Efficacité relative du calage PLS par rapport aux calages Ridge et ACP

Sara Nahchel [1] & Jelloul Allal [2] & Zoubir Zarrouk [3]

[1] *Faculté des Sciences, Université Mohamed 1er, BV Mohammed VI BP 717, 60000, Oujda, Maroc - s.nahchel@ump.ac.ma*

[2] *Faculté des Sciences, Université Mohamed 1er, BV Mohammed VI BP 717, 60000, Oujda, Maroc - jell.allal@gmail.com*

[3] *Faculté des Sciences Juridiques, Economiques et Sociales, Université Mohamed 1er, Complexe universitaire, Hay Al Qods BP 724, Oujda, Maroc - zoubire@hotmail.com*

Résumé. Le calage en tant que méthode d'ajustement des échantillons permet d'améliorer la précision des estimateurs. Néanmoins, la présence de forte multicolinéarité entre les variables de calage engendre une perte de precision, voir même une impossibilité de calculer les poids de calage. En vue de remédier à ce problème, plusieurs techniques ont été proposées dans la littérature: le calage pénalisé (Barseley et Chambers, 1983), le calage sur composantes principales (Goga,Shehzad et Vanheuverzwyn, 2011) et plus récemment, le calage sur composantes PLS (Nahchel et al., 2014). Dans ce papier nous étudions l'efficacité relative du calage sur composantes PLS par rapport aux autres techniques citées précédemment.

Mots-clés. Multicolinéarité, calage sur composantes PLS, calage sur composante principales et calage pénalisé.

Introduction

Le calage vise à ajuster la distribution d'un échantillon s de taille n suivant quelques variables présentant une relation plausible avec la variable d'intérêt. Pour ce faire, des poids w_k sont affectés à chaque individu. Le calcul de ces poids, s'effectue par résolution du programme de minimisation sous contrainte suivant (Deville et Särndal, 1992) :

$$\begin{cases} \min \sum_{k\in s} G_k(w_k, d_k) \\ \sum_{k\in s} w_k x_k = \sum_{k\in U} x_k \end{cases} \tag{1}$$

où G_k est la fonction de distance définie pour chaque $w > 0$ ainsi que vérifiant des propriétés proposées par Deville et Särndal (1992), les w_k sont les poids de calage, les d_k sont les poids de sondage utilisé, $x_k = (x_{k1}, ..., x_{km})'$ est le vecteur comportant les valeurs des variables auxiliaires pour le $k$ième individu et U est la population de taille N.

L'équivalence asymptotique entre l'estimateur de calage et l'estimateur par régression (Deville et Särndal, 1992) permet d'écrire les poids de calage sous la forme:

$$w_k = d_k + d_k q_k x_k^{'} \left(\sum_{k \in s} d_k q_k x_k x_k^{'} \right)^{-1} \left(\sum_{k \in U} x_k - \sum_{k \in s} d_k x_k \right) \tag{2}$$

avec $q_k = \dfrac{1}{\sigma_k^2}$ et σ_k^2 est la variance de la variable d'intérêt y_k suivant le modèle de régression $\xi \ : \ y = X\beta + \varepsilon$.

Par la suite l'estimateur par calage est donnée par:

$$\hat{t}_y = \sum_{k \in s} w_k y_k = \sum_{k \in s} d_k y_k + \left(\sum_{k \in U} x_k - \sum_{k \in s} d_k x_k \right) \hat{\beta} \tag{3}$$

où $\hat{\beta} = \left(\sum_{k \in s} d_k x_k x_k^{'} \right) \sum_{k \in s} d_k q_k x_k y_k$.

La pécision de $\hat{t}_y$ est évaluée suivant l'estimateur de la variance asymptotique qui suit:

$$\hat{AV} = \sum_{l \in s} \sum_{k \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \left(\frac{y_k - x_k^{'} \hat{\beta}_{OLS}}{\pi_k} \right) \left(\frac{y_l - x_l^{'} \hat{\beta}_{OLS}}{\pi_l} \right) \tag{4}$$

tel que $\pi_k = Prob(k \in s)$ et $\pi_{kl} = Prob(k, l \in s)$.

En présence de multicolinéarité, l'estimateur de calage perd sa précision et le calcul des poids devient impossible. De ce fait, plusieurs remèdes ont été proposés dans la litterature: le calage pénalisé (Bardsley et Chambers, 1983), le calage sur composantes principales (Goga, Shehzad et Vanheuverzwyn, 2011) et plus récemment le calage sur composantes PLS (Nahchel et al., 2014). Après une étude des propriétés statistiques des différentes méthodes de calage, nous testons l'efficacité relative du calage PLS par rapport au calage pénalisé et au calage sur composantes principales et nous confirmons les résultats sur des données réelles fournit gracieusement par la société Marocmetrie en 2014.

1 Calage pénalisé

Le calage pénalisé (Bardsley et Chambers, 1983), se distingue du calage ordinaire par la pénalisation des contraintes. Cette pénalisation a pour objet de corriger le mauvais conditionnement de la matrice des variables auxiliaires. En se réferant à Beaumont et Bocci (2008), le calage pénalisé se définit par la minimisation de

$$\sum_{k \in s} G_k(w_k, d_k) + \frac{(1 - \lambda^*)}{\lambda^*} \left(\sum_{k \in s} w_k x_k - \sum_{k \in U} x_k \right)^{'} C \left(\sum_{k \in s} w_k x_k - \sum_{k \in U} x_k \right) \tag{5}$$

où $0 \leq \lambda^* < 1$ et $C = diag(c_1, ..., c_m)$ est la matrice des coûts avec $c_i \geq 0 \ \ \forall i = 1...m$.
Les poids de calage suivant cette méthode ont pour expression:

$$w_{Ridge,k} = d_k + d_k q_k x_k^{'} \left(\sum_{k \in s} d_k q_k x_k x_k^{'} + \frac{\lambda^*}{1-\lambda^*} C^{-1} \right)^{-1} \left(\sum_{k \in U} x_k - \sum_{k \in s} d_k x_k \right) \quad (6)$$

La précision de l'estimateur de calage qui en résulte est estimé par l'expression de la variance asymptotique suivante:

$$\hat{AV}_{Ridge} = \sum_{l \in s} \sum_{k \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \left(\frac{y_k - x_k^{'} \hat{\beta}_{Ridge}}{\pi_k} \right) \left(\frac{y_l - x_l^{'} \hat{\beta}_{Ridge}}{\pi_l} \right) \quad (7)$$

avec $\hat{\beta}_{Ridge} = \left(\sum_{k \in s} \frac{x_k x_k^{'}}{\sigma_k^2 \pi_k} + \lambda C^{-1} \right)^{-1} \sum_{k \in s} \frac{x_k y_k}{\sigma_k^2 \pi_k}$.

Quant au biais suivant le modèle ξ est exprimé par (Shehzad, 2012):

$$\begin{aligned} Bias_\xi(\hat{Y}_{Ridge}) &= E_\xi \left(\sum_{k \in s} w_{Ridge,k} y_k - \sum_{k \in U} y_k \right) \\ &= -\lambda \left(\sum_{k \in U} x_k^{'} - \sum_{k \in s} \frac{x_k^{'}}{\pi_k} \right) \left(\sum_{k \in s} d_k q_k x_k x_k^{'} + \frac{\lambda^*}{1-\lambda^*} C^{-1} \right)^{-1} C^{-1} \beta \quad (8) \end{aligned}$$

2 Calage sur composantes principales

Le calage sur composante principales (Goga, Shehzad et Venheuverzwyn, 2011) permet également de corriger la multicolinéarité entre les variables auxiliaires en remplaçant ces dernières par un nombre r adéquat des composantes principales relatives à l'information auxiliaire. Ainsi, en notant par $z_k = (z_{k1}, ..., z_{kr})^{'}$ le vecteur contenant les valeurs des r premières composantes principales pour le $k$ième individu, les poids de calage dans ce cas sont donnés par:

$$w_{PC,k} = d_k + d_k q_k z_k^{'} \left(\sum_{k \in s} d_k q_k z_k z_k^{'} \right)^{-1} \left(\sum_{k \in U} z_k - \sum_{k \in s} d_k z_k \right) \quad (9)$$

La précision de l'estimateur de calage $\hat{Y}_{pc} = \sum_{k \in s} w_{pc,k} y_k$ est mesurée par l'estimateur de la variance asymptotique qui suit:

$$\hat{AV}_{PC} = \sum_{l \in s} \sum_{k \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \left(\frac{y_k - z_k^{'} \hat{\beta}_{PC}}{\pi_k} \right) \left(\frac{y_l - z_l^{'} \hat{\beta}_{PC}}{\pi_l} \right) \quad (10)$$

avec $\hat{\beta}_{PC} = \left(\sum_{k\in s} \frac{z_k z_k^{'}}{\sigma_k^2 \pi_k}\right)^{-1} \sum_{k\in s} \frac{z_k y_k}{\sigma_k^2 \pi_k}$.

En posant $X = (x_k^{'})_{k\in U}$, il existe une matrice A tel que $Z_m = XA$.

Soit A_r le restriction de A sur les r premières colonnes. Alors, le biais de $\hat{Y}_{pc}$ sous le modèle ξ est donné par (Shehzad, 2012):

$$Bias(\hat{Y}_{pc}) = -\left(\sum_{k\in s} \frac{z_k^{'}}{\pi_k}\right)\left(A_r^{'} - A^{'}\right)\beta \tag{11}$$

3 Calage sur composante PLS

Le calage sur composantes PLS (Nahchel et al., 2016) consiste à caler sur les v premières composantes PLS. Ces dernières permettent de résumer l'information fournie par les variables auxiliaires tout en exploitant la relation entre celles-ci et la variable d'intérêt. En outre, cette méthode de calage présente l'avantage de traiter un bloc de variables d'intérêt simultanément par rapport à un bloc de variables auxiliaires (PLS2 De Jong, 1993)

Soit $l_k = (l_{k1}, ..., l_{kv})^{'}$ le vecteur contenant les valeurs des v composantes PLS pour le $k$ième individu. Alors, la minimisation de

$$\begin{cases} \min \sum_{k\in s} G_k(w_k, d_k) \\ \sum_{k\in s} w_k l_k = \sum_{k\in U} l_k \end{cases} \tag{12}$$

fournit des poids de calage ayant pour expression:

$$w_{PLS,k} = d_k + d_k q_k l_k^{'} \left(\sum_{k\in s} d_k q_k l_k l_k^{'}\right)^{-1} \left(\sum_{k\in U} l_k - \sum_{k\in s} d_k l_k\right) \tag{13}$$

L'expression matricielle de la formule précédente s'écrit comme suit:

$$w_{PLS} = d_s^{'} + \left(1_U^{'} L_v - d_s^{'} L_{v,s}\right)\left(L_{v,s}^{'} \tilde{\Pi}_s^{-1} L_{v,s}\right)^{-1} L_{v,s}^{'} \tilde{\Pi}_s^{-1} \tag{14}$$

où $d_s = (d_1, ..., d_n)^{'}$, $1_U = (\underbrace{1, ..., 1}_{N \quad fois})^{'}$, $L_v = (l_k^{'})_{k\in U}$ tel qu'il existe une matrice M_v vérifiant $L_v = XM_v$, $L_{v,s} = (l_k^{'})_{k\in s} = X_s M_v$ et $\tilde{\Pi}_s = diag(\sigma_k^2 \pi_k)_{k\in s}$.

L'estimateur de calage sera de la forme:

$$\hat{Y}_{PLS} = w_{PLS} y_s$$

$$= d_s^{'} y_s + \left(1_U^{'} L_v - d_s^{'} L_{v,s}\right) \left(L_{v,s}^{'} \tilde{\Pi}_s^{-1} L_{v,s}\right)^{-1} L_{v,s}^{'} \tilde{\Pi}_s^{-1} y_s$$

$$= \hat{Y}_{HT} + \left(1_U^{'} L_v - d_s^{'} L_{v,s}\right) \hat{\beta}_{PLS} \tag{15}$$

avec $y_s = (y_k)_{k\in s}^{'}$, $\hat{Y}_{HT} = d_s^{'} y_s$ est l'estimateur d'Horvitz Thompson (Horvitz et Thompson, 1952) et $\hat{\beta}_{PLS} = \left(L_{v,s}^{'} \tilde{\Pi}_s^{-1} L_{v,s}\right)^{-1} L_{v,s}^{'} \tilde{\Pi}_s^{-1} y_s$.

Autrement dit, l'estimateur de calage qui en résulte s'écrit sous la forme d'un estimateur par régression (Särndal, Swensson et Wretman, 1992). Par conséquent, l'estimteur par calage sur composantes PLS est caractérisé par les propriétés statistiques suivantes:

1- si$\dfrac{m}{n} \to 0$, la variance asymptotique est donné par

$$AV_{PLS} = \sum_{k\in U} \sum_{i\in U} \frac{\pi_{ki} - \pi_k \pi_h}{\pi_{ki}} \left(\frac{y_k - l_k^{'} \hat{\beta}_{PLS}}{\pi_k}\right) \left(\frac{y_i - l_i^{'} \hat{\beta}_{PLS}}{\pi_i}\right) \tag{16}$$

et est estimée par:

$$\hat{AV}_{PLS} = \sum_{k\in s} \sum_{i\in s} \frac{\pi_{ki} - \pi_k \pi_i}{\pi_{ki}} \left(\frac{y_k - l_k^{'} \hat{\beta}_{PLS}}{\pi_k}\right) \left(\frac{y_i - l_i^{'} \hat{\beta}_{PLS}}{\pi_i}\right) \tag{17}$$

2- Le biais sous le modèle ξ de $\hat{Y}_{PLS}$ est exprimé par:

$$E_\xi \left(\hat{Y}_{PLS} - t_y\right) = \left(1_U^{'} L_v - d_s^{'} L_{v,s}\right) \left(M_v^{'} - M^{'}\right) \beta \tag{18}$$

4 Efficacité relative du calage sur composantes PLS

On sait par ailleurs que les estimateurs issus des trois méthodes de calage sont tous ξ-biaisés. Par contre les estimateurs issus de la régression multiple sont non biaisés. Par conséquent l'efficacité des estimateurs des trois méthodes sera comparée par rapport aux estimateurs de la régression multiple.

- L'utilisation du résultat démontré par De Jong (1993) et Hoog et Phatak (2002) qu'est

$$\|L_{v,s} \hat{\beta}_{PLS} - X_s \hat{\beta}\| \leq \|Z_{r,s} \hat{\beta}_{PC} - X_s \hat{\beta}\| \tag{19}$$

ainsi que des formules (10) et (17) permettent de déduire que

$$\hat{AV}(\hat{Y}_{PLS}) \leq \hat{AV}(\hat{Y}_{pc}) \tag{20}$$

- D'une autre part, le calcul de $L_{v,s}\hat{\beta}_{PLS}$ en remplaçant $\hat{\beta}_{PLS}$ par son expression aboutit à

$$\begin{aligned} L_{v,s}\hat{\beta}_{PLS} &= X_s M_v \hat{\beta}_{PLS} \\ &= X_s M_v \left(L'_{v,s}\tilde{\Pi}_s^{-1}L_{v,s}\right)^{-1} L'_{v,s}\tilde{\Pi}_s^{-1}y_s \\ &= X_s M_v \left(M'_v X'_s \tilde{\Pi}_s^{-1} X_s M_v\right)^{-1} M'_v X'_s \tilde{\Pi}_s^{-1} y_s \end{aligned}$$

  Soit $(^+)$ signe du pseudo-inverse. Puisque le pseudo-inverse de M_v est supposé existant suite à l'algorithme proposé par De Jong (1993) pour la régression PLS, on trouve que

$$\begin{aligned} L_{v,s}\hat{\beta}_{PLS} &= X_s M_v \left(M'_v X'_s \tilde{\Pi}_s^{-1} X_s M_v\right)^{+} M'_v X'_s \tilde{\Pi}_s^{-1} y_s \\ &= X_s \left({M'_v}^{+} M'_V X'_s \tilde{\Pi}_s^{-1} X_s M_v M_v^{+}\right)^{+} X'_s \tilde{\Pi}_s^{-1} y_s \\ &= X_s \left(X'_s \tilde{\Pi}_s^{-1} X_s\right)^{-1} X'_s \tilde{\Pi}_s^{-1} y_s \\ &= X_s \hat{\beta} \end{aligned}$$

  Hoerl et Kennard (1970), ont montré que $\hat{\beta}_{Ridge} = \left(I_m + \lambda (X'_s X_s)^{-1}\right)^{-1} \hat{\beta}$.

  Alors, $$X_s \hat{\beta}_{Ridge} = X_s \left(I_m + \frac{\lambda^*}{1-\lambda^*}(X'_s X_s)^{-1}\right)^{-1} \hat{\beta}$$

  Par la suite,

$$\|L_{p,s}\hat{\beta}_{PLS} - X_s\hat{\beta}\| \leq \|X_s\hat{\beta}_{Ridge} - X\hat{\beta}\| \tag{21}$$

  D'où,

$$\|L_{p,s}\hat{\beta}_{PLS} - y\| \leq \|X_s\hat{\beta}_{Ridge} - y\| \tag{22}$$

  Les formules (7) et (17) permettent ainsi de conclure que

$$\hat{AV}(\hat{Y}_{PLS}) \leq \hat{AV}(\hat{Y}_{Ridge}) \tag{23}$$

En Conséquence des formules (20) et (23), le calage sur composantes PLS est bien meilleur en terme de précision que le calage sur composantes principale et que le calage pénalisé.

5 Simulation

En vue de confirmer l'efficacité du calage sur composantes PLS par rapport aux autres techniques (Ridge et ACP), nous nous sommes basé sur des données réelles de taille 10121 fournie par la société Marocmetrie, qui était chargé des mesures de l'audience TV

sur le territoire Marocain en 2014. La base de données comporte 23 variables auxiliaires (6 qualitatives et 17 quantitatives) représentant des informations personnelles de chaque individu du panel.

Afin d'étudier les propriétés statistiques des estimateurs issus des différentes techniques de calage, nous avons décidé de travailler sur un échantillon de 1013 individus (soit 10% de la population). Cet échantillon a été extrait 3000 fois par tirage alétoire simple.

$$Soit, \qquad \pi_k = \frac{n}{N} \quad \forall\ k \in s \qquad et \qquad \pi_{kl} = \frac{n(n-1)}{N(N-1)} \quad if\ k \neq l$$

En conséquence, les expressions de la variance asymptotique des différentes techniques de calage (7) , (10) et (17) deviennent

$$\hat{AV}_{Ridge} = \sum_{k\in s} \frac{N}{n}\left(\frac{N-n}{n}\right)\left(y_k - x_k^{'}\hat{\beta}_{Ridge}\right)^2 + \sum_{k\neq l}\frac{N}{n}\left(\frac{n-N}{n(n-1)}\right)\left(y_k - x_k^{'}\hat{\beta}_{Ridge}\right)\left(y_l - x_l^{'}\hat{\beta}_{Ridge}\right) \tag{24}$$

$$\hat{AV}_{PC} = \sum_{k\in s} \frac{N}{n}\left(\frac{N-n}{n}\right)\left(y_k - z_k^{'}\hat{\beta}_{PC}\right)^2 + \sum_{k\neq l}\frac{N}{n}\left(\frac{n-N}{n(n-1)}\right)\left(y_k - z_k^{'}\hat{\beta}_{PC}\right)\left(y_l - z_l^{'}\hat{\beta}_{PC}\right) \tag{25}$$

et

$$\hat{AV}_{PLS} = \sum_{k\in s} \frac{N}{n}\left(\frac{N-n}{n}\right)\left(y_k - l_k^{'}\hat{\beta}_{PLS}\right)^2 + \sum_{k\neq i}\frac{N}{n}\left(\frac{n-N}{n(n-1)}\right)\left(y_k - l_k^{'}\hat{\beta}_{PLS}\right)\left(y_i - l_i^{'}\hat{\beta}_{PLS}\right) \tag{26}$$

Le tableau ci-après, résume les résultats de la simulation et fournit la moyenne des estimations calculées sur les 3000 répliques.

Ces résultats, confirment la performance du calage sur composantes PLS par rapport

Table 1: Tableau des résultats de la simulation

	Ridge calibration	PC calibration	PLS-calibration
Total estimation	114333	114326	114342
Variance of the total estimator	11180522.45	9851007.58	9845052.59
Bias	-5	-12	4
Relative Bias	-3.5	-0.0001	0.00004
Weights variance	2.8	2.7	2.7

au calage sur composantes principales et au calage pénalisé. En effet, la variance de $\hat{Y}_{PLS}$ est en moyenne plus petite que la variance de $\hat{Y}_{PC}$ qui est à son tour plus petite que la variance de $\hat{Y}_{Ridge}$ ($\hat{AV}(\hat{Y}_{PLS}) < \hat{AV}(\hat{Y}_{PC}) < \hat{AV}(\hat{Y}_{Ridge})$). Aussi, le biais de $\hat{Y}_{PC}$ est plus petit que le biais de $\hat{Y}_{PC}$ qui est à son tour plus petit que le biais de $\hat{Y}_{Ridge}$ ($R.Bias(\hat{Y}_{PLS}) < R.Bias(\hat{Y}_{PC}) < R.Bias(\hat{Y}_{Ridge})$) . En outre la variance des poids du calage sur composantes PLS et du calage sur composantes principales est plus petites de la variance des poids du calage pénalisé ($Var(W_{PLS}) = Var(W_{PC}) < Var(W_{Ridge})$).

Conclusion

Les résultats empiriques de la simulation permettent de confirmer les résultats théoriques relatives à la comparaison des estimateurs issus des trois techniques. Il en résulte que le calage sur composantes PLS permet d'estimer le total de la variable d'intérêt avec plus de précision tout en conservant le plus possible la structure initiale de l'échantillon. Nous pensons que le calage sur composantes PLS reste le meilleur remède au traitement de la multicolinéarité en théorie de sondages.

References

[1] Nahchel, S., Allal, J., Zarrouk, Z. and Alami, Y. (2014), 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, Huitième colloque francophone sur les sondages, Bourgogne-Dijon.

[2] Bardsley, P. and Chambers, R.L. (1984), Multipurpose estimation from unbalanced samples, Applied Statistics, Volume 33, 290-299.

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

[4] Chun H. and Keleş S. (2010), Sparse partial least squares regression for simultaneous dimension reduction and variable selection, Journal of the Royal Statistical Society, Volume 72, Part 1, 3-25.

[5] De Hoog, F. and Phatak, A. (2002), Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS, Chemometrics Journal, Volume 16, 361-367.

[6] De Jong, S. (1993), PLS fits closer than PCR, Chemometrics Journal, Volume 7, 551-557.

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

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

[9] Hoerl, A.E. and Kennard, R.W. (1970), Ridge regression: biased estimation for nonorthogonal problems, Technometrics, Volume 12, 55-67.

[10] Horvitz, D.G. and Thompson, D.J. (1952), Ageneralization of sampling without replacement from a finite universe. Jounal of the American Statistical Association, Volume 47, 663-685.

[11] Särndal, C.E., Swensson, B. and Wretman, J. (1992), Model-assisted Survey Sampling, Springer-Verlag, New York.

[12] Shehzad, M.A. (2012), Pénalisation et réduction de la dimension des variables auxiliaires en théorie de sondages, Thesis of doctorat grade.

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