10 \\ 1 & 1 & 1 & 0 & 0 & 00 & 10 & 01 \\ 1 & 1 & 0 & 1 & 1 & 00 & 10 & 10 \\ 1 & 1 & 0 & 1 & 0 & 10 & 00 & 01 \\ 1 & 1 & 0 & 0 & 1 & 10 & 01 & 00 \\ 1 & 0 & 1 & 1 & 1 & 00 & 01 & 01 \\ 1 & 0 & 1 & 1 & 0 & 10 & 10 & 00 \\ 1 & 0 & 1 & 0 & 1 & 10 & 00 & 10 \\ 1 & 0 & 0 & 1 & 0 & 01 & 01 & 10 \\ 1 & 0 & 0 & 0 & 1 & 01 & 10 & 01 \\ 1 & 0 & 0 & 0 & 0 & 00 & 00 & 00 \end{pmatrix}$$

Les produits obtenus ont été présentés à 263 consommateurs, qui ont tous répondu par un classement total.

3.2.2. Les résultats et leur interprétation

La reconstitution de l'ordre des préférences se mesure avec le carré du coefficient de corrélation multiple R^2. Lorsque ce coefficient est faible il traduit soit une inadéquation du modèle d'utilités additives, soit une incohérence des réponses de l'enquêté. C'est en général cette dernière hypothèse qui est retenue puisque le modèle est considéré comme réaliste. On éliminera alors les enquêtés pour lesquels le R^2 est inférieur à un seuil fixé par l'utilisateur.

Pour le cas de notre exemple le R^2 est très proche de 1 pour presque tous les individus dans le cas d'ordre total, comme le montre le graphique 1, ce qui montre que le modèle explique bien le choix des consommateurs.



GRAPHIQUE 1
R^2 dans le cas d'ordre total

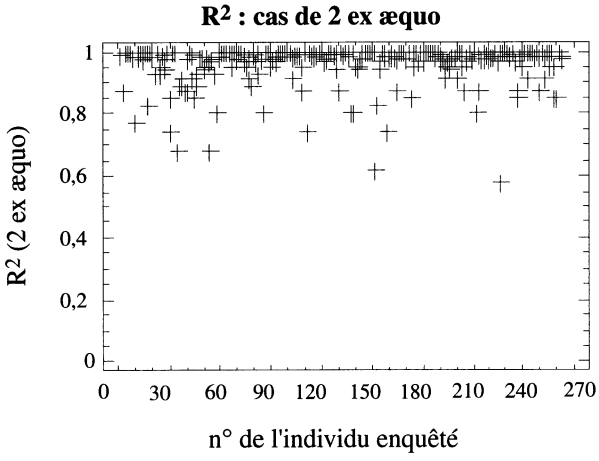
3.2.3. Les simulations

Nous allons tout d'abord établir l'influence du nombre de scénarios non classés sur la qualité globale de l'ajustement.

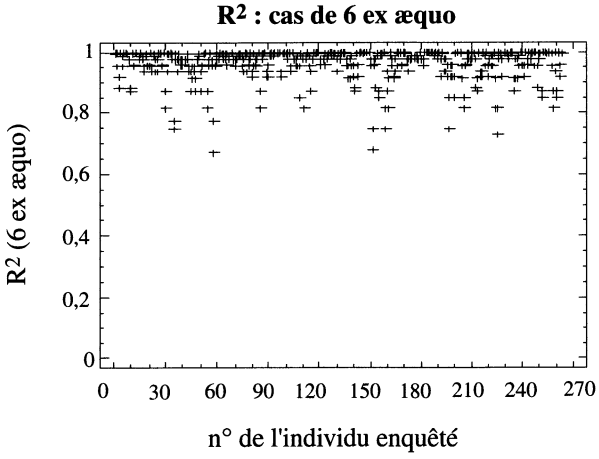
Pour cela nous avons simulé des classements partiels graduels, en remplaçant successivement, sur les données de l'ordre total, les rangs des derniers scénarios classés par des valeurs manquantes. Celles-ci sont transformées par la suite en ex æquo, ainsi nous avons des ordres partiels avec i valeurs manquantes ce qui se traduit par $(i-1)$ ex æquo, $i = 1, \ldots, 12$. Le cas $i = 1$ nous redonne le cas classique d'ordre total.

3.2.3.1. Cas où les valeurs manquantes sont remplacées par le rang du dernier classé +1

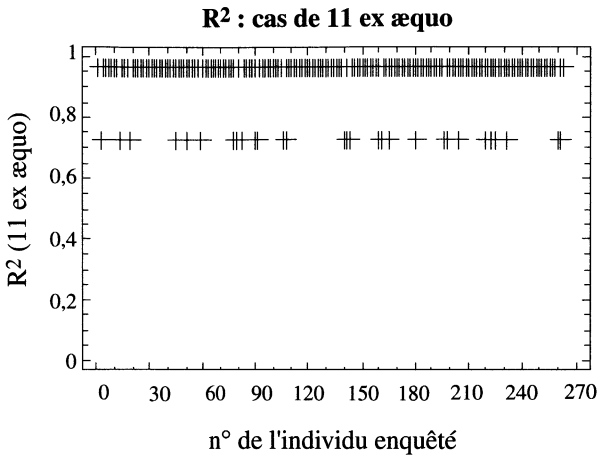
Le nombre de valeurs prises par le R^2 diminue lorsque le nombre d'ex æquo augmente comme le montrent les graphiques 2 , 3 et 4.



GRAPHIQUE 2



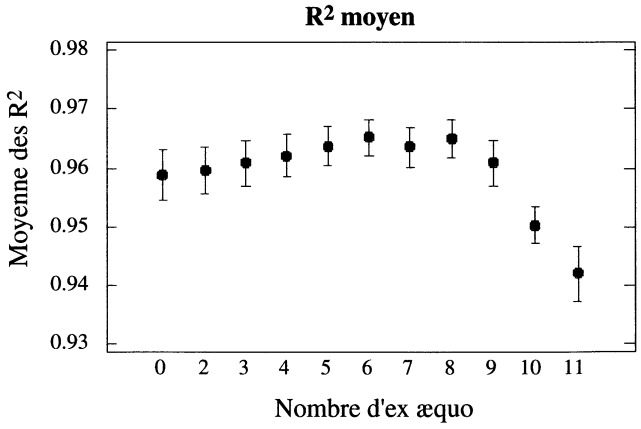
GRAPHIQUE 3



GRAPHIQUE 4

En poussant jusqu'à l'absurde la simulation avec 11 ex æquo, on ne trouve plus que deux valeurs du R^2, cas sans utilité dans la pratique.

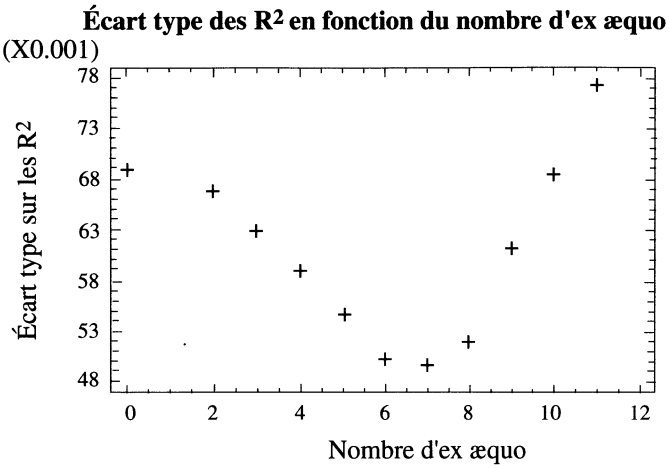
Le graphique 5 donne la moyenne des R^2 sur les 263 régressions (avec une barre d'erreur de $\pm$ 1 écart type)



GRAPHIQUE 5
Moyenne des R^2
(Les valeurs manquantes sont remplacées par le rang du dernier classé +1)

Le R^2 moyen croît légèrement lorsque le nombre d'ex æquo est entre 2 et 6 et décroît entre 7 et 11 ex æquo (graphique 5).

Le graphique 6 donne l'écart type sur les R^2 associes aux 263 régressions



GRAPHIQUE 6
Écart type des R^2
(Les valeurs manquantes sont remplacées par le rang du dernier classé +1)

L'écart type est donc minimal pour 7 ex æquo.

3.2.3.2. Cas où les valeurs manquantes sont remplacées par la moyenne des rangs non classés

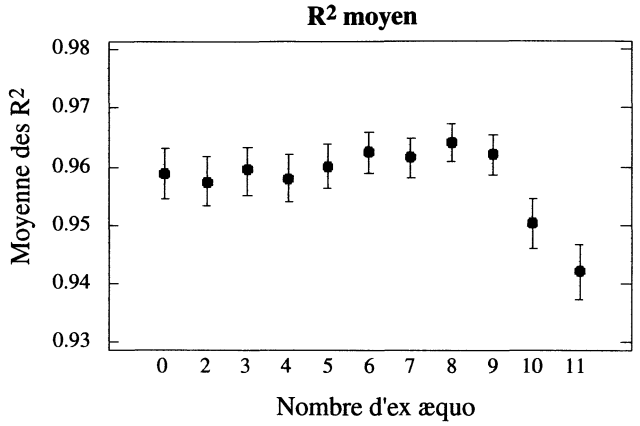
De même que pour le cas précédent, le nombre de valeurs prises par le R^2 diminue lorsque le nombre d'ex æquo augmente.

Le graphique 7 donne la moyenne des R^2 sur les 263 régressions (avec une barre d'erreur de $\pm$ 1 écart type). Le R^2 moyen croît légèrement lorsque le nombre d'ex æquo est entre 2 et 7 et décroît entre 8 et 11 ex æquo (graphique 7).

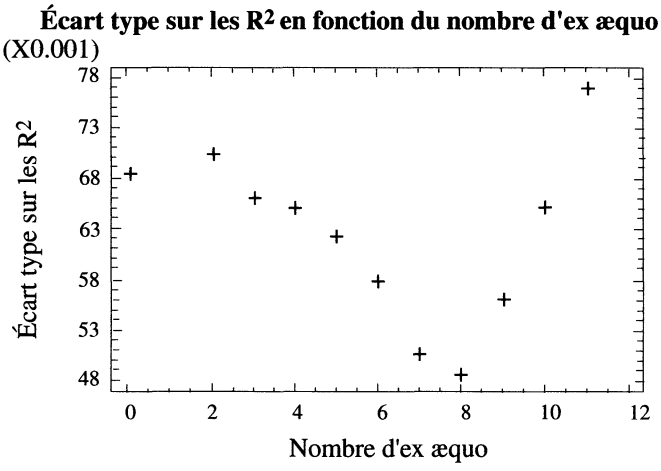
Le graphique 8 donne l'écart type des R^2 associés aux 263 régressions.

L'écart type décroît entre 2 et 8 ex æquo et remonte à partir de neuf ex æquo comme le montre le graphique 8. Le R^2 moyen reste supérieur à 0.94 avec un écart type très faible.

Nous retrouvons, le même comportement pour les R^2, avec un léger décalage de la moyenne et de l'écart type vers un nombre d'ex æquo autour de sept. Ce décalage est dû simplement au fait que la moyenne par excès, des rangs non classés pour un nombre d'ex æquo égal à 6 (qui est égale à 10) est la même pour un nombre d'ex æquo égal à 7.



GRAPHIQUE 7
Moyenne des R^2
(Les valeurs manquantes sont remplacées par la moyenne des rangs non classés)

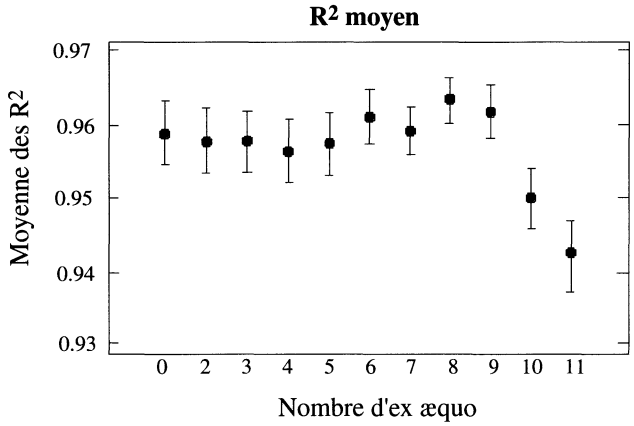


GRAPHIQUE 8
Écart type des R^2
(Les valeurs manquantes sont remplacées par la moyenne des rangs non classés)

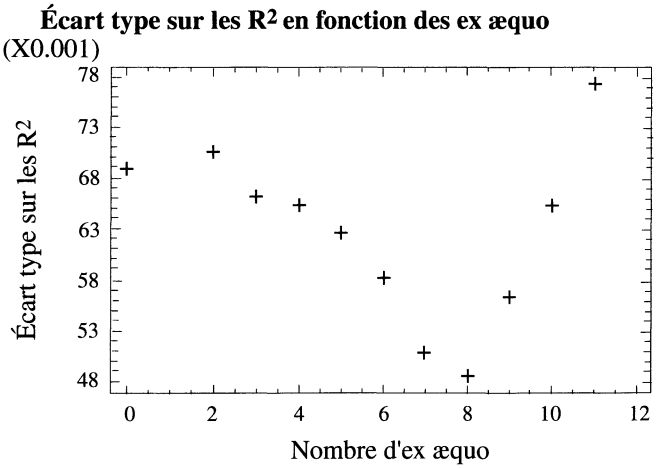
3.2.3.3. Cas où les valeurs manquantes sont remplacées par le rang maximum

De même que pour les cas précédents, le nombre de valeurs prises par le R^2 diminue lorsque le nombre d'ex æquo augmente.

Le graphique 9 donne la moyenne des R^2 sur les 263 régressions (avec une barre d'erreur de $\pm$ 1 écart type)



GRAPHIQUE 9
Moyenne des R^2
(Les valeurs manquantes sont remplacées par le rang maximum)



GRAPHIQUE 10
Écart type des R^2
(Les valeurs manquantes sont remplacées par le rang maximum)

L'écart type décroît entre 2 et 8 ex æquo et remonte à partir de neuf ex æquo comme le montre le graphique10.

Le R^2 moyen reste supérieur à 0.94 avec un écart type très faible, et qui a la même allure que les cas précédents.

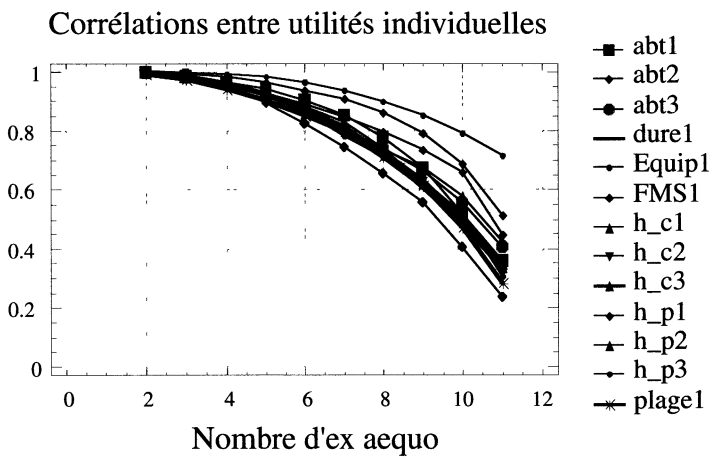
Le R^2 ne mesure pas en réalité l'effet du nombre d'ex æquo sur la qualité des résultats. Il est plus important de voir l'effet sur l'estimation des paramètres.

3.2.4. Étude des utilités individuelles

L'étude empirique montre que les utilités individuelles restent stables lorsque le nombre d'ex æquo n'est pas trop important (de l'ordre de 6). Cette stabilité se perd lorsque le nombre d'ex æquo devient élevé. Ceci traduit une très forte corrélation entre les utilités calculées sans ex æquo et celles avec un faible nombre d'ex æquo (donc un nombre de valeurs manquantes faible).

Les utilités pour les facteurs à deux niveaux étant égales en valeur absolue mais de signes opposés, nous représentons uniquement les corrélations de celles associées au premier, et celles associées aux trois niveaux des autres facteurs sur les graphiques 11 et 12.

3.2.4.1. Cas où les valeurs manquantes sont remplacées par le rang du dernier classé +1

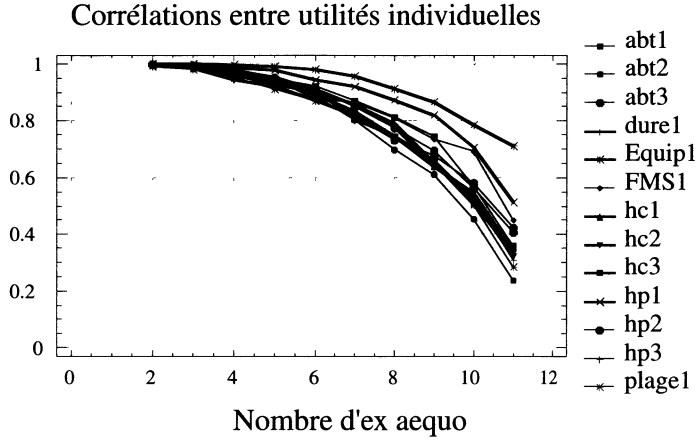


GRAPHIQUE 11
Corrélation entre utilités individuelles dans le cas d'ordre total et différents ordres partiels (les valeurs manquantes sont remplacées par le rang du dernier classé +1)

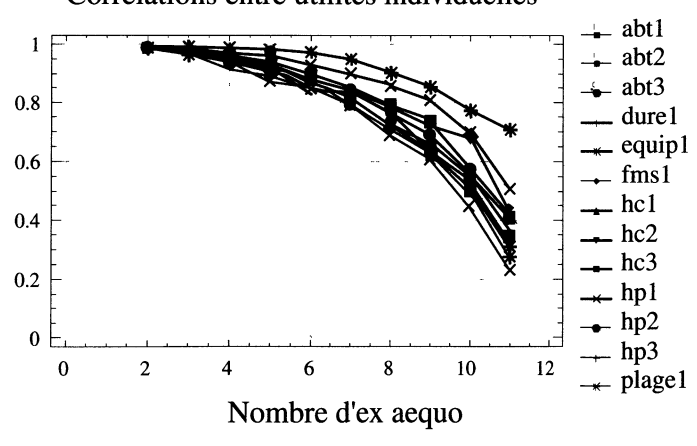
Les corrélations entre utilités individuelles sont très proches de 1 (de l'ordre de 0.99) pour tous les niveaux des facteurs, lorsqu'on a seulement deux ou trois ex æquo et décroissent à des vitesses différentes selon les attributs, et restent supérieures à 0.8 pour tous les attributs lorsque le nombre d'ex æquo ne dépasse pas 6.

3.2.4.2. Cas où les valeurs manquantes sont remplacées par la moyenne des rangs non classés

Les corrélations entre utilités individuelles restent très proches de 1 (de l'ordre de 0.99) pour tous les niveaux des facteurs, lorsqu'on a seulement deux ou trois ex æquo et décroît à des vitesses différentes selon les attributs, mais elle reste supérieure à 0.82 pour tous les attributs lorsque le nombre d'ex æquo ne dépasse pas 6 ou 7.



GRAPHIQUE 12
Corrélation entre utilités individuelles dans le cas d'ordre total et différents ordres partiels (les valeurs manquantes sont remplacées par la moyenne des rangs non classés)



GRAPHIQUE 13
Corrélation entre utilités individuelles dans le cas d'ordre total et différents ordres partiels (les valeurs manquantes sont remplacées par le rang maximum)

3.2.4.3. Cas où les valeurs manquantes sont remplacées par le rang maximum

Les corrélations entre utilités individuelles restent ici aussi, très proches de 1 (de l'ordre de 0.99) pour tous les niveaux des facteurs, lorsqu'on a seulement deux ou trois ex æquo et décroissent à des vitesses différentes selon les attributs, mais elles restent supérieure à 0.8 pour tous les attributs lorsque le nombre d'ex æquo ne dépasse pas 6 ou 7.

3.2.5. Hiérarchie des facteurs

3.2.5.1. Cas où les valeurs manquantes sont remplacées par le rang du dernier classé +1

L'étude des importances des utilités, dans le cas de notre exemple, montre une certaine stabilité de ces dernières lorsque le nombre d'ex æquo n'est pas très important (ici, si le classement se fait jusqu'au sixième scénario). Cette stabilité se perd lorsque le nombre d'ex æquo devient élevé, ce qui dénote une forte corrélation entre ces importances lorsque le nombre d'ex æquo est faible donc un nombre de valeurs manquantes faible (de l'ordre de $q/2$, q étant le nombre de scénarios) avec le cas d'ordre total, comme le montrent le tableau 1 et le graphique 14.

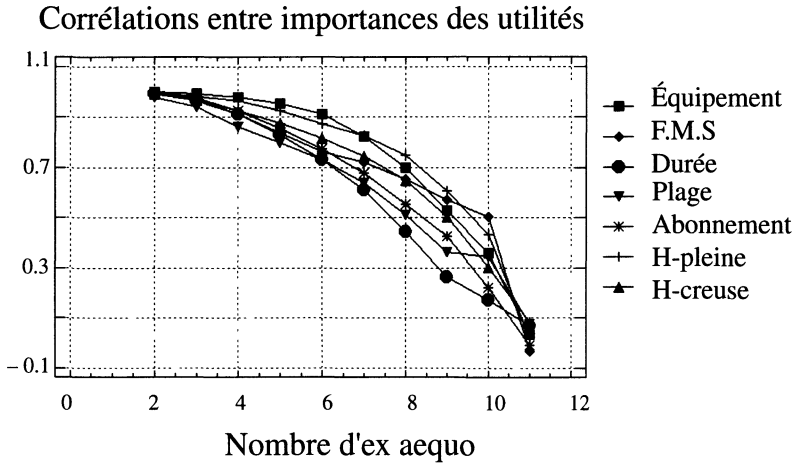
TABLEAU 1
Corrélation entre importances des facteurs dans le cas d'ordre total et différents ordres partiels

Nombre d'ex æquo	équipement	FMS	durée	plage	abonnement	h-pleine	h-creuse
2	0.99816	0.99346	0.99184	0.97789	0.99424	0.99521	0.99086
3	0.99211	0.97092	0.96390	0.94017	0.97338	0.98316	0.96629
4	0.97838	0.91102	0.90942	0.86295	0.92560	0.96391	0.92637
5	0.95408	0.83698	0.83045	0.79723	0.85450	0.92451	0.87554
6	0.91276	0.76302	0.73114	0.73028	0.77237	0.87226	0.81394
7	0.82371	0.71935	0.61059	0.63803	0.67610	0.82709	0.74301
8	0.69778	0.65336	0.44328	0.50988	0.55305	0.74954	0.64980
9	0.52875	0.57248	0.26386	0.36207	0.42579	0.60485	0.50500
10	0.35712	0.50498	0.16924	0.34297	0.22038	0.43167	0.30110
11	0.03341	0.03470	0.06879	0.07652	-0.01084	0.02027	0.08234

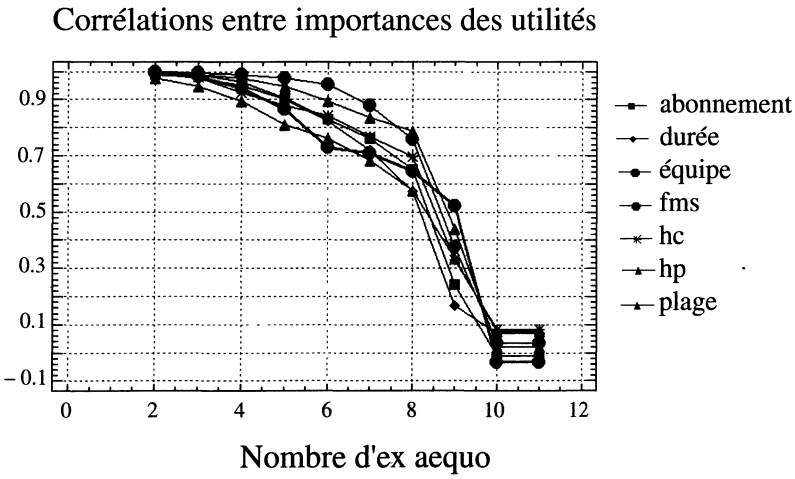
3.2.5.2. Cas où les valeurs manquantes sont remplacées par la moyenne des rangs non classés

L'étude montre une certaine stabilité des importances des utilités lorsque le nombre d'ex æquo n'est pas très important (ici, si le classement se fait jusqu'au

septième scénario, pour les mêmes raisons qu'au § 3.2.3.2). Cette stabilité se perd lorsque le nombre d'ex æquo devient élevé, comme le montre le graphique 15.



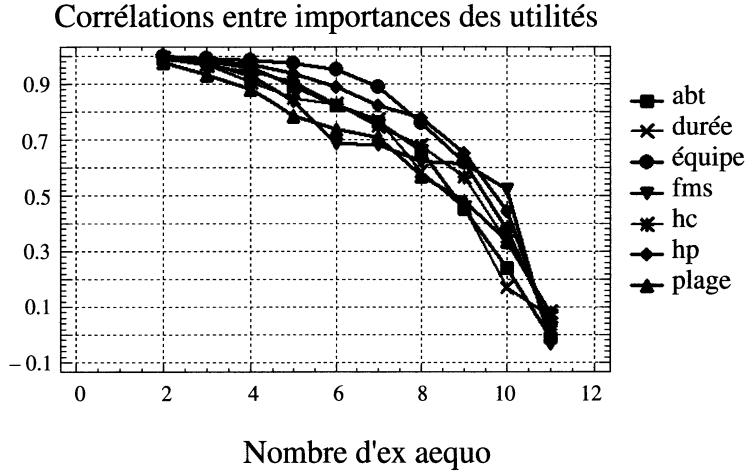
GRAPHIQUE 14
Corrélation entre importances des facteurs dans le cas d'ordre total et différents ordres partiels (les valeurs manquantes sont remplacées par le rang du dernier classé +1)



GRAPHIQUE 15
Corrélation entre utilités individuelles dans le cas d'ordre total et différents ordres partiels (les valeurs manquantes sont remplacées par la moyenne des rangs non classés)

3.2.5.3. Cas où les valeurs manquantes sont remplacées par le rang maximum

L'étude montre, dans ce cas aussi, une certaine stabilité des importances des utilités lorsque le nombre d'ex æquo n'est pas très important (ici, si le classement se fait jusqu'au septième scénario). Cette stabilité se perd lorsque le nombre d'ex æquo devient élevé, comme le montre le graphique 16.



GRAPHIQUE 16
Corrélation entre importances des facteurs dans le cas d'ordre total et différents ordres partiels (les valeurs manquantes sont remplacées par le rang maximum)

Le seuil de valeurs manquantes pour lequel les données restent utilisables semble être de l'ordre de $q/2$, q étant le nombre de scénarios, hypothèse que nous ne pouvons confirmer faute de données.

4. Modèle avec perturbations aléatoires

La partie précédente représente une expérience sur des données réelles, mais d'une seule expérience on ne peut tirer de généralités.

Nous avons donc décidé de simuler des classements de manière plus systématique. Pour partir d'un modèle réaliste, nous avons fait le choix d'un système d'utilité β et simulé des classements compatibles et cohérents en ajoutant des bruits ε de variances données à $X\beta$.

4.1 Étude du R^2

Nous partons d'un vecteur β qui est la moyenne des 263 coefficients β du modèle associé aux données de l'exemple du § 3.2. On perturbe ensuite $X\beta$, où X est la matrice d'expérience associée, par une variable normale $\varepsilon \rightarrow N(0, \sigma^2 \mathrm{Id_q})$, où

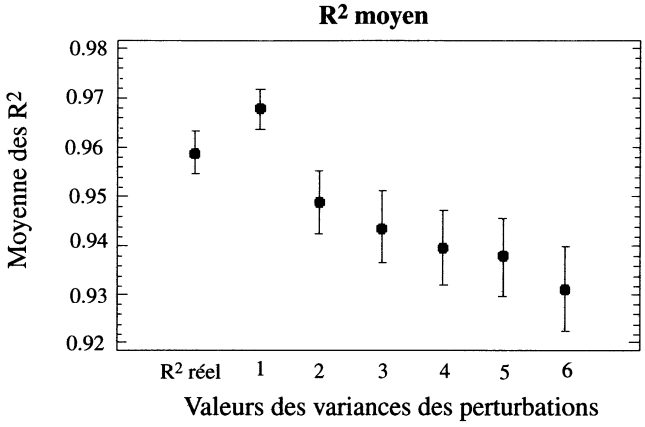
$q = 12$ est le nombre de scénarios, et on remplace $y = X\beta + e$ (e étant une réalisation de ε) par les rangs de classement. On obtient ainsi en tirant 100 fois des valeurs de ε, 100 classements différents. On fait varier σ^2 de 1 à 6.

La simulation des rangs de classement à partir de coefficients d'un modèle sur des données réelles, nous garantit la cohérence des données simulées avec un coefficient de corrélation multiple proche de 1.

En effet nous obtenons un R^2 moyen très proche de 1 et du cas réel (pour toutes les valeurs de σ^2), avec un écart type qui augmente légèrement avec σ^2 comme le montre le tableau 2 et le graphique 17.

TABLEAU 2
Statistiques sur les R^2 pour $\sigma^2 = 1, \ldots, 6$

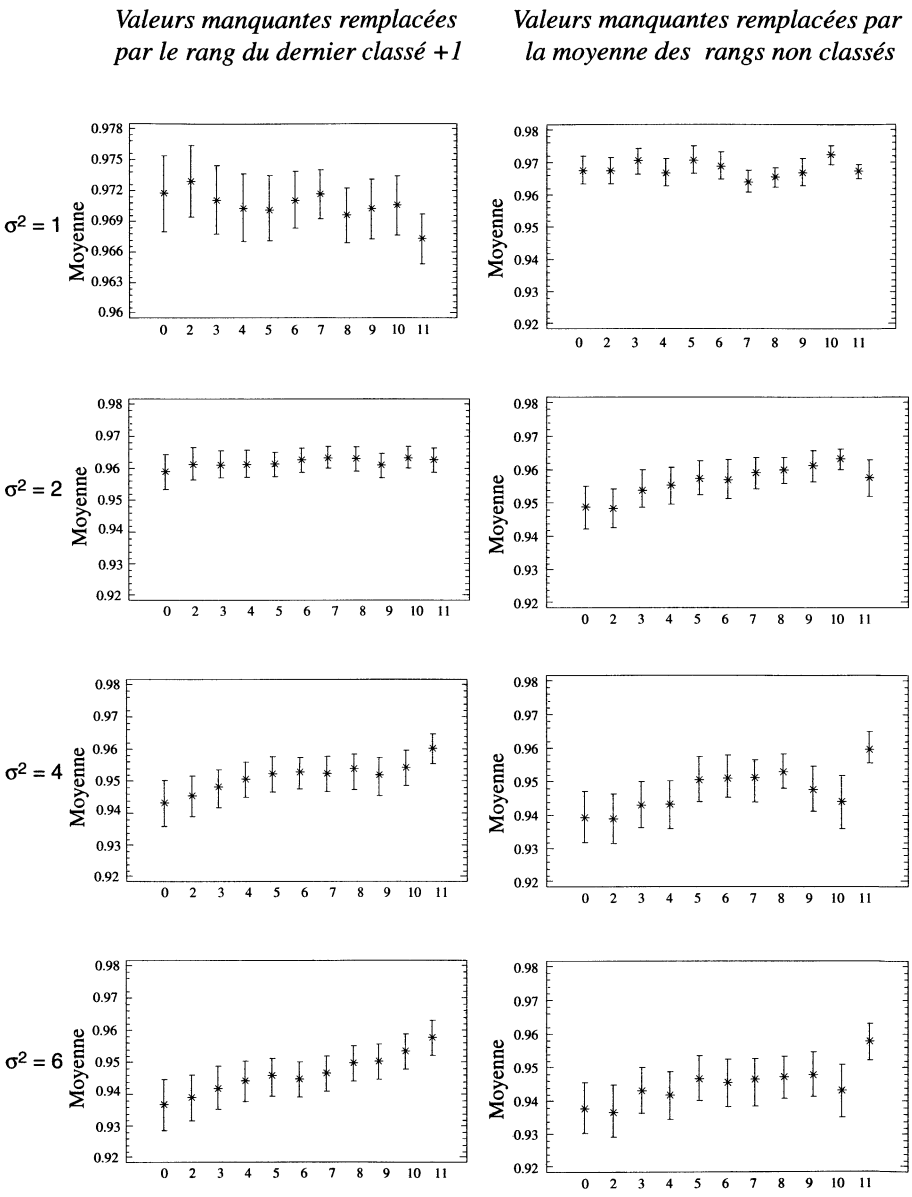
σ^2	effectif	R^2 moyen	Écart type des R^2	R^2 minimal	R^2 maximal
Cas réel	263	0.958866	0.00688944	0.552448	1.00000
1	100	0.967669	0.00427224	0.801088	1.00000
2	100	0.948563	0.00655709	0.719503	1.00000
3	100	0.943520	0.00701494	0.588967	1.00000
4	100	0.939145	0.00753101	0.588967	1.00000
5	100	0.937467	0.00775796	0.588967	1.00000
6	100	0.931064	0.00878767	0.552448	1.00000



GRAPHIQUE 17
Moyenne des R^2 pour $\sigma^2 = 1, \ldots, 6$. La barre d'erreur est de $\pm$ 1 écart type

Nous introduisons ensuite des valeurs manquantes, corrigées par des ex æquo selon les procédés décrits au § 3.1 pour les différentes valeurs de σ^2.

Le graphique 18 nous donne la moyenne des R^2 sur les 100 régressions (avec une barre d'erreur de $\pm$ 1 écart type), pour les différentes valeurs de σ^2.



GRAPHIQUE 18
Moyennes des R^2 sur les 100 régressions $\sigma^2 = 1, 2, 4, 6$

4.2 Étude des utilités individuelles

Les utilités pour les facteurs à deux niveaux étant égales en valeur absolue mais de signes opposés, nous représentons uniquement les corrélations de celles associées au premier, et celles associées aux trois niveaux des autres facteurs .

Le graphique 19 nous donne les corrélations entre utilités individuelles pour différentes valeurs de σ^2 (à gauche les valeurs manquantes sont remplacées par le rang du dernier classé +1, à droite les valeurs manquantes sont remplacées par la moyenne des rangs non classés).

Les utilités individuelles restent stables tant que le nombre d'ex æquo n'est pas trop important (de l'ordre de 6). Cette stabilité se perd lorsque le nombre d'ex æquo devient élevé. Ceci traduit une très forte corrélation entre les utilités calculées sans ex æquo et celles avec un faible nombre d'ex æquo. Cette corrélation augmente légèrement lorsque σ^2 augmente (0.8 pour $\sigma^2 = 1$ à 0.89 pour $\sigma^2 = 6$) pour les deux méthodes de calcul des ex æquo.

Nous remarquons que pour un nombre d'ex æquo égal à 6 la corrélation entre utilités individuelles qui était de l'ordre de 0.81 pour le cas réel, devient de l'ordre de 0.80 pour $\sigma^2 = 1$, 0.82 pour $\sigma^2 = 2$, 0.85 pour $\sigma^2 = 3$, 0.86 pour $\sigma^2 = 4$, 0.87 pour $\sigma^2 = 5$ et de l'ordre de 0.89 pour $\sigma^2 = 6$, donc croit légèrement lorsque σ^2 augmente, avec une légère différence entre les deux méthodes (de l'ordre de 0.01).

4.3. Étude des importances des utilités

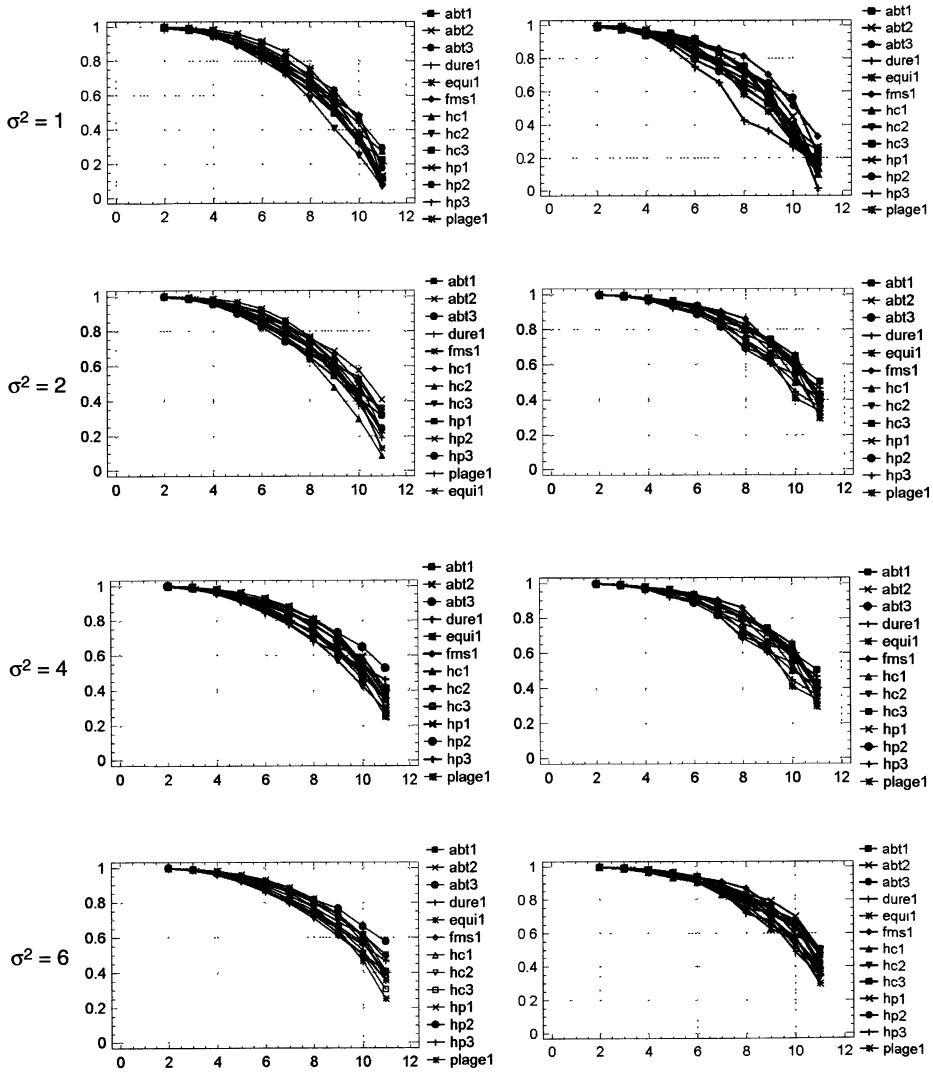
Le graphique 20 nous donne les corrélations entre importances des utilités des facteurs pour différentes valeurs de σ^2 (à gauche les valeurs manquantes sont remplacées par le rang du dernier classé +1, à droite les valeurs manquantes sont remplacées par la moyenne des rangs non classés).

Nous remarquons que dans le cas où les valeurs manquantes sont remplacées par le rang du dernier classé +1, pour un nombre d'ex æquo égal à 6, la corrélation entre importances des utilités qui se situait entre 0.7 et 0.9 (selon le facteur) pour le cas réel, se situe entre 0.4 et 0.9 pour $\sigma^2 = 1$, entre 0.5 et 0.9 pour $\sigma^2 = 2$, entre 0.6 et 0.9 pour $\sigma^2 = 3$, entre 0.61 et 0.91 pour $\sigma^2 = 4$, entre 0.65 et 0.9 pour $\sigma^2 = 5$, et entre 0.65 et 0.92 pour $\sigma^2 = 6$. Dans le cas où les valeurs manquantes sont remplacées par la moyenne des rangs manquants, pour un nombre d'ex æquo égal à 6 cette corrélation se situe entre 0.6 et 0.9 pour $\sigma^2 = 1$, entre 0.7 et 0.91 pour $\sigma^2 = 2$, entre 0.75 et 0.9 pour $\sigma^2 = 3$, entre 0.75 et 0.91 pour $\sigma^2 = 4$, entre 0.75 et 0.9 pour $\sigma^2 = 5$, et entre 0.8 et 0.9 pour $\sigma^2 = 6$.

L'étude des importances des utilités, montre une certaine stabilité de ces dernières lorsque le nombre d'ex æquo n'est pas très important (ici, si le classement se fait jusqu'au sixième scénario). Cette stabilité se perd lorsque le nombre d'ex æquo devient élevé, ce qui dénote une forte corrélation entre ces importances lorsque le nombre d'ex æquo est faible donc un nombre de valeurs manquantes faible (de l'ordre de $q/2$, q étant le nombre de scénarios) avec le cas d'ordre total. La corrélation est d'autant plus importante que σ^2 augmente. Avec une légère différence entre les deux méthodes de calcul pour certains facteurs.

Valeurs manquantes remplacées par le rang du dernier classé +1

Valeurs manquantes remplacées par la moyenne des rangs non classés

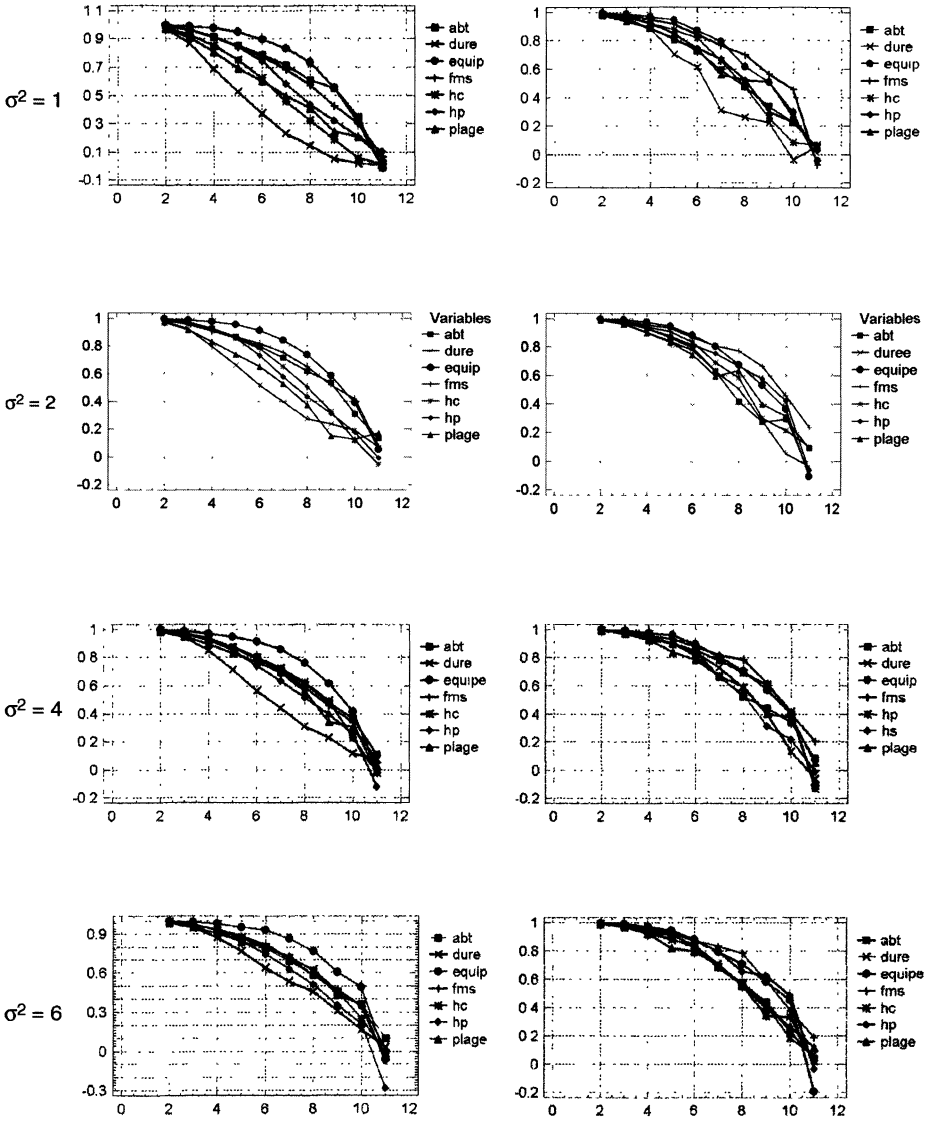


GRAPHIQUE 19

Corrélations entre utilités individuelles pour $\sigma^2 = 1, 2, 4, 6$

Valeurs manquantes remplacées par le rang du dernier classé +1

Valeurs manquantes remplacées par la moyenne des rangs non classés



GRAPHIQUE 20

Corrélations entre importances des utilités des facteurs pour $\sigma^2 = 1, 2, 4, 6$

5. Conclusion

Le cas de classement partiel des données en analyse conjointe est assez fréquent surtout lorsque le nombre de produits proposés dépasse la dizaine. Ces données ne sont pas complètement inutilisables si l'enquêté classe la majorité des produits. Nous avons montré sur un exemple en comparant le classement total et divers classements partiels, selon différentes méthodes d'estimation, que les utilités et les importances dans le cas partiel sont assez proches de celles de l'ordre total, pourvu que le nombre de valeurs manquantes ne soit pas trop important : nos expériences ont montré une bonne stabilité des résultats jusqu'à la moitié des scénarios classés.

Bibliographie

[1] AGHA A.K. (1991), *Régression et analyse canonique sous contraintes linéaires*, Thèse Université Paris-Dauphine.

[2] DROESBEKE J.J., FINE J., SAPORTA G. (éditeurs), *Plans d'Expériences, Applications à l'entreprise*, Éditions Technip.

[3] EVRARD Y., PRAS B., ROUX E. (2000), «*Market*», Dunod.

[4] GREEN P.E., SRINIVASAN V. (1990), *Conjoint analysis in marketing : new developments with implications for research and practice*, Journal of Marketing, 3-19.

[5] KRUSKAL J.B. (1965), *Analysis of factorial experiments by estimating monotone transformations of the data*, JRSS, B, 27, 251-263.

[6] SAS Institute Inc., SAS Technical Report R-109 (1993), *Conjoint Analysis Examples*, Cary, NC : SAS Institute Inc.

[7] TENENHAUS M. (1988), *Canonical analysis of two polyedral convex cones*, Psychometrika, 53, 4, 503-524.

Annexe : Rappel sur les plans d'expérience

Soient p variables qualitatives $X_1, X_2, \ldots, X_p$ (dites encore facteurs ou attributs) à $m_1, m_2, \ldots, m_p$ modalités (ou niveaux) respectivement. On notera $k = \sum_{i=1}^{p} m_i$ le nombre total de modalités.

On définit un produit (ou scénario) par une combinaison des modalités de chaque variable X_i (ou d'attributs). Un scénario est donc caractérisé par un seul niveau de chacun des attributs, lesquels sont décrits par des indicatrices. On construit q scénarios différents, à l'aide de techniques de plans d'expériences, Nous en rappelons ici les plus importants.

Plan complet

Soit X, la matrice formée par q scénarios résultant de combinaisons des attributs des p facteurs, le nombre potentiel de scénarios associés à X est égal au nombre de combinaisons possibles entre les attributs, soit $\prod_{i=1}^{p} m_i$, les présenter tous constitue un plan factoriel complet.

Le nombre de produits augmentant de façon exponentielle en fonction du nombre des facteurs rend cette démarche irréaliste. On fait alors appel aux plans fractionnaires optimaux [2] qu'on utilise plutôt en pratique, d'autant plus que le consommateur n'arrive pas à mémoriser tous les scénarios proposés dés que leur nombre dépasse la dizaine.

Comment choisir les scénarios et leur nombre de manière optimale?

La théorie des plans d'expériences permet de répondre à cette question : on utilisera lorsque cela est possible des fractions orthogonales du plan complet [2] qui permettent de bien séparer les influences des facteurs. Les constructions de plans orthogonaux peuvent se faire à la main dans des cas simples ou avec des tables toutes faites; une bonne pratique et de l'habileté sont toutefois nécessaires. Sinon les plans D-optimaux constituent l'outil presque absolu pour résoudre le problème.

Plans orthogonaux fractionnaires et D-optimalité

Trouver les meilleurs plans d'expériences à nombre de scénarios fixés suppose que l'on a défini un critère d'optimalité et que l'on dispose d'un algorithme efficace de recherche [2]. Le critère usuel est la D-optimalité.

La définition de critères d'optimalité renvoie au modèle linéaire. Il s'agit ici d'estimer le mieux possible les coefficients du modèle :

Soit $\mathbf{Y}$ le vecteur des rangs de classement (de taille q pour un individu donné), ou une transformation monotone de ce dernier, et soit $\mathbf{X_0}$ la matrice des variables explicatives qui sont ici les indicatrices des niveaux des facteurs (avec une indicatrice de moins que de niveaux pour assurer l'inversibilité de $\mathbf{X_0'X_0}$) les utilités sont les solutions $\hat{b}$ des équations normales :

$$\hat{\mathbf{b}} = (\mathbf{X_0'X_0})^{-1}\mathbf{X_0'Y}$$

Les utilités sont estimées sans biais mais avec une variabilité qui est définie par la matrice de variance des résidus et qui vaut :

$$\mathbf{V}(\hat{\mathbf{b}}) = \sigma^2(\mathbf{X_0'X_0})^{-1}$$

Les estimations seront donc d'autant meilleures que la matrice $(\mathbf{X_0'X_0})^{-1}$ sera «petite». Plusieurs définitions sont alors possibles, celle que l'on retient le plus souvent à la suite de Fédorov [2], consiste à avoir un déterminant minimal pour $(\mathbf{X_0'X_0})^{-1}$ et donc maximal pour $(\mathbf{X_0'X_0})$; c'est la ***D-optimalité*** : la recherche

des plans D-optimaux impose de recourir à des logiciels spécialisés, le plus souvent conçus pour le contrôle industriel de la qualité.

Les algorithmes consistent à rechercher parmi l'ensemble des solutions candidates, en général en partant du plan factoriel complet, le ou les sous-ensembles de scénarios de taille fixée maximisant le déterminant de $(\mathbf{X}_0'\mathbf{X}_0)$.