REVUE DE STATISTIQUE APPLIQUÉE

G. HUDON

Une comparaison des résultats de modèles log-linéaires et de généralisations de l'analyse des correspondances

Revue de statistique appliquée, tome 38, nº 2 (1990), p. 43-53

<http://www.numdam.org/item?id=RSA_1990__38_2_43_0>

Rev. Statistique Appliquée, 1990, **XXXVIII** (2), 43-53

UNE COMPARAISON DES RÉSULTATS DE MODÈLES LOG-LINÉAIRES ET DE GÉNÉRALISATIONS DE L'ANALYSE DES CORRESPONDANCES

G. HUDON

Département de mathématiques et d'informatique
Université du Québec à Montréal

1. Introduction

Le but de cet article est de comparer les résultats de trois méthodes d'analyse avec un même ensemble de données. Nous choisissons un exemple très simple, un tableau $2 \times 2 \times 2$, ce qui permettra une analyse plus facile qu'un gros ensemble de données. On veut comparer les résultats obtenus par les modèles log-linéaires, l'analyse des correspondances multiples et une autre généralisation de l'analyse des correspondaces.

2. Les Données

Voici un tableau contenant des données sur 13832 meurtres commis aux Etats-Unis en 1970 :

Type de meurtre

Race	*Sexe*	*Fusil*	*Couteau*	*Total*
Blanche	*Homme*	3910	808	4718
	Femme	1050	234	1284
Noire	*Homme*	5218	1385	6603
	Femme	929	298	1227
Total		11107	2725	13832

Dans cet exemple, nous avons trois variables qualitatives : la race, le sexe et le type de meurtre. Ce qui nous intéresse, c'est d'examiner les relations qui existent entre les trois variables.

3. Modèles log-linéaires

Soit x_{ijk} la fréquence observée pour la cellule ijk. Par exemple x_{111} serait la fréquence observée (3910) lorsque Race = Blanche, Sexe = Homme et Type = Fusil. On note les variables par leur première lettre, R = Race, S = Sexe, T = Type et on note aussi $\mu_{ijk} = E[x_{ijk}]$, la fréquence espérée pour la cellule ijk. Le premier modèle que nous examinons est le modèle d'indépendance :

$$log(\mu_{ijk}) = u + u_{R(i)} + u_{S(j)} + u_{T(k)} \qquad i = 1,2 \quad j = 1,2 \quad k = 1,2$$

avec les contraintes

$$\sum_{u_{R(i)}} = \sum_{u_{S(j)}} = \sum_{u_{T(k)}} = 0$$

Ce type de modèle est en fait un modèle multiplicatif. On veut exprimer la fréquence attendue comme un produit de termes qui ne dépendent que des variables prises individuellement. On obtient une estimation de μ_{ijk} par des méthodes bien connues et on la note m_{ijk}. On mesure le degré de divergence entre les fréquences observées et les fréquences attendues par le khi-deux :

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} \frac{(x_{ijk} - m_{ijk})^2}{m_{ijk}}$$

On note ce modèle par les variables, soit le modèle R, S, T. Dans notre exemple, nous avons $\chi^2 = 117.02$, ce qui indique que les trois variables qualitatives ne sont pas indépendantes. Voici maintenant le tableau des fréquences observées avec les m_{ijk} entre parenthèses.

		Type de meurtre	
Race	*Sexe*	*Fusil*	*Couteau*
Blanche	*Homme*	3910(3944, 64)	808(967.78)
	Femme	1050(874.92)	234(214.66)
Noire	*Homme*	5218(5146.04)	1385(1262.53)
	Femme	929(1141.39)	298(280.04)

Une fois qu'on a trouvé qu'il y a un lien entre les trois variables qualitatives, il est important de pouvoir spécifier la nature du lien qui existe entre ces variables. Nous regardons maintenant le modèle saturé, c'est-à-dire un modèle qui a autant des paramètres que de cases. On postule

$$\begin{aligned} log(\mu_{ijk}) = {} & u + u_{R(i)} + u_{S(j)} + u_{T(k)} + u_{RS(ij)} + u_{RT(ik)} \\ & + u_{ST(jk)} + u_{RST(ijk)} \end{aligned}$$

avec les contraintes (le + indique une somme sur cet indice)

$$u_{R(+)} = u_{S(+)} = u_{T(+)} = u_{RS(i+)} = u_{RS(+j)} = u_{RT(i+)} =$$
$$u_{RT(+k)} = u_{ST(j+)} = u_{ST(+k)} = u_{RST(ij+)} = u_{RST(+jk)} = u_{RST(i+k)} = 0$$

On note ce modèle RST. Il contient l'interaction triple RST et les interactions doubles RS, RT et ST. C'est toujours un modèle multiplicatif, sauf qu'il est saturé, c'est-à-dire que l'on obtient des fréquences attendues égales aux fréquences observées. On a aussi $\chi^2 = 0$. Si on pose $u_{RST(ijk)} = 0$. On obtient un modèle qui ne contient pas l'interaction triple RST mais qui possède les interactions doubles RS, RT et ST. On postule pour ce modèle :

$$log(\mu_{ijk}) = u + u_{R(i)} + u_{S(j)} + u_{T(k)} + u_{RS(ij)} + u_{RT(ik)} + u_{ST(jk)}$$

avec les contraintes appropriées. On note ce modèle RS, RT, ST. On obtient le tableau des fréquences attendues m_{ijk} :

Type de meurtre

Race	*Sexe*	*Fusil*	*Couteau*
Blanche	*Homme*	3919.4603	798.53965
	Femme	1040.5397	243.46035
Noire	*Homme*	5208.5397	1394.4603
	Femme	938.46035	288.53965

Pour ce modèle, nous avons $\chi^2 = 1.08$, avec un niveau de signification de 0,2997, ce qui indique que le modèle s'ajuste très bien aux données. Les fréquences attendues peuvent s'écrire sous la forme d'un produit, puisque c'est un modèle multiplicatif. On obtient :

Effet

Race	*Sexe*	*Type*	*Attendu*		*R*	*S*	*T*	*RS*	*RT*	*ST*
Blanche	*Homme*	*Fusil*	3919.4603	=(1057.1)	(0.893)	(2.066)	(1.999)	(0.908)	(1.071)	(1.035)
		Couteau	798.53965	=(1057.1)	(0.893)	(2.066)	(0.500)	(0.908)	(0.934)	(0.966)
	Femme	*Fusil*	1040.5397	=(1057.1)	(0.893)	(0.484)	(1.999)	(1.102)	(1.071)	(0.966)
		Couteau	243.4603	=(1057.1)	(0.893)	(0.484)	(0.500)	(1.102)	(0.934)	(1.035)
Noire	*Homme*	*Fusil*	5208.5397	=(1057.1)	(1.120)	(2.066)	(1.999)	(1.102)	(0.934)	(1.035)
		Couteau	1394.4603	=(1057.1)	(1.120)	(2.066)	(0.500)	(1.102)	(1.071)	(0.966)
	Femme	*Fusil*	938.46035	=(1057.1)	(1.120)	(0.484)	(1.999)	(0.908)	(0.934)	(0.966)
		Couteau	288.53965	=(1057.1)	(1.120)	(0.484)	(0.500)	(0.908)	(1.071)	(1.035)

On est maintenant en mesure d'examiner les effets. L'interaction RS nous donne un effectif plus élevé pour une femme blanche et pour un homme noir que les autres, toutes proportions gardées. On voit que l'interaction la plus forte est RS (1.102), la deuxième est RT(1.071) et la plus faible est ST (1.035). Pour vérifier ces résultats, voici les résultats de l'ajustement de plusieurs modèles à ces données :

Modèle	*dl*	χ^2	*niveau de signification*
RST	0	0	------
RS, RT, ST	1	1.08	0.2997
RS, RT	2	7.57	0.0227
RS, ST	2	39.92	0.0000
RT, ST	2	78.00	0.0000
R, S, T	4	117.02	0.0000

Pour le modèle RS, RT, on postule :

$$log\,(\mu_{ijk}) = u + u_{R(i)} + u_{S(j)} + u_{T(k)} + u_{RS(ij)} + u_{RT(ik)}$$

avec les contraintes appropriées. Les modèles RS, ST et RT, ST sont construits de la même façon. Le modèle le plus plausible est RS, RT, ST. Par contre le facteur ST ne semble pas être très significatif, puisque le modèle RS, RT donne un niveau de signification de 0.0227.

4. Analyse des correspondances multiples

Nous allons maintenant examiner les données avec l'analyse des correspondances multiples. On obtient le tableau disjontif complet suivant (après regroupement des lignes identiques) :

Etiquette

Rb	Rn	Sh	Sf	Tf	Tc
3910	0	3910	0	3910	0
808	0	808	0	0	808
1050	0	0	1050	1050	0
234	0	0	234	0	234
0	5218	5218	0	5218	0
0	1385	1385	0	0	1385
0	929	0	929	929	0
0	298	0	298	0	298

Sur ce tableau, nous appliquons l'analyse des correspondances. Il y a d'autres façons d'utiliser l'analyse des correspondances multiples, mais nous utilisons ici la méthode classique. On obtient les valeurs propres suivantes :

No	*Inertie*	*%*
1	0.36073	36.073
2	0.33882	33.882
3	0.30046	30.046

Voici les coordonnées (×100) et les contributions relatives pour les deux premiers axes :

Axe 1			Axe 2		
Etiquette	*Coord*	*Ctrel*	*Etiquette*	*Coord*	*Ctrel*
Rb	87.50	30.7	*Rb*	−8.3	0.3
Rn	−67.1	23.5	*Rn*	6.3	0.2
Sh	−28.8	6.3	*Sh*	−26.6	5.7
Sf	129.7	28.2	*Sf*	120.1	25.8
Tf	17.3	2.2	*Tf*	−41.2	13.4
Tc	−70.50	9.1	*Tc*	167.8	54.6

On obtient le graphique suivant des deux premiers axes :

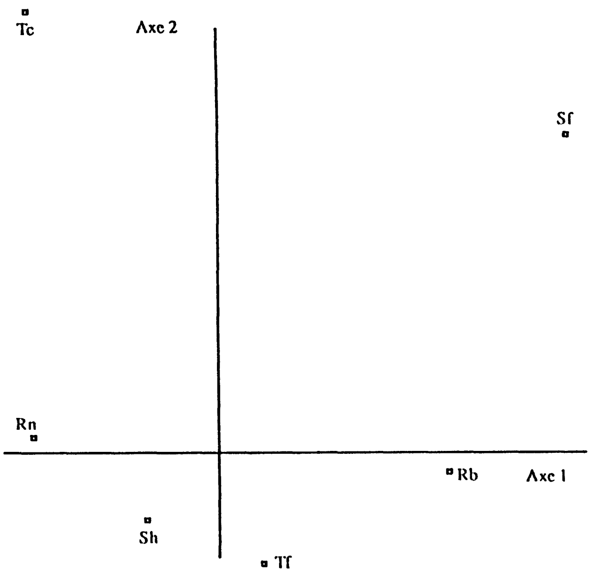

FIGURE 1

Il est clair que l'axe 2 est l'effet ST, les femmes préférant les couteaux. L'axe 1 est la superposition des deux effets RS et RT. Il est problématique de noter que l'effet ST est beaucoup plus faible que les effets RS et RT, mais l'axe 2 représente quand même 33.882 % de l'inertie totale. Il est aussi troublant de noter que si l'axe 1 est formé des deux effets RS et RT, on doit conclure que les femmes sont opposées aux couteaux sur cet axe. On obtient alors une contradiction qui semble difficile à résoudre.

Il faut aussi noter que l'analyse des correspondaces multiples n'est pas en mesure d'évaluer les interactions triples ou quadruples dans un ensemble de données. L'exemple que nous venons d'étudier est clairement un exemple bivarié, car on ne retient pas l'interaction triple. L'analyse des correspondances multiples devrait donc bien fonctionner dans un contexte semblable. Pour examiner plus en détail les interactions doubles, supposons que l'on enlève l'effet ST des données. C'est-à-dire que nous voulons étudier la matrice des fréquences attendues sous le modèle RS, RT. On a les fréquences suivantes pour ce modèle :

		Type de meurtre		
Race	*Sexe*	*Fusil*	*Couteau*	*Total*
Blanche	*Homme*	3898.9	819.1	4718
	Femme	1061.1	222.9	1284
Noire	*Homme*	5183.7	1419.3	6603
	Femme	963.3	263.7	1227
Total		11107	2725	13832

Nous appliquons l'analyse des correspondances multiples sur le tableau disjonctif complet. On obtient les valeurs propres suivantes :

N°	*Inertie*	%
1	0.36388	36.388
2	0.33214	33.214
3	0.30398	30.398

Voici les coordonnées ($\times 100$) et les contributions relatives pour les deux premiers axes :

	Axe 1			Axe 2	
Etiquette	*Coord*	*Ctrel*	*Etiquette*	*Coord*	*Ctrel*
Rb	-83.5	27.7	Rb	-1.7	0.0
Rn	64.0	21.3	Rn	1.3	0.0
Sh	28.5	6.1	Sh	26.9	6.0
Sf	-129.0	27.5	Sf	-121.0	26.9
Tf	-21.6	3.4	Tf	40.5	14.0
Tc	88.1	14.0	Tc	-165.0	53.9

Le deuxième axe demeure un axe ST, en effet qui a été enlevé du modèle. Le premier axe demeure aussi un axe RS, RT avec la même contradiction concernant Sf et Tc. Les femmes sont opposées aux couteaux sur l'axe 1, elles sont favorables aux couteaux sur l'axe 2, mais conditionnellement à la race, il n'y a pas de lien entre S et T avec ce tableau de données. Ces résultats sont pratiquement identiques à ceux de l'analyse précédente, ce qui est logique puisque l'interaction ST est peu significative.

5. Analyse des correspondances mutivariées

Escofier (1983) a généralisé l'analyse des correspondances pour examiner la différence entre deux matrices de nombres non-négatifs. Van Der Heijden (1985) a utilisé les tableaux multiples et la généralisation d'Escofier pour décomposer la différence entre deux matrices, chacune distribuée selon un modèle log-linéaire. Dans ses analyses, chaque matrice devait avoir les mêmes marginales.

Ici nous utilisons un cas particulier de la généralisation d'Escofier pour étudier la différence entre la matrice de données et la matrice des fréquences attendues sous l'hypothèse d'indépendance. Si on considère par exemple le tableau $R \times ST$, on cherche à comparer le tableau des x_{ijk} au tableau des $x_{i++}x_{+j+}x_{++k}/n^2$ (où + indique une somme sur l'indice correspondant et n est le nombre d'observations) avec les pondérations x_{i++}/n pour i et $x_{+j+}x_{++k}/n^2$ pour le couple (j,k), ce qui permet de définir les profils et les poids des lignes i et des colonnes (j,k) des deux tableaux que l'on compare.

Rappelons qu'avec les pondérations précédentes, la comparaison de ces deux tableaux revient à analyser les deux nuages de points suivants :

1) Nuage $N(I)$des points i de $\mathbf{R}_{JK}$de coordonnées :

$$\frac{x_{ijk} - x_{i++}x_{+j+}x_{++k}/n^2}{x_{i++}/n}$$

i étant affecté de la masse x_{i++}/n, et l'espace $\mathbf{R}_{JK}$ étant muni de la métrique de matrice la matrice diagonale des inverses des $X_{+j+}X_{++k}/n^2$.

2) Nuage $N(JK)$ des points (j,k) de $\mathbf{R}_I$ de coordonnées :

$$\frac{x_{ijk} - x_{i++}x_{+j+}x_{++k}/n^2}{x_{+j+}x_{++k}/n^2}$$

le point (j,k) étant affecté du poids $x_{+j+}x_{++k}/n^2$, et l'espace $\mathbf{R}_1$ étant muni de la métrique de matrice, la matrice diagonale des inverses des x_{i++}/n.

On peut noter que ces deux analyses se font sans centrer (i.e. par rapport à l'origine). Comme les deux tableaux comparés ont même marge sur I, et des marges différentes sur JK, le centre de gravité du nuage $N(JK)$ est à l'origine, alors que le centre de gravité de $N(I)$ qui a pour coordonnées $x_{+jk} - x_{+j+}x_{++k}/n$, n'y est pas. Il en résulte, puisque l'on a Cardl = 2, que l'on a deux axes factoriels non triviaux (et non pas un). Nous appellerons cette méthode "Analyse des correspondances multivariées".

Pour le tableau qui nous intéresse, le χ^2 d'indépendance est donné par la formule suivante :

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \frac{\left(x_{ijk} - x_{i++}x_{+j+}x_{++k}/n^2\right)^2}{x_{i++}x_{+j+}x_{++k}/n^2}$$

On calcule les termes suivants qui interviennent dans le calcul du χ^2 :

$$\frac{x_{ijk} - x_{i++}x_{+j+}x_{++k}/n^2}{\sqrt{x_{i++}x_{+j+}x_{++k}}}$$

et on forme une matrice X avec ces termes, en croisant deux des variables. Par exemple, le tableau multiple $RS\times T$ serait

3910	808
1050	234
5218	1385
929	298

et la matrice X serait une matrice 4×2 formée des termes appropriés. Le tableau multiple $R \times ST$ serait :

3910	808	1050	234
5218	1385	929	298

De la matrice $X'X$,nous calculons les valeurs propres. La somme des valeurs propres donne χ^2/n^2. C'est donc une décomposition du χ^2 d'indépendance. Chaque valeur propre représente l'inertie d'un axe. Dans notre cas particulier nous aurons deux axes. Les coordonnées sont la projection du profil d'une ligne de la matrice X sur les vecteurs propres (de norme 1) de la matrice $X'X$ ou d'une colonne de la matrice X sur les vecteurs propres (de norme 1) de la matrice XX'. Il est à remarquer que les valeurs propres non-nulles de XX' et $X'X$ sont les mêmes. Les contributions aux axes sont le carré des coordonnées pondérées par la masse de la ligne ou de la colonne.

Les valeurs propres que nous calculons dépendent du tableau multiple utilisé. Dans notre exemple, le χ^2 d'indépendance donne 117.02. Voici la décomposition de ce χ^2 selon le tableau multiple utilisé.

	Axe 1		Axe 2	
Tableau	χ^2	%	χ^2	%
$RS \times T$	77.32	66.1	39.70	33.9
$R \times ST$	112.73	96.3	4.29	3.7
$RT \times S$	81.33	69.5	35.69	30.5

Le tableau $R \times ST$ contient pratiquement toute son inertie sur l'axe 1. Il serait donc préférable d'utiliser le tableau $RS \times T$ ou le tableau $RT \times S$. Voici les coordonnées ($\times 100$) et les contributions relatives (%) pour les deux premiers axes pour le tableau $RS \times T$:

	Axe 1			Axe 2	
Etiquette	*Coord*	*Ctrel*	*Etiquette*	*Coord*	*Ctrel*
$RbSh$	2.23	3.2	$RbSh$	−7.03	61.1
$RbSf$	−18.37	47.50	$RbSf$	0.35	0.0
$RnSh$	−2.09	3.6	$RnSh$	3.97	25.5
$RnSf$	15.77	45.7	$RnSf$	6.11	13.4
Tf	−8.18	96.1	Tf	−1.19	4.0
Tc	−3.35	3.9	Tc	11.82	96.0

On obtient le graphique suivant des deux premiers axes :

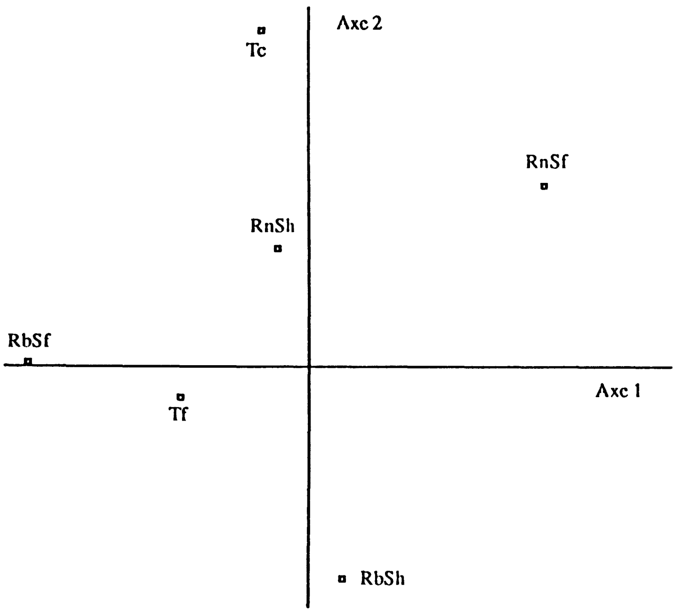

FIGURE 2

L'axe 1 est formé principalement de l'effet RS (opposition manifeste entre $RbSf$ et $RnSf$) tandis que l'axe 2 nous révèle l'effet $RT(RnSh, RnSf$ et Tc sont tous positifs). Pour compléter la comparaison, étudions la matrice des fréquences attendues sous le modèle RS, RT, comme nous avons fait pour l'analyse des correspondances multiples. On a le tableau des fréquences attendues :

		Type de meurtre		
Race	*Sexe*	*Fusil*	*Couteau*	*Total*
Blanche	*Homme*	3898.9	819.10	4718
	Femme	1061.1	222.9	1284
Noire	*Homme*	5183.7	1419.3	6603
	Femme	963.3	263.7	1227
Total		11107	2725	13832

On fait l'analyse avec le tableau $RS \times T$. le χ^2 d'indépendance donne 111.79. Voici la décomposition de ce χ^2 :

	Axe 1		Axe 2	
Tableau	χ^2	%	χ^2	%
$RS \times T$	74.92	67.0	36.87	33.0

Voici les coordonnées ($\times 100$) et les contributions relatives (%) pour les deux premiers axes pour le tableau $RS \times T$:

	Axe 1			Axe 2	
Etiquette	*Coord*	*Ctrel*	*Etiquette*	*Coord*	*Ctrel*
$RbSh$	3.80	9.5	$RbSh$	−5.75	44.1
$RbSf$	−18.02	47.3	$RbSf$	−6.45	12.3
$RnSh$	−2.91	7.2	$RnSh$	4.72	38.8
$RnSf$	13.77	36.0	$RnSf$	3.52	4.8
Tf	−7.45	82.4	Tf	−2.42	17.6
Tc	−6.96	17.6	Tc	10.56	82.4

On voit encore plus clairement l'interprêtation des axes. L'axe 1 représente l'effet RS et l'axe 2 reflète l'effet RT. On voit bien que l'analyse des correspondances multivariées donne un résultat qui concorde avec celui obtenu par le modèle log-linéaire RS, RT.

Pour terminer notre comparaison, nous pouvons faire l'analyse des correspondances usuelles du tableau multiple $RS \times T$. Nous obtenons un seul axe factoriel avec les coordonnées ($\times 100$) et les contributions relatives (%) suivantes :

Etiquette	*Coord*	*Ctrel*
$RbSh$	-6.5	44.3
$RbSf$	-3.7	4.0
$RnSh$	3.2	15.2
$RnSf$	11.5	36.5
Tf	-2.8	19.7
Tc	11.5	80.3

Cet axe reflète les interactions RT et ST que l'on retrouve dans le tableau des fréquences attendues du modèle RS, RT, ST :

– opposition entre Tc et $RbSh$: coefficients d'interaction $TcRb(0.934)$ et $TcSh(0.966)$ plus petits que 1.

– similarité entre Tc et $RnSf$: coefficients d'interaction $TcRn$ (1.071) et $TcSf(1.035)$ plus grands que 1.

6. Références

BENZÉCRI, J.-P. (1973) : L'Analyse des Données, Tome 2 : L'Analyse des Correspondances, Paris, *Dunod.*

BISHOP, Y.M.M., FIENBERG, S.E. & HOLLAND P.W. (1975) : Discrete Multivariate Analysis. *The M.I.T. Press, Cambridge, Massachusetts.*

ESCOFIER, B. (1979) : Traitement Simultané de Variables Qualitatives et Quantitatives en Analyse Factorielle. *Cahiers de l'Analyse des Données,* 4, 137-146.

ESCOFIER, B. (1983) : Analyse de la différence entre deux mesures sur le produit de deux mêmes ensembles. *Cahiers de l'Analyse des Données,* 3, 325-329.

GREENACRE M.J. (1984) : Theory and Applications of Correspondence Analysis. London, Academic Press.

GREENACRE, M.J. and HASTIE, T. (1987) : The Geometric Interpretation of Correspondence Analysis. Journal of the American Statistical Association, 82, 437-447.

LEBART,L., MORINEAU, A. and WARWICK, K. (1984) : Multivariate Descriptive Statistical Analysis, New York, *John Wiley.*

TENENHAUS, M. and YOUNG, F.W. (1985) : An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Data, *Psychometrika,* 50,91-119.

VAN DER HEIJDEN, P.G.M. (1985) : Correspondence Analysis used complementary to Loglinear Analysis, *Psychometrika,* 50, 429-447.