

## CALAGES SUCCESSIFS, CALAGES ITERES

Jean-Claude Deville<sup>1</sup>

### RÉSUMÉ

Un estimateur linéaire sera dit calé sur un ensemble de variables si leurs totaux sont estimés 'parfaitement', c'est à dire avec une variance nulle. Ceci peut être réalisé au stade de l'échantillonnage (stratification, équilibrage) ou par une procédure de calage 'classique' ou utilisant certains instruments. Il arrive souvent qu'on utilise plusieurs calages successifs dans une procédure d'estimation (stratification puis poststratification par exemple). Or le second calage détruit partiellement, voire complètement, les améliorations de précision recherchées par le premier. On peut même obtenir une détérioration, comme certains exemples le montreront.

On peut aussi être tenté d'itérer les calages en recommençant la procédure sur le premier ensemble de variables, puis le second en s'attendant à ce que la convergence du processus amène à bénéficier de l'information apportée par les deux ensembles de variables auxiliaires. C'est ce qui se passe avec l'algorithme classique du raking-ratio.

On établit une condition suffisante simple pour qu'il en soit ainsi. Une condition nécessaire et suffisante plus complexe mais facile à vérifier sur les données permet de déceler les cas où les choses ne se passent pas bien: les itérations ne convergent pas et la variance générée par les systèmes de poids successifs peut devenir infinie.

MOTS CLÉS : Calage; calage généralisé, pondérations; raking-ratio.

### ABSTRACT (12 points)

A linear estimator is said to be calibrated on a group of variables if their totals are estimated "perfectly", that is that the variance is null. This can be performed at the sampling stage (stratification r balanced sampling) or by a classical calibration procedure using certain tools. It is often the case that we use many successive calibrations in the estimation procedure (stratification and post-stratification for example). In this case, the second calibration can partially or completely remove the gains acheived in the first. As several examples show, it can even result in a deterioration of the first calibration.

We can also attempt to iterate teh calibrations by starting the procedure on teh first set of variables, and then the second expecting that the convergence of the two processes will lead to improvements from the two groups of auxiliary variables. This is what happens with the classic raking-ratio algorithm.

KEY WORDS: Calibrations, Generalized calibration, Raking ratio, Weighting.

### 1. INTRODUCTION : POSITION DES PROBLÈMES

Il arrive souvent, dans la pratique des enquêtes, qu'on soit amené à modifier les poids d'extrapolation plusieurs fois de suite. Généralement les poids initiaux sont ceux de Horvitz-Thompson, inverses des probabilités d'inclusion. Ce n'est pas obligatoire cependant. On peut, par exemple, utiliser des poids sans biais issus à d'un échantillonnage (méthode de partage des poids), voire des poids déjà ajustés par un procédé quelconque. On considérera que ces poids sont calés sur certaines variables s'ils permettent une extrapolation parfaite du total de ces variables, c'est à dire de variance nulle ou négligeable. Cette propriété résulte généralement d'une stratification ou d'un échantillonnage équilibré et concerne des quantités simples comme la taille de la population (total de la variable 'unité'), la taille des strates ou parfois d'autres variables. Pour tenir compte d'informations supplémentaires données sous la forme de totaux de variables dites auxiliaires, on modifie ces poids (calage) de façon à annuler la variance d'estimation des totaux auxiliaires. Assez souvent cette phase est précédée d'ajustements (parfois plusieurs) destinés à corriger les biais dus à la non-réponse. On assimilera ces ajustements à une phase de calage, ce qui est justifié pour la majorité des cas pratiques.

---

<sup>1</sup> Jean-Claude Deville, ENSAI/CREST, Laboratoire de Statistique d'Enquête, Campus de Ker-Lann-35170-BRUZ, France; deville@ensai.fr

Chacune de ces modifications consiste, à partir des poids courants, éventuellement obtenus par des calages antérieurs, à se caler sur une nouvelle information auxiliaire, en utilisant, éventuellement, certaines variables comme ‘instruments’ du calage. Malheureusement, sauf cas exceptionnel, ces nouveaux poids ne sont plus calés sur les variables auxquelles les poids initiaux étaient ajustés. Autrement dit la variance de ces variables est passée de zéro à une quantité positive, ce qui implique que la précision de certaines variables d’intérêt a pu diminuer à cause de cette opération.

Exemple 1: Un exemple typique et simple est celui d’un sondage aléatoire simple où la somme des poids (initiaux) est calée sur le total  $N$  de population. Si on dispose d’une information  $t_X = \sum_U x_k$ , total sur la population d’une auxiliaire  $x$ , on peut utiliser un estimateur par ratio dont les poids sont  $w_1 = t_X / \hat{t}_X w_0$  ( pour alléger les notations on supprimera, dorénavant, les indices  $k$  des individus quand le contexte le permet) avec  $\hat{t}_X = \sum_s x w_0$ . La somme des nouveaux poids vaut maintenant  $N t_X / \hat{t}_X$ , le calage sur la variable ‘unité’ est détruit et l’estimation du total de la population devient incertain.

Cette situation est générale. On se propose d’évaluer les modifications que subissent les précisions de variables de calage  $U$  (à  $q$  composantes) et  $X$  (à  $p$  composantes) pour des calages successifs. Ceci prend plusieurs aspects : évolution des pondérations, qualité du calage ‘résiduel’ et variance d’estimation obtenue pour le total d’une variable d’intérêt quelconque  $y$ .

On regardera aussi ce qui se passe quand on itère la procédure de calage : partant de poids  $w_0$  calés sur  $U$ , on obtient des poids calés sur  $X$ ,  $w_1$ . Ces derniers n’étant plus calés sur  $U$ , on refait le calage sur  $U$  partant des  $w_1$ , ce qui donne des poids  $w_2$ , et ainsi de suite. Cette procédure itérative peut converger (calage, variance, et poids). Si c’est le cas, les poids finaux convergés sont calés simultanément sur  $U$  et  $X$ .

Il y a des cas où le calage converge mais ni les poids ni la variance qui oscillent indéfiniment entre deux valeurs. Des oscillations de poids, de variances et de calages s’observent aussi par exemple dans le cas où les deux calages successifs sont des estimations par ratio.

Exemple 1 (suite) : si on veut maintenant renormaliser la somme des poids à  $N$  par ‘règle de trois’, c’est-à-dire en utilisant un estimateur par ratio sur la variable unité, on retombe sur les poids initiaux.

La divergence, enfin, est possible quoiqu’un peu pathologique. La seconde partie de cet article classe les cas et donne des exemples et des conditions de convergence.

Un exemple et un résultat bien connu concernent l’algorithme du raking-ratio. Dans cette technique on dispose de deux (ou plus) informations auxiliaires concernant les effectifs des modalités de variables qualitatives : répartition par groupes d’âge d’une part, par régions d’autre part, par exemple. L’une d’entre elles (la région souvent), a déjà servi pour une stratification initiale. Sauf pathologie, cet algorithme converge et les pondérations limites calent sur les « marges » du tableau (inconnu en général) croisant les deux variables. On sait (Deville, Sarndal, Sautory, 1993) que ces pondérations peuvent être obtenues par un calage direct et que les variances sont celle des résidus de la variable d’intérêt dans un ajustement de type analyse de variance à effets additifs sans interaction.

On trouvera au paragraphe 5 des exemples élémentaires en dimension un illustrant presque tous les cas possibles. Avant cela, nous allons revoir les propriétés utiles des estimateurs par calage généralisé, puis comment les choses se passent pour deux calages successifs.

## 2. QUELQUES RAPPELS SUR LE CALAGE GENERALISE

On travaille avec des poids initiaux  $d$  sans biais où asymptotiquement sans biais (Deville, Sarndal (1992) par exemple). Les totaux de variables d’intérêt  $y$  sont estimés par  $\hat{t}_y = \sum_s d_k y_k = \sum dy$ . Pour ne pas surcharger, les sommes portent toujours sur toute la population mais les poids sont nuls hors échantillon. La quantité  $\hat{t}_y$  estime donc sans biais

(approximativement éventuellement)  $t_y = \sum y$ . La variance de cet estimateur est donnée par une forme quadratique connue  $Q(y)$ . Pour tout échantillon  $s$ , on dispose d'une forme quadratique  $Q_s(y_s)$  qui estime 'proprement'  $Q(y)$ . Les poids initiaux peuvent être ceux de Horvitz-Thompson ou n'importe quels autres, par exemple déjà calés ou redressés pour non-réponse. Un  $p$ -vecteur de totaux  $t_x = \sum x$  constitue l'information auxiliaire. Par tradition, on notera  $X$  la matrice  $N \times p$  des  $x_k$  de sorte que, par exemple,  $t_x = X' \mathbf{1}$  où  $\mathbf{1}$  est la variable unité dont toutes les coordonnées valent 1. On utilise une 'fonction de calage'  $F$  de  $\mathbf{R}^p$  dans  $\mathbf{R}$ , croissante et aussi régulière qu'il est nécessaire vérifiant  $F(0)=1$  et  $F'(0)=1$ ;  $z_k$  est un  $p$ -vecteur de variables 'instrumentales', et on note  $Z$  la matrice qui empile les  $z_k$ ; enfin,  $\lambda$  est un  $p$ -vecteur d'ajustement.

On cherche de nouveaux poids  $w_k = dF(z_k' \lambda)$  vérifiant les *équations de calage*:  $t_x = \sum dx F(z' \lambda)$  où  $\lambda$  dépend de  $s$  et est donc aléatoire. Si les  $z_k$  sont uniformément bornés (y compris en un sens asymptotique –Deville, Sarndal-1992), et si la matrice  $Z'X$  est inversible, les équations de calage ont une unique solution avec une probabilité qui tend vers 1 et on a  $\lambda = \sum (dxz')^{-1}(t_x - \hat{t}_x) = (Z'X)^{-1}(t_x - \hat{t}_x)$  à des infiniment petits d'ordre supérieur près. L'estimateur calé vaut, toujours à des infiniment petits d'ordre supérieur près,  $\hat{t}_y^w = \hat{t}_y + B'(t_x - \hat{t}_x)$  où  $B$  est le vecteur des coefficients de régression de  $y$  sur les  $x$  avec les instruments  $z$ . Autrement dit  $B$  est solution des équations normales  $Z'(y - XB) = 0$  soit  $B = (Z'X)^{-1}Z'y$ . La variance de cet estimateur est donnée (approximativement, c'est à dire à partir d'arguments asymptotiques que nous ne développerons pas pour ne pas nous emmerder ni le monde avec) par  $Q(e)$ , où  $e_k = y_k - x_k' B$  est le résidu (où l'erreur de prédiction) dans la régression de  $y$  sur les  $x$  avec instruments  $z$ . Le vecteur des résidus  $e = y - XB$  vaut  $(I - P)y$  où  $P$  est le projecteur dans  $\mathbf{R}^N$   $P = X(Z'X)^{-1}Z'$  sur  $Im(X)$  (sous espace engendré par les colonnes de  $X$ ) le long de  $Z^\perp$ , orthogonal de  $Im(Z)$  pour la métrique canonique de  $\mathbf{R}^N$ .

La variance de  $\hat{t}_y^w$  vaut donc  $Q((I - P)y) = y'(I - P)'Q(I - P)y$  en assimilant forme quadratique et matrice symétrique. L'estimation de variance s'effectue de façon consistante en utilisant  $Q_s(\hat{e}_s)$  où  $\hat{e}$  est l'ensemble des résidus empiriques.

Remarque: les auxiliaires sont connues sur l'échantillon et on connaît leur total. En revanche les instruments n'ont à être connus que sur l'échantillon. Naturellement, des auxiliaires peuvent servir d'instruments. C'est d'ailleurs le cas pour le calage classique quand les 'poids  $q$ ' sont tous égaux.

### 3. CALAGES SUCCESSIFS

On s'intéresse maintenant à ce qui se passe quand on fait deux calages successifs. On passe des poids initiaux  $w_0$  à des poids  $w_1$  par calage sur le  $p$ -vecteur  $X$  avec des instruments  $Z$ . Puis on cale sur  $U$  en utilisant des instruments  $V$ , présents dans l'échantillon. A priori, les  $X$  et les  $U$  peuvent comporter un sous espace de variables communes. C'est souvent le cas de la constante (variable 'unité'), comme dans le cas où  $X = (\mathbf{1}, x)$  –estimation par régression –, et où  $U$  est l'ensemble des indicatrices d'une variable catégorielle utilisée pour une poststratification.

Il se peut aussi que les instruments  $Z$  et  $V$  aient un sous espace commun. Naturellement, il peut se faire aussi que qu'une bonne partie des instruments soit obtenue simplement à partir des auxiliaires comme c'est le cas pour l'estimation par régression ou le calage ordinaire. Dans ce paragraphe ceci ne pose pas véritablement de problème, contrairement à ce qui se passe dans les itérations du calage dont il sera traité plus loin. Avec une fonction de calage  $G$  pour le second calage, on aura les équations de calage  $t_U = \sum w_1 u G(v' \mu)$  où  $\mu$  est un  $q$ -paramètre d'ajustement. Pour les propriétés au premier ordre, on pourra admettre que  $G$  et  $F$  sont linéaires, et que les nouveaux poids valent à peu près  $w_2 = w_1(1 + v' \mu) = w_0(1 + z' \lambda)(1 + v' \mu) = w_0(1 + z' \lambda + v' \mu) + \text{termes d'ordres supérieurs}$ .

On a maintenant un résultat important :

Résultat 1: La variance asymptotique de  $\sum w_2 y$ , estimateur muni de poids issus de deux calages successifs, vaut  $Q((I - P_X)(I - P_U)y)$  où  $P_X$  est le projecteur dans  $\mathbf{R}^N$   $P_X = X(Z'X)^{-1}Z'$  sur  $Im(X)$  (sous espace engendré par les colonnes de  $X$ ) le long de  $Z^\perp$  et  $P_U$  le projecteur  $P_U = U(V'U)^{-1}V'$  sur  $Im(U)$  le long de  $V^\perp$ .

En effet,  $\sum w_2 y$  a pour variance  $Q_i((1-P_U) y)$  où  $Q_i(y) = \text{Var}(\sum w_1 y) = Q((1-P_X) y)$ .

On l'obtient donc en utilisant les résidus de  $y$  régressé sur les  $U$  (avec les instruments  $V$ ), puis en régressant ces résidus sur les  $X$  (instruments  $Z$ ) ; ces nouveaux résidus sont alors portés dans la forme quadratique de variance initiale. L'estimation de variance se fait de façon parallèle en utilisant les résidus empiriques et l'estimation initiale de variance pour un estimateur non calé (poids  $w_0$ ).

Exemple: un sondage de Bernoulli a une variance proportionnelle à  $\sum y^2$ . Si on cale sur  $N$  avec instrument  $I$ , elle devient proportionnelle à  $\sum (y - \bar{y})^2$ . Après calage sur une variable  $x$  et instrument  $x$ , elle devient proportionnelle à  $\sum (y - \bar{y} - ax)^2$  avec  $a = \sum (x - \bar{x})(y - \bar{y}) / \sum x^2$ .

On va examiner maintenant la question du défaut de calage. Le premier calage annule la variance de  $\hat{t}_x^1 = \sum w_1 x$ . L'estimateur calé vaut, pour une variable courante  $y$  :  $\hat{t}_y^2 = \hat{t}_y^1 + B'_{y|U|V}(t_U - \hat{t}_U^1)$  où la régression est sur  $U$  le long de  $V^\perp$ . Après le second calage, on a donc, en appliquant ce résultat à chacune des coordonnées de  $X$  :  $\hat{t}_x^2 = \hat{t}_x^1 + B'_{XU}(t_U - \hat{t}_U^1)$ , et, comme  $\hat{t}_x^1 = t_x$  (estimateur calé !), on obtient le résultat :

Résultat 2 : Le défaut de calage sur  $X$  après un second calage sur  $U$  vérifie :

$$\hat{t}_x^2 - t_x = B'_{XU}(t_U - \hat{t}_U^1) \quad (1)$$

où  $B_{XU} = (V'U)^{-1}V'X$  est la matrice  $q \times p$  dont les colonnes sont formées de coefficients de régression des variables de  $X$  sur les variables de  $U$  avec les instruments  $V$ .

Les  $X$  ne restent calés parfaitement que s'ils n'ont aucune corrélation avec les  $U$ .

De la même façon, on obtient que  $\hat{t}_U^1 - \hat{t}_U^0 = B'_{UX}(t_x - \hat{t}_x^0)$ . S'il se trouve que l'estimateur initial (poids  $w_0$ ) lui-même est calé sur  $U$  (variables de stratification où d'équilibrage), on obtient :

$$\hat{t}_x^2 - t_x = B'_{XU}B'_{UX}(\hat{t}_x^0 - t_x) \quad (2)$$

Cette égalité permet d'étudier le défaut de calage sur  $X$  après un calage intermédiaire sur  $U$ . En particulier l'opérateur  $B_{UX}B_{XU} = (Z'X)^{-1}Z'U(V'U)^{-1}V'X$  fait apparaître les projecteurs  $P_U$  et  $P_X$ . On a par exemple la relation utile  $X B_{UX}B_{XU} = P_X P_U X$ . L'opérateur en question donne les coordonnées sur  $X$  du produit des deux projecteurs appliqué à un vecteur de  $\text{Im}(X)$  muni de la base constituée des colonnes de cette matrice. Toute la structure de l'opération 'double calage' dépend donc clairement de la restriction à  $\text{Im}(X)$  du produit de ces deux projecteurs.

Conclusion provisoire et extension légère: Lors de calages successifs, seul le dernier calage est acquis ; les calages antérieurs sont détruits de façon plus ou moins radicale. Si ces calages se font avec des auxiliaires  $X_i$  et des instruments  $Z_i$ ,  $P_i = X_i(Z_i'X_i)^{-1}Z_i'$  est le projecteur sur  $X_i$  orthogonalement à  $Z_i$ . Pour  $n$  calages successifs la variance d'estimation du total de  $y$  est donnée approximativement par  $Q(e_n)$  où  $e_n = (\prod_{i=1}^n (1 - P_i))y$ . Ces quantités sont susceptibles d'évoluer de façon assez erratique et peu fiable. Dans le cas de deux calages tout dépend du produit  $P_1 P_2$ .

#### 4. CALAGES ITÉRÉS : GÉNÉRALITÉS

Le contexte est toujours celui de la partie précédente, mais nous nous intéressons maintenant à ce qui se passe quand on itère le calage en s'ajustant sur les  $X$  aux étapes impaires, puis sur les  $U$  aux étapes paires. Pour obtenir la convergence des calages eux-mêmes c'est assez simple.

Résultat 3 : Une condition suffisante de convergence du calage en  $X$  et  $U$  est que la matrice  $B_{UX}B_{XU}$  ait toutes ses valeurs propres inférieures à 1 en valeur absolue.

La relation (2) permet en effet d'écrire :

$$\hat{t}_X^{2n} - t_X = ((B_{UX} B_{XU})')^n (\hat{t}_X^0 - t_X) \quad (3)$$

Il en résulte simplement qu'une condition suffisante de convergence du calage en  $X$  pour les étapes paires est que la matrice  $B_{UX} B_{XU}$  ait toutes ses valeurs propres inférieures à 1 en valeur absolue. Comme les poids des étapes impaires sont calés, cela assure bien la convergence. Cela assure aussi la convergence du calage sur  $U$  aux étapes impaires à cause de la relation (1). Par ailleurs, la suite des paramètres d'ajustements  $\lambda$  et  $\mu$  est majorée (en norme) par une série géométrique. On en déduit que la suite des poids  $w_n$  converge également vers une limite dont la partie principale est de la forme  $1 + z'\lambda + v'\mu$ . Cette quantité n'est autre que la partie principale des poids de calage sur  $t_X$  et  $t_U$  simultanément en utilisant les instruments  $Z$  et  $V$  simultanément. Par suite le calage final donne un estimateur dont la variance est équivalente à celle de l'estimateur par régression sur  $(X, U)$  utilisant les instruments  $(Z, V)$ .

Inversement, si  $B_{UX} B_{XU}$  admet une valeur propre plus grande que 1 en valeur absolue le calage diverge car une au moins de ses coordonnées tend vers l'infini en valeur absolue (sauf événement de probabilité négligeable).

Tout semble parfait si nous ignorons qu'il puisse exister des valeurs propres égales à 1 (en valeur absolue qui plus est !). Malheureusement, ce type d'événement arrive nécessairement dans le cas fréquent où les auxiliaires ont un ensemble commun de variables, très souvent au moins la constante. Si des variables sont communes aux  $X$  et aux  $V$ , on peut toujours les faire apparaître explicitement (cas des indicatrices d'une poststratification pour la constante) et cela produit un bloc unité dans les matrices  $B_{UX}$  et  $B_{XU}$ . Ces deux matrices et leur produit vont donc avoir autant de valeur propre égales à 1 que de variables communes. De même, l'existence d'instruments communs (souvent encore la constante) peut produire des valeurs propres unité pour le produit des deux matrices. Enfin, on peut montrer que l'usage d'instruments arbitraires peut mener à des valeurs propres arbitraires, éventuellement complexes de module 1.

Pour essayer de comprendre ce qui peut arriver nous allons regarder attentivement ce qui se passe quand tous les calages se font sur une seule variable.

## 5. LE CAS DE LA DIMENSION UN

Les matrices  $B_{UX}$  et  $B_{XU}$  se réduisent à des scalaires  $B_{UX} = \sum vx / \sum vu$  et  $B_{XU} = \sum zu / \sum zx$ . Sans restreindre la généralité on peut convenir que  $U=I$ , et  $B_{UX}$  est la moyenne de  $x$  pondérée par  $v$ , tandis que  $B_{XU}$  est l'inverse de la moyenne de  $x$  pondérée par  $z$ . Les équations de calage paires s'écrivent  $N = \sum w_{2n}(1 + v\mu)$ . Les calages impairs s'écrivent  $X = \sum w_{2n+1}(1 + z\lambda)$ . Les quantités  $\sum v$  et  $\sum zx$  diffèrent de zéro à cause de l'hypothèse d'inversibilité des matrices  $V'U$  et  $Z'X$ . On a donc :  $B_{UX} B_{XU} = \frac{\sum vx}{\sum vu} \frac{\sum zu}{\sum zx} = \frac{\bar{x}_v}{\bar{x}_z}$  (4).

Ce ratio est susceptible de prendre n'importe quelle valeur réelle. Il y a convergence si sa valeur absolue est plus petite que 1, divergence si elle est supérieure à 1, conformément à l'analyse générale.

Le cas particulier d'une valeur (propre) de module un peut provenir de plusieurs cas particuliers.

Cas 1 : Si  $x=u$  (même auxiliaire) les deux facteurs valent un et les calages sur l'auxiliaire sont évidemment assurés à toutes les étapes de l'itération. Cependant la variance est, aux étapes paires, celle de  $y - x \frac{\sum zy}{\sum zx}$ , résidus de  $y$  régressé

sur  $x$  avec les instruments  $z$ , aux étapes impaires celle de  $y - x \frac{\sum uy}{\sum ux}$ , estimateur par ratio si  $u=I$ . Bien que les calages

soient toujours valides, le système de poids et la variance ne convergent pas : on observe une oscillation due au fait que les instruments sont différents. Si  $x=I$  les résidus ne sont pas autre chose que des écarts à la moyenne, pondérée par  $z$  ou  $u$  selon la parité.

Cas 2 : Naturellement, si de plus  $z=v$  il n'y a de fait qu'un seul calage et d'une certaine façon, la convergence a lieu dès la première itération.

Cas 3 : La quantité (4) vaut 1 pour toute  $x$  et toute  $u$  quand  $z=v$  c'est à dire quand l'instrument est commun aux deux calages. Géométriquement cela correspond au cas où l'orthogonal des instruments (communs, dans le cas général)

coupe le sous espace des explicatives ( $ImU+ImX$ ) selon un sous espace non réduit à zéro. Ici, si on fait  $v=z=I$ , on retrouve l'exemple élémentaire des ratios successifs. Dans ce cas, il n'y a pas de convergence, mais oscillation selon la parité des itérations des poids, des résidus et des calages eux-mêmes.

Cas 4 : Si en plus  $x=u$  on est dans le cas 2.

Cas 5 : La quantité (4) vaut 1 en module fortuitement.

On a généralement pas de convergence mais des oscillations. Il est facile de voir que si on étudiait des modèles de dimension 2 (ou plus !) les matrices mises en jeu seraient presque arbitraires et que les itérations pourraient, en choisissant bien les instruments se représenter par des rotations (par exemple !) dans des espaces appropriés.

Remarquons enfin que si on a  $z=x$  et  $v=u$ , ce qui correspond à la régression par moindres carrés ordinaires, ou, plus généralement  $z=Mx$  et  $v=Mu$  pour une matrice symétrique positive -moindres carrés généralisés- (ou même  $M$  matrice inversible quelconque), la quantité (4) vaut  $\frac{(x'Mu)^2}{(x'Mx)(u'Mu)}$ . Dès que  $x \neq u$  cette quantité est plus petite que

1. Si  $x=u$ , on est en fait dans le cas (2) et donc on a toujours convergence.

## 6. DIFFICULTES LIEES A DES VARIABLES COMMUNES

La classification du paragraphe précédent peut paraître simpliste, mais, en fait, elle permet de comprendre la façon dont les choses se passent dans le cas général. La démonstration complète est assez fastidieuse bien qu'elle ne repose que sur un argument assez simple utilisé de façon répétitive, et, in fine, sur la décomposition de Jordan d'une matrice carrée générale. Nous nous contenterons ici d'énoncer les résultats qui généralisent les constatations faites dans le cas de la dimension un. Introduisons quelques notations.  $C$  désigne le sous espace des variables communes à  $X$  et  $U$ , et, par abus de notation, la matrice  $N \times c$  de ces variables quand on les a écrites de façon explicite. De même  $I$  désigne l'espace des instruments communs à  $Z$  et  $V$  de même que la matrice  $N \times i$  obtenue en mettant les variables communes sous forme explicite.

Résultat 4 : Si  $c < i$  et que  $B_{UX}B_{XU}$  n'admet aucune valeur propre plus grande que 1 en valeur absolue, il y a convergence du calage mais la variance et les poids oscillent indéfiniment entre deux valeurs.

Résultat 5 : Si  $c > i$  et que  $B_{UX}B_{XU}$  n'admet aucune valeur propre plus grande que 1 en valeur absolue, il y a oscillations du calage, des poids et de la variance.

Dans le cas où  $c=i$  on peut montrer facilement le lemme suivant :

Lemme 1 : On peut trouver un changement de variable  $X=(C, X_I)$ ,  $U=(C, U_I)$ ,  $Z=(I, Z_I)$ ,  $V=(I, V_I)$  vérifiant les conditions  $I'X_I = I'U_I = Z_I' C = V_I' C = 0$  et  $I' C = V_I' U_I = Z_I' X_I = I$  (matrice identité de dimension convenable). Les auxiliaires  $X_I$  et  $V_I$  sont disjointes ainsi que les instruments  $Z_I$  et  $V_I$ .

Il en résulte que itérations de calages peuvent être faites indépendamment pour  $(C, I)$  et pour le reste des variables et des instruments. Pour ce dernier ensemble de variables, le problème est bien conditionné, et un second lemme technique permet de bien identifier le comportement des itérations. On supposera  $q < p$ .

Lemme 2 : Si les auxiliaires  $X$  et  $U$  d'une part, et les instruments  $Z$  et  $V$ , d'autre part n'ont de sous espace communs que zéro (ils sont 'disjointes'), on peut trouver une base de  $Im(X)$  et une base de  $Im(U)$  telles que :

- la matrice de  $P_X$  restreinte à  $Im(U)$  s'écrit :  $\begin{pmatrix} I_q & I_q^* & 0 \\ 0 & 0 & 0 \end{pmatrix}$  où  $I_q^*$  est une matrice carrée diagonale de ne comportant que des 1 dans sa partie nord-ouest et des 0 ailleurs

- la matrice de  $P_U$  restreinte à  $Im(X)$  s'écrit :  $\begin{pmatrix} 0 & 0 & 0 \\ I_p & L & L_0 \end{pmatrix}$  où  $L$  et  $L_0$  sont des matrices arbitraires.

Si  $x$  est alors un vecteur arbitraire de  $Im(X)$  on a  $P_X P_U x = Lx$  ce qui permet d'étudier facilement les itérations. On obtient la conclusion suivante :

**Résultat 6** : La suite des calages, des pondérations et des variances converge si et seulement si  $c=i$  et que toutes les valeurs propre de  $L$  sont strictement inférieures à 1 en module. L'estimateur calé est équivalent à un estimateur par régression sur  $(C, X_I, U_I)$  (soit toutes les variables non colinéaires) utilisant les instruments  $(I, Z_I, U_I)$ .

En dépit des apparences cette condition nécessaire et suffisante est relativement facile à vérifier car elle n'utilise que des conditions algébriques simples. Toutefois, il faudra remplacer les matrices par des estimations en espérant que les variations de poids au cours des itérations ne changent pas la structure du problème.

## 7. UNE CONDITION SUFFISANTE

L'usage d'instruments distinct des auxiliaires constitue une extension très utile des méthodes de calage, en particulier quand il s'agit de les utiliser pour diminuer le biais de réponse. Cependant, comme on vient de le voir, on rencontre des difficultés sérieuses quand les choisit de façon trop indépendante des auxiliaires. Il est bien connu qu'on observe une perte d'efficacité par rapport au calage ordinaire quand on veut réduire la variance d'estimation. On vient de voir aussi qu'on peut avoir une totale inconsistance dans le cas de calages successifs ou itérés, ou, assez souvent, des oscillations lors d'itérations successives du calage.

Une partie de ces difficultés disparaissent quand les instruments dépendent des auxiliaires par une transformation linéaire inversible. D'abord les auxiliaires communes ont la même dimension que les instruments communs, ce qui assure la première partie de la condition nécessaire et suffisante de convergence du résultat 6.3.

On a vu aussi qu'en dimension un, cette condition assurait quasi magiquement la convergence du calage itéré. Le cas général est justiciable de la condition suffisante suivante :

**Résultat 7** : Si les instruments sont déterminés à partir d'un principe de moindres carrés (éventuellement pondérés) identique pour les deux calages, le calage itéré converge vers le calage simultané.

**Remarque et commentaire 1** : Ce résultat ne fonctionne pas même si les deux calages sont des calages ordinaires associés à des poids de régression différents. Voir le contre exemple ci dessous.

**Commentaire 2** : Le principe des moindres carrés ordinaires conduit à  $Z=X$  et  $V=U$ . Celui des moindres carrés pondérés à multiplier les auxiliaires par une matrice diagonale à éléments strictement positifs. Le résultat ci-dessus demeure valide si on obtient les instruments par multiplication des auxiliaires par une matrice symétrique définie positive. Cette remarque peut s'avérer utile dans certaines configurations.

**Preuve** : Soit  $M$  la matrice métrique symétrique positive du principe de moindres carrés utilisé : identité pour les MCO, une matrice diagonale de poids pour les moindres carrés pondérés. Comme on a  $Z=MX$  et  $V=MU$  le projecteur  $P_U = U(V'U)^{-1}V'$  et  $P_X = X(M'X)^{-1}M'$  est un projecteur orthogonal dans la métrique  $M$  ( $M$  symétrique). Il en va de même de  $P_V$ . Le produit de ces deux projecteurs a une norme inférieure à 1. Il n'a de valeurs propres égales à un que si  $X$  et  $V$  ont un sous espace commun. Le résultat 6 s'applique donc.

**Exemple** : Dans l'estimation par raking-ratio, les deux poststratifications alternées peuvent être vues comme des régressions de type MCO sur les indicatrices des modalités des variables utilisées. Il existe une auxiliaire commune qui est la constante, et un instrument commun qui est aussi la constante. La convergence est donc assurée d'après le résultat 7 (sauf les pathologies habituelles qui se traduisent par une violation du bon conditionnement du problème).

**Exemple et contre exemple** : Supposons un calage sur deux variables, l'unité et  $x$ , associé à une régression des moindres carrés ordinaires  $y=ax+b$ .

On peut être tenté de dissocier les deux équations normales de la régression, où, en terme de calage, alterner un calage sur  $N$  qui dit que la somme des poids vaut  $N$ ,  $\sum w = N$ , et un calage sur  $x$  qui dit que  $\sum wx = t_x$ . Ces deux calages appellent naturellement à l'idée de règle de trois sur les poids, ou d'estimateur par ratio.

On voit facilement qu'on ne peut pas alterner les règles de trois. Ce sont en effet des régressions qui toutes les deux utilisent le même instrument qui est l'unité. Conformément au résultat du paragraphe 5-cas (3)-, les poids, les variances et les résidus oscillent.

Supposons qu'on commence par normaliser les poids à un par règle de trois, le second calage utilisera l'instrument des MCO c'est à dire  $x$  lui même. L'équation de calage sera  $t_x = \sum w_x(I + \lambda x)$ . En vertu du résultat 7, la suite des calages converge vers un estimateur équivalent à l'estimateur par régression simple.

Supposons maintenant qu'on tienne au ratio, et donc qu'on commence par un calage sur  $x$  avec l'instrument  $I$ . On ne peut guère envisager que de caler sur  $N$  avec l'instrument  $x$ , c'est à dire en utilisant le calage  $N = \sum w(I + \lambda x)$  associé à la régression sur la constante (moyenne !) pondérée par les  $x$ . Malheureusement cette procédure d'estimation est inconsistante car elle diverge. On remarquera que ce deuxième calage peut être compris, si tous les  $x_k$  sont positifs, comme une estimation par régression avec poids  $I/x_k$ . D'où le commentaire 1 ci-dessus.

Remarque : Les instruments n'agissent que par le sous espace qu'ils engendrent ; néanmoins, le fait de scinder ce sous espace en deux sous espaces supplémentaires ne peut pas se faire indépendamment d'une scission cohérente du sous espace des auxiliaires.

## 8. CONCLUSIONS

L'utilisation de calages successifs est une pratique assez risquée. Sauf exception, en effet, le second calage altère les effets du premier, soit assez légèrement, soit totalement. Quand on itère les calages, les difficultés éventuelles deviennent plus critiques.

Certes, dans le cas où les deux (ou plus) calages successifs font appel aux même type de régression en utilisant les mêmes poids, l'itération des calages converge vers un estimateur du même type équivalent à un estimateur par régression sur l'ensemble des variables indépendantes. En revanche, dès que l'on n'utilise pas les mêmes poids, ou qu'on utilise des instruments choisis indépendamment, le risque d'instabilité ou de divergence des estimations existe et doit être étudié sérieusement.

De fait, il est beaucoup plus sur, quand on le peut, de caler en une seule fois sur toutes les auxiliaires disponibles, y compris pour corriger de la non-réponse. Le choix des instruments reste une affaire assez délicate, mais, à tous les moins, le risque d'aboutir à une estimation inconsistante n'existe plus et les performances de l'estimateur calé en une unique étape sont bien établies, bref, on sait ce qu'on fait.

## RÉFÉRENCES

- Deville, J.C, et Särndal, C.E (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association*, Vol 87, p 376-382
- Deville, J.C, Särndal, C.E. et Sautory, O. (1993), Generalized raking procedures in survey sampling, *Journal of the American Statistical Association*, Vol 88, pp 1013-1020
- Deville, J.C. (1998), La correction de la non-réponse par calage ou par échantillonnage équilibré, *Actes de la rencontre annuelle de la Société de Statistique du Canada*, Sherbrooke, Canada.
- Deville, J.C. (2000), Generalized Calibration and Application to Weighting for Non-response, *Actes du Colloque COMPSTAT 2000*, Utrecht, Pays-Bas, Springer
- Deville, J.C. (2002), La correction de la non-réponse par calage généralisé, *Journées de Méthodologie Statistique de L'INSEE*, Paris, France sur <http://jms.insee.fr>

Deville, J.C. , Tillé, Y.(2004), Efficient balanced sampling: The cube method , *Biometrika*, 91,4, p. 893-912.

Hidiroglou,M..A.(2001), L'échantillonnage double, *Survey Methodology*, Vol 27 n°2.

Lavallée, P. (2002), Le Sondage Indirect ou la Méthode Généralisée de Partage des Poids, *Ellipses*, Bruxelles.

Le Guennec, J. , Sautory, O. (2002) « Une nouvelle version de la macro CALMAR de redressement d'échantillon par calage », *Journées de Méthodologie Statistique de L'INSEE*, Paris ,France sur <http://jms.insee.fr>

Särndal,C.E, Lundström, S. (2005),*Estimation in Surveys with Nonresponse*, *Wiley*, New-York.