Math. & Sci. hum., (40e année, n° 160, *2002*, p. 13-45)

RÉGRESSION ET ANALYSE GÉOMÉTRIQUE DES DONNÉES : RÉFLEXIONS ET SUGGESTIONS

Henry ROUANET [1], Frédéric LEBARON [2], Viviane LE HAY [3], Werner ACKERMANN [4], Brigitte LE ROUX [5]

RÉSUMÉ – *Les données multivariées sont souvent traitées par les méthodes de régression d'une part, l'Analyse Géométrique des Données (*ACP, AC*...) d'autre part. Nous nous proposons de montrer sur des exemples, à partir de la communauté des structures mathématiques, comment on peut intégrer les méthodes de régression dans l'analyse géométrique, et visualiser les effets de structure. Il n'y a pas lieu d'opposer des méthodes statistiques qui seraient par essence « explicatives » à d'autres qui seraient par essence « descriptives ».*

MOTS CLÉS – Effet de structure, Régression, Analyse Géométrique des Données, Descriptif versus explicatif.

SUMMARY – Regression and Geometric Data Analysis : Reflections and Suggestions
*Multivariate data are often treated with regression methods on one hand, Geometric Data Analysis methods (*PCA, CA*...) on the other hand. We intend to show, thanks to the mathematical structures common to the two methods, illustrated by examples, how one can integrate regression methods in geometric analysis. Geometric analysis allows a visualization of structural effects. There is no ground to oppose « explanatory » and « descriptive » statistical methods.*

KEYWORDS – Structural effects, Regression, Geometric Data Analysis, Descriptive versus explanatory.

1. INTRODUCTION

Les données multivariées sont souvent traitées par les méthodes de régression d'une part, les méthodes d'Analyse Géométrique des Données (ACP, AC...) d'autre part. Nous nous proposons de montrer, à partir de la communauté de structures mathématiques, comment les méthodes de régression peuvent être intégrées dans l'analyse géométrique.

Le plan de l'article est le suivant : Effet de structure et régression (§2.) ; représentations gométriques, exemple « Ouvrier » (§3.) ; régression linéaire et ACP, Dossier « Biscuits » (§4.) ; commentaires finaux (§5.).

[1] CRIP5, Université René Descartes, rouanet@math-info.univ-paris5.fr
[2] CSE-IRESCO, Université de Picardie Jules Verne, frlebaron@wanadoo.fr
[3] Doctorante, OSC (IEP de Paris), viviane.lehay@caramail.com
[4] CSO, FNSP/CNRS, w.ackermann@cso.cnrs.fr
[5] MAP5, FRE 2428, CNRS, Université René Descartes, lerb@math-info.univ-paris5.fr

2. EFFET DE STRUCTURE ET RÉGRESSION

2.1. Le « paradoxe des lycées »

L'exemple construit du « paradoxe des lycées » [Rouanet, 1978] nous servira de situation de mise en train. Dans un débat télévisé sur « les filles et l'éducation », on commente les résultats au bac des élèves de la ville de Kingborn l'année passée (nombre de réussites et d'échecs) (Tableau 1).

	+(réussite)	−(échec)	
Garçons	24	36	60
Filles	36	24	60

Tableau 1. Paradoxe des lycées : tableau de base.

Les taux de succès sont 24/60 = 40% pour les garçons, et 36/60 = 60% pour les filles, soit une différence de 20% en faveur des filles. Les filles réussissent donc mieux que les garçons à Kingborn. Mais un participant signale qu'à Kingborn il y a deux lycées (Roméo et Juliette) et présente les résultats pour chacun d'eux (Tableau 2). Au lycée Roméo, les taux de succès sont de 30% pour les garçons, de 10% pour les filles ; au lycée Juliette, les taux sont de 90% pour les garçons, de 70% pour les filles. Ainsi, dans chacun des deux lycées, on a une différence de 20% en faveur des garçons. Les garçons réussissent donc mieux que les filles à Kingborn !

	Roméo			Juliette		
	+	−		+	−	
Garçons	15	35	50	9	1	10
Filles	1	9	10	35	15	50

Tableau 2. Paradoxe des lycées. Effectifs dans les deux lycées.

Stupeur : les deux conclusions sont opposées. Comment cela est-il possible ? Il n'y a pas d'erreur dans les chiffres. En additionnant les deux tableaux case par case, on retrouve bien le tableau de base. Dans ces conditions, lequel des deux effets est l'« effet vrai » du sexe sur la réussite ? Et pour commencer, y a-t-il un « effet vrai » ?

2.2. Effet de structure

L'exemple des lycées illustre le cas paradoxal du phénomène appelé classiquement *effet de structure*. L'effet du facteur[6] Sexe sur la Réussite n'est pas le même selon qu'on le considère au niveau *global*, ou *conditionnellement* au facteur Lycée, parce que les facteurs Sexe et Lycée sont *corrélés* (Roméo est en gros un « lycée de garçons », Juliette un « lycée de filles ») ; avec des facteurs non–corrélés (orthogonaux), l'effet conditionnel serait égal à l'effet global.

[6]Dans ce texte, nous entendons par facteur (par analogie avec des facteurs d'un plan d'expérience), une variable indépendante codée en un nombre fini de modalités, à distinguer d'une variable principale (variable factorielle) issue d'une analyse géométrique.

L'effet de structure présente un caractère d'ubiquité dans les données d'observation, et se rencontre dès l'examen de tableaux croisés à plusieurs variables [Novi, 1998]. Il est familier aux démographes et aux économistes ; les commentaires de Simiand et de Halbwachs, rappelés par Desrosières [1982], sont célèbres, avec la fameuse parabole du renne et du chameau. Eliminer l'effet de structure revient à raconter comment vivrait un chameau si, restant chameau, il était transporté dans les régions polaires (et *mutatis mutandis* pour le renne)[7].

Notre propos sera non pas d'éliminer les effets de structure mais de les analyser[8]. Lorsqu'on passe de l'effet global (« main effect ») d'un facteur sur une variable à un effet conditionnellement à d'autres facteurs, on peut avoir *accentuation* de l'effet, avec à la limite *émergence* de l'effet ; *atténuation*, avec à la limite *disparition* de l'effet ; *stabilité* de l'effet (effet conditionnel égal à l'effet global) ; *renversement* de l'effet (changement de signe), comme dans le cas paradoxal des lycées.

Nous illustrerons les divers cas sur plusieurs exemples artificiels inspirés du paradoxe des lycées (avec des valeurs numériques légèrement modifiées pour faciliter les calculs) ; il nous suffira de prendre deux facteurs à deux modalités : A (qu'on appellera encore facteur Lycée), de modalités a (Roméo) et a' (Juliette), et B (qu'on appellera encore facteur Sexe), de modalités b (garçon) et b' (fille), avec le tableau croisé des effectifs suivant (Tableau 3) :

	b (Garçons)	b' (Filles)	
a (Roméo)	40	10	50
a' (Juliette)	10	40	50
	50	50	100

Tableau 3. Effet de structure. Tableau croisé des effectifs.

Le croisement $A \times B$ est marginalement équilibré (même nombre d'élèves dans les deux lycées, et autant de garçons que de filles), mais il est non–orthogonal. Le coefficient de corrélation r (point–tétrachorique) entre A et B vaut 0.6, son carré Φ^2 (carré moyen de contingence) vaut .36. Considérons maintenant une variable de réussite y codée en deux modalités (succès $+$, échec $-$). En prenant le facteur Sexe B comme facteur d'intérêt, on considérera l'effet global (noté B), différence des fréquences de succès globales entre Garçons et Filles, et l'effet conditionnel, noté B/A, différence des fréquences de succès conditionnellement au facteur A. Les exemples construits ci–après illustrent les situations remarquables suivantes : Renversement, Disparition, Stabilité, Emergence. Pour centrer l'étude sur l'effet de structure, les exemples présentent les deux propriétés suivantes. D'abord, dans chaque situation, la fréquence générale de succès est toujours .50. Ensuite, on a égalité de l'effet

[7] De nos jours, on se demanderait à coup sûr quel serait le salaire d'une femme si restant femme, elle avait un âge, niveau de qualification, secteur, grade, etc. égaux à ceux des hommes.

[8] La démarche suivie ici pour étudier l'effet de structure pourra être rapprochée de celle qu'on trouve dans les articles consacrés aux comparaisons des inégalités dans la *Revue Française de Sociologie*, avec notamment Combessie [1984], Barbut [1984], Grémy [1984], Prévot [1985], Merllié [1985], Vallet [1988], ainsi que dans le numéro spécial de *Mathématiques et Sciences humaines* [93, 1986].

de B pour a (différence des fréquences de succès entre garçons et filles dans le lycée a) et de l'effet de B pour a' (différence des fréquences de succès dans le lycée a') ; en pareil cas, l'effet conditionnel B/A est défini sans ambiguïté comme la valeur commune[9] des deux effets pour a et a'. Dans ce qui suit, nous détaillerons la situation Renversement, et nous donnerons simplement les résultats pour les trois situations Stabilité, Disparition, Émergence.

2.3. Situation Renversement (« Reversal »)

Le Tableau 4 illustre la situation Renversement :

	+	−		
ab	12	28	40	
ab'	1	9	10	
$a'b$	9	1	10	
$a'b'$	28	12	40	
	50	50	100	Fréquence générale de + : 50/100=.50
b	21	29	50	Fréquence de + pour b : $21/50 = .42$
b'	29	21	50	Fréquence de + pour b' : $29/50 = .58$

Tableau 4. Renversement : Effectifs de base et effectifs pour B.

On en déduit l'effet global du facteur B : $.42 - .58 = -.16$

Conclusion : Globalement, les filles réussissent mieux que les garçons.

Examinons maintenant les fréquences de réussite pour b (garçons) et pour b' (filles) conditionnellement à la modalité a (Roméo) et à la modalité a' (Juliette) du facteur Lycées (Tableau 5).

	Lycée a		Lycée a'
	+		+
ab	.30	$a'b$	.90
ab'	.10	$a'b'$	.70

Tableau 5. Renversement. Fréquences de succès pour b et b' conditionnellement à a et conditionnellement à a', et effets conditionnels B/a : $.30 - .10 = +.20$ et $B/a' = .90 - .70 = +.20$.

Les deux effets du facteur B (Sexe) conditionnellement au lycée a et au lycée a' sont égaux ; leur valeur commune est par définition l'effet du facteur B (Sexe) conditionnellement au facteur A (Lycées), noté B/A. On a donc Effet $B/A = +.20$.

Conclusion : Conditionnellement au facteur Lycée, les garçons réussissent mieux que les filles. L'effet conditionnel est de signe opposé à l'effet global : *renversement* !

[9]Lorsque les effets conditionnels sont inégaux, l'effet de B conditionnellement à A est défini comme une certaine moyenne des effets conditionnels, ainsi que nous l'illustrerons sur l'exemple « Ouvrier » au §3.4. L'homogénéité des effets conditionnels est souvent appelée « absence d'interaction ». L'effet de structure n'a rien à voir avec l'effet d'interaction : dans tous les exemples de ce §2. qui illustrent l'effet de structure, l'effet d'interaction est nul.

2.4. Régressions simple et multiple

Les effets globaux et conditionnels s'étudient classiquement par les procédures de régression. Le *modèle–cadre* d'une régression est défini par un ensemble de variables, par le statut de *variable à prédire* (on dit aussi « variable dépendante ») donné à l'une des variables, par celui de *variables prédictrices* (on dit aussi « variables indépendantes ») donné aux autres variables[10], ainsi que par la *fonction de lien* entre les deux types de variables : linéaire (variables numériques, éventuellement après codage), logistique (variables catégorisées), etc.

Pour la discussion, nous considérerons une variable à prédire y (notation consacrée : « la variable y de la régression »), et nous envisagerons d'une part la régression linéaire simple de y sur une variable, d'autre part la régression de y sur deux variables. Les formules de base de la régression linéaire sont rappelées au §2.7. (p. 21). Dans nos exemples, chacun des deux facteurs (Lycée et Sexe) est à deux modalités. En prenant comme variables prédictrices, codées en (1, 0), la variable x_a indicatrice du lycée Roméo, et la variable x_b indicatrice de Garçon (autrement dit en prenant comme modalités de référence Juliette et Fille), les deux régressions linéaires pertinentes pour l'étude de l'effet du facteur B sur la variable dépendante Réussite y (elle aussi codée en 1 et 0) sont les suivantes :

— *Régression simple* $\widetilde{y}^b$ de y sur la variable x_b (Sexe) : $\widetilde{y}^b = ux_b + u_0$.

Le coefficient u correspond à l'effet global du facteur B (Sexe).

— *Régression multiple* $\widetilde{y}^{a+b}$ de y sur x_a et x_b[11] : $\widetilde{y}^{a+b} = u_a x_a + u_b x_b + u_0$.

Les coefficients de régression partiels u_a et u_b correspondent respectivement aux effets conditionnels A/B (effet de A conditionnellement à B) et B/A (effet de B conditionnellement à A). Le coefficient R^2, carré de la corrélation multiple R (corrélation entre y et $\widetilde{y}^{a+b}$), exprime la qualité de l'ajustement de la variable dépendante y par la régression multiple.

Dans la situation de renversement, on a les équations suivantes[12] :

— *Régression simple sur* x_b : $\widetilde{y}^b = -0.16x_b + 0.58$.

La valeur 0.58 est la valeur prédite de y pour fille. Le coefficient de régression simple -0.16 est l'effet global du facteur Sexe. (Par ailleurs, on a la régression simple

[10] *Variable dépendante, variables indépendantes* : ces appellations, ainsi que celle d'effet, renvoient à la *méthodologie expérimentale*, dans laquelle les variables indépendantes sont les variables sous contrôle direct de l'expérimentateur ; transportées aux données d'observation, ces appellations sont métaphoriques. Leur intérêt est précisément d'être des marqueurs de l'allégeance de la méthodologie statistique des données d'observation vis-à-vis de la méthodologie expérimentale, allégeance bien analysée par Passeron [1991, p. 129], qui parfois, il faut le dire, confine à l'hyper-expérimentalisme. Dans maintes recherches économétriques, on trouve des traces de la méthodologie expérimentale, toutes les fois que sont privilégiées des variables « sur lesquelles on pourrait agir » ; ainsi, pour déterminer les facteurs de la variation du taux de chômage dans différents pays, on se tourne vers des variables sur lesquelles la politique économique, la législation ou les acteurs sociaux nationaux peuvent agir, comme des indicateurs de flexibilité du marché du travail ou des variables résumant les « chocs macroéconomiques ». Exemple : Fitoussi & *al.* [2000, p. 47-49].

[11] L'indice « $a + b$ » a pour but de souligner la propriété additive de la régression multiple.

[12] Dans cet exemple, la variable y, ainsi que les variables x_a et x_b ont même variance (à savoir $1/4$), donc les coefficients de régression simples sont égaux aux coefficients de corrélation, et les coefficients de régression partiels égaux aux coefficients réduits (les « beta-weights »).

sur x_a : $-0.48x_a + 0.74$, d'où l'effet global du facteur Lycée : -0.48).

— *Régression multiple sur x_a et x_b* : $\tilde{y}^{a+b} = -0.60x_a + 0.20x_b + 0.70$.

Le coefficient -0.60 est l'effet A/B, valeur commune de $A/b = .30 - .90$ et $A/b' = .10 - .70$; la valeur 0.70 est la valeur prédite de y pour (Juliette, Fille). Le coefficient de régression partiel $+0.20$ est l'effet du facteur Sexe conditionnellement au facteur Lycée : cf. Tableau 5 (p. 16). On passe donc de -0.16 à $+0.20$: renversement de l'effet.

En termes de proportion de variance prise en compte (cf. §2.7.), on passe pour le facteur Sexe de $(0.16)^2 = 0.0256$ (régression simple) à $R^2 = (-0.60) \times (-0.48) + (+0.20) \times (-0.16) = 0.288 - 0.032 = 0.256$[13].

2.5. Autres situations remarquables

Les trois autres situations remarquables évoquées plus haut (disparition, stabilité et émergence) sont rapidement présentées ci–après. Pour chacune d'entre elles, nous indiquons simplement l'effet global et l'effet conditionnel du facteur B, les deux équations de régression (simple et multiple), et la conclusion. Le Tableau 6 présente les effectifs de base des trois situations.

	+	−	
ab	8		40
ab'	2		10
$a'b$	8		10
$a'b'$	32		40
	50		100

6a. Disparition

	+	−	
ab	8		40
ab'	8		10
$a'b$	2		10
$a'b'$	32		40
	50		100

6b. Stabilité

	+	−	
ab	16		40
ab'	1		10
$a'b$	9		10
$a'b'$	24		40
	50		100

6c. Emergence

Tableau 6. Trois situations remarquables : effectifs de base.

2.5.1. Disparition de l'effet (« Vanishing ») (Tableau 6a)

• Effet global du facteur B (Sexe) : $(8+8)/(40+10) - (2+32)/(10+40) = -0.36$. Globalement, les filles réussissent mieux que les garçons.

• Effet conditionnel : $B/a = (8/40 - 2/10) = 0$, $B/a' = (8/10 - 32/40) = 0$, d'où $B/A = 0$. Conditionnellement au Lycée, garçons et filles réussissent aussi bien.

Conclusion : il y a effet global mais il n'y a pas d'effet conditionnel : *disparition.*

• Régression simple : $\tilde{y}^b = -0.36x_b + 0.68$.

• Régression multiple : $\tilde{y}^{a+b} = -0.60x_a + 0x_b + 0.80$, avec $R^2 = 0.360$.

[13] La décomposition de R^2 est additive (on a un terme pour chaque variable indépendante) ; mais elle n'est pas monotone : le deuxième terme est ici négatif (ce qui correspond au renversement de l'effet). Le premier terme de la somme (ici $.288 > R^2 = .256$) ne saurait être regardé comme un « R^2 partiel » exprimant la proportion de variance prise en compte par le facteur Lycée, pas plus que le second terme -0.032 ne saurait être regardé comme un « R^2 partiel » exprimant la proportion de variance prise en compte par le seul facteur Sexe.

2.5.2. Stabilité de l'effet (« Steady ») (Tableau 6b)

• Effet global du facteur B (Sexe) : -0.60. Globalement, les filles réussissent mieux que les garçons.
• Effet conditionnel : -0.60. Conditionnellement, les filles réussissent mieux que les garçons.

Conclusion : l'effet conditionnel est égal à l'effet global : *stabilité*.

• Régression simple : $\widetilde{y}^b = -0.60x_b + 0.80$.
• Régression multiple : $\widetilde{y}^{a+b} = 0x_a - 0.60x_b + 0.80$, avec $R^2 = 0.360$.

2.5.3. Emergence de l'effet (Tableau 6c)

• Effet global du facteurs B (Sexe) : 0. Globalement, les filles réussissent aussi bien que les garçons.
• Effet conditionnel : $+0.30$. Conditionnellement, les garçons réussissent mieux que les filles.

Conclusion : il n'y a pas d'effet global mais il y a un effet conditionnel : *émergence*.

• Régression simple : $\widetilde{y}^b = 0x_b + 0.50$.
• Régression multiple : $\widetilde{y}^{a+b} = -0.50x_a + 0.30x_b + 0.60$, avec $R^2 = 0.160$.

2.6. Prédiction et schéma explicatif

À ce point, quelques premières réflexions.

2.6.1. Prédiction

L'essence de la régression est la prédiction : à partir de ce qu'on connaît (les variables prédictrices), on cherche à se prononcer sur ce qu'on voudrait connaître (la variable à prédire y). On peut chercher à prédire la réussite d'un élève d'abord à partir du sexe, puis en ajoutant le lycée, puis le milieu socio-économique de la famille, etc. Dans la seule perspective de la prédiction, d'une part la liste des variables prédictrices du modèle–cadre n'est pas limitative, et les liaisons éventuelles entre ces variables (quasi-colinéarités, etc.) sont peu gênantes ; d'autre part, ce qui importe c'est la variable prédite $\widetilde{y}$ « en extension », c'est-à-dire définie par l'ensemble des valeurs prédites, avec bien sûr la valeur du coefficient multiple R (ou de son carré R^2), qui en tant qu'indice de qualité de l'ajustement, se doit d'être aussi élevé que possible[14]. Toujours dans cette seule perspective, l'idée de regarder un effet conditionnel comme un « effet vrai, toutes choses égales par ailleurs », ne trouve guère sa place.

[14] Cf. Riandey [1991] : « Quel institut affirmerait l'absence de demande latente pour le type d'estimation suivant : proportion d'intentions de vote pour le candidat X, de la part des femmes de 18 à 25 ans, actives, titulaires du bac, résidant dans les agglomérations d'au moins 200000 habitants ? »

2.6.2. Schéma explicatif

La régression met en jeu un schéma explicatif dès lors qu'on donne à la variable à prédire le statut de « variable à expliquer », et aux variables prédictrices celui de « variables explicatives »[15]. Dans cette perspective, on cherchera à expliquer la réussite des élèves à partir de plusieurs variables explicatives, disons le Sexe (variable d'intérêt) et de diverses variables d'environnement : Lycées, etc. Intuitivement, on cherche les « poids relatifs » qui reviennent aux diverses variables dans le schéma explicatif, et le coefficient de régression d'une variable explicative est interprété selon la phraséologie de l'« effet vrai, *toutes choses égales par ailleurs* » – c'est–à–dire de fait, conditionnellement aux autres variables retenues dans le modèle–cadre (cf. p. 17). Les deux cas de prédilection pour cette phraséologie sont la situation Disparition et la situation Stabilité.

Dans la situation *Disparition*, on dira : « Il n'y a pas d'effet vrai du facteur Sexe » ; l'effet global « les filles réussissent mieux que les garçons » n'est qu'« apparent », car il est « expliqué » par les facteurs d'environnement (Lycée, etc.)[16].

Dans la situation *Stabilité*, la meilleure réussite des filles sera réputée « effet vrai » : l'effet subsiste quand on tient compte de l'environnement (Lycée, etc.).

La situation *Emergence* semble beaucoup moins présente dans les discussions méthodologiques. Dans la logique précédente on devrait conclure encore à un « effet vrai », mais ne correspondant pas à un « effet apparent ».

Enfin, la situation *Renversement* – où l'effet vrai s'oppose à l'effet apparent – attire généralement l'attention par son caractère paradoxal (cf. par exemple [Vallet, Caille, 1995]).

Dans la perspective explicative, outre le R^2, la spécification des variables retenues dans le modèle–cadre, avec leurs liaisons éventuelles (quasi–colinéarités, etc.), et la définition « en compréhension » de la variable prédite (c'est–à–dire son expression en fonction des variables prédictrices) deviennent cruciales. Le cas d'émergence dans lequel « il y a un effet vrai sans avoir d'effet apparent »[17] pose de façon aigüe le problème du choix des variables explicatives. Pour qu'on puisse trouver un effet émergent,

[15] Il est clair que la procédure statistique de régression est en soi neutre vis–à–vis de la notion de schéma explicatif. On peut régresser la longueur d'une barre de métal sur la température – renvoyant au schéma explicatif de la dilatation – mais aussi bien régresser la température sur la longueur de la barre (simple perspective prédictive). Parler de « variable à expliquer, variables explicatives » revient à dissymétriser la problématique, donc à poser, ne serait ce que *de facto*, un schéma explicatif.

[16] *Exemple psychométrique.* Dans une école maternelle (avec des enfants d'âges différents), une tâche est mieux réussie par les enfants les plus grands ; l'effet de la taille disparaît quand on tient compte de l'âge ; on dira que « le facteur âge explique l'effet (apparent) de la taille sur la performance ». *Exemple sociologique* (cf. la controverse autour du mythe de la « bell curve » : [Fischer & *al.*, 1998]) ; on trouve un « main effect » de la variable « ethnic differences » sur le QI ; et lorsqu'on régresse le QI sur les deux variables « ethnic differences » et « environment », le coefficient de régression de « ethnic differences » vient après « environment » ; c'est donc que le facteur « environment » explique (au moins en partie) les variations dues au facteur « ethnic differences ».

[17] En psychométrie, une variable émergente est appelée « variable supprimante de Horst ». Selon Faverge [1966, tome 2, p. 204] : « Le second test permet d'éliminer du premier une partie qui n'est pas valide » (pour la prédiction de la variable y). Pour un exemple de quasi–émergence, cf. infra l'exemple des biscuits.

il faut évidemment que la variable correspondante fasse partie du modèle–cadre de la régression ; si le choix des variables a été guidé par l'existence d'effets globaux importants, on risque de laisser de côté une telle variable. Dans les applications, pour pouvoir qualifier un effet conditionnel d'« effet vrai, toutes choses égales par ailleurs », il faut se sentir bien assuré que le schéma explicatif dans lequel a été inscrite la procédure de régression contient bien toutes les variables pertinentes ! Ces difficultés sont bien connues des spécialistes[18].

L'usage de la régression en sciences sociales, même s'il remonte à une lointaine époque, est aujourd'hui devenu massif et calqué sur son usage en économétrie (« modèle économétrique » en vient à désigner tout modèle de régression même sans lien avec l'économie). Dans la hâte de trouver les « effets vrais », on en vient à *oublier* les effets globaux, qui pourtant donnent une information, sans doute imparfaite, mais plus sûre que les effets conditionnels, qui dépendent de façon cruciale du modèle-cadre[19].

Les réflexions précédentes invitent à une démarche plus prudente ; elles suggèrent de toujours calculer les effets globaux et de les comparer aux effets conditionnels, c'est-à-dire d'examiner les effets de structure. Nous allons voir maintenant comment ces comparaisons peuvent être rendues tout à fait intuitives, grâce à la *représentation géométrique des effets* (§3.).

2.7. Annexe. Rappel des formules de base de la régression linéaire

2.7.1. Variables réduites

Soient z, z_a et z_b des variables réduites ; r_a et r_b les coefficients de corrélation de z avec z_a et z_b respectivement, et r le coefficient de corrélation entre z_a et z_b. On a :

• Régression simple de z sur z_b : $\widetilde{z} = r_b z_b$.

• Régression multiple de z sur z_a et z_b : $\widetilde{z}^{a+b} = \beta_a z_a + \beta_b z_b$, avec $\beta_a = (r_a - r r_b)/(1-r^2)$ et $\beta_b = (r_b - r r_a)/(1-r^2)$. Qualité de l'ajustement $R^2 = \beta_a r_a + \beta_b r_b$.

2.7.2. Cas général

Étant données les variables y, x_a et x_b d'écarts-types respectifs σ, σ_a et σ_b :

• Régression simple $\widetilde{y} = u\, x_b + u_0$ avec $u = (\sigma/\sigma_b) r_b$ et $u_0 = \overline{y} - u\overline{x}_b$.

• Régression multiple : $\widetilde{y}^{a+b} = u_a x_a + u_b x_b + u_0$, avec $u_a = (\sigma/\sigma_a)\beta_a$, $u_b = (\sigma/\sigma_b)\beta_b$ et $u_0 = \overline{y} - u_a\overline{x}_a - u_b\overline{x}_b$.

Le rapport Effet conditionnel / Effet global est $u/u_b = \beta_b/r_b$[20].

[18] Nous parlerions de « pièges de la régression »... si la notion de « piège statistique » faisait partie de notre bagage intellectuel ! Relisons plutôt les pages magistrales de Malinvaud [1981, p. 236-237] sur les erreurs de spécification et les quasi–colinéarités ; on songe à la pensée de Pascal : « L'omission d'un principe mène à l'erreur ! »

[19] Par exemple, dans Fitoussi & *al.* (*op. cit.*, p. 48), on lit que le coefficient de Δec (variation du degré de coordination entre employeurs) passe de 2.22 dans l'équation de base à -2.40 dans l'équation augmentée de la variation du chômage (variable estimée par une autre équation).

[20] Dans l'exemple du renversement on a $r = +0.6$; $r_a = -0.48$, $r_b = -0.16$; $\sigma = \sigma_a = \sigma_b = 0.5$.

3. REPRÉSENTATIONS GÉOMÉTRIQUES

Dans cette partie, nous illustrons, sur les exemples précédents, puis sur l'exemple « Ouvrier », des représentations géométriques de la régression, pour deux variables prédictrices numériques, en liaison avec l'Analyse en Composantes Principales (ACP) ; puis nous proposons une méthode générale.

3.1. Représentation géométrique pour deux variables prédictrices

Cette représentation est classique : voir par exemple Kendall et Stuart [1973, Vol. 3, p. 340] ; sa justification est que dans la formalisation linéaire l'espace des variables centrées admet la covariance pour produit scalaire et l'écart-type pour norme [Rouanet, Le Roux, 1993, p. 59]. Deux variantes sont utilisées : celle des variables centrées, et celle des variables réduites ; ci–après nous détaillons cette dernière[21].

Traçons un cercle de centre O et de rayon unité (cercle des corrélations), et représentons d'abord les deux variables réduites z_a et z_b associées aux variables prédictrices x_a et x_b par deux rayons du cercle, et d'angle θ tel que $\cos^2\theta = \Phi^2 = 0.36$, soit donc $\Phi = 0.6$, d'où $\theta = 53°16$ (Figure 1). D'où l'axe $\mathcal{A}$ « Lycée » avec les deux points A (Roméo) et A′ (Juliette), et l'axe $\mathcal{B}$ « Sexe » avec les deux points B (Garçon) et B′ (Fille). L'axe $\mathcal{A}_\perp$ perpendiculaire à $\mathcal{B}$ est la « ligne de démarcation » séparant les effets qui vont « du côté des garçons » de ceux « du côté des filles ».

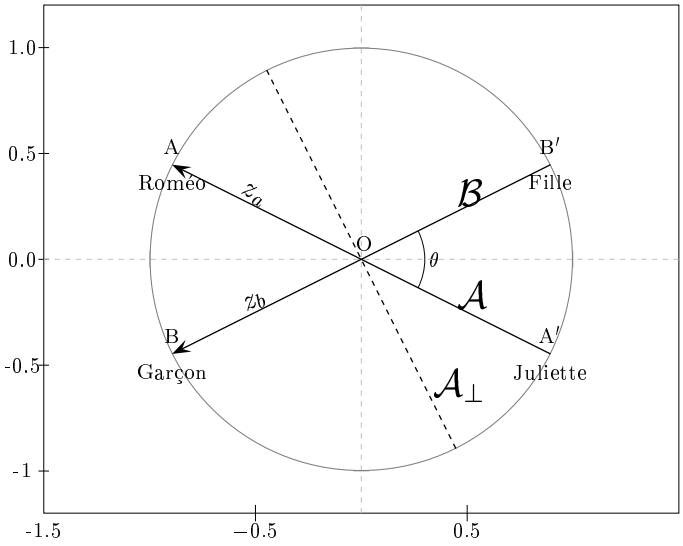

Figure 1. Cercle des corrélations et variables de base z_a et z_b ($\theta = 53°16$).

La variable réduite z associée à la variable à prédire (Réussite) correspond à un vecteur dont l'extrémité est sur la sphère de rayon unité, et qui se projette orthogonalement dans le plan selon le point M à l'intérieur du cercle de rayon unité. D'où la construction illustrée pour le cas « Renversement » par la Figure 2, p. 23.

[21] Dans l'exemple fictif des Lycées, les deux représentations sont indistinguables du fait que les variables ont toutes même moyenne (1/2) et même écart–type (1/2).

On construit sur l'axe $\mathcal{A}$ le point M_a de coordonnée -0.48 (effet global du facteur A), et sur l'axe $\mathcal{B}$ le point M_b de coordonnée -0.16 (effet global du facteur B). Le point M représentant la variable Réussite est l'intersection de la perpendiculaire à $\mathcal{A}$ menée par M_a et de la perpendiculaire à $\mathcal{B}$ menée par M_b ; les deux effets globaux $(-0.48, -0.16)$ sont les deux *coordonnées droites* du point M. Par ailleurs, projetons le point M d'une part sur l'axe $\mathcal{A}$ parallèlement à l'axe $\mathcal{B}$, d'où le point M'_a ; d'autre part sur l'axe $\mathcal{B}$ parallèlement à A, d'où le point M'_b ; on obtient les deux effets conditionnels $(-0.60, +0.20)$, comme *coordonnées obliques* du point M.

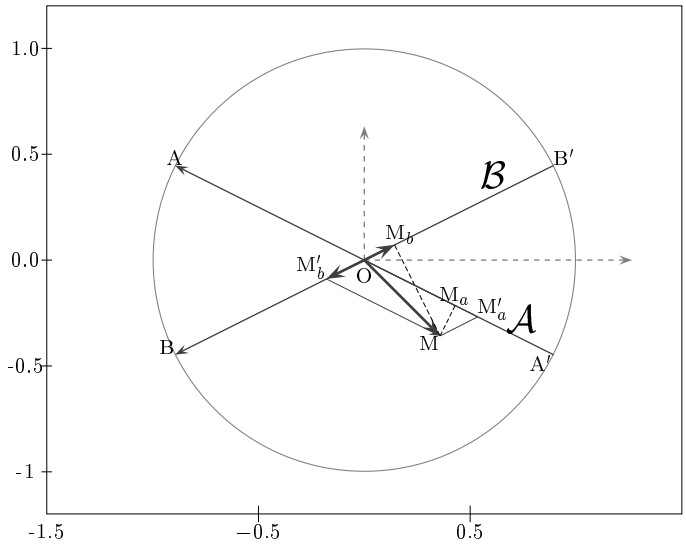

Figure 2. Situation Renversement. Effet global $B = -0.16$ (point M_b), effet conditionnel $B/A = +0.20$ (point M'_b), de signes opposés (avec effet global A (point M_a) et effet A/B (point M'_a)).

Autrement dit : l'effet global du facteur B s'obtient en projetant le point M sur l'axe $\mathcal{B}$ orthogonalement – d'où -0.16, du côté de B′ (filles) – et l'effet conditionnel, en projetant le point M sur l'axe $\mathcal{B}$ parallèlement à l'axe $\mathcal{A}$ – d'où $+0.20$, du côté de B (garçons). Comparer les deux effets revient à comparer les deux projections $\overrightarrow{\mathrm{OM}'_b}$ (oblique) et $\overrightarrow{\mathrm{OM}_b}$ (droite). On a ici $\overrightarrow{\mathrm{OM}'_b}/\overrightarrow{\mathrm{OM}_b} = (+0.20)/(-0.16) = -1.25$.

Le coefficient $R^2 = .256$ est égal au produit des coordonnées obliques par les coordonnées droites[22] ; sa racine carrée $R = .506$ (coefficient de corrélation multiple) est le rapport de la longueur OM au rayon du cercle (comme on peut le vérifier graphiquement). Plus le point M est éloigné du point-origine O, meilleure est la qualité de l'ajustement par la régression multiple.

Propriété d'orthogonalité (cf. Figure 3 p. 24). Construisons l'axe $\mathcal{B}_\perp$ orthogonal à $\mathcal{A}$, d'où les points $\mathrm{B}_\perp$ et $\mathrm{M}_\perp$ projections orthogonales respectives des points B et M sur cet axe. La droite $\mathrm{MM}_\perp$ est parallèle à $\mathcal{A}$, donc le rapport $\overrightarrow{\mathrm{OM}'_b}/\overrightarrow{\mathrm{OB}}$ (coefficient de régression partielle) et le rapport $\overrightarrow{\mathrm{OM}_\perp}/\overrightarrow{\mathrm{OB}_\perp}$ sont égaux.

[22]Le cercle de rayon unité est le lieu des points dont la somme des produits des coordonnées obliques et droites est inférieur ou égal à 1.

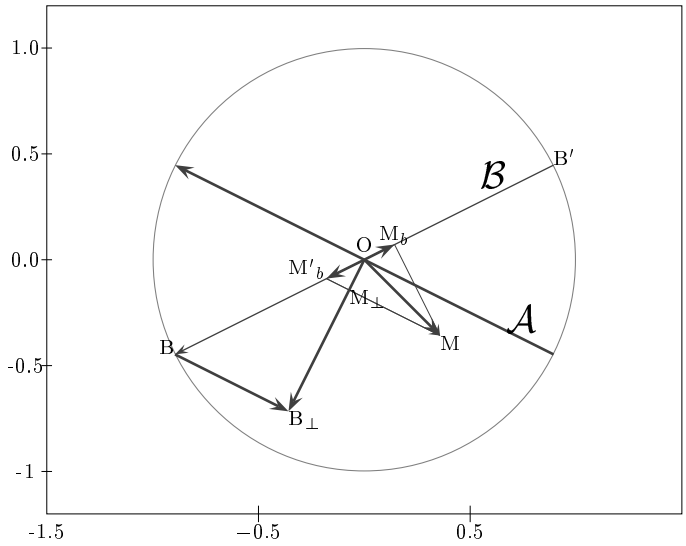

Figure 3. Situation Renversement : propriété d'orthogonalité.

3.2. La « rose des vents » des effets

En procédant de même pour les autres situations remarquables, on obtient leurs représentations géométriques, soit en termes du facteur B : Disparition de l'effet (Figure 4), Stabilité (Figure 5), Émergence (Figure 6, p. 25).

L'ensemble des points M possibles est l'intérieur du cercle des corrélations, et le rapport $\overrightarrow{\mathrm{OM}'_b}/\overrightarrow{\mathrm{OM}_b}$ = effet conditionnel/effet global est le même pour tous les points M d'un rayon du cercle, seule varie la qualité de l'ajustement. En parcourant (dans le sens des aiguilles d'une montre) le cercle à partir du cas–limite d'émergence (point E sur la droite $\mathcal{A}_\perp$) (Figure 7, p. 26), on obtient trois zones successives[23]. On a d'abord une zone d'*accentuation* (rapport des effets supérieur à 1) ; puis en passant par le cas intermédiaire de stabilité (rapport égal à 1 pour le point B′), on a une zone d'*atténuation* (rapport inférieur à 1 ; effets de même signe) ; puis, traversant le cas–limite de disparition (rapport nul pour le point A), on a une zone de *renversement* (rapport négatif). Et symétriquement de l'autre côté de la ligne $\mathcal{A}_\perp$. D'où la « rose des vents » des effets pour le facteur B.

En procédant de même pour le facteur A, on obtient la « double rose des vents » pour les deux effets A et B avec les divers cas (Figure 8, p. 26) : double accentuation ; double atténuation ; accentuation pour un facteur et renversement pour l'autre (zone paradoxale en gris).

3.2.1. Effet de structure et corrélation des variables de base

L'étude a été menée jusqu'ici pour la même valeur de Φ, à savoir 0.6. Or l'effet de structure est d'autant plus marqué que la liaison entre les facteurs A et B est forte. Les Figures 9 et 10 (p. 27) illustrent deux cas proches des cas extrêmes.

[23] On retrouvera ces zones en étudiant pour r (ou Φ) fixé (ici 0.6) la variation de la fonction $\beta_b/r_b = (1 - rr_a/r_b)/(1 - r^2)$ selon les valeurs du rapport r_a/r_b .

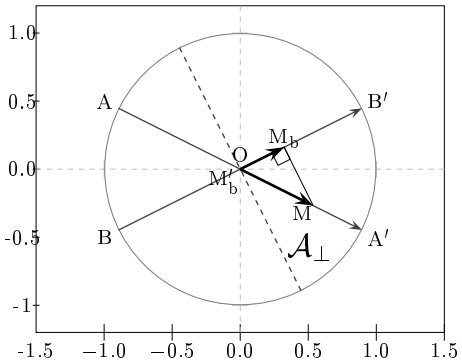

Figure 4. Situation Disparition. Effet $B = -0.36$ (point M_b) ; effet $B/A = 0$ (point $\mathrm{M}'_b = \mathrm{O}$)

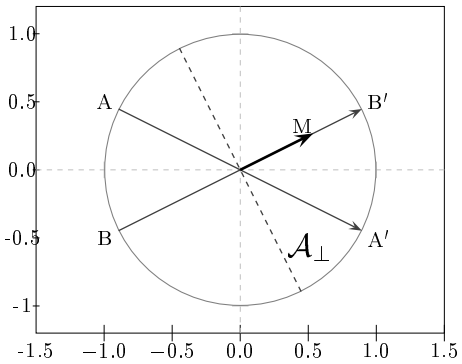

Figure 5. Situation Stabilité. Effet $B = -0.6 =$ effet B/A (point M).

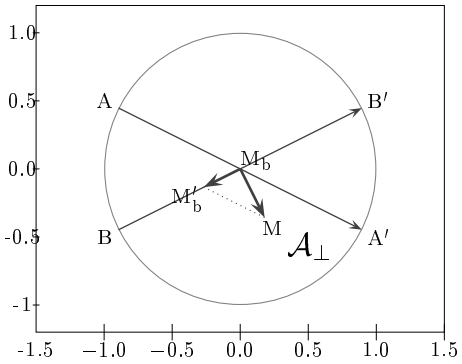

Figure 6. Situation Emergence. Effet $B = 0$ (point $\mathrm{M}_b = \mathrm{O}$) ; effet $B/A = -0.6$ (point M'_b).

– *Quasi-orthogonalité* (le « bon cas ») ; Φ est proche de 0, θ proche d'un angle droit : d'une part les deux zones de double accentuation et de double atténuation sont prédominantes, et la zone paradoxale réduite ; d'autre part l'accentuation et l'atténuation sont faibles.

– *Quasi-colinéarité* (le « mauvais cas ») ; Φ est proche de 1, θ proche de 0 : la zone de double accentuation et la zone de double atténuation sont réduites ; les zones paradoxales sont prédominantes.

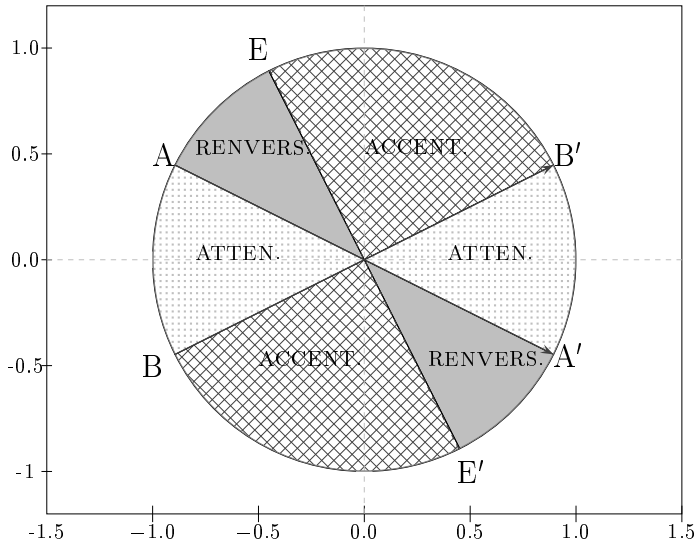

Figure 7. Rose des vents pour l'effet du facteur B

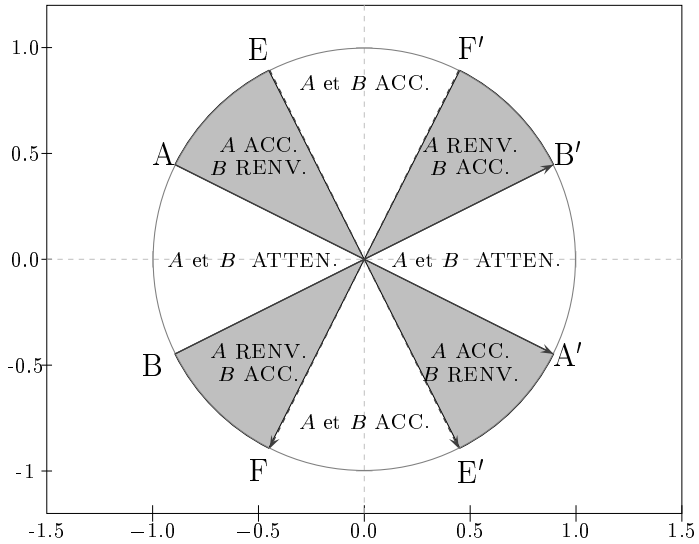

Figure 8. Rose des vents pour les effets de A et B

3.3. Régression et Analyse en Composantes Principales (ACP)

3.3.1. Analyse en Composantes principales

Sur les données « Lycées », on peut procéder à une ACP standard des deux variables prédictrices (indicatrices de Roméo et de Garçon) prises comme variables actives. Le premier axe ($\lambda_1 = 1.6$) oppose ab (« garçons dans lycée de garçons ») à $a'b'$ (« filles dans lycée de filles ») ; le deuxième axe ($\lambda_2 = 0.4$) oppose $a'b$ (« garçons dans lycée de filles ») à ab' (« filles dans lycée de garçons »). Cette interprétation apparaît à l'évidence sur le nuage des individus : cf. Figure 11, p. 28, à comparer avec la Figure 2, p. 23, (avec les variables principales en pointillés gris). La première variable principale est en somme une variable de « concordance » Sexe-Lycée, la deuxième une variable de « discordance » Sexe-Lycée.

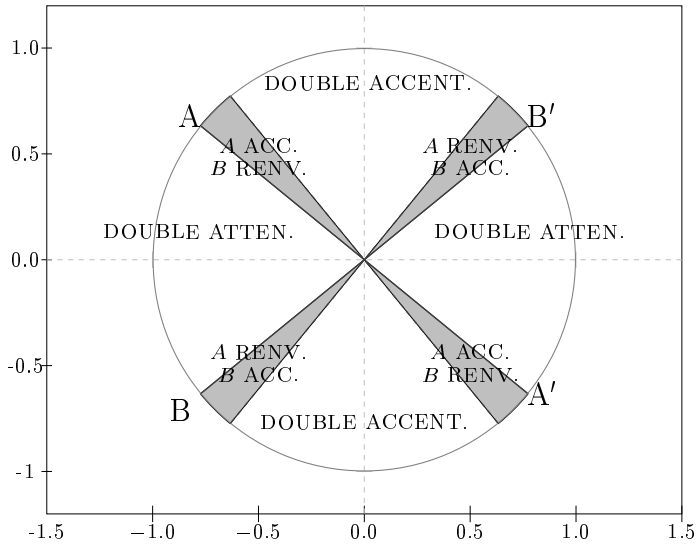

Figure 9. Quasi-orthogonalité

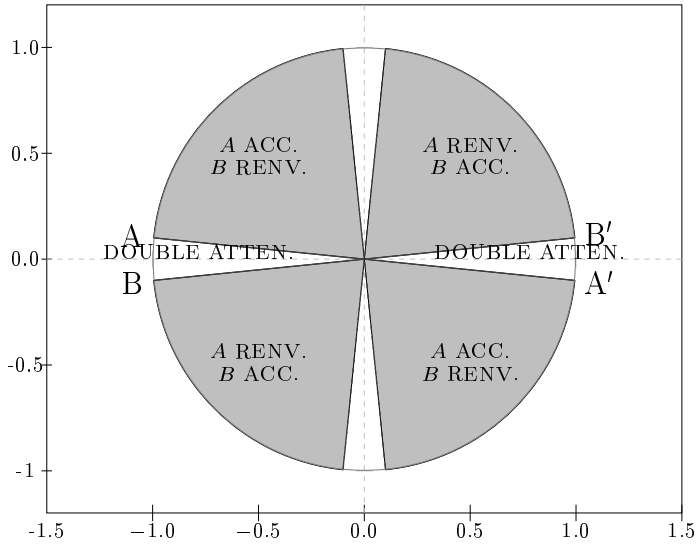

Figure 10. Quasi-colinéarité

3.3.2. Variables supplémentaires et régressions sur les variables principales

Si l'on met en variable supplémentaire la variable dépendante y (Réussite), les deux coordonnées principales de y sont (en ACP standard) les coefficients de corrélation de y avec chacune des variables principales.

Pour la situation de Renversement (cf. Figure 2, p. 23), les coordonnées principales du point M (variable y réduite) sont $(+0.358; -0.358)$. La qualité de représentation de la variable y est $R^2 = .256$, carré de la corrélation multiple $R = .506$ (longueur du segment OM comme dit plus haut). Les coordonnées principales du point M sont aussi les coefficients de régression de la variable dépendante réduite y sur les variables principales réduites[24]. En termes d'effets, on peut donc inter-

[24]Le premier axe est donc un axe global de réussite opposant les filles (qui réussissent mieux

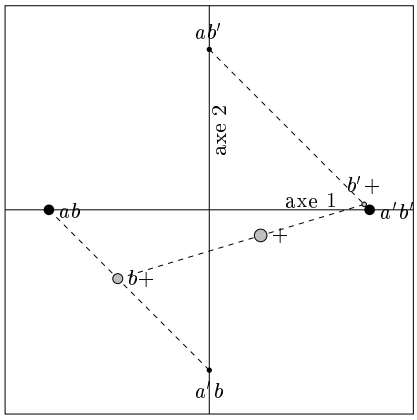

Figure 11. Situation Renversement. Nuage des individus : point moyen global de Succès (+) ; point moyen de Succès chez les garçons ($b+$) ; point moyen de Succès chez les filles ($b'+$).

préter les effets de chacune des variables principales sur la variable y comme suit : – effet de la première variable principale (concordance) : meilleures réussites des « filles dans lycée de filles » que des « garçons dans lycée de garçons » ; – effet de la deuxième variable principale (discordance) : meilleures réussites des « garçons dans lycée de filles » que des « filles dans lycée de garçons ».

Ces énoncés constituent la traduction du phénomène de renversement en termes d'analyse géométrique. Remarquons que puisque les variables principales sont non-corrélées (et partant non soumises à l'effet de structure), l'« effet »(tout court) d'une variable principale sur la variable dépendante est aussi bien l'effet global que l'effet conditionnellement aux autres variables principales (« effet vrai » ?)

3.3.3. Régressions sur les variables initiales

Dans le plan principal, on visualise les régressions simples ($\overrightarrow{\mathrm{OM}_a}$ et $\overrightarrow{\mathrm{OM}_b}$) et la régression multiple sur A et B ($\overrightarrow{\mathrm{OM}'_a}$ et $\overrightarrow{\mathrm{OM}'_b}$), comme dit plus haut (cf. Figure 2, p. 23) ; d'où la comparaison des effets global et conditionnel (rapports de vecteurs), conduisant aux conclusions d'accentuation pour A, de renversement pour B.

3.4. Exemple « Ouvrier »

3.4.1. Les données

Comme première application à des données réelles (très simplifiées), nous prendrons l'*exemple « Ouvrier »*, emprunté à la recherche classique de Michelat & Simon (1971). Nous prendrons comme *variables prédictrices* dichotomisées la *Profession A* (a Ouvrier O, a' Non–Ouvrier NO) et l'*Origine B* (b Père Ouvrier PO, b' Père Non–Ouvrier PNO), et comme *variable à prédire C* dichotomisée le sentiment d'*Appartenance* à la classe ouvrière (c *Oui*, c' *Non*) ; les *effectifs de base* (reconstitués d'après Grémy, 1979) sont donnés par le Tableau 7.

que les garçons, avec un taux de réussite de 58%) et le lycée Juliette (qui est meilleur que le lycée Roméo, avec un taux de réussite de 74%) d'une part, aux garçons et au lycée Roméo d'autre part.

				Appartenance *Oui*	*Non*	
O	*PO*	a	b	84	74	158
	PNO	a	b'	51	58	109
NO	*PO*	a'	b	33	84	117
	PNO	a'	b'	45	281	326
				213	497	710

Tableau 7. Exemple « Ouvrier » : données de base.

3.4.2. Effets globaux et conditionnels

Comme dans l'exemple des Lycées, à partir des données de base on dérive (voir Tableau 8) : d'une part les effectifs du croisement (non orthogonal) Profession×Origine, avec le coefficient $\Phi = .3258$; d'autre part, les fréquences des *Oui* conditionnellement aux quatre catégories définies par le croisement $A \times B$ (Exemple, pour ab : $84/158 = .5316$), ainsi que les fréquences selon les catégories A et B (Exemple, pour a : $\frac{84+51}{158+109} = \frac{135}{267} = .5056$) ; enfin la fréquence générale (moyenne pondérée des fréquences) $f_c = 213/710 = 0.30$. On en déduit les *effets globaux et conditionnels* pour chacun des deux facteurs A (Profession) et B (Origine).

			Origine B *PO* b	*PNO* b'	
Profession A	*O*	a	158	109	267
	NO	a'	117	326	443
			275	435	710

$f_a = .3761$
$f_b = .3873$

	Oui
ab	.5316
ab'	.4679
$a'b$	.2821
$a'b'$	.1380
Moy	.3000

	Oui
a	.5056
a'	.1761
b	.4255
b'	.2207

$r_a = .3483$
$r_b = .2177$

Tableau 8. Effectifs du croisement Profession×Origine ; fréquences de a et b, fréquences du Sentiment d'appartenance et coefficients de corrélations de *Appartenance* avec *Profession* et *Origine*.

• *Effets liés à la Profession*
Effet global de la profession : $A = .5056 - .1761 = +0.3295$.
Effet de la profession pour l'origine PO : $A/b = .5316 - .2821 = +0.2495$.
Effet de la profession pour PNO : $A/b' = .4679 - .1380 = +0.3299$.

Les effets A/b et A/b' sont tous deux positifs, mais différents[25]. On définira l'effet de la Profession conditionnellement à l'Origine comme la moyenne des deux effets, pondérée selon la pondération harmonique[26]. On obtient ainsi $A/B = +0.2936$. D'où le rapport des effets Conditionnel/Global : $\frac{A/B}{A} = +.2936/.3295 = +0.89$.
Conclusion : *faible atténuation.*

• *Effets liés à l'Origine*
Effet global de l'origine : $B = .4255 - .2207 = +0.2048$.
Effet de l'origine pour O : Effet $B/a = .5316 - .4679 = +0.0637$.
Effet de l'origine pour NO : Effet $B/a' = .2821 - .1380 = +0.1441$.

[25]La différence des deux effets définit l'*effet d'interaction* entre A et B : $.2495 - .3299 = -.0804$.
[26]Pondération harmonique : les poids de A/b et A/b' sont proportionnels à $\frac{1}{\frac{1}{158}+\frac{1}{117}} = 67.22$ et $\frac{1}{\frac{1}{109}+\frac{1}{326}} = 81.69$. Avec les poids élémentaires $158 + 117 = 275$ et $109 + 326 = 435$, on trouverait 0.2988 au lieu de 0.2936. De même pour le calcul de B/A.

La moyenne pondérée selon la pondération harmonique définit l'effet de l'Origine conditionnellement à la Profession $B/A = +0.1097$. D'où le rapport Conditionnel/Global : $\frac{B/A}{B} = .1097/.2048 = +0.54$.
Conclusion : *forte atténuation.*

3.4.3. Régressions

Ici encore, on retrouve les effets globaux et conditionnels en procédant aux régressions sur les variables indicatrices.

Régressions simples $\widetilde{y}_a = 0.3295\,\delta_a + 0.1761$; $\widetilde{y}_b = 0.2048\,\delta_b + 0.2207$.
Régression multiple $\widetilde{y}_{a+b} = 0.2936\,\delta_a + 0.1097\,\delta_b + 0.138$
Coefficient $R^2 = .1335$ (voir §2.7., p.21) ; d'où $R = .365$.

3.4.4. Représentations géométriques

La représentation des variables initiales (indicatrices) et celle des variables réduites sont ici distinctes ; nous détaillerons la première, afin de souligner les invariants affins. Dans l'une et l'autre représentations, on construit les axes $\mathcal{A}$ (Profession) et $\mathcal{B}$ (Origine) d'angle θ tel que $\cos\theta = +.3258 = \Phi$, soit $\theta = 71°$: voir Figure 12.

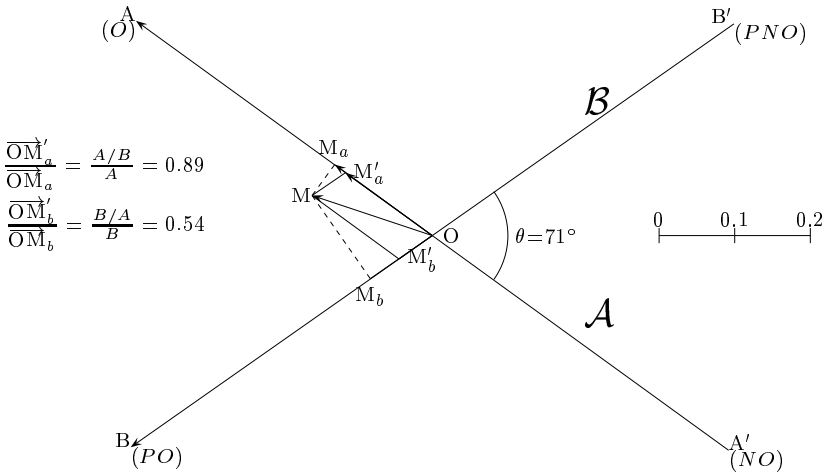

Figure 12. Exemple « Ouvrier », modèle symétrique. Régression simple sur A : M_a (+0.3295) ; Régression simple sur B : M_b (+.2048). Régressions multiples M'_a (+.2936), M'_b (+.1097). Rapport des Effets Conditionnel/Global : $\frac{A/B}{A} = 0.89$, $\frac{B/A}{B} = 0.54$.

Pour représenter les *variables indicatrices*, on construit les points A (O) et A′ (NO) sur l'axe $\mathcal{A}$, avec $\mathrm{OA} = \mathrm{OA}' = \sigma_a$ (avec $\sigma_a = \frac{\sqrt{267\times443}}{710} = .4844$), et les points B (PO) et B′ (PNO) sur l'axe $\mathcal{B}$, avec $\mathrm{OB} = \mathrm{OB}' = \sigma_b$ (avec $\sigma_b = \frac{\sqrt{275\times435}}{710} = .4871$). Dans la base $(\overrightarrow{\mathrm{OA}}, \overrightarrow{\mathrm{OB}})$, le point M (Appartenance) a pour coordonnées orthogonales les coefficients de régression simples $(0.3295, 0.2048)$, d'où les points M_a et M_b, et pour coordonnées obliques les coefficients de régression partiels $(0.2936, 0.1097)$, d'où les points M'_a, M'_b, avec la décomposition vectorielle $\overrightarrow{\mathrm{OM}} = \overrightarrow{\mathrm{OM}'_a} + \overrightarrow{\mathrm{OM}'_b}$. Les rapports

des effets (invariants affins) $\frac{A/B}{A}$ et $\frac{B/A}{B}$ sont représentés par les rapports vectoriels respectifs $\overrightarrow{\mathrm{OM}'_a}/\overrightarrow{\mathrm{OM}_a} = 0.89$ et $\overrightarrow{\mathrm{OM}'_b}/\overrightarrow{\mathrm{OM}_b} = 0.54$.

Pour représenter les *Variables réduites*, on construirait le point ayant pour coordonnées orthogonales les coefficients de corrélation $r_a = 0.3483$ et $r_b = 0.2177$, et pour coordonnées obliques les coefficients de régression partiels réduits (les « beta–weights ») 0.3103 $(= 0.2936\sqrt{\frac{.3761(1-.3761)}{0.3(1-0.3)}})$ et 0.1166 $(= 0.1097\sqrt{\frac{.3873(1-.3873)}{0.3(1-0.3)}})$; les rapports vectoriels représentent encore les rapports Conditionnel/Global $\frac{A/B}{A}$ et $\frac{B/A}{B}$.

3.4.5. Modèle symétrique et modèle hiérarchique

D'un simple point de vue prédictif, rien n'empêche de résumer les analyses de régression effectuées par les deux effets conditionnels (avec la « double atténuation des effets »), l'effet conditionnel de la Profession étant en gros le triple de l'effet de l'Origine. Mais ce résumé, basé sur un *modèle symétrique* faisant jouer des rôles symétriques aux deux variables Profession et Origine, ne renvoie à aucun schéma explicatif ; il serait peu sensé de dire que l'« effet vrai » de la profession est 0.29 et celui de l'origine 0.11 « toutes choses égales par ailleurs ».

Un autre résumé des analyses, plus proche d'un schéma explicatif, est celui du *modèle hiérarchique*,qui consiste à prendre les deux effets suivants :

- *Effet global* de la Profession : 0.3295 (comme précédemment).
- *Effet résiduel* de l'Origine par rapport à la Profession : cet effet est égal à l'effet conditionnel $B/A = 0.1097$ du modèle symétrique ; mais conceptuellement, c'est l'effet de la variable résiduelle de la régression de l'Origine sur la Profession.

En effet, si l'on régresse δ_b sur δ_a, et qu'on considère la variable résiduelle $\delta_b^\perp$, la variable $\widetilde{y}_{a+b}$ peut s'exprimer comme la régressée sur $\delta_a + \delta_b^\perp$:

$$\widetilde{y}_{a+b} = 0.3295\,\delta_a + 0.1097\,\delta_b^\perp + 0.138$$

(cf. propriété d'orthogonalité : le coefficient de δ_a est le coefficient de régression simple, celui de $\delta_b^\perp$ est le coefficient de régression partiel). Le modèle hiérarchique peut également être représenté géométriquement, en construisant le vecteur correspondant à la variable résiduelle $\delta_b^\perp$; la projection M'_b sur $\mathcal{B}$ parallèlement à $\mathcal{A}$ est remplacée par la projection orthogonale sur $\mathcal{B}^\perp$ toujours parallèlement à $\mathcal{A}$ (cf. Figure 13). Au lieu de la décomposition vectorielle oblique $\overrightarrow{\mathrm{OM}} = \overrightarrow{\mathrm{OM}'_a} + \overrightarrow{\mathrm{OM}'_b}$ (Figure 12), on a la décomposition vectorielle orthogonale $\overrightarrow{\mathrm{OM}} = \overrightarrow{\mathrm{OM}_a} + \overrightarrow{\mathrm{OM}_b}^\perp$ (Figure 13).

3.5. Méthode générale proposée

3.5.1. Régression complète et ACP

Soit, dans un modèle–cadre de régression, un ensemble de variables numériques (en nombre quelconque) susceptibles de servir de variables indépendantes, sur lesquelles on procède à une ACP, par exemple une ACP standard[27]. Mettre une variable dépendante y en variable supplémentaire, c'est procéder à la régression complète de y

[27] L'ACP standard (ou des corrélations) va avec la représentation des variables réduites, l'ACP simple (ou des covariances), avec celle des variables quelconques.

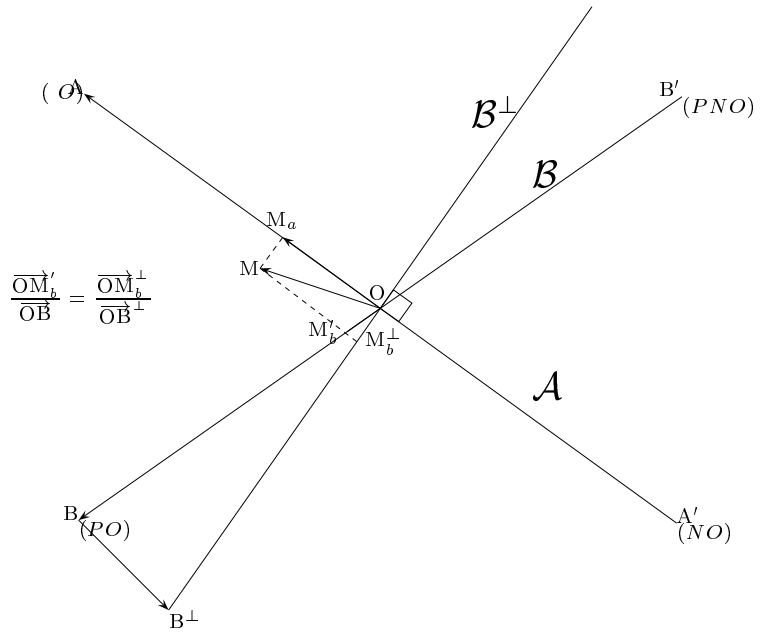

Figure 13. Exemple « Ouvrier », modèle hiérarchique. Régression simple sur A : M_a ; régression résiduelle par rapport à A : $M_b^{\perp}$.

sur l'espace des variables indépendantes (variables actives de l'ACP). La qualité de représentation de la variable y sur les variables actives n'est autre que le carré du coefficient de corrélation multiple (R^2) de la régression de y sur toutes les variables indépendantes. En considérant un sous-espace principal (par exemple un plan principal), on obtiendra une représentation principale de cette régression complète. En outre, chaque coordonnée principale de y peut s'interpréter comme l'effet de la variable principale correspondante sur la variable dépendante y : effet global mais aussi bien effet conditionnellement aux autres variables principales[28].

3.5.2. Régressions sur les variables initiales

Régressions simples. On peut représenter la régression d'une variable dépendante sur une des variables indépendantes initiales, dans un plan principal approprié ; tout en notant que, du fait que la projection principale ne conserve pas l'orthogonalité (dans le cas de plus de deux variables indépendantes), la régression simple ne se lira pas, en général, comme une projection orthogonale.

Régression multiple. La représentation géométrique se généralise à toute régression d'une variable dépendante sur une partie des variables indépendantes. En construisant les variables régressées dans un plan principal approprié, on pourra comparer effets conditionnels et effets globaux, c'est-à-dire étudier géométriquement les effets

[28]Dans le rude langage évoqué plus haut, on pourrait donc dire que l'effet d'une variable principale est un « effet vrai, toutes choses égales par ailleurs ». Cette remarque ne devrait-elle pas interpeller ceux qui voient dans la régression le moyen d'atteindre les « effets vrais », et leur suggérer d'étendre le privilège des « effets vrais » aux variables principales issues d'une analyse géométrique ?

de structure. Toute projection conservant les rapports, chaque rapport effet conditionnel /effet global s'interprète directement en projection.

3.5.3. Régression sur variables résiduelles

Dans la régression multiple d'une variable dépendante y sur un ensemble de variables indépendantes, on peut distinguer une variable indépendante particulière x_0 avec son coefficient de régression partiel u_{x_0} ; par ailleurs, on peut régresser la variable x_0 sur les autres variables indépendantes, d'où la variable résiduelle x'' de cette régression. Si on régresse la variable dépendante y sur cette variable résiduelle, le coefficient de régression (simple) de y sur x'' est égal à u_{x_0} (cf. propriété d'orthogonalité). Les régressions sur les variables résiduelles peuvent, elles aussi, être représentées géométriquement dans les sous-espaces principaux.

4. DOSSIER BISCUITS

Dans cette partie, nous appliquons la méthode proposée aux données d'une recherche dans le domaine de la nutrition, portant sur la *perception du caractère sucré d'aliments*[29].

4.1. Les données

Pour 39 biscuits du commerce, on dispose, pour chaque biscuit, de sa teneur (pourcentage en poids) en Sucre, Gras, Eau, Amidon, Protéine et Fibres, ainsi que de l'évaluation du Sucre, par 102 sujets, sur une échelle de 1 (très peu sucré) à 9 (très sucré). Les données de base du Dossier Biscuits se présentent ainsi sous la forme d'un tableau 39×7 (Tableau 9, p. 34) ; les 39 biscuits sont décrits selon le *Sucre perçu* (moyenne des évaluations des sujets)[30], et selon les 6 teneurs. Nous prendrons le Sucre perçu comme variable dépendante, et les 6 variables de teneur comme variables indépendantes. Le Tableau 10 (p. 34) donne, pour les 7 variables, les coefficients de corrélation deux à deux.

4.2. Analyse en composantes principales

Nous avons procédé à une ACP standard (ACP des corrélations) en prenant les six teneurs comme variables actives, avec la variable Sucre perçu réduite (SucP) en variable supplémentaire (on dit aussi illustrative). Trois valeurs propres sont supérieures à 1 : $\lambda_1 = 2.488$, $\lambda_2 = 1.397$ et $\lambda_3 = 1.102$. Les coordonnées principales sont données dans le Tableau 11, p. 35.

La Figure 14 (p. 35) représente les six variables actives et la variable supplémentaire, avec le cercle des corrélations. La première variable principale est corrélée avec

[29] Abdallah, Chabert, Le Roux, Louis-Sylvestre [1998]. Nous renvoyons à cet article pour une discussion sur le schéma physiologique explicatif de la perception du sucre, et le rôle de la teneur en gras et en eau dans cette perception.

[30] Sans chercher ici à discuter du bien-fondé de la procédure de moyennage entre sujets, disons simplement que tous les sujets ont utilisé l'étendue de l'échelle proposée de 1 à 9.

$n°$	Sucre perçu	Suc	Gra	Eau	Ami	Pro	Fib	Valeurs prédites SP1	SP2	SP3
4	7.75	51.5	2.1	16.2	23.2	3.7	3.3	6.8	6.1	6.3
22	7.17	50.8	4.4	20.2	13.1	4.7	6.8	6.7	6.2	6.7
7	7.05	44.8	29.3	1.4	11.3	5.4	7.8	6.3	7.2	7.1
35	6.91	40.7	31.4	.5	13.9	7.5	6.0	5.9	7.0	6.9
25	6.85	30.9	20.0	14.2	18.3	4.2	12.4	5.1	5.3	5.8
1	6.84	56.5	2.6	15.0	18.6	4.5	2.8	7.2	6.6	6.8
34	6.81	49.6	8.1	13.7	13.7	4.5	10.4	6.6	6.3	6.5
17	6.70	46.0	27.8	4.0	10.0	6.7	5.5	6.4	7.2	7.3
15	6.62	41.7	16.3	5.6	21.0	5.4	10.0	6.0	6.1	6.0
39	6.54	39.0	34.0	1.5	19.0	4.5	2.0	5.8	7.0	7.0
8	6.52	43.0	13.2	14.5	17.5	4.1	7.7	6.1	6.0	6.4
3	6.51	41.3	17.0	16.6	15.8	3.6	5.7	6.0	6.1	6.6
18	6.46	47.1	12.9	10.8	19.7	3.7	5.8	6.4	6.4	6.5
12	6.46	54.0	7.0	9.4	23.1	4.0	2.5	7.0	6.7	6.6
32	6.44	27.2	19.3	2.0	41.9	4.7	4.9	4.8	4.9	4.7
33	6.43	37.5	26.4	2.3	20.9	6.7	6.2	5.7	6.3	6.3
26	6.22	56.0	1.1	8.2	21.2	5.8	7.7	7.2	6.5	6.3
5	6.05	58.4	4.0	3.5	24.5	7.5	2.1	7.3	6.9	6.5
2	5.99	40.0	20.0	.9	33.0	4.7	1.4	5.9	6.2	5.9
16	5.95	37.0	17.0	14.0	17.1	7.0	7.9	5.6	5.7	6.1
27	5.93	43.6	2.4	18.1	24.0	2.2	9.7	6.2	5.4	5.7
30	5.93	23.5	22.9	2.5	35.2	6.6	9.3	4.6	4.8	4.7
21	5.85	29.2	25.4	2.0	27.1	6.7	9.6	5.0	5.5	5.4
14	5.82	27.5	25.0	1.8	32.5	6.5	6.7	4.9	5.3	5.2
37	5.78	30.7	37.7	1.5	23.1	4.4	2.6	5.1	6.4	6.5
36	5.71	35.8	20.7	1.2	31.6	6.0	4.7	5.5	5.8	5.6
20	5.61	25.0	18.0	1.0	44.3	8.0	3.7	4.7	4.6	4.4
19	5.18	44.1	8.3	2.0	36.6	6.3	2.7	6.2	5.8	5.4
11	4.88	57.5	6.0	7.7	17.8	8.5	2.5	7.3	6.9	6.8
13	4.77	31.0	16.6	2.5	36.4	6.8	6.7	5.2	5.1	4.9
44	4.77	43.0	17.7	1.8	24.5	6.4	6.6	6.1	6.3	6.1
40	4.68	26.9	18.7	4.6	36.0	6.0	7.8	4.8	4.9	4.8
29	4.67	21.5	28.5	1.4	39.4	5.5	3.7	4.4	5.0	4.9
45	3.99	19.2	11.3	1.9	53.1	8.5	6.0	4.2	3.7	3.4
31	3.93	27.4	11.9	1.4	45.3	7.5	6.5	4.9	4.5	4.1
28	2.74	7.2	13.3	29.0	37.9	8.9	3.7	3.2	2.7	3.8
41	2.45	23.4	4.5	5.5	50.4	10.1	6.1	4.5	3.6	3.4
10	1.70	9.1	3.0	3.6	67.7	10.4	6.2	3.4	2.2	1.8
9	1.31	3.1	18.7	3.5	61.5	8.2	5.0	2.9	2.6	2.6
Moy	5.59	36.5	16.0	6.9	28.7	6.1	5.9	5.59	5.59	5.59

Tableau 9. Biscuits. Données de base : Sucre perçu et 6 variables de composition (variables de base en pourcentages) ; les biscuits sont ordonnés selon les valeurs décroissantes du sucre perçu. Valeurs prédites du Sucre Perçu(SP) pour le Modèle I (SP1), le Modèle II (SP2) et le Modèle III (SP3).

	SucP	Suc	Gra	Eau	Ami	Pro	Fib
Sucre Perçu	1.000						
Sucre	.744	1.000					
Gras	.154	−.298	1.000				
Eau	.141	.186	−.495	1.000			
Amidon	−.843	−.789	−.151	−.324	1.000		
Protéines	−.731	−.528	−.058	−.311	.608	1.000	
Fibres	.116	−.110	.004	.165	−.150	−.108	1.000

Tableau 10. Biscuits. Corrélations entre les 6 variables de composition et Sucre perçu.

le Sucre ; le premier axe oppose des biscuits très sucrés (à gauche) à des biscuits peu sucrés (à droite). Le deuxième axe oppose des biscuits peu gras à forte teneur en eau à des biscuits très gras.

Le carré de la corrélation multiple (R^2) entre la variable supplémentaire SucP et les 6 variables principales est égal à .832 ; le carré de la corrélation multiple entre SucP et les deux premières variables principales vaut .823 (somme des carrés des corrélations simples : $(.807)^2+(.414)^2$) ; la racine carrée est .907 (longueur du vecteur Sucre perçu). Nous prendrons donc ce plan pour représenter géométriquement les régressions et étudier les effets de structure.

	y_1	y_2	y_3
Sucre	-.849	-.080	−.374
Gras	.235	−.896	.216
Eau	−.555	.630	.213
Amidon	.880	.347	−.047
Protéines	.779	.254	−.097
Fibres	−.143	.076	.927
SucP	−.807	−.414	.004

Tableau 11. Biscuits. Coefficients de régression des 6 variables réduites sur les variables principales réduites (cf. représentation principale dans le plan 1-2).

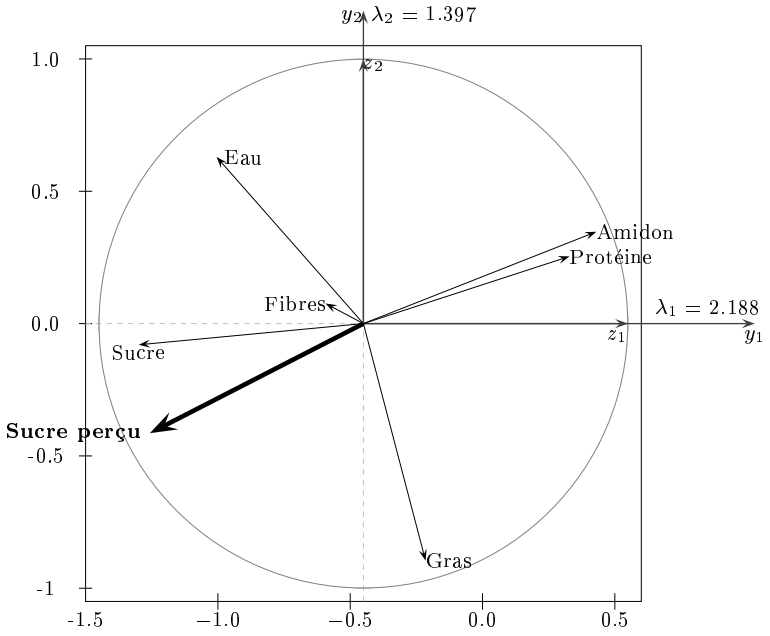

Figure 14. Biscuits. Plan principal 1-2 : cercle des corrélations de l'ACP standard des 6 variables de teneur, et Sucre Perçu en supplémentaire (vecteur en gras).

4.3. Régressions

Plusieurs régressions ont été effectuées sur les variables réduites. Les coefficients de régression se trouvent dans le Tableau 12 (p. 36).

4.3.1. Régression complète et ACP

Le carré de la corrélation multiple est $R^2 = .832$[31]. La valeur .832 fournit une limite supérieure pour la qualité des prédictions des diverses régressions ultérieures. Si l'on prend les 6 variables indépendantes, on a indétermination des 6 coefficients de régression, du fait que les 6 teneurs sont liées par une relation linéaire (colinéarité)[32].

[31] On retrouve la valeur obtenue dans l'ACP puisque mettre une variable en supplémentaire en prenant toutes les variables principales équivaut à une régression complète.

[32] La variable régressée sur les 6 variables indépendantes est parfaitement définie « en extension », c'est-à-dire par l'ensemble des valeurs prédites, mais elle est indéterminée « en compréhension », c'est-à-dire qu'on ne peut pas déterminer de manière unique les coefficients de régression, autrement dit « les parts qui reviennent » à chacune de ces variables indépendantes.

Si l'on écarte la variable Fibres, on n'a plus colinéarité (stricte), et on trouve les 5 coefficients reproduits dans le Tableau 12 (p. 36) c'est-à-dire la prédiction :

SP5 = −0.062 Sucre −0.150 Gras −0.295 Eau −0.763 Amidon −0.406 Protéines

À partir des coefficients de régression des variables sur la première variable principale, on retrouve la première coordonnée principale de SucP : $(-0.062) \times (-0.849) + \ldots + (-0.406) \times (+0.779) = -0.807$. (Et de même pour la deuxième variable principale)[33]. Nous ne tenterons pas d'interpréter les coefficients de la régression complète (cf. commentaire ultérieur sur la quasi–colinéarité) ; notons simplement la valeur infime du coefficient de la variable Sucre.

	Sucre	Gras	Eau	Amidon	Protéines	Fibres	R^2 coef.
Régression Complète	−0.062	−0.150	−0.295	−0.763	−0.406	//	**R^2 = .832**
Régressions simples	+0.744 [.**553**] Suc1	+0.154 [.024] Gra1	+0.141 [.020] Eau1	−0.843 [.711] Ami1	−0.738 [.543] Pro1	+0.116 [.013] Fib1	
Régression double Suc+Gra	+0.867 Suc2	+0.412 Gra2					**R^2 = .708**
Régression triple Suc+Gra+Eau	+0.856 Suc3	+0.530 Gra3	+0.244 Eau3				**R^2 = .753**

Tableau 12. Biscuits. Coefficients des diverses régressions. Pour les régressions simples, on a indiqué entre crochets les carrés des coefficients de corrélation. Exemple $(0.744)^2 = .553$ (qualité de représentation pour le Modèle I).

4.3.2. Régressions simples

La question est ici : « Dans quelle mesure peut-on prédire le Sucre perçu à partir de chacune des teneurs prises séparément ? » Les 6 coefficients de régression simples sont reproduits dans le Tableau 12. On a ainsi : Suc1= 0.744 Sucre ; Gra1 = +0.154 Gras, etc. Les 6 vecteurs correspondants ont été tracés sur la Figure 15 (p. 37).

Les régressions qui suivent visent à valider des schémas explicatifs du physiologiste, liés à l'existence de récepteurs sensoriels pour le sucre, le gras et l'eau.

4.3.3. Régression simple sur la teneur en Sucre (Modèle I)

Question : « Dans quelle mesure peut-on prédire le Sucre perçu à partir de la seule teneur en Sucre ? »

Cette régression, qu'on notera Suc1, n'est autre que la régression simple de SucP sur Sucre déjà effectuée (Figure 15) ; on a donc : Suc1= 0.744 Sucre. La qualité de l'ajustement de cette régression est $(0.744)^2 = .553$; valeur assez éloignée de $R^2 = .832$. La simple teneur en Sucre est insuffisante pour prédire le Sucre perçu ; ce qui motive les modèles qui suivent.

[33]En ajoutant vectoriellement ces composantes, on pourrait effectuer une construction géométrique de la régression complète, comme pour les régressions multiples présentées ci-après.

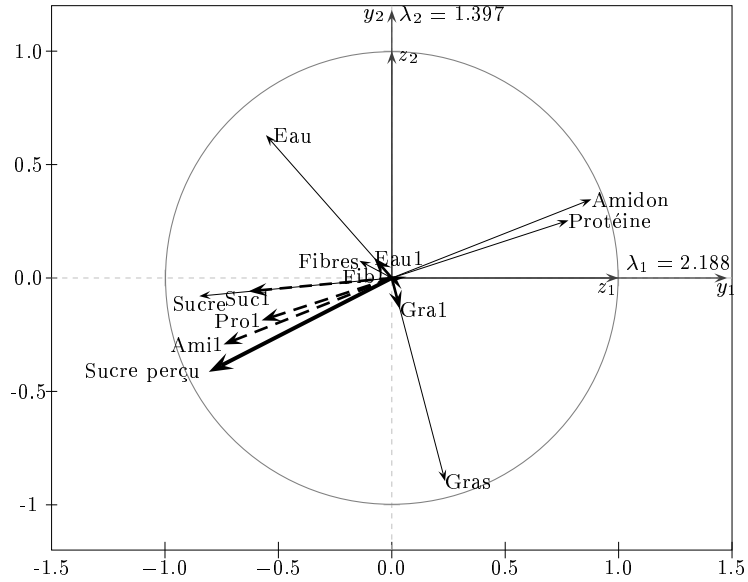

Figure 15. Plan principal 1-2. Régressions simples (en pointillés) du Sucre perçu sur chacune des 6 variables de teneur (réduites). Exemple : Suc1 = 0.744 Sucre, qualité de l'ajustement $(.744)^2 = .553$.

4.3.4. Régression double sur les teneurs en Sucre et en Gras (Modèle II)

Question : « Dans quelle mesure peut-on prédire le Sucre perçu à partir de la teneur en Sucre et de la teneur en Gras ? »

En notant cette régression double SP2, on trouve : SP2 = 0.867 Sucre + 0.412 Gras (=Suc2 + Gras2). D'où $R^2 = 0.867 \times 0.744 + 0.412 \times 0.154 = 0.708$ (somme des produits des coefficients de régression partiels par les corrélations). On est encore loin de .832. La représentation géométrique des modèles I et II est donnée d'une part dans le plan (Sucre, Gras) – avec conservation des angles droits par la régression simple, Figure 16 (p. 38) –, d'autre part en projection principale sur le plan 1-2 (Figure 17, p. 38).

Dans l'un et l'autre cas, on observe les rapports des effets conditionnel/global (0.867/0.744 pour Sucre, 0.412/0.154 pour Gras),qui sont des invariants affins (comme dit plus haut).

4.3.5. Régression triple sur les teneurs en Sucre, Gras et Eau

Question : « Dans quelle mesure peut-on prédire le Sucre perçu à partir des trois teneurs en Sucre, en Gras et en Eau ? »

On trouve : SP3 = 0.856 Sucre + 0.530 Gras + 0.244 Eau (= Suc3 + Gra3 + Eau3). D'où $R^2 = 0.753$. Voir Figure 18 (p. 39).

4.4. Effets de structure

4.4.1. Principaux résultats

La comparaison des coefficients de régression partiels avec les coefficients de régression simple illustre les effets de structure. Pour le Sucre, les coefficients passent de

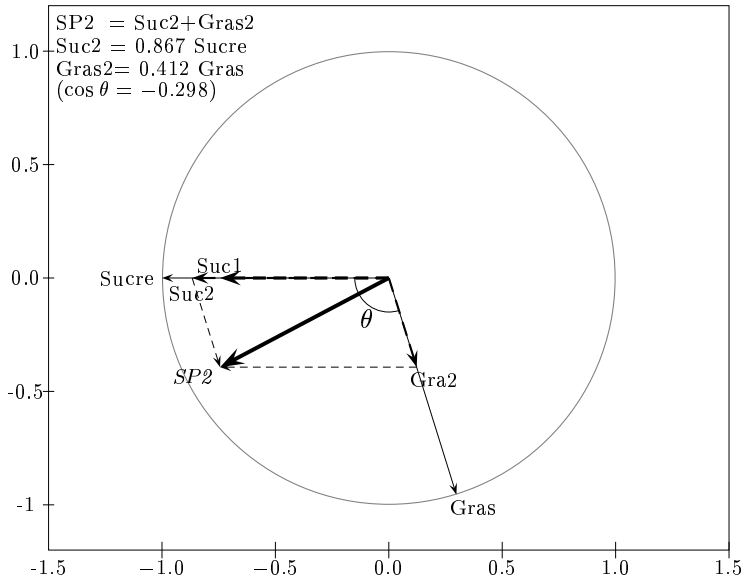

Figure 16. *Plan (Sucre, Gras).* Modèle I (Suc1= 0.744 Sucre) et Modèle II : Régression du Sucre Perçu sur Sucre + Gras (SP2 = 0.867 Sucre + 0.412 Gras, R^2=.708)

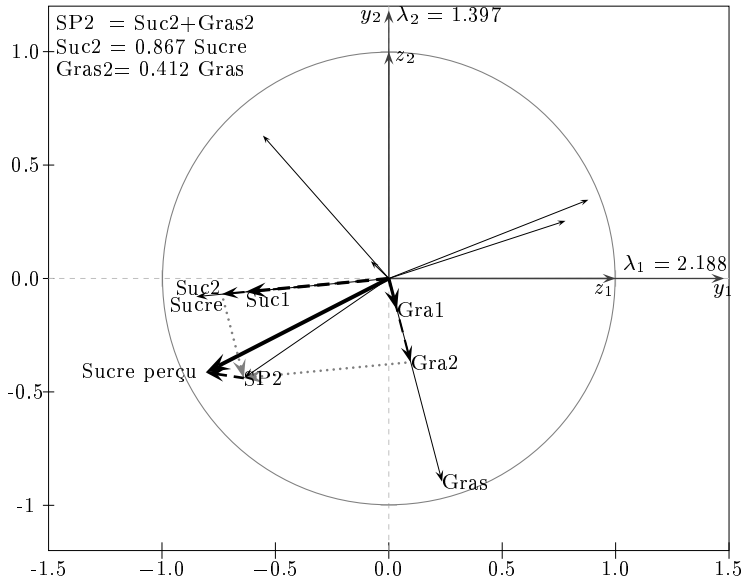

Figure 17. *Plan principal 1-2.* Modèle II : régression du Sucre Perçu sur Sucre + Gras (SP2 = 0.867 Sucre + 0.412 Gras ; R^2=.708).

0.744 (modèle I) à 0.867 (modèle II) puis 0.856 (modèle III), donc accentuation sensible puis à peu près stabilité (voir Tableau 12 p. 36). Pour la variable Gras, les coefficients sont : +0.154 (prédiction de Sucre perçu à partir de Gras seulement), +0.412 (Modèle II), +0.530 (Modèle III). On a donc deux accentuations successives. Pour la régression double, le rapport des effets conditionnel/global vaut 0.412/0.154 = 2.7 ; pour la régression triple, il vaut 0.530/0.154 = 3.4. On est proche de la

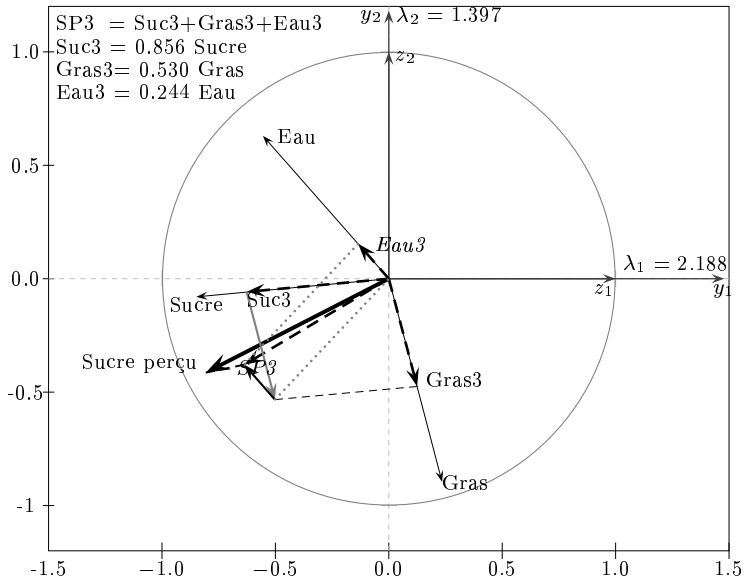

Figure 18. *Plan principal 1-2.* Modèle III : Régression du Sucre Perçu sur Sucre+Gras+Eau (SP3=0.856 Sucre + 0.530 Gras + 0.244 Eau ; $R^2 = .753$).

situation d'émergence[34].

4.5. Modèles hiérarchiques

4.5.1. Modèle IIbis

Pour prédire le Sucre perçu à partir de Sucre et Gras, on peut d'abord régresser la variable Gras sur la variable Sucre.

$$\widetilde{\text{Gras}} = -0.298\,\text{Sucre}$$

D'où la décomposition orthogonale de la variable Gras :

$$\text{Gras} = -0.298\,\text{Sucre} + \text{Gras}^{\perp}$$

où $\text{Gras}^{\perp}$ désigne la variable résiduelle de Gras par rapport à Sucre. En substituant cette expression de Gras dans la régression du modèle II, on obtient la régression de Sucre Perçu sur Sucre + $\text{Gras}^{\perp}$ (Modèle hiérarchique IIbis).

$$SP2 = 0.867\,\text{Sucre} + 0.412\,(-0.298\,\text{Sucre} + \text{Gras}^{\perp})$$

$$SP2 = 0.744\,\text{Sucre} + 0.412\,\text{Gras}^{\perp}$$

On retrouve le coefficient du Sucre du Modèle I (0.744) et le coefficient partiel de Gras du Modèle II (0.412).Géométriquement (Figure 19 p. 40), on construit le vecteur représentant la variable résiduelle $\text{Gras}^{\perp}$ puis les composantes Suc1 = 0.744 Sucre et 0.412 $\text{Gras}^{\perp}$. Le point SP2 est invariant, le vecteur correspondant est exprimé par rapport à la base orthogonale (Sucre, $\text{Gras}^{\perp}$).

[34]Rappelons que ces rapports sont des invariants affins, donc seraient les mêmes si les régressions étaient effectuées sur les variables initiales.

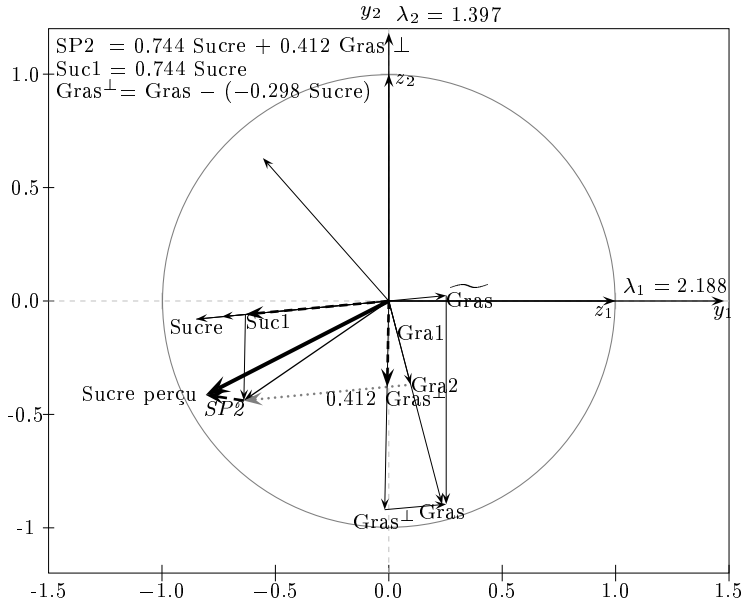

Figure 19. Modèle hiérarchique IIbis. Régression du Sucre Perçu sur Sucre + Gras ⊥ : 0.744 Sucre + 0.412 Gras ⊥. Coefficient R^2=.708.

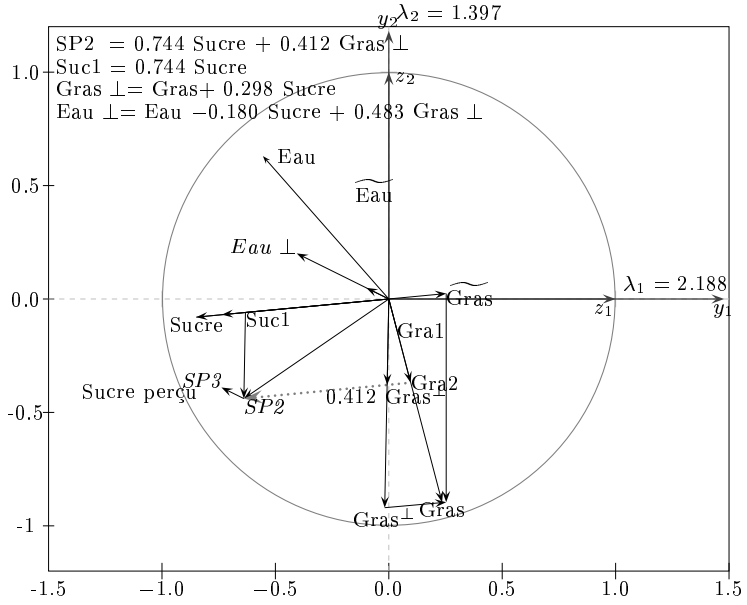

Figure 20. Modèle hiérarchique IIIbis. Régression du Sucre Perçu sur Sucre + Gras ⊥ + Eau ⊥ : 0.744 Sucre + 0.412 Gras ⊥ + 0.244 Eau ⊥. Coefficient R^2=.753.

4.5.2. Modèle IIIbis

Régressons maintenant la variable Eau sur Sucre + Gras :

$$\widetilde{\text{Eau}} = +0.042\,\text{Sucre} - 0.604\,\text{Gras} = +0.042\,\text{Sucre} - 0.604\,(-0.298\,\text{Sucre} + \text{Gras}^{\perp})$$

$$\widetilde{\text{Eau}} = +0.180\,\text{Sucre} - 0.483\,\text{Gras}^{\perp}$$

D'où la décomposition orthogonale de la variable Eau en $\widetilde{\text{Eau}}$ et $\text{Eau}^{\perp}$:

$$\textit{Eau} = +0.180\,\text{Sucre} - 0.483\,\text{Gras}^{\perp} + \text{Eau}^{\perp}$$

D'où en substituant, dans le modèle III :

$$\text{SP3} = 0.856\,\text{Sucre} + 0.530\,\text{Gras} + 0.244\,\text{Eau} = 0.744\,\text{Sucre} + 0.412\,\text{Gras}^{\perp} + 0.244\,\text{Eau}^{\perp}$$

On retrouve le coefficient du Sucre du Modèle I, le coefficient partiel de Gras du modèle II et le coefficient partiel de l'eau du modèle III. Voir Figure 20 (p. 40).

4.5.3. Commentaire sur la quasi–colinéarité

Le choix des trois modèles examinés, à partir du modèle–cadre des six variables de composition, a été guidé par le schéma explicatif du spécialiste concernant la perception du caractère sucré, schéma centré d'abord sur la teneur en sucre, puis sur le rôle du gras, enfin sur celui de l'eau. Les résultats corroborent la principale hypothèse physiologique : « Le gras renforce la perception du sucre ».

Faute d'un tel guide, on aurait pu, à partir des 6 variables de composition, procéder à toutes sortes de régressions, depuis les régressions simples jusqu'à la régression complète déjà commentée. On aurait alors trouvé que dans la plupart des régressions, c'est l'amidon qui joue un rôle prédominant, et en conséquence été tenté de conclure que c'est l'amidon qui régit la perception du sucre. En réalité la variable Amidon n'a aucune valeur explicative et se trouve, dans l'ensemble des biscuits étudiés, en quasi–colinéarité avec le Sucre. Ce qui appelle des commentaires analogues à ceux (déjà évoqués) de Malinvaud dans le contexte économique.

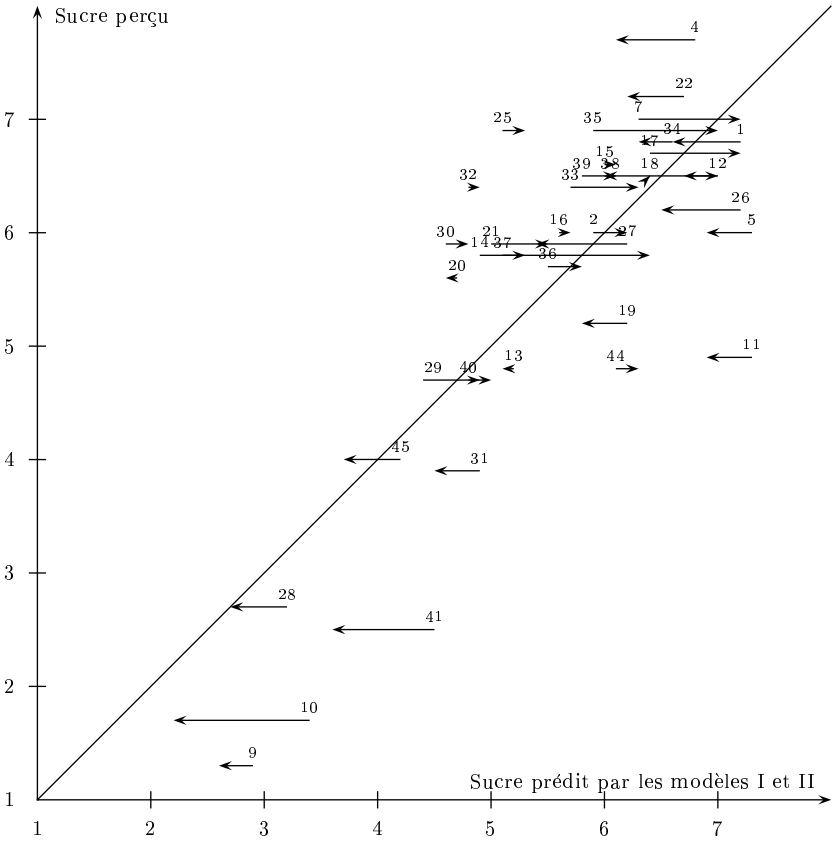

Figure 21. Régression des 39 biscuits (individus), modèles I et II ; les flèches vont du Modèle I vers le Modèle II. Exemple : Pour le biscuit 10, le sucre perçu est surestimé par les deux modèles (point 10 à droite de la première bissectrice) ; la surestimation est très forte pour le Modèle I.

4.5.4. Diagrammes individuels

Nous avons centré l'article sur les représentations dans l'espace des variables, avec une seule brève allusion à l'espace des individus (Figure 11, p. 28) ; mais une représentation géométrique ne saurait être complète sans l'examen des données individuelles, appelant tout naturellement pour l'interprétation le langage des « groupes d'individus » : voir notamment Rouanet, Ackermann, Le Roux [2000] et Chiche,

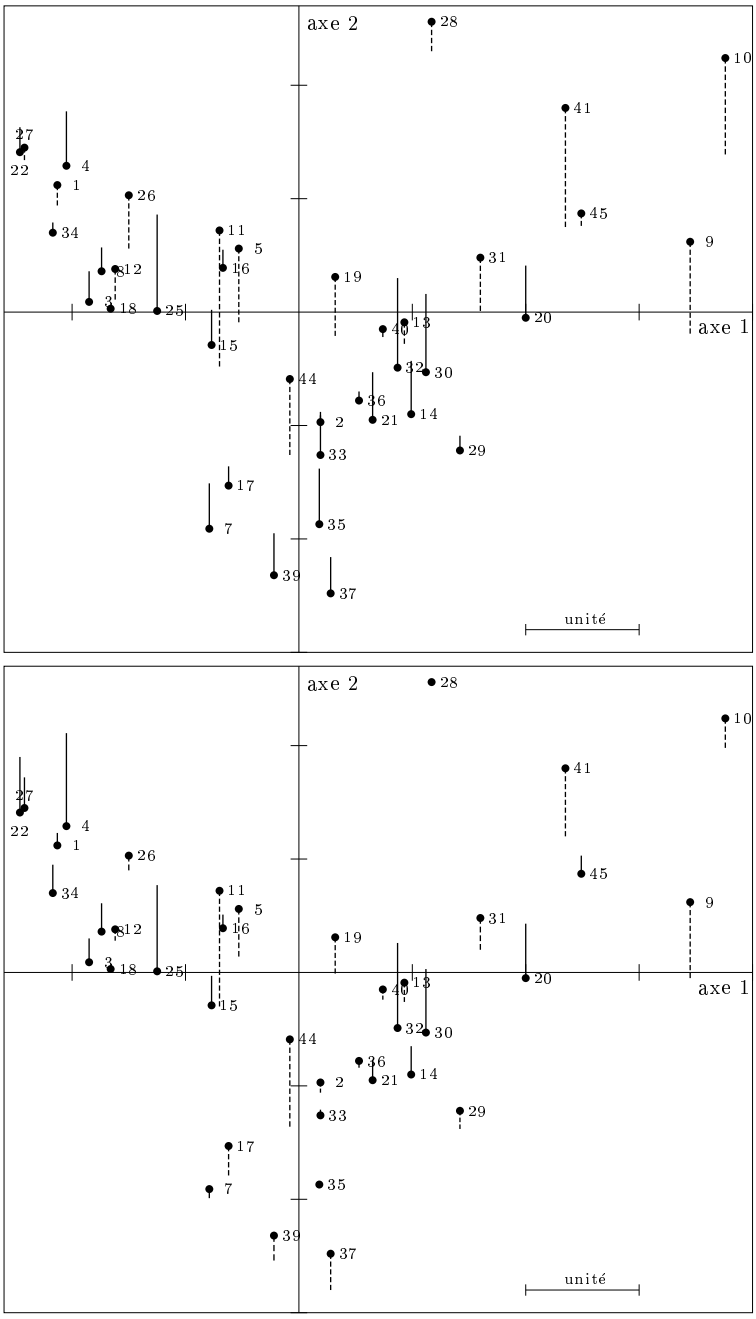

Figure 22. Représentation des 39 biscuits dans le plan principal 1-2 ; l'écart entre le sucre perçu et la valeur prédite par la régression (modèle I : Figure du haut, modèle II : Figure du bas) est figuré par un trait plein lorsque l'écart est positif et par un trait pointillé lorsque l'écart est négatif (la longueur de l'écart est égale à la moitié de celle de la Figure 21).

Le Roux, Perrineau, Rouanet [2000]. Cette remarque vaut tout autant pour la régression que pour l'ACP. Les Figures 21 et 22 visent à évoquer l'intérêt de telles représentations.

• *Régression.* Dans la Figure 21 (p. 41), on a représenté l'ensemble des 39 biscuits (individus) avec les prédictions individuelles du Modèle I et du Modèle II. Le graphique montre, pour chaque biscuit, l'amélioration apportée quand on passe du Modèle I au Modèle II.

•*ACP.* Dans la Figure 22 (p. 42), on a représenté le nuage des 39 biscuits dans le plan principal, en figurant pour chaque biscuit l'écart du Modèle I et celui du Modèle II. On constate notamment que pour les biscuits gras (partie inférieure du diagramme), le Modèle I tend à sous–estimer le Sucre perçu, alors que le Modèle II, globalement meilleur, tend à le surestimer.

5. CONCLUSIONS

Esquissons quelques conclusions qui se dégagent de cette étude.

1) Il est possible de conjuguer les procédures de régression avec les méthodes d'Analyse Géométrique des Données, en allant plus loin que la technique classique des éléments supplémentaires, et en représentant géométriquement les effets conditionnels aussi bien que les effets globaux des variables de la régression. L'utilisation conjointe de la régression linéaire et de l'ACP est directe, du fait de l'identité de structure (espace linéaire de variables). Plus généralement, on peut intégrer la régression linéaire sur variables indicatrices, et même la régression logistique, dans l'Analyse des Correspondances Multiples (ACM)[35]. Nous nous proposons de développer ces extensions dans un travail ultérieur.

2) Souvent sur un même ensemble de variables, on procède à diverses régressions (cf. dossier Biscuits). En construisant un modèle géométrique des données à partir d'un ensemble de variables assez ample pour couvrir les divers modèles de régressions, on pourra étudier géométriquement en projection principale chacune des régressions, ce qui permettra de s'assurer immédiatement, pour commencer, qu'il y a bien, pour chaque variable dépendante, une « variance à expliquer » (coefficient R^2 assez élevé)[36], et pour chaque modèle, d'examiner les effets de structure (effets conditionnels/globaux).

3) Les individus – à la différence des variables – sont porteurs de toute l'information. Une analyse géométrique n'est complète que si elle comporte le nuage des individus, de même qu'une analyse de régression n'est complète que si elle comporte l'examen des écarts individuels des valeurs obbservées aux valeurs prédites par le modèle. Dans l'analyse du dossier Biscuits, nous avons effectué des représentations qui intègrent les deux méthodes au niveau individuel.

[35] Le cas de deux variables dichotomiques (comme dans l'« exemple des lycées ») peut clairement se formuler aussi bien en termes d'ACM ou d'ACP.

[36] « If you want to explain something, you must have something to explain ! » (aphorisme attribué à J.W. Tukey).

4) Lorsqu'on adopte une rhétorique explicative en statistique (« variables explicatives », etc.), on voit mal pourquoi le privilège des « effets toutes choses égales par ailleurs » accordé aux variables d'une régression ne serait pas étendu aux variables principales d'une analyse géométrique. La régression n'a pas le monopole de l'« explication »[37] !

5) Le rappel « Corrélation n'est pas causalité » est une constante de la méthodologie statistique (voir par exemple [Boudon, 1967]). N'opposons pas des méthodes statistiques qui seraient (en quelque sorte par essence) « explicatives », à d'autres qui seraient vouées (toujours par essence) à être seulement « descriptives » ou « exploratoires ». Lorsque pour examiner un schéma explicatif (physique, biologique, sociologique...), on utilise des méthodes statistiques, le plus sage peut-être est de laisser la rhétorique explicative hors de la statistique, et de s'en tenir au principe : « La statistique n'explique rien – mais elle fournit des éléments potentiels d'explication » [Lebart & *al.*, 1995, p. 209].

Remerciements : Ce texte reprend un exposé fait en Mars 2002 au Séminaire organisé par Bruno Cautrès et Jean Chiche dans le cadre du DEA de sciences politiques dirigé par Richard Balme ; nous les remercions vivement. Nous remercions également Marc Barbut pour ses commentaires sur une première version du texte.

BIBLIOGRAPHIE

[1] ABDALLAH L., CHABERT M. , LE ROUX B. , LOUIS-SYLVESTRE J., « Is pleasantness of biscuits and cakes related to their actual or their perceived sugar and fat content ? », *Appetite*, 30, 1998, p. 309-324.

[2] BARBUT M., « Note sur quelques indicateurs globaux de l'inégalité : C.Gini, V. Pareto, P. Lévy », *Revue Française de Sociologie*, XXV, 4, 1984, p. 609-622.

[3] *Mathématiques et Sciences humaines*, « Sur la mesure et la comparaison des inégalités sociales », numéro spécial, 93, 1986, p. 5-69.

[4] BOUDON R., *L'analyse mathématique des faits sociaux*, Paris, Plon, 1967.

[5] CHICHE J., LE ROUX B., PERRINEAU P., ROUANET H., « L'espace politique des électeurs français à la fin des années 1990 », *Revue française de science politique*, Vol.50, 3, 2000, p. 463-487.

[6] COMBESSIE J.C., « L'évolution comparée des inégalités : problèmes statistiques », *Revue Française de Sociologie*, XXV, 2, 1984, p. 233-254.

[7] DESROSIÈRES A., « Un essai de mise en relation des histoires récentes de la statistique et de la sociologie », *Actes de la Journée d'études Sociologie et Statistique*, INSEE, Société française de sociologie, Paris, Octobre 1982.

[8] FAVERGE J.M., *Méthodes statistiques en psychologie appliquée*, Paris, Presses Universitaires de France, 1966.

[9] FISCHER C.S., HOUT M., SANCHEZ J.M., LUCAS S.R., SWIDLER A., VOS K., *Inequality by design : Cracking the Bell Curve Myth*, Princeton University Press, 1998.

[37] Dans la tradition psychométrique, c'est l'Analyse Factorielle qui est réputée « explicative », par opposition à la régression regardée comme « pragmatique ».

[10] FITOUSSI J.P. & al., *Réductions du chômage, les réussites en Europe*, Paris, La Documentation Française, 2000.
[11] GRÉMY J.P., *Introduction à la lecture des tableaux statistiques*, Paris, LEMTAS, 1979.
[12] GRÉMY J.P., « Sur les différences entre pourcentages et leur interprétation », *Revue Française de Sociologie*, XXV, 3, 1984, p. 396-420.
[13] KENDALL M., STUART A., *The Advanced Theory of Statistics*, London, Griffin, 1973.
[14] LEBART L., MORINEAU A., PIRON M., *Statistique exploratoire multidimensionnelle*, Paris, Dunod, 1995.
[15] MALINVAUD E., *Méthodes statistiques de l'économétrie*, Paris, Dunod, 1981.
[16] MERLLIÉ D., « Analyses de l'interaction entre variables. Problème statistique ou sociologique ? », *Revue Française de Sociologie*, XXVI, 2, 1985, p. 629-652.
[17] MICHELAT G., SIMON M., « Classe sociale objective, classe sociale subjective et comportement électoral », *Revue Française de Sociologie*, XII, 4, 1971.
[18] NOVI M., *Pourcentages et tableaux statistiques*, Paris, Presses Universitaires de France, 1998.
[19] PASSERON J.-C., *Le raisonnement sociologique. L'espace non-poppérien du raisonnement naturel*, Paris, Nathan, 1991.
[20] PRÉVOT J., « À propos d'indices et de comparaisons de proportions », *Revue Française de Sociologie*, XXVI, 4, 1985, p. 601-628.
[21] RIANDEY B., « L'utilisation de la régression logistique dans les enquêtes », *Bulletin de Méthodologie sociologique*, 33, 1991, p. 79-84.
[22] ROUANET H., « Barouf à Bombach », *Echo des Messaches*, 1978, p. 30.
[23] ROUANET H., LE ROUX B., *Analyse des données multidimensionnelles*, Paris, Dunod, 1993.
[24] ROUANET H., ACKERMANN W., LE ROUX B., « The geometric analysis of questionnaires : the lesson of Bourdieu's La Distinction », *Bulletin de Méthodologie Sociologique*, 65, 2000, p. 5-16.
[25] VALLET L.A., « L'évolution de l'inégalité des chances devant l'enseignement : un point de vue de modélisation statistique », *Revue Française de Sociologie*, XXIX, 3, 1988, p. 395-423.
[26] VALLET L.A., CAILLE J.P., « Les carrières scolaires au collège des élèves étrangers issus de l'immigration », *Éducation et Formations*, 40, 1995, p. 5-14.