Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Calage et poids restreints

par Alain Théberge

Juin 2000

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2000
Vol. 26, N° 1, pp. 113-122
Statistique Canada, N° 12-001 au catalogue

Calage et poids restreints

Alain Théberge [1]

Résumé

Pour mieux comprendre l'impact de l'imposition d'une région de restriction sur les poids de calage, on examine le comportement asymptotique de ceux-ci. On donne des conditions nécessaires et suffisantes pour l'existence d'une solution à l'équation de calage avec des poids à l'intérieur d'intervalles donnés. Une formulation plus générale du problème de calage permet de faire un compromis entre le besoin de satisfaire l'équation de calage, et le désir d'obtenir des poids qui sont près des poids de Horvitz-Thompson. Si on relâche les exigences vis-à-vis l'équation de calage, alors diverses méthodes d'estimation avec poids restreints peuvent être utilisées. Les estimateurs présentés ont habituellement les mêmes propriétés asymptotiques que l'estimateur par calage sans restrictions sur les poids, et certains ont des poids qui peuvent être calculés explicitement, sans procédure itérative. On montre comment ces estimateurs peuvent être adaptés pour tirer parti d'un estimateur synthétique. Une stratégie semblable à celle utilisée pour restreindre les poids est appliquée aux données aberrantes.

Mots clés : Petits domaines; inverse de Moore-Penrose; solutions d'inéquations; propriétés asymptotiques; données aberrantes.

1. Introduction

L'estimateur par calage possède de bonnes propriétés asymptotiques. Mais pour de petites tailles d'échantillon, ou si le calage se fait au niveau de domaines dont certains ont peu d'observations, les poids de cet estimateur peuvent prendre des valeurs extrêmes. Une façon de pallier ce problème consiste à utiliser la méthode de calage avec des mesures de distance qui font que les poids des observations sont restreints à certains intervalles autour des poids de sondage. Cette approche a été développée par Deville et Särndal (1992). D'autres méthodes qui visent à obtenir des estimations robustes qui respectent l'équation de calage sont données dans Duchesne (1999). Cet article contient une bonne bibliographie sur les estimateurs robustes. On n'est toutefois pas assuré de l'existence d'une solution à l'équation de calage avec des poids restreints. Même si de tels poids existent, il se peut que le statisticien préfère résoudre le problème de poids extrêmes en relâchant quelque peu ses exigences vis-à-vis l'équation de calage, plutôt qu'en resserrant les contraintes sur les poids en utilisant une mesure de distance plus « contraignante ». On présentera ici, une formulation du problème de calage qui offre plus de flexibilité au statisticien. Il s'agit en fait d'un problème de minimisation du même ordre que celui rencontré lors d'une régression avec coefficients de coûts (« ridge regression »). Bardsley et Chambers (1984) ont rencontré le même problème de minimisation dans leur recherche d'estimateurs basés sur des modèles. Cette formulation du problème de calage peut être utilisée pour restreindre les poids sans utiliser de mesure de distance spéciale entre les poids calés et les poids de Horvitz-Thompson. Rao et Singh (1997) ont eux combiné cette approche avec des méthodes itératives utilisant des mesures de distance. D'autres façons de restreindre les poids seront aussi examinées.

Dans la section suivante on présente la méthode de calage en l'absence de bornes sur les valeurs des poids. Le problème de calage qui est posé ne présume pas de l'existence d'une solution à l'équation de calage. Les propriétés asymptotiques des poids calés sont discutées. Ces propriétés sont d'intérêt pour le comportement asymptotique des estimateurs dont les poids sont restreints. La section 3 donne des conditions nécessaires et suffisantes pour l'existence de poids restreints qui satisfont l'équation de calage. À la section 4, on discute d'une façon de poser le problème d'estimation en dosant l'importance qu'on accorde à l'équation de calage. La section 5 donne différentes façons de restreindre les poids qui ne reposent pas sur l'utilisation d'une distance particulière. On propose à la section 6, un estimateur avec poids restreints qui est utile pour de petits domaines. Finalement, la section 7 aborde les données aberrantes en développant une méthode semblable à celle utilisée pour traiter les poids extrêmes.

2. Calage

Soient $\boldsymbol{Y} \in \mathbb{R}^{N\times d}$ une matrice de d variables d'intérêt pour une population de taille N, et $\boldsymbol{c} \in \mathbb{R}^{N}$ un vecteur de constantes connues, on tire un échantillon s de taille n et on notera à l'aide de l'indice s les sous-vecteurs ou les sous-matrices qui correspondent à l'échantillon. Il s'agit d'estimer $\boldsymbol{Y}'\boldsymbol{c}$ par $\boldsymbol{Y}_s'\boldsymbol{w}_s$, où $\boldsymbol{w}_s \in \mathbb{R}^{n}$ est un vecteur de poids pour les unités échantillonnées. Pour un vecteur $\boldsymbol{v}$ et une matrice diagonale positive $\boldsymbol{F}$ de même dimension, on définit $\|\boldsymbol{v}\|_{\boldsymbol{F}}^2 = \boldsymbol{v}'\boldsymbol{F}\boldsymbol{v}$. Pour une matrice d'information

1. Alain Théberge, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario) K1A 0T6 Canada.

auxiliaire $\boldsymbol{X} \in \mathbb{R}^{N\times p}$, $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ la matrice diagonale des poids de sondage, des matrices diagonales positives données $\boldsymbol{U}_s \in \mathbb{R}^{n\times n}$ et $\boldsymbol{T} \in \mathbb{R}^{p\times p}$, on cherche parmi les vecteurs de poids $\boldsymbol{W}_s \in \mathbb{R}^n$ qui minimisent $\|\boldsymbol{X}_s'\boldsymbol{w}_s - \boldsymbol{X}'\boldsymbol{c}\|_{\boldsymbol{T}}^2$, celui qui minimise $D_s(\boldsymbol{w}_s) = \|\boldsymbol{w}_s - \boldsymbol{A}_s\boldsymbol{c}_s\|_{\boldsymbol{U}_s}^2$. Cette formulation du problème qui ne présume pas de l'existence de poids satisfaisant l'équation de calage, $\boldsymbol{X}_s'\boldsymbol{w}_s = \boldsymbol{X}'\boldsymbol{c}$, est donnée dans Théberge (1999). La solution recherchée constitue le vecteur des poids calés $\boldsymbol{w}_{\text{cal}}$. On a

$$\boldsymbol{w}_{\text{cal}} = \boldsymbol{A}_s\boldsymbol{c}_s + \boldsymbol{U}_s^{-1}\boldsymbol{X}_s\boldsymbol{T}^{1/2}(\boldsymbol{T}^{1/2}\boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s\boldsymbol{T}^{1/2})^{\dagger} \boldsymbol{T}^{1/2}(\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s), \quad (1)$$

où $\boldsymbol{F}^{\dagger}$ dénote l'inverse de Moore-Penrose de la matrice $\boldsymbol{F}$.

Afin de mieux étudier les propriétés asymptotiques d'estimateurs par calage avec poids restreints, on examinera maintenant le comportement de $\boldsymbol{w}_{\text{cal}}$ lorsque $n \to \infty$. On suppose l'existence d'un cadre asymptotique où la taille de la population et la taille de l'échantillon tendent vers l'infini, voir par exemple Isaki et Fuller (1982), et pour lequel on a

$$\begin{aligned} \boldsymbol{Y}'\boldsymbol{c} &= O_p(N^{\gamma})\ (\gamma \geq 0) \\ \boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s &= O_p(n^{-1/2}N^{\gamma}) \qquad (2) \\ \boldsymbol{T}^{1/2}\boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s\boldsymbol{T}^{1/2} &= O_p(n). \end{aligned}$$

Il s'ensuit que $(\boldsymbol{T}^{1/2}\boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s\boldsymbol{T}^{1/2})^{\dagger} = O_p(n^{-1})$, puisqu'une des propriétés de l'inverse de Moore-Penrose d'une matrice $\boldsymbol{F}$ est $\boldsymbol{F}^{\dagger}\boldsymbol{F}\boldsymbol{F}^{\dagger} = \boldsymbol{F}^{\dagger}$. Habituellement, on peut s'attendre à avoir $\gamma = 1$ lorsque chaque élément du vecteur $\boldsymbol{c}$ vaut 1 (estimation d'un total), et $\gamma = 0$ lorsque chaque élément de $\boldsymbol{c}$ vaut $1/N$ (estimation d'une moyenne). Sous les conditions (2) on a donc,

$$\begin{aligned} \boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\boldsymbol{c}_s &= \boldsymbol{U}_s^{-1}\boldsymbol{X}_s\boldsymbol{T}^{1/2}(\boldsymbol{T}^{1/2}\boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s\boldsymbol{T}^{1/2})^{\dagger} \\ &\quad \boldsymbol{T}^{1/2}(\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s) \\ &= O_p(n^{-1})O_p(n^{-1/2}N^{\gamma}) \\ &= O_p(n^{-3/2}N^{\gamma}). \qquad (3) \end{aligned}$$

Donc $\boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\boldsymbol{c}_s$ converge en probabilité vers $\boldsymbol{0}$, si

$$\lim_{n,N\to\infty} n^{-3/2}N^{\gamma} = 0.$$

Pour un cadre asymptotique tel celui de Brewer (1979) où la fraction de sondage n/N est constante, ou tout cadre pour lequel la fraction de sondage converge vers un nombre positif, cette condition est vérifiée si $\gamma < 3/2$.

Si on écrit $\boldsymbol{w}_{\text{cal}} = \boldsymbol{A}_s\boldsymbol{c}_s + \boldsymbol{U}_s^{-1}\boldsymbol{X}_s\boldsymbol{T}^{1/2}\boldsymbol{H}_s^{\dagger}\boldsymbol{T}^{1/2}(\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s)$, où $\boldsymbol{H}_s = \boldsymbol{T}^{1/2}\boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s\boldsymbol{T}^{1/2}$, on a

$$\begin{aligned} D_s(\boldsymbol{w}_{\text{cal}}) &= (\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s)'\boldsymbol{T}^{1/2}\boldsymbol{H}_s^{\dagger}\boldsymbol{H}_s\boldsymbol{H}_s^{\dagger} \\ &\quad \boldsymbol{T}^{1/2}(\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s) \\ &= (\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s)'\boldsymbol{T}^{1/2}\boldsymbol{H}_s^{\dagger}\boldsymbol{T}^{1/2}(\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s) \\ &= O_p(n^{-1/2}N^{\gamma})O_p(n^{-1})O_p(n^{-1/2}N^{\gamma}) \\ &= O_p(n^{-2}N^{2\gamma}). \qquad (4) \end{aligned}$$

Toujours pour un cadre asymptotique où la fraction de sondage converge vers un nombre positif, on a $D_s(\boldsymbol{w}_{\text{cal}})$ converge en probabilité vers 0, si $\gamma < 1$. Il y a donc des cas, notamment pour l'estimation d'un total, où $\boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\boldsymbol{c}_s$ converge en probabilité vers $\boldsymbol{0}$, mais où $D_s(\boldsymbol{w}_{\text{cal}}) = \|\boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\boldsymbol{c}_s\|_{\boldsymbol{U}_s}^2$ ne converge pas vers 0.

3. Solutions à l'équation de calage et poids restreints

Même lorsqu'il n'y a pas de restrictions sur les poids, il est possible qu'il n'y ait pas de solution à l'équation de calage. En appliquant Graybill (1983, 113) au problème de calage, on obtient que l'équation de calage $\boldsymbol{X}_s'\boldsymbol{w}_s = \boldsymbol{X}'\boldsymbol{c}$ a une solution si et seulement si $(\boldsymbol{X}_s^{\dagger}\boldsymbol{X}_s)'\boldsymbol{X}'\boldsymbol{c} = \boldsymbol{X}'\boldsymbol{c}$. Si une solution existe, il est possible que les poids calés soient négatifs ou exceptionnellement grands. Deville et Särndal (1992) ont proposé d'utiliser diverses mesures de distance autres qu'une somme pondérée de carrés pour mesurer la distance entre les poids de Horvitz-Thompson et les poids calés, afin de restreindre les poids à certains intervalles tout en satisfaisant l'équation de calage. Cette approche ne fonctionnera que s'il existe des poids dans ces intervalles qui satisfont l'équation de calage. Le but principal de cette section est de trouver des conditions nécessaires et suffisantes pour l'existence d'un vecteur de poids $\boldsymbol{w}_s$ à l'intérieur de bornes données, et tel que les estimations des totaux pour les variables auxiliaires sont également bornées. C'est-à-dire qu'on cherche des conditions pour l'existence d'un vecteur $\boldsymbol{w}_s$ tel que $\boldsymbol{w}^{(L)} \leq \boldsymbol{w}_s \leq \boldsymbol{w}^{(H)}$ et $\boldsymbol{t}^{(L)} \leq \boldsymbol{X}_s'\boldsymbol{w}_s \leq \boldsymbol{t}^{(H)}$, où $\boldsymbol{w}^{(L)}$, $\boldsymbol{w}^{(H)}$, $\boldsymbol{t}^{(L)}$ et $\boldsymbol{t}^{(H)}$ sont donnés. En particulier, en posant $\boldsymbol{t}^{(L)} = \boldsymbol{t}^{(H)} = \boldsymbol{X}'\boldsymbol{c}$, on obtiendrait des conditions pour l'existence de poids restreints aux intervalles $\boldsymbol{w}^{(L)} \leq \boldsymbol{w}_s \leq \boldsymbol{w}^{(H)}$, qui satisfont l'équation de calage.

Une première étape est fournie par le théorème suivant de Fan (1956). Il est énoncé ici pour une matrice $\boldsymbol{M}$ de dimension finie, quoique la preuve donnée par Fan tienne également pour une matrice de dimension infinie. Le théorème fait appel au noyau de $\boldsymbol{M}'$, $N(\boldsymbol{M}')$, défini comme l'ensemble des vecteurs $\boldsymbol{\alpha}$ tels que $\boldsymbol{M}'\boldsymbol{\alpha} = \boldsymbol{0}$.

Théorème : Soient $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ et $\boldsymbol{l} \in \mathbb{R}^m$, $\exists \boldsymbol{w} \in \mathbb{R}^n$ tel que $\boldsymbol{M}\boldsymbol{w} \geq \boldsymbol{l}$ si et seulement si pour tout $\boldsymbol{\lambda} \geq \boldsymbol{0}$ dans $N(\boldsymbol{M}')$, on a $\boldsymbol{l}'\boldsymbol{\lambda} \leq 0$.

Corollaire : Soient $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ et $\boldsymbol{l}, \boldsymbol{h} \in \mathbb{R}^m$, $\exists \boldsymbol{w} \in \mathbb{R}^n$ tel que $\boldsymbol{l} \le \boldsymbol{M}\boldsymbol{w} \le \boldsymbol{h}$ si et seulement si premièrement $\boldsymbol{l} \le \boldsymbol{h}$ et deuxièmement $\boldsymbol{\lambda} \in N(\boldsymbol{M}') \Rightarrow -\boldsymbol{l}'\boldsymbol{\lambda}_- \le \boldsymbol{h}'\boldsymbol{\lambda}_+$, où $\boldsymbol{\lambda}_+ = \max(\boldsymbol{\lambda}, \boldsymbol{0})$ et $\boldsymbol{\lambda}_- = \min(\boldsymbol{\lambda}, \boldsymbol{0})$ les extrema étant pris élément par élément.

Le corollaire est obtenu en posant

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M} \\ -\boldsymbol{M} \end{pmatrix},\ \boldsymbol{l} = \begin{pmatrix} \boldsymbol{l} \\ -\boldsymbol{h} \end{pmatrix} \text{ et } \boldsymbol{\lambda} = \begin{pmatrix} -\boldsymbol{\lambda}_- \\ \boldsymbol{\lambda}_+ \end{pmatrix}$$

dans le théorème.

On note p la dimension de $N(\boldsymbol{M}')$. Si p est égal à zéro, alors $\boldsymbol{\lambda} \in N(\boldsymbol{M}')$ implique $\boldsymbol{\lambda} = \boldsymbol{0}$, et la condition du théorème (ou la condition similaire du corollaire) est évidemment satisfaite. Si p est égal à un, alors $\boldsymbol{\lambda} \in N(\boldsymbol{M}')$ implique que $\boldsymbol{\lambda}$ est un multiple d'un vecteur z, et il est suffisant de vérifier la condition pour $\boldsymbol{\lambda} = z$ et $\boldsymbol{\lambda} = -z$. En utilisant la propriété $(-\boldsymbol{\lambda})_- = -(\boldsymbol{\lambda})_+$, le problème posé au début de la section peut maintenant être résolu si $\boldsymbol{X}_s$ est un vecteur. Le corollaire avec

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{I}_n \\ \boldsymbol{X}_s' \end{pmatrix},\ \boldsymbol{l} = \begin{pmatrix} \boldsymbol{w}^{(L)} \\ \boldsymbol{t}^{(L)} \end{pmatrix},\ \boldsymbol{h} = \begin{pmatrix} \boldsymbol{w}^{(H)} \\ \boldsymbol{t}^{(H)} \end{pmatrix},$$

et le fait que

$$z = \begin{pmatrix} -\boldsymbol{X}_s \\ 1 \end{pmatrix}$$

engendre $N(\boldsymbol{M}')$, donne les conditions nécessaires et suffisantes

$$\begin{aligned} \boldsymbol{w}^{(L)} &\le \boldsymbol{w}^{(H)} \\ \boldsymbol{t}^{(L)} &\le \boldsymbol{t}^{(H)} \\ (\boldsymbol{X}_s)_+'\boldsymbol{w}^{(L)} + (\boldsymbol{X}_s)_-'\boldsymbol{w}^{(H)} &\le \boldsymbol{t}^{(H)} \\ \boldsymbol{t}^{(L)} &\le (\boldsymbol{X}_s)_+'\boldsymbol{w}^{(H)} + (\boldsymbol{X}_s)_-'\boldsymbol{w}^{(L)}. \end{aligned} \tag{5}$$

La troisième inégalité dans (5) affirme que le total pondéré de la variable auxiliaire ne doit pas être supérieur à $\boldsymbol{t}^{(H)}$, lorsque le plus petit poids possible $\boldsymbol{w}^{(L)}$ est donné aux unités où la variable auxiliaire prend une valeur positive, et lorsque le plus grand poids possible $\boldsymbol{w}^{(H)}$, est donné aux unités où la variable auxiliaire prend une valeur négative. La quatrième inégalité dans (5) affirme que le total pondéré de la variable auxiliaire ne doit pas être inférieur à $\boldsymbol{t}^{(L)}$, lorsque le plus grand poids possible est donné aux unités où la variable auxiliaire prend une valeur positive, et lorsque le plus petit poids possible est donné aux unités où la variable auxiliaire prend une valeur négative.

Même lorsque $p > 1$, il est suffisant de vérifier la condition du corollaire pour un nombre fini de valeurs de $\boldsymbol{\lambda}$. Soit $\boldsymbol{V} \in \mathbb{R}^{m\times p}$, $2 \le p \le m$ une matrice dont les colonnes forment une base pour $N(\boldsymbol{M}')$. Il est toujours possible de construire $\boldsymbol{V}$ de sorte que p de ses lignes, $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_m$, soient les vecteurs unités de $\mathbb{R}^p$, et on supposera que $\boldsymbol{V}$ est de cette forme. On démontre dans l'annexe A qu'il est suffisant de vérifier la condition du corollaire pour les vecteurs $\boldsymbol{\lambda} = \boldsymbol{V}\boldsymbol{\varphi}$ et $\boldsymbol{\lambda} = -\boldsymbol{V}\boldsymbol{\varphi}$, où $\boldsymbol{\varphi} = (\varphi_1, ..., \varphi_p)'$ est un vecteur non nul et satisfait $\boldsymbol{v}_i'\boldsymbol{\varphi} = 0$ pour un sous-ensemble de $(p-1)$ vecteurs $\boldsymbol{v}_i$ linéairement indépendants. On doit donc vérifier la condition pour au plus $\binom{m}{p-1}$ vecteurs $\boldsymbol{\varphi}$, donc au plus $2\binom{m}{p-1}$ valeurs de $\boldsymbol{\lambda}$.

En utilisant le corollaire avec

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{I}_n \\ \boldsymbol{X}_s' \end{pmatrix},\ \boldsymbol{l} = \begin{pmatrix} \boldsymbol{w}^{(L)} \\ \boldsymbol{t}^{(L)} \end{pmatrix},\ \boldsymbol{h} = \begin{pmatrix} \boldsymbol{w}^{(H)} \\ \boldsymbol{t}^{(H)} \end{pmatrix},$$

et en notant que les colonnes de

$$\boldsymbol{V} = \begin{pmatrix} -\boldsymbol{X}_s \\ \boldsymbol{I}_p \end{pmatrix}$$

forment une base pour $N(\boldsymbol{M}')$, on obtient les conditions nécessaires et suffisantes suivantes pour l'existence d'une solution au problème mentionné au début de cette section lorsque $\boldsymbol{X}_s \in \mathbb{R}^{n\times p}$ avec $p > 1$. On doit avoir $\boldsymbol{w}^{(L)} \le \boldsymbol{w}^{(H)}$, $\boldsymbol{t}^{(L)} \le \boldsymbol{t}^{(H)}$, et pour chaque sous-ensemble de $(p-1)$ lignes de

$$\boldsymbol{V} = \begin{pmatrix} -\boldsymbol{X}_s \\ \boldsymbol{I}_p \end{pmatrix}$$

linéairement indépendantes,

$$\begin{aligned} &(\boldsymbol{X}_s\boldsymbol{\varphi})_+'\,\boldsymbol{w}^{(L)} - \boldsymbol{\varphi}_-'\,\boldsymbol{t}^{(L)} \le -(\boldsymbol{X}_s\boldsymbol{\varphi})_-'\,\boldsymbol{w}^{(H)} + \boldsymbol{\varphi}_+'\,\boldsymbol{t}^{(H)} \\ &-(\boldsymbol{X}_s\boldsymbol{\varphi})_-'\,\boldsymbol{w}^{(L)} + \boldsymbol{\varphi}_-'\,\boldsymbol{t}^{(L)} \le (\boldsymbol{X}_s\boldsymbol{\varphi})_+'\,\boldsymbol{w}^{(H)} - \boldsymbol{\varphi}_-'\,\boldsymbol{t}^{(H)} \end{aligned} \tag{6}$$

pour un vecteur non nul $\boldsymbol{\varphi} \in \mathbb{R}^p$ de direction perpendiculaire à chaque ligne du sous-ensemble. La deuxième inéquation de (6) est obtenue de la première en changeant le signe de $\boldsymbol{\varphi}$.

Si $\boldsymbol{V}_{\text{sub}} \in \mathbb{R}^{p\times p}$ est une matrice non singulière dont les lignes sont des lignes de $\boldsymbol{V}$, alors chaque colonne de $\boldsymbol{V}_{\text{sub}}^{-1}$ est un vecteur perpendiculaire à $(p-1)$ lignes de $\boldsymbol{V}$ linéairement indépendantes. D'où le résultat suivant :

Il existe un vecteur de poids $\boldsymbol{w}_s$ tels que $\boldsymbol{w}^{(L)} \le \boldsymbol{w}_s \le \boldsymbol{w}^{(H)}$ et $\boldsymbol{t}^{(L)} \le \boldsymbol{X}_s'\boldsymbol{w}_s \le \boldsymbol{t}^{(H)}$ si et seulement si $\boldsymbol{w}^{(L)} \le \boldsymbol{w}^{(H)}$, $\boldsymbol{t}^{(L)} \le \boldsymbol{t}^{(H)}$ et

$$\begin{aligned} &(\boldsymbol{X}_s\boldsymbol{V}_{\text{sub}}^{-1})_+'\,\boldsymbol{w}^{(L)} - (\boldsymbol{V}_{\text{sub}}^{-1})_-'\,\boldsymbol{t}^{(L)} \le \\ &\qquad -(\boldsymbol{X}_s\boldsymbol{V}_{\text{sub}}^{-1})_-'\,\boldsymbol{w}^{(H)} + (\boldsymbol{V}_{\text{sub}}^{-1})_+'\,\boldsymbol{t}^{(H)} \\ &-(\boldsymbol{X}_s\boldsymbol{V}_{\text{sub}}^{-1})_-'\,\boldsymbol{w}^{(L)} + (\boldsymbol{V}_{\text{sub}}^{-1})_+'\,\boldsymbol{t}^{(L)} \le \\ &\qquad (\boldsymbol{X}_s\boldsymbol{V}_{\text{sub}}^{-1})_+'\,\boldsymbol{w}^{(H)} - (\boldsymbol{V}_{\text{sub}}^{-1})_-'\,\boldsymbol{t}^{(H)} \end{aligned} \tag{7}$$

pour toutes les matrices non singulières $\boldsymbol{V}_{\text{sub}} \in \mathbb{R}^{p\times p}$ dont les lignes sont des lignes de

$$\boldsymbol{V} = \begin{pmatrix} -\boldsymbol{X}_s \\ \boldsymbol{I}_p \end{pmatrix}.$$

Ces conditions sont dans une certaine mesure redondantes. Par exemple, si les inégalités (7) sont satisfaites pour $\boldsymbol{V}_{\text{sub}} = \boldsymbol{V}_1$, alors elles sont nécessairement satisfaites pour toute matrice $\boldsymbol{V}_2$ obtenue de $\boldsymbol{V}_1$ par une permutation de lignes.

Un autre exemple est fourni par la pondération d'observations dans un tableau de contingence. Soient $\hat{N}_{ij} = n_{ij} w_{ij}$ $(i = 1, 2, ..., R;\ j = 1, 2, ..., C)$, où n_{ij} est le nombre d'observations dans la cellule (i, j) du tableau de contingence et w_{ij} est le poids de chacune de ces observations, on veut savoir s'il existe des poids w_{ij} tels que $\hat{N}_{ij}$ satisfait certaines contraintes. Par exemple, motivé par le problème de la convergence de la procédure du quotient réitéré (raking ratio), Bacharach (1965) donne des conditions nécessaires et suffisantes pour l'existence de poids w_{ij} tels que $\hat{N}_{ij} \geq 0$, $\sum_{i=1}^{R} \hat{N}_{ij} = N_{.j}\ (j = 1, ..., C)$, $\sum_{j=1}^{C} \hat{N}_{ij} = N_{i.}\ (i = 1, ..., R)$, où les valeurs de $N_{.j}$ et $N_{i.}$ sont données. Le résultat suivant qui est démontré à l'annexe B, est plus général.

Pour des constantes arbitraires $N_{ij}^{(L)}$, $N_{ij}^{(H)}$, $N_{.j}^{(L)}$, $N_{.j}^{(H)}$, $N_{i.}^{(L)}$, $N_{i.}^{(H)}$, $N_{..}^{(L)}$, et $N_{..}^{(H)}$, il existe des $\hat{N}_{ij}$ tels que

$$\begin{aligned}
&N_{ij}^{(L)} \leq \hat{N}_{ij} \leq N_{ij}^{(H)} && i = 1, ..., R;\ j = 1, ..., C; \\
&N_{.j}^{(L)} \leq \sum_{i=1}^{R} \hat{N}_{ij} \leq N_{.j}^{(H)} && j = 1, ..., C; \\
&N_{i.}^{(L)} \leq \sum_{j=1}^{C} \hat{N}_{ij} \leq N_{i.}^{(H)} && i = 1, ..., R; \\
&N_{..}^{(L)} \leq \sum_{i=1}^{R} \sum_{j=1}^{C} \hat{N}_{ij} \leq N_{..}^{(H)},
\end{aligned} \tag{8}$$

si et seulement si

$$\begin{aligned}
&N_{ij}^{(L)} \leq N_{ij}^{(H)} && i = 1, ..., R;\ j = 1, ..., C; \\
&N_{.j}^{(L)} \leq N_{.j}^{(H)} && j = 1, ..., C; \\
&N_{i.}^{(L)} \leq N_{i.}^{(H)} && i = 1, ..., R; \\
&N_{..}^{(L)} \leq N_{..}^{(H)}
\end{aligned}$$

$$\begin{aligned}
&\sum_{j \in T} \left(N_{.j}^{(L)} - \sum_{i \notin S} N_{ij}^{(H)} \right) \\
&\qquad \leq \sum_{i \in S} \left(N_{i.}^{(H)} - \sum_{j \notin T} N_{ij}^{(L)} \right) \\
&\sum_{i \in S} \left(N_{i.}^{(L)} - \sum_{j \notin T} N_{ij}^{(H)} \right) \\
&\qquad \leq \sum_{j \in T} \left(N_{.j}^{(H)} - \sum_{i \notin T} N_{.j}^{(L)} \right) \\
&N_{..}^{(L)} + \sum_{j \notin T} \left(N_{.j}^{(H)} - \sum_{i \notin S} N_{ij}^{(H)} \right) \\
&\qquad \leq \sum_{i \in S} \left(N_{i.}^{(H)} - \sum_{j \in T} N_{ij}^{(L)} \right) + \sum_{j=1}^{J} N_{.j}^{(H)} \\
&\sum_{i=1}^{I} N_{i.}^{(L)} + \sum_{j \notin T} \left(N_{.j}^{(L)} - \sum_{i \notin S} N_{ij}^{(H)} \right) \\
&\qquad \leq \sum_{i \in S} \left(N_{i.}^{(L)} - \sum_{j \in T} N_{ij}^{(L)} \right) + N_{..}^{(H)}
\end{aligned} \tag{9}$$

pour tout $S \subseteq \{1, 2, ..., R\}$, $T \subseteq \{1, 2, ..., C\}$

On peut réduire le nombre d'inéquations à vérifier. Par exemple, plutôt que de vérifier

$$\sum_{j \in T} \left(N_{.j}^{(L)} - \sum_{i \notin S} N_{ij}^{(H)} \right) \leq \sum_{i \in S} \left(N_{i.}^{(H)} - \sum_{j \notin T} N_{ij}^{(L)} \right)$$

pour tout $S \subseteq \{1, 2, ..., R\}$, et $T \subseteq \{1, 2, ..., C\}$, on montre facilement qu'il est équivalent de vérifier que

$$\sum_{j \in T} N_{.j}^{(L)} \leq \sum_{i=1}^{R} \min\left(\left(N_{i.}^{(H)} - \sum_{j \notin T} N_{ij}^{(L)} \right), \sum_{j \in T} N_{ij}^{(H)} \right)$$

pour tout $T \subseteq \{1, 2, ..., C\}$.

4. Calage mitigé

On peut ne pas être satisfait avec l'approche en deux temps de l'estimation par calage, où on cherche d'abord les vecteurs de poids qui satisfont au mieux l'équation de calage, puis parmi cet ensemble de vecteurs, on cherche celui qui se rapproche le plus des poids de Horvitz-Thompson. Pour de petits échantillons, cette méthode peut conduire à des poids que le statisticien considère trop éloignés des poids de Horvitz-Thompson.

On pourrait préférer doser l'importance qu'on accorde à l'équation de calage par rapport à la norme de $\boldsymbol{w}_s - \boldsymbol{A}_s \boldsymbol{c}_s$. Ainsi, on peut désirer trouver un vecteur de poids $\boldsymbol{w}_s$ qui minimise

$$\left\| \begin{pmatrix} \boldsymbol{w}_s - \boldsymbol{A}_s \boldsymbol{c}_s \\ \boldsymbol{X}_s' \boldsymbol{w}_s - \boldsymbol{X}' \boldsymbol{c} \end{pmatrix} \right\|_V^2,$$

où

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{U}_s & \boldsymbol{0} \\ \boldsymbol{0} & \alpha\boldsymbol{T} \end{pmatrix}$$

et $\alpha \geq 0$. On minimise alors

$$\|\boldsymbol{w}_s - \boldsymbol{A}_s\boldsymbol{c}_s\|^2_{\boldsymbol{U}_s} + \alpha\|\boldsymbol{X}_s'\boldsymbol{w}_s - \boldsymbol{X}'\boldsymbol{c}\|^2_{\boldsymbol{T}} = D_s(\boldsymbol{w}_s) + \alpha\|\boldsymbol{X}_s'\boldsymbol{w}_s - \boldsymbol{X}'\boldsymbol{c}\|^2_{\boldsymbol{T}}.$$

Un problème de minimisation semblable est rencontré lors d'une régression avec coefficients de coûts (ridge regression). Pour $\alpha = 0$ la solution est donnée par les poids de Horvitz-Thompson $\boldsymbol{w}_s = \boldsymbol{A}_s\boldsymbol{c}_s$. Tandis que pour $\alpha > 0$, on cherche $\boldsymbol{w}_s(\alpha)$ qui minimise $\|\boldsymbol{K}(\boldsymbol{w}_s - \boldsymbol{A}_s\boldsymbol{c}_s) - \boldsymbol{b}\|^2_{\boldsymbol{V}}$, où $\boldsymbol{K} = (\boldsymbol{I}_n, \boldsymbol{X}_s)'$, $\boldsymbol{b} = (\boldsymbol{0}_{1\times n}, (\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s)')'$ et $\boldsymbol{0}_{1\times n} \in \mathbb{R}^n$ est un vecteur-ligne de 0. Ben-Israel et Greville (1980) donne

$$\boldsymbol{w}_s(\alpha) - \boldsymbol{A}_s\boldsymbol{c}_s = (\boldsymbol{K}'\boldsymbol{V}\boldsymbol{K})^{-1}\boldsymbol{K}'\boldsymbol{V}\boldsymbol{b}. \qquad (10)$$

Donc on obtient en substituant les valeurs de $\boldsymbol{K}$, $\boldsymbol{V}$, et $\boldsymbol{b}$

$$\boldsymbol{w}_s(\alpha) = \boldsymbol{A}_s\boldsymbol{c}_s + \alpha(\boldsymbol{U}_s + \alpha\boldsymbol{X}_s\boldsymbol{T}\boldsymbol{X}_s')^{-1}\boldsymbol{X}_s\boldsymbol{T}(\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s). \qquad (11)$$

On montre facilement que

$$\alpha(\boldsymbol{U}_s + \alpha\boldsymbol{X}_s\boldsymbol{T}\boldsymbol{X}_s')^{-1}\boldsymbol{X}_s\boldsymbol{T} = \boldsymbol{U}_s^{-1}\boldsymbol{X}_s(\alpha^{-1}\boldsymbol{T}^{-1} + \boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s)^{-1},$$

d'où

$$\boldsymbol{w}_s(\alpha) = \boldsymbol{A}_s\boldsymbol{c}_s + \boldsymbol{U}_s^{-1}\boldsymbol{X}_s(\alpha^{-1}\boldsymbol{T}^{-1} + \boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s)^{-1}(\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s). \qquad (12)$$

L'estimateur $\boldsymbol{Y}_s'\boldsymbol{w}_s(\alpha)$ prend donc la forme $\hat{\boldsymbol{Y}}'\boldsymbol{c} + (\boldsymbol{Y}_s - \hat{\boldsymbol{Y}}_s)'\boldsymbol{A}_s\boldsymbol{c}_s$, où $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_s(\alpha)$ et

$$\hat{\boldsymbol{\beta}}_s(\alpha) = (\boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s + \alpha^{-1}\boldsymbol{T}^{-1})^{-1}\boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{Y}_s.$$

Le vecteur des coefficients de régression est donc celui qu'on obtient lors d'une régression avec coefficients de coûts (ridge regression). Tout comme la méthode d'estimation par calage et la méthode d'estimation par régression généralisée telle que décrite dans Särndal, Swensson et Wretman (1992), conduisent aux mêmes estimateurs, un parallèle semblable peut être fait entre le calage mitigé et la régression avec coefficients de coûts.

À partir de l'équation (12), on peut également utiliser Ben-Israel et Greville (1980), et le fait que $\boldsymbol{F}^\dagger = \boldsymbol{F}'(\boldsymbol{F}\boldsymbol{F}')^\dagger$ avec $\boldsymbol{F} = \boldsymbol{T}^{1/2}\boldsymbol{X}_s'\boldsymbol{U}_s^{-1/2}$, pour montrer que

$$\lim_{\alpha\to\infty} \boldsymbol{w}_s(\alpha) = \boldsymbol{w}_{\text{cal}}.$$

C'est le résultat auquel on devait s'attendre, puisque trouver le vecteur $\boldsymbol{w}_s(\alpha)$ qui minimise $D_s(\boldsymbol{w}_s) + \alpha\|\boldsymbol{X}_s'\boldsymbol{w}_s - \boldsymbol{X}'\boldsymbol{c}\|^2_{\boldsymbol{T}}$ lorsque $\alpha \to \infty$, revient à trouver le vecteur de poids qui minimise $D_s(\boldsymbol{w}_s)$ parmi ceux qui minimisent $\|\boldsymbol{X}_s'\boldsymbol{w}_s - \boldsymbol{X}'\boldsymbol{c}\|^2_{\boldsymbol{T}}$.

Rao et Singh (1997) ont défini des tolérances pour chacune des p contraintes de l'équation de calage, et ils ont établi une relation entre ces tolérances et la matrice $\alpha\boldsymbol{T}$.

Pour $\alpha \in [0, \infty[$ la fonction $\boldsymbol{w}_s(\alpha)$ décrit une courbe dans $\mathbb{R}^n$ qui relie le point $\boldsymbol{A}_s\boldsymbol{c}_s$ au point $\boldsymbol{w}_{\text{cal}}$. Si $p = 1$, c'est-à-dire si $\boldsymbol{X}$ est un vecteur, cette courbe est en fait un segment de droite. En effet, dans ce cas la matrice $(\alpha^{-1}\boldsymbol{T}^{-1} + \boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s)^{-1}$ et le vecteur $\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s$ sont des scalaires, les poids $\boldsymbol{w}_s(\alpha)$ donnés par (12) sont donc égaux aux poids de Horvitz-Thompson plus un multiple du vecteur $\boldsymbol{U}_s^{-1}\boldsymbol{X}_s$. Puis toujours pour $p = 1$, on a

$$\lim_{\alpha\to\infty} \boldsymbol{w}_s(\alpha) = \boldsymbol{w}_{\text{cal}} = \boldsymbol{A}_s\boldsymbol{c}_s + [(\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s)/(\boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s)]\,\boldsymbol{U}_s^{-1}\boldsymbol{X}_s$$

ce qui donne l'estimateur

$$\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}} = \boldsymbol{Y}_s'\boldsymbol{A}_s\boldsymbol{c}_s + [(\boldsymbol{Y}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s)/(\boldsymbol{X}_s'\boldsymbol{U}_s^{-1}\boldsymbol{X}_s)](\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s)$$

Si on prend $\boldsymbol{U} = \boldsymbol{A}^{-1}\,\text{diag}(\boldsymbol{X})$, on obtient l'estimateur par le quotient

$$\boldsymbol{Y}_s'\boldsymbol{A}_s\boldsymbol{c}_s + [(\boldsymbol{Y}_s'\boldsymbol{A}_s\boldsymbol{1}_{n\times 1})/(\boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{1}_{n\times 1})](\boldsymbol{X}'\boldsymbol{c} - \boldsymbol{X}_s'\boldsymbol{A}_s\boldsymbol{c}_s),$$

où $\boldsymbol{1}_{a\times b} \in \mathbb{R}^{a\times b}$ est une matrice de 1.

Ben-Israel et Greville (1980, 111, exercice 15) montrent que $D_s(\boldsymbol{w}_s(\alpha))$ est une fonction monotone croissante de α. Il faut cependant noter que pour une unité $k \in s$, $|w_k(\alpha) - a_k c_k|$ n'est pas nécessairement une fonction monotone de α. Lorsque α augmente, le vecteur de poids $\boldsymbol{w}_s(\alpha)$ s'éloigne du vecteur des poids de Horvitz-Thompson, mais ce n'est pas nécessairement le cas pour chaque coordonnée de ce vecteur.

Dans cet article on utilisera le calage mitigé pour restreindre les poids, donc lorsque la taille d'échantillon est relativement petite. Il est cependant facile de montrer que pour un cadre asymptotique qui satisfait (2) et pour lequel $\hat{\boldsymbol{\beta}}_s(\alpha) - \boldsymbol{\beta}(\alpha) \to \boldsymbol{0}$ en probabilité, avec

$$\boldsymbol{\beta}(\alpha) = (\boldsymbol{X}'\boldsymbol{U}^{-1}\boldsymbol{X} + \alpha^{-1}\boldsymbol{T}^{-1})^{-1}\boldsymbol{X}'\boldsymbol{U}^{-1}\boldsymbol{Y},$$

on a $\boldsymbol{Y}_s'\boldsymbol{w}_s(\alpha)$ est un estimateur asymptotiquement non biaisé dont la variance asymptotique est

$$(\boldsymbol{Y} - \boldsymbol{Y}^*)'\,\text{diag}(\boldsymbol{c})(\boldsymbol{A}\prod\boldsymbol{A} - \boldsymbol{1}_{N\times N})\,\text{diag}(\boldsymbol{c})(\boldsymbol{Y} - \boldsymbol{Y}^*),$$

où $\boldsymbol{Y}^* = \boldsymbol{X}\boldsymbol{\beta}(\alpha)$, $\prod$ est la matrice des probabilités d'inclusion de second ordre, et $\text{diag}(\boldsymbol{c})$ est la matrice diagonale formée à partir du vecteur $\boldsymbol{c}$.

5. Méthodes d'estimation avec poids restreints

Afin d'éviter d'obtenir des poids ayant des valeurs extrêmes, on peut vouloir forcer le vecteur de poids à être à l'intérieur d'une région déterminée. On supposera que cette région de restriction est convexe et fermée, et que $\boldsymbol{A}_s\boldsymbol{c}_s$ est un point de cette région. Par exemple, si $\boldsymbol{w}^{(L)} < \boldsymbol{A}_s\boldsymbol{c}_s < \boldsymbol{w}^{(H)}$, on pourrait vouloir restreindre les poids à la région $R_{\boldsymbol{w}} = \{\boldsymbol{w}_s : \boldsymbol{w}^{(L)} \le \boldsymbol{w}_s \le \boldsymbol{w}^{(H)}\}$. On supposera que

$$\lim_{n\to\infty} \boldsymbol{w}^{(L)} - \boldsymbol{A}_s\boldsymbol{c}_s < \boldsymbol{0} \text{ et } \lim_{n\to\infty} \boldsymbol{w}^{(H)} - \boldsymbol{A}_s\boldsymbol{c}_s > \boldsymbol{0}.$$

L'approche mentionnée à la section 3 consiste à choisir une mesure de distance entre les poids calés et les poids de Horvitz-Thompson qui donnera des poids satisfaisant l'équation de calage qui sont dans la région de restriction, ceci pourvu que de tels poids existent. L'approche qui sera étudiée dans cette section consiste à tempérer l'exigence de satisfaire l'équation de calage lorsque le vecteur des poids de calage $\boldsymbol{w}_{\text{cal}}$ est à l'extérieur de la région de restriction. Différentes façons de tempérer cette exigence conduisent à différentes méthodes de pondération.

Lorsque $\boldsymbol{w}_{\text{cal}}$ est à l'extérieur de la région de restriction, on pourrait par exemple, chercher les points de la courbe $\boldsymbol{w}_s(\alpha)$ paramétrisée par $\alpha \ge 0$ qui sont sur la frontière de cette région. Ces points ont la propriété d'être une solution du problème de minimisation décrit dans la section 4 pour les valeurs de α correspondantes, donc à travers la matrice $\boldsymbol{T}$, on peut pondérer l'importance de chaque équation de calage. Avec l'exemple de la région de restriction donné plus haut, si

$$\boldsymbol{w}_{\text{cal}} = \lim_{\alpha\to\infty} \boldsymbol{w}_s(\alpha)$$

est à l'intérieur de cette région, alors on peut prendre comme vecteur de poids restreints $\boldsymbol{w}_{\text{res1}} = \boldsymbol{w}_{\text{cal}}$, sinon on peut prendre $\boldsymbol{w}_{\text{res1}} = \boldsymbol{w}_s(\alpha)$ pour $\alpha < \infty$ tel que $\boldsymbol{w}_s(\alpha)$ est à la frontière de la région de restriction. Si le cadre asymptotique est tel que les conditions (2) sont respectées avec $\gamma < 3/2$ alors pour n suffisamment grand, la probabilité que $\boldsymbol{w}_{\text{cal}}$ soit à l'intérieur de la région de restriction est égale à un. En effet, on a $\boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\boldsymbol{c}_s$ converge en probabilité vers $\boldsymbol{0}$. Les propriétés asymptotiques de l'estimateur qui utilise les poids restreints, $\boldsymbol{w}_{\text{res1}}$, sont donc les mêmes que celles de l'estimateur par calage. Il est important de noter que puisque $|w_k - a_k c_k|$ n'est pas nécessairement une fonction monotone de α, il est possible que $\boldsymbol{w}_s(\alpha)$ soit à la frontière de la région de restriction pour plusieurs valeurs de α, même si la région de restriction est convexe. Il n'est pas nécessairement simple de trouver toutes ces valeurs, et il faut décider laquelle utiliser.

Une autre option pour restreindre les poids consisterait à se donner comme région de restriction, les poids $\boldsymbol{w}_s$ qui satisfont $D_s(\boldsymbol{w}_s) \le l$ pour une borne $l > 0$. On prend ensuite comme vecteur de poids restreints $\boldsymbol{w}_{\text{res2}} = \boldsymbol{w}_{\text{cal}}$, si $\boldsymbol{w}_{\text{cal}}$ est dans la région de restriction, sinon on cherche $\alpha > 0$ tel que $D_s(\boldsymbol{w}_s(\alpha)) = l$. Cette valeur de α est unique et peut être trouvée de façon itérative. On calcule ensuite les poids $\boldsymbol{w}_{\text{res2}} = \boldsymbol{w}_s(\alpha)$ qui correspondent à cette valeur de α à l'aide de l'équation (12). Si le cadre asymptotique est tel que les conditions (2) sont respectées avec $\gamma < 1$, et si l ne varie pas avec n, alors pour n suffisamment grand, la probabilité que $\boldsymbol{w}_{\text{cal}}$ soit à l'intérieur de la région de restriction est égale à un. En effet, on a $D_s(\boldsymbol{w}_{\text{cal}})$ converge en probabilité vers 0. Les propriétés asymptotiques de l'estimateur qui utilise les poids restreints, $\boldsymbol{w}_{\text{res2}}$, sont alors les mêmes que celles de l'estimateur par calage. Malheureusement, lorsqu'on estime un total il faut s'attendre à avoir $\gamma = 1$. Pour pallier cet inconvénient, on peut utiliser $l\sqrt{n}$ comme borne, plutôt que l. On peut justifier cette borne par le fait que la longueur de la diagonale principale d'un hypercube de $\mathbb{R}^n$ est égale au diamètre de la boule qui circonscrit cet hypercube, par contre, le diamètre de la boule inscrite dans ce même hypercube est plus petit par un facteur de $\sqrt{n}$. Il reste que le statisticien peut être inconfortable avec l'utilisation d'un cadre asymptotique où la borne croit avec la taille de l'échantillon. Il y a aussi le fait que cette approche ne permette pas de limiter individuellement les poids des observations. Seule la distance entre le vecteur des poids restreints et le vecteur des poids de Horvitz-Thompson est contrôlée.

Avec les méthodes décrites plus haut, on cherche les points de la courbe $\boldsymbol{w}_s(\alpha)$ qui sont sur la frontière de la région de restriction. La valeur de α pour laquelle $\boldsymbol{w}_s(\alpha)$ est à la frontière de la région de restriction doit souvent être trouvée de façon itérative. On pourrait plus simplement remplacer la courbe $\boldsymbol{w}_s(\alpha)$ par le segment de droite reliant $\boldsymbol{A}_s\boldsymbol{c}_s$ à $\boldsymbol{w}_{\text{cal}}$. Pour la région de restriction $R_{\boldsymbol{w}}$, on aurait comme vecteur de poids restreints $\boldsymbol{w}_{\text{res3}} = \boldsymbol{w}_{\text{cal}}$, si $\boldsymbol{w}_{\text{cal}}$ est dans la région de restriction, et sinon $\boldsymbol{w}_{\text{res3}}$ serait égal au point où le segment de droite traverse la frontière de la région de restriction, c'est-à-dire

$$\boldsymbol{w}_{\text{res3}} = \boldsymbol{A}_s\boldsymbol{c}_s + \xi(\boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\boldsymbol{c}_s),$$

où

$$\xi = \min_k \{\max[(\boldsymbol{w}^{(L)} - \boldsymbol{A}_s\boldsymbol{c}_s)/(\boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\boldsymbol{c}_s), (\boldsymbol{w}^{(H)} - \boldsymbol{A}_s\boldsymbol{c}_s)/(\boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\boldsymbol{c}_s)]\},$$

la division des vecteurs se faisant élément par élément, le maximum des deux vecteurs étant pris élément par élément, et $\min_k$ donnant l'élément minimum. On pourrait aussi considérer le vecteur de poids de la région de restriction, $\boldsymbol{w}_{\text{res4}}$,

qui est le plus près de $\boldsymbol{w}_{\text{cal}}$. Toujours pour la région de restriction $R_{\boldsymbol{w}}$, on aurait

$$\boldsymbol{w}_{\text{res}\,4} = \min[\max(\boldsymbol{w}_{\text{cal}},\ \boldsymbol{w}^{(L)}),\ \boldsymbol{w}^{(H)}].$$

Les propriétés asymptotiques des estimateurs qui utilisent les poids restreints $\boldsymbol{w}_{\text{res}\,3}$ ou $\boldsymbol{w}_{\text{res}\,4}$ sont les mêmes que celles de l'estimateur par calage, pourvu que $\boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\,\boldsymbol{c}_s$ converge en probabilité vers $\boldsymbol{0}$, ce qui est normalement le cas.

Une propriété intéressante de toutes les approches discutées dans cette section est que quelle que soit la région de restriction, l'existence de poids restreints est assurée. Ce n'est pas toujours le cas avec une approche basée sur des mesures de distance. On va maintenant présenter un exemple simple pour comparer quelques-unes des méthodes.

On veut estimer un total à partir d'un échantillon aléatoire simple de taille 2 d'une population de taille 20. C'est-à-dire $\boldsymbol{c} = \boldsymbol{1}_{20\times 1}$ et $\boldsymbol{a} = 10(\boldsymbol{1}_{20\times 1})$. On utilise le vecteur d'information auxiliaire $\boldsymbol{X} = (1, 2, 3, ..., 20)'$, on suppose que l'échantillon obtenu est $s = \{2, 12\}$ et on choisit de prendre $\boldsymbol{U}$ une matrice diagonale avec $u_{kk} = x_k = k$. On se donne une région de restriction rectangulaire à l'aide des points $\boldsymbol{w}^{(L)} = (0,\ 0)'$ et $\boldsymbol{w}^{(H)} = (20,\ 13)'$. C'est-à-dire que le poids de la première unité de l'échantillon doit être supérieur à 0 et inférieur à 20, tandis que le poids de la seconde unité de l'échantillon doit être supérieur à 0 et inférieur à 13.

Sous ces conditions, les poids calés $\boldsymbol{w}_{\text{cal}} = (15,\ 15)'$ sont à l'extérieur de la région de restriction. Puisque $p = 1$, les poids $\boldsymbol{w}_s(\alpha)$ sont sur le segment de droite qui relie $\boldsymbol{A}_s\,\boldsymbol{c}_s = (10,\ 10)'$ à $\boldsymbol{w}_{\text{cal}}$. On a donc $\boldsymbol{w}_{\text{res}1} = \boldsymbol{w}_{\text{res}\,3}$, c'est-à-dire que les deux méthodes donnent le même résultat. Dans ce cas-ci on a $\boldsymbol{w}_{\text{res}1} = \boldsymbol{w}_{\text{res}\,3} = (13,\ 13)'$. La méthode qui consiste à choisir le point de la région de restriction le plus près des poids calés donne $\boldsymbol{w}_{\text{res}\,4} = (15,\ 13)'$. Par contre, si on cherche $\boldsymbol{w}_{\text{res}\,5}$, les poids restreints obtenus en continuant d'exiger que l'équation de calage soit satisfaite et en utilisant une mesure de distance qui prend une valeur infinie à l'extérieur de la région de restriction, alors il n'y a pas de solution. En effet, pour tout poids dans la région de restriction $\boldsymbol{X}_s'\,\boldsymbol{w}_s \le 196$, tandis que $\boldsymbol{X}'\boldsymbol{c} = 210$. Si on avait, disons $\boldsymbol{w}^{(H)} = (30,\ 13)'$, alors avec $D_s(\boldsymbol{w}_s)$ comme mesure de distance à l'intérieur de la région de restriction on aurait $\boldsymbol{w}_{\text{res}\,5} = (27,\ 13)'$. Ces poids sont plutôt éloignés de $\boldsymbol{w}_{\text{cal}} = (15,\ 15)'$ et de $\boldsymbol{A}_s\,\boldsymbol{c}_s = (10,\ 10)'$. C'est le prix qu'il faut payer si on tient à avoir des poids qui respectent l'équation de calage.

6. Estimateurs pour domaines avec une composante synthétique

On utilise des poids restreints à cause des propriétés de l'estimateur par calage pour de petites tailles d'échantillon. Pour de grandes tailles d'échantillon, on sait qu'on a normalement $\boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\,\boldsymbol{c}_s$ qui converge en probabilité vers $\boldsymbol{0}$, donc des poids qui ne sont pas problématiques. Un statisticien qui est confronté au problème de poids extrêmes doit donc vraisemblablement faire face à un autre problème lié aux petites tailles d'échantillon, à savoir l'estimation pour de petits domaines. On présentera dans cette section, un estimateur dont les propriétés asymptotiques sont celles de l'estimateur par calage, mais qui utilise des poids restreints et tire profit d'un estimateur synthétique.

Soit $\tilde{\boldsymbol{Y}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}_s$ une estimation synthétique pour $\boldsymbol{Y}$, on a

$$\begin{aligned}
\tilde{\boldsymbol{Y}}'\boldsymbol{w}_s &= (\boldsymbol{X}_s\tilde{\boldsymbol{\beta}}_s)'\boldsymbol{w}_s \\
&= \tilde{\boldsymbol{\beta}}_s'\boldsymbol{X}_s'\boldsymbol{w}_s \\
&\simeq \tilde{\boldsymbol{\beta}}_s'\boldsymbol{X}'\boldsymbol{c} \\
&= (\boldsymbol{X}\tilde{\boldsymbol{\beta}}_s)'\boldsymbol{c} \\
&= \tilde{\boldsymbol{Y}}'\boldsymbol{c} \qquad (13)
\end{aligned}$$

avec égalité pour la troisième expression de droite si les poids satisfont l'équation de calage $\boldsymbol{X}_s'\boldsymbol{w}_s = \boldsymbol{X}'\boldsymbol{c}$. Les poids $\boldsymbol{w}_{\text{cal}}$ donnés par (1) minimisent $||\boldsymbol{X}_s'\boldsymbol{w}_s - \boldsymbol{X}'\boldsymbol{c}||_{\boldsymbol{T}}^2$. On peut donc estimer $\boldsymbol{Y}'\boldsymbol{c}$ par

$$\hat{\boldsymbol{\tau}} = (\boldsymbol{Y}_s - \tilde{\boldsymbol{Y}}_s)'\boldsymbol{w}_{\text{res}} + \tilde{\boldsymbol{Y}}'\boldsymbol{c}. \qquad (14)$$

Il y aura égalité entre cet estimateur et l'estimateur $\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}}$ dès que l'échantillon sera suffisamment grand pour que l'équation de calage soit satisfaite et pour que $\boldsymbol{w}_{\text{cal}}$ soit dans la région de restriction, c'est-à-dire dès que $\boldsymbol{w}_{\text{res}} = \boldsymbol{w}_{\text{cal}}$. Les propriétés asymptotiques de ces deux estimateurs sont donc les mêmes sous certaines conditions discutées dans la section précédente. L'avantage de l'estimateur $\hat{\boldsymbol{\tau}}$ est qu'il donne une estimation synthétique lorsque des colonnes de $\boldsymbol{Y}_s$ et $\tilde{\boldsymbol{Y}}_s$ sont nulles.

7. Données aberrantes

On pourrait traiter les données aberrantes d'une façon similaire aux poids extrêmes. La stratégie est la suivante : on se donne une région de restriction pour $\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}}$, on montre que pour n suffisamment grand $\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}}$ est à l'intérieur de cette région de restriction, et on se donne un estimateur plus « raisonnable » pour remplacer $\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}}$ dans les cas où $\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}}$ est à l'extérieur de la région de restriction. Dans le cas d'un échantillon stratifié on aurait normalement une région de restriction par strate.

On a montré à la section 2 que sous certaines conditions sur le cadre asymptotique, $\boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\,\boldsymbol{c}_s = O_p(n^{-3/2}N^{\gamma})$. On a donc $\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}} - \boldsymbol{Y}_s'\boldsymbol{A}_s\,\boldsymbol{c}_s = O_p(n^{-1/2}N^{\gamma})$, et si on suppose que

$$\boldsymbol{Y}_s'\boldsymbol{A}_s\,\boldsymbol{c}_s - \boldsymbol{Y}'\boldsymbol{c} = O_p(n^{-1/2}N^{\gamma}), \qquad (15)$$

alors $\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}} - \boldsymbol{Y}'\boldsymbol{c} = O_p(n^{-1/2}N^{\gamma})$. Un expert (ou un groupe d'experts) peut déterminer à partir d'informations indépendantes des données d'enquête, qu'il ne serait pas

raisonnable d'avoir $\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}}$ à l'extérieur d'une certaine région. Si $\boldsymbol{Y}'\boldsymbol{c}$ est dans la région de restriction (c'est-à-dire si l'expert ne juge pas déraisonnable une estimation du paramètre qui serait égale à la vraie valeur, $\boldsymbol{Y}'\boldsymbol{c}$, du paramètre), si $\gamma = 0$, et si la région de restriction ne varie pas avec n ou N (ou si $\gamma = 1$, et la région de restriction varie proportionnellement à N), alors pour n suffisamment grand, la probabilité que $\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}}$ soit à l'intérieur de la région de restriction est égale à un. Dans les cas où $\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}}$ est à l'extérieur de la région de restriction, on pourrait utiliser comme estimation le point de la région de restriction qui est le plus près de $\boldsymbol{Y}_s'\boldsymbol{w}_{\text{cal}}$ ou on pourrait poser égal à un le poids des quelques observations qui sont jugées aberrantes, et répartir leurs poids originaux (moins le nombre d'observations aberrantes) sur les observations non aberrantes. Les propriétés asymptotiques de cet estimateur modifié pour traiter les valeurs aberrantes sont alors les mêmes que celles de l'estimateur non modifié.

Dans le cas d'un échantillon non stratifié cette méthode est relativement facile à appliquer. Si par contre l'échantillon est stratifié, et si des contraintes sont imposées aux estimations de chaque strate, alors on doit faire face à deux problèmes additionnels. Premièrement, si le cadre asymptotique est tel que le nombre de strates augmente de façon proportionnelle à la taille d'échantillon, alors la supposition donnée en (15) ne tient pas, puisque la taille moyenne d'échantillon par strate reste constante alors que $n \to \infty$. Il s'agit de savoir s'il est raisonnable de se donner un cadre asymptotique où le nombre de strates est constant (ou croît moins vite que n). Un tel cadre asymptotique est moins plausible si le nombre d'observations par strate est petit. Le deuxième problème est la difficulté pour l'expert d'imposer des contraintes aux estimations pour chacune des strates. Plus il y a de strates, plus le risque est grand que $\boldsymbol{Y}'\boldsymbol{c}$ ne soit pas dans la région de restriction définie par l'expert. En fait, dans le cas d'un échantillon stratifié, il est préférable que l'expert utilise les informations indépendantes des données d'enquêtes, afin de s'assurer de l'homogénéité des strates, avant que la stratification soit finalisée. En d'autres mots il vaut mieux utiliser l'information qui est disponible avant l'enquête, pour prévenir les données aberrantes, plutôt que pour les corriger. Si l'information a été utilisée de sorte qu'avant l'enquête, on n'a aucune raison de croire qu'il y a dans quelque strate une observation non représentative, alors on n'a aucune justification pour prétendre le contraire après la collecte des données.

8. Conclusion

Si pour de grandes tailles d'échantillons, les poids calés demeurent à l'intérieur d'une région de restriction, alors les propriétés asymptotiques de l'estimateur avec poids restreints sont bien sûr identiques à celles de l'estimateur par calage. Pour un cadre asymptotique donné, on peut habituellement s'attendre à avoir $\boldsymbol{w}_{\text{cal}} - \boldsymbol{A}_s\boldsymbol{c}_s$ qui converge en probabilité vers $\boldsymbol{0}$, donc pour des tailles d'échantillons suffisamment grandes, les poids calés $\boldsymbol{w}_{\text{cal}}$ demeureront à l'intérieur de la région de restriction $R_{\boldsymbol{w}}$ si $\boldsymbol{A}_s\boldsymbol{c}_s$ est à l'intérieur de $R_{\boldsymbol{w}}$. Cependant, on a vu que pour l'estimation d'un total, on n'a pas nécessairement convergence vers 0 de $D_s(\boldsymbol{w}_{\text{cal}})$. Une région de restriction définie par $\|\boldsymbol{w}_s - \boldsymbol{A}_s\boldsymbol{c}_s\|_{U_s}^2 \le l$ est donc à éviter, à tout le moins si on estime un total plutôt qu'une moyenne.

On a donné des conditions nécessaires et suffisantes pour l'existence de poids restreints à des intervalles qui satisfont l'équation de calage. Si de tels poids n'existent pas, il faut alors abandonner l'idée de satisfaire exactement l'équation de calage. On peut reformuler le problème de calage avec poids restreints, de manière à ce qu'une solution soit toujours possible. Quelques-unes des approches présentées dans cet article permettent d'obtenir une solution sans avoir recours à des méthodes itératives. Il s'agit de méthodes simples, faciles à interpréter. Les propriétés asymptotiques de ces estimateurs sont habituellement identiques à celles de l'estimateur par calage sans restrictions sur les poids.

Le problème de poids extrêmes survient avec des tailles d'échantillon qui sont petites, donc le problème d'estimation pour de petits domaines devrait être envisagé en même temps. Il est possible de tirer profit d'estimateurs synthétiques tout en ayant un estimateur avec des poids restreints qui a de bonnes propriétés asymptotiques.

On peut aussi modifier l'estimateur par calage, ou tout autre estimateur asymptotiquement convergent, pour traiter les données aberrantes. Les conditions pour que cet estimateur modifié ait les mêmes propriétés asymptotiques que l'estimateur non modifié ne sont pas facilement vérifiables, tout comme il est difficile de vérifier qu'une donnée aberrante est effectivement aberrante (non représentative). Cependant, ces conditions permettent d'identifier les facteurs qui font qu'un estimateur corrigé pour données aberrantes peut être statistiquement valable.

Remerciements

L'auteur tient à remercier un éditeur associé et un arbitre pour leurs remarques constructives qui ont permis d'améliorer cet article.

Annexe A

On veut vérifier que $\Omega(\boldsymbol{\varphi}) = \boldsymbol{l}'(\boldsymbol{V}\boldsymbol{\varphi})_- - \boldsymbol{h}'(\boldsymbol{V}\boldsymbol{\varphi})_+$ est inférieur ou égal à zéro. Premièrement, il est facile de montrer que ceci est vrai pour un vecteur $\boldsymbol{\varphi}$, si et seulement si c'est vrai pour un vecteur $k\boldsymbol{\varphi}$ avec $k > 0$ arbitraire. Seulement la direction de $\boldsymbol{\varphi}$ importe. Il est donc suffisant de vérifier la condition pour $\boldsymbol{\varphi}$ de norme égale à un. Pour la preuve, on utilisera la norme 1_1 de $\|\boldsymbol{\varphi}\|_{l_1} = \sum_{i=1}^{p} |\varphi_i|$. Les vecteurs $\boldsymbol{\varphi}$ avec $\|\boldsymbol{\varphi}\|_{l_1} = 1$ sont situés dans des hyperplans dont les intersections sont à des points perpendiculaires aux vecteurs unité, c'est-à-dire des points dont au

moins une des coordonnées est zéro. La fonction Ω varie linéairement sauf à des points $\boldsymbol{\varphi}$ perpendiculaires à une ou plusieurs rangées de $\boldsymbol{V}$. Même lorsque le domaine de Ω est restreint aux vecteurs $\boldsymbol{\varphi}$ avec $\|\boldsymbol{\varphi}\|_{l_1}=1$ qui sont perpendiculaires à $0 \le j < (p-1)$ rangées de $\boldsymbol{V}$ linéairement indépendantes, la fonction Ω varie encore linéairement sauf à des points perpendiculaires à d'autres rangées de $\boldsymbol{V}$ ou perpendiculaires à des vecteurs unité (lesquels sont également des rangées de $\boldsymbol{V}$. Le maximum de Ω pour $\|\boldsymbol{\varphi}\|_{l_1}=1$ est donc atteint en un point $\boldsymbol{\varphi}$ perpendiculaire à $(p-1)$ rangées de $\boldsymbol{V}$ linéairement indépendantes. Il est donc suffisant de vérifier la condition pour deux vecteurs de direction opposée qui sont perpendiculaires à $(p-1)$ rangées de $\boldsymbol{V}$ linéairement indépendantes et ce, pour chaque sous-ensemble de $(p-1)$ rangées de $\boldsymbol{V}$ linéairement indépendantes.

Annexe B

On note vec$(\boldsymbol{F})$ le vecteur obtenu en empilant les colonnes successives de la matrice $\boldsymbol{F} \in \mathbb{R}^{a\times b}$ avec la première colonne au-dessus, et on définit le produit de Kronecker de deux matrices $\boldsymbol{F}$ et $\boldsymbol{G}$ comme

$$\boldsymbol{F} \otimes \boldsymbol{G} = \begin{pmatrix} f_{11}\boldsymbol{G} & \cdots & f_{1n}\boldsymbol{G} \\ \vdots & & \vdots \\ f_{m1}\boldsymbol{G} & \cdots & f_{mn}\boldsymbol{G} \end{pmatrix}. \qquad \text{(B1)}$$

Le résultat découle du corollaire de la section 3 avec

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{I}_{RC} \\ \boldsymbol{I}_R \otimes \mathbf{1}_{1\times C} \\ \mathbf{1}_{1\times R} \otimes \boldsymbol{I}_C \\ \mathbf{1}_{1\times RC} \end{pmatrix}, \quad \boldsymbol{w} = \text{vec}((\hat{N}_{ij})'),$$

$$\boldsymbol{l} = \begin{pmatrix} \text{vec}((N_{ij}^{(L)})') \\ N_{1.}^{(L)} \\ \vdots \\ N_{R.}^{(L)} \\ N_{.1}^{(L)} \\ \vdots \\ N_{.C}^{(L)} \\ N_{..}^{(L)} \end{pmatrix}, \boldsymbol{h} = \begin{pmatrix} \text{vec}((N_{ij}^{(H)})') \\ N_{1.}^{(H)} \\ \vdots \\ N_{R.}^{(H)} \\ N_{.1}^{(H)} \\ \vdots \\ N_{.C}^{(H)} \\ N_{..}^{(H)} \end{pmatrix}. \qquad \text{(B2)}$$

Seul un ensemble fini de conditions doit être vérifié, premièrement en notant que les colonnes de

$$\boldsymbol{V} = \begin{pmatrix} -\boldsymbol{I}_R \otimes \mathbf{1}_{C\times 1} & -\mathbf{1}_{R\times 1} \otimes \boldsymbol{I}_C & -\mathbf{1}_{RC\times 1} \\ \boldsymbol{I}_R & \mathbf{0}_{R\times C} & \mathbf{0}_{R\times 1} \\ \mathbf{0}_{C\times R} & \boldsymbol{I}_C & \mathbf{0}_{C\times 1} \\ \mathbf{0}_{1\times R} & \mathbf{0}_{1\times C} & 1 \end{pmatrix} \qquad \text{(B3)}$$

forment une base pour $N(\boldsymbol{M}')$. C'est-à-dire que $\boldsymbol{M}'\boldsymbol{V}=\boldsymbol{0}$, les colonnes de $\boldsymbol{V}$ sont linéairement indépendantes, et $N(\boldsymbol{M}')$ est de dimension $R+C+1$. En notant également que les dernières $R+C+1$ lignes de $\boldsymbol{V}$ sont les vecteurs unité. Et finalement, en vérifiant les conditions du corollaire pour tous les vecteurs $\boldsymbol{\lambda}=\boldsymbol{V}\boldsymbol{\varphi}$ et $\boldsymbol{\lambda}=-\boldsymbol{V}\boldsymbol{\varphi}$, où $\boldsymbol{\varphi}$ est perpendiculaire à $R+C$ lignes de $\boldsymbol{V}$ linéairement indépendantes. Cette dernière étape est élaborée plus en détails dans le paragraphe suivant.

Un sous-ensemble arbitraire de $R+C$ lignes de $\boldsymbol{V}$ linéairement indépendantes qui inclut la dernière ligne de $\boldsymbol{V}$ sera noté L, et le sous-ensemble de toutes les lignes de $\boldsymbol{V}$ qui sont des combinaisons linéaires de lignes de L sera noté L^+. Si L^+ inclut la ligne $RC+i$ $(i=1, ..., R)$ si et seulement si $i \notin S \subseteq \{1, 2, ..., R\}$, et inclut la ligne $RC+R+j$ $(j=1, ..., C)$ si et seulement si $j \notin T \subseteq \{1, 2, ..., C\}$, alors on pose $\boldsymbol{\varphi}=(\boldsymbol{\varphi}_S', -\boldsymbol{\varphi}_T', 0)'$, où le $i^{\text{ième}}$ élément de $\boldsymbol{\varphi}_S \in \mathbb{R}^R$ est égal à 1 si $i \in S$ et à zéro sinon, et le $j^{\text{ième}}$ élément de $\boldsymbol{\varphi}_T \in \mathbb{R}^C$ est égal à un si $j \in T$ et à zéro sinon. Alors

$$\boldsymbol{V}\boldsymbol{\varphi} = ((-\boldsymbol{\varphi}_S \otimes \mathbf{1}_{C\times 1} + \mathbf{1}_{R\times 1} \otimes \boldsymbol{\varphi}_T)', \boldsymbol{\varphi}_S', -\boldsymbol{\varphi}_T', 0)',$$

donc $\boldsymbol{\varphi}$ est perpendiculaire à toutes les lignes de L^+, et à plus forte raison, $\boldsymbol{\varphi}$ est perpendiculaire à toutes les lignes de L. De même, le vecteur $\boldsymbol{\varphi}^* = (\boldsymbol{\varphi}_S', \boldsymbol{\varphi}_T', -1)'$ est perpendiculaire à toutes les lignes d'un sous-ensemble de $R+C$ lignes de $\boldsymbol{V}$ linéairement indépendantes qui inclut la ligne $RC+i$ $(i=1, ..., R)$ si et seulement si $i \notin S$, et inclut la ligne $RC+R+j$ $(j=1, ..., C)$ si et seulement si $j \notin T$, mais qui n'inclut pas la dernière ligne de $\boldsymbol{V}$. La condition $-\boldsymbol{l}'\boldsymbol{\lambda}_- \le \boldsymbol{h}'\boldsymbol{\lambda}_+$ avec $\boldsymbol{\lambda}=\boldsymbol{V}\boldsymbol{\varphi}$ donne le cinquième ensemble d'inéquations dans (9). De même, en posant $\boldsymbol{\lambda}$ égal à $-\boldsymbol{V}\boldsymbol{\varphi}$, $\boldsymbol{V}\boldsymbol{\varphi}^*$ et $-\boldsymbol{V}\boldsymbol{\varphi}^*$ on obtient les trois derniers ensembles d'inéquations dans (9).

Bibliographie

Bacharach, M. (1965). Estimating nonnegative matrices from marginal data. *International Economic Review*, 6, 294-310.

Bardsley, P., et Chambers, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

Ben-Israel, A., et Greville, T.N.E. (1980). *Generalized Inverses*: *Theory and Applications*. Huntington, New York : Robert E. Krieger Publishing Company.

Brewer, K.R.W. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Duchesne, P. (1999). Estimateurs de calage robustes. *Techniques d'enquête*, 25, 47-60.

Fan, K. (1956). On systems of linear inequalities. *Annals of Mathematics Studies*, (Éds. H. W. Kuhn, et A. W. Tucker), 38, 99-156.

Graybill, F.A. (1983). *Matrices with Applications in Statistics*, (Deuxième édition). Belmont, California: Wadsworth Publishing.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Rao, J.N.K., et Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.