

REVUE DE STATISTIQUE APPLIQUÉE

X. BRY

Extension de l'analyse en composantes thématiques univariée au modèle linéaire généralisé

Revue de statistique appliquée, tome 54, n° 3 (2006), p. 5-32

http://www.numdam.org/item?id=RSA_2006__54_3_5_0

© Société française de statistique, 2006, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

EXTENSION DE L'ANALYSE EN COMPOSANTES THÉMATIQUES UNIVARIÉE AU MODÈLE LINÉAIRE GÉNÉRALISÉ

X. BRY

*Département des Sciences Mathématiques, Université de Montpellier II
xbry@math.univ-montp2.fr*

RÉSUMÉ

Une variable endogène y est supposée dépendre de plusieurs variables explicatives latentes sous-jacentes à autant de groupes de variables exogènes. Cet article rappelle tout d'abord le principe de l'analyse en composantes thématiques (ACT) dans le cadre d'un modèle linéaire ordinaire de y . Un assouplissement de la méthode est proposé afin de pouvoir dépister d'éventuels faisceaux de variables exogènes. L'ACT est ensuite étendue à la modélisation linéaire généralisée. Cette extension est enfin appliquée à l'analyse de données biographiques.

Mots-clés : *Analyse en Composantes Thématiques, Modèle Linéaire Généralisé, Moindres Carrés Partiels, Régression PLS, Rotation quartimax.*

ABSTRACT

We consider an endogenous variable y depending upon several latent variables, each of them underlying in a different exogenous predictor group. This paper first recalls the principles of thematic components analysis (TCA) when y 's model is an ordinary linear one. We propose a way to make TCA more flexible, so as to be able to track down possible exogenous variable bundles. Then, TCA is extended to generalized linear modelling. This extension is applied to life history data analysis.

Keywords : *Generalized Linear Model, Partial Least Squares, PLS Regression, Quartimax rotation, Thematic Components Analysis.*

Sigles et acronymes :

ACP : Analyse en Composantes Principales

ACT : Analyse en Composantes Thématiques

MLG : Modèle Linéaire Généralisé

PLS : Partial Least Squares

Notations

X désigne indifféremment un groupe de J variables décrivant n individus et la matrice (n, J) dont ces variables sont les vecteurs colonnes.

x, y désignent indifféremment des variables numériques et les vecteurs de \mathbf{R}^n qui les codent.

$\langle x|y \rangle$ désigne le produit scalaire canonique des vecteurs x et y .

On désigne par $\langle X \rangle$ le sous-espace de \mathbf{R}^n engendré par les colonnes de X .

On notera $\Pi_{\langle X \rangle}$ le projecteur orthogonal sur le sous-espace $\langle X \rangle$.

N.B. Les variables numériques sont, sauf mention contraire, prises centrées réduites.

1. Introduction

On considère n individus décrits par un certain nombre de variables. Ces variables sont regroupées en *groupes thématiques*, ou plus simplement *thèmes*, selon leur appariement sur le plan conceptuel. Par exemple : si les individus sont des logements, le thème «coût» pourra comprendre le prix du logement et le loyer mensuel; le thème «taille», la surface habitable, la surface du terrain, le nombre de pièces d'habitation, etc.; le thème «qualité du quartier», la présence ou non de diverses commodités à proximité (axe routier, parc, commerces...).

Les différents thèmes présentent *a priori* des relations de dépendance causale dont procèdent des liaisons statistiques. Dans notre exemple, le coût du logement dépend (entre autres) de sa taille et de la qualité du quartier. L'ensemble des thèmes structuré par les liaisons entre eux sera appelé *modèle thématique*.

Dans le sillage de l'approche PLS [Tenenhaus 1999], nous avons proposé dans [Bry 2003] une méthode apparentée permettant l'exploration d'un modèle thématique, l'Analyse en Composantes Thématiques (ACT). L'ACT suppose que chaque thème est structuré autour d'un petit nombre de variables latentes qui le résumant efficacement, et que les variables latentes des différents thèmes entretiennent des relations statistiques correspondant aux relations de dépendance logique des thèmes. Les relations statistiques considérées par l'ACT sont celles d'un modèle linéaire ordinaire. L'ACT estime chaque variable latente par un *facteur*, ou *composante thématique* (on confondra par commodité la variable latente et le facteur qui l'estime, lorsque cela ne prête pas à conséquence). Dans l'exemple des logements, l'ACT permet ainsi dans un premier temps d'obtenir un facteur de coût linéairement modélisé en fonction d'un facteur de taille et d'un facteur de qualité du quartier. L'ACT peut dans un deuxième temps calculer, dans chaque thème que l'on souhaite explorer de façon multidimensionnelle, un second facteur décorrélé avec le premier, et intégrer ces facteurs dans le modèle statistique final. L'ACT a été étendue aux modèles comportant des effets d'interaction entre variables latentes explicatives [Bry 2004 a].

La technique d'estimation de l'ACT réalise un compromis entre deux contraintes n'allant pas *a priori* dans le même sens. La première est que chaque facteur doit être structurellement fort dans son thème, *i.e.* en capter une part non négligeable de la

variance totale. La seconde est que les facteurs doivent fournir une bonne qualité d'ajustement du modèle statistique qui les relie. Nous allons ici assouplir les critères utilisés par ces deux contraintes. L'assouplissement du critère de force structurelle des facteurs leur permettra de dépister des phénomènes plus locaux de corrélation entre variables dans un thème (typiquement, l'existence de faisceaux de variables très corrélées). Concernant la qualité d'ajustement du modèle, le critère utilisé par l'ACT correspond à la modélisation linéaire ordinaire : il s'agit du R^2 . Nous allons ici étendre de façon très simple et directe le mécanisme de cette méthode de sorte à traiter la modélisation linéaire généralisée, ce qui nous conduira à utiliser, lors de l'estimation, la maximisation d'une vraisemblance. Le cadre dans lequel nous procéderons à cette extension est volontairement simplifié afin de ne pas alourdir le formalisme : on considèrera un thème dépendant unique, ne contenant qu'une variable y . Ce modèle thématique sera dit *univarié*, au sens où il ne contient qu'une variable dépendante. Il s'agit de modéliser cette variable à partir de plusieurs thèmes explicatifs (cf. figure 1).

2. L'Analyse en Composantes Thématiques univariée (ACT1)

2.1. Le modèle

On considère le modèle illustré par la figure 1. Une variable numérique observée y dépend de déterminants latents F_1, \dots, F_R , chaque F_r étant sous-jacent à un groupe explicatif X_r de variables observées.

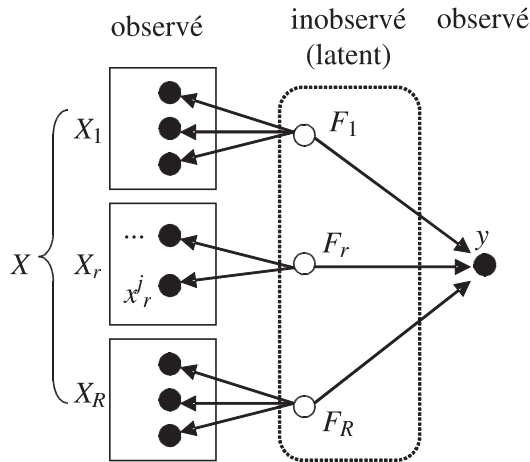


FIGURE 1

Modèle d'une variable observée en fonction de variables latentes

L'analyse en composantes thématiques considère qu'à une perturbation ϵ près, y dépend linéairement des F_r :

$$\sum_{r=1}^R b_r F_r + \epsilon, \text{ où } \forall r, F_r \perp \epsilon$$

Les variables latentes F_1, \dots, F_R sont supposées satisfaire au mieux une double contrainte :

- Chaque F_r est «structurellement forte» dans son groupe X_r , c'est-à-dire capte une part non résiduelle de la variance du groupe.
- L'ensemble des F_r fournit une «bonne» prédiction de y .

C'est cette double contrainte qui détermine la stratégie d'estimation.

Note 1 : On prendra soin d'examiner le sens des flèches sur la figure 1. Plusieurs flèches arrivant sur une même variable signifie que cette dernière est modélisable à partir d'une combinaison linéaire des variables se trouvant à l'origine des flèches. Dans ce modèle, seule y est conditionnée par plusieurs variables, les F_r . La liaison entre y et chaque F_r est donc *partielle*. Par contre, les variables explicatives x_r^j du groupe X_r sont déterminées par une seule et même variable latente, F_r . La liaison entre chaque x_r^j et F_r est donc *globale*.

Note 2 : En réalité, l'hypothèse d'unicité de la variable latente de chaque groupe sera relâchée dans un deuxième temps. Nous proposerons un schéma d'emboîtement approprié à l'exploration multidimensionnelle de chaque groupe explicatif dans le cadre de la modélisation de y .

2.2. Résultantes : objectif, définition et extensions

Nous avons étudié dans [Bry 2001] et [Bry 2004 a] une manière simple et souple de focaliser un facteur sur les structures «fortes» d'un groupe de variables : l'opérateur de résultante linéaire. Nous avons montré dans [Bry 2004 a] que ces structures fortes étaient celles que matérialisent les composantes principales du groupe. Nous allons ici assouplir cette notion de structure forte pour qu'elle puisse prendre en compte l'existence de faisceaux de variables, et étendre l'opérateur de résultante de sorte à permettre le dépistage de telles structures.

2.2.1. Résultante linéaire

1. Définitions

(a) Résultante linéaire :

- Soit X un groupe de J variables numériques et M une matrice symétrique positive de dimension J . Soit y une variable numérique. Nous avons défini dans [Bry 2001] la *résultante* de y sur X muni de la métrique M comme la variable : $R_{X,M}y = XM^Xy$. Nous avons montré que celle-ci était colinéaire au facteur $F = XMu$ solution du programme :

$$Max_{u'} Mu=1 < XMu|y >$$

Dans le cas $M = I$, elle correspond à la première composante de l'analyse inter-batteries de Tucker [Tucker 1958], qui est aussi la première composante de la régression PLS.

(b) Degré d'une résultante :

Nous avons introduit dans [Bry 2004 a] la *résultante de degré* α , α étant un réel ≥ 0 . Son opérateur est :

$$R_{X,M}^\alpha = (R_{X,M})^\alpha = (XMX')^\alpha$$

2. Comportement

– Si F^k est le k -ième facteur normé d'ACP de X avec la métrique M , et λ_k la valeur propre associée, on a :

$$R_{X,M}^\alpha y = (XMX')^\alpha y = \sum_k \lambda_k^\alpha \langle F_k | y \rangle F_k$$

On obtient donc une somme des composantes de y sur la base factorielle, chaque composante étant pondérée par l'inertie captée par le facteur, i.e. la force structurale de celui-ci, mise à la puissance α .

En conséquence, y est rapprochée des composantes principales de X en proportion de la puissance α de l'inertie qu'elles captent, et dans la mesure où y présente avec elles une corrélation non nulle. L'exposant α permet de régler le degré de prise en compte de la force des structures dans le calage de y (illustration figure 2).

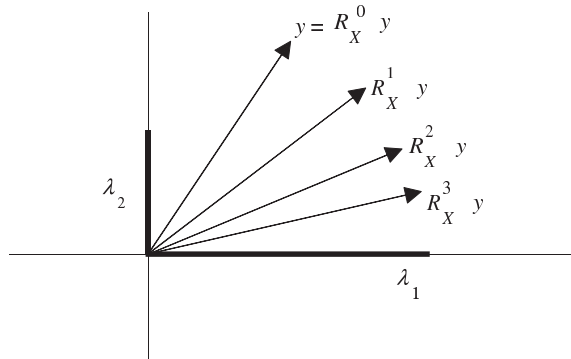


FIGURE 2

Résultante selon son degré

N.B. λ_1 et λ_2 sont les deux premières valeurs propres d'ACP de X

On a : $\forall \lambda_k > 0, \lim_{\alpha \rightarrow 0} \lambda_k^\alpha = 1$. Par conséquent : $\lim_{\alpha \rightarrow 0} R_{X,M}^\alpha y = \Pi_{\langle X \rangle} y$: lorsque α tend vers 0, les structures internes de X sont de moins en moins prises en compte, jusqu'à ne plus l'être du tout.

- Lorsque $M = (X'X)^{-1}$, on a : $R_{X,M} = X(X'X)^{-1}X' = \Pi_{\langle X \rangle}$.
- La résultante $R_{X,M}^\alpha$, dont l'opérateur est linéaire, sera dorénavant appelée *résultante linéaire*, car l'extension qu'on en propose ci-dessous sort de la linéarité.

2.2.2. Résultante non linéaire

1. Pourquoi chercher à dépasser les résultantes linéaires?

Considérons un groupe X de variables numériques centrées réduites, et affectons lui la métrique I . Quel que soit $\alpha > 0$, l'opérateur $R_{X,I}^\alpha$ attirera toujours la variable y vers les composantes principales les plus fortes de X, I avec lesquelles y a une corrélation non nulle, même – et l'on peut y voir un problème – si ces composantes sont éloignées de y (c'est-à-dire lui sont faiblement corrélées). Ceci est illustré par la figure 2, sur laquelle y est plus proche de la seconde composante principale, et se trouve pourtant attirée vers la première, parce que celle-ci est plus forte. Si l'on fait l'hypothèse que la structure de X est essentiellement unidimensionnelle, on ne peut que se satisfaire d'un tel comportement : hors de la première composante principale, on n'est censé trouver que du bruit, dont il faut s'éloigner pour se rapprocher de la structure. Mais la plupart des situations impliquent des groupes structurellement multidimensionnels. Dans ce cas, le fait que seule la force des structures intervienne dans le positionnement d'un facteur, et non leur proximité, peut nuire au dépistage de celles-ci. Considérons par exemple la figure 3, qui montre un groupe constitué de deux faisceaux de variables d'égale importance, et positivement mais faiblement corrélés.

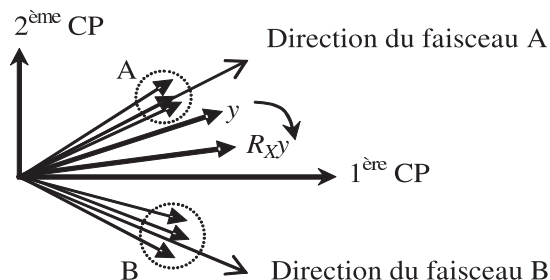


FIGURE 3

Faisceaux, composantes principales et résultantes

La variable y est proche du faisceau A, et n'a pas grand chose à voir avec B (qui n'est que faiblement corrélé avec A). Pourtant, la résultante linéaire attire y vers la première composante, et par conséquent, vers B. Pour que y soit attirée vers A, il est nécessaire d'introduire, dans le calcul de la résultante, une «prime de proximité», ce qui en détruira généralement la linéarité.

2. Résultantes non-linéaires

(a) Formules

– Soit X un groupe de variables centrées réduites et y une variable numérique réduite. On a :

$$R_{X,My} = \sum_j \langle y|x^j \rangle x^j$$

Soit $\beta \in R^+$. Nous introduisons une «prime à la proximité» dans le calcul ci-dessus, de la façon suivante :

$$\sum_j |\langle y|x^j \rangle|^\beta \langle y|x^j \rangle x^j$$

En pratique, il pourra paraître suffisant d'utiliser $\beta = 2k$, où k est un entier naturel, que nous appellerons *ordre* de la résultante non linéaire. On notera alors cette résultante non linéaire :

$$S_{X,k}y = \sum_j \langle y|x^j \rangle^{2k+1} x^j \quad (1)$$

– Généralisons la situation précédente en considérant un groupe de variables numériques X partitionné en R sous-groupes $X_1, \dots, X_r, \dots, X_R$. Posons :

$$M = \text{Diag}(\{(X_r'X_r)^{-1}\}_r)$$

La résultante linéaire de y sur ce groupe est :

$$R_{X,My} = XM X' y = \sum_r X_r (X_r' X_r)^{-1} X_r' y = \sum_r \Pi_{\langle X_r \rangle y}$$

Introduisons à présent notre prime à la proximité :

$$S_{X,k}y = \sum_r \cos^2(y, \langle X_r \rangle) \Pi_{\langle X_r \rangle y} = \sum_r \|\Pi_{\langle X_r \rangle y}\|^{2k} \Pi_{\langle X_r \rangle y} \quad (2)$$

La formule (2) généralise (1) et permet de traiter les variables qualitatives. Elle peut s'écrire :

$$S_{X,k}y = XM_{X,y,k} X' y$$

où : $M_{X,y,k} = \text{Diag}(\{\|\Pi_{\langle X_r \rangle y}\|^{2k} (X_r' X_r)^{-1}\}_r)$

$M_{X,y,k}$ est une matrice symétrique positive incluant la prime à la proximité, de sorte qu'elle dépend de y , ce qui détruit la linéarité. La matrice $M_{X,y,k}$

est une métrique euclidienne locale. Donc, la matrice $S_{X,y,k} = XM_{X,y,k}X'$ est un opérateur local de résultante. Tout comme les opérateurs de résultante linéaire, il peut être mis à n'importe quelle puissance $\alpha \in R^+$:

$$S_{X,y,k}^\alpha = (XM_{X,y,k}X')^\alpha$$

Le paramètre α est interprété de la même façon que pour les résultantes linéaires.

(b) Comportement :

– Quand $\alpha = 0$, on obtient la projection orthogonale sur $\langle X \rangle$.

– Considérons à présent que $\alpha > 0$:

Quand $k = 0$, on obtient la résultante linéaire, la prime à la proximité étant nulle.

Quand $k > 0$, les sous-groupes de variables engendrant les sous-espaces dont y est plus proche ont plus de poids.

Quand $k \rightarrow \infty$, le sous-groupe engendrant le sous-espace dont y est le plus proche (i.e. dont le $\cos(y, \langle X_r \rangle)$ est le plus grand) est prépondérant : $S_{X,k}^\alpha(y)$ est colinéaire à la projection orthogonale de y sur ce sous-espace.

Quand chaque sous-groupe est réduit à une seule variable, $S_{X,k}^\alpha(y)$ est colinéaire à la variable la plus corrélée à y .

Afin d'illustrer ce comportement, nous avons utilisé un générateur de nombres au hasard pour fabriquer un groupe X constitué de deux faisceaux de variables (A et B) faisant un angle d'environ $\pi/4$. Le faisceau A contient 4 variables (a^1, \dots, a^4), calculées en additionnant 4 bruits indépendants à une même variable. Le faisceau B ne contient que 2 variables, engendrées selon le même procédé. Le faisceau B est «plus léger» que A . Puis, plusieurs variables y^j sont engendrées par combinaison linéaire des variables de X . Enfin, les résultantes non-linéaires des y^j sur X (pondéré par I) sont calculées pour les ordres 0 à 6. Toutes les résultantes sont de degré $\alpha = 1$. La Résultante d'ordre k sera notée en abrégé $S_k y$. Toutes les variables sont projetées sur le premier plan d'ACP de X .

La figure 4 montre ce qu'il advient d'une variable (y^7) située entre les faisceaux A et B , mais plus près de B , selon la valeur de l'ordre k .

La figure 5 montre les résultantes d'ordre 0 (linéaire) et 6 (fortement non-linéaire) de toutes les variables y^1 à y^7 . On constate rapidement que toutes les résultantes S_6 se trouvent dans un voisinage des faisceaux, contrairement aux résultantes S_0 . Les variables y^4, y^5 et y^7 , qui se trouvaient dans une relative proximité du faisceau B , ont été attirées vers lui par S_6 , alors que y^5 et y^7 ont été rapprochées de A (plus proche de la première CP) par S_0 . D'autre part, y^1, y^2, y^3 et y^6 , plus proches de A , ont bien été attirées vers lui par S_6 .

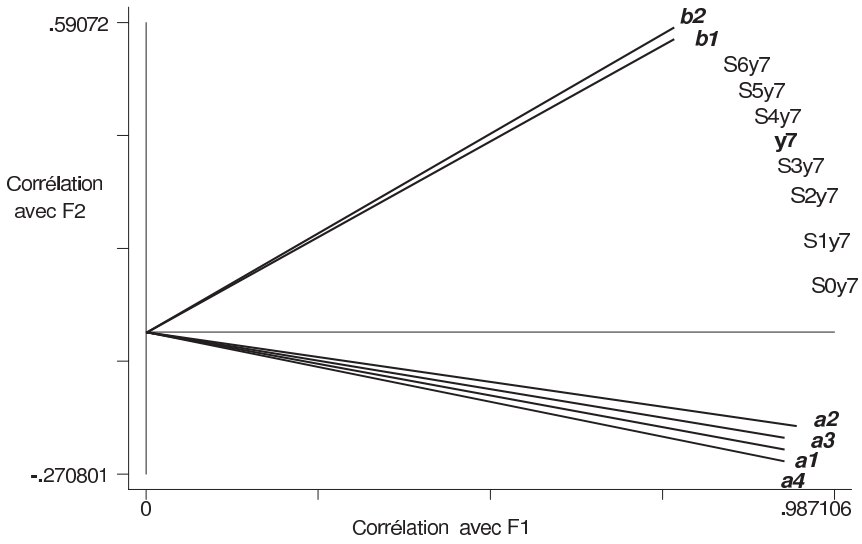


FIGURE 4

Attractions possibles d'une variable «intermédiaire» entre deux faisceaux
 N.B. On rappelle que le degré $\alpha = 1$ et que l'ordre $k = 0$ à 6

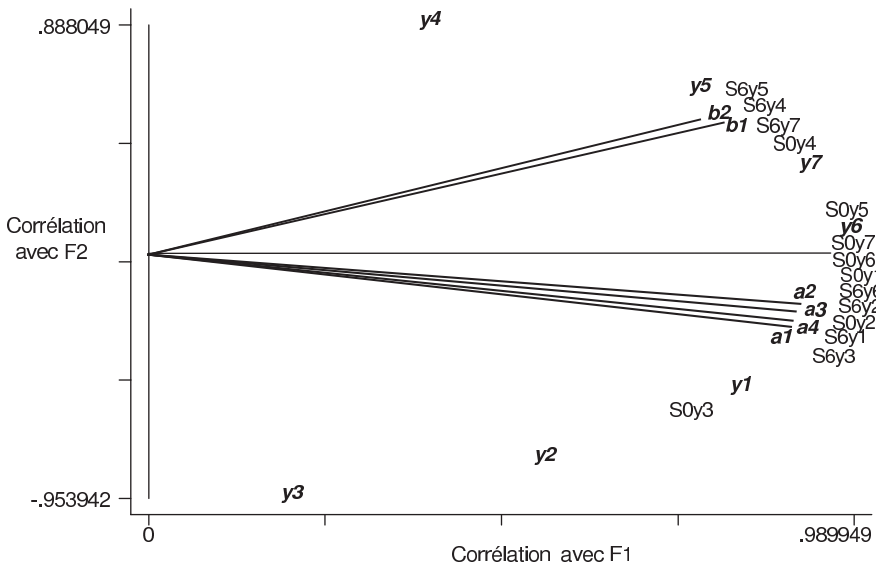


FIGURE 5

Résultantes S_0 et S_6

(c) Lien avec la rotation quartimax

Un lien peut être trouvé entre la résultante non linéaire d'ordre 1 et la rotation *quartimax*. Cette rotation a pour objectif de rapprocher un ensemble de H facteurs orthonormés d'éventuels faisceaux de variables. Elle a été construite par plusieurs auteurs [Ferguson 1954] [Carroll 1953] [Neuhaus & Wrigley 1954] [Saunders 1960] sur la base de critères distincts mais équivalents. Elle résulte par exemple du programme suivant :

$$\max_{F^1, \dots, F^H \text{ orthonormés}} \sum_{h=1}^H \sum_j \cos^4(x^j, F^h)$$

L'heuristique qui fonde ce programme peut être directement étendue à toute puissance paire supérieure à 4. On obtient ainsi, pour $k \geq 2$:

$$\max_{F^1, \dots, F^H \text{ orthonormés}} \sum_{h=1}^H \sum_j \cos^{2k}(x^j, F^h)$$

On peut mettre ce programme sous la forme :

$$\max_{F^1, \dots, F^H \text{ orthonormés}} \sum_{h=1}^H \left\langle \sum_j \cos^{2k-1}(x^j, F^h) x^j \middle| F^h \right\rangle$$

Soit :

$$\max_{F^1, \dots, F^H \text{ orthonormés}} \sum_{h=1}^H \langle S_{X, k-1}(F^h) | F^h \rangle$$

Dans chaque produit scalaire $\langle S_{X, k-1}(F^h) | F^h \rangle$, deux éléments sont pris en compte : la corrélation de chaque facteur avec sa résultante non linéaire, et la norme de cette dernière. La résultante non linéaire d'un facteur F rapprochant celui-ci d'une structure forte et proche, le facteur lui-même sera d'autant plus proche de cette structure qu'il est corrélé à cette résultante. D'autre part, la norme de cette résultante sera d'autant plus élevée que F est proche de la structure. Le critère ci-dessus est ainsi facilement interprétable.

2.3. L'algorithme ACT1

On considère le modèle illustré par la figure 1. Chaque groupe X_r est muni d'un opérateur de résultante R_{X_r} adapté au regard qu'on pose sur les variables du groupe.

Pour une résultante linéaire, on choisira une métrique M_r et un degré α . Brièvement : lorsque X_r est constitué de variables quantitatives, celles-ci seront

centrées réduites. Si l'on veut utiliser leurs structures de corrélations internes pour attirer les facteurs, on prendra $M_r = I$. L'opérateur de résultante $X_r X_r'$ a pour directions propres les facteurs d'ACP réduite de X_r , associées aux valeurs propres correspondantes. Si l'on ne veut pas tenir compte des corrélations internes à X_r , on prendra $M_r = (X_r' X_r)^{-1}$. L'opérateur de résultante associé est alors le projecteur orthogonal sur $\langle X_r \rangle$; ses directions propres sont toutes celles de $\langle X_r \rangle$, associées à la valeur propre 1; aucune n'est donc plus forte qu'une autre.

Si l'on veut tenir compte des structures fortes de corrélation, mais pour attirer les facteurs vers des faisceaux de variables proches, on emploiera une résultante non linéaire (en pouvant en faire varier l'ordre et le degré).

2.3.1. Algorithme

N.B. Cet algorithme étend aux résultantes non linéaires l'algorithme originel de l'ACT1 présenté dans [Bry 2003].

1. Détermination des facteurs de rang 1

Notations : F_r^n désigne le facteur de rang n du groupe X_r ; F_{-r}^n désigne l'ensemble des facteurs de rang n à l'exception de celui du groupe X_r ;

Itération 0 (initialisation) :

– On prend pour valeur initiale F_r^1 du premier facteur de chaque groupe X_r :

$$F_r^1 = S_{X_r, k} y \text{ normée.}$$

Itération $k + 1, k \geq 0$:

Pas a : Pour chaque groupe X_r la nouvelle valeur du facteur F_r^1 est ainsi calculée : on régresse y sur $\{X_r, F_{-r}^1\}$, on prend la composante sur $\langle X_r \rangle$ de la partie prédite.

Pas b : On calcule la résultante de cette composante sur le groupe. On norme le résultat.

Si les facteurs trouvés à l'itération $k + 1$ sont suffisamment proches de ceux fournis par l'itération k , on s'arrête. On note $F_1^1, \tau \dots \tau, \tau F_R^1$ les facteurs de rang 1 finaux.

L'itération courante est illustrée sur la figure 6.

Note : Selon la résultante utilisée lors du pas b , les résultats des modèles finaux seront différents; il est utile de les comparer en termes de variance captée par les facteurs ainsi qu'en termes de qualité d'ajustement du modèle.

2. Facteurs de rang $n \geq 2$

Chaque groupe X_r dont on désire poursuivre l'exploration est régressé sur ses facteurs F_r^1, \dots, F_r^{n-1} . On note X_r^n le groupe des résidus de cette régression.

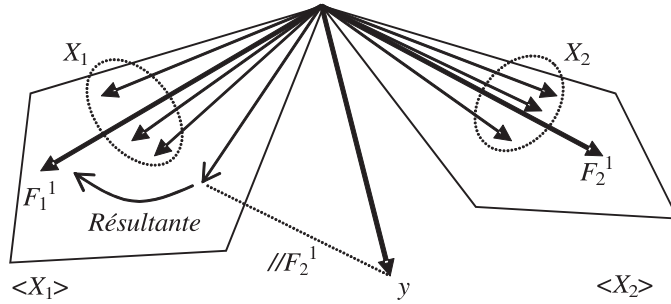


FIGURE 6

Illustration de l'étape 1 de l'ACT1 (cas de deux groupes explicatifs)

Recalcul de F_1^1 : y est projeté sur $\langle X_1 \rangle$ parallèlement à F_2^1 pour déterminer la composante sur $\langle X_1 \rangle$ de la régression de y sur $\{X_1, F_2^1\}$. Puis, on calcule la résultante de cette composante sur X_1 . On procède symétriquement pour calculer la valeur courante de F_2^1 .

Ceci fait, on reconduit l'algorithme utilisé pour le calcul des facteurs de rang 1 en prenant pour groupes explicatifs les groupes de résidus X_r^n ainsi que chacun des facteurs de rang inférieur à n obtenus dans tous les groupes.

2.3.2. Propriétés

On rappelle ici deux propriétés importantes de l'algorithme lorsque la résultante utilisée est linéaire.

Propriété P1 : lorsque le groupe X_r n'a pas de structure (ou qu'on choisit d'ignorer celle-ci), l'algorithme n'utilise ce groupe qu'en tant que sous-espace $\langle X_r \rangle$.

Ne pas tenir compte de la structure interne d'un groupe X_r , c'est lui adjoindre la métrique $M_r = (X_r^T X_r)^{-1}$. L'opérateur de résultante linéaire associé est alors le projecteur orthogonal sur $\langle X_r \rangle$. Cela revient, dans l'algorithme, à supprimer le pas b de l'itération courante. Celle-ci est ainsi réduite au pas a , qui n'utilise X_r qu'en tant que sous-espace prédictif. Il est important de noter que dès lors, l'itération courante, en régressant y sur $\{X_r, F_{-r}^1(k)\}$, trouve le facteur F_r^1 qui maximise l'ajustement du modèle de y . Ne pas vérifier la propriété P1 impliquerait qu'une contrainte autre que l'ajustement du modèle est artificiellement introduite par l'algorithme dans le calcul du facteur F_r courant.

Propriété P2 : lorsque l'on ne tient pas compte des structures de corrélations internes des groupes X_r , le R^2 de régression de y sur les F_r ne peut que croître à chaque itération et à chaque étape.

Cette propriété est une conséquence de P1. La preuve est donnée dans [Bry 2004 b] par exemple.

Lorsque l'on ne tient pas compte de l'existence de structures fortes à l'intérieur des groupes, seule la qualité d'ajustement du modèle fondé sur les variables latentes doit intervenir dans l'estimation de ces dernières; il paraît donc nécessaire de garantir au moins la croissance systématique de cette qualité au cours de l'algorithme.

3. Extension de l'ACT1 au modèle linéaire généralisé

3.1. L'extension ACT1G

Considérons de nouveau le modèle de la figure 1. La variable dépendante y est supposée suivre ici, conditionnellement aux groupes X_r , une loi P_θ , où θ est un vecteur de paramètres pouvant s'exprimer en fonction d'une combinaison linéaire des variables latentes. On posera $\theta = g(Fb)$, où $F = (F_1, \dots, F_R)$ et b est un vecteur de R paramètres scalaires.

L'extension de l'ACT1 à un modèle linéaire généralisé de y est directe. Il suffit d'y remplacer, dans le pas a de l'itération courante, la régression linéaire de y sur $Z = \{X_r, F_{-r}^1(k)\}$ par la régression linéaire généralisée adaptée au modèle de y et utilisant la maximisation de vraisemblance pour l'estimation. Cette régression fournit une estimation $Z\hat{c}$, dont la composante sur $\langle X_r \rangle$ est notée $X_r\hat{c}_r$. Le pas b prend pour nouvelle valeur de F_r^1 la résultante sur X_r de $X_r\hat{c}_r$.

Nous appellerons ACT1G cette extension.

3.2. Propriétés

La propriété P1 est trivialement vérifiée par ACT1G. La propriété P2 est étendue en :

Propriété P2' : lorsque l'on ne tient pas compte des structures de corrélations internes des groupes X_r , la vraisemblance du modèle de y en fonction des F_r ne peut que croître à chaque itération et à chaque étape.

Preuve : au pas a de l'itération courante, la régression linéaire généralisée de y sur $Z = \{X_r, F_{-r}^1(k)\}$ maximise sur $F_r \in \langle X_r \rangle$ la vraisemblance du modèle, les facteurs $F_s, s \neq r$ étant fixés. Ce maximum étant atteint, on fixe F_r et l'on permet à un autre facteur F_s de varier dans $\langle X_s \rangle$ afin de maximiser la vraisemblance sur F_s . La vraisemblance croît donc par rapport à la valeur précédente. Et ainsi de suite.

Une fois stabilisés les facteurs d'un rang donné, l'introduction de facteurs de rang supérieur élargit le sous-espace de maximisation (ne pas introduire ces facteurs dans le modèle équivaut simplement à imposer la nullité de leur coefficient). La vraisemblance croît donc là encore.

Note : L'utilisation d'ACT1G avec un modèle gaussien ordinaire (*i.e.* à perturbations indépendantes et homoscédastiques) pour y continue redonne trivialement l'ACT1 originelle, la maximisation de la vraisemblance correspondant dans ce cas à la régression ordinaire.

3.3. Cas particulier de la régression PLS

– Lorsque l'on n'a qu'un groupe $X = \{x^j\}_{j=\tau_1, \dots, J}$ de prédicteurs (centrés réduits) d'une variable numérique y , et qu'on utilise la résultante linéaire associant à X la métrique I , l'ACT1 revient à faire la régression PLS univariée (notée PLS1) de y sur X . En effet, le pas a projette orthogonalement y sur $\langle X \rangle$ et le pas b calcule sa résultante $R_{X,I}\Pi_{\langle X \rangle}y$. Or : $R_{X,I}\Pi_{\langle X \rangle}y = XX'\Pi_{\langle X \rangle}y = X(\Pi_{\langle X \rangle}X)'y = XX'y$. Le facteur F^1 est donc la première composante PLS de y . Le procédé étant ensuite reconduit sur les résidus de régression de X sur F^1 , on trouve la deuxième composante PLS. Etc.

– Supposons à présent que y suive un modèle linéaire généralisé, conditionnellement à X . L'ACT1G devient dans ce cas une extension de PLS1 à la modélisation linéaire généralisée. Cette extension consiste, pour le facteur F^1 , à faire la régression linéaire généralisée de y sur X , puis à calculer la résultante sur X de la prédiction $X\hat{b}$ obtenue. Nous désignerons cette variante par le sigle : PLS1G.

Une autre extension de PLS1, nommée PLSGLM (PLS for Generalized Linear Models) a été proposée dans [Bastien *et al.* 2004]. Dans cette dernière, le facteur de rang 1 est ainsi calculé :

– On estime prenant tour à tour un modèle linéaire généralisé ayant pour unique variable explicative chacune des variables x^j . Ceci fournit autant d'estimations $\hat{b}_j x^j$.

– On pose alors : $F^1 = \sum_{j=1, \dots, J} \hat{b}_j x^j$ normé.

Le facteur de rang 2 est obtenu en reconduisant cette démarche sur les résidus de régression de X sur F^1 , et ainsi de suite.

Cette méthode possède un avantage évident sur l'ACT1G en cas de multicolinéarité parfaite dans X , puisqu'elle ne procède à aucune régression sur X . Cet avantage nous semble toutefois à relativiser. En effet, d'une part, les multicolinéarités de X n'empêchent nullement la régression sur X , mais seulement son calcul *direct* (les bons logiciels éliminent d'ailleurs les prédicteurs redondants afin de procéder à cette régression). D'autre part, on montre facilement que PLSGLM ne possède pas la propriété P1. Dans le cas où X n'a pas de structure principale de corrélation ($X'X = I$), il n'y a aucune raison que le premier facteur fourni par cette extension soit l'estimateur du maximum de vraisemblance de Xb . Supposons en effet, par exemple, que X soit constitué de deux variables centrées réduites décorrélées : ces variables sont I -orthogonales. Supposons en outre que y suive un modèle linéaire gaussien à perturbations ϵ hétéroscédastiques. La matrice de variance Ω de ces perturbations n'est pas proportionnelle à I . La maximisation de la vraisemblance de tout modèle gaussien ayant cette structure de perturbation est obtenue en projetant Ω -orthogonalement y sur le sous-espace des prédicteurs. Or, comme Ω n'est pas proportionnelle à I , on aura en général :

$$\Pi_{\langle x^1 \rangle}^{\Omega-\perp} y + \Pi_{\langle x^2 \rangle}^{\Omega-\perp} y \neq \Pi_{\langle x^1, x^2 \rangle}^{\Omega-\perp} y$$

Le fait que P1 ne soit pas vérifiée provient d'un manque de séparation, dans cette technique d'estimation, entre les deux contraintes d'estimation que sont : l'ajustement du modèle de prédiction (par maximisation de la vraisemblance) et la proximité de

chaque facteur à des structures fortes de X (obtenue par calcul de résultante). Une telle séparation permet à l'ACTIG de retrouver la pure maximisation de vraisemblance lorsqu'on décide d'ignorer les structures internes des groupes de prédicteurs.

4. Exemple d'application

4.1. Présentation

Nous avons appliqué la méthode proposée à l'étude d'un événement – le divorce – dans la vie d'un homme dakarais de la génération ayant 25 à 45 ans à la date de l'enquête. Nous utilisons les données de l'enquête biographique réalisée par l'IRD-IFAN en 2001 à Dakar [Antoine *et al.* 2002]. On recueille rétrospectivement, le jour de l'enquête, l'état de chaque individu de l'échantillon relativement à un certain nombre de variables au cours de sa vie. Les dates des changements d'état (appelés aussi événements) sont relevées. Le jour de l'enquête représente pour chaque individu enquêté l'instant où il sort d'observation. L'indépendance stochastique entre la sortie d'observation et les autres événements est une condition nécessaire de l'analyse biographique.

Nous recherchons les facteurs influençant le plus le risque instantané de divorcer. Une échelle du temps étant choisie pour chaque individu – ici, son âge – on note T la durée aléatoire écoulée au moment où l'individu connaît le divorce (événement noté E). Le risque instantané de connaître E à la date t est noté $h(t)$ et défini comme :

$$h(t) = \lim_{dt \rightarrow 0} \frac{1}{dt} P(t \leq T < t + dt | t \leq T)$$

Le vecteur des caractéristiques d'un individu à la date t étant noté x_t , le risque instantané encouru par cet individu de connaître E à t conditionnellement à ces caractéristiques est noté $h(t|x_t)$. La dépendance du risque à ses déterminants est formalisée à l'aide d'un modèle de Cox. Celui-ci suppose que chaque déterminant a un effet multiplicatif constant sur le risque instantané de connaître l'événement étudié E , les autres déterminants restant fixés. Notant β un vecteur de coefficients de même taille que x_t , le modèle de Cox pose ainsi :

$$h(t|x_t) = h_0(t)e^{\beta' x_t}$$

La régression de Cox utilisée par chaque itération dans notre extension de l'ACT procède à une maximisation de vraisemblance qui intègre le phénomène de sortie d'observation. Notre étude est fondée sur un échantillon de 137 hommes dont 23 ont connu le divorce au jour de l'enquête. Les variables explicatives retenues (*cf.* tableau 1) sont relativement nombreuses, eu égard à la taille modeste de l'échantillon : 53 variables quantitatives ou indicatrices de modalités. La régression de Cox sur l'ensemble des variables, qui n'ont pas été présélectionnées en fonction de leur relative décorrélation, fait courir de nombreux risques de multicollinéarité. De fait, très peu de modalités explicatives ont un effet statistiquement significatif sur le risque. Si l'on ne

retient que les modalités interprétables, il n'y en a que trois (*Autre ethnie* n'a pas un contenu clair) qui influencent nettement, toutes choses égales par ailleurs, le risque de divorcer rapidement : la religion Mouride, l'absence de logement autonome et l'existence d'un lien de parenté du côté maternel avec l'épouse. (cf. tableau 1).

Nous allons alors procéder successivement à deux analyses exploratoires de ces déterminants. En premier lieu, nous lancerons une analyse non thématique utilisant l'extension de PLS à laquelle se réduit l'ACT lorsque le groupe prédictif est unique (cf. section 3.3). Dans un deuxième temps, l'ensemble des déterminants sera structuré en quatre groupes thématiques : un groupe de caractéristiques culturelles, un groupe de variables éducationnelles et parentales, un groupe de caractéristiques économiques, et enfin un groupe de variables caractérisant l'union (cf. tableau 1 et figure 7). Nous verrons à quel modèle factoriel et à quelle sélection de prédictifs chacune de ces deux méthodes conduit.

N.B. Toutes les résultantes utilisées seront de degré $\alpha = 1$.

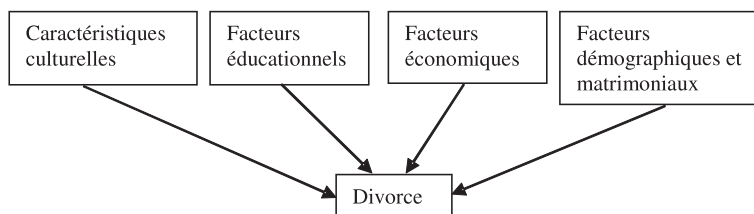


FIGURE 7

Modèle thématique du risque de premier mariage

4.2. Analyse non-thématique (PLS1G)

Exploration et interprétation des facteurs

On utilise dans un premier temps la résultante linéaire; puis, la résultante non linéaire d'ordre $k = 1$.

N.B. On dira qu'un facteur est bien illustré par une variable de son thème quand il est fortement corrélé à celle-ci. Il sera dit richement illustré s'il est bien illustré par plusieurs variables du thème.

PLS1G avec résultante linéaire :

Une fois obtenus les premiers facteurs, on estime le modèle de Cox fondé sur eux. On utilise empiriquement la probabilité critique liée à l'estimation de l'effet de chaque facteur sur le risque pour mesurer la force de la liaison partielle entre ce facteur et le risque dans le modèle (sur le plan de l'inférence, cette utilisation est évidemment illicite, les facteurs n'étant pas exogènes). Il apparaît que selon ce critère, seul le

TABLEAU 1

Régression de Cox sur la totalité des facteurs du divorce

N.B. Le symbole ** indique que le fait que le multiplicateur du risque estimé est inférieur (resp. supérieur) à 1 est significatif au niveau 1%; * au niveau 5%.

Groupe	Variable	Modalités	Etiquettes	Multiplicateur du risque	Significativité	Répartition en % (homme-mois)
Facteurs culturels	Ethnie	Wolof	wolof	ref		43
		Poular	alpoular	1,38		28
		Serer	serer	0,25		14
		Diola	diola	1,72		2
		Autre ethnie	aut eth	0,01	**	13
	Religion	Autre musulman	aut_musu	ref		25
		Mouride	mouride	36,55	*	47
		Tidiane Chrétien	tidiane chrétien	3,94 13,77		25 3
	Lieu de naissance	Dakar	ln_dakar	ref		66
		milieu rural milieu urbain	ln_rural ln urbain	1,14 4,48		19 14
	Lieu socialisation	Dakar	se_dakar	ref		63
		en milieu rural en milieu urbain	se_rural se urbain	1,85 0,01		20 17
Facteurs éducationnels et parentaux	Instruction mère	Non scolarisé	mnonsc	ref		92
		Primaire	mprim	0,18		7
		Secondaire et plus	msecp	7,04		1
	Instruction père	Non scolarisé	pnonsc	ref		71
		Primaire	pprim	0,99		14
		Secondaire et plus	psecp	4,42		15
Instruction	Non scolarisé	nonscol	ref		31	
	Primaire	primaire	0,63		33	
	Secondaire et plus	second_p	2,11		36	
Facteurs économiques	Logement	Autonome	logauto	ref		55
		Jamais logement autonome	jamloau	23,86	**	45
	Période d'activité	Informel	informel	ref		26
		Patron	patron	2,93		17
		Salarie	salarie	2,46		47
		Apprenti ou élève	app_elev	0,66		7
		Chômeur	chomeur	1,81		3
	Activité épouse	Inactive	cfnonact	ref		67
		Employé	cfemploy	9,56		2
		Vendeuse	cfvente	0,16		19
Domestique		cfdomest	0,90		13	
Facteurs démographiques et matrimoniaux	Descendance	Pas d'enfant	pasenf	ref		26
		1 enfant	enf1	2,11		25
		2 enfants	enf2	0,23		21
		3 enfants et plus	enf3p	0,36		29
	Âge au mariage	Avant 21 ans	am_av20	ref		5
		21 - 25 ans	am21a25	0,29		37
		26 - 29 ans	am26a29	0,25		33
		30 - 34 ans	am30a34	2,92		19
		Après 35 ans	am_ap35	3,91		7
	Choix conjoint	Par parent	chxpar	ref		12
		Par l'intéressé	chxmoim	2,20		88
	Parenté conjoint	Non apparenté	nonpart	ref		54
		Côté paternel	parpat	0,82		25
		Côté maternel	parmat	0,13	*	21
Expérience divorce du conjoint	Cjt jamais divorcé auparavant	cjlnodiv	ref		97	
	Cjt déjà divorcé	cjlexdiv	0,05		3	
Nature de l'union	Monogame	monogame	ref		94	
	Polygame	polygame	5,74		6	

premier facteur est fortement lié au risque, le deuxième présentant aussi une certaine liaison, mais beaucoup moins nette.

Le facteur 1 est assez hétérogène et remarquablement mal illustré, aucune variable ne faisant avec lui une corrélation allant jusqu'à 0.5. Il est difficilement interprétable.

Le facteur 2 met en relief un unique phénomène : le statut de l'union (monogame / polygame). Outre que l'illustration de l'axe n'est pas riche, elle n'est pas d'une très grande qualité (la corrélation de polygame avec le facteur 2 n'est que de 0.79).

Au final, convenons que les facteurs pointent peu de prédicteurs, et que le seul facteur très lié au risque, le premier, est peu interprétable car aucune variable n'en est proche.

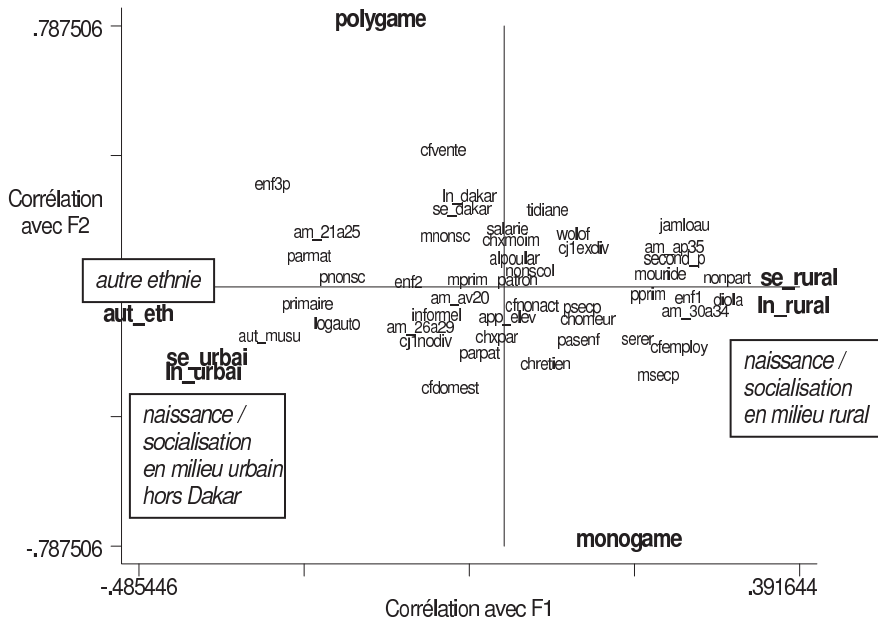


FIGURE 8

Plan explicatif (1,2) fourni par PLS1G avec résultante linéaire

PLS1G avec résultante non linéaire d'ordre 1 :

En utilisant la résultante non linéaire d'ordre 1, on espère rapprocher le facteur 1 de variables pour le rendre plus facile à interpréter. On constate (cf. figure 9) que la variable qui lui est la plus corrélée – *autre ethnie*, prédicteur dépisté par la régression de Cox sur toutes les variables – a fortement attiré le premier facteur. Celui-ci se trouve à présent précisément illustré, mais ne représente plus qu'une seule variable. Le facteur 2 s'est, de même, rapproché de la variable qui lui est la plus corrélée : *polygame/monogame*.

En définitive, si l'illustration des facteurs est plus nette, elle reste ici pauvre.

Sélection de prédicteurs et modèle parcimonieux

Selon les résultats de PLS1G, on est fondé à retenir les prédicteurs suivants : *autre ethnie* (sans interprétation) et *polygame*. Mis ensemble dans un même modèle et

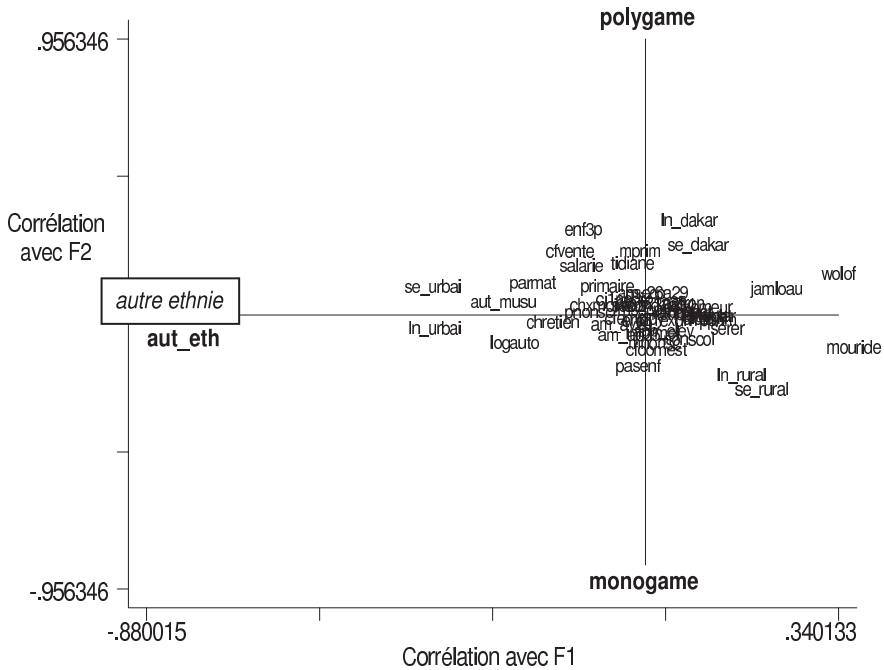


FIGURE 9

Plan explicatif (1,2) fourni par PLS1G avec résultante non linéaire d'ordre 1

à l'exclusion d'autres prédicteurs, aucune de ces variables n'a d'effet statistiquement significatif. Dans le modèle complet, cependant, *autre ethnie* avait un effet partiel assez fortement réducteur du risque de divorce.

Diagnostic

Une ACP de la totalité des variables explicatives fait apparaître que la décroissance des valeurs propres est lente : il n'y a pas de structures fortes nettement identifiables dans cet ensemble de variables. Par conséquent, cet ensemble agira globalement comme sous-espace : le premier facteur est essentiellement estimé par l'ajustement d'une combinaison linéaire des explicatives à la variable dépendante ; il est donc tout à fait normal de trouver un premier axe très lié à la variable dépendante, mais éloigné des variables explicatives. Il est alors non moins normal de trouver des facteurs de rang 2 et ultérieur peu liés à la variable dépendante : le premier facteur a pris toute la puissance prédictive.

4.3. Analyse thématique (ACTIG)

Exploration et interprétation des facteurs

L'ACTIG, partant du modèle thématique de la figure 7, s'intéresse d'emblée à la séparation des effets des quatre thèmes explicatifs. Ici aussi, on utilise dans un premier temps la résultante linéaire; puis, la résultante non linéaire d'ordre 1.

ACTIG avec résultante linéaire :

Dans cette analyse, seul le facteur de rang 1 de chaque groupe explicatif a un effet partiel très net sur le risque de divorce (cf. tableau 2, avec la mise en garde précédente concernant l'usage des probabilités critiques). Nous allons néanmoins donner une représentation de chaque thème dans le plan de ses deux premiers facteurs.

TABLEAU 2

Régression de Cox sur les premières composantes thématiques

N.B. Le facteur j du groupe Xk est noté XkFj. Les facteurs étant réduits, l'augmentation d'une unité correspond à une augmentation d'un écart-type. On constate que chacun des facteurs de rang 1 est amplificateur du risque.

Facteurs	Multiplicateur du risque (effet partiel)	Probabilité critique
X1F1	4.88	$< 10^{-3}$
X1F2	0.83	0.48
X2F1	1.82	$< 10^{-2}$
X2F2	1.03	0.87
X3F1	3.89	$< 10^{-3}$
X3F2	1.16	0.5
X4F1	2.98	$< 10^{-3}$
X4F2	1.21	0.37

Facteurs culturels :

Le facteur 1 est médiocrement corrélé aux variables du thème, donc assez mal illustré, principalement par la *naissance / socialisation en milieu rural*, et l'appartenance à la confrérie *mouride* (cf. figure 10).

Facteurs éducationnels :

Le facteur 1 du thème éducationnel met en relief le niveau de scolarisation secondaire et supérieur de l'individu comme de ses parents (cf. figure 11).

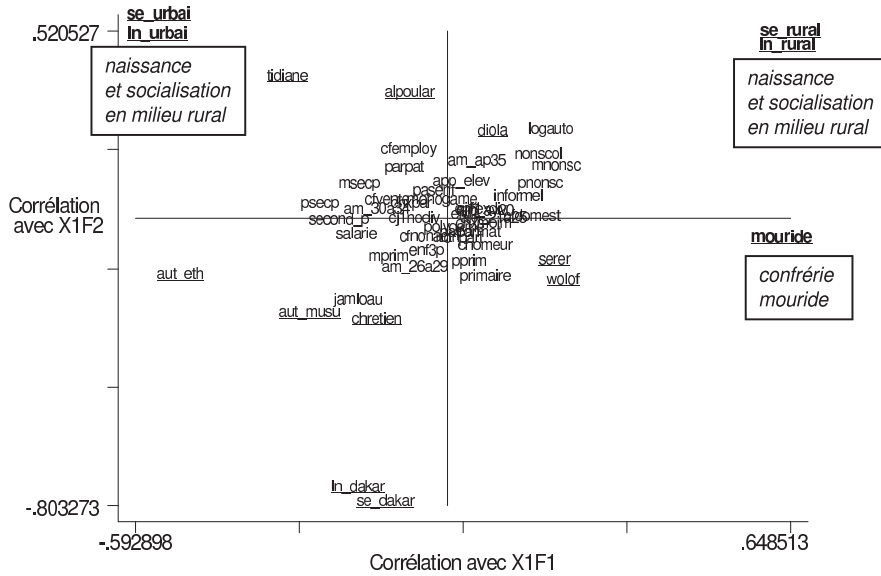


FIGURE 10

Plan (1,2) des facteurs culturels

N.B. Les variables du thème culture sont celles qui sont soulignées.

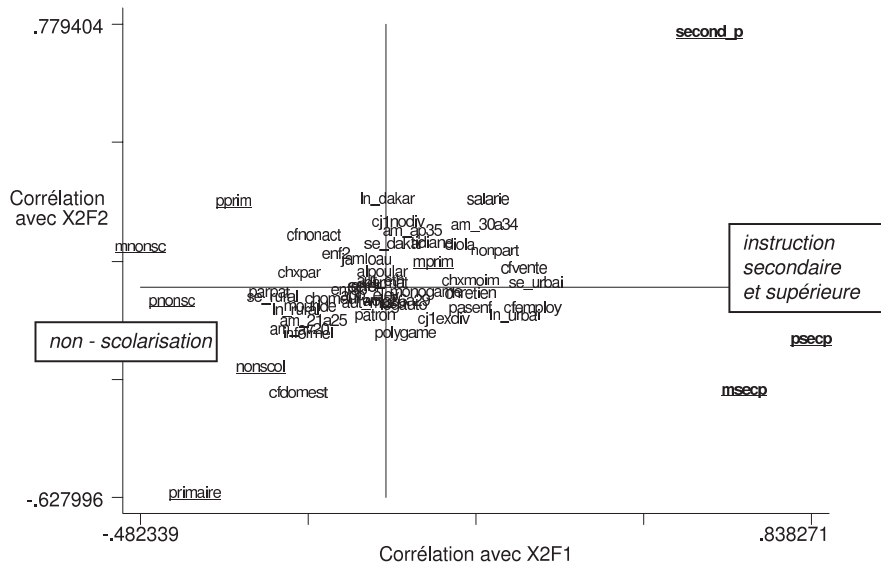


FIGURE 11

Plan (1,2) des facteurs économiques

N.B. Les variables du thème éducation sont celles qui sont soulignées.

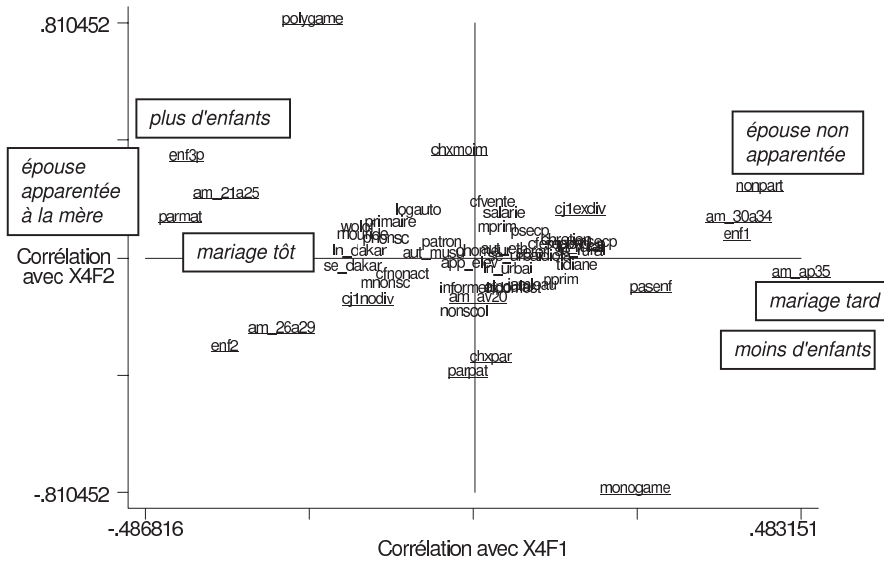


FIGURE 13

Plan (1,2) des facteurs démographiques et familiaux
 Les variables du thème démographie/famille sont celles qui sont soulignées.

TABLEAU 3

Régression de Cox sur les premières composantes thématiques.

Facteurs	Multiplicateur du risque (effet partiel)	Probabilité critique
X1F1	3.85	$< 10^{-3}$
X1F2	0.81	0.42
X2F1	1.99	$< 10^{-2}$
X2F2	1.26	0.34
X3F1	2.78	$< 10^{-2}$
X3F2	1.29	0.5
X4F1	3.49	$< 10^{-2}$
X4F2	1.4	0.05

Facteurs culturels :

Le facteur 1 est cette fois très bien illustré par l'appartenance à la confrérie *mouride* (corrélation de 0.90; cf. figure 14).

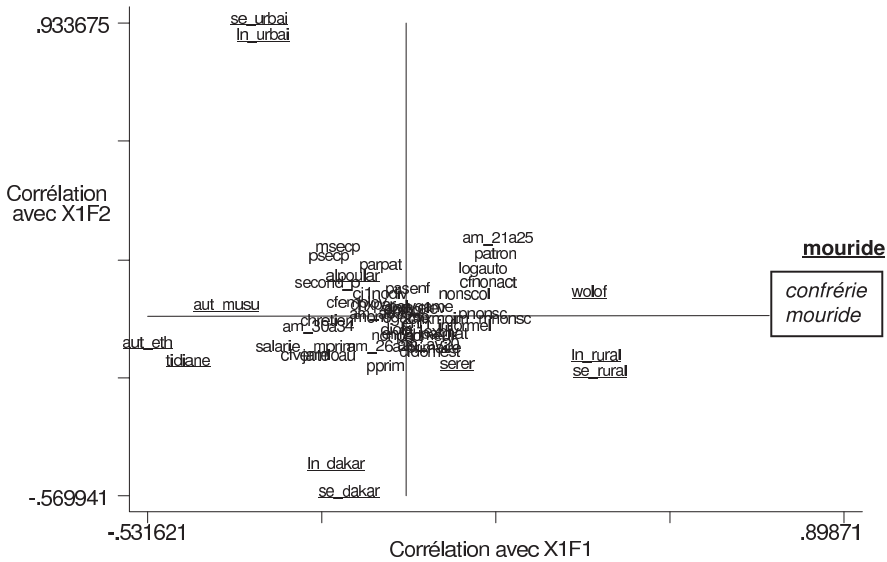


FIGURE 14

Plan (1,2) des facteurs culturels

N.B. Les variables du thème culture sont celles qui sont soulignées.

Facteurs éducationnels :

Le facteur 1 du thème éducationnel met comme précédemment en relief le niveau de scolarisation supérieur de l'individu comme de ses parents, mais reproduit un peu mieux une gradation allant de la non-scolarisation au niveau secondaire et supérieur (cf. figure 15). On peut pratiquement l'interpréter comme un niveau d'instruction contextuel de l'individu.

Facteurs économiques :

Le facteur 1 du thème économique s'est rapproché de la variable : *autonomie de logement* (cf. figure 16). Le facteur est à présent bien illustré par cette variable.

Facteurs démographiques et familiaux :

Le facteur 1 de ce thème est bien mieux illustré que celui fourni par la résultante linéaire. Son interprétation devient possible : d'une part, il oppose l'apparement de l'épouse au côté maternel de l'homme à son absence d'apparement; d'autre part il reproduit en partie la gradation de l'âge au mariage. Le facteur 2 est celui du statut polygame de l'union, mais on a vu que son effet n'était pas aussi net que celui des autres.

Bilan :

L'usage d'une résultante non linéaire, en rapprochant les composantes thématiques des variables, a permis de les rendre plus aisées à interpréter. Ce faisant, elle facilite grandement la sélection finale des prédicteurs.

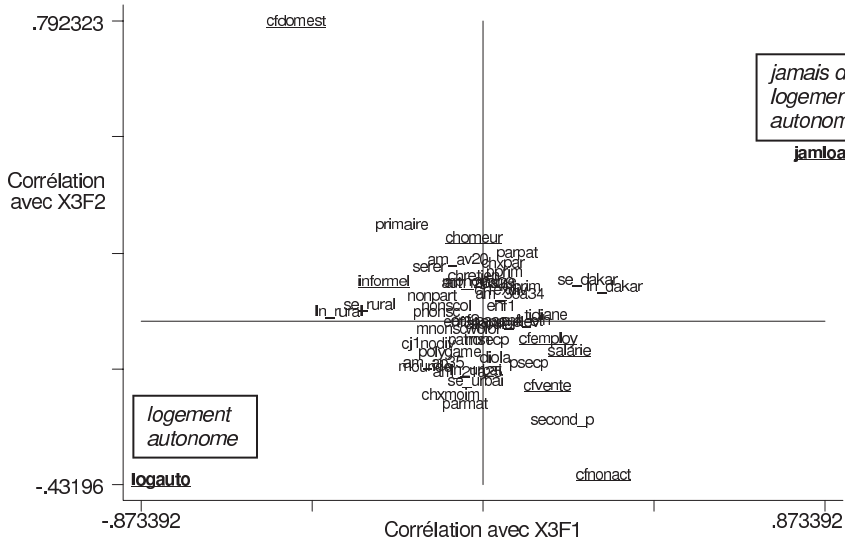


FIGURE 16

Plan (1,2) des facteurs économiques

N.B. Les variables du thème économie sont celles qui sont soulignées.

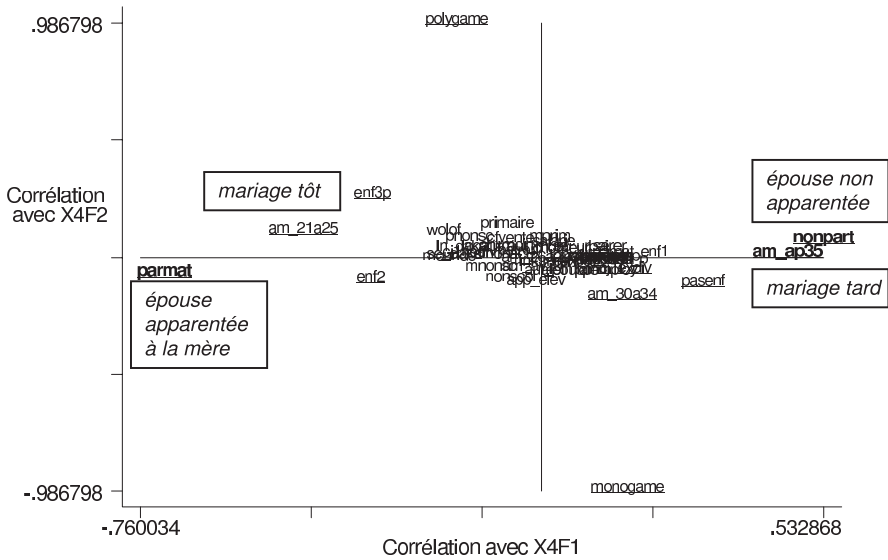


FIGURE 17

Plan (1,2) des facteurs démographiques et familiaux

N.B. Les variables du thème démographie/famille sont celles qui sont soulignées.

TABLEAU 4
Régression de Cox sur les prédicteurs sélectionnés

Variable/Modalité	Etiquette	Multiplicateur du risque	Prob. crit. unilatérale
Religion mouride	<i>mouride</i>	7.99	0.000
Jamais de logt autonome	<i>jamloau</i>	1.34	0.006
Épouse non apparentée	<i>nonpart</i>	3.81	0.002
Niveau scolaire (var. continue)	<i>niscola</i>	2.63	0.037
Âge au mariage (var. continue)	<i>agemar</i>	1.51	0.042

4.4. Conclusion

Dans une situation où les variables sont nombreuses et les colinéarités non contrôlées, la régression de Cox sur l'ensemble des variables occulte certaines dimensions explicatives importantes. L'utilisation de la généralisation de PLS aux MLG n'a pas été d'un grand secours, car l'unique groupe de prédicteurs s'est avéré faiblement structuré, de sorte que la seule dimension prédictive fournie par la méthode est restée pratiquement impossible à interpréter, et n'a pas conduit à une sélection de prédicteurs efficaces. L'extension de l'ACT, fondée sur un modèle explicatif thématique, a quant à elle nettement amélioré le dépistage. L'utilisation de la résultante linéaire a fourni des facteurs restant en demi-teinte, aucun des thèmes explicatifs n'étant très fortement structuré intérieurement. Mais l'emploi d'une résultante non linéaire a permis de rapprocher les facteurs de variables particulières qui se sont révélées, rassemblées dans un modèle, avoir un effet partiel fort sur le risque étudié.

Remerciements

Je remercie ici Philippe Antoine pour les données et sa collaboration, et Pierre Cazes pour son attentive relecture et ses conseils.

Références

- ANTOINE P. et ABDOU SALAM FALL (sd de) (2002), *Crise, passage à l'âge adulte et devenir de la famille dans les classes moyennes et pauvres à Dakar*, Rapport d'étape pour le CODESRIA, IRD-IFAN, Dakar, Mars 2002, 118p + 22p annexes.
- ANTOINE P., DIAL FATOU B. (2003), Mariage, divorce et remariage à Dakar et Lomé. Journées scientifiques de l'AUF *Familles du Nord, Familles du Sud*, Marseille 23-26 juin 2003, 22 p. (à paraître)

- BASTIEN P., ESPOSITO VINZI V., TENENHAUS M. (2005), *PLS generalized linear regression*, CSDA, vol. 48, pp. 17-46.
- BRY X. (2001), *Une autre approche de l'Analyse Factorielle : l'Analyse en Résultantes Covariantes*, RSA vol. 49(3), pp. 5-38.
- BRY X. (2003), *Une méthode d'estimation empirique d'un modèle à variables latentes : l'Analyse en Composantes Thématiques*, RSA vol. 51(2), pp. 5-45 .
- BRY X. (2004 a), *Estimation empirique d'un modèle à variables latentes avec interactions*, RSA, vol 52(3), pp. 5-35.
- BRY X. (2004 b), *Une méthodologie exploratoire pour l'analyse et la synthèse d'un modèle explicatif : l'analyse en composantes thématiques*, Thèse de doctorat, Laboratoire LISE-CEREMADE, Université de Paris IX-Dauphine.
- BRY X., ANTOINE P. (2004), *Explorer l'explicatif : application à l'analyse biographique*, Population F, 59(6), pp.909-946.
- CARROLL J.-B. (1953), *An analytical solution for approximating simple structure in factor analysis*, Psychometrika, vol. 18, pp. 23-38.
- FERGUSSON G. A. (1954), *The concept of parsimony in factor analysis*, Psychometrika, vol. 19, pp. 281-290.
- LOHMÖLLER J.-B. (1989), *Latent Variables Path Modelling with Partial Least Squares*, Physica-Verlag, Heidelberg.
- MOREAU J., DOUDIN P.-A., CAZES P. (2000), *L'analyse des correspondances et les techniques connexes - Approches nouvelles pour l'analyse statistique des données*, Springer, coll. Mathématiques et applications 32.
- MORRISON D. F. (1967), *Multivariate Statistical Methods*, McGraw-Hill Series in Probability and Statistics.
- NEUHAUS J., WRIGLEY C. (1954), *The Quartimax method : An analytical approach to orthogonal simple structure*, British Journal of Statistical Psychology, vol.7 pp. 81-91.
- SAUNDERS D. R. (1960), *A computer program to find the best-fitting orthogonal factors for a given hypothesis*, Psychometrika, vol.25, pp. 207-210.
- TENENHAUS M. (1998), *La régression PLS, théorie et pratique*, Technip.
- TENENHAUS M. (1999), *L'approche PLS*, Revue de Statistique Appliquée, vol. 47(2), pp. 5-40.
- THURSTONE L. L. (1945), *Multiple Factor Analysis*, University of Chicago Press.
- TUCKER L. R. (1958), *An inter-battery method of factor analysis*, Psychometrika, nr 23, pp. 111-136.
- VIVIEN M. (2002), *Approches PLS linéaires et non linéaires – Théorie et pratique*, Thèse de doctorat, Université Montpellier I.
- WOLD H. (1985), *Partial Least Squares*, Encyclopedia of Statistical Sciences, John Wiley & Sons, pp. 581-591.