Réduction de la dimension des variables auxiliaires par calage sur les composantes PLS: une application au panel de mesure de l'audience TV de Marocmetrie.

Sara Nahchel*, Zoubir Zarrouk**, Jelloul Allal*, Younes Alami***

*Faculté des Sciences – Université Mohamed 1er, BV Mohammed VI BP 717 60000 Oujda
**Faculté des Sciences Juridiques, Economiques et Sociales - Université Mohamed 1er,
Complexe universitaire - Hay Al Qods BP 724 Oujda.
*** Marocmetrie 179, bd Hassan 1er, 2ème étage, Casablanca

s.nahchel@gmail.com
zarrouk.zoubir@gmail.com
jell.allal@gmail.com
younes.alami@marocmetrie.ma

Résumé : Le calage sur marges est une technique qui consiste à ajuster l'échantillon de départ sur la marge des critères de contrôle, marge que l'on connait sur l'ensemble de la population (Deville et Särndal, 1992). L'idée de base du calage est de pouvoir utiliser l'information auxiliaire non prise en compte en vue d'augmenter la précision des estimateurs. Cependant l'amélioration apportée par l'estimateur de calage reste conditionnée par le choix et le nombre des variables auxiliaires utilisées dans le processus de calage. En effet, la variance de l'estimateur par calage risque de devenir importante quand on utilise un très grand nombre de variables auxiliaires, notamment en présence d'une forte multicolinéarité. Les principales solutions utilisées dans la littérature sont la réduction des dimensions des variables auxiliaires par calage sur les composantes ACP (Goga, Shehzad et Vanheuverzwyn, 2011) et le calage pénalisé basé sur la régression RIDGE (Beaumont et Bocci, 2008). Dans cette communication nous proposons une nouvelle technique de réduction de la dimension des variables auxiliaires basée sur les composantes PLS. Les résultats obtenus sur les données du panel de mesure de l'audience TV de Marocmetrie montrent l'efficacité du calage PLS par rapport au calage classique et au calage basé sur les composantes ACP.
Mots-clés. Calage sur marges, information auxiliaire, régression PLS, calage sur composantes PLS, Régression RIDGE, calage sur composantes ACP.

Abstract: Calibration is a technique that consists of adjusting the initial sample on a margin that is supposed known for whole the population (Deville and Särndal, 1992). The basic idea of calibration is to use auxiliary information that is not taken into account, in order to increase the estimator precision. However, the improvement provided by the calibration estimator depends on the choice of the auxiliary variables used in the calibration process as well as their number. In fact, the variance of the calibration estimator may become significant when we use a large number of auxiliary variables especially with the presence of strong multi-colinearity. Many solutions have been used in the literature such as reducing the dimension of these variables by calibrating on principal components (Goga, Shehzad and Vanheuverzwyn, 2011) and calibration based on RIDGE regression (Beaumont and Bocci, 2008). The result of the analysis applied on data provided by Marocmetrie a Morrocan company specialized on Tv audience measurement, shows the effectiveness of the PLS calibration compared to the classical calibration, and the calibration based on the PCA components.
Keywords: Calibration, auxiliary information, PLS regression, calibration on PLS components, RIDGE regression.

I. Introduction

En théorie de sondages, la prise en compte de l'information auxiliaire s'avère d'une grande utilité car elle permet l'amélioration de la précision des estimateurs. Cette information auxiliaire peut être utilisée, en amont, pour la construction du plan d'échantillonnage (stratification, sondage à probabilités inégales, sondage équilibré, etc.) ou en aval au niveau du calcul des estimateurs en utilisant l'une des méthodes classiques de redressement : post-stratification, estimation par le quotient, calage sur marge. Le calage sur marge est en effet une technique qui consiste à ajuster l'échantillon de départ sur la marge des critères de contrôle, marge que l'on connait sur l'ensemble de la population (Deville et Särndal, 1992).

L'idée de base du calage est de pouvoir utiliser l'information auxiliaire non prise en compte afin d'augmenter la précision des estimateurs. Cependant l'amélioration apportée par l'estimateur de calage reste conditionnée par le choix des variables auxiliaires utilisées dans le processus de calage (El Haj Tirari, 2012). En pratique, la variance de l'estimateur par calage peut devenir importante quand on utilise un très grand nombre de variables auxiliaires, notamment en présence d'une forte multicolinéarité. Plusieurs remèdes sont utilisés dans la littérature:

- La sélection de variables selon le critère de variance (El Haj Tirari, 2012).
- La régression RIDGE qui peut également s'interpréter comme un calage pénalisé (Beaumont et Bocci, 2008).
- Le calage par réduction de la dimension basé sur les composantes principales ACP (Goga, Shehzad et Vanheuverzwyn, 2011).

Cet article a pour objet de proposer une nouvelle technique de réduction de la dimension des variables auxiliaires basée sur les composantes PLS. Ce travail comporte trois parties : la première aborde les fondements théoriques de l'estimateur par calage, la deuxième partie est dédiée à la description de l'algorithme PLS et au choix des composantes PLS dans le contexte du sondage, tandis que la troisième partie est réservée aux applications numériques sur les données du panel Marocmetrie, entreprise marocaine spécialisée dans la mesure de l'audience TV.

II. Le calage sur marges

Le calage vise à ajuster la distribution de l'échantillon suivant quelques variables présentant une relation plausible avec la variable d'intérêt et dont les valeurs sont connues au niveau de l'ensemble de la population. Pour cela on affecte des poids de redressement à tous les individus de l'échantillon afin de pouvoir construire des estimateurs linéaires des variables d'intérêts. On impose néanmoins à ces poids d'être aussi proches que possible du poids de sondage. Le calcul des nouveaux poids, se traduit par la résolution du programme de minimisation sous contrainte suivant (Deville et Särndal, 1992) :

$$\begin{cases} \min \sum_{i \in S} H\left(d_i, w_i\right) \\ \sum_{i \in s} x_{ki} w_i = X_k \quad \forall k \in \{1, ..., K\} \end{cases}$$

où $H\left(.,.\right)$ est une pseudo-distance sur $\Re$ qui peut aussi être définie par $H\left(d_i, w_i\right) = d_i G\left(\frac{w_i}{d_i}\right)$ telle que $G\left(\ \right)$ est une fonction distance convexe, les d_i sont les poids de sondages et les w_i sont les poids de calage. Nous notons dans ce qui suit $r = \frac{w}{d}$.

Dans la littérature on distingue principalement 5 fonctions de distance :

La méthode linéaire : elle est équivalente à une méthode classique d'estimation dite estimation par régression, car sa fonction distance est définie par $G(r)=\frac{1}{2}(r-1)^2,\ r\in\Re$ (Deville, Särndal et Sautrory, 1993).

La méthode raking-ratio : elle est connue également sous le nom *« Iterative Proportional Fitting »* (IPF), discriminée par la fonction $G(r)=r\log(r)-r+1\quad ,r>0$ (Ireland and Kullback, 1968).

La méthode logit : elle peut être considérée comme semblable à la méthode du raking-ratio et est basée sur la fonction :

$$G(r)=\left[(r-L)\log\left(\frac{r-L}{1-L}\right)+(U-r)\log\left(\frac{U-r}{U-1}\right)\right]\frac{1}{A}\qquad ,si\ L<r<U\ \ (et\ +\infty\ \text{sinon})$$

$$avec\ A=\frac{U-L}{(1-L)(U-1)}$$ (Deville et Särndal, 1992).

La méthode linéaire tronquée : elle prend pour fonction de distance

$G(r)=\frac{1}{2}(\text{r-1})^2\quad ,si\ L<r<U\ \ (et\ +\infty\ \text{sinon})$ (Husain 1969).

La méthode sinus-hyperbolique : méthode plus récente, elle n'est disponible que sur SAS, sa fonction distance $G_\alpha(r)=\frac{1}{2\alpha}\int_1^r sh\left[\alpha\left(t-\frac{1}{t}\right)\right]dt\quad ,\alpha>0$, elle est assez particulière et est paramétrée par un alpha contrôlant l'étendue de la distribution des poids.

La problématique de calage soulève deux questions importantes :

a) Le choix de la fonction de distance.

Il n'existe pas de réponse tranchée, mais il faut noter que cette dernière conditionne la distribution des poids calculés après le processus de redressement (Deville et Särndal, 1992).

b) Le choix et le nombre de variables de calage.

A priori, l'augmentation du nombre de variables de calage permet de contrôler davantage de variables et donc d'augmenter la précision des estimateurs (Ardilly, 1994). Cependant si le nombre de variables auxiliaires devient très grand, les contraintes seront nombreuses (équations de calage) et il est possible que la mise en œuvre opérationnelle impose une trop grande dispersion en termes de poids calculés de façon à ce que toutes les équations puissent être satisfaites simultanément, ou soulève un véritable problème de convergence de l'algorithme de calage. Le nombre de contraintes n'est pas le seul inconvénient de ce processus, car on peut rencontrer également le phénomène de multicolinéarité des variables de calage en raison de la nature des données.

Il en résulte des estimations erronées des poids de calage.

La sélection des variables de calage passe par un examen minutieux de la dispersion des poids et par le choix de la fonction de distance appropriée. Cependant, il existe d'autres solutions dans la littérature, on peut citer la réduction de la dimension des variables de calage (Goga, Shehzad et Vanheuverzwyn, 2011) et la pénalisation par la régression RIDGE (Beaumont et Bocci, 2008). Dans ce qui suit nous proposons une nouvelle technique de réduction de la dimension de la matrice de calage en se basant sur les composantes PLS.

Calage sur les composantes PLS

Le calage sur composantes PLS, comme son nom l'indique, fait référence à une combinaison entre le calage sur marge et l'algorithme PLS. La régression PLS « *Partial Least Squares régression* » tire son origine des sciences sociales (H. Wold, 1966), et devient très populaire en chimie grâce aux travaux du fils de son concepteur, Svante Wold (1983).

Elle permet de relier une ou plusieurs variables de réponse *y* à un ensemble de variables prédictives $x_{1,}$ $x_{2.....}x_{k,,}$ dans des conditions où la régression multiple fonctionne mal ou plus du tout (forte multicolinéarité, plus de variables que d'observations, présence de données manquantes). En fait, cette méthode généralise et combine les caractéristiques de l'analyse en composantes principales et de la régression multiple. Plus précisément, elle cherche des composantes, appelées *variables latentes*, liées à X (la matrice de variables explicatives) et à Y (la matrice des variables à expliquer), servant à exprimer la régression de Y sur ces variables et finalement de Y sur X.

Notons, Y le vecteur de la variable d'étude et X la matrice de dimension $n \times p$ des p variables auxiliaires quantitatives observées sur n individus. Cette dernière matrice est supposée centrée.

L'algorithme de la régression PLS peut être décrit comme suit :

Etape 1 : On construit la première composante t_1 comme une combinaison linéaire des p variables explicatives x_j.

$$t_1 = w_{11}x_1 + ... + w_{1p}x_p = Xw_1$$

Les coefficients $w_1^{'} = \left(w_{11},...,w_{12},...,w_{1p}\right)$ de cette combinaison linéaire, calculés suivant la formule $w_{1j} = \dfrac{\operatorname{cov}\left(x_j, y\right)}{\sqrt{\sum_{j=1}^{p} \operatorname{cov}^2\left(x_j, y\right)}} = \dfrac{X^{'}y}{\left\|X^{'}y\right\|}$, cherchent à résumer au mieux les variables explicatives x_j et à expliquer la variable y.

Une régression simple de y sur $t_1 : y = c_1t_1 + y_1$ est ensuite effectuée, donnant lieu au calcul de y_1 le vecteur des résidus et de c_1 le coefficient de régression dont l'expression: $c_1 = \dfrac{\operatorname{cov}\left(y, t_1\right)}{\sigma_{t_1}^2} = \dfrac{y^{'}t_1}{t_1^{'}t_1}$.

On en déduit, une première équation de régression $y = c_1w_{11}x_1 + ... + c_1w_{1p}x_p + y_1$.

Etape 2 : une deuxième composante t_2 est construite de sorte à ce qu'elle ne soit pas corrélée à t_1 et à ce qu'elle puisse bien expliquer le résidu y_1. La composante t_2 est en effet une combinaison linéaire des résidus x_{1j} des régressions simples des variables x_j sur t_1 :

$t_2 = w_{21}x_{11} + ... + w_{2p}x_{1p} = X_1w_2$ avec $w_{2j} = \dfrac{\operatorname{cov}(x_{1j}, y_1)}{\sqrt{\sum_{j=1}^{p} \operatorname{cov}^2(x_{1j}, y_1)}} \Rightarrow w_2 = \dfrac{X_1^{'}y_1}{\left\|X_1^{'}y_1\right\|}$.

Le calcul des résidus x_{1j} nécessite la réalisation de la régression linéaire de toutes les variables x_j sur t_1 : $x_j = p_{1j}t_1 + x_{1j}$ où $p_{1j} = \dfrac{\operatorname{cov}\left(x_j, t_1\right)}{\sigma_{t_1}^2} = \dfrac{X^{'}t_1}{t_1^{'}t_1}$ est le coefficient de régression. On en déduit la valeur

des résidus par une simple soustraction $x_{1j} = x_j - p_{1j}t_1$ écrite également sous forme matricielle de la façon suivante : $X_1 = X - t_1 p_1^{'}$ avec $X_1 = \left(x_{11},...,x_{1p}\right)$.

Ensuite, on effectue une régression de y sur t_1 et t_2 : $y = c_1 t_1 + c_2 t_2 + y_2$, où c_1 est le coefficient de régression de la première étape, c_2 le coefficient de la régression simple de y_1 sur t_2 et y_2 le vecteur des résidus de cette régression.

$$y_1 = c_2 t_2 + y_2 \qquad\qquad c_2 = \frac{\mathrm{cov}\left(y_1, t_2\right)}{\sigma_{t_2}^2} = \frac{y_1^{'} t_2}{t_2^{'} t_2}$$

Etapes suivantes : cette procédure itérative peut être poursuivie en utilisant de la même manière les résidus y_2 et $x_{21},...,x_{2p}$.

Critère d'arrêt : les composantes PLS sont choisies sur la base du critère AIC (Akaike, 1973). En effet, l'augmentation brusque de la valeur du critère implique un arrêt du processus d'itération, car elle traduit la non significativité de la dernière composante incorporée dans le modèle de régression. Les composantes PLS calculées lors des itérations qui précédent le stade d'arrêt sont alors les seules retenues.
Les premières composantes PLS étant choisies, un calage classique sera fait à la lumière de ce qui est fait dans le cas du calage sur composantes principales. La procédure de calage est achevée par la récupération des poids de calage selon la fonction distance choisie.

Application

Pour mettre en œuvre le calage sur composantes PLS, nous avons utilisé les données de l'enquête de cadrage réalisée par Marocmetrie, société spécialisée dans la mesure de l'audience TV. La base de données ayant servi au calage comporte 26 variables observées sur 4800 individus, dont 6 variables caractérisent les foyers et le reste caractérise les individus. En vue de comparer les résultats avec la méthode de calage sur les composantes ACP, seules les variables quantitatives ont été sélectionnées.
En respectant les démarches de la méthode PLS décrites précédemment, nous avons appliqué la régression PLS entre la variable d'intérêt (y) qui est l'audience de télévision et la matrice X des variable de calage de taille 4800 x 17 à l'aide du logiciel R. Deux composantes PLS résumant toute l'information contenue dans les variables auxiliaires ont été extraites sur la base du critère AIC.

En vue de montrer l'intérêt d'une telle technique de calage, les résultats obtenus ont été comparés à la méthode de calage sur les composantes principales ACP (Goga, Shehzad et Vanheuverzwyn, 2011) et à la méthode de calage classique basée sur la régression (Deville et Särndal, 1992). Les résultats des trois techniques de calage, selon le choix de la fonction distance, sont résumés dans le tableau et les graphiques ci-après.

Tableau résumant les variations des coefficients de redressement

Fonction distance		Méthode de calage	Maximum des coefficients de redressement	Minimum des coefficients de redressement	Longueur de l'intervalle de variation des coefficients de redressement	Ecart- type
Linéaire		classique	6,52	-0,97	7,49	0,60
		sur composantes ACP	6,27	-0,85	7,12	0,54
		sur composantes PLS	4,82	-1,06	5,88	0,46
Logit	Bornes [0,13]	classique	9,36	0,01	9,35	0,63
		sur composantes ACP	10,46	0,1	10,36	0,58
		sur composantes PLS	8,84	0,09	8,75	0,49
	Bornes [0,5]	sur composantes ACP	4,99	0,05	4,94	0,56
		sur composantes PLS	4,81	0,06	4,75	0,47
	Bornes [0.2,5]	sur composantes PLS	3	0,2	2,8	0,48
Raking-ratio		sur composantes ACP	15,55	0,12	15,43	0,62
		sur composantes PLS	13,1	0,12	12,98	0,53
Sinus-hyperbolique		sur composantes PLS	3,59	0,31	3,28	0,49

Cas de la fonction de distance linéaire :

Dans le cas du calage sur composante PLS, la dispersion des poids est plus faible (0,46) et présente une distribution d'allure presque gaussienne (Figure1) en comparaison de la méthode du calage sur composantes principales qui présente une dispersion de (0,54) et la méthode de calage par régression ordinaire.

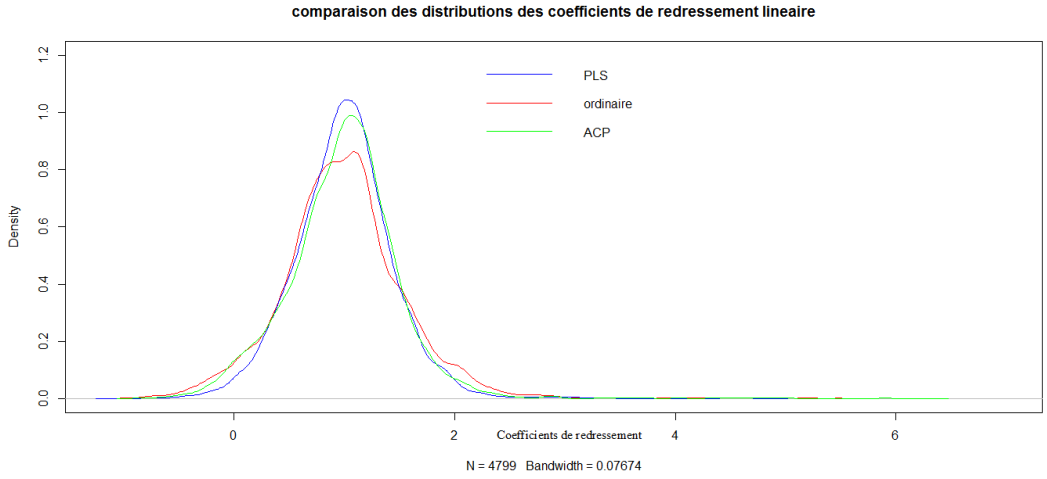

Figure 1 : Cas de la distance linéaire

Cas des fonctions de distances bornées (Logit)

Dans le cas de l'intervalle [0,13], le calage sur composante PLS présente une faible distribution des poids (0,49) et est d'allure presque normale (Figure2) en comparaison avec la méthode par calage sur composantes principales (0,58) et la méthode de calage ordinaire (0,63)

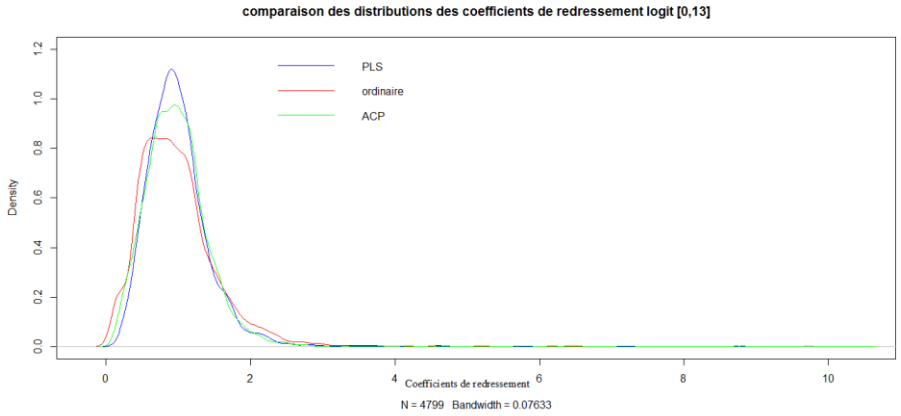

Figure 2 : Cas de la distance bornée logit [0 ;13]

Si on réduit l'intervalle de variation à [0 ; 0,5], la PLS et la méthode par composante ACP conduisent à une dispersion de poids plus faible mais deviennent presque équivalentes à la méthode par calage sur composantes principales (la méthode ordinaire ne converge pas). Par contre si on réduit considérablement l'intervalle de variation à [0,2 ; 3], seule la méthode PLS converge.

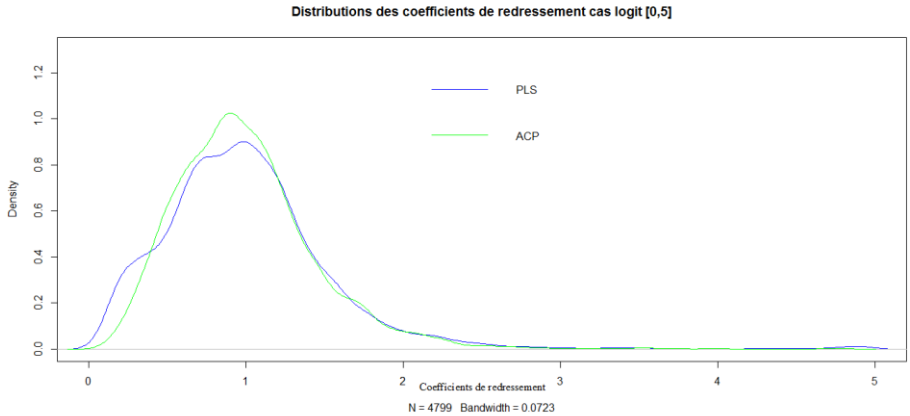

Figure 3 : Cas de la distance bornée logit [0 ;5]

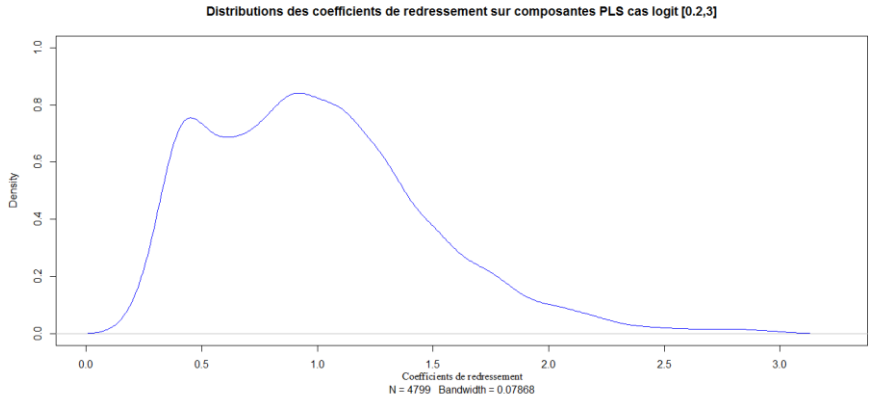

Figure 4 : Cas de la distance bornée logit [0.2 ;3]

Cas de la fonction de distance raking ratio

Dans le cas du calage sur composantes PLS, la distribution des poids est d'allure normale avec une dispersion très faible (0,53) en comparaison de la méthode de régression sur composantes principales, qui présente une dispersion de 0,62. Ces deux méthodes restent pratiquement équivalentes alors que la méthode ordinaire ne converge pas.

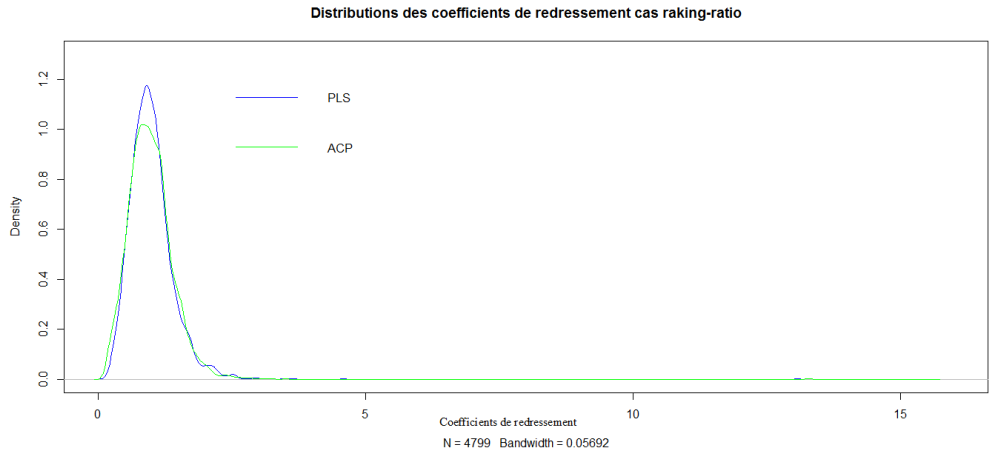

Figure 5 : Cas de la distance raking-ratio

Cas du sinus hyperbolique :

La supériorité de l'approche PLS est bien prouvée, la dispersion de distribution reste faible (0,49) et les autres méthodes ne convergent même pas.

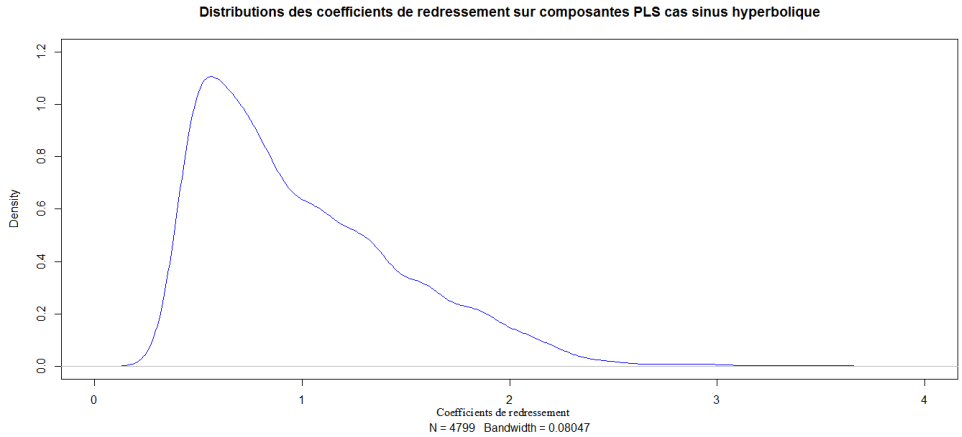

Figure 6 : Cas de la distance sinus-hyperbolique

Conclusion et discussion.

Ces résultats appellent les remarques suivantes :

Quel que soit le choix de la fonction de distance (linéaire, raking-ratio, logit ou sinus hyperbolique), la convergence du redressement a été bien assurée dans le cas de la régression PLS alors que la méthode de calage sur composantes principales n'a pas pu aboutir à une solution dans le cas de la distance sinus hyperbolique. Tandis que le calage ordinaire ne donne des résultats que dans le cas de les fonctions de distance linéaire et logit. Les observations à ce sujet ne s'arrêtent pas là mais il parait que dans le cas des fonctions de distances bornées, comme la fonction logit, la PLS réduit

considérablement l'intervalle de variation des coefficients de redressement par rapport à la situation ordinaire et celle intégrant l'ACP.

En outre, il parait que le calage sur composantes PLS et le calage sur composantes ACP tendent à normaliser la distribution des coefficients de redressement pour approcher soit la loi gaussienne ou la loi log-normale. Dès lors, l'explication la plus adéquate à cette situation est de dire que les perturbations causées par la multicolinéarité des variables auxiliaires ont pu être éliminées en faveur de l'utilisation de la régression PLS ou de l'ACP tout en gardant la quasi-totalité de l'information auxiliaire.

Limites et recommandations

Il reste à comparer les résultats de PLS avec la méthode dite de calage pénalisé.

Bibliographie

[1] Akaike H. (1973), Information theory and an extension of the maximum Likelihood Principle, in *Selected Papers of Hirotugu Akaike, Springer (1998), Parzen E., Tanabe K., Kitagawa G. eds.*

[2] Ardilly P. (1994), Les Techniques de Sondages, *Edition Technip*, Paris

[3] Beaumont J-F., Bocci C. (2008), Another look at ridge calibration, *International Journal of Statistics, vol. LXVI, n. 1, pp. 5-20*

[4] Deville, J.C, and Särndal, C-E (1992), Calibration estimators in survey sampling, *Journal of American Statistical Association*, Vol. 87, pp. 376-382

[5] Deville J-C., Särndal, C-E., Sautory,O. (1993), Generalized raking procedures in survey sampling, *Journal of American Statistical Association*, Vol. 88, pp. 1013-1020

[6] El Haj Tirari M. (2012), Critère du choix des variables auxiliaires à utiliser dans l'estimateur par calage, *Septième colloque francophone sur les sondages*, Rennes

[7] Goga C., Shehzad M.-A., et Vanheuverzwyn A. (2011), Principal component regression with survey data application on the French media audience, *Proceedings of the 58th ISI World Statistics Congress, Dublin*

[8] Ireland C.T. and Kullback S. (1968), Contingency tables with given marginals, *Biometrika*, Vol. 55, 1, pp. 179-188

[9] Husain. M (1969), construction of Regression Weights for estimation in sample surveys, *Thèse de maitrise non publiée, Iowa State University*, Iowa.

[10] Tenenhaus M. (1998), La régression PLS : Théorie et pratique, *Editions Technip*, Paris

[11] Wold. H (1966), Estimation of principal components and related models by iterative least squares, *In P.R.Krishnaiaah (ed.) Multivariate Analysis, Academic Press, New York.*

[12] Wold. S, Martens. H, Wold. H (1983), The multivariate calibration problem in chemistry solved by the PLS method, Springer Berlin Heidelberg Ed.