Prise en compte des équilibrages dans l'estimation de la variance transversale EEC-TH

version définitive

Karim MOUSSALLAM(*)

(*) Insee, Unité de Méthodologie Statistique-Ménages

Introduction

L'étude explore comment faire pour mettre en pratique la version 'optimale' de l'approximation de la variance d'un sondage équilibré proposée par Deville et Tillé [DT].

Le tableau 1 des tests de calcul de variance de l'article [DT] suggère assez clairement une supériorité empirique de la technique d'approximation qui identifie les coefficients diagonaux de la forme quadratique approximante avec ceux de l'estimateur sans biais de la variance. En effet, la pondération ainsi optimisée fournit la plus faible sous-estimation 'maximale', en un sens défini dans l'article, pour 6 des 9 plans de sondage simulés, et la plus petite surestimation maximale pour 4 cas. De plus, elle n'est la pire méthode pour aucun des contextes testés, sur ces deux critères. Les résultats obtenus par A.Matei et Y.Tillé pour un tirage de taille fixe à entropie maximale [AMYT] paraissent plus mitigés. Cependant, ils ne sont pas défavorables à cette méthode. Elle évite les surestimations et sous-estimations les plus extrêmes.

Compte tenu de la généralisation de l'équilibrage dans les plans de sondage des ménages réalisés par l'Insee, et de l'absence d'alternative connue à ce mode d'estimation de la variance d'un sondage équilibré, il parait utile d'étudier les moyens de mettre en œuvre la technique d'optimisation de la pondération, et d'évaluer sa qualité pour l'estimation de variance sur les données réelles d'un sondage particulier.

Le principe de choisir la pondération des résidus qui rapproche la diagonale de la forme quadratique approximante de celle de l'estimateur sans biais de la variance reste éclairant pour améliorer la qualité de l'estimation de variance dans le contexte étudié. Pour l'appliquer, l'algorithme récursif est simple, rapide et produit la pondération la plus minimisante sur les données étudiées, par rapport aux alternatives testées. Au final toutefois, l'amélioration de la qualité de l'estimation de variance par cette optimisation ressort moins clairement que dans les tests de [DT]. L'absence de prise en compte directe de la variance d'atterrissage (cf [GC]) contribue sans doute à limiter l'enjeu de la pondération pour la qualité de l'estimation de variance et complique l'évaluation de sa méthode.

La première partie de l'étude s'intéresse aux propriétés théoriques de l'approximation de variance et à la technique d'optimisation de la pondération. La seconde partie évalue l'efficacité des choix envisagés sur des simulations de l'échantillonnage de première phase de l'enquête Emploi dans la Taxe d'Habitation (appelée EEC-TH dans la suite). Le plan de ce sondage est décrit dans [VL]. Les notations utilisées dans la suite sont explicitées en Annexe G.

1 rappel du principe d'estimation de la variance d'un sondage équilibré

– Pour un sondage sans remise, la variance de l'estimateur d'Horvitz-Thompson du total Y d'une variable d'intérêt y sur une population $\mathscr{P}$ est une forme quadratique en $\frac{y}{\pi}(\mathscr{P})$:

$$Var\left(\widehat{Y}\right) = \sum_{i,j\in\mathscr{P}} \frac{y}{\pi}(i)\frac{y}{\pi}(j)\left(\pi(i,j) - \pi(i)\pi(j)\right) \qquad \textbf{(e1.1)}$$

$$(\text{où } \pi(i) = p(i \in s) > 0)$$

Si la probabilité d'inclusion double ($\pi(i,j) = p(i,j \in s)$) ne s'annule pas sur $\mathscr{P}^2$ alors l'estimateur sans biais (ESB) de la variance est la forme quadratique en $\frac{y}{\pi}(s)$ [1] :

$$\widehat{Var}_{HT}\left(\widehat{Y}\right) = \widehat{V}_{HT} = \sum_{i,j\in s} \frac{y}{\pi}(i)\frac{y}{\pi}(j)\frac{\pi(i,j) - \pi(i)\pi(j)}{\pi(i,j)} \qquad \textbf{(e1.2)}$$

$$= Q^{HT}\left(\frac{y}{\pi}(s)\right)$$

– La forme quadratique (**e1.1**) est positive, car exactement égale à $Var\left[\sum_s \frac{y}{\pi}\right]$. Par contre, ce n'est pas toujours le cas pour l'estimateur sans biais (**e1.2**).

→ Si l'échantillonnage est de taille fixe et de probabilités d'inclusion doubles non nulles et si la condition de Yates-Grundy ($\forall i \neq j, \pi(i)\pi(j) \geqslant \pi(i,j)$) est satisfaite alors :

$$\widehat{V}_{YG} = \frac{1}{2}\sum_{i,j\in s}\left(\frac{y}{\pi}(i) - \frac{y}{\pi}(j)\right)^2 \frac{\pi(i)\pi(j) - \pi(i,j)}{\pi(i,j)} \qquad \textbf{(e1.3)}$$

est positif et estime sans biais la variance. En général $\widehat{V}_{YG} \neq \widehat{V}_{HT}$, sauf pour un SAS [2]. La diagonale de la forme quadratique de $\widehat{V}_{YG}$ en fonction de y/π a pour coefficient générique :

$$1 - \pi_i + \sum_{j\in s}\frac{\pi_i\pi_j}{\pi_{i,j}} - |s| \quad (\geqslant 1 - \pi_i \text{ si taille fixe, vu que } \pi_j \geqslant \pi_{ij})$$

– Pour un sondage équilibré d'entropie maximale, c'est-à-dire un sondage poissonnien p^* conditionné par l'équation d'équilibrage $\left(p(s) = p^*\left[s\middle|\widehat{X} = x(+)\right]\right)$, la variance asymptotique est approximée en régressant la variable d'intérêt par les variables d'équilibrage avec la pondération $c^* = \frac{\pi^*(1-\pi^*)}{\pi}$:

$$Var_{p^*}\left[\widehat{Y}\middle|\widehat{X} = x(+)\right] \cong Var_{p^*}\left(\widehat{Y} - \widehat{X}'\beta^*\right) \text{ avec}$$

$$\beta^* \in \underset{\beta}{\operatorname{argmin}}\, Var_{p^*}\left(\widehat{Y} - \widehat{X}'\beta\right), \text{ et comme } p^* \text{ est poissonnien :}$$

$$Var_{p^*}\left[\widehat{Y}\middle|\widehat{X} = x(+)\right] \cong \sum_{\mathscr{P}}\left(\frac{y}{\pi} - \frac{x'}{\pi}\beta^*\right)^2 \pi^*(1-\pi^*) \text{ où :}$$

$$\beta^* \in \underset{\beta}{\operatorname{argmin}} \sum_{\mathscr{P}}\left(\frac{y}{\pi} - \frac{x'}{\pi}\beta\right)^2 \pi^*(1-\pi^*)$$

Soit M^* la mesure sur $\mathscr{P}$ pondérée par $\pi^*(1-\pi^*)$ et $\widehat{M^*}$ la mesure empirique associée au plan de sondage effectif (telle que $E_p\left(\widehat{M^*}\right) = M^*$) : $\widehat{M^*}(z) = \sum_s z\frac{\pi^*(1-\pi^*)}{\pi}$.

---

1. C'est la seule forme quadratique en $y(s)$ à coefficients fixes $/s$ qui soit sans biais pour toute variable $y(\mathscr{P})$.
2. Pour un sondage aléatoire simple (échantillons de tailles fixes et équiprobables) :
$$\widehat{V}_{HT}(\widehat{Y}) = \frac{N(N-n)}{n}\mathcal{S}^2_s(y) = \frac{N(N-n)}{n}\frac{1}{n-1}\sum_s (y-\overline{y}_s)^2 = \frac{1}{2}\frac{N(N-n)}{n-1}\frac{1}{n^2}\sum_{i,j\in s}(y_i - y_j)^2$$

→ L'estimation approximative de la variance équilibrée consiste à remplacer la mesure inconnue $\widehat{M^*}$ par une approximation $\widetilde{M^*}$ calculable sur l'échantillon sous la forme : $\widetilde{M^*}(z) = \sum_s cz$. Par la méthode de linéarisation [JCD], l'approximation de variance est formulée ainsi :

$$\widehat{Var}\left(\widehat{Y}\right) = \widetilde{M^*}\left[\left(\frac{y}{\pi} - \frac{x}{\pi}'\widehat{\beta}\right)^2\right] = Q\left(\frac{y}{\pi}(s)\right), \text{ avec :} \quad \textbf{(e1.4)}$$

$$\widehat{\beta} \in \underset{\beta}{\operatorname{argmin}}\, \widetilde{M^*}\left[\left(\frac{y}{\pi} - \frac{x}{\pi}'\beta\right)^2\right]$$

⇒ En pratique, la variance est estimée par la somme pondérée des carrés des résidus du ratio de la variable d'intérêt sur la probabilité d'inclusion régressé par le vecteur d'équilibrage. La pondération d'équilibrage est c, et le résidu d'équilibrage est $(y - x\widehat{\beta})/\pi$.

Il faut souligner que la justification méthodologique de ce type d'estimateur de variance est asymptotique. De plus, l'approximation est affectée d'un biais fonction de la proximité entre le plan de sondage effectif et celui équilibré d'entropie maximale. Une approximation moins forte de l'estimation de variance est possible, en estimant par simulation les probabilités d'inclusion double [GC], au lieu d'approcher seulement les probabilités d'inclusion simple.

→ Il est clair que l'enjeu méthodologique essentiel dans le cadre de l'approximation [DT] de la variance est le choix de la pondération $c(s)$ des résidus d'équilibrage. C'est l'objet de l'étude.

2 propriétés de la forme quadratique approximante

2.1 expression comme fonction de la pondération

– L'estimateur de variance défini par (**e1.4**) est encore une forme quadratique en $\frac{y}{\pi}(s)$. Avec les notations $u = \frac{x}{\pi}$ et $z = \frac{y}{\pi}$, il peut se formuler ainsi :

$$\widehat{Var}\left(\widehat{Y}\right) = \min_{\beta}\left\|\frac{y}{\pi} - \frac{x}{\pi}'\beta\right\|_c^2 = \min_{\beta}\|z - u_s\beta\|_c^2 \quad \textbf{(e2.5)}$$

$$= \left\|\text{proj}_{\text{Im}(u_s)}^{c^\perp}(z)\right\|_c^2 \ (c \text{ est confondue avec la métrique diagonale correspondante, } *c)$$

$$= z^s \text{proj}_{\text{Im}(u_s)}^{c^\perp}{}' * c \text{proj}_{\text{Im}(u_s)}^{c^\perp} z_s = z^s * c \text{proj}_{\text{Im}(u_s)}^{c^\perp} z_s = z^s\left(*c - c\text{proj}_{\text{Im}(u_s)}^{c}\right) z_s$$

$$= z^s\left(*c - cu_s\left(u^s c u_s\right)^- u^s c\right) z_s \text{ d'où la formule explicite :}$$

$$\widehat{Var}\left(\widehat{Y}\right) = z^s \Delta(c)\, z_s \quad \textbf{(e2.6)}$$

$$\text{avec : } \Delta(c) = \text{diag}(c) - (cu)_s\left[\textstyle\sum_s cuu'\right]^-(cu)^s \quad \textbf{(e2.7)}$$

– La formule (**e2.7**) peut également se déduire de l'équation normale :
$\min_{\beta}\|z - u_s\beta\|_c^2 = \left\|z - u_s\widehat{\beta}\right\|_c^2$ où $\widehat{\beta}$ est déterminée [3] par :

$$\left\langle z - u_s\widehat{\beta}, u_s\right\rangle_c = 0 \text{ (équation normale)} \quad \textbf{(e2.8)}$$

d'où : $\left\|z - u_s\widehat{\beta}\right\|_c^2 = \left\langle z - u_s\widehat{\beta}, z\right\rangle_c = \|z\|_c^2 - \left\langle u_s\widehat{\beta}, z\right\rangle_c$

$$= z' * cz - z' * cu_s\left(\sum cuu'\right)^- u^s * cz = z' * c\left(\text{id} - u_s\left(\sum cuu'\right)^- u^s * c\right) z$$

– Plus explicitement, l'estimateur de la variance équilibrée peut se calculer ainsi :

$$\widehat{V} = \sum_s c\left(\frac{y}{\pi}\right)^2 - \sum_s \frac{yx'}{\pi^2}\left[\sum_s c\frac{xx'}{\pi^2}\right]^- \sum_s c\frac{yx}{\pi^2} \quad \textbf{(e2.9)}$$

3. de manière unique si et seulement si u_s est de plein rang colonne.

– Dans la suite, pour l'indicatrice de l'élément $i \in s$, $\mathbb{1}_i$, le coefficient de la régression est noté $\widehat{\beta}_i$ ou $\widehat{\beta}_c(\mathbb{1}_i)$.

– Vu (**e2.7**) et (**e2.5**), les coefficients diagonaux $\delta(c)$ de la forme quadratique approximante se formulent de deux façons, la première utile au calcul et la seconde pour l'étude de leurs propriétés :

$$\delta_i(c) = \Delta_{i,i}(c) = c\left(1 - cu'\left(\sum cuu'\right)^- u\right)(i) = c_i(1 - u_i'\widehat{\beta}_i) \quad \textbf{(e2.10)}$$

$$= \min_\beta \|\mathbb{1}_i - u_s\beta\|_c^2 = \left\|\mathbb{1}_i - u_s\widehat{\beta}_i\right\|_c^2 = \sum_s c(\mathbb{1}_i - u'\widehat{\beta}_i)^2 \quad \textbf{(e2.11)}$$

$$= \|\mathbb{1}_i\|_c^2 - \left\|u_s\widehat{\beta}_i\right\|_c^2 = c_i - \left\|u_s\widehat{\beta}_i\right\|_c^2 \quad \textbf{(e2.12)}$$

– Il découle de ces égalités que $c_i(1 - u_i'\widehat{\beta}_i)u_i'\widehat{\beta}_i = \sum_{s \neq i} c\left[u'\widehat{\beta}_i\right]^2$ et donc $0 \leqslant u_i'\widehat{\beta}_i \leqslant 1$.

Cet encadrement de la ième coordonnée de la projection de $\mathbb{1}_i$ se construit de manière plus satisfaisante ainsi :

$$u_i'\widehat{\beta}_i = \left\langle \mathbb{1}_i, \text{proj}^c_{\text{Im}(u_s)}[\mathbb{1}_i]\right\rangle = \frac{1}{c_i}\left\langle \mathbb{1}_i, \text{proj}^c_{\text{Im}(u_s)}[\mathbb{1}_i]\right\rangle_c = \frac{1}{c_i}\left\|\text{proj}^c_{\text{Im}(u_s)}[\mathbb{1}_i]\right\|_c^2 \leqslant \frac{1}{c_i}\|\mathbb{1}_i\|_c^2 \leqslant 1$$

– L'expression (**e2.10**) entraîne que $u'\left(\sum cuu'\right)^- u$ est constant sur les inverses généralisées de $\sum cuu'$. Il est utile pour la suite de l'étude d'en analyser le mécanisme.

• La démonstration la plus directe utilise le fait que $\left[\sqrt{c}u\right]_s\left(\sum cuu'\right)^-\left[\sqrt{c}u\right]^s$ est la matrice de la projection orthogonale sur $\text{Im}\left\{\left[\sqrt{c}u\right]_s\right\}$(**eC.65**), donc ne dépend pas du choix de l'inverse généralisée.

• $\text{Im}\left(\sum cuu'\right) = \text{Im}\left\{\left[\sqrt{c}u\right]^s\left[\sqrt{c}u\right]_s\right\} = \text{Im}\left\{\left[\sqrt{c}u\right]^s\right\}$ (par (**eA.57**)). D'où :

$$\text{Im}\left(\sum cuu'\right) = \text{Im}\left[u^s * \sqrt{c}\right] = \text{Im}(u^s) \text{ (vu que } \text{Im}\left\{\text{diag}\left[\sqrt{c}\right]\right\} = \mathbb{R}_s \text{ si } c > 0)$$

$$\Rightarrow \forall c > 0, \text{Im}\left(\sum cuu'\right) = \text{Im}(u^s) = \text{Im}\left(\sum uu'\right) \quad \textbf{(e2.13)}$$

→ (**e2.13**)$\Rightarrow \forall c > 0, \text{rang}\left(\sum cuu'\right) = \text{rang}(u^s)$ $(= \text{rang}(u_s))$

⇒ si $c > 0$ alors (**e2.13**)$\Rightarrow u_i \in \text{Im}\left(\sum cuu'\right)$ car $u_i = u^s\mathbb{1}_i \in \text{Im}(u^s) = \text{Im}\left(\sum cuu'\right)$

$$\Rightarrow \text{ si } c > 0, \exists a_i \,/\, u_i = \sum cuu'a_i \quad \textbf{(e2.14)}$$

(soit synthétiquement : $\exists A \,/\, u_s = A\sum cuu'$)

$$\Rightarrow u_i'\left(\sum cuu'\right)^- u_j = a_i'\left(\sum cuu'\right)\left(\sum cuu'\right)^-\left(\sum cuu'\right)a_j = a_i'\left(\sum cuu'\right)a_j = \sum c(a_i'u)(a_j'u)$$

donc ne dépend pas du choix de l'inverse généralisée.

→ La forme quadratique approximante (**e2.5**) est manifestement positive. De plus elle s'annule pour $z_s \in \text{Im}(u_s) \Leftrightarrow \exists\beta \,/\, \frac{y}{\pi}(s) = \frac{x}{\pi}(s)\beta \Leftrightarrow y_s \in \text{Im}(x_s)$. C'est une bonne propriété de l'approximation de variance, puisque la variance du total estimé d'une variable qui vérifie cette condition pour tout s est effectivement nulle si l'équilibrage est exact :

$$\sum_s \frac{y}{\pi} = \sum_s \left(\frac{x}{\pi}\right)'\beta = x(+)'\beta = y(+)$$

– L'approximation de variance est une fonction de la pondération c notée $\varphi(c) = \varphi_z(c) = \min_\beta \|z - u_s\beta\|_c^2$. Cette fonction peut être définie ainsi sur $\mathbb{R}_s^+$. Outre la positivité, elle possède plusieurs propriétés intéressantes. L'étude de celles-ci est également utile pour $\Delta(c)$ et $\delta(c)$.

– Comme $\text{Im}(u_s)$ est un sous-espace vectoriel de dimension finie, le minimum $\varphi(c) = \min_\beta \|z - u_s\beta\|_c^2$ est toujours atteint par la projection orthogonale de z_s sur l'image de u_s pour la métrique c. Donc il existe au moins un $\widehat{\beta}_c \in \text{argmin}_\beta \|z - u_s\beta\|_c^2$ et avec un tel vecteur $\varphi(c) = \left\|z - u_s\widehat{\beta}_c\right\|_c^2 = \|z\|_c^2 - \left\|u_s\widehat{\beta}_c\right\|_c^2$. Cette formule de la fonction ne dépend pas du vecteur choisi dans $\text{argmin}_\beta \|z - u_s\beta\|_c$ (Annexe C). (Lorsque c'est nécessaire, la notation plus précise $\widehat{\beta}_c(z)$ est utilisée.)

– La fonction φ est croissante au sens où :

$$c \leqslant \widetilde{c} \quad (\Leftrightarrow \forall i, c_i \leqslant \widetilde{c}_i) \Rightarrow \forall\beta \|z - u_s\beta\|_c^2 \leqslant \|z - u_s\beta\|_{\widetilde{c}}^2 \Rightarrow \varphi(c) \leqslant \varphi(\widetilde{c})$$

– Homogénéité de degré 1 :

$$\forall \lambda \geqslant 0, \varphi(\lambda c) = \min_{\beta} \|z - u_s\beta\|^2_{\lambda c} = \min_{\beta} \lambda \|z - u_s\beta\|^2_c = \lambda \min_{\beta} \|z - u_s\beta\|^2_c = \lambda\varphi(c)$$

Un intérêt pratique de cette propriété est qu'à partir d'une pondération c telle que $\sum_s \delta(c) > 0$, il est toujours possible de construire une pondération $\widetilde{c}$ dont la somme $\sum_s \delta(\widetilde{c})$ est égale à la trace de l'estimateur sans biais, simplement par un facteur multiplicatif.

– $\varphi(c)$ est concave sur $\mathbb{R}^+_s$:

$$\forall \alpha \in [0,1], \forall \beta, \|z - u_s\beta\|^2_{\alpha c + (1-\alpha)\widetilde{c}} = \alpha \|z - u_s\beta\|^2_c + (1-\alpha) \|z - u_s\beta\|^2_{\widetilde{c}}$$
$$\geqslant \alpha\varphi(c) + (1-\alpha)\varphi(\widetilde{c})$$

– Cette propriété équivaut à la convexité en c de $\left\|u_s\widehat{\beta}_c\right\|^2_c$ $\left(= \sum cu'z\left(\sum cuu'\right)^- \sum cuz\right)$, comme somme de deux fonctions convexes : $\left\|u_s\widehat{\beta}_c\right\|^2_c = \|z\|^2_c - \varphi(c)$ [4].

– si $a \geqslant 0$ et $c \geqslant 0$ alors $\varphi(a+c) \geqslant \varphi(a) + \varphi(c)$ car $\varphi(a+c) = \left\|z - u_s\widehat{\beta}_{c+a}\right\|^2_{a+c} = \left\|z - u_s\widehat{\beta}_{c+a}\right\|^2_a + \left\|z - u_s\widehat{\beta}_{c+a}\right\|^2_c \geqslant \left\|z - u_s\widehat{\beta}_a\right\|^2_a + \left\|z - u_s\widehat{\beta}_c\right\|^2_c$

2.2 continuité sur $\mathbb{R}^+_s$:

– semi-continuité supérieure : $\varphi(c) = \left\|z - u_s\widehat{\beta}_c\right\|^2_{\widetilde{c}} + \left\|z - u_s\widehat{\beta}_c\right\|^2_{c-\widetilde{c}} \geqslant \varphi(\widetilde{c}) + \left\|z - u_s\widehat{\beta}_c\right\|^2_{c-\widetilde{c}}$
Comme ce dernier terme est une fonction linéaire de $c - \widetilde{c}$, $\overline{\lim}\,\varphi(c) \leqslant \varphi(c)$ ($\overline{\lim}$ désigne la limite supérieure).

– semi-continuité inférieure : $\underline{\lim}\,\varphi(c) \geqslant \varphi(c)$ est vrai pour $c = 0$, car $\varphi(0) = 0$. Supposons $c \neq 0$. Alors pour $\|\widetilde{c} - c\|$ assez petit, $\widetilde{c}[\{c > 0\}] > 0$, et sur ce voisinage de c dans $\mathbb{R}^+_s$:

$$\varphi(c) \leqslant \left\|z - u_s\widehat{\beta}_{\widetilde{c}}\right\|^2_c \leqslant \left\|z - u_s\widehat{\beta}_{\widetilde{c}}\right\|_{\widetilde{c}} \max_{c>0}\left(\frac{c}{\widetilde{c}}\right) = \varphi(\widetilde{c}) \max_{c>0}\left(\frac{c}{\widetilde{c}}\right)$$

La limite inférieure de la dernière expression égale $\underline{\lim}\,\varphi(c)$, car $\lim\limits_{\widetilde{c}\longrightarrow c} \max\limits_{c>0}\left(\frac{c}{\widetilde{c}}\right) = 1 > 0$.

• Une démonstration alternative est éclairante, bien que limitée à un ouvert :

– D_+ l'ensemble des matrices définies positives est un ouvert : si $\min\limits_{\|x\|=1} x'Ax > 0$ alors $\min\limits_{\|x\|=1} x'Bx \geqslant \min\limits_{\|x\|=1} x'Ax - \|B - A\|$ parce que $x'(B-A)x \geqslant -\|B-A\|\,\|x\|^2$ [5]

→ $\varphi(A) = \min\limits_{\beta} \|z - u_s\beta\|^2_A$ est continue sur D_+ :

▷ Comme A est définie positive et que toutes les normes sont équivalentes en dimension finie, $\exists \lambda(A) > 0 \,/\, \forall x, \|x\|^2 \leqslant \lambda(A)\|x\|^2_A$ (avec $\breve{A} = (A + A')/2$, $\lambda(A) = \left\|\breve{A}^{-1}\right\|$ est une solution vu que $\|x\|^2 = x'\breve{A}^{\frac{1}{2}}\breve{A}^{-1}\breve{A}^{\frac{1}{2}}x \leqslant \left\|\breve{A}^{-1}\right\| \left\|\breve{A}^{\frac{1}{2}}x\right\|^2 = \left\|\breve{A}^{-1}\right\| x'Ax$)

▷ soit $\mathcal{V}(A)$ un voisinage de A dans D_+ tel que $\left\|\mathcal{V}(A) - A\right\| < 1/\lambda(A)$

▷ pour $\widetilde{A} \in \mathcal{V}(A)$ et pour tout β :

$$\left|\|z - u_s\beta\|^2_A - \|z - u_s\beta\|^2_{\widetilde{A}}\right| = \left|(z - u_s\beta)'(A - \widetilde{A})(z - u_s\beta)\right|$$
$$\leqslant \left\|\widetilde{A} - A\right\| \|z - u_s\beta\|^2 \leqslant \left\|\widetilde{A} - A\right\| \lambda(A) \|z - u_s\beta\|^2_A$$
$$\Rightarrow \|z - u_s\beta\|^2_A \left[1 - \left\|\widetilde{A} - A\right\| \lambda(A)\right] \leqslant \|z - u_s\beta\|^2_{\widetilde{A}} \leqslant \|z - u_s\beta\|^2_A \left[1 + \left\|\widetilde{A} - A\right\| \lambda(A)\right]$$
$$\Rightarrow \varphi(A)\left[1 - \left\|\widetilde{A} - A\right\| \lambda(A)\right] \leqslant \varphi(\widetilde{A}) \leqslant \varphi(A)\left[1 + \left\|\widetilde{A} - A\right\| \lambda(A)\right] \text{ (car } 1 - \left\|\widetilde{A} - A\right\| \lambda(A) > 0)$$
$$\Rightarrow \left|\varphi(\widetilde{A}) - \varphi(A)\right| \leqslant \left\|\widetilde{A} - A\right\| \lambda(A) \varphi(A)$$

---

4. Incidemment, avec $z = \mathbb{1}_i$, il en résulte que $c_i^2 u_i' \left(\sum cuu'\right)^- u_i \left(= \left\|u_s\widehat{\beta}_c(\mathbb{1}_i)\right\|^2_c\right)$ est convexe en c_s.

5. Si $\|\|$ désigne la norme euclidienne canonique, $|x'(B-A)x| \leqslant \|x\|\,\|(B-A)x\| \leqslant \|x\|\,\|B-A\|\,\|x\|$, vu (**eE.69**).

- majoration explicite de la variation de $\delta_i(c)$, pour $c > 0$:

$$(\min\{\widetilde{c}/\,c\} - 1)\,\delta_i(c) \leqslant \delta_i(\widetilde{c}) - \delta_i(c) \leqslant (\max\{\widetilde{c}/\,c\} - 1)\,\delta_i(c) \text{ donc :}$$
$$|\delta_i(\widetilde{c}) - \delta_i(c)| \leqslant \max[|\widetilde{c}/\,c - 1|]\,\delta_i(c)$$

2.3 dérivabilité sur $\mathbb{R}_s^{+*}$

Pour appliquer à $\delta(c)$ un algorithme de minimisation utilisant le gradient, il faut que cette fonction soit dérivable. Le calcul de la dérivée $\dot{\delta}(c)$ est utilisé non seulement pour programmer ces algorithmes mais également pour étudier l'existence et l'unicité de solutions.

- La démonstration est effectuée plus généralement pour $\varphi(c) = \left\|z - u_s\widehat{\beta}_c\right\|_{A(c)}$ où $A(c)$ est une fonction dérivable à valeur dans l'ensemble des matrices symétriques définies positives.

hypothèses • $A(c)$ dérivable en c

• A est symétrique définie positive sur $\mathcal{V}(c)$ [6]

soit $c + h \in \mathcal{V}(c)$

$$\begin{aligned}
\varphi(c+h) - \varphi(c) &= \left\|z - u_s\widehat{\beta}_{c+h}\right\|^2_{A(c+h)} - \left\|z - u_s\widehat{\beta}_c\right\|^2_{A(c)} \\
&= \left\|z - u_s\widehat{\beta}_c\right\|^2_{A(c+h)} - \left\|u_s(\widehat{\beta}_{c+h} - \widehat{\beta}_c)\right\|^2_{A(c+h)} - \left\|z - u_s\widehat{\beta}_c\right\|^2_{A(c)} \qquad \textbf{(e2.15)} \\
&\quad \text{(par Pythagore)} \\
\left\langle u_s(\widehat{\beta}_{c+h} - \widehat{\beta}_c), u_s\right\rangle_{A(c+h)} &= \left\langle z - u_s\widehat{\beta}_c, u_s\right\rangle_{A(c+h)} \text{ (équation normale pour } \widehat{\beta}_{c+h}) \\
&= \left\langle z - u_s\widehat{\beta}_c, u_s\right\rangle_{A(c+h)-A(c)} \text{ (équation normale pour } \widehat{\beta}_c) \\
&= \left\langle A(c+h)^{-1}(A(c+h) - A(c))(z - u_s\widehat{\beta}_c), u_s\right\rangle_{A(c+h)} \\
\Rightarrow \left\|u_s(\widehat{\beta}_{c+h} - \widehat{\beta}_c)\right\|^2_{A(c+h)} &\leqslant \left\|A(c+h)^{-1}(A(c+h) - A(c))(z - u_s\widehat{\beta}_c)\right\|_{A(c+h)} \left\|u_s(\widehat{\beta}_{c+h} - \widehat{\beta}_c)\right\|_{A(c+h)} \\
&\quad \text{(par l'inégalité de Cauchy-Schwartz)} \\
\Rightarrow \left\|u_s(\widehat{\beta}_{c+h} - \widehat{\beta}_c)\right\|_{A(c+h)} &\leqslant \left\|A(c+h)^{-1}(A(c+h) - A(c))(z - u_s\widehat{\beta}_c)\right\|_{A(c+h)} \\
&= \sqrt{(z - u_s\widehat{\beta}_c)'(A(c+h) - A(c))A(c+h)^{-1}(A(c+h) - A(c))(z - u_s\widehat{\beta}_c)} \\
&\leqslant \left\|z - u_s\widehat{\beta}_c\right\| \|A(c+h) - A(c)\| \sqrt{\left\|A(c+h)^{-1}\right\|} \\
&= O(h) \text{ [7]}
\end{aligned}$$

car le premier terme est constant /h, le second est un $O(h)$ par la dérivabilité de $A(c)$ et le dernier terme converge vers $\sqrt{\left\|A(c)^{-1}\right\|} < +\infty$ par les continuités de la racine carrée, de $A(c)$ et de l'inversion sur les matrices inversibles. Comme $\left\|u_s(\widehat{\beta}_{c+h} - \widehat{\beta}_c)\right\|_{A(c+h)} = O(h) \Rightarrow \left\|u_s(\widehat{\beta}_{c+h} - \widehat{\beta}_c)\right\|^2_{A(c+h)} = o(h)$, et la dérivabilité du premier terme de (**e2.15**) qui s'ensuit entraine celle de $\varphi(c)$, ainsi que l'expression de la dérivée.

→ La dérivée de φ peut se formuler ainsi :

$$\dot{\varphi}(c) = \left\|z - u_s\widehat{\beta}_c\right\|^2_{dc} \text{ où } \widehat{\beta}_c \in \underset{\beta}{\operatorname{argmin}} \left\|z - u_s\beta\right\|_c \qquad \textbf{(e2.16)}$$

– Plus simplement, la formule de la dérivée peut s'obtenir en traitant $\widehat{\beta}_c$ comme une fonction dérivable de c (c'est effectivement le cas avec le choix $\widehat{\beta}_c = \left(\sum cuu'\right)^+ \sum cuz$) :

$$\dot{\varphi}(c) = \left\|z - u_s\widehat{\beta}_c\right\|_{dc} + 2\left\langle z - u_s\widehat{\beta}_c, u_s d\widehat{\beta}_c\right\rangle_c = \left\|z - u_s\widehat{\beta}_c\right\|_{dc} \text{ (vu l'équation normale)}$$

---

6. donc $\|\|_{A(c)}$ est une norme euclidienne

7. $o(h) : \lim_{h\longrightarrow 0} \frac{\|o(h)\|}{\|h\|} = 0$ et $O(h) : \overline{\lim_{h\longrightarrow 0}} \frac{\|O(h)\|}{\|h\|} < +\infty$.

(C'est donc l'orthogonalité entre $z - u_s\widehat{\beta}_c$ et u_s qui fait disparaitre $d\widehat{\beta}_c$ de l'expression de $\dot{\varphi}(c)$. Il n'est pas rigoureux de considérer que $d\widehat{\beta}_c$ est négligeable en soit.)

– La norme de la dérivée peut s'expliciter ainsi, pour la norme du 'sup' $\|\|_\infty$:

$$\|\dot{\varphi}(c)\| = \sup_{\|h\|_\infty=1} \left|\sum \left[z - u'\widehat{\beta}_c\right]^2 h\right| \qquad \textbf{(e2.17)}$$

$$= \sum_s \left[z - u_s\widehat{\beta}_c\right]^2 \left(\leqslant \frac{\varphi(c)}{\min(c)} \text{ et } \geqslant \varphi(1)\right)$$

– L' Annexe D prouve que φ est deux fois dérivable sur $\{c > 0\}$, et même C^∞.

– Puisque pour tout y, la fonction $\varphi(c) = y'\Delta(c)\,y$ est dérivable, toutes les composantes de la matrice $\Delta(c)$ sont dérivables en c, et donc Δ est dérivable. Par suite sa diagonale δ est dérivable sur $\{c > 0\}$. Une formule explicite de la dérivée de $\delta(c)$ se déduit de (**e2.16**) :

$$\dot{\delta}_i(c) = \left\|\mathbb{1}_i - u_s\widehat{\beta}_c(\mathbb{1}_i)\right\|_{dc}^2 \qquad \textbf{(e2.18)}$$

$$\dot{\delta}_i(c)^j = \left(\mathbb{1}_i^j - u_j'\left(\sum cuu'\right)^- c_i u_i\right)^2 \qquad [8]$$

$$\dot{\delta}(c) = \left(\mathrm{id} - (cu)_s\left(\sum cuu'\right)^- u^s\right)^{*2} \qquad \textbf{(e2.19)}$$

La dérivée se formule aussi : $\dot{\delta}_s(c) = \left\|\mathbb{1}_s - u\widehat{\beta}_c(\mathbb{1}_s)\right\|_{dc}^2$ (l'indice intra-norme est occulté [9]).

→ expression des dérivées partielles de la diagonale (en utilisant (**e2.10**)) :

$$\frac{\partial\delta_i(c)}{\partial c_i} = \left(1 - u_i'\widehat{\beta}_c(\mathbb{1}_i)\right)^2 = \left(1 - u_i'\left(\sum cuu'\right)^- c_i u_i\right)^2 = \left[\frac{\delta_i(c)}{c_i}\right]^2 \qquad \textbf{(e2.20)}$$

$$i \neq j : \frac{\partial\delta_i(c)}{\partial c_j} = \left(u_j'\widehat{\beta}_c(\mathbb{1}_i)\right)^2 = \left(u_j'\left(\sum cuu'\right)^- c_i u_i\right)^2 \qquad \textbf{(e2.21)}$$

– Pour comparer les deux dérivées ci-dessus, il peut être utile d'utiliser les formules :

$$1 - u_i'\widehat{\beta}_i = \left\langle \mathbb{1}_i, \mathrm{proj}^{c^\perp}_{\mathrm{Im}(u_s)}[\mathbb{1}_i]\right\rangle = \frac{1}{c_i}\left\|\mathrm{proj}^{c^\perp}_{\mathrm{Im}(u_s)}[\mathbb{1}_i]\right\|_c^2 \qquad \textbf{(e2.22)}$$

$$\text{si } j \neq i, u_j'\widehat{\beta}_i = \left\langle \mathbb{1}_j, \mathrm{proj}^{c^\perp}_{\mathrm{Im}(u_s)}[\mathbb{1}_i]\right\rangle = \frac{1}{c_j}\left\langle \mathbb{1}_j, \mathrm{proj}^{c^\perp}_{\mathrm{Im}(u_s)}[\mathbb{1}_i]\right\rangle_c$$

$$= \frac{1}{c_j}\left\langle \mathrm{proj}^{c^\perp}_{\mathrm{Im}(u_s)}[\mathbb{1}_j], \mathrm{proj}^{c^\perp}_{\mathrm{Im}(u_s)}[\mathbb{1}_i]\right\rangle_c \qquad \textbf{(e2.23)}$$

→ de (**e2.20**) et (*l3.2*, ci-après) on déduit :

$$\text{si } \delta_i(1) > 0 \text{ alors } \delta_i(c) \text{ croit strictement en fonction de } c_i > 0 \qquad \textbf{(e2.24)}$$

• dérivabilité à droite en $c \geqslant 0$:

Pour $c \geqslant 0$ (et $c \neq 0$), la fonction peut encore se formuler ainsi :

$$\varphi(c) = \min_\beta \|z - u_s\beta\|_c^2 = \left\|z - u_s\widehat{\beta}_c\right\|_c^2$$

et pour $c + h \geqslant 0$ il reste vrai que :

$$\varphi(c+h) = \left\|z - u_s\widehat{\beta}_{c+h}\right\|_{c+h}^2 = \left\|z - u_s\widehat{\beta}_c\right\|_{c+h}^2 - \left\|u_s\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right)\right\|_{c+h}^2$$

d'où $\left\|u_s\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right)\right\|_{c+h}^2 = -\varphi(c+h) + \varphi(c) + \left\|z - u_s\widehat{\beta}_c\right\|_h^2$, qui tend vers 0 lorsque $h \longrightarrow 0$ et $c + h \geqslant 0$, vu la continuité de φ sur $\mathbb{R}_s^+$ et la linéarité en h du dernier terme. Il en découle que pour $c \geqslant 0$: $\lim_{\substack{h \longrightarrow 0 \\ c+h \geqslant 0}} \left\|u_s\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right)\right\|_{c+h}^2 = 0.$

- ou bien $c_i > 0$ et alors $\left|u_i'\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right)\right| \overset{h\longrightarrow 0}{\longrightarrow} 0$ (car $\leqslant \left\|u_s\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right)\right\|_{c+h} / (c_i + h_i)$)
- ou bien $c_i = 0$:

→ si $u_i \in \mathrm{Im}\left(u^{c>0}\right)$ alors comme $u_{c>0}\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right) \overset{h\longrightarrow 0}{\longrightarrow} 0$, $u_i'\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right) \overset{h\longrightarrow 0}{\longrightarrow} 0$

8. Si $\exists i \neq j \,/\, c_i \neq c_j$ et $u_j'\left(\sum cuu'\right)^- u_i \neq 0$ alors $\dot{\delta}_i(c)^j \neq \dot{\delta}_j(c)^i$ et $\dot{\delta}(c)$ n'est pas une matrice symétrique.

9. une solution pour expliciter le double indiçage $(s \times s)$: $\dot{\delta}_s(c) = \left\|\mathbb{1}_s - u_{(s)}\widehat{\beta}_c(\mathbb{1}_s)\right\|_{dc_{(s)}}^2$

→ si $u_i \notin \mathrm{Im}\left(u^{c>0}\right)$ alors nécessairement $z_i - u_i'\widehat{\beta}_{c+h} = 0$ (si $h_i > 0$), parce qu'il existe γ tel que $\begin{cases} u_i'\gamma = z_i \\ u_{c>0}\gamma = 0 \end{cases}$ (**e3.41**). Pour la même raison, $\widehat{\beta}_c$ peut toujours être choisi tel que $z_i - u_i'\widehat{\beta}_c = 0$, d'où $u_i'(\widehat{\beta}_{c+h} - \widehat{\beta}_c) = 0$.

⇒ Dans les deux sous-cas, $\left[u_i'\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right)\right]^2 h_i = o(h)$. Ceci prouve la dérivabilité à droite de φ en $c \geqslant 0$ parce que :

$$\left\|u_s\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right)\right\|_{c+h}^2 = \left\langle u_s\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right), u_s\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right)\right\rangle_{c+h} = \left\langle z - u_s\widehat{\beta}_c, u_s\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right)\right\rangle_{c+h}$$
$$= \left\langle z - u_s\widehat{\beta}_c, u_s\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right)\right\rangle_h = \sum h_i\left(z_i - u_i'\widehat{\beta}_c\right)u_i'\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right) = o(h)$$

L'expression de la dérivée partielle en $c_i = 0$ est :

$$\frac{\partial\varphi(c)}{\partial c_i}(c_i = 0) = \left\{z_i - u_i'\widehat{\beta}_c(z)\right\}^2 \text{ si } u_i \in \mathrm{Im}\left(u^{c>0}\right) \text{ et } 0 \text{ sinon} \qquad \textbf{(e2.25)}$$

3 cible et contraintes de l'optimisation

3.1 critère à minimiser

– Le cadre d'approximation de la variance [DT] est appliqué en choisissant une pondération qui rapproche le plus possible la forme quadratique approximante (**e2.5**) de l'estimateur sans biais (**e1.2**). Pour concrétiser cette démarche, l'idée proposée par [DT] est d'exploiter l'information connue sur celui-ci : la diagonale de la forme quadratique $Q^{HT}\left(\frac{y}{\pi}(s)\right)$ égale $(1-\pi)(s)$. Vu autrement, pour $z = \mathbb{1}_i$, (**e2.6**) donne $\widehat{V} = \delta_i(c)\,\mathbb{1}_i$, qui doit estimer $Var\left(\widehat{\pi\mathbb{1}_i}\right) = Var\left[\mathbb{1}_i\right] = \pi_i(1-\pi_i)$, et comme $\delta_i(c)$ est constante sur les échantillons, $E\left\{\delta_i(c)\,\mathbb{1}_i\right\} = \delta_i(c)\,\pi_i$.

L'équation de pondération est (**e3.26**). Comme le montre [DT], il est équivalent [10] de résoudre les deux équations :

$$\delta(c) = 1 - \pi \qquad \textbf{(e3.26)}$$

$$\text{et } \frac{\delta(c)}{\sum_s \delta(c)} = \frac{1-\pi}{\sum_s 1-\pi}, \text{ soit en posant } w = \frac{c}{\sum_s c} \text{ (pour } c > 0) :$$

$$\frac{\delta(w)}{\sum_s \delta(w)} = \frac{1-\pi}{\sum_s 1-\pi}, w > 0, \sum_s w = 1 \qquad \textbf{(e3.27)}$$

En effet, une solution c de (**e3.26**) donne une solution $w = c/\sum c$ de (**e3.27**). Inversement, si w est solution de (**e3.27**) alors $c = w\sum_s 1 - \pi/\sum_s \delta(w)$ est solution de (**e3.26**). Et ces deux applications sont inverses l'une de l'autre sur les deux ensembles de solutions.

– expressions de la diagonale 'réduite' :

$$F_i(w) = \frac{\delta_i(w)}{\sum_s \delta(w)} = \frac{\min_\beta \left\|\mathbb{1}_i - u_s\beta\right\|_w^2}{\sum_{j\in s}\min_\beta \left\|\mathbb{1}_j - u_s\beta\right\|_w^2} = \frac{w_i(1-u_i'\widehat{\beta}_i)}{\sum_{j\in s} w_j(1-u_j'\widehat{\beta}_j)} = \frac{w_i(1-w_iu_i'\left[\sum_s wuu'\right]^- u_i)}{\sum_{j\in s} w_j(1-w_ju_j'\left[\sum_s wuu'\right]^- u_j)} \qquad \textbf{(e3.28)}$$

La deuxième expression montre que cette fonction est le rapport de deux fonctions concaves. Sa convexité est donc indéterminée.

– Il n'a pas été possible d'étendre le résultat de [DT] sur la borne supérieure de $F(w)$, en raison de la présence dans la formule de la matrice $\left[\sum_s wuu'\right]^-$, au lieu d'une constante. A fortiori, cette formulation n'a pas permis d'expliciter une solution exacte en fonction de $1 - \pi_s$.

10. à la condition implicite que $\forall c > 0, \sum_s \delta(c) > 0$

→ La présente étude a privilégié l'équation (**e3.26**) parce que la fonction δ est plus simple.

Une pondération est qualifiée d'optimale si elle vérifie l'équation (**e3.26**), ou si elle s'en approche au maximum sous les contraintes imposées à la minimisation et pour une distance à déterminer. Cependant, cette approche naturelle [11] présuppose le choix d'un critère de proximité entre les deux diagonales, si l'égalisation stricte n'est pas réalisable pour le plan de sondage considéré.

Comme exemple pour illustrer la portée de la proximité des diagonales, les deux formes quadratiques de même diagonale $Q = \begin{pmatrix} a & c \\ c & a \end{pmatrix}$ et $\widetilde{Q} = \begin{pmatrix} a & -c \\ -c & a \end{pmatrix}$ sont positives si $a \geqslant 0$ et $c^2 \leqslant a^2$ (Annexe A) . Alors $\sup_{\|u\| \leqslant 1} \left|Q - \widetilde{Q}\right|(u) = 2|c|$ peut être arbitrairement grand. Néanmoins, l'écart reste limité relativement à $\sup_{\|u\| \leqslant 1} |Q|(u) = a + |c| \geqslant 2|c|$.

– Si seuls sont pris en compte les éléments diagonaux des deux formes quadratiques Q et Q^{HT}, l'écart ainsi tronqué entre les deux estimateurs pour une variable z vaut :

$$\sum_{i \in s} \delta_i(c) z_i^2 - \sum_{i \in s} (1 - \pi_i) z_i^2 = \sum_s a z^2 \qquad \textbf{(e3.29)}$$

(en posant $a = \delta(c) - (1 - \pi)$)

→ La pondération 'calée sur la trace' $c_2 = (1 - \pi)\left(\sum_s 1 - \pi\right) \Big/ \sum_s \delta(1 - \pi)$ présente trois avantages théoriques :

- Elle minimise l'écart précédent en valeur absolue sur les z tel que z^2 est constante.
- Elle vérifie la propriété attrayante que si $\pi_i = 1$ alors $\delta_i(c_2) = 0 = Q^{HT}_{i,i}$.
- $c_2 \geqslant 1 - \pi$ car $\delta(1 - \pi) \leqslant 1 - \pi$ donc le facteur de calage est $\geqslant 1$.

→ La pondération c_2 exploite une seule information sur la diagonale de l'estimateur sans biais. Comme pour le calage sur marge, exploiter l'information multidimensionnelle connue pourrait rendre plus précise l'estimation de variance.

– Pour minimiser en c le critère $\max_{\sum z^2 \leqslant 1} \left(\sum a z^2\right)^2$, l'optimum est :

$$\operatorname*{argmin}_c \max_s a^2 = \operatorname*{argmin}_c \max_s \left(\delta(c) - (1 - \pi)\right)^2$$

Cette option présente l'intérêt d'assurer la qualité minimale la meilleure possible (selon le critère 'tronqué' retenu), dans la classe des approximations retenues, pour toutes les variables d'intérêt. Mais elle n'a pas été testée ici, parce qu'elle n'est pas implémentable par les algorithmes programmés pour cette étude.

– Un dernier critère de proximité vise à minimiser la moyenne équipondérée de l'écart (**e3.29**) au carré sur toutes les variables dont les coordonnées sont comprises entre -1 et 1 :

$$\int_{-1}^{1} \left(\sum a z^2\right)^2 dz \frac{1}{2^{|s|}} = \frac{1}{2^{|s|}} \int_{-1}^{1} \left(\sum (a z^2)^2 + \sum_{i \neq j} a_i z_i^2 a_j z_j^2\right) dz = \sum a^2 \frac{1}{5} + \sum_{i \neq j} a_i a_j \frac{1}{3}\frac{1}{3}$$

$$= \frac{1}{45}\left(4 \sum a^2 + 5\left(\sum a\right)^2\right) \qquad \textbf{(e3.30)}$$

(L'intégrale ci-dessus est multiple, puisque $z \in [-1, 1]_s$.)

→ Cette troisième approche suggère le critère à minimiser $\sum \left(\delta(c) - (1 - \pi)\right)^2$, qui est le carré de la norme euclidienne de l'écart entre les deux diagonales [12]. Dans la suite, cette fonction objectif est notée $\tau(c) = \|\delta(c) - (1 - \pi)\|^2$ et $f(c) = \delta(c) - (1 - \pi)$. Avec ce choix, une pondération optimale est une solution du programme $\min_c \|\delta(c) - (1 - \pi)\|^2$.

---

11. Le rapprochement est discutable si la forme quadratique définie par l'estimateur sans biais n'est pas positive.

12. En fait, le critère a été retenu avant l'analyse ci-dessus. Le critère de proximité alternatif peut se formuler : $\alpha_0 \|\delta(c) - (1 - \pi)\|^2 + \beta_0 \left\{\sum \delta(c) - \sum (1 - \pi)\right\}^2 = \|A(\delta(c) - (1 - \pi))\|^2$ avec $A = \alpha\,\mathrm{id} + \beta 1_s 1_s'$.

→ La concavité de $\varphi(c)$ entraîne que pour tout $i \in s$, $1-\pi_i-\delta_i(c)$ est convexe en c. Comme $x \longmapsto x^2$ est convexe et croissante sur $\mathbb{R}^+$, pour tout $i \in s$, $\{\delta_i(c)-(1-\pi_i)\}^2$ est convexe en c sur un voisinage convexe de $1-\pi$ tel que $\delta(c) \leqslant 1-\pi$. Parce qu'une somme de fonctions convexes est convexe, la fonction objectif $\tau(c)$ est convexe en c sur un tel voisinage de $1-\pi$.

→ dérivée partielle de la fonction objectif (pour $c>0$) :

$$\frac{\partial\tau(c)}{\partial c_i} = \frac{\partial\left\|\delta(c)-(1-\pi)\right\|^2}{\partial c_i} = \sum_{\neq i} 2(\delta(c)-(1-\pi))\frac{\partial\delta(c)}{\partial c_i} + 2(\delta_i(c)-(1-\pi_i))\frac{\partial\delta_i(c)}{\partial c_i}$$

$$= \sum_{\neq i} 2(\delta(c)-(1-\pi))\left(u_i'\left(\sum cuu'\right)^- cu\right)^2 + 2(\delta_i(c)-(1-\pi_i))\left(1-u_i'\left(\sum cuu'\right)^- c_i u_i\right)^2 \quad \textbf{(e3.31)}$$

$$= \sum_{\neq i} 2(\delta(c)-(1-\pi))\left(u_i'\left(\sum cuu'\right)^- cu\right)^2 + 2(\delta_i(c)-(1-\pi_i))\left[\frac{\delta_i(c)}{c_i}\right]^2 \quad \textbf{(e3.32)}$$

$\Rightarrow$ $\dfrac{\partial\tau(1-\pi)}{\partial c_i} \leqslant 0$: C'est cohérent avec $\delta(1-\pi) \leqslant 1-\pi$.

– Si $c_i = 0, c_s \neq 0$ et $u_i \in \mathrm{Im}\left(u^{c>0}\right)$ alors $\dfrac{\partial\delta_i(c)}{\partial c_i} = 1$ (vu (**e2.25**) et $\widehat{\beta}_c(\mathbb{1}_i) = 0$ dans ce cas). Mais même ces conditions ne suffisent sans doute pas à assurer que $c_i > 0$ pour une solution minimisante, parce qu'il semble possible que $\sum(\delta(c)-(1-\pi))_{\neq i}\left[u_i'\widehat{\beta}_{\neq i}\right]^2 > 1-\pi_i$ pour $c_i = 0$ et $u_i \in \mathrm{Im}\left(u^{c>0}\right)$.

• La logique de construction de la pondération par rapprochement de l'estimateur sans biais n'est pas la seule envisageable. La pondération de Hajek, ou corrigée de la perte de degrés de liberté, est formulée ainsi :

$$c_h = (1-\pi)|s| / (|s| - \mathrm{rang}(x_s))$$

Le coefficient correcteur de la perte de degrés de liberté peut s'interpréter par un modèle tel que conditionnellement à $\widetilde{u}_s = \sqrt{\dfrac{\pi^*(1-\pi^*)}{\pi}}\dfrac{x}{\pi}(s)$, $\widetilde{z}_s = \sqrt{\dfrac{\pi^*(1-\pi^*)}{\pi}}\dfrac{y}{\pi}(s)$ est d'espérance $\widetilde{u}_s\beta^*$ et de variance diagonale $\sigma^2\,\mathrm{id}_s$. Dans ce cadre : $E\left\{\sum_s\left(\widetilde{z}_s-\widetilde{u}_s\beta^*\right)^2\middle|\widetilde{u}_s\right\} = \sigma^2|s|$. Alors avec l'estimateur : $\widehat{\beta}^s = \left\{\sum_s \dfrac{\pi^*(1-\pi^*)}{\pi}\dfrac{xx'}{\pi^2}\right\}^- \sum_s \dfrac{\pi^*(1-\pi^*)}{\pi}\dfrac{yx}{\pi^2}$,
et $\widehat{\sigma^2} = \sum_s\left(\widetilde{z}_s-\widetilde{u}_s\widehat{\beta}\right)^2 = \left\|\mathrm{proj}^{\perp}_{\mathrm{Im}(\widetilde{u}_s)}(\widetilde{z}_s)\right\|^2$, on a : $E\left\{\widehat{\sigma^2}\middle|\widetilde{u}_s\right\} = \sigma^2(|s|-\mathrm{rang}(x_s))$. Donc dans ce modèle $\sum_s \dfrac{\pi^*(1-\pi^*)}{\pi}\left(\dfrac{y}{\pi}-\dfrac{x}{\pi}'\beta^*\right)^2$ peut s'estimer par : $\dfrac{|s|}{|s|-\mathrm{rang}(x_s)}\sum_s \dfrac{\pi^*(1-\pi^*)}{\pi}\left(\dfrac{y}{\pi}-\dfrac{x}{\pi}'\widehat{\beta}\right)^2$.

3.2 bornes de la pondération c

– Si c est considéré comme une approximation de $\dfrac{\pi^*(1-\pi^*)}{\pi}$ alors il conviendrait d'imposer $0 < c \leqslant \dfrac{1}{4\pi}$ [13]. Pour les algorithmes à gradient, cette borne supérieure a permis de limiter les difficultés informatiques d'inversion de matrice [14]. Par ailleurs, ces contraintes de calcul numérique ont amené à annuler les poids c_i des unités telles que $\delta_i(1) \leqslant 10^{-6}$.

13. Ce majorant peut être multiplié par $|s|/(|s|-\mathrm{rang}(x_s))$ pour une correction pour perte de degrés de liberté.

14. La fonction *ginv* de SAS/IML échoue lorsque certains coefficients sont trop grands, par exemple pour $(10^{18}\quad 10^{18}//0\quad 10^{18})$. Cette erreur stoppe l'exécution du module IML en cours. L'utilisation des instructions IML *call push/resume/pause* permet de gérer cette erreur. Un calcul alternatif de l'inverse a été programmé pour contourner ce problème, ainsi que pour améliorer la qualité de l'inverse, pour une matrice symétrique. Mais il n'a pas été retenu car plus lent que *ginv* et surtout dans certains cas de valeurs propres plus instables d'une itération d'un calcul d'une même inverse à l'autre.

– L'équation (**e2.11**) fournit un encadrement de $\delta(c)$:

$$0 \leqslant \delta_i(c) = \min_{\beta} \|\mathbb{1}_i - u_s\beta\|_c^2 \leqslant \|\mathbb{1}_i\|_c^2 = c_i \text{ d'où :}$$

$$0 \leqslant \delta(c) \leqslant c \qquad \textbf{(e3.33)}$$

• $c_i = 0 \Rightarrow \delta_i(c) = 0$. Donc une solution exacte de l'équation de pondération vérifie nécessairement $c_i > 0$, pour i tel que $\pi_i < 1$.

• $\delta_i(c) = 1 - \pi_i \Rightarrow c_i \geqslant 1 - \pi_i$ donc si c est une solution exacte alors $c \geqslant 1 - \pi$. Ceci suggère de borner la minimisation à $\{c \geqslant 1 - \pi\}$.

– Le cas où $\delta_i(c) = c_i$ est exclu si $cu(i) \neq 0$:

$$\delta_i(c) = c_i \Leftrightarrow 0 \in \operatorname*{argmin}_{\beta} \|\mathbb{1}_i - u_s\beta\|_c^2 \Leftrightarrow \langle \mathbb{1}_i - u'0, u_s \rangle_c = 0 \Leftrightarrow c_i u_i = 0$$

$$\rightarrow \text{ si } c_i u_i \neq 0 \text{ alors } \delta_i(c) < c_i \qquad \textbf{(e3.34)}$$

→ Cette propriété implique que si $c > 0$ est une solution de $\delta(c) = 1 - \pi$ et si le vecteur de calage ne s'annule pas sur s alors $\forall i \in s \,/\, \pi_i < 1, c_i > 1 - \pi_i$. En effet, comme $\delta_i(c) \leqslant c_i$ une solution vérifie nécessairement $c_i \geqslant 1 - \pi_i$ et de plus $c_i = 1 - \pi_i \Rightarrow \delta_i(c) = c_i \Rightarrow c_i u_i = 0$.

→ La croissance de la fonction δ et son homogénéité de degré 1 implique l'encadrement :

$$\min(c)\delta(1) \leqslant \delta(c) \leqslant \max(c)\delta(1) \qquad \textbf{(e3.35)}$$

– Soit c une solution exacte, donc telle que $\delta(c) = 1 - \pi$. Si $\delta(1) > 0$, cet encadrement implique que :

$$\min(c) \leqslant \min\left[\frac{1-\pi}{\delta(1)}\right] \text{ et } \max(c) \geqslant \max\left[\frac{1-\pi}{\delta(1)}\right]$$

→ Cette dernière inégalité prouve que la pondération maximale d'une solution (**e3.26**) est d'autant plus grande que le minimum de $\delta(1)$ sur $\{\pi < 1\}$ est petit.

• Si $u_i \neq 0$ alors la croissance du i-ième coefficient diagonal en fonction du ième poids est bornée, à autres poids fixés et finis :

$$\lim_{c_i \longrightarrow +\infty} \delta_i(c) = \min_{u_i'\beta=1} \sum_{\neq i} c(u_s\beta)^2 = \min_{u_i'\beta=1} \|u_{\neq i}\beta\|_{c_{\neq i}}^2 \qquad \textbf{(e3.36)}$$

[15]

↳ $\delta_i(c)$ est fonction croissante de c_i, et ce coefficient est majoré par $\min_{u_i'\beta=1} \sum_{\neq i} c(u_s\beta)^2$.

↳ Pour tout $\eta > 0$, $\delta_i(c) \geqslant \eta \wedge \min_{|1-u_i'\beta| \leqslant \sqrt{\frac{\eta}{c_i}}} \sum_{\neq i} c(u_s\beta)^2$ (minimum des deux termes).

En effet, pour tout β, ou bien $(1 - u_i'\beta)^2 \geqslant \dfrac{\eta}{c_i}$ et alors $c_i(1 - u_i'\beta)^2 + \sum_{\neq i} c(u_s\beta)^2 \geqslant \eta$, ou bien $(1 - u_i'\beta)^2 < \dfrac{\eta}{c_i}$ et alors $\sum_{\neq i} c(u_s\beta)^2 \geqslant \min_{|1-u_i'\beta| \leqslant \sqrt{\frac{\eta}{c_i}}} \sum_{\neq i} c(u_s\beta)^2$. La suite suppose le choix $\eta < c_i$.

↳ $$\min_{|1-u_i'\beta| \leqslant \sqrt{\frac{\eta}{c_i}}} \sum_{\neq i} c(u_s\beta)^2 = \min_{\begin{cases} u_i'\beta = 1 \\ |1-\alpha| \leqslant \sqrt{\frac{\eta}{c_i}} \end{cases}} \sum_{\neq i} c(u_s\beta)^2 \alpha^2 = \min_{u_i'\beta=1} \sum_{\neq i} c(u_s\beta)^2 \left[1 - \sqrt{\frac{\eta}{c_i}}\right]^2$$

• pour $j \neq i$:

$$\lim_{c_i \longrightarrow +\infty} \delta_j(c) = c_j \left[1 - c_j u_j' \text{proj}_{u_i}^{\perp} \left[\sum_{\neq i} c \text{proj}_{u_i}^{\perp} u u' \text{proj}_{u_i}^{\perp}\right]^{+} \text{proj}_{u_i}^{\perp} u_j\right] \qquad \textbf{(e3.37)}$$

↳ On peut écrire $u_s = \lambda_s u_i' + v_s$ avec $v_s u_i = 0$ (donc $v_i = 0$ et $\lambda_i = 1$).

↳ Comme $\beta = \alpha u_i / \|u_i\|^2 + h$ avec $h \perp u_i$:

$$\delta_j(c) = \min_{\beta} \|\mathbb{1}_j - u_s\beta\|_c^2 = \min_{\alpha \in \mathbb{R}, h \perp u_i} \|\mathbb{1}_j - \alpha\lambda_s - v_s h\|_c^2$$

$$= \left\|\mathbb{1}_j - \widehat{\alpha}\lambda_s - v_s\widehat{h}\right\|_c^2 = \left\|\mathbb{1}_j - \widehat{\alpha}\lambda_s - v_s\widehat{h}\right\|_{c_{\neq i}}^2 + c_i\widehat{\alpha}^2$$

15. $\sum_{\neq i}$ signifie $\sum_{j \in s, j \neq i}$.

↳ Donc $c_i\widehat{\alpha}^2 \leqslant \delta_j(c) \leqslant c_j < +\infty$ et par suite $\lim\limits_{c_i \longrightarrow \infty} \widehat{\alpha} = 0$.

↳ Soit $\widetilde{h} \in \operatorname{argmin}_{h \perp u_i} \left\| \mathbb{1}_j - v_s h \right\|_{c_{\neq i}}$

↳ $\sqrt{\delta_j(c)} \leqslant \left\| \mathbb{1}_j - v_s \widetilde{h} \right\|_{c_{\neq i}}$ (vu en prenant $\beta = \widetilde{h}$)

↳ $\sqrt{\delta_j(c)} \geqslant \left\| \mathbb{1}_j - \lambda_s \widehat{\alpha} - v_s \widehat{h} \right\|_{c_{\neq i}} \geqslant \left\| \mathbb{1}_j - v_s \widehat{h} \right\|_{c_{\neq i}} - \left\| \lambda_s \widehat{\alpha} \right\|_{c_{\neq i}} \geqslant \left\| \mathbb{1}_j - v_s \widetilde{h} \right\|_{c_{\neq i}} - \left\| \lambda_s \right\|_{c_{\neq i}} |\widehat{\alpha}|$

↳ Comme le dernier terme ci-dessus tend vers 0, $\lim\limits_{c_i \longrightarrow \infty} \sqrt{\delta_j(c)} = \left\| \mathbb{1}_j - v_s \widetilde{h} \right\|_{c_{\neq i}}$ □

$\Rightarrow \lim\limits_{c_i \longrightarrow \infty} \tau = \left[\min\limits_{u_i'\beta = 1} \sum\limits_{\neq i} c \left(u_s \beta \right)^2 - (1 - \pi_i) \right]^2 + \sum\limits_{\neq i} \left[c \left(1 - cv' \left[\sum\limits_{\neq i} cvv' \right]^+ v \right) - (1 - \pi) \right]^2$: la fonction objectif tend vers une limite finie lorsqu'un seul coefficient de pondération tend vers l'infini.

Plus généralement :

lemme l3.1 *limite à l'infini de δ*

si $\delta(1) > 0$ [16]*, $1 - \pi_s > 0$, u_s ne s'annule pas, $c^{\mathbb{N}}$ est une séquence de pondérations convergente vers $c^\infty \in]0, +\infty]_s$ et $\delta_s(c^{\mathbb{N}})$ est bornée dans $\mathbb{R}_s$ alors, avec $s_\infty = \{ i \in s \,/\, c_i^\infty = +\infty \}$, $\Delta s = s \setminus s_\infty$ son complémentaire et $s_\infty^{\hat{i}} = s_\infty \setminus \{i\}$:*

▷ $\Delta s \neq \varnothing$ *(ie au moins un poids fini)*

▷ $\forall i \in s_\infty, \lim \delta_i(c^{\mathbb{N}}) = \min\limits_{\substack{u_i'\beta = 1 \\ u_{s_\infty^{\hat{i}}}\beta = 0}} \sum\limits_{\Delta s} c^\infty \left(u'\beta \right)^2 = \min\limits_{\substack{u_i'\beta = 1 \\ u_{s_\infty^{\hat{i}}}\beta = 0}} \left\| u_{\Delta s}\beta \right\|_{c_\infty}^2$ **(e3.38)**

▷ $\forall j \in \Delta s, \lim \delta_j(c^{\mathbb{N}}) = \min\limits_{\beta} \left\| \mathbb{1}_j - v_{\Delta s}\beta \right\|_{c_{\Delta s}^\infty}^2$ *, où* $v^{\Delta s} = \operatorname{proj}^{\perp}_{\operatorname{Im}(u^{s_\infty})}(u^{\Delta s})$ **(e3.39)**

↳ Si $\Delta s = \varnothing$ alors $\delta_s(1 - \pi) > 0$ implique que $\lim\limits_{\lambda \longrightarrow \infty} \lambda \delta(1 - \pi) = (+\infty)_s$, et donc $\lim \delta_s(c^{\mathbb{N}}) = (+\infty)_s$, ce qui contredit l'hypothèse que $\delta_s(c^{\mathbb{N}})$ est bornée.

↳ Pour $i \in s_\infty$ $\delta_i(c^n) \leqslant \min\limits_{\substack{u_i'\beta = 1 \\ u_{s_\infty^{\hat{i}}}\beta = 0}} \sum\limits_{\Delta s} c^n \left(u'\beta \right)^2$. Comme $c_{\Delta s}^{\mathbb{N}} \longrightarrow c_{\Delta s}^\infty$ et $0 < c_{\Delta s}^\infty < +\infty$ par hypothèse, pour tout $0 < \epsilon < 1$ et n assez grand $c_{\Delta s}^\infty(1 - \epsilon) \leqslant c_{\Delta s}^n \leqslant c_{\Delta s}^\infty(1 + \epsilon)$, d'où on déduit que : $\overline{\lim}\, \delta_i(c^{\mathbb{N}}) \leqslant \min\limits_{\substack{u_i'\beta = 1 \\ u_{s_\infty^{\hat{i}}}\beta = 0}} \sum\limits_{\Delta s} c^\infty \left(u'\beta \right)^2 (< +\infty)$.

↳ On peut décomposer : $\widehat{\beta}^n = \widehat{\beta}_{c^n}(\mathbb{1}_i) = u^{s_\infty}\widehat{\gamma}^n + \Delta\widehat{\beta}^n$ avec $\Delta\widehat{\beta}^n \perp \operatorname{Im}(u^{s_\infty})$.

$\overline{\lim}\, c_i^{\mathbb{N}}(1 - u_i'u^{s_\infty}\widehat{\gamma}^{\mathbb{N}})^2 + \sum\limits_{s_\infty^{\hat{i}}} c^{\mathbb{N}} \left(u'u^{s_\infty}\widehat{\gamma}^{\mathbb{N}} \right)^2 \leqslant \overline{\lim}\, \delta_i(c^{\mathbb{N}}) < +\infty \Rightarrow \lim u_{s_\infty} u^{s_\infty}\widehat{\gamma}^{\mathbb{N}} = \begin{pmatrix} 1 \\ 0_{s_\infty^{\hat{i}}} \end{pmatrix}$

↳ minoration :

$\delta_i(c^n) = \min\limits_{\Delta\beta \in \operatorname{Im}(u^{s_\infty})^\perp} \sum\limits_{s} c^n \left\{ \mathbb{1}_i - u'(u^{s_\infty}\widehat{\gamma}^n + \Delta\beta) \right\}^2 \geqslant \min\limits_{\Delta\beta \in \operatorname{Im}(u^{s_\infty})^\perp} \sum\limits_{\Delta s} c^n \left[u'(u^{s_\infty}\widehat{\gamma}^n + \Delta\beta) \right]^2$

Ce minorant est manifestement une fonction continue en $(c_{\Delta s}^n, u^{s_\infty}\widehat{\gamma}^n)$.

↳ $\lim u_{s_\infty} u^{s_\infty}\widehat{\gamma}^{\mathbb{N}} = \begin{pmatrix} 1 \\ 0_{s_\infty^{\hat{i}}} \end{pmatrix} \Rightarrow \begin{pmatrix} 1 \\ 0_{s_\infty^{\hat{i}}} \end{pmatrix} \in \operatorname{Im}(u_{s_\infty})$ (car ce sous-espace vectoriel est fermé),

$\Rightarrow \exists \gamma^\infty \,/\, u_{s_\infty} u^{s_\infty}\gamma^\infty = \begin{pmatrix} 1 \\ 0_{s_\infty^{\hat{i}}} \end{pmatrix}$ Comme u_{s_∞} est linéaire et injective sur $\operatorname{Im}(u^{s_\infty})$ (de dimension finie), $\lim u_{s_\infty} u^{s_\infty}\widehat{\gamma}^{\mathbb{N}} = u_{s_\infty} u^{s_\infty}\gamma^\infty \Rightarrow \lim u^{s_\infty}\widehat{\gamma}^{\mathbb{N}} = u^{s_\infty}\gamma^\infty$.

16. La démonstration fonctionne également pour $\widetilde{s} \subset s$ tel que $\delta_{s \setminus \widetilde{s}}(1) = 0$, $\delta_{\widetilde{s}}(1) > 0$ et $1 - \pi_{\widetilde{s}} > 0$.

↳ d'où $\underline{\lim}\,\delta_i(c^{\mathbb{N}}) \geqslant \min_{\Delta\beta\in \mathrm{Im}(u^{s\infty})^{\perp}} \sum_{\Delta s} c^{\infty}\left[u'(u^{s\infty}\gamma^{\infty}+\Delta\beta)\right]^2 \geqslant \min_{u_s\beta=\begin{pmatrix}1\\0_{s^{\infty}_{\infty}}\end{pmatrix}} \sum_{\Delta s} c^{\infty}(u'\beta)^2$

⇒ $\delta_i(c^{\mathbb{N}})$ est convergente vers cette limite.

↳ La propriété (**e3.39**) se démontre comme (**e3.37**), en écrivant $u_s = \Lambda u_{s_\infty} + v_s \,/\, v_s u^{s\infty} = 0$ et $\beta = \alpha + \gamma$ avec $\alpha \in \mathrm{Im}(u^{s\infty})$, $\gamma \in \mathrm{Im}(u^{s\infty})^{\perp}$. □

→ En général, ces éléments théoriques n'assurent pas que la minimisation ne fasse pas tendre des coefficients de la pondération vers l'infini (voir une condition suffisante pour la pondération 'récursive').

3.3 existence de solutions

– Le lemme *l3.2* exclut l'existence d'une solution exacte de (**e3.26**) si l'indicatrice d'une des unités échantillonnées est dans l'image des vecteurs d'équilibrage rapportés à la probabilité d'inclusion. Ceci correspond à une unité atypique pour le vecteur d'équilibrage.

lemme l3.2 *condition de constance de* δ_i

$$\begin{aligned}&\delta_i(1) = 0 \\ &\Leftrightarrow \mathbb{1}_i \in \mathrm{Im}(u_s) \Leftrightarrow \exists c>0 \,/\, \delta_i(c) = 0 \Leftrightarrow \forall c>0, \delta_i(c) = 0\end{aligned} \qquad \textbf{(e3.40)}$$

↳ Ce lemme découle immédiatement du fait que pour $c > 0$ $\delta_i(c)$ est le carré d'une distance entre $\mathbb{1}_i$ et $\mathrm{Im}(u_s)$, selon (**e2.11**).

– Le cas $\delta_i(1) = 0$ est exclu lorsque $u_s = 1_s$ et $|s| > 1$, vu qu'alors $\delta_i(1) = 1 - 1/\sum_s 1 = 1 - 1/|s| > 0$. Il n'apparait donc pas dans les contextes d'une seule contrainte d'équilibrage pour un échantillonnage de taille fixe.

– La condition $\exists i \in s \,/\, \mathbb{1}_i \in \mathrm{Im}(u_s)$ ne semble pas très forte, comme $\mathbb{1}_s$ est une base de $\mathbb{R}_s$ dont $\mathrm{Im}(u_s)$ est un sous-espace vectoriel. Elle peut être précisée ainsi :

$$\exists\beta \,/\, u_s\beta = \mathbb{1}_i \Leftrightarrow \begin{cases} u_{\neq i}\beta = 0 \\ u_i'\beta = 1 \end{cases} \Leftrightarrow \begin{cases} \beta' u^{\neq i} = 0 \\ u_i'\beta \neq 0 \end{cases} \Leftrightarrow \begin{cases} \beta \in \mathrm{Im}(u^{\neq i})^{\perp} \\ u_i'\beta \neq 0 \end{cases} \Leftrightarrow u_i \notin \mathrm{Im}(u^{\neq i}) \qquad \textbf{(e3.41)}$$

En effet $u_i \notin \mathrm{Im}(u^{\neq i}) \Leftrightarrow u_i \notin \mathrm{Im}(u^{\neq i})^{\perp\perp} \Leftrightarrow \exists\beta \in \mathrm{Im}(u^{\neq i})^{\perp} \,/\, u_i'\beta \neq 0$. Donc :

$$\mathbb{1}_i \in \mathrm{Im}(u_s) \Leftrightarrow u_i \notin \mathrm{Im}(u^{\neq i}) \qquad \textbf{(e3.42)}$$

Cette condition peut s'exprimer ainsi : x_i n'est pas une combinaison linéaire des vecteurs d'équilibrage du reste de l'échantillon. En effet : $\mathrm{Im}(u^{\neq i}) = \mathrm{Im}(x^{\neq i})\,\mathrm{diag}\left[\left(\frac{1}{\pi}\right)^{\neq i}\right] = \mathrm{Im}(x^{\neq i})$.

→ Il y a donc un lien via ces unités atypiques entre la difficulté d'optimiser la pondération et celle de l'équilibrage : si $\begin{cases} x_i \notin \mathrm{Im}(x^{\mathscr{P}\setminus i}) \\ \pi_i < 1 \end{cases}$ alors il n'est pas possible d'équilibrer exactement le plan de sondage, puisque pour un échantillon $s \not\ni i$, $\sum_s \frac{x}{\pi} \neq \sum_{\mathscr{P}} x$.

– Les unités concernées par (**e3.42**) n'interviennent pas dans les coefficients diagonaux des autres unités :

$$\mathbb{1}_i \in \mathrm{Im}(u_s) \Rightarrow \forall j \neq i, \left\langle \mathbb{1}_j - u_s\widehat{\beta}_j, \mathbb{1}_i \right\rangle = 0 \Rightarrow u_i'\widehat{\beta}_c(\mathbb{1}_{\neq i}) = 0 \text{ (si } c>0) \qquad \textbf{(e3.43)}$$

$$u_i \notin \mathrm{Im}(u^{\neq i}) \Rightarrow \forall j \neq i, u_j'\left(\sum cuu'\right)^{-} u_j = u_j'\left[\sum_{\neq i} cuu'\right]^{-} u_j \qquad \textbf{(e3.44)}$$

La deuxième propriété découle de ce que pour tout β, il existe γ tel que $u_{\neq i}\gamma = 0$ et $u_i'\gamma = -u_i'\beta$, vu (**e3.41**) [17]. Il s'ensuit que $\forall j \neq i, \forall\beta, \left\|\mathbb{1}_j - u_s\beta\right\|_c^2 \geqslant \left\|\mathbb{1}_j - u_s\beta - u_s\gamma\right\|_c^2 = \left\|\mathbb{1}_j - u_s\beta\right\|_{c_{\neq i}}^2$.

17. deux autres preuves : $\mathbb{1}_i \in \mathrm{Im}(u_s) \Rightarrow \left\|\mathrm{proj}^{c^{\perp}}_{\mathrm{Im}(u_s)}[\mathbb{1}_i]\right\|_c^2 = c_i(1-u_i'\widehat{\beta}_i)^2 + \sum_{\neq i} c\left[u'\widehat{\beta}_i\right]^2 = 0$ d'une part et $\Rightarrow \forall j \neq i, \left\langle \mathbb{1}_i, \mathrm{proj}^{c^{\perp}}_{\mathrm{Im}(u_s)}[\mathbb{1}_j]\right\rangle_c = 0$ d'autre part.

– La condition $\exists i \in s \,/\, \delta_i(1) = 0$ (**e3.40**) est vérifiée pour une strate du tirage de première phase de l'enquête Emploi de taille d'échantillon égale à la dimension d'équilibrage, cas de la Corse [18], ou lorsque des variables d'équilibrage sont très concentrées. C'est le cas par exemple pour l'indicatrice de commune rurale en Ile-de-France. Ces deux régions sont les seules où le nombre d'unités $\left|\left\{\delta(1) \leqslant 10^{-6}\right\}\right|$ dépasse 3 pour au moins une des 10 000 simulations. Pour la Corse, toutes les unités sont dans ce cas dans 80% des simulations.

Un troisième cas de coefficients diagonaux quasiment nuls est observé pour le plan de sondage étudié, pour des régions où deux unités $i \neq j$ représentent la totalité d'une variable d'équilibrage sur l'échantillon. En effet, l'équilibrage du complémentaire de l'échantillon implique qu'une sous-matrice de u_s est du type :

$$u_{i,j}^{k,l} = \begin{pmatrix} \dfrac{\alpha}{\pi_i} & \dfrac{\alpha}{1-\pi_i} \\ \dfrac{\beta}{\pi_j} & \dfrac{\beta}{1-\pi_j} \end{pmatrix} \text{ et } u_{\{i,j\}^c}^{k,l} = 0$$

qui est de déterminant non nul si $\pi_i \neq \pi_j$ et $\alpha\beta \neq 0$. Alors $\mathbb{1}_i$ et $\mathbb{1}_j \in \mathrm{Im}\left(u_s^{k,l}\right) \subset \mathrm{Im}\left(u_s\right)$.

– Les coordonnées i telles que $\delta_i(1) = 0$ ne sont pas prises en compte dans l'approximation de variance [19]. C'est donc un facteur de biais de celle-ci, si une unité concernée est de probabilité d'inclusion $0 < \pi_i < 1$.

– Un deuxième problème soulevé par la condition (**e3.40**) est d'inciter l'algorithme de minimisation à la hausse des pondérations d'autres unités, lorsque l'égalité à 0 est approchée. En effet pour $i \neq j$ la dérivée partielle croisée est positive (**e2.21**).

Ce phénomène de compensation est accentué lorsque la pondération est plafonnée. Il explique sans doute que les coefficients diagonaux ne soient pas tous inférieurs à 1 pour les minimisations testées [20]. Par contre, la pondération simple $c_1 = 1 - \pi$ vérifie cette propriété souhaitable, parce que $\delta(1-\pi) \leqslant 1 - \pi$ (vu (**e3.33**)).

– Comme vu précédemment, pour une unité i telle que $\delta_i(1) = 0$:

- $\forall c > 0, \begin{cases} u_i' \widehat{\beta}_c(\mathbb{1}_i) = 1 \\ u_{\neq i}' \widehat{\beta}_c(\mathbb{1}_i) = 0 \end{cases}$
- pour $j \neq i$, $u_i' \widehat{\beta}_c(\mathbb{1}_j) = 0$

$\Rightarrow$ (**e3.31**) entraine alors que $\forall c > 0, \dfrac{\partial \tau(c)}{\partial c_i} = 0$. Donc pour une telle unité, si l'égalité du coefficient à 0 est exacte, le poids peut rester constant dans une minoration.

– Les cas où $\delta_i(1)$ est très proche de 0 mais pas exactement nul induisent en pratique une erreur bloquante dans l'inversion de la matrice $\sum cuu'$ au cours de la minimisation, parce qu'ils mènent rapidement à des poids très élevés, au delà d'un million [21]. C'est pourquoi ils ont été systématiquement écartés des minimisations par une méthode de gradient, ce qui correspond à forcer leur poids c à 0. (Ces poids ont été mis à la valeur $1 - \pi$ pour effectuer l'estimation de variance, afin d'éviter que la PROC REG ne s'interrompt pour cause de poids nuls. Ceci ne doit pas avoir d'incidence sur l'estimation de variance.) Théoriquement, la mise à l'écart de

18. Pour la Corse, $|s| = \mathrm{rang}\left(x_s\right)$ dans la plupart des simulations de tirage. L'observation de la constance de $\delta(c)$ qui en découle, et donc de l'absence d'effet des algorithmes d'optimisation de la pondération, a mené à la découverte du lemme de constance du coefficient diagonal.

19. Pour ces cas, comme $\Delta \geqslant 0, \forall j \in s, |\Delta_{i,j}(c)| \leqslant \sqrt{\delta_i(c)\,\delta_j(c)} = 0$. Donc l'estimateur de variance ne comprend pas de terme en y_i. Par suite, il est impossible que l'estimateur soit sans biais $\forall y\ (\mathscr{P})$.

20. S'il n'y avait pas de compensation et si $\delta_i(c) > 1$ alors baisser c_i rapprocherait de $1 - \pi_i$, vu que $\dfrac{\partial \delta_i(c)}{\partial c_i} = \left(1 - u_i'\left(\sum cuu'\right)^- c_i u_i\right)^2 = \left[\delta_i(c) / c_i\right]^2 > 0$, sans modifier $\delta_{\neq i}(c)$, et il serait possible ainsi d'atteindre $1 - \pi_i$ puisque δ est continue et que pour $c_i = 0$, $\delta_i(c) = 0$ (vu (**e2.10**)).

$\widetilde{s} = \{i \in s \,/\, \delta_i(1) \cong 0\} = s_{\delta(1)=0}$ devrait être suffisante, parce que, pour $j \in s \setminus \widetilde{s}$:

$$\begin{cases} \forall i \in \widetilde{s}, u_i \notin \mathrm{Im}\left(u^{\neq i}\right) \\ u_j \in \mathrm{Im}\left(u^{\neq j}\right) \end{cases} \Rightarrow u_j \in \mathrm{Im}\left(u^{\in \widetilde{s}, \neq j}\right)$$

En effet, si $u_j = u_{\widetilde{s}}\lambda_{\widetilde{s}} + u^{s\setminus\widetilde{s},\neq j}\beta$ et $\exists i \in \widetilde{s} \,/\, \lambda_i \neq 0$ alors $u_i \in \mathrm{Im}\left(u^{\neq i}\right)$.

– Du fait sans doute de l'incertitude numérique, des cas (3/2000 en moyenne) sont observés où $\delta_i(1) > 10^{-6}$ mais $\delta_i(1-\pi) = 0$. Plus surprenant, il se trouve également que $\delta_i(1-\pi) > 10^{-6}$ mais $\delta_i(c) = 0$ pour c grand.

– Même lorsque $\delta(1)$ n'a pas de coordonnée nulle, l'existence de $c \geqslant 0$ tel que $\delta(c) = 1-\pi$ peut dépendre du vecteur $(1-\pi)_s$.

- L' Annexe E montre que :
  - ▷ Le rang de $\dot{\delta}(c)$ est constant et égal à $\dim\{\mathrm{Vect}[h^2, h \in \mathrm{Im}(x_s)^{\perp}]\}$.
  - ▷ Le rang de $\dot{\delta}(c)$ est $\geqslant |s| - \mathrm{rang}(x_s)$.
  - ▷ Pour que $\dot{\delta}(c)$ soit inversible, il suffit qu'il existe une sous-matrice de la matrice d'équilibrage x_s de $\mathrm{rang}(x_s)$ colonnes, composée d'une première sous-matrice carrée de dimension et de rang égaux à $\mathrm{rang}(x_s)$ et la deuxième sous-matrice telle que la suppression d'une ligne quelconque donne une matrice de rang égal à $\mathrm{rang}(x_s)$. Cette condition suffisante peut s'interpréter approximativement comme celle d'une taille d'échantillon suffisamment grande par rapport au nombre de contraintes d'équilibrage.

– Vu l'étude de la dérivée 'à droite' de $\tau(c) = \|\delta(c) - (1-\pi)\|$, il ne semble pas évident qu'une solution minimisante donne toujours des poids non nuls aux unités i telles que $\delta_i(1) > 0$.

→ Pour une solution minimisante telle que $c > 0$, la dérivée de la fonction objectif est nécessairement nulle. Le i-ème coefficient de cette dérivée peut s'exprimer ainsi :

$$2\sum_j [\delta_j(c) - (1-\pi_j)]\left\{\mathbb{1}_i^j - u_i'\left(\sum cuu'\right)^- c_j u_j\right\}^2 \qquad \textbf{(e3.45)}$$

→ Cet expression suggère que si le minimum n'est pas une solution exacte et si $\delta(1) > 0$ alors le signe de $\delta(c) - (1-\pi)$ n'est pas constant sur les coordonnées du minimum.

– Plus rigoureusement, si $\delta(c) \geqslant 1-\pi$ et s'il existe i tel que $\delta_i(c) > 1-\pi_i$ alors : $\dfrac{\partial\tau}{\partial c_i} \geqslant 2(\delta_i(c)-(1-\pi_i))\left(1 - u_i'\left(\sum cuu'\right)^- c_i u_i\right)^2 = 2(\delta_i(c)-(1-\pi_i))\left[\dfrac{\delta_i(c)}{c_i}\right]^2 > 0$. Il est alors possible de diminuer strictement la fonction objectif en réduisant c_i.

– Au final, il n'a pas été possible de déterminer des conditions suffisantes d'existence d'une solution $0 < c < +\infty$ qui minimise $\|\delta(c) - (1-\pi)\|$. La difficulté est de prouver qu'il est possible de limiter la minimisation à un compact, autrement dit que la procédure de minimisation ne fait pas tendre des coordonnées de la pondération vers l'infini ou vers zéro [22].

4 comment optimiser la pondération des résidus d'équilibrage

4.1 algorithmes de minimisation de $\|\delta(c) - (1-\pi)\|$

4.1.1 un algorithme récursif

– Une procédure de minimisation de mise en œuvre simple recherche un point fixe de la fonction $\xi(c) = c - \delta(c) + 1 - \pi$ [23]. Son calcul ne nécessite donc pas la dérivée de δ. Cette

21. La probabilité d'inclusion du sondage étudié est comprise entre 1.1% et 4.2%, avec une médiane de 1.6%. S'il existe $i \in s \,/\, \delta_i(1) = 10^{-6}$ et $\pi = 1\%$ alors le poids maximal d'une solution exacte dépasse 990 000.

22. Plus précisément, il suffirait que le rapport du poids maximal sur le poids minimal soit borné (si $(\overline{c} \,/\, \underline{c}) \leqslant (1/\epsilon) < +\infty : \lambda(c) = \langle \delta(c), 1-\pi\rangle / \|\delta(c)\|^2$ est définie pour $c \in [\epsilon, 1]_s$, si $\delta(1) > 0$, et $\|\delta[\lambda(c)\, c] - (1-\pi)\|$ est une fonction continue sur ce compact.)

méthode est identique à celle décrite par [AMYT] page 553, à l'initialisation et au traitement en cas de non convergence près.

→ $\forall c \geqslant 0, \xi(c) \geqslant 1-\pi$, vu que $\delta(c) \leqslant c$. Cette propriété est satisfaisante, parce qu'elle est vérifiée par une solution exacte (éventuelle), et utile pour étudier $\xi^{\mathbb{N}}$.

→ Si $\pi_i = 1$ alors l'initialisation de c à $1-\pi$ donne $\left[\xi^{\mathbb{N}}(1-\pi)\right]_i = 0$.

→ si $\delta_i(1) = 0$ alors $\xi_i(c) = c_i + 1 - \pi_i$ donc $\xi_i^n(1-\pi) = (n+1)(1-\pi_i)$

▹ Il peut être préférable d'écarter ces unités de la procédure récursive, si leur pondération croissante à l'infini pose un problème numérique.

– $\xi_i(c) > c_i \Leftrightarrow \delta_i(c) < 1-\pi_i$ donc à chaque itération, la procédure :

▹ augmente (strictement) le poids des $i \,/\, \delta_i(c) < 1-\pi_i$

▹ diminue le poids des $i \,/\, \delta_i(c) > 1-\pi_i$

▹ s'arrête si la nouvelle pondération ne fait pas strictement baisser la fonction objectif

lemme l4.3 *condition suffisante de majoration des poids récursifs*

(pour une infinité d'itérations)

$$\textit{si } u_i \neq 0 \textit{ et } \min_{u_i'\beta=1} \left\|u_{\neq i}\beta\right\|^2_{1-\pi_{\neq i}} > 1-\pi_i \textit{ alors } \begin{cases} \overline{\lim}\, \xi_i^{\mathbb{N}}(1-\pi) < +\infty \\ \overline{\lim}\, \delta_i^{\mathbb{N}}(1-\pi) < +\infty \end{cases} \qquad \textbf{(e4.46)}$$

Posons $\delta_i^{\mathbb{N}} = \delta_i\left[\xi^{\mathbb{N}}(1-\pi)\right]$ et $\xi_i^{\mathbb{N}} = \xi_i^{\mathbb{N}}(1-\pi)$.

↳Supposons que : $\overline{\lim}\, \xi_i^{\mathbb{N}} = +\infty$ **(e4.47)**

– soit $\sigma_1(\mathbb{N}) \subset \mathbb{N} \,/\, \lim \xi_i^{\sigma_1(\mathbb{N})} = +\infty$

→ $\sigma_2(n) = \sup\left\{k \leqslant \sigma_1(n) \,/\, \delta_i^k \leqslant 1-\pi_i\right\}$ est bien définie car $\delta_i^0 = \delta_i(1-\pi) \leqslant 1-\pi_i$.

→ si $\sigma_2(n) < \sigma_1(n)$ alors ξ_i^k décroit pour k allant de $\sigma_2(n)+1$ à $\sigma_1(n)$, parce que $\delta_i^k > 1-\pi_i \Rightarrow \xi_i^{k+1} < \xi_i^k$, donc $\xi_i^{\sigma_2(n)+1} \geqslant \xi_i^{\sigma_1(n)}$ et $\xi_i^{\sigma_2(n)+1} = \xi_i^{\sigma_2(n)} + 1 - \pi_i - \delta_i\left[\xi^{\sigma_2(n)}\right] \leqslant \xi_i^{\sigma_2(n)} + 1 - \pi_i$

→ donc dans tous les cas $\xi_i^{\sigma_2(n)} \geqslant \xi_i^{\sigma_1(n)} - (1-\pi_i)$ et donc $\lim \xi_i^{\sigma_2(\mathbb{N})} = +\infty$

→ $\delta_i\left[\xi^{\sigma_2(\mathbb{N})}\right] \geqslant \min_{\beta} \left\|\mathbb{1}_i - u_s\beta\right\|^2_{\xi_i^{\sigma_2(n)},1-\pi_{\neq i}}$ donc $\underline{\lim}\, \delta_i\left[\xi^{\sigma_2(\mathbb{N})}\right] \geqslant \min_{u_i'\beta=1} \left\|u_{\neq i}\beta\right\|^2_{1-\pi_{\neq i}}$, en appliquant (**e3.36**). Par la condition (**e4.46**), ce minorant est strictement supérieur à $1-\pi_i$, alors que $\delta_i\left[\xi^{\sigma_2(\mathbb{N})}\right] \leqslant 1-\pi_i$, d'où une contradiction. Donc $\overline{\lim}\, \xi_i^{\mathbb{N}} = +\infty$ est impossible sous la condition (**e4.46**).

↳ $\overline{\lim}\, \delta_i\left(\xi^{\mathbb{N}}\right) \leqslant \overline{\lim}\, \xi_i^{\mathbb{N}} < +\infty$ □

→ La condition suffisante de majoration des poids (**e4.46**) peut être explicitée ainsi :

$$\forall i \in s, u_i \neq 0 \text{ et } (1-\pi_i) u_i' \left\{u^{\neq i} * (1-\pi)_{\neq i} u_{\neq i}\right\}^- u_i < 1 \qquad \textbf{(e4.48)}$$

Cette formulation suggère l'interprétation qu'aucune unité ne doit être prépondérante ou trop singulière selon le vecteur de calage.

→ La condition (**e4.46**) constitue une sorte de généralisation de celle en haut de la page 580 de [DT], qui correspond au cas où $u_s = 1_s$.

→ Pour une unité i qui vérifie (**e4.46**), $\exists! \overline{c}_i \,/\, \delta_i\left[\overline{c}_i, 1-\pi_{\neq i}\right] = 1-\pi_i$ et $\xi_i^{\mathbb{N}} \leqslant \overline{c}_i + 1 - \pi_i$

↳ La fonction $f(c_i) = \delta(c_i, 1-\pi_{\neq i})$ est continue sur $\mathbb{R}^{+*}$ et strictement croissante (parce que $\delta_i(1) > 0$), $f(1-\pi_i) < 1-\pi_i$ (si $(1-\pi_i)u_i \neq 0$) et $\lim_{+\infty} f > 1-\pi_i$ par hypothèse.

↳ La propriété $\xi_i^n \leqslant \overline{c}_i + 1 - \pi_i$ est vraie pour $n = 0$. Supposons la vraie pour un $n \in \mathbb{N}$. Ou bien $\xi_i^n > \overline{c}_i$ et alors $\delta_i(\xi^n) \geqslant \delta_i\left[\overline{c}_i, 1-\pi_{\neq i}\right] = 1-\pi_i$ donc $\xi_i^{n+1} = \xi_i^n - \delta_i(\xi^n) + 1 - \pi_i \leqslant \xi_i^n \leqslant \overline{c}_i + 1 - \pi_i$. Ou bien $\xi_i^n \leqslant \overline{c}_i$ et alors $\xi_i^{n+1} = \xi_i^n - \delta_i(\xi^n) + 1 - \pi_i \leqslant \xi_i^n + 1 - \pi_i \leqslant \overline{c}_i + 1 - \pi_i$.

◃ Cependant, la condition (**e4.46**) s'avère trop forte pour le sondage étudié. En effet, sur le premier échantillon simulé, elle n'est vérifiée que pour l'Ile-de-France, même après avoir écarté les unités de $\delta_i(1) \leqslant 10^{-6}$. Même lorsque les variables d'équilibrage du complémentaire sont omises, la condition de majoration des poids récursifs n'est vérifiée que pour 8 régions.

23. Alternativement, si l'objectif est de minimiser $\|A[\delta(c) - (1-\pi)]\|$, avec par exemple $A = \alpha\,\mathrm{id} + \beta 1_s 1_s'$, la fonction contractante de récursion pourrait être du type : $\xi(c) = c - \lambda A(\delta(c) - (1-\pi))$, avec $\lambda > 0$.

• Si $\xi^{\mathbb{N}}$ converge vers $c^{\infty} \in [1-\pi, +\infty[_s$ alors c^{∞} est une solution exacte.

↳ Comme ξ est continue sur $\mathbb{R}_s^+$, $\xi(\lim \xi^{\mathbb{N}}) = \lim \xi^{\mathbb{N}}$ donc $\xi(c^{\infty}) = c^{\infty}$ ce qui équivaut à $\delta(c^{\infty}) = 1-\pi$.

– Une condition de baisse de la fonction objectif par récurrence sur ξ doit assurer :

$$\|\delta[\xi(c)] - (1-\pi)\| = \|\delta[c - \delta(c) + 1 - \pi] - (1-\pi)\| \leqslant \|\delta(c) - (1-\pi)\| \qquad \textbf{(e4.49)}$$

L'intérêt de cette propriété est que la fonction objectif baisserait en suivant la trajectoire $\xi^{\mathbb{N}}(c)$, tant qu'elle est vérifiée.

– Or :

▷ $c - \xi(c) = \delta(c) - (1-\pi)$

▷ $\delta[\xi(c)] - (1-\pi) = \delta[\xi(c)] - \xi(c) + \xi(c)(1-\pi) = (\delta - \mathrm{id})(\xi(c)) - (\delta - \mathrm{id})(c)$

Donc la condition de baisse par récurrence sur ξ peut s'écrire comme une contraction de $\mathrm{id} - \delta$:

$$\begin{aligned} &\|\delta[\xi(c)] - (1-\pi)\| \leqslant \|\delta(c) - (1-\pi)\| \\ \Leftrightarrow\ &\|(\mathrm{id}-\delta)(\xi(c)) - (\mathrm{id}-\delta)(c)\| \leqslant \|\xi(c) - c\| \end{aligned} \qquad \textbf{(e4.50)}$$

– Une formulation alternative de la condition de baisse :

$$\begin{aligned} &\|\delta[\xi(c)] - (1-\pi)\|^2 \leqslant \|\delta(c) - (1-\pi)\|^2 \\ \Leftrightarrow\ &\|\delta[\xi(c)] - \delta(c)\|^2 \leqslant -2\langle \delta[\xi(c)] - \delta(c), \delta(c) - (1-\pi)\rangle \end{aligned} \qquad \textbf{(e4.51)}$$

→ Il y a baisse de la fonction objectif entre c et $\xi(c)$ si le raisonnement suivant est valide : si • $\|\delta(c) - (1-\pi)\|$ est assez petit pour approximer $\delta[\xi(c)] - \delta(c)$ par $\dot{\delta}(c)[1 - \pi - \delta(c)]$

• Les termes non diagonaux de cette dérivée sont négligeables pour $1 - \pi_s - \delta_s(c)$ alors :

$$\begin{aligned} &\|\delta[\xi(c)] - \delta(c)\|^2 + 2\langle \delta[\xi(c)] - \delta(c), \delta(c) - (1-\pi)\rangle \\ &\cong \sum_{i\in s}\left\{\frac{\partial \delta_i(c)}{\partial c_i}(1 - \pi_i - \delta_i(c))\right\}^2 - 2\sum_{i\in s}\frac{\partial \delta_i(c)}{\partial c_i}[1 - \pi_i - \delta_i(c)]^2 \leqslant 0 \\ &\text{car } \frac{\partial \delta_i(c)}{\partial c_i} = \left[\frac{\delta_i(c)}{c_i}\right]^2 \leqslant 1 \end{aligned}$$

– La condition de baisse (**e4.50**) est vérifiée si $\left\|id - \dot{\delta}\right\| \leqslant 1$ [24] sur $[c, c - \delta(c) + 1 - \pi]$, c'est-à-dire que cette dérivée soit contractante sur cet intervalle :

$$\|(\mathrm{id}-\delta)(\xi(c)) - (\mathrm{id}-\delta)(c)\| \leqslant \sup_{\widetilde{c}\in[c,\xi(c)]} \left\|(\mathrm{id}-\dot{\delta})(\widetilde{c})(\xi(c) - c)\right\| \qquad \textbf{(e4.52)}$$

$$\leqslant \sup_{[c,\xi(c)]} \left\|\mathrm{id}-\dot{\delta}\right\| \|\xi(c) - c\| \qquad \textbf{(e4.53)}$$

• L' Annexe E montre que $\mathrm{id} - \dot{\delta}$ est contractante en c constante, mais pas généralement.

– Néanmoins, l'algorithme récursif possède les propriétés avantageuses suivantes :

▷ baisse de la fonction objectif par rapport à la pondération $1 - \pi$ (Annexe F)

▷ La pondération produite est minorée par $1 - \pi$, et donc ne s'annule pas.

▷ Sous la condition (**e4.46**) l'algorithme ne fait pas tendre de poids vers l'infini.

– Cependant :

◁ Il n'est pas démontré que l'algorithme réduise la fonction objectif indéfiniment.

◁ L'optimalité de la pondération produite n'est pas assurée, même lorsqu'il y a une solution exacte [25].

◁ La condition suffisante de majoration des poids s'avère trop forte pour le sondage étudié.

→ Au final, la justification de cet algorithme repose essentiellement sur son efficacité constatée empiriquement à diminuer la fonction objectif, dans le contexte d'un plan de sondage où il ne semble pas exister de solution exacte à l'équation de pondération.

---

24. La norme d'une application linéaire est définie en Annexe E.

25. Une tentative pour appliquer un raisonnement analogue à celui de [CDL] p.468 n'a pas abouti.

La simplicité de l'algorithme récursif, sans calcul de la dérivée, a pour contrepartie notamment une croissance lente des poids maximaux. Par suite, cette version bute le plus souvent sur le nombre maximal d'itérations. Une première solution à cet inconvénient est d'utiliser une fonction de récursion multiplicative. L'autre solution est le recours à un algorithme 'à gradient'.

4.1.2 algorithme récursif multiplicatif

Pour accélérer ou au moins prolonger la croissance de la pondération des unités de petits $\delta_i(1)$, la fonction de récursion 'multiplicative' $\widetilde{\xi}(c) = (1-\pi)/(\delta(c)/c)$ est envisageable [26]. Effectivement, elle s'avère fournir un algorithme quatre fois plus rapide que la fonction 'additive' pour une pondération plus minimisante pour la majorité des régions.

Cependant, son paramétrage est délicat. Si un incrément maximal de la pondération par itération n'est pas imposé ou trop élevé, l'algorithme s'arrête très rapidement par interruption de la baisse de la fonction objectif. En ce cas, le performance de la minimisation est très variable selon la région, avec des baisses de la distance amples par rapport à l'algorithme précédent (>10%), mais également des augmentations (modérées). Ces résultats laissent supposer qu'il reste une marge d'amélioration sensible de la minimisation.

D'autre part, la version multiplicative de l'algorithme récursif accroit les coefficients diagonaux excessifs ($\delta(1) > 1$) si l'incrément toléré est supérieur à 10. Une interprétation est que les variations entre itérations sont trop fortes pour un ajustement fin. L'ajout d'une séquence additive ne change pas sensiblement le résultat. Si le blocage sur un point final problématique est dû au fait que l'itération suivant n'envisage qu'une agrégation de modifications dont le bilan est défavorable, une autre piste serait de n'effectuer qu'une partie des variations de poids $\xi(c) - c$. Mais les résultats obtenus dans cette direction n'ont pas été probants.

L'arrêt de l'algorithme récursif est déclenché par l'interruption de la baisse de la fonction objectif plutôt que par le nombre maximal d'itérations. De plus il n'est pas en général causé par l'impossibilité de calculer la fonction. La dérivée de la fonction objectif, calculée pour cette pondération sur le premier échantillon simulé, n'est quasiment nulle que pour l'Ile-de-France. C'est également le cas pour la version multiplicative. Ce résultat suggère que l'algorithme récursif n'aboutit pas à une solution minimisante.

4.1.3 algorithmes utilisant le gradient

Trois types de méthode de minimisation utilisant le gradient, donc uniquement une dérivée première, ont été testés. A la k-ième étape du processus un algorithme de minimisation fournit un incrément $\widetilde{\delta}$ tel que $\left\| f\left(c_k + \widetilde{\delta}\right)\right\| < \left\| f\left(c_k\right)\right\|$.

- descente du gradient (GDA) : Comme $\left\| f\left(c + \widetilde{\delta}\right)\right\|^2 = \left\| f(c)\right\|^2 + 2\left\langle f(c), \dot{f}(c)\,\widetilde{\delta}\right\rangle + o\left(\widetilde{\delta}\right)$, à la condition que $\dot{f}(c)'f(c)$ ne soit pas nul, il est assuré qu'il existe $\lambda > 0$ tel que : $\left\| f\left\{c - \lambda \dot{f}(c)'f(c)\right\}\right\|^2 < \left\| f(c)\right\|^2$, et de plus ce vecteur donne la direction de baisse maximale en c, au sens où $\dfrac{\dot{f}(c)'f(c)}{\left\|\dot{f}(c)'f(c)\right\|} \in \underset{\|u\|=1}{\operatorname{argmax}} \dfrac{\partial \left\| f(c+\lambda u)\right\|^2}{\partial \lambda}(\lambda = 0)$.

Au voisinage d'une solution, $\dot{f}(c)'f(c)$ se rapproche de 0. Il s'avère en pratique que les baisses de la fonction objectif obtenues par l'algorithme GDA sont fortes pour les premières itérations, mais s'affaiblissent rapidement avant l'arrivée à une solution aussi bonne que d'autres algorithmes de temps d'exécution comparable.

---

26. Cette approche correspond à la transformation logarithmique effectuée dans [CDL] avant la récursion.

• Gauss-Newton (GNA) : Si f était deux fois dérivable en c alors le développement limité de la fonction objectif serait :

$$\left\|f\left(c+\widetilde{\delta}\right)\right\|^2 = \|f(c)\|^2 + 2\left\langle f(c), \dot{f}(c)\widetilde{\delta}\right\rangle + \left\langle \dot{f}(c)\widetilde{\delta}, \dot{f}(c)\widetilde{\delta}\right\rangle + \left\langle f(c), \ddot{f}(c)\widetilde{\delta}^2\right\rangle + o\left(\left\|\widetilde{\delta}\right\|^2\right)$$

La méthode GNA choisit l'incrément qui minimise la fonction convexe suivante :

$$\widetilde{\delta}^* \in \underset{\widetilde{\delta}}{\operatorname{argmin}}\, 2\left\langle f(c), \dot{f}(c)\widetilde{\delta}\right\rangle + \left\langle \dot{f}(c)\widetilde{\delta}, \dot{f}(c)\widetilde{\delta}\right\rangle$$

$$\Rightarrow \dot{f}(c)'\dot{f}(c)\widetilde{\delta}^* = -\dot{f}(c)'f(c) \Rightarrow \widetilde{\delta}^* = -\left\{\dot{f}(c)'\dot{f}(c)\right\}^{-}\dot{f}(c)'f(c)$$

C'est bien une direction de baisse au sens large, parce que :

$$\dot{f}(c)\widetilde{\delta}^* = -\mathrm{proj}_{\mathrm{Im}[\dot{f}(c)]}[f(c)] \Rightarrow \left\langle f(c), \dot{f}(c)\widetilde{\delta}\right\rangle = -\left\|\mathrm{proj}_{\mathrm{Im}[\dot{f}(c)]}[f(c)]\right\|^2 \leqslant 0$$

Bien que cet incrément soit encore nul lorsque le gradient $\dot{f}(c)'f(c)$ s'annule, la méthode tend à gonfler l'incrément dans certaines directions où la dérivée est petite, ce qui peut accélérer la baisse de la fonction objectif. Un peu plus précisément, en posant $A = \dot{f}(c)$, soit v une base orthonormée de vecteurs propres de AA', donc une base de $\mathrm{Im}(A)$ (**eA.57**). Pour tout x, $\|A'x\|^2 = \sum \lambda(v)\langle x, v\rangle^2$ alors que $\left\|\mathrm{proj}_{\mathrm{Im}(A)}(x)\right\|^2 = \sum \langle x, v\rangle^2$. Donc si $\mathrm{proj}_{\mathrm{Im}(A)}(x)$ est dans une somme d'espaces propres de $AA' = \dot{f}(c)'\dot{f}(c)$ de valeurs propres inférieures à 1 alors $\left\|\mathrm{proj}_{\mathrm{Im}(A)}(x)\right\| \geqslant \|A'x\|$. Or $f(c)$ est susceptible d'être dans ce cas si $\|A'f(c)\|$ est petit par rapport à $\|f(c)\|$, donc là où la performance de GDA est faible. A noter que si $A'f(c)$ n'est pas nul alors $x = \mathrm{proj}_{\mathrm{Im}(A)}[f(c)]$ non plus, parce que $\mathrm{Ker}\left\{\mathrm{proj}_{\mathrm{Im}(A)}\right\} = \mathrm{Im}(A)^{\perp} = \mathrm{Ker}(A')$ (**eA.56**).

• Levenberg-Marquardt (LMA) : Ce troisième algorithme vise à combiner l'avantage de la quasi-assurance d'une baisse de la norme par la méthode du gradient avec la plus grande rapidité de la méthode de Newton au voisinage d'une solution. La formule de l'incrément diffère de celle de l'algorithme GNA par le produit d'un facteur d'atténuation (variable) α avec la diagonale de la matrice à inverser, dans l'équation déterminant l'incrément $\widetilde{\delta}$:

$$\left\{\dot{f}(c)'\dot{f}(c) + \alpha\,\mathrm{diag}\left\{\dot{f}(c)'\dot{f}(c)\right\}\right\}\widetilde{\delta} = -\dot{f}(c)'f(c)$$

→ Parmi les algorithmes à gradient étudiés, il a semblé plus efficace de combiner une première étape de minimisation par descente du gradient d'au plus 10 itérations et tant que le taux de décroissance de la fonction objectif excède 0.1%, avec une deuxième étape de minimisation par GNA. En particulier, ce type de programme a paru plus rapide qu'un algorithme de Levenberg et Marquardt, dans la version testée pour celui-ci.

4.1.4 nécessité d'imposer des contraintes à la minimisation

Sans contraintes, les algorithmes à gradient sont rapidement bloqués par un problème numérique d'inversion de matrice. De ce fait, il s'est avéré nécessaire d'imposer un plafonnement à la pondération. Une alternative intéressante est de contrôler que $\delta < 1$. L'intérêt est que cette contrainte s'introduit naturellement compte tenu de l'objectif de rapprochement de $1 - \pi_s$, donc moins arbitraire. Mais sans algorithme qui intègre efficacement ce type de contrainte, l'étude ne l'a testée (par la version n_3 ci-après) que par rejet des incréments qui ne satisfont pas la contrainte. Ceci mène rapidement à l'arrêt de l'algorithme. La non-optimalité de cette méthode est indiquée par une dérivée ($\left\|\dot{f}(c)'f(c)\right\|$) plus éloignée de 0 que les autres méthodes à gradient.

4.1.5 changement de variable pour plafonner la pondération

Pour mettre en œuvre une contrainte de plafonnement de la pondération, un changement de variable est apparu nettement préférable à une restriction de l'incrément.

– Si la contrainte est $c \in]\underline{c}, \overline{c}[$, avec la restriction que $-\infty < \underline{c} < \overline{c} < +\infty$, une fonction bijective (et C^1) de $\mathbb{R}$ vers $]\underline{c}, \overline{c}[$ est : $\varphi(t) = \underline{c} + (\overline{c} - \underline{c})\dfrac{1}{1+e^{-2t}}$ [27]. La dérivée du changement de variable se calcule ainsi :

$$\dot{\varphi}(t) = (\overline{c} - \underline{c})\left[-\frac{1}{(1+e^{-2t})^2}\right](-2e^{-2t}) = (\overline{c} - \underline{c})\left[\frac{1}{1+e^{-2t}} - \frac{1}{(1+e^{-2t})^2}\right]2$$

$$= 2(c - \underline{c}) - 2\frac{(c - \underline{c})^2}{\overline{c} - \underline{c}}$$

L'inverse du changement de variable vaut : $\varphi^{-1}(c) = -\dfrac{1}{2} log\left(\dfrac{\overline{c} - c}{c - \underline{c}}\right)$.

note 1: Pour $\underline{c} = 0, \overline{c} = \dfrac{1}{4\pi}$ et π prenant les valeurs 0.01, 0.02, 0.03 et 0.04, le taux d'écart entre $\varphi^{-1}\varphi\varphi^{-1}\varphi(t)$ et t est inférieur à 0.05% pour t allant de -50 à 16. Au delà de cette borne, l'imprécision numérique devient plus sensible. Ces résultats suggèrent que le changement de variable ne pose pas de problèmes informatiques, à condition d'éviter les extrémités.

– sans borne supérieure pour la pondération c :

$$\varphi(t) = \underline{c} + e^t \qquad \dot{\varphi}(t) = e^t = c - \underline{c} \qquad \varphi^{-1}(c) = log(c - \underline{c})$$

• Dans la suite, les résultats chiffrés sont issus de $10\,000$ simulations du tirage de première phase de l'enquête Emploi. Ce sondage de grappes de logements est stratifié par région et équilibré sur 32 variables, dont la moitié pour équilibrer l'échantillon complémentaire.

27. Elle peut se construire à partir de la tangente hyperbolique, vu que $\dfrac{1}{1+e^{-2t}} = \dfrac{\text{th}(t)+1}{2}$. Cette fonction présente l'intérêt d'être bijective de $\mathbb{R}$ dans $]0,1[$.

4.1.6 Comparaison de l'efficacité de la minimisation selon l'algorithme

La combinaison de premières étapes GDA avec des étapes finales par GNA parait clairement préférable à l'algorithme de pure descente du gradient pour l'objectif opérationnel à minimiser (Tableau 1), tout en étant plus rapide :

- dérivée plus proche de 0
- sensibilité nettement moindre à l'initialisation (à une région près)
- minimisation plus poussée (distance et dérivée plus petites, pour toutes les régions)

Tableau 1 – Comparaison des performances pour la minimisation de deux algorithmes 'à gradient', sur la première simulation d'échantillonnage, par région

	région	distance finale			norme de la dérivée		taux impact initialis.	
		GDA	GDA +GNA	taux écart	GDA	GDA +GNA	GDA	GDA +GNA
11	Île-de-France	0,984	0,984	0,0	3,02E-04	1,54E-08	0,000	0,000
21	Champagne-Ardenne	1,877	1,870	-0,4	1,65E-02	2,12E-05	-0,189	0,000
22	Picardie	1,217	1,199	-1,5	2,19E-02	8,92E-07	-0,146	0,000
23	Haute-Normandie	1,782	1,770	-0,6	3,03E-02	4,82E-05	-0,183	0,000
24	Centre	1,580	1,568	-0,8	3,29E-02	4,75E-06	-0,096	0,000
25	Basse-Normandie	1,531	1,518	-0,9	3,53E-02	6,37E-06	-0,073	0,000
26	Bourgogne	2,158	2,151	-0,3	2,57E-02	1,19E-06	-0,146	-0,412
31	Nord-Pas-de-Calais	1,021	1,013	-0,8	1,43E-02	4,72E-07	-0,124	0,000
41	Lorraine	1,682	1,669	-0,8	4,18E-02	4,25E-06	-0,125	0,000
42	Alsace	1,703	1,689	-0,8	2,24E-02	2,73E-06	-0,090	0,000
43	Franche-Comté	1,858	1,841	-0,9	2,88E-02	1,46E-05	-0,204	0,000
52	Pays de la Loire	1,178	1,169	-0,7	2,64E-02	2,14E-07	-0,110	0,000
53	Bretagne	0,886	0,876	-1,2	1,21E-02	1,21E-05	-0,111	0,000
54	Poitou-Charentes	2,025	2,014	-0,5	4,04E-02	8,49E-06	-0,085	0,000
72	Aquitaine	1,798	1,787	-0,6	1,90E-02	1,18E-05	-0,081	0,000
73	Midi-Pyrénées	1,444	1,440	-0,3	1,62E-02	3,04E-06	-0,016	0,000
74	Limousin	2,156	2,151	-0,3	1,26E-02	7,74E-06	-0,090	0,000
82	Rhône-Alpes	0,877	0,872	-0,6	1,02E-02	4,21E-08	-0,185	0,000
83	Auvergne	2,303	2,296	-0,3	2,69E-02	4,15E-06	-0,113	0,000
91	Languedoc-Roussillon	0,986	0,971	-1,5	2,93E-02	1,40E-06	-0,128	0,000
93	Provence-Alpes-Côte d'Azur	0,431	0,414	-3,9	1,04E-02	3,84E-06	-0,195	0,000

Note :

– *Un maximum de 200 itérations est programmé pour l'algorithme de descente du gradient et 100 pour la combinaison d'algorithmes GDA+GNA. La durée d'exécution est de 1'45" pour le premier programme contre moins de 1'05" pour le second. La lenteur relative des algorithmes à gradient testés ici pourrait être affectée par la partie du programme qui effectue une recherche itérative du premier incrément qui fait baisser la fonction objectif. Une autre possibilité est que le plafonnement de la pondération imposée à ces méthodes induit un mouvement vers une frontière du domaine contraint.*

– *Les deux premières colonnes fournissent la valeur finale de la fonction objectif (*$\|\delta(c) - (1-\pi)\|$*).*

– *La norme de la dérivée est* $\left\|\dot{f}(c)' f(c)\right\|$*. Elle est calculée sans les coordonnées telles que* $\delta_i(c) \leqslant 10^{-6}$*, qui sont exclues des itérations de la minimisation. Mais celles-ci sont prises en compte dans la distance finale.*

– *Le taux d'impact de l'initialisation est calculé entre les distances finales obtenues avec les initialisations respectivement à* $1-\pi$ *et* $\min\left[\frac{1-\pi}{\delta(1)}, \frac{1}{4\pi}\right]$.

L'algorithme récursif a été appliqué avec une contrainte de décroissance minimale de la fonction objectif (0.001%) et une limite de 500 itérations. Ainsi paramétré, cet algorithme n'a pas nécessité de plafonner directement la pondération [28] ou d'annuler des poids pour éviter les problèmes numériques. C'est avantageux par rapport aux autres méthodes de minimisation testées. Il bénéficie également de la simplicité de sa programmation (pas de calcul du gradient $\dot{\delta}(c)$). La condition de poursuite de l'algorithme tant que la fonction objectif baisse est nécessaire aussi pour les autres algorithmes. Elle présente l'avantage de ne pas être arbitraire, contrairement à l'imposition de bornes à c ou $\delta(c)$. Cette technique s'avère empiriquement très efficace pour optimiser la pondération, au sens où elle mène plus rapidement que les autres algorithmes testés à des baisses plus amples de $\|\delta(c) - (1-\pi)\|$. Sur les 10 000 simulations, les algorithmes récursifs minimisent davantage la fonction objectif que les algorithmes à gradient testés pour toutes les régions (Tableau 2). En contrepartie, la pondération maximale est beaucoup plus élevée.

Tableau 2 – Médiane calculée sur les 10 000 simulations, pour la fonction objectif de la minimisation et pour le poids maximal, par région

	distance $\|\delta(c) - (1-\pi)\|$							pondération $\max(c)$								
région	v1	v2	n1	n2	n3	ren	rem	v1	v2	n1	n2	n3	ren	ren *	rem	$\frac{1-\pi^*}{\delta(1)}$
11	1,89	1,78	1,14	1,16	1,26	1,13	1,11	1,0	1,0	12,5	5,0	2,7	225	21	71	2,1
21	3,81	3,05	1,91	2,00	3,02	1,83	1,76	1,0	1,7	13,4	5,5	1,9	434	433	65 182	23,9
22	4,17	3,21	2,07	2,21	3,38	2,00	1,87	1,0	2,2	20,3	5,9	2,4	392	392	32 047	20,1
23	3,91	3,06	1,82	1,94	3,13	1,77	1,68	1,0	1,8	19,4	5,5	2,0	414	414	47 052	15,4
24	3,50	2,66	1,40	1,50	2,57	1,37	1,32	1,0	1,5	19,0	5,2	1,6	400	400	71 487	10,3
25	4,20	3,38	2,32	2,44	3,55	2,21	2,08	1,0	2,2	18,7	6,0	2,5	423	405	43 611	39,7
26	4,11	3,22	2,09	2,22	3,37	2,02	1,91	1,0	2,1	19,3	5,8	2,3	414	413	40 157	27,6
31	2,99	2,41	1,14	1,22	2,01	1,12	1,11	1,0	1,2	17,3	5,1	1,5	374	374	95 871	7,3
41	3,62	2,73	1,49	1,60	2,70	1,45	1,38	1,0	1,6	18,8	5,3	1,7	399	399	64 378	10,7
42	3,76	2,98	1,75	1,86	2,97	1,71	1,64	1,0	1,6	18,3	5,4	1,8	438	435	64 090	20,2
43	4,20	3,39	2,22	2,34	3,55	2,12	2,01	1,0	2,2	15,5	6,0	2,5	412	411	42 035	27,3
52	3,04	2,34	1,20	1,27	2,01	1,18	1,17	1,0	1,3	18,1	5,1	1,5	399	398	89 271	9,8
53	3,18	2,38	1,11	1,20	2,14	1,09	1,06	1,0	1,3	18,3	5,1	1,5	359	359	87 802	6,7
54	4,01	3,21	2,12	2,24	3,31	2,05	1,92	1,0	2,0	20,4	5,7	2,1	475	373	46 809	13,8
72	3,18	2,48	1,31	1,38	2,20	1,29	1,27	1,0	1,3	18,4	5,1	1,5	430	416	97 736	14,1
73	3,40	2,67	1,63	1,70	2,51	1,60	1,51	1,0	1,4	18,8	5,2	1,6	459	400	81 885	15,1
74	3,93	3,10	2,12	2,18	3,17	2,00	1,92	1,0	1,9	11,1	5,7	2,2	486	401	72 921	15,0
82	2,44	1,93	0,59	0,66	1,29	0,56	0,56	1,0	1,1	14,9	5,0	1,5	238	238	78 327	3,5
83	4,38	3,73	2,71	2,85	3,94	2,54	2,36	1,0	2,7	22,9	6,5	3,0	298	274	22 037	34,4
91	3,32	2,58	1,42	1,50	2,40	1,40	1,36	1,0	1,4	18,2	5,2	1,5	409	409	79 751	11,6
93	2,54	2,12	0,89	0,95	1,59	0,87	0,87	1,0	1,1	17,1	5,0	1,4	331	331	101 251	5,2

Note :

- *Les codes des méthodes sont explicités en note de bas du Tableau 9, sauf* `rem` *: récursif de fonction de récursion multiplicative. Les unités i telles que $\delta_i(c) \leqslant 10^{-6}$ sont prises en compte dans le calcul de distance.*
- *La durée d'exécution moyenne du programme récursif est de 37" pour la version additive et de 8" pour la multiplicative (comparé à 1' pour GDA+GNA).*
- *L'ajout d'un contrôle sur la qualité de l'inversion de matrice n'a pas d'effet sensible sur cette minimisation.*
- ** : hors unités de $\delta_i(1) \leqslant 10^{-6}$*

28. De fait, le nombre maximal d'itérations $\overline{n}$ borne la pondération récursive à $\overline{n}(1-\pi)$. Avec 10 000 itérations sans conditions d'arrêt, la pondération maximale dépasse 7 000.

• La méthode récursive donne la distance minimale dans la grande majorité des simulations croisées avec les régions (Tableau 3), et jamais la pire.

Tableau 3 – Rang des méthodes selon la distance obtenue, par simulation et région, hors Corse

	1	2	3	4	5	6	7
v1					50	347	209 708
v2				5	88 095	122 005	
n1	2 422	989	204 020	2 645	29		
n2	85	101	2 686	207 191	42		
n3			37	30	121 888	87 753	397
ren	25 581	181 718	2 593	213			
rem	182 017	27 297	769	21	1		

La baisse de la distance $\|\delta(c) - (1-\pi)\|$ entre les pondérations v_2 et ren n'est pas seulement observée en médiane des simulations, mais sur toutes les simulations. En fait, elle est d'au moins 15% pour toutes les simulations et toutes les régions (hors Corse).

Avec un autre critère de qualité de la minimisation, l'algorithme récursif parait très largement préférable selon le cumul des coefficients diagonaux supérieurs à 1 (Tableau 4).

Tableau 4 – Cumul des coefficients diagonaux supérieurs à 1 sur 10 000 simulations

	v2	n1	n2	n3	ren	rem
$\sum(\delta-1)^+$	1 856 162	630 236	720 287	1 631 884	41 338	69 444
$\max(\delta)$	3,05	2,22	2,50	3,12	1,70	2,85

C'est remarquable comme la pondération ren semble plus élevée que celle de v_2 (vu le Tableau 2), et que δ est croissante. En fait, si la pondération maximale de la méthode récursive est très supérieure à celle de v_2 et la pondération moyenne également (+18), pour la majorité des unités la pondération 'récursive' est plus basse : $c_{ren} < c_2 - 10^{-6}$ pour 16 millions de poids individuels sur les 24 millions simulés. Alors que le calage sur la trace augmente uniformément la pondération de départ, l'optimisation fournit une pondération plus modulée.

• Sur 500 itérations de l'algorithme récursif sur le premier échantillon simulé, sans condition d'arrêt, $\delta_i^{\mathbb{N}}$ dépasse $1-\pi_i$ d'au plus 6%, quelque soit la région.

• De plus, la distance $\|\delta(\xi^{\mathbb{N}}) - (1-\pi)\|$ baisse d'une itération à l'autre pour tous les individus et presque toutes les 500 itérations.

• L'évolution des poids d'une itération à l'autre n'apparait pas erratique (Annexe F).

• Plus précisément, sur les données étudiées, l'algorithme récursif parvient en au plus 500 itérations à rapprocher $\delta_i(c)$ de $1-\pi_i$ à un taux d'écart de moins de 0.4% pour les trois quarts des unités de l'ensemble des simulations (Tableau 5). La majorité des unités ont un coefficient ν supérieur à 0.2. Pour cette tranche, 95% des unités ont un taux d'écart de moins de 2.1%. Le fait que cette catégorie comporte également des unités de $\delta_i(c)$ quasiment nul, contrairement à leur coefficient ν, s'explique sans doute par une incertitude numérique.

Tableau 5 – Distribution sur 10 000 simulations du taux d'écart entre $\delta_i(c)$ et $1-\pi_i$ selon le coefficient $\nu_i = \min_{u_i'\beta=1} \|u^{\neq i}\beta\|^2_{1-\pi_{\neq i}} - (1-\pi_i)$

ν	n	q1	median	q3	p90	p95	max
<-0.2	3 164 930	0,7	25,1	60,6	86,0	100,0	100
0.2-<0	724 824	0,0	0,0	0,0	2,2	5,3	100
0-<0.2	677 321	0,0	0,0	0,0	2,0	4,9	100
>=0.2	19 047 551	0,0	0,0	0,0	0,9	2,1	100
ensemble	23 614 626	0,0	0,0	0,4	6,0	42,7	100

⇒ Sur les données étudiées, l'algorithme récursif apparait nettement préférable aux autres pondérations testées pour l'objectif de minimiser la distance choisie entre les deux diagonales, du moins si l'incertitude numérique sur cet indicateur est ignorée (voir ci-dessous). Il semble capable d'égaliser pratiquement les deux diagonales pour les unités dont le coefficient ν n'est pas trop petit, en un nombre limité d'itérations.

4.2 les coefficients diagonaux contrôlent les résidus

Les poids très élevés paraissent susceptibles d'accroitre la sensibilité de l'estimation de variance à des points extrêmes, dans le sens d'une surestimation de la variance [29]. Mais le contrôle des coefficients diagonaux permettent de borner l'impact de la valeur individuelle d'une variable sur l'estimateur de variance. Donc les unités de poids très élevé mais de faible coefficient diagonal doivent peser peu dans la variance estimée.

C'est effectivement démontrable, sous certaines conditions :

$$\widehat{\epsilon} = \left(\mathrm{id} - \mathrm{proj}^{c}_{\mathrm{Im}(u_s)}\right) z \Rightarrow c\widehat{\epsilon} = c\left(\mathrm{id} - \mathrm{proj}^{c}_{\mathrm{Im}(u_s)}\right) z = \Delta\left(c\right) z$$

$$\text{donc} : |c_i \widehat{\epsilon}_i| = \left| \sum_j \Delta_{i,j}\left(c\right) z_j \right| \qquad \textbf{(e4.54)}$$

$$\leqslant \sum_j |z\left(j\right)| \sqrt{\delta_i\left(c\right) \delta_j\left(c\right)} \text{ (car } \Delta \geqslant 0 \text{ (eA.58))}$$

$$\text{soit} : |c_i \widehat{\epsilon}_i| \leqslant \sqrt{\delta_i\left(c\right)} \sum_s |z| \sqrt{\delta\left(c\right)} \qquad \textbf{(e4.55)}$$

Cette majoration du résidu en fonction des coefficients diagonaux implique que même lorsque le poids d'une unité c_i est très élevé, sa contribution à l'estimation de variance $(c_i \widehat{\epsilon}_i^2)$ est faible si $\delta_i\left(c\right)$ est petit et que le deuxième terme de ce majorant n'est pas trop grand.

4.3 calage final

Pour les pondérations produites par minimisation, la non-convergence des coefficients diagonaux pour les unités de faible coefficient ν peut être contrecarrée par un calage final sur la trace de l'estimateur sans biais, c'est-à-dire $\sum_s 1 - \pi$.

Pour la version retenue de l'algorithme récursif, le critère de proximité des diagonales a été privilégié. C'est pourquoi ce calage final n'a été effectué dans la version récursive `ren` que s'il baissait la fonction objectif. Ce choix est discutable : pour cette procédure, l'arrêt après un nombre fini d'itérations a sans doute principalement pour effet de sous-pondérer certaines unités (selon leur coefficient diagonal). En moyenne sur les simulations, l'écart (négatif) entre la somme des coefficients diagonaux et celle de $1 - \pi$ dépasse 10% pour 7 régions et 20% pour une région (Auvergne).

⇒ Un calage final de la pondération d'équilibrage peut être prudent pour prémunir d'une sous-estimation de la variance.

---

29. L'idée est qu'augmenter le poids relatif d'une unité atypique dans une moyenne pondérée de carrés de résidus accroit celle-ci, bien que l'unité tire davantage l'estimation du coefficient β :

$$\frac{\partial \dfrac{M\left[\omega\left(y - x'\widehat{\beta}\right)^2\right]}{M\left(\omega\right)}}{\partial \omega_i} = \frac{1}{M\left(\omega\right)} \left\{ \left(y_i - x_i'\widehat{\beta}\right)^2 - \frac{M\left(\omega\left(y - x'\widehat{\beta}\right)^2\right)}{M\left(\omega\right)} \right\}$$

4.4 effet du nombre de variables d'équilibrage sur la minimisation

– La réduction de la norme par l'algorithme de minimisation est sensible même lorsque les variables d'équilibrage du complémentaire de l'EEC-TH sont prises en compte dans la fonction objectif (Tableau 6), mis à part la Corse. Cependant il n'y a plus de région pour lesquelles une solution quasiment exacte est trouvée, alors que sans les variables complémentaires la norme est presque annulée pour 7 régions.

Tableau 6 – Réduction de l'objectif $\|f(c)\|$ par rapport à $\|f(1-\pi_1)\|$ selon la prise en compte des variables d'équilibrage de l'échantillon complémentaire de l'EEC-TH

	sans les var.complémentaires			avec les var.complémentaires					$\min[\delta(1)]$	
région	distance initiale	distance finale	taux di-min %	distance initiale	distance finale	taux di-min %	$\|s\|$	$rang(x_s)$	sans	avec
11	1,1670	0,9844	−15,6	1,8645	0,9846	−47,2	535	25	0,0E+00	0,0E+00
21	1,9304	0,0226	−98,8	3,8652	1,8703	−51,6	63	28	0,41	0,03
22	2,1368	0,2749	−87,1	4,0142	1,1915	−70,3	51	28	0,34	0,10
23	1,9943	0,0912	−95,4	3,9615	1,7997	−54,6	58	28	0,43	0,04
24	1,7369	0,0000	−100,0	3,5548	1,5702	−55,8	76	28	0,57	0,14
25	2,0147	0,1293	−93,6	4,0035	1,5197	−62,0	53	28	0,40	0,01
26	2,0110	0,0843	−95,8	4,0316	2,1936	−45,6	53	28	0,41	0,07
31	1,3730	0,0000	−100,0	3,0112	1,0151	−66,3	138	28	0,70	0,14
41	1,7816	0,0000	−100,0	3,6465	1,6839	−53,8	73	28	0,48	0,10
42	1,8961	0,0467	−97,5	3,8392	1,7591	−54,2	65	28	0,45	0,07
43	2,1276	0,0195	−99,1	4,2186	1,8638	−55,8	48	28	0,42	0,07
52	1,4864	0,2660	−82,1	3,1284	1,1710	−62,6	117	28	0,40	0,10
53	1,5527	0,0000	−100,0	3,2522	0,8944	−72,5	101	28	0,56	0,16
54	2,0397	0,7308	−64,2	3,9911	2,0343	−49,0	58	28	0,18	4,4E-11
72	1,5825	0,4929	−68,9	3,3243	1,7952	−46,0	111	28	0,30	0,03
73	1,6794	0,4019	−76,1	3,3315	1,4453	−56,6	90	28	0,33	0,01
74	2,0240	0,7901	−61,0	3,8823	2,1557	−44,5	59	28	0,11	0,0E+00
82	1,1507	0,0000	−100,0	2,5801	0,8654	−66,5	207	28	0,68	0,12
83	2,3102	0,6565	−71,6	4,4218	2,3521	−46,8	42	28	0,22	9,7E-10
91	1,5786	0,0000	−100,0	3,3111	0,9824	−70,3	97	28	0,63	0,15
93	1,1515	0,0000	−100,0	2,6027	0,3912	−85,0	226	28	0,65	0,31
94	3,0975	3,0975	0,0	3,0975	3,0975	0,0	10	10	0,0E+00	0,0E+00

Note :

– *La distance est la norme de la fonction à annuler* $\|f(c)\| = \sqrt{\sum_{i\in s}\left[\delta_c(i) - (1-\pi_1(i))\right]^2}$, *pour la première simulation du tirage de l'échantillon de première phase de l'EEC-TH.*

– $\operatorname{rang}\left(\sum uu'\right) = \operatorname{rang}\left(\sum_s \frac{xx'}{\pi^2}\right) = \operatorname{rang}\left[\left(\frac{x}{\pi}\right)_s\right] = \operatorname{rang}(x_s)$, *où* x *est le vecteur d'équilibrage. Il a été calculé via la fonction* `homogen` *de SAS/IML, qui fournit une base du noyau (*$\operatorname{Ker}\left(\sum uu'\right)$*). Le nombre total de variables d'équilibrages est de 32.*

– *Les deux dernières colonnes fournissent le minimum régional du coefficient diagonal pour la pondération par 1. (Les cinq coefficients très proches de 0 ne sont pas stables d'une itération du calcul par* `ginv` *à l'autre, tout en restant quasiment nuls.)*

– *L'algorithme utilisé pour ce tableau est récursif. Les unités de* $\delta_i(1)$ *très petit n'ont pas été exclues.*

4.5 incertitude numérique

– Le Tableau 2 suggère une relation entre la distance minimisée et le niveau maximal de la pondération, avec un ordre similaire $ren/n_1/n_2/n_3/v_2$. C'est particulièrement clair pour les méthodes n_1 et n_2 : l'abaissement du plafond de la pondération de $\frac{1}{4\pi}$ à 5 toutes choses égales par ailleurs induit une croissance allant de 2% à 12% de la distance obtenue en médiane. Or la qualité de l'inversion de la matrice $\sum cuu'$ semble décroître en fonction de $\max(c)$, ou sans doute plutôt de la dispersion du poids d'équilibrage. Par exemple, pour la Picardie, la pondération maximale augmente en fonction de l'itération de l'algorithme GDA+GNA et la qualité (absolue) de l'inverse tend à se dégrader corrélativement (Tableau 7).

– Il a été envisagé d'introduire des contraintes qui stoppent l'algorithme lorsque la qualité de l'inverse tombe en dessus d'un seuil. Les tests effectués n'ont pas mené à retenir cette option pour les méthodes à gradient, au profit d'un plafonnement de la pondération , dont le choix est peut-être moins arbitraire.

→ Il reste cependant que l'incertitude numérique induite par les poids élevés relativise les performances de minimisation et affecte la comparaison des différentes méthodes. En outre, la relation entre l'importance de la diminution de la distance entre les diagonales et la qualité de l'estimation de variance est inconnue. Il est donc indispensable de compléter la comparaison des pondérations d'équilibrage sur des simulations de l'estimation de variance.

Tableau 7 – Dégradation de la qualité de l'inverse de $A = \sum cuu'$ en fonction de l'itération de l'algorithme de minimisation, pour la Picardie (reg=22), sur la première simulation

itération	$\|f(c)\|$	$\|\dot{f}(c)' f(c)\|$	$\max(c)$	$\|A - AA^{-}A\|$	taux diminution distance
	0,4601		10,1	1,7	
1	0,4010	0,3094	10,1	6,2	-12,8
2	0,3794	0,1743	10,1	4,0	-5,4
3	0,3775	0,1408	10,1	12,6	-0,5
4	0,3701	0,1404	10,1	7,6	-1,9
5	0,3572	0,1499	10,1	11,0	-3,5
6	0,3537	0,0454	10,1	3,4	-1,0
7	0,3507	0,0457	10,1	12,1	-0,8
8	0,3481	0,0469	10,1	6,0	-0,8
9	0,3460	0,0497	10,1	10,3	-0,6
10	0,3439	0,0529	10,2	7,1	-0,6
11	0,3395	0,0581	15,3	11,3	-1,3
12	0,3315	0,0574	17,0	137,1	-2,3
13	0,3312	0,0947	17,1	20,6	-0,1
14	0,3258	0,0956	17,1	80,1	-1,6
15	0,3218	0,1090	17,1	28,5	-1,2
16	0,3215	0,1162	17,1	53,6	-0,1
17	0,3148	0,1162	17,1	37,8	-2,1
18	0,3148	0,1308	17,1	332,4	0,0
19	0,2952	0,1308	17,1	114,7	-6,2
20	0,2817	0,1293	17,1	918,4	-4,6
21	0,2778	0,1139	17,1	445,9	-1,4
22	0,2458	0,1077	17,1	369,5	-11,5
23	0,2442	0,0212	17,1	419,6	-0,7
24	0,2441	0,0020	17,1	368,4	0,0
25	0,2441	0,0009	17,1	602,4	0,0
26	0,2441	0,0009	17,1	384,1	0,0
27	0,2441	0,0000	17,1	450,5	0,0

5 enjeu de la pondération des résidus d'équilibrage pour la variance de première phase de l'enquête emploi

Le résultat le plus sensible sur la qualité simulée de l'estimation de la variance de première phase de l'enquête Emploi est que les pondérations des résidus les plus simples, par $1-\pi$ ou par équipondération sous-estiment plus fréquemment et plus largement la variance (Tableau 9). En effet, pour trois des quatre variables testées, ces méthodes sous-estiment l'écart quadratique moyen (calculé sur les 10 000 simulations) dans plus de la moitié des simulations. La sous-estimation par la pondération $1-\pi$ (version v_1) est conforme aux attentes théoriques, vu que $\delta(1-\pi) < 1-\pi$, si $(1-\pi)x \neq 0$ [30]. Il semble donc prudent de choisir une pondération plus élaborée.

L'équipondération avec calage sur la trace de l'ESB, v_a, est indistinguable de la pondération $1-\pi$ calée pareillement (v_2). Cette faible sensibilité de l'approximation de variance $\left(\left\|z-u_s\widehat{\beta}_c(z)\right\|_c^2\right)$ aux petites fluctuations de la pondération peut être rapprochée de la formule de sa dérivée en fonction de la pondération : $\left\|z-u_s\widehat{\beta}_{c+\Delta c}(z)\right\|_{c+\Delta c}^2 \cong \left\|z-u_s\widehat{\beta}_c(z)\right\|_c^2 + \left\|z-u_s\widehat{\beta}_c(z)\right\|_{\Delta c}^2$ et $\left|\left\|z-u_s\widehat{\beta}_c(z)\right\|_{\Delta c}^2\right| \ll \left\|z-u_s\widehat{\beta}_c(z)\right\|_c^2$ si $|\Delta c| \ll c$.

Le classement des méthodes selon la qualité de l'estimation de variance dépend de la variable testée. Autrement dit, aucune des méthodes testées ne permet de réduire sensiblement les sous-estimations médianes sans accroitre les surestimations médianes.

– Les deux dernières variables, dont l'estimation de variance est moins précise, ont un écart quadratique moyen qui se stabilise plus lentement (Tableau 8). L'ampleur de la surestimation de la dernière variance parait reliée à la baisse sensible de l'écart quadratique moyen mesuré respectivement sur 2 000 et 10 000 simulations.

Tableau 8 – Stabilisation de la racine carrée de l'écart quadratique moyen (reqm)

	stabilisation reqm simulation	reqm 10000/1000	reqm 10000/2000
foy rev	3 691	0,8	0,9
nbpi2	1 315	-1,7	0,0
mar	7 348	2,9	0,3
isf	5 820	-3,8	-3,0

Note : La première colonne donne le nombre de simulations nécessaires pour stabiliser le reqm à moins de 1% de sa limite calculée sur les 10 000 *simulations.*

– L'estimation de variance de la première variable est très largement sous- ou sur-estimée dans un faible nombre de simulations, pour toutes les pondérations testées. L'ampleur de l'écart entre les indicateurs des écarts-types et des variances estimées signifie que le coefficient de variation entre les simulations de l'estimation de l'écart-type est élevé (vu que $Var_{sim}(\widehat{\sigma}) = \frac{\sum\widehat{\sigma}^2}{|sim|} - \left(\frac{\sum\widehat{\sigma}}{|sim|}\right)^2$). Cette variabilité de l'estimateur de la première variance est également observée sur l'intervalle inter-quartile, nettement plus ample que pour les autres variables. Enfin c'est la seule variable pour laquelle l'écart entre la moyenne et la médiane des estimations de variance est notable.

30. Un autre argument, sans référence à l'estimateur sans biais, est que si $\pi^* = \pi$ alors l'estimateur $\sum_s(1-\pi)\left(\frac{y}{\pi} - \frac{x}{\pi}\widehat{\beta}\right)^2$ est toujours inférieur à l'estimateur 'théorique' $\sum_s(1-\pi)\left(\frac{y}{\pi} - \frac{x}{\pi}\beta^*\right)^2$, donc susceptible de sous-estimer systématiquement la variance.

→ notes de lecture du Tableau 9 :

– *Le taux d'erreur est mesuré par rapport à l'erreur moyenne (rmse) calculée sur les* 10 000 *simulations du tirage de première phase* $\left(\sigma_+^2\left(\widehat{Y}\right) = \sum_{sim}\left(\widehat{Y}-Y\right)^2 \Big/ |sim|\right)$.

– *La colonne 'moy' donne la moyenne des taux d'erreurs de l'écart-type estimé :* $\sum_{sim} \frac{\widehat{\sigma}\left(\widehat{Y}\right)-\sigma_+\left(\widehat{Y}\right)}{\sigma_+\left(\widehat{Y}\right)} \Big/ |sim|$*, égal au taux d'erreur de la moyenne des estimations d'écart-type.*

– *Dans la colonne 'moy var' figure la racine carrée de la moyenne des erreurs sur la variance estimée* $\sqrt{\sum_{sim}\widehat{Var}\left(\widehat{Y}\right) \Big/ |sim| \Big/ \sigma_+^2\left(\widehat{Y}\right)} - 1$*, en %.*

– *La dernière colonne fournit le rang selon la médiane du taux d'erreur.*

– *Pour toutes les méthodes, le poids d'une unité de* $c \leqslant 0$ *est forcé à* $c = 1-\pi$.

– $v_1 : c = 1-\pi$

– $v_0 : c = 1$

– $v_2 : c = (1-\pi)\sum_s (1-\pi) \Big/ \sum_s \delta\,(1-\pi)$

– $v_a : c = \sum_s (1-\pi) \Big/ \sum_s \delta\,(1)$ *(équipondération avec calage sur la trace de l'ESB)*

– $v_b : c = \sum_s \delta\,(1)\,(1-\pi) \Big/ \sum_s \delta\,(1)^2$ *(valeur* ν *qui minimise* $\|\nu\delta\,(1)-(1-\pi)\|$*)*

– $v_h : c = (1-\pi)\,|s| \,/\, |s| - \operatorname{rang}(x_s)$

– $n_1 : c < \frac{1}{4\pi}, \min_c \|\delta\,(c)-(1-\pi)\|_{\delta(1)}$ [31]*, calage final sur la trace de l'ESB*

– $n_2 : c < 5, \min_c \|\delta\,(c)-(1-\pi)\|_{\delta(1)}$*, calage sur la trace de l'ESB*

– n_3 *: pas de plafonnement de* c *mais contrainte* $\delta\,(c) < 1$, $\min_c \|\delta\,(c)-(1-\pi)\|_{\delta(1)}$*, calage final sur la trace de l'ESB*

– ren *: algorithme récursif d'au plus 500 itérations* [32]*. Un calage final sur la trace de l'ESB est effectué s'il diminue strictement la fonction objectif* [33]*.*

– rel *: Cet algorithme récursif diffère du précédent uniquement par les variables d'équilibrage complémentaire, qui ne sont pris en compte que pour les régions de taille d'échantillon supérieure à 100, pour le calcul à la fois de* c *et des résidus d'équilibrage.*

– cen *: calage systématique de* ren *sur la trace de l'ESB*

– *L'indexation de la table contenant les* 10 000 *échantillons simulés joue crucialement sur la performance du programme. A défaut, la rapidité de lecture d'un échantillon se détériore avec l'indice de sa position. Par exemple, pour* $i = 4\,532$*, l'indexation réduit la durée de lecture d'un échantillon simulé de 2 minutes et 48 secondes à 0.10 seconde.*

– *Les variables étudiées sont issues de la base de sondage, donc de totaux connus :*

- `foy_rev`=*revenu du foyer fiscal*
- `nb_pi2`=*nombre de logements principaux TH de deux pièces*
- `mar`=*(définition incertaine) nombre de foyers de couples mariés*
- `isf`=*nombre de foyers redevables de l'ISF*

---

31. La pondération de la norme par $\delta\,(1)$ vise à limiter l'augmentation de poids c_i qui accroit $\delta_j\,(c)$, où $j \neq i$, lorsque c_j atteint le taquet $\overline{c}$.

32. Les autres critères d'arrêt sont : $\left(\sum cuu'\right)^-$ pas calculable par *ginv*, $\|\delta\,(c)-(1-\pi)\| < 10^{-8}$ et $\|\Delta\delta\,(c)\| < 10^{-8}$ (la variation du vecteur des coefficients diagonaux entre les deux dernières itérations).

33. Le calage final sur la trace est effectué ainsi, pour une strate h : $\widetilde{c} = c\sum_{s_h}(1-\pi) \,/\, \sum_{s_h}\delta\,(c)$. Il tend à augmenter l'estimation de variance.

Tableau 9 – Distribution sur 10 000 simulations du taux d'écart à l'erreur quadratique moyenne

variable		moy	moyvar	min	p5	p10	q1	médian	q3	p90	p95	max	rang
foy_rev	v1	4,7	8,5	-40,5	-28,5	-24,1	-15,1	-2,0	16,2	43,6	65,2	155,9	1
foy_rev	v0	5,5	9,3	-40,0	-27,9	-23,5	-14,5	-1,2	17,1	44,8	66,5	157,9	2
foy_rev	v2	10,1	13,9	-37,0	-24,4	-19,9	-10,4	3,3	22,3	50,0	72,0	163,7	11
foy_rev	va	10,1	13,9	-37,0	-24,4	-19,9	-10,4	3,3	22,3	50,1	72,0	163,7	10
foy_rev	vb	9,0	12,8	-37,7	-25,3	-20,7	-11,3	2,2	21,2	48,8	70,8	162,5	4
foy_rev	vh	11,4	15,3	-36,1	-23,3	-18,8	-9,2	4,6	23,9	51,8	73,7	165,5	12
foy_rev	n1	9,3	13,0	-37,3	-24,7	-20,1	-10,7	2,9	21,4	48,7	70,0	160,1	8
foy_rev	n2	9,3	13,0	-37,3	-24,7	-20,1	-10,7	2,8	21,4	48,7	70,0	160,1	7
foy_rev	n3	9,4	13,1	-37,1	-24,7	-20,2	-10,7	2,8	21,6	48,8	70,5	160,5	6
foy_rev	ren	8,8	12,5	-37,7	-25,1	-20,6	-11,2	2,3	20,8	48,0	69,5	159,4	5
foy_rev	cen	9,3	13,0	-37,3	-24,7	-20,1	-10,6	2,9	21,4	48,7	70,0	160,1	9
foy_rev	rel	7,8	11,5	-38,3	-26,0	-21,4	-12,3	1,2	19,8	47,1	68,9	159,1	3
nbpi2	v1	-7,8	-7,8	-16,6	-11,4	-10,6	-9,3	-7,8	-6,3	-4,9	-4,1	0,7	1
nbpi2	v0	-7,1	-7,0	-15,9	-10,7	-9,8	-8,6	-7,1	-5,6	-4,2	-3,4	1,5	2
nbpi2	v2	3,0	3,1	-6,5	-0,9	-0,1	1,4	3,0	4,6	6,1	7,1	13,5	6
nbpi2	va	3,0	3,1	-6,5	-0,9	-0,1	1,4	3,0	4,6	6,1	7,1	13,5	7
nbpi2	vb	0,5	0,5	-8,7	-3,4	-2,5	-1,1	0,5	2,1	3,6	4,5	10,3	4
nbpi2	vh	5,8	5,8	-4,3	1,7	2,6	4,1	5,8	7,5	9,0	10,0	17,4	12
nbpi2	n1	4,3	4,3	-5,7	0,2	1,0	2,5	4,3	6,0	7,6	8,6	14,1	10
nbpi2	n2	4,2	4,2	-5,7	0,1	1,0	2,5	4,2	5,9	7,5	8,4	14,0	9
nbpi2	n3	3,5	3,5	-6,0	-0,5	0,4	1,8	3,5	5,1	6,7	7,6	14,3	8
nbpi2	ren	2,9	3,0	-6,6	-1,1	-0,3	1,2	2,9	4,6	6,2	7,2	12,3	5
nbpi2	cen	4,3	4,3	-5,7	0,2	1,1	2,6	4,3	6,0	7,6	8,6	14,1	11
nbpi2	rel	-0,4	-0,3	-9,9	-4,3	-3,5	-2,0	-0,4	1,3	2,8	3,7	9,0	3
mar	v1	-18,1	-18,1	-24,8	-20,9	-20,3	-19,3	-18,1	-16,9	-15,9	-15,3	-9,7	1
mar	v0	-17,4	-17,4	-24,2	-20,3	-19,6	-18,6	-17,4	-16,3	-15,2	-14,6	-9,0	2
mar	v2	-6,0	-6,0	-13,4	-9,2	-8,6	-7,4	-6,0	-4,6	-3,4	-2,7	3,5	5
mar	va	-6,0	-6,0	-13,4	-9,2	-8,6	-7,4	-6,0	-4,6	-3,4	-2,7	3,5	6
mar	vb	-8,9	-8,9	-16,1	-12,0	-11,4	-10,2	-8,9	-7,6	-6,4	-5,7	0,3	4
mar	vh	-2,8	-2,8	-10,2	-6,2	-5,5	-4,3	-2,8	-1,4	-0,1	0,6	6,9	12
mar	n1	-3,9	-3,9	-12,5	-7,5	-6,8	-5,5	-4,0	-2,4	-1,0	-0,2	5,9	10
mar	n2	-4,0	-4,0	-12,5	-7,6	-6,9	-5,6	-4,1	-2,6	-1,2	-0,3	6,1	9
mar	n3	-5,3	-5,3	-12,9	-8,6	-7,9	-6,7	-5,3	-3,9	-2,6	-1,9	4,4	8
mar	ren	-5,6	-5,6	-13,8	-9,1	-8,4	-7,1	-5,6	-4,1	-2,7	-1,9	4,2	7
mar	cen	-3,9	-3,9	-12,4	-7,5	-6,8	-5,5	-4,0	-2,4	-1,0	-0,1	5,9	11
mar	rel	-9,8	-9,8	-17,7	-13,1	-12,4	-11,2	-9,9	-8,5	-7,1	-6,4	-0,4	3
isf	v1	13,6	13,7	-9,7	3,5	5,6	9,3	13,3	17,6	21,8	24,3	39,9	1
isf	v0	14,4	14,6	-9,0	4,3	6,4	10,1	14,2	18,6	22,7	25,2	41,0	2
isf	v2	20,2	20,4	-3,4	10,0	12,2	15,9	20,0	24,4	28,5	31,0	47,0	8
isf	va	20,2	20,4	-3,4	10,0	12,2	15,9	20,0	24,4	28,5	31,0	47,0	7
isf	vb	18,9	19,0	-4,7	8,7	10,8	14,5	18,7	23,0	27,2	29,7	45,7	4
isf	vh	21,9	22,1	-1,7	11,7	13,8	17,6	21,7	26,1	30,2	32,7	48,8	12
isf	n1	20,7	20,9	-3,5	10,3	12,5	16,3	20,4	25,0	29,1	31,7	48,6	10
isf	n2	20,7	20,8	-3,6	10,2	12,4	16,3	20,4	25,0	29,0	31,6	48,6	9
isf	n3	20,1	20,3	-3,8	9,8	12,0	15,8	19,9	24,4	28,4	31,1	48,0	6
isf	ren	20,0	20,2	-4,2	9,6	11,8	15,6	19,7	24,3	28,4	31,0	48,0	5
isf	cen	20,7	20,9	-3,5	10,3	12,5	16,3	20,4	25,0	29,1	31,7	48,6	11
isf	rel	18,7	18,8	-5,6	8,2	10,4	14,2	18,4	23,0	27,1	29,7	46,4	3

– La variance estimée diffère selon que les variables d'intérêt sont régressées séparément par la procédure REG de SAS ou simultanément. L'estimation est toujours plus élevée avec la régression 'simultanée' des 4 variables. Ceci affecte sensiblement les pondérations à valeurs élevées. Par exemple, le taux d'erreur maximal sur la variance du revenu par la méthode ren atteint 1353%. Plutôt qu'un signal de non robustesse, cette anomalie est causée par les observations de valeur manquante sur l'une des variables dépendantes qui sont exclues par la PROC REG du calcul du coefficient ($\widehat{\beta}$), mais pas de celui des résidus. Seule la variable `isf` a des valeurs manquantes dans la base de sondage [34], d'où l'absence de différence entre les deux calculs de l'approximation de variance pour cette variable.

→ Du fait de poids maximaux plus élevés, la pondération optimisée est beaucoup plus sensible à cette incohérence entre les champs du calcul du coefficient et du résidu de la régression. Il faut donc veiller à l'éviter.

La méthode d'optimisation de la pondération par récursion (ren) donne une estimation de variance préférable à la meilleure non-optimisée (v_2) dans plus de la moitié des simulations pour les quatre variables (Tableau 10). Mais le bilan est plus mitigé pour les autres pondérations optimisées. Donc il n'y a pas de relation simple entre la hiérarchie des méthodes selon la minimisation de la distance entre diagonales obtenue (Tableau 2) et celle de la qualité de l'estimation de variance.

Tableau 10 – Nombre de simulations où l'estimation de variance est préférable à celle de v_2

variable	v_h	n1	n2	n3	ren	cen	rel
foy_rev	4 399	5 625	5 595	5 549	5 335	5 617	5 179
nbpi2	1 038	1 152	1 137	1 049	5 424	1 151	4 382
mar	9 986	9 984	9 985	9 986	7 193	9 984	0
isf	3	3 100	3 211	6 944	6 975	3 071	9 400

Note : Pour une simulation, une méthode d'estimation de la variance m_1 est jugée préférable à m_2 si m_1 surestime l'écart quadratique moyen (calculé sur l'ensemble des simulations) et m_2 soit le surestime encore davantage soit le sous-estime strictement, ou alors m_1 sous-estime le reqm et m_2 encore plus.

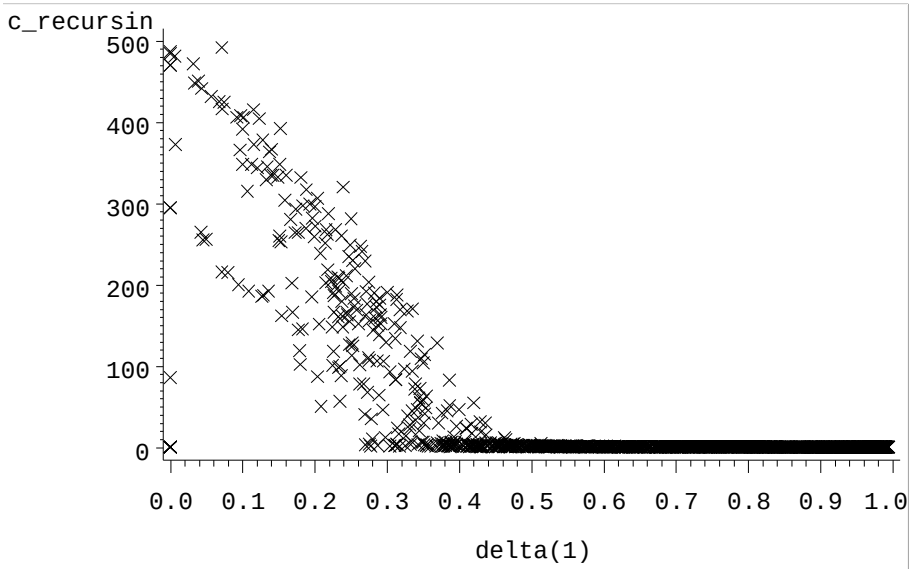
Au total, dans le cas de l'échantillon de première phase d'EEC-TH, la minimisation de la distance entre les diagonales de l'approximation de variance et de l'estimateur sans biais fournit des pondérations de qualité de l'estimation de variance acceptable par rapport au simple calage sur la trace. L'algorithme récursif, qui 'pousse' cette minimisation plus loin, semble fournir une estimation meilleure, en médiane des simulations. De plus, il s'avère possible de moduler cet algorithme. Par exemple, pour le critère de proximité $\|\delta(c) - (1-\pi)\|^2 + (\sum \delta(c) - \sum 1-\pi)^2/|s|$, une variante de l'algorithme récursif a fourni une distribution des précisions sur les 10 000 simulations très proche de n_2 [35], mais avec une pondération maximale encore plus élevée que ren. Un calage final systématique de la pondération récursive donne des estimations presque indistinguables de celles de n_1. C'est vrai également pour la version multiplicative.

L'incidence de l'optimisation de la pondération est limitée, selon la comparaison des versions v_h et ren. L'écart entre les médianes représente 2 à 3% de la racine carrée de l'écart quadratique moyen, c'est-à-dire de l'objectif à estimer pour le calcul de variance. Mais cette différence n'est pas négligeable en comparaison au taux d'erreur de cette estimation pour 3 des 4 variables testées.

34. Il n'y a qu'une dizaine d'unités concernées dans la base de sondage.
35. On se serait attendu plutôt à des résultats intermédiaires entre les versions v_2 et ren.

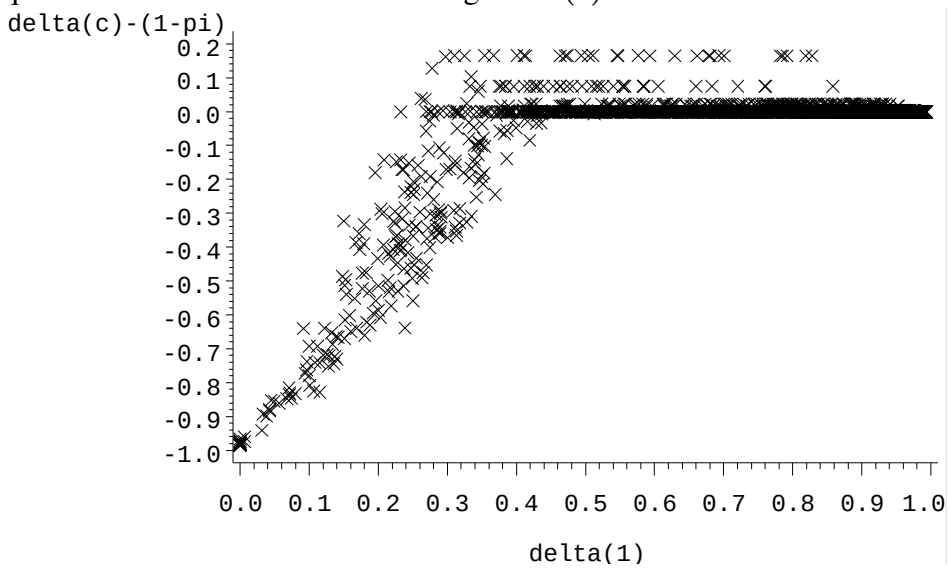
Le rôle du coefficient $\delta(1)$ dans les pondérations élevées produites par les méthodes de minimisation est illustré de manière claire sur la première simulation de la méthode récursive (Graphique 1). Les pondérations de plus de 100 correspondent à des coefficients $\delta(1) \leqslant 0.4$.

Graphique 1 – Pondération c 'minimisante' de la méthode récursive en fonction du coefficient $\delta(1)$, pour la première simulation d'échantillonnage



Même la méthode récursive ne parvient pas à pousser la minimisation suffisamment pour les plus petits coefficients $\delta(1)$, c'est-à-dire les unités d'indicatrice proche de $\mathrm{Im}(u_s)$ (Graphique 2). De plus, l'augmentation de poids de ces unités a une incidence à la hausse des coefficients diagonaux des autres unités, si la compensation par le poids de celles-ci n'est pas suffisante ou est bloquée par le rejet de l'itération suivante. Ce mécanisme explique sans doute les coefficients diagonaux excédant $1 - \pi$. Ce phénomène traduit en partie l'absence de solution exacte, si l'algorithme fonctionne correctement [36], ou le nombre insuffisant d'itérations.

Graphique 2 – Écart entre le coefficient diagonal $\delta(c)$ et $1 - \pi$ en fonction du coefficient $\delta(1)$



36. Des coefficients diagonaux excessifs ou insuffisants sont vraisemblablement inéluctables dès qu'il n'y a pas de solution exacte à l'équation de pondération mais que la dérivée est nulle pour une solution minimisante, vu (**e3.45**).

– La faiblesse des tailles d'échantillon de certaines strates régionales, ainsi que la dégradation de la qualité numérique de la diagonale calculée lorsque la pondération augmente, pourrait limiter l'avantage de la pondération optimisée (ren) dans le contexte étudié. En effet, la taille d'échantillon comparée à la dimension de l'espace d'équilibrage joue notamment sur la fréquence des unités i telles que $\mathbb{1}_i \in \mathrm{Im}\,(u_s)$, donc $\delta_i\,(1) = 0$.

– La variable testée de variance la plus surestimée par la méthode des résidus, `isf`, a le coefficient de variation le plus élevé (Tableau 11), et inversement pour la plus sous-estimée (`mar`). Si le plan de sondage est assimilé à un SAS, on s'attendrait à ce que l'approximation poissonnienne surestime davantage les variables de ratio $\hat{\mathcal{V}}^2\,(y) \Big/ \dfrac{\sum y^2}{|\mathscr{P}|}$ plus bas, mais l'inverse est observé.

Tableau 11 – Distribution des variables testées dans la base de sondage

statistique	foy rev	nbpi2	mar	isf
CV	74,7	69,8	57,9	186,5
moyenne	5 198 633	25	78,1	2,4
$100\hat{\mathcal{V}}^2\,(y) \Big/ \dfrac{\sum y^2}{\lvert\mathscr{P}\rvert}$	35,8	32,7	25,1	77,7
R^2	0,77	0,84	0,98	0,52
reqm/total vrai, en %	1,1	0,9	0,3	2,4

– Le lien de la variable d'intérêt avec les variables de l'équilibrage pourrait expliquer la sous-estimation de l'écart quadratique moyen de la variable `mar`. Le R^2 très élevé de cette variable pourrait correspondre à un poids relatif de la variance d'atterrissage plus important, selon G. Chauvet [GC]. Cependant, cette sous-estimation porte sur un reqm très petit relativement à la moyenne de la variable.

– L'écart quadratique moyen augmente sensiblement entre $1\,000$ et $10\,000$ simulations pour la variable `mar`, alors qu'il diminue amplement pour `isf` (Tableau 8). Cette instabilité est aussi observée en comparant l'écart quadratique moyen calculé sur les deux moitiés des simulations (Tableau 12). Ces résultats suggèrent qu'une autre piste d'explication des différences observées entre les variables pour la qualité d'estimation de la variance est que l'écart quadratique lui-même serait sur- ou sous-évalué sur les $10\,000$ simulations.

Tableau 12 – Taux d'écart entre les reqm des $5\,000$ premières et dernières simulations

reg	total	11	21	22	23	24	25	26	31	41	42	43	52	53	54	72	73	74	82	83	91	93	94
foy rev	2,0	2,0	-5,5	-2,9	-0,6	3,9	-0,3	10,0	3,4	-2,3	1,8	1,9	5,3	-9,8	8,2	6,9	-1,9	-2,8	-7,8	0,9	-4,0	2,8	0,6
nbpi2	-0,7	2,8	5,0	4,4	-6,6	2,3	0,4	-1,4	2,1	-3,2	0,0	-4,5	-1,7	-5,1	1,3	-2,3	-0,1	2,8	-0,2	2,3	-3,6	1,5	-0,8
mar	3,4	3,6	-2,4	-1,4	-4,4	3,2	5,9	3,0	1,5	0,4	1,2	-2,1	4,1	-2,9	-2,7	-2,2	3,8	-0,4	3,4	0,7	5,3	-2,9	2,0
isf	-4,6	-2,7	0,7	-1,1	-10,6	-3,8	-3,0	-1,3	2,2	4,2	-2,5	-3,6	-0,9	0,7	2,4	-2,4	3,1	-5,0	1,3	-2,7	6,2	2,2	7,9

5.1 efficacité pour les quatre plus grandes régions

La comparaison des algorithmes pour l'Ile-de-France, Rhône-Alpes, Provence-Alpes-Côte d'Azur et le Nord-Pas-de-Calais parait utile pour évaluer le rôle de la faiblesse de la taille d'échantillon régional dans l'incidence limitée de l'optimisation des poids des résidus d'équilibrage sur la qualité de l'approximation de variance. En effet, l'échantillon francilien est de plus de 500 pour la première phase de l'enquête Emploi. Parallèlement, les algorithmes de minimisation annulent presque la distance $\delta(c) - (1 - \pi)$ pour cette région, hors unités de très faible $\delta(1)$ le cas échéant. Pour les trois régions, le poids des coefficients diagonaux supérieurs à 1 est quasiment nulle pour l'algorithme récursif. Donc ces régions pourraient permettre d'évaluer le rôle de l'absence de solution exacte à l'équation de pondération dans les résultats limités de l'optimisation de la pondération.

En fait, les deux pondérations retenues ici ne sont pas clairement hiérarchisées par la qualité de l'estimation de variance pour ces trois régions (Tableau 13). (Les résultats obtenus pour la réunion des trois plus grandes régions sont proches de ceux de l'Ile-de-France.)

Tableau 13 – Distribution de l'erreur d'estimation de la variance pour les quatre régions

région	variable		moy	moyvar	min	p5	p10	q1	médian	q3	p90	p95	max
IdF	foy_rev	v2	13,2	18,5	-41,7	-26,2	-21,5	-10,8	4,5	26,5	63,4	89,6	202,0
	foy_rev	ren	12,1	17,4	-42,2	-26,9	-22,1	-11,6	**3,7**	25,4	61,2	87,1	197,8
	nbpi2	v2	8,2	8,3	-6,5	1,4	2,9	5,4	8,1	10,9	13,5	15,2	27,6
	nbpi2	ren	8,4	8,5	-6,6	1,6	3,1	5,6	8,3	11,2	13,8	15,5	28,7
	mar	v2	4,0	4,1	-11,6	-2,1	-0,9	1,4	4,0	6,5	9,0	10,5	18,4
	mar	ren	4,9	5,0	-11,3	-1,6	-0,3	2,1	4,8	7,5	10,2	11,8	19,6
	isf	v2	36,1	36,4	0,7	21,1	24,3	29,7	36,0	42,4	48,3	51,9	72,8
	isf	ren	35,4	35,8	-0,3	20,3	23,5	29,0	**35,2**	41,8	47,7	51,4	73,4
RA	foy_rev	v2	-9,9	2,8	-60,2	-48,0	-45,3	-39,2	-26,1	-2,3	42,0	84,3	322,0
	foy_rev	ren	-10,5	1,3	-60,5	-47,9	-45,3	-39,1	**-26,1**	-1,9	38,5	80,2	304,8
	nbpi2	v2	3,9	4,2	-19,4	-8,4	-6,0	-1,9	3,2	8,7	14,5	18,5	44,0
	nbpi2	ren	4,9	5,3	-19,5	-7,8	-5,4	-1,1	4,2	10,0	16,3	20,3	48,7
	mar	v2	1,2	1,4	-20,8	-9,2	-7,1	-3,4	0,9	5,4	9,7	12,5	31,8
	mar	ren	3,4	3,7	-20,1	-8,0	-5,6	-1,7	2,9	8,0	13,1	16,6	44,8
	isf	v2	2,9	4,3	-41,9	-22,7	-18,2	-9,2	1,5	13,9	25,7	32,9	69,9
	isf	ren	4,0	5,4	-42,2	-22,4	-17,6	-8,6	2,3	15,4	28,0	35,6	77,5
PACA	foy_rev	v2	-3,8	4,7	-47,8	-30,7	-27,2	-21,0	-12,7	-0,8	15,4	34,6	315,8
	foy_rev	ren	-3,9	4,2	-47,5	-30,8	-27,4	-21,0	**-12,6**	-0,7	15,4	35,0	306,4
	nbpi2	v2	6,9	7,1	-16,3	-3,7	-1,6	2,3	6,7	11,3	15,6	18,4	37,2
	nbpi2	ren	7,6	7,8	-16,7	-3,5	-1,2	2,7	7,5	12,2	16,6	19,7	38,3
	mar	v2	-3,8	-3,6	-27,2	-15,1	-12,8	-9,0	-4,3	0,9	5,8	8,7	25,8
	mar	ren	-1,7	-1,3	-26,8	-14,0	-11,5	-7,4	**-2,3**	3,3	9,2	13,1	35,4
	isf	v2	1,7	3,4	-43,0	-22,5	-18,6	-11,0	-1,6	9,9	27,9	40,3	99,7
	isf	ren	2,8	4,6	-42,6	-22,5	-18,3	-10,6	**-0,7**	11,8	30,8	42,3	101,5
NPDC	foy_rev	v2	-6,6	-5,0	-43,9	-29,4	-26,1	-19,1	-9,4	2,6	16,0	25,1	83,7
	foy_rev	ren	-5,7	-4,1	-42,9	-29,3	-25,8	-18,6	**-8,6**	4,4	17,3	26,8	97,4
	nbpi2	v2	-1,5	-0,9	-32,0	-17,3	-14,4	-8,9	-2,2	5,0	12,6	17,2	37,6
	nbpi2	ren	-0,4	0,2	-31,8	-17,0	-13,9	-8,2	**-1,3**	6,3	14,5	19,5	42,3
	mar	v2	-8,3	-8,1	-34,1	-19,4	-16,9	-12,9	-8,5	-3,8	0,4	3,1	20,7
	mar	ren	-7,3	-7,0	-33,2	-19,1	-16,5	-12,2	**-7,5**	-2,5	2,1	5,1	22,0
	isf	v2	-12,6	-8,5	-58,8	-46,1	-41,8	-32,8	-18,5	1,0	27,6	45,3	112,1
	isf	ren	-9,0	-4,3	-59,2	-45,3	-40,8	-31,3	**-15,8**	6,1	34,4	51,0	132,2

Conclusion

– Sur le plan théorique, l'étude a mis en évidence une condition nécessaire d'existence d'une solution exacte à l'équation de pondération, qui égalise les deux diagonales. Elle reflète que les unités atypiques entravent le rapprochement des diagonales, comme l'équilibrage. Parce que la condition nécessaire n'est pas toujours vérifiée lorsque la dimension d'équilibrage est supérieure à 1, il a semblé judicieux d'élargir l'objectif à la minimisation d'une distance entre les deux diagonales. Dans cette optique, l'algorithme récursif assure une baisse très sensible de la distance inter-diagonales par rapport à la pondération par $1 - \pi$, même calée sur la trace. Le deuxième résultat méthodologique est une condition suffisante pour qu'une pondération produite par cette procédure soit finie.

– En revanche, l'existence d'une solution minimisant la distance entre les diagonales des deux estimateurs de variance n'a pas pu être démontrée. La convergence de l'algorithme vers une solution exacte n'est pas assurée non plus. De plus, les propriétés de la pondération minimisante (éventuelle) restent en partie inconnues. Il n'y a pas de garantie théorique qu'une telle pondération soit non nulle et finie et a fortiori qu'elle soit préférable pour la qualité de l'estimation de variance. Une autre piste de recherche méthodologique est d'améliorer le critère de proximité, par exemple en prenant en compte la trace. En effet, le choix de distance joue un rôle crucial dans le compromis entre les 'excès' et les 'défauts' des coefficients diagonaux, en l'absence de solution exacte à l'équation de pondération.

– Les résultats des différentes versions testées suggèrent qu'il est encore possible d'améliorer sensiblement l'algorithme de minimisation de la distance entre diagonales.

– Empiriquement, selon l'objectif technique de rapprochement des diagonales, l'algorithme récursif ressort clairement comme le meilleur choix parmi ceux évalués pour optimiser la pondération. Cependant, l'ampleur et la régularité de la minimisation par cet algorithme restent inexpliquées. Donc il n'est pas sûr que la méthode soit aussi performante sur d'autres données.

– Pour l'objectif final d'améliorer la qualité de l'estimation de variance, les résultats obtenus confortent globalement le choix d'optimiser la pondération selon le principe de minimisation de la distance des diagonales. Cette méthode réduit l'erreur d'estimation de la variance de manière relativement sensible. Les poids maximaux élevés ne paraissent pas affecter la robustesse de la méthode. Mais l'enjeu par rapport à la pondération simple 'calée sur la trace' est limité relativement à la variance, sur les données étudiées. La supériorité de la pondération optimisée ressort donc moins clairement que dans les tests de [DT], peut-être du fait du contexte d'absence de solution exacte. L'écart entre la pondération optimisée et celle de Hajek est de 1 à 3% de l'écart-type à estimer. Le contrôle des coefficients diagonaux peut importer pour l'estimation de variance d'un sondage en deux phases.

La conclusion opérationnelle de l'étude est qu'il faut éviter de pondérer l'estimateur de variance par 1 ou par $1 - \pi$, parce que cela mène en général à une sous-estimation sensible de la variance. Le choix parmi les autres pondérations est moins tranché. La pondération optimisée par la procédure récursive constitue une alternative crédible aux pondérations plus 'simples' et peut être préférée au calage de $1 - \pi$ sur la trace pour le plan de sondage étudié. Cependant, la justification théorique et les résultats obtenus dans la présente étude peuvent être jugés insuffisants pour préférer ce choix à la pondération de Hajek préconisée par [AMYT]. Celle-ci limite davantage les sous-estimations, au prix de surestimations accentuées par ailleurs.

Références

[DT] Variance approximation under balanced sampling, JC Deville, Y Tillé, Journal of statistical planning and inference 2005

[AMYT] Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size, A.Matei, Y.Tillé, Journal of Official Statistics vol.21 N° 4, 2005

[JCD] Variance estimation for complex statistics and estimators : linearisation and residual techniques, JC Deville, Survey Methodology, vol.25 N° 2, 1999

[VL] La construction du nouvel échantillon de l'enquête emploi en continu à partir des fichiers de la taxe d'habitation, V.Loonis, Actes des JMS 2009

[CDL] Weighted finite population sampling to maximize entropy, XH Chen, AP Dempster, JS Liu, Biometrika, 1994

[GC] On variance estimation for the French master sample, Journal of Official Statistics, vol.27 N° 4, 2011

Annexe A matrices symétriques

– Pour une matrice A d'un espace vectoriel $E(=\mathbb{R}_s)$ dans un espace F :

• $\text{rang}(A) + \dim[\text{Ker}(A)] = \dim(E)$: Il existe un sous-espace $\Delta E \,/\, E = \text{Ker}(A) \oplus \Delta E$ et alors A est bijective de ΔE sur $\text{Im}(A)$.

• $$\text{Im}(A) = \text{Ker}\left(A'\right)^{\perp} \qquad \textbf{(eA.56)}$$

car $\text{Im}(A)^{\perp} = \{\langle AE, \rangle = 0\} = \left\{\left\langle E, A'\right\rangle = 0\right\} = \left\{A' = 0\right\} = \text{Ker}\left(A'\right)$ (Il en découle que $\text{rang}\left(A'\right) = \dim\left[F - \text{Ker}\left(A'\right)\right] = \dim\left[\text{Ker}\left(A'\right)^{\perp}\right] = \text{rang}(A)$.)

• $$\text{Im}(A) = \text{Im}\left(AA'\right) \qquad \textbf{(eA.57)}$$

↳ $\text{Im}\left(AA'\right) \subset \text{Im}(A)$

↳ $\text{Ker}\left(AA'\right) = \text{Ker}\left(A'\right)$ car $\forall h, h'AA'h = \left\|A'h\right\|^2$

↳ $\Rightarrow \text{rang}\left(AA'\right) = \text{rang}\left(A'\right) = \text{rang}(A)$

– Si A est une matrice symétrique ($A' = A$) alors :

• $\text{Im}(A) = \text{Ker}(A)^{\perp}$

• A est bijective sur $\text{Im}(A)$ (car surjective et $\text{Im}(A) \cap \text{Ker}(A) = 0 \Rightarrow A$ injective sur $\text{Im}(A)$)

– Si A est une matrice symétrique alors elle est diagonalisable dans une base orthonormée. Donc il existe une matrice orthogonale U, c'est-à-dire que $UU' = U'U = \text{id}$, et une matrice D diagonale telles que $A = UDU'$.

– Par définition, une matrice carrée est positive si $\forall x, x'Ax \geqslant 0$ et définie positive si $\forall x \neq 0, x'Ax > 0$. Pour une matrice définie positive A, $\|x\|_A = \sqrt{x'Ax}$ est une norme [37].

– Pour deux matrices symétriques A et B, $A \leqslant B \overset{\text{déf}}{\Leftrightarrow} \forall x, x'Ax \leqslant x'Bx$

– Si tous les mineurs principaux d'une matrice M (non nulle) symétrique sont positifs, alors le déterminant s'écrit $\left|M - \lambda\,\text{id}\right| = \sum_k (-\lambda)^k a_k$ avec $a_k \geqslant 0$, et $a \neq 0$. Donc ce polynôme en λ n'admet pas de racines strictement négatives. (Par ailleurs, le critère de Sylvester, plus fort, énonce que si les n mineurs principaux du type $\left|M_{\leqslant k}\right|$ sont > 0 alors M est définie positive.)

37. A est (définie) positive si la matrice symétrique $A + A'$ est (définie) positive, car $x'Ax = x'A'x = x'(A + A')x/2$ et donc $\|x\|_A = \|x\|_{(A+A')/2}$. Néanmoins, l'extension de la définition peut être utile parce que l'ensemble des matrices symétriques n'est pas un ouvert.

– Si Δ est une matrice symétrique et positive alors pour tous les indices i, j :

$$\forall x, (x\mathbb{1}_i + \mathbb{1}_j)'\Delta(x\mathbb{1}_i + \mathbb{1}_j) \geqslant 0$$

$$\text{donc } \forall x, x^2\Delta_{i,i} + 2\Delta_{i,j}x + \Delta_{j,j} \geqslant 0$$

$$\text{donc le discriminant est négatif : } \Delta_{i,j}^2 - \Delta_{i,i}\Delta_{j,j} \leqslant 0 \qquad \textbf{(eA.58)}$$

→ Cette condition n'est pas suffisante si $|s| > 2$: soit $a_s > 0$ et $\Delta_{i,j} = (2\mathbb{1}_{i=j} - 1)\sqrt{a_i}\sqrt{a_j}$

$$\sum_i \sqrt{a_i}^2 \frac{1}{\sqrt{a_i}}\frac{1}{\sqrt{a_i}} - \sum_{i\neq j} \sqrt{a_i}\sqrt{a_j}\frac{1}{\sqrt{a_i}}\frac{1}{\sqrt{a_j}} = |s| - |s|(|s| - 1) = |s|(2 - |s|) < 0$$

Annexe B inverses généralisées

– A^- désigne une inverse généralisée de la matrice carrée A, définie par la propriété : $AA^-A = A$. Cette définition est telle que si $b \in \text{Im}(A)$ alors $a = A^-b$ est une solution de l'équation $Aa = b$.

→ Si $B[\text{Im}(A)] \subset \text{Ker}(A)$ alors $A^- + B$ est une inverse généralisée.

– Deux inverses généralisées de A, A^- et $\widetilde{A^-}$, diffèrent par une matrice B telle que $B[\text{Im}(A)] \subset \text{Ker}(A)$, vu que $A(A^- - \widetilde{A^-})A = 0$.

– La qualité d'une inverse généralisée est mesurée dans la présente étude par $\|A - AA^-A\|$, avec la norme euclidienne, c'est-à-dire la racine carrée de la somme des carrés des écarts entre les coefficients des deux matrices.

– Pour une matrice A, A^+ est l'inverse généralisée de Moore-Penrose (MP). Elle est définie par les quatre propriétés :

$$AA^+A = A \qquad \text{(mp1)}$$

$$A^+AA^+ = A^+ \qquad \text{(mp2)}$$

$$(A^+A)' = A^+A \qquad \text{(mp3)}$$

$$(AA^+)' = AA^+ \qquad \text{(mp4)}$$

$$\text{Im}(A) \overset{(mp1)}{\subset} \text{Im}(AA^+) \overset{(mp4)}{=} \text{Im}\left[(A^+)'A'\right] \subset \text{Im}\left[(A^+)'\right] \overset{(mp2)}{\subset} \text{Im}\left[(AA^+)'\right] \overset{(mp4)}{=} \text{Im}(AA^+) \subset \text{Im}(A)$$

$$\Rightarrow \text{Im}(A) = \text{Im}\left[(A^+)'\right] = \text{Ker}(A^+)^\perp \qquad \textbf{(eB.59)}$$

- La première égalité de (**eB.59**) entraine que $\text{rang}[A^+] = \text{rang}(A)$.

→ Comme A et A^+ jouent des rôles symétriques dans les propriétés de Moore-Penrose, on peut les permuter dans (**eB.59**), d'où l'image de l'inverse MP :

$$\text{Im}(A^+) = \text{Im}[A'] = \text{Ker}(A)^\perp \qquad \textbf{(eB.60)}$$

⇒ Si A est une matrice symétrique alors $\text{Im}(A^+) = \text{Im}(A)$ et $\text{Ker}(A^+) = \text{Ker}(A)$.

- unicité : Si A^+ et $\widetilde{A^+}$ vérifient les quatre conditions de Moore-Penrose, alors elles ont même noyau $(\text{Im}(A)^\perp$, vu **eB.59**). Par conséquent, il existe une matrice inversible B telle que $\widetilde{A^+} = BA^+$.

(mp1) $\Leftrightarrow ABA^+A = A$. D'autre part, (mp3)$\Rightarrow (BA^+A)' = BA^+A$ soit : $A^+AB' = BA^+A \Rightarrow AA^+AB' = ABA^+A \Rightarrow AB' = ABA^+A = A$, d'après la première égalité. Donc $AB' = A \Rightarrow A^+AB' = A^+A \overset{(mp3)}{\Rightarrow} BA^+A = A^+A$ et en multipliant à droite par A^+, il vient finalement $A^+ = BA^+ = \widetilde{A^+}$

⇒ Les conditions de Moore-Penrose définissent une inverse unique (si elle existe). Il en découle :

- $(A^+)' = (A')^+$
- si $\alpha \neq 0$ alors $(\alpha A)^+ = \frac{1}{\alpha}A^+$
- si A et B sont deux matrices symétriques et si $AB = 0 = BA$ alors :

$$(A + B)^+ = A^+ + B^+ \qquad \textbf{(eB.61)}$$

$$\begin{cases} \text{Im}(A^+) = \text{Im}(A) \subset \text{Ker}(B) = \text{Ker}(B^+) \\ \text{Im}(B^+) = \text{Im}(B) \subset \text{Ker}(A) = \text{Ker}(A^+) \end{cases} \Rightarrow \begin{cases} A^+B = 0 \\ B^+A = 0 \\ BA^+ = 0 \\ AB^+ = 0 \end{cases} \Rightarrow \begin{cases} (A^+ + B^+)(A + B) = A^+A + B^+B \\ (A + B)(A^+ + B^+) = AA^+ + BB^+ \end{cases}$$

Les six dernières identités entraînent que $A^+ + B^+$ vérifient les propriétés de Moore et Penrose pour $A + B$.

→ Si A est une matrice symétrique alors l'inverse de Moore-Penrose est donnée par $A^+ = UD^+U'$, où D^+ est la matrice diagonale constituée de l'inverse des coefficients non nuls de D (une diagonalisation de A) et de ses coefficients nuls. L'inverse généralisée peut également dans ce cas se construire à partir du fait qu'une matrice symétrique A est bijective sur $\mathrm{Im}(A)$ et nulle sur $\mathrm{Im}(A)^\perp$.

– si $A = \alpha uu'$ avec $\alpha \neq 0$ et $\|u\| \neq 0$ alors $A^+ = \dfrac{1}{\alpha}\dfrac{uu'}{\|u\|^4}$

– décroissance de l'inversion de Moore et Penrose :

Si A et B sont deux matrices symétriques de même image :

$$\text{si } A \leqslant B \text{ alors } B^+ \leqslant A^+ \qquad \textbf{(eB.62)}$$

↳ Un même changement de base orthonormée permet de se ramener à deux matrices symétriques et inversibles :

$$A = U\begin{pmatrix}\widetilde{A} & 0\\ 0 & 0\end{pmatrix}U' \text{ et } B = U\begin{pmatrix}\widetilde{B} & 0\\ 0 & 0\end{pmatrix}U'$$

$$\widetilde{A} \leqslant \widetilde{B} \Rightarrow \widetilde{B}^{-\frac{1}{2}}\widetilde{A}\widetilde{B}^{-\frac{1}{2}} \leqslant \mathrm{id} \text{ (car } A \leqslant B \Rightarrow \forall C, CAC' \leqslant CBC')$$

$$\Rightarrow \widetilde{B}^{\frac{1}{2}}\widetilde{A}^{-1}\widetilde{B}^{\frac{1}{2}} \geqslant \mathrm{id} \text{ (en utilisant la propriété précédente)}$$

$$\Rightarrow \widetilde{A}^{-1} \geqslant \widetilde{B}^{-1} \Rightarrow A^+ \geqslant B^+$$

Annexe C projection orthogonale

Un projecteur M-orthogonal, p, est défini par les deux propriétés :

$$\rightarrow (\mathrm{id} - p)'Mp = 0 \qquad \textbf{(eC.63)}$$

$$\rightarrow p^2 = p \text{ (idempotence)} \qquad \textbf{(eC.64)}$$

$\Rightarrow \mathrm{Ker}(p) = \{p = 0\} = \{\|p\|_M = 0\} = \{p'Mp = 0\} = \{x \,/\, x'p'Mp = 0\} \overset{\text{(eC.63)}}{=} \{x \,/\, x'Mp = 0\} = \mathrm{Im}(p)^{\perp_M}$

– $\mathrm{proj}_{\mathrm{Im}(u_s)}(z) = u_s\beta$ tel que $\langle z - u_s\beta, u_s\rangle = 0$ soit $\langle u_s'z - u_s'u_s\beta, \rangle = 0$ et donc $u_s'u_s\beta = u_s'z$. $\beta = \left(u_s'u_s\right)^- u_s'z$ est une solution, d'où finalement :

$$\mathrm{proj}_{\mathrm{Im}(u_s)} = u_s\left(u_s'u_s\right)^- u_s' \qquad \textbf{(eC.65)}$$

– AA^+ est le projecteur orthogonal sur $\mathrm{Im}(A)$ car $\mathrm{Im}(AA^+) = \mathrm{Im}(A)$, $AA^+AA^+ = AA^+$ et $(AA^+)' = AA^+$.

– u_s est constante sur l'ensemble $\mathrm{argmin}_\beta \|z - u_s\beta\|_c$, de valeur $\mathrm{proj}^c_{\mathrm{Im}(u_s)}(z)$:

si $\widehat{\beta}, \widetilde{\beta} \in \underset{\beta}{\mathrm{argmin}} \|z - u_s\beta\|_c$ alors :

$$\left\|z - u_s\widehat{\beta}\right\|_c^2 = \left\|z - u_s\widetilde{\beta}\right\|_c^2 = \left\|z - u_s\widehat{\beta}\right\|_c^2 + \left\|u_s\widehat{\beta} - u_s\widetilde{\beta}\right\|_c^2 \Rightarrow \left\|u_s\widehat{\beta} - u_s\widetilde{\beta}\right\|_c^2 = 0$$

Annexe D dérivabilité infinie de $\delta(c)$

• $\mathrm{proj}^c_{\mathrm{Im}(u_s)}(z)$ est C^∞ sur $\{c > 0\}$:

Soit $\widetilde{u}_s$ une sous-matrice de u_s de plein rang colonne et surjective sur $\mathrm{Im}(u_s)$. Alors $\mathrm{proj}^c_{\mathrm{Im}(u_s)}(z) = \widetilde{u}_s\widetilde{\widetilde{\beta}}$ avec $\left\langle z - \widetilde{u}_s\widetilde{\widetilde{\beta}}, \widetilde{u}_s\right\rangle_c = 0$ et $\varphi(c) = \left\|z - \widetilde{u}_s\widetilde{\widetilde{\beta}}\right\|_c^2$

– Une démonstration de la dérivabilité infinie ($\delta \in C^\infty(\mathbb{R}_+^*)$) par le théorème d'inversion locale utilise le fait que la fonction $\Phi\left(c, \widetilde{\beta}\right) = \left\langle z - \widetilde{u}_s\widetilde{\beta}, \widetilde{u}_s\right\rangle_c = \left\langle z - \widetilde{u}_s\widetilde{\beta}, c\widetilde{u}_s\right\rangle$ est C^∞ (comme produit scalaire de deux fonctions linéaires), et comme $\widetilde{u}_s$ est de plein rang colonne et $c > 0$, $\dfrac{\partial\Phi}{\partial\widetilde{\beta}}(c, \widetilde{\beta}) = -\langle\widetilde{u}_s, \widetilde{u}_s\rangle_c$ est inversible pour tout $c > 0$. Donc $\Psi\left(c, \widetilde{\beta}\right) = \left(c, \Phi\left(c, \widetilde{\beta}\right)\right)$ est un difféomorphisme local, et $\widetilde{\widetilde{\beta}}(c) = \mathrm{proj}_{\mathrm{Vect}(\beta)}\Psi^{-1}(c, 0)$ est C^∞.

– Une démonstration plus spécifique :

$\widehat{\widetilde{\beta}} = \left[\widetilde{u}^s * c\widetilde{u}_s\right]^{-1} \widetilde{u}^s * cz$ est le produit matriciel de deux fonctions C^∞, parce que la fonction $A \longmapsto A^{-1}$ est C^∞ sur l'ensemble des matrices inversibles. (Si A est inversible et si $\|H\| < 1 / \|A^{-1}\|$ alors $(A+H)^{-1} = A^{-1}(\mathrm{id} + A^{-1}H)^{-1} = A^{-1}\sum(-A^{-1}H)^{\mathbb{N}}$).

→ Il s'ensuit que $\mathrm{proj}^c_{\mathrm{Im}(u_s)}(z) = \widetilde{u}_s\widehat{\widetilde{\beta}} = u_s\widehat{\beta}_c$ est C^∞ sur $\{c>0\}$ et par suite $\varphi(c) = \left\langle \left[z - u_s\widehat{\beta}_c\right]^2, c\right\rangle = \left\langle z - u_s\widehat{\beta}_c, cz\right\rangle$ aussi (comme produit scalaire de deux fonctions C^∞).

- $\left(\sum cuu'\right)^+$ est dérivable (et C^∞) en c sur $\mathbb{R}^*_+(s)$:

↳ Comme $\|z\|_c^2 - \varphi_z(c) = \left\|u_s\widehat{\beta}_c(z)\right\|_c^2$ est dérivable (car différence de deux dérivables) :

$$\begin{aligned}\left\|u_s\widehat{\beta}_c(z)\right\|_c^2 &= \sum_j c_j\left[u_j'\left(\sum cuu'\right)^-\sum czu\right]^2\\ &= \left(\sum czu'\right)\left(\sum cuu'\right)^-\left(\sum cuu'\right)\left(\sum cuu'\right)^-\left(\sum czu\right)\\ &= \left(\sum czu'\right)\left(\sum cuu'\right)^-\left(\sum czu\right)\\ &= \sum_{i,j} c_ic_jz_iz_ju_i'\left(\sum cuu'\right)^- u_j\end{aligned}$$

est dérivable en $c>0$ pour tout z.

⇔ $\forall i,j\in s, u_i'\left(\sum cuu'\right)^- u_j$ est dérivable, donc $\sum uu'\left(\sum cuu'\right)^-\sum uu'$ est dérivable.

↳ Il existe une (unique) matrice Φ inversible telle que :

$$\begin{cases}\sum uu'\Phi = \mathrm{id} = \Phi\sum uu' \text{ sur } \mathrm{Im}\left(\sum uu'\right)\\ \Phi = \mathrm{id} \text{ sur } \mathrm{Ker}\left(\sum uu'\right)\end{cases} \qquad \textbf{(eD.66)}$$

Ceci découle de ce que comme $\sum uu'$ est symétrique, $E = \mathrm{Im}\left(\sum uu'\right)\oplus\mathrm{Ker}\left(\sum uu'\right)$ et $\sum uu'$ est bijective sur $\mathrm{Im}\left(\sum uu'\right)$.

↳ La matrice Φ est symétrique, du fait de l'unicité de la solution de (**eD.66**).

↳ Comme Φ ne dépend pas de c, il en découle que $\Gamma(c) = \Phi'\sum uu'\left(\sum cuu'\right)^-\sum uu'\Phi$ est dérivable en c.

▷ sur $\mathrm{Ker}\left(\sum uu'\right)$, $\Gamma(c) = 0$ car $\sum uu'\Phi\left[\mathrm{Ker}\left(\sum uu'\right)\right] = \sum uu'\mathrm{Ker}\left(\sum uu'\right) = \{0\}$ donc $\Gamma(c) = \left(\sum cuu'\right)^+$ sur ce sous-espace vectoriel (car $\mathrm{Ker}\left(\sum uu'\right) = \mathrm{Ker}\left(\left(\sum uu'\right)^+\right)$).

▷ Pour $x\in\mathrm{Im}\left(\sum uu'\right)$:

$$\begin{aligned}\Gamma(c)\,x &= \Phi'\sum uu'\left(\sum cuu'\right)^+\sum uu'\Phi x\\ &= \Phi\sum uu'\left(\sum cuu'\right)^+ x \text{ (car } \textstyle\sum uu'\Phi = \mathrm{id} \text{ sur } \mathrm{Im}\left(\sum uu'\right))\\ &= \left(\sum cuu'\right)^+ x\end{aligned}$$

(car $\mathrm{Im}\left[\left(\sum cuu'\right)^+\right] = \mathrm{Im}\left(\sum cuu'\right) = \mathrm{Im}\left(\sum uu'\right)$ et $\Phi\sum uu' = \mathrm{id}$ sur $\mathrm{Im}\left(\sum uu'\right)$)

⇒ $\Gamma(c) = \left(\sum cuu'\right)^+$, et cette dernière fonction de c est donc également dérivable.

- une formulation de la dérivée seconde de φ :

$$\ddot{\varphi}(c) = \frac{d}{dc}\left\|z - u_s\widehat{\beta}_c\right\|_{dc} = 2\left\langle z - u_s\widehat{\beta}_c, -d(u_s\widehat{\beta}_c)\right\rangle_{dc}$$

$$\begin{aligned}\left\langle u_s\left(\widehat{\beta}_c - \widehat{\beta}_{c+h}\right), u_s\right\rangle_{c+h} &= \left\langle z - u_s\widehat{\beta}_{c+h}, u_s\right\rangle_{c+h} - \left\langle z - u_s\widehat{\beta}_c, u_s\right\rangle_{c+h}\\ &= -\left\langle z - u_s\widehat{\beta}_c, u_s\right\rangle_{c+h} \text{ (car } \widehat{\beta}_{c+h} \text{ vérifie l'équation normale pour } \langle\rangle_{c+h})\\ &= -\left\langle z - u_s\widehat{\beta}_c, u_s\right\rangle_h \text{ (car } \widehat{\beta}_c \text{ vérifie l'équation normale pour } \langle\rangle_c) \text{ d'où :}\\ \left\langle u_s\left(\widehat{\beta}_{c+h} - \widehat{\beta}_c\right), u_s\right\rangle_{c+h} &= \left\langle z - u_s\widehat{\beta}_c, u_s\right\rangle_h \qquad \textbf{(eD.67)}\end{aligned}$$

(**eD.67**) permet de formuler la dérivée du projecteur ainsi :

$$\left\langle d(u_s\widehat{\beta}_c), u_s\right\rangle_c = \left\langle z - u_s\widehat{\beta}_c, u_s\right\rangle_{dc} \qquad \textbf{(eD.68)}$$

(Cette équation détermine effectivement $d(u_s\widehat{\beta}_c)$: $d(u_s\widehat{\beta}_c)h = u_s\left(u^s * cu_s\right)^- u^s * h\left[z - u\widehat{\beta}_c\right]_s$).

Il découle de (**eD.68**) :

$$\begin{aligned}
\left\langle z-u_s\widehat{\beta}_c, u_s\widehat{\beta}_{c+h}\right\rangle_{dc} &= \left\langle d(u_s\widehat{\beta}_c), u_s\widehat{\beta}_{c+h}\right\rangle_c \\
\Rightarrow \left\langle z-u_s\widehat{\beta}_c, d(u_s\widehat{\beta}_c)\right\rangle_{dc} &= \left\langle d(u_s\widehat{\beta}_c), d(u_s\widehat{\beta}_c)\right\rangle_c \text{ d'où :} \\
\ddot{\varphi}(c) &= -2\left\langle z-u_s\widehat{\beta}_c, d(u_s\widehat{\beta}_c)\right\rangle_{dc} = -2\left\langle d(u_s\widehat{\beta}_c), d(u_s\widehat{\beta}_c)\right\rangle_c \\
\ddot{\varphi}(c)\,hh &= -2\left[z-u\widehat{\beta}_c\right]^s hu_s\left(u^s * cu_s\right)^- u^s h\left[z-u\widehat{\beta}_c\right]_s \\
&= -2\left\|\mathrm{proj}^c_{\mathrm{Im}(u_s)}\left\{\frac{h}{c}\left[z-u\widehat{\beta}_c\right]_s\right\}\right\|_c^2
\end{aligned}$$

Le signe de cette dernière formule permet de retrouver la concavité de φ.

Annexe E rang de $\dot{\delta}(c)$ et contraction de $\mathrm{id}-\dot{\delta}(c)$

E.1 norme d'une fonction linéaire

Si $f \in \mathfrak{L}(E,F)$ [38], E de dimension finie de base e_s et F normé alors :

$$\|f\| \overset{\text{déf}}{=} \sup_{h\neq 0} \|fh\| / \|h\| \qquad \textbf{(eE.69)}$$

→ $\|f\| < +\infty$:

↳ première approche :

$\sup_{h\neq 0} \|fh\| / \|h\| = \sup_{\|h\|=1} \|fh\| < \infty$ car f est continue et $\{\|\| = 1\}$ compact parce que $\dim(E) < \infty$

↳ deuxième approche (sans admettre la continuité de f) :

$$\begin{aligned}
\sup_{\|x\|=1} \|f(x)\| &\leqslant \sup_{\|x\|=1} \sum |x_i| \|f(e_i)\| \\
&\leqslant \nu(x) \sum \|f(e_i)\| \text{ (où } \nu(x) = \sup(x_i, i \in s)) \\
&\leqslant \eta \|x\| \text{ avec } \eta < +\infty \text{ (en admettant l'équivalence des normes)} \\
&\Rightarrow \sup_{\|x\|=1} \|f(x)\| < +\infty \\
&\Rightarrow f \text{ continue : } \|f(y)-f(x)\| = \|f(y-x)\| \leqslant \|f\| \|y-x\|
\end{aligned}$$

– $\forall h, \left|h'Ah\right| \leqslant \|h\| \|Ah\| \leqslant \|A\| \|h\|^2$

• si $f \in \mathfrak{L}(\mathbb{R}_s)$ muni de la distance euclidienne canonique alors $\|f\| = \sup\sqrt{\Lambda(f'f)}$ [39]. En effet, soit v une base orthonormée de vecteurs propres de la matrice symétrique positive $f'f$. Pour $\|x\| = 1$, $x'f'fx = \sum \lambda x^2 \leqslant \sup[\lambda] \sum x^2 = \sup \Lambda(f'f) \Rightarrow \|f'f\| = \sup \Lambda(f'f)$.

→ $\|f\|^2 = \sup_{\|h\|=1} (\|fh\|)^2 = \sup_{\|h\|=1} h'f'fh = \|f'f\|$ donc :

$$\|f\|^2 = \|f'f\| = \sup \Lambda(f'f) \qquad \textbf{(eE.70)}$$

• Les espaces propres de ff' de valeurs propres non nulles sont isomorphes à ceux de $f'f$: Soit $\lambda \in \Lambda(f'f) \setminus 0$

→ f est injective sur $E_\lambda(f'f)$, vu que $E_\lambda(f'f) \cap \mathrm{Ker}(f) = 0$ pour $\lambda \neq 0$

→ $f\left[E_\lambda(f'f)\right] \subset E_\lambda(ff')$: pour $h \in E_\lambda(f'f) \setminus 0$, $f(h) \neq 0$ et $ff'f(h) = \lambda f(h)$ donc $f(h) \in E_\lambda(ff')$

→ par symétrie, $f'E_\lambda(ff') \subset E_\lambda(f'f)$

→ $\boxed{f\left[E_\lambda(f'f)\right] = E_\lambda(ff')}$ car f' est injective sur $E_\lambda(ff')$ (par symétrie de l'injectivité de f sur $E_\lambda(f'f)$), donc $\dim\left[E_\lambda(ff')\right] \leqslant \dim\left[E_\lambda(f'f)\right]$ et symétriquement $\dim\left[E_\lambda(f'f)\right] \leqslant \dim\left[E_\lambda(ff')\right]$. Par suite $\dim\left[E_\lambda(f'f)\right] = \dim\left[E_\lambda(ff')\right]$.

---

38. $\mathfrak{L}(E,F)$ est l'ensemble des applications linéaires entre les espaces vectoriels E et F.
39. Λ désigne l'ensemble des valeurs propres et VecP celui des vecteurs propres.

$\Rightarrow \quad \|f\| = \sup\sqrt{\Lambda(f'f)} = \sup\sqrt{\Lambda(ff')} = \|f'\|$

Cette propriété peut se retrouver par l'expression :

$$\|f\| = \sup_{k,h\neq 0} \frac{|\langle k, fh\rangle|}{\|k\|\,\|h\|}$$

- si f est symétrique alors $\|f\| = \sup|\Lambda(f)|$: si v est une base orthonormée de $\mathrm{VecP}(f)$ alors $(\|fh\|)^2 = \sum \lambda^2 h^2 \leqslant \sup\Lambda(f)^2$
- si f (une fonction non linéaire) est dérivable sur $[x,y]$ alors

$$\|f(y) - f(x)\| \leqslant \sup_{z\in[x,y]} \left\|\dot f(z)(y-x)\right\| \leqslant \sup_{[x,y]} \left\|\dot f\right\| \|y-x\|$$

E.2 contractions

- Pour $f \in \mathcal{L}(E)$, f contractante $\overset{\text{déf}}{\Leftrightarrow} \|f\| \leqslant 1 \Leftrightarrow \forall h \in E, \|f(h)\| \leqslant \|h\|$

Dans $\mathbb{R}_s$ muni de la distance euclidienne canonique :

- f contractante $\Leftrightarrow \sup\Lambda\left(f'f\right) \leqslant 1$ (vu (**eE.70**))
- $\mathrm{id} - f$ contractante $\Leftrightarrow f' + f \geqslant f'f \Leftrightarrow f' + f \geqslant ff'$:

$$\begin{aligned} &\forall h, \|(\mathrm{id}-f)h\| \leqslant \|h\| \\ \Leftrightarrow\ &(\mathrm{id}-f)'(\mathrm{id}-f) \leqslant \mathrm{id} \\ \Leftrightarrow\ &f'f - f' - f \leqslant 0 \end{aligned}$$

  - Comme $\left\|\mathrm{id} - f'\right\| = \|\mathrm{id} - f\|$, $\mathrm{id} - f$ contractante $\Leftrightarrow \mathrm{id} - f'$ contractante
- $\mathrm{id} - f$ contractante $\Leftrightarrow \forall h, h'(id-f)'(id-f)h \leqslant h'h \Leftrightarrow \forall h, h'f'fh - 2h'fh \leqslant 0$
  - → $\mathrm{id} - f$ contractante $\Leftrightarrow \forall h, \|f(h)\|^2 \leqslant 2h'f(h)$
    - ⇝ donc une condition nécessaire est : $\forall h, h'f(h) \geqslant 0$
- si $\forall h, h'fh \geqslant 0$ alors pour $\lambda > 0$ assez petit, $\mathrm{id} - \lambda f$ est contractante : (cas $f \neq 0$)
  - → $\min_{\begin{cases} h \in \mathrm{Ker}(f)^\perp \\ \|h\| = 1 \end{cases}} h'fh = \eta > 0$ car fonction continue strictement positive sur un compact
  - → pour $\lambda > 0, h \in \mathrm{Ker}(f)^\perp$ et $\|h\| = 1$:

$$(\lambda)^2 h'f'fh - 2\lambda h'fh \leqslant (\lambda)^2 \left\|f'f\right\| - 2\lambda\eta = \lambda\left[\lambda\left\|f'f\right\| - 2\eta\right] < 0 \text{ pour } \lambda < \frac{2\eta}{\left\|f'f\right\|}$$

- si $h \in \mathrm{Ker}(f) \setminus 0$ alors $\|(id-f)h\| = \|h\|$ donc $\|\mathrm{id} - f\|^2 = \sup\Lambda\left\{(id-f)'(id-f)\right\} \geqslant 1$, si $\mathrm{Ker}(f) \neq 0$

E.3 application à $\mathrm{id} - \dot\delta(c)$

→ Pour la convergence de l'algorithme récursif, il est utile d'étudier la condition $\left\|\mathrm{id} - \dot\delta(c)\right\| \leqslant 1$. Ça ne parait pas facile à déterminer sur la matrice de cette application linéaire :

$$\begin{aligned} \mathrm{id} - \dot\delta(c) &= \mathrm{id} - \left(\mathrm{id} - (cu)_s\left(\sum cuu'\right)^- u^s\right)^{*2} \\ &= -2\mathrm{diag}\left[(cu)_s\left(\sum cuu'\right)^- u^s\right] + \left((cu)_s\left(\sum cuu'\right)^- u^s\right)^{*2} \end{aligned}$$

- une formule explicite de $\left\|\mathrm{id} - \dot\delta(c)\right\|$ pour la norme euclidienne :

$$\left\|\mathrm{id} - \dot\delta(c)\right\|^2 = \sup_{h\neq 0} \frac{\left\|h_s - \left\|\mathbb{1}_s - u\widehat{\beta}_c(\mathbb{1}_s)\right\|_h^2\right\|^2}{\|h\|^2} = \sup_{h\neq 0} \frac{\sum_i \left\{h_i - \sum_{j\in s}\left\{\mathbb{1}_i^j - u'_j\widehat{\beta}_c(\mathbb{1}_i)\right\}^2 h_j\right\}^2}{\sum h^2} \qquad \textbf{(eE.71)}$$

- La première expression permet de vérifier que $\dfrac{\left\|\left(\mathrm{id} - \dot\delta(c)\right)c\right\|^2}{\|c\|^2} \leqslant 1$:

$$c_i - \left\|\mathbb{1}_i - u_s\widehat{\beta}_c(\mathbb{1}_i)\right\|_c^2 = c_i - \left(c_i - \left\|u_s\widehat{\beta}_c(\mathbb{1}_i)\right\|_c^2\right) = \left\|u_s\widehat{\beta}_c(\mathbb{1}_i)\right\|_c^2 \in \left[0, \|\mathbb{1}_i\|_c^2\right] = [0, c_i]$$

– pour la norme du sup :

$$\begin{aligned}\left\|\mathrm{id}-\dot{\delta}(c)\right\|_\infty &= \sup_i \sup_{\|h\|_\infty=1} \left|h_i - \left\|\mathbb{1}_i - u_s\widehat{\beta}_i(c)\right\|_h^2\right| \\ &= \sup_i 1 - \left\{1-u_i'\widehat{\beta}_i(c)\right\}^2 + \sum_{j\neq i}\left\{u_j'\widehat{\beta}_i(c)\right\}^2 \\ &= 1 + \left\|\mathbb{1}_i - u_s\widehat{\beta}_i(c)\right\|_1^2 - 2\left[\frac{\delta_i(c)}{c_i}\right]^2\end{aligned}$$

Donc $\mathrm{id}-\dot{\delta}(c)$ est contractante pour cette norme si $\forall i \in s, \left[\frac{\delta_i(c)}{c_i}\right]^2 > \frac{1}{2}\left\|\mathbb{1}_i - u_s\widehat{\beta}_i(c)\right\|_1^2 = \frac{1}{2}\delta_i(1)$

⇒ Pour c constante, $\mathrm{id}-\dot{\delta}(c)$ est contractante ssi $\forall i \in s, \delta_i(1) > \sqrt{2}/2$.

E.3.1 expression de $\dot{\delta}(c)$ en fonction d'un projecteur

Dans la suite, $q = q_c$ désigne le projecteur orthogonal (pour le produit scalaire) canonique sur $\mathrm{Im}\left[\sqrt{c}u_s\right]^\perp$, q^{*2} est la matrice composée des carrés des coefficients de q, les matrices diagonales sont confondues avec les vecteurs correspondants, et la commutativité du produit de matrices diagonales est utilisée.

- expression de la dérivée en fonction de la projection orthogonale canonique :

$$\begin{aligned}\mathrm{proj}^{c^\perp}_{\mathrm{Im}(u_s)} &= \mathrm{id} - u_s\left(\sum cuu'\right)^- u^s * c = \left[\sqrt{c}\right]^{-1} q * \sqrt{c} \\ \dot{\delta}(c)_{i,j} &= \left[\mathbb{1}_i^j - u_j'\widehat{\beta}_c(\mathbb{1}_i)\right]^2 = \left[\mathrm{proj}^{c^\perp}_{\mathrm{Im}(u_s)}[\mathbb{1}_i]_j\right]^2 \\ &= \left\{\left[\sqrt{c_i}\right]^{-1} q_{i,j}\sqrt{c_j}\right\}^2 = (c_i)^{-1} q_{i,j}^2 c_j \\ \Rightarrow \dot{\delta}(c) &= c^{-1}q^{*2} * c \qquad \textbf{(eE.72)}\end{aligned}$$

- soit v_K une base orthonormée (BON) de $\mathrm{Im}(q)$ alors $q = \sum_K vv'$ et :

$$\begin{aligned}h'c^{-1}q^{*2} * ch &= \sum_{i,j} h_i \frac{1}{c_i}\sum_{k,l} v_k^i v_k^j v_l^i v_l^j c_j h_j \\ &= \sum_{k,l}\left\langle h, \frac{v_k v_l}{c}\right\rangle \langle h, cv_k v_l\rangle\end{aligned}$$

→ Cette approche permet de prouver (avec $c = 1$) que la matrice des coefficients au carré d'une somme de vecteurs colonnes multipliés par leurs transposés, c'est-à-dire d'une matrice symétrique positive quelconque, est positive, puisqu'il apparait ainsi une somme de carrés [40].

→ En outre, elle prouve qu'il n'est pas toujours vrai que $c^{-1}q^{*2} * c \geqslant 0$. En effet, si $\mathrm{Im}(q) = \mathbb{R}v$ et si $(\frac{v^2}{c}, cv^2)$ ne sont pas colinéaires, alors il existe h tel que $\left\langle h, \frac{v^2}{c}\right\rangle \langle h, cv^2\rangle < 0$. Or $(\frac{v^2}{c}, cv^2)$ ne sont pas colinéaires s'il existe $i \neq j$ tel que $v_i^2 v_j^2\left(\frac{c_j}{c_i} - \frac{c_i}{c_j}\right) \neq 0$, c'est-à-dire deux coordonnées de v non nulles pour lesquelles la pondération n'est pas identique. C'est possible même si c est très proche d'une constante, sauf si v a une seule coordonnée non nulle.

---

40. Plus généralement, si M et $\widetilde{M}$ sont des matrices symétriques positives de même dimension alors $M * \widetilde{M} \geqslant 0$:

$$h'M * \widetilde{M}h = \sum_{i,j,k,l} h_i\sqrt{M}_{i,k}\sqrt{M}_{j,k}\sqrt{\widetilde{M}}_{i,l}\sqrt{\widetilde{M}}_{j,l}h_j = \sum_{k,l}\left\langle h, \sqrt{M}^k\sqrt{\widetilde{M}}^l\right\rangle^2$$

- majoration de q^{*2} :

$$\sum_{i,j}(q_{i,j})^2 h_i h_j = \sum_{i,j,k,l} h_i h_j v_k^i v_k^j v_l^i v_l^j = \sum_{k,l}\langle h, v_k v_l\rangle^2$$
$$= \sum_{k,l}\langle hv_k, v_l\rangle^2 = \sum_{k\in K}(hv_k)'\sum_{l\in K} v_l v_l'(hv_k) = \sum_{k\in K}(hv_k)'q(hv_k) = \sum_{k\in K}(hv_k)'q^2(hv_k)$$
$$= \sum_K \|q(hv)\|^2 \leqslant \sum_K \|hv\|^2 = \sum_{k\in K}\sum_i \left(h_i v_k^i\right)^2$$
$$\leqslant \sum_i h_i^2 \sum_{k\in K}\left(v_k^i\right)^2 = \sum_i h_i^2 q_i \qquad \text{d'où un majorant :}$$
$$h'q^{*2}h \leqslant h'\mathrm{diag}\,(q)\,h$$

⇒ Les deux résultats précédents mènent à un encadrement de la matrice des carrés des coefficients d'un projecteur orthogonal canonique :

$$0 \leqslant q^{*2} \leqslant \mathrm{diag}\,(q) \leqslant \mathrm{id} \qquad \textbf{(eE.73)}$$

E.3.2 rang de $\dot{\delta}(c)$

- Le rang de $\dot{\delta}(c)$ est constant sur $\{c>0\}$:

$$\mathrm{rang}\left[\dot{\delta}(c)\right] = \mathrm{rang}\left[(q_1)^{*2}\right] = \dim\left\{\mathrm{Vect}\left[h^2, h\in \mathrm{Im}\,(x_s)^\perp\right]\right\} \qquad \textbf{(eE.74)}$$

↳ Comme la matrice diagonale c est bijective, $\mathrm{rang}\left(c^{-1}q^{*2}*c\right) = \mathrm{rang}\left(q^{*2}\right)$

$q^{*2}h = 0 \Leftrightarrow h'q^{*2}h = 0$ (car q^{*2} est symétrique et positive vu (**eE.73**))

$h'q^{*2}h = \sum_{k,l\in K}\langle h, v_k v_l\rangle^2 = 0$ (où v_K est toujours une BON de $\mathrm{Im}\,(q)$)

$\Leftrightarrow \forall k,l\in K, \langle h, v_k v_l\rangle = 0 \Leftrightarrow \forall x,y\in \mathrm{Im}\,(q), \langle h, xy\rangle = 0$ (car v_K base de $\mathrm{Im}\,(q)$)

$\Rightarrow \mathrm{Ker}\left(q^{*2}\right) = \{xy, x, y\in \mathrm{Im}\,(q)\}^\perp$

$\Rightarrow \mathrm{Im}\left(q^{*2}\right) = \mathrm{Vect}\left[xy, x, y\in \mathrm{Im}\,(q)\right]$ (car q^{*2} est symétrique donc $\mathrm{Im}\left(q^{*2}\right) = \mathrm{Ker}\left(q^{*2}\right)^\perp$)

$\Rightarrow \mathrm{Im}\left(q^{*2}\right) = \mathrm{Vect}\left[x^2, x\in \mathrm{Im}\,(q)\right]$ car $xy = \left((x+y)^2 - x^2 - y^2\right)/2$

$\Rightarrow \mathrm{rang}\left(q^{*2}\right) = \dim\left\{\mathrm{Vect}\left[h^2, h\in \mathrm{Im}\,(q)\right]\right\}$ où $\mathrm{Im}\,(q) = \mathrm{Im}\left[\sqrt{c}u_s\right]^\perp$

↳ $(q_c)^{*2}$ et $(q_1)^{*2}$ ont même rang :

$$h\in \mathrm{Im}\left[\sqrt{c}u_s\right]^\perp \Leftrightarrow u_s'*\sqrt{c}h = 0$$
$$\Leftrightarrow \sqrt{c}h \in \mathrm{Im}\,(u_s)^\perp$$
$$\left\{h^2, h\in \mathrm{Im}\left[\sqrt{c}u_s\right]^\perp\right\} = \left\{\left[\sqrt{c}h\right]^2, h\in \mathrm{Im}\,(u_s)^\perp\right\} = \left\{ch^2, h\in \mathrm{Im}\,(u_s)^\perp\right\}$$
$$= c\left\{h^2, h\in \mathrm{Im}\,(u_s)^\perp\right\}$$
$$\Rightarrow \mathrm{Vect}\left\{h^2, h\in \mathrm{Im}\left[\sqrt{c}u_s\right]^\perp\right\} = c\mathrm{Vect}\left[h^2, h\in \mathrm{Im}\,(u_s)^\perp\right]$$
$$\Rightarrow \mathrm{Im}\left[(q_c)^{*2}\right] = c\mathrm{Im}\left[(q_1)^{*2}\right] \text{ qui sont de mêmes dimensions car } c>0$$

↳ $\dim\left\{\mathrm{Vect}\left[h^2, h\in \mathrm{Im}\,(u_s)^\perp\right]\right\} = \dim\left\{\mathrm{Vect}\left[h^2, h\in \mathrm{Im}\,(x_s)^\perp\right]\right\}$:

$$\mathrm{Im}\,(u_s)^\perp = \left\{h \,/\, h'\frac{x}{\pi}(s) = 0\right\} = \left\{\pi_s\frac{h}{\pi} \,/\, \frac{h}{\pi}'x_s = 0\right\} = \pi_s\mathrm{Im}\,(x_s)^\perp$$
$$\Rightarrow \left[\mathrm{Im}\,(u_s)^\perp\right]^{*2} = (\pi_s)^{*2}\left[\mathrm{Im}\,(x_s)^\perp\right]^{*2}$$
$$\Rightarrow \mathrm{Vect}\left\{\left[\mathrm{Im}\,(u_s)^\perp\right]^{*2}\right\} = (\pi_s)^{*2}\,\mathrm{Vect}\left\{\left[\mathrm{Im}\,(x_s)^\perp\right]^{*2}\right\}$$

- détermination du rang de $\dot{\delta}(c) = \dim\left\{\mathrm{Vect}\left[h^2, h\in \mathrm{Im}\,(u_s)^\perp\right]\right\}$:
  - Il existe deux sous-ensembles K d'indices colonnes et $s_K \subset s$ de cardinaux égaux au rang de u_s et tels que $u^K_{s_K}$ soit inversible et $\exists B \,/\, u_s = u_s^K B$. Dans la suite, Δs désigne le complémentaire de s_K dans s.

– Il est possible de paramétrer les éléments de $\mathrm{Im}\,(u_s)^\perp$ par leurs coordonnées sur le sous-ensemble des indices Δs : soit $x \in \mathrm{Im}\,(u_s)^\perp$

$$x' u_s = 0 \Leftrightarrow x' u_s^K = 0 \text{ car } u_s = u_s^K B$$
$$\Leftrightarrow x'_{s_K} u_{s_K}^K + x'_{\Delta s} u_{\Delta s}^K = 0$$
$$\Leftrightarrow x_{s_K} = -\left[u_K^{s_K}\right]^{-1} u_K^{\Delta s} x_{\Delta s}$$
$$\Rightarrow x = \begin{pmatrix} \mathrm{id}_{\Delta s} \\ A \end{pmatrix} x_{\Delta s} \text{ où } A = -\left[u_K^{s_K}\right]^{-1} u_K^{\Delta s}$$

– détermination de l'orthogonal de $\mathrm{Vect}\left[x^2, x \in \mathrm{Im}\,(u_s)^\perp\right]$:

$$h' x^2 = h'\left[\begin{pmatrix} x_{\Delta s} \\ A x_{\Delta s} \end{pmatrix}^2\right]$$
$$= \sum_{j \in \Delta s} h_j (x_j)^2 + \sum_{i \in s_K} h_i \left[\sum_{j \in \Delta s} A_{i,j} x_j\right]^2$$
$$= \sum_{j \in \Delta s} (x_j)^2 \left(h_j + \sum_{i \in s_K} h_i (A_{i,j})^2\right) + \sum_{i \in s_K} h_i \sum_{j \neq \tilde{j}} A_{i,j} A_{i,\tilde{j}} x_j x_{\tilde{j}}$$

La condition pour que ceci soit nul pour tout $x_{\Delta s}$ est :

$$\begin{cases} \forall j \in \Delta s, h_j = -\sum_{i \in s_K} h_i (A_{i,j})^2 \\ \forall j_1 \neq j_2 \in \Delta s, \sum_{i \in s_K} h_i A_{i,j_1} A_{i,j_2} = 0 \end{cases}$$

$\Rightarrow$ $\dim\left\{\left\{x^2, x \in \mathrm{Im}\,(u_s)^\perp\right\}^\perp\right\} \leqslant \mathrm{rang}\,(u_s)$ $(= |s_K|)$

$\Rightarrow$ $\mathrm{rang}\left[\dot\delta\,(c)\right] \geqslant |s| - \mathrm{rang}\,(u_s)$

$\Rightarrow$ Le rang de $\dot\delta\,(c)$ est déterminé par celui de la famille de (produits de) vecteurs $\left(A^{j_1} A^{j_2}\right)^{j_1 < j_2 \in \Delta s}$.

• Si $\forall j_1 \in \Delta s, u_{\Delta s \neq j_1}^K$ est de rang $= |K|$ alors $\mathrm{rang}\left[\dot\delta\,(c)\right] = |s|$ et $\dot\delta\,(c)$ est inversible. En effet :

→ $\mathrm{rang}\left(A^{\Delta s \neq j_1}\right) = \mathrm{rang}\left[u_K^{\Delta s \neq j_1}\right]$, car la première matrice est le produit de la seconde par une matrice inversible.

→ si $\forall j_1 \in \Delta s, \mathrm{rang}\left(A^{\Delta s \neq j_1}\right) = |K| = |s_K|$ alors $\sum_{i \in s_K} h_i A_{i,j_1} A_i^{\Delta s \neq j_1} = 0 \Rightarrow \sum_{i \in s_K} h_i A_{i,j_1} = 0$

$(\forall j_1 \in \Delta s) \Rightarrow \sum_{i \in s_K} h_i A_i = 0 \Rightarrow h_{s_K} = 0$ (En particulier, pour une seule contrainte d'équilibrage, δ est un difféomorphisme local si toutes les coordonnées de u_s sont non nulles pour tout échantillon.)

E.3.3 conditions pour que $\mathrm{id} - \dot\delta\,(c)$ soit contractante

• si c est constante alors $\mathrm{id} - \dot\delta\,(c) = \mathrm{id} - \dot\delta\,(1)$ est contractante pour la norme euclidienne :

– $\dot\delta\,(c) = c^{-1} q^{*2} * c = q^{*2}$: En effet, comme c est une constante, elle peut être factorisée.

– $q_c = q_1$ est évident car $\mathrm{Im}\left[\sqrt{c} u_s\right] = \mathrm{Im}\,(u_s)$ si c constante > 0

– $\mathrm{id} - \dot\delta\,(c)$ contractante $\Leftrightarrow \forall h, 2h' \dot\delta\,(c)\, h \geqslant \left\|\dot\delta\,(c)\, h\right\|^2 \Leftrightarrow 2\dot\delta\,(c) \geqslant \dot\delta\,(c)' \dot\delta\,(c)$ (par définition) donc :

$$\mathrm{id} - \dot\delta\,(c) \text{ contractante} \Leftrightarrow 2c^{-1} q^{*2} * c \geqslant c q^{*2} * c^{-2} q^{*2} * c$$

→ Si c est constante, la contraction revient donc à : $2q^{*2} \geqslant q^{*2} q^{*2}$. Cette propriété est vérifiée, parce que (vu (**eE.73**)) $q^{*2} \leqslant \mathrm{id} \Rightarrow q^{*2} q^{*2} \leqslant q^{*2}$. Cette implication découle de ce que q^{*2} est symétrique et positive, donc admet une racine carrée.

– Plus précisément, $\mathrm{id} - \dot\delta\,(1)$ est strictement contractante sur $\mathrm{Ker}\left[\dot\delta\,(1)\right]^\perp$: Comme $\dot\delta\,(1) = q^{*2}$, qui est symétrique et positive, si $q^{*2} h \neq 0$ alors $h' q^{*2} h \neq 0$ et donc $h' q^{*2} q^{*2} h \leqslant h' q^{*2} h < 2h' q^{*2} h$.

• si c_0 est constante alors il existe un voisinage $\mathcal{V}(c_0)$ tel que $\mathrm{id} - \dot\delta\,(c)$ soit contractante 'sur la majeure partie de E' (en un sens limité apparaissant ci-dessous) :

– Comme $\mathrm{id} - \dot{\delta}(1)$ est strictement contractante sur $\mathrm{Ker}\left[\dot{\delta}(1)\right]^{\perp}$ et que $\left\{h \in \mathrm{Ker}\left[\dot{\delta}(1)\right]^{\perp}, \|h\| = 1\right\}$ est compact :

$$\sup_{\substack{h \in \mathrm{Ker}\left[\dot{\delta}(1)\right]^{\perp} \\ \|h\| = 1}} \left\|(\mathrm{id} - \dot{\delta}(1))h\right\| = \eta < 1$$

– Par la continuité de la fonction $\mathrm{id} - \dot{\delta}(c)$ sur $\{c > 0\}$, pour $\eta < \nu < 1$ il existe $\mathcal{V}(c_0)$ tel que :

$$\sup_{\substack{h \in \mathrm{Ker}\left[\dot{\delta}(1)\right]^{\perp} \\ \|h\| = 1}} \left\|(\mathrm{id} - \dot{\delta}(c))h\right\| \leqslant \nu < 1$$

– $\mathrm{Ker}\left[\dot{\delta}(1)\right]^{\perp} \cap \mathrm{Ker}\left[\dot{\delta}(c)\right] = 0$, puisque pour $h \neq 0$ dans $\mathrm{Ker}\left[\dot{\delta}(1)\right]^{\perp}$, $\left\|(\mathrm{id} - \dot{\delta}(c))h\right\| \leqslant \nu \|h\|$, donc $\dot{\delta}(c)\, h = 0$ impossible.

– Comme $\dim\left\{\mathrm{Ker}\left[\dot{\delta}(c)\right]\right\} = \dim\left\{\mathrm{Ker}\left[\dot{\delta}(1)\right]\right\}$, $E = \mathrm{Ker}\left[\dot{\delta}(1)\right]^{\perp} \oplus \mathrm{Ker}\left[\dot{\delta}(c)\right]$.

• En général, la condition $\dot{\delta}(c) \geqslant 0$ (nécessaire pour une contraction) a peu de chances d'être vérifiée (sur tout l'espace E) lorsque c n'est pas constante et $\dot{\delta}(c)$ n'est pas inversible ($\Rightarrow \mathrm{Ker}\left(q^{*2}\right) \neq 0$) :

$$h'\dot{\delta}(c)\, h = h'c^{-1}q^{*2} * ch$$

$$= (k_1 + k_2)' c^{-2} q^{*2} k_1 \text{ où } ch = k_1 + k_2 \text{ et } k_1 \in \mathrm{Im}\left(q^{*2}\right), k_2 \in \mathrm{Im}\left(q^{*2}\right)^{\perp}$$

$$\forall h, h'\dot{\delta}(c)\, h \geqslant 0 \Leftrightarrow \forall k_1 \in \mathrm{Im}\left(q^{*2}\right), k_2 \in \mathrm{Im}\left(q^{*2}\right)^{\perp}, k_1' c^{-2} q^{*2} k_1 + k_2' c^{-2} q^{*2} k_1 \geqslant 0$$

$$\Rightarrow \forall k_1 \in \mathrm{Im}\left(q^{*2}\right), k_2 \in \mathrm{Im}\left(q^{*2}\right)^{\perp}, k_2' c^{-2} q^{*2} k_1 = 0 \text{ (sinon ce terme pourrait être dilaté vers } -\infty \text{ via } k_2)$$

$$\Rightarrow \forall k_1 \in \mathrm{Im}\left(q^{*2}\right), k_2 \in \mathrm{Im}\left(q^{*2}\right)^{\perp}, k_2' c^{-2} k_1 = 0 \text{ (parce que } q^{*2}\mathrm{Im}\left(q^{*2}\right) = \mathrm{Im}\left(q^{*2}\right))$$

Cette dernière condition de préservation de l'orthogonalité est très forte. Elle est vérifiée ssi $\mathrm{Im}\left(q^{*2}\right)$ est stable par c^{-2}. En effet, si f est une application linéaire et p un projecteur orthogonal canonique, f conserve l'orthogonalité entre $\mathrm{Im}(p)$ et $\mathrm{Im}(p)^{\perp}$ ssi $\forall x, y, [f(\mathrm{id} - p)x]' fpy = 0 \Leftrightarrow (\mathrm{id} - p) f' fp = 0$. Si $u \in \mathrm{Im}(p)$ et $f'fu \notin \mathrm{Im}(p)$ alors $(\mathrm{id} - p) f' fu \neq 0$ donc $(\mathrm{id} - p) f' fpu \neq 0$. Donc $[f\mathrm{Im}(p)] \perp [f\mathrm{Im}(\mathrm{id} - p)] \Leftrightarrow (\mathrm{id} - p) f' fp = 0 \Leftrightarrow f'f\mathrm{Im}(p) \subset \mathrm{Im}(p)$.

→ Pour expliquer l'efficacité de la minimisation par l'algorithme récursif , il est sans doute nécessaire de recourir à une propriété moins forte que la contraction 'pure'.

• En conclusion de cette étude théorique de la contraction de $\mathrm{id} - \dot{\delta}(c)$, celle-ci est contractante lorsque c est constante et 'généralement contractante' pour c assez proche d'une constante. Il n'a pas été possible de mettre en évidence plus clairement la raison de l'efficacité de la minimisation observée pour l'algorithme récursif.

→ Empiriquement, pour valider et compléter l'étude de la contraction de $\mathrm{id} - \delta$, un calcul a été effectué de la plus grande valeur propre de :

$$\left[\mathrm{id} - \dot{\delta}(c)\right]'\left(\mathrm{id} - \dot{\delta}(c)\right) = \left[\mathrm{id} - \left(\mathrm{id} - (cu)_s \left(\sum cuu'\right)^{-} u^s\right)^{*2}\right]' \left[\mathrm{id} - \left(\mathrm{id} - (cu)_s \left(\sum cuu'\right)^{-} u^s\right)^{*2}\right]$$

→ Sur les 10 000 échantillons simulés, les pondérations $c = 1$ et $c = 1 - \pi$ donnent pour cette matrice des valeurs propres maximales de l'ordre de 1 [41], à $4\,10^{-6}$ près.

→ En revanche pour la pondération $c = \dfrac{1 - \pi}{\delta(1)}$ (pour les unités telles que $\delta_i(1) > 10^{-6}$, et $c_i = 1 - \pi_i$ sinon), la valeur propre maximale est plus élevée, sans toutefois dépasser à 1.05 (à l'exception de 315 des 10 000 simulations).

→ Sur le premier échantillon, lorsqu'un poids de 2 est attribué à la moitié de chaque échantillon régional, et 1 pour le reste, la valeur propre propre maximale dépasse sensiblement 1 pour 4 régions, dont Pays de la Loire et PACA.

⇒ Sur les données du sondage étudié, la fonction ξ n'est plus contractante lorsque la pondération est trop dispersée, en fonction de la taille de l'échantillon et de la spécificité régionale. Ce résultat est conforme aux attentes théoriques. Cependant, la condition de contraction dans toutes les directions n'est pas nécessaire pour l'efficacité relative de l'algorithme récursif.

41. La valeur 1 est atteinte dès que $\mathrm{Ker}\left[\dot{\delta}(c)\right] \neq 0$.

Annexe F compléments sur l'algorithme récursif

• Tant que $\delta(c) \leqslant 1-\pi$ et $\delta[\xi(c)] \leqslant 1-\pi+1-\pi-\delta(c)$ et qu'il existe $i \in s$ tel que ces deux inégalités soient strictes et $\delta_i(1) > 0$, la condition de poursuite de l'algorithme est vérifiée. En effet, ces conditions impliquent que $\xi(c) \geqslant c$ et $\delta(c) \leqslant \delta[\xi(c)] \leqslant 1-\pi+1-\pi-\delta(c)$ (inégalités vectorielles) donc $-(1-\pi-\delta(c)) \leqslant \delta[\xi(c)] - (1-\pi) \leqslant 1-\pi-\delta(c)$ et ces deux inégalités sont strictes pour i [42], ce qui implique que $\|1-\pi-\delta[\xi(c)]\| < \|1-\pi-\delta(c)\|$ (pour la norme euclidienne).

⇒ Si $\exists i \in s\,/(1-\pi_i)u_i \neq 0$ et $\delta_i(1) > 0$ alors la première étape de l'algorithme récursif fonctionne toujours, c'est-à-dire que :

$$\|\delta[\xi(1-\pi)] - (1-\pi)\| < \|\delta(1-\pi) - (1-\pi)\| \qquad \textbf{(eF.75)}$$

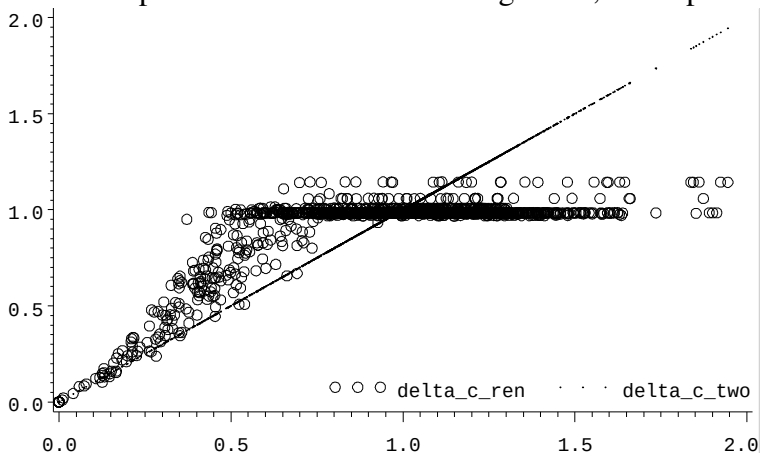
↳ Comme $\forall c > 0, \delta(c) \leqslant c$, $\delta(1-\pi) \leqslant 1-\pi$, et cette inégalité est stricte pour la coordonnée i car $(1-\pi_i)u_i \neq 0$.

↳ de même $\delta[\xi(1-\pi)] \leqslant \xi(1-\pi) = 1-\pi+1-\pi-\delta(1-\pi)$. L'inégalité est stricte pour i car $\xi_i(1-\pi)\,u_i \neq 0$ vu que $u_i \neq 0$ et $\xi_i(1-\pi) \neq 0$ car $1-\pi_i > 0$ d'une part et $\delta_i(1-\pi) < 1-\pi_i$ d'autre part.

⇒ La pondération récursive est préférable à la pondération par $1-\pi$ selon la distance inter-diagonales, à la seule condition qu'il existe $i \in s\,/(1-\pi_i)u_i \neq 0$ et $\delta_i(1) > 0$.

• Le Graphique 3 montre que la pondération récursive comparée au calage sur la trace de l'ESB donne plus de poids aux unités 'atypiques', celles de plus faibles coefficients diagonaux. D'autre part le maximum du coefficient diagonal sur l'ensemble des simulations est de 1.70 pour la pondération récursive et la médiane de 0.985 , contre 3.05 et 1.014.

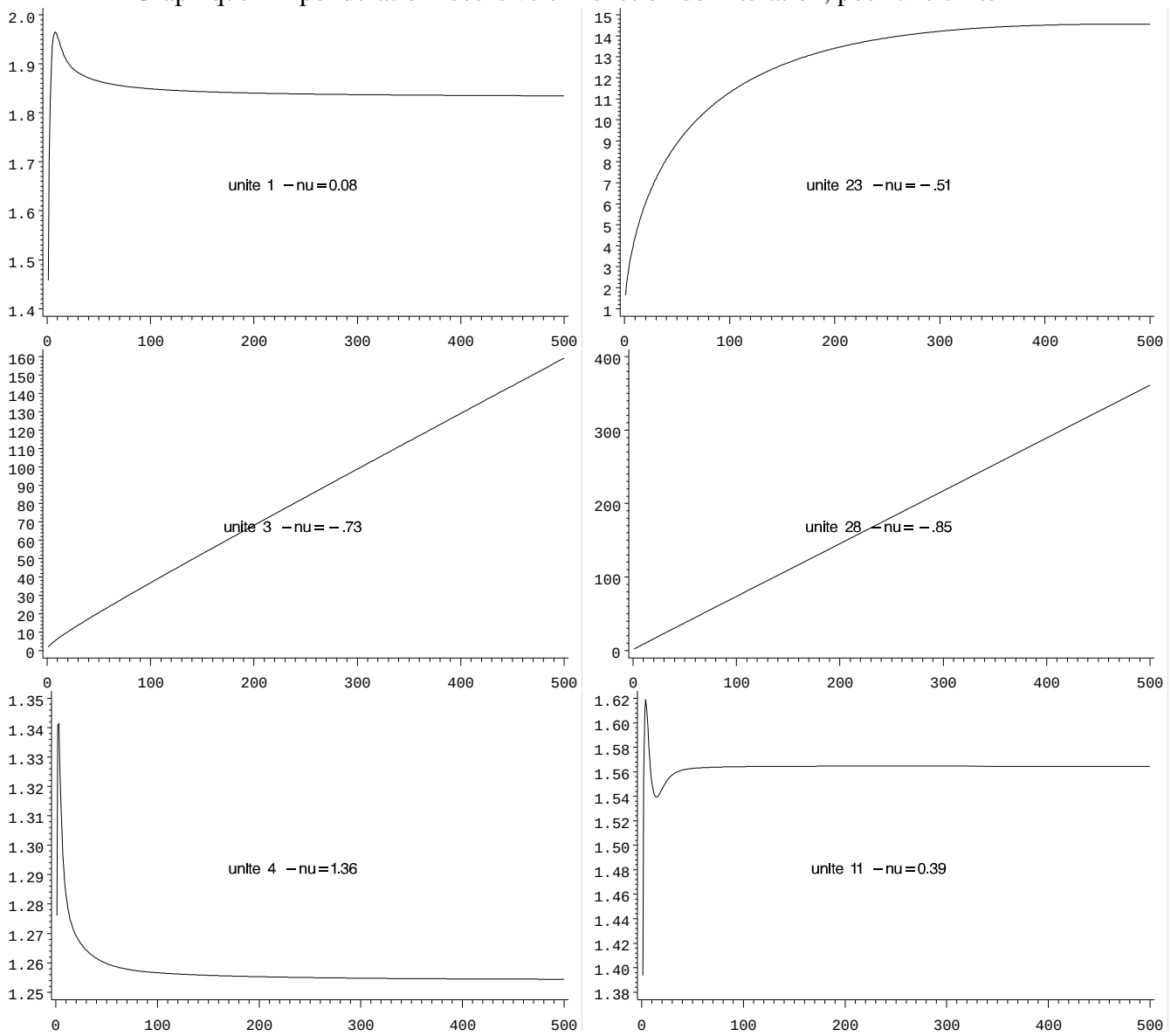
Graphique 3 – Comparaison des coefficients diagonaux, sur le premier échantillon



• Sur l'exemple des 58 unités d'une région (Poitou-Charentes), l'évolution des $\xi_i(\mathbb{N})$ parait très régulière. Trois profils peuvent être distingués (Graphique 4) : stabilisation rapide, croissance puis stabilisation progressive et enfin croissance régulière.

42. en utilisant : $\begin{cases} \xi(c) \geqslant c \\ \xi_i(c) < c_i \Rightarrow \delta_i(c) < \delta_i[\xi(c)], \text{ vu (e2.24)} \\ \delta_i(1) > 0 \end{cases}$

Graphique 4 – pondération récursive en fonction de l'itération, pour une unité



Note : La pondération est produite par l'algorithme récursif sur le premier échantillon simulé, pour la région Poitou-Charentes. Les unités représentées sont représentatives des différents profils d'évolution de la pondération selon l'itération. Ils se distinguent par le paramètre $\nu_i = \min_{u_i'\beta=1} \left\| u^{\neq i}\beta \right\|^2_{1-\pi_{\neq i}} - (1-\pi_i)$, *ainsi que par* $\delta_i(1)$.

Annexe G notations

- Le complémentaire d'un sous-ensemble $A \subset \Omega$ est noté $A^c = \Omega \setminus A$.
- L'ensemble des applications de l'ensemble A dans B est noté B^A ou B_A selon la représentation de ses éléments, respectivement en ligne (donc indicée en colonne) ou en vecteur-colonne
- Pour une matrice A, la transposée est notée A', A^- est une inverse généralisée et A^+ est l'inverse généralisée de Moore et Penrose (**??**).
- $\mathrm{Im}(A)$ est l'image d'une matrice A ($\mathrm{Im}(A) = AE$, où E est l'espace de départ), $\mathrm{Ker}(A)$ est son noyau ($\mathrm{Ker}(A) = \{A = 0\}$), $|A|$ son déterminant et $\mathrm{rang}(A) = \dim[\mathrm{Im}(A)]$
- x_s ou $x(s)$ désigne une matrice indicée en ligne par l'ensemble s. Celle-ci est confondue avec la famille de vecteurs lignes $\left(\left(x_i'\right)_{i\in s}\right)$ et avec l'application linéaire $\left(\beta \longmapsto \left(x_i'\beta\right)_{i\in s}\right)$. Si nécessaire, l'ensemble des indices des colonnes est précisé sous la forme $x_s(K)$.

– x^s désigne une matrice indicée en colonne par l'ensemble s (donc $x^s = (x_s)'$) (Une entorse à la logique de la notation des indiçages est faite dans le cas d'un indice unique : u_i ou $u(i)$ désignera un vecteur colonne et u_i' un vecteur ligne, pour éviter la confusion possible avec la puissance i.)

– Les indices sont omis lorsque le contexte n'est pas ambigu.

– Pour un ensemble B :

- $\mathrm{Vect}(B)$ est le plus petit sous-espace vectoriel qui contient B
- $B^\perp = \{x \in E \,/\, \langle x, B\rangle = 0\} = \{x \in E \,/\, \forall y \in Bx, y = 0\}$ est le sous-espace vectoriel orthogonal à B

– s est l'échantillon tiré dans l'univers $\mathscr{P}$. Alternativement, en fonction du contexte, cette notation désigne l'indicatrice de présence dans l'échantillon ($s \in \{0,1\}^{\mathscr{P}}$).

– $|s|$ est le cardinal de l'ensemble s, soit ici la taille de l'échantillon

– $\mathbb{1}_i$ est le vecteur de ième coordonnée égale à 1 et dont les autres coordonnées sont nulles. ($\mathbb{1}_s$ est donc la matrice formée de ces indicatrices, à ne pas confondre avec le vecteur de coordonnées égales à 1, noté 1_s)

– L'espace vectoriel d'intérêt pour l'échantillon ($\ni y(s)$) est noté $E = \mathbb{R}_s$.

– Pour deux vecteurs x et y du même espace vectoriel $\mathbb{R}_s$:

- $x \leqslant y$ si $\forall i, x_i \leqslant y_i$
- $x < y$ si $\forall i, x_i < y_i$ (en particulier, $c > 0$ signifie : $\forall i \in s, c_i > 0$)

– Le produit de deux vecteurs est le vecteur constitué des produits de leurs coordonnées : $(xy)_i = x_i y_i$ et en particulier $x^2 = \left((x_i)^2\right)_{i \in s}$. Parallèlement, la matrice de diagonale x est confondue avec le vecteur x.

– Plus généralement deux types de produits termes à termes de matrices sont à distinguer :

→ produit de deux matrices A et B de mêmes dimensions : $(A * B)_{i,j} = A_{i,j} B_{i,j}$ (Cette notation vise à éviter la confusion avec le produit matriciel usuel.)

→ produit d'un vecteur par une matrice :

- en lignes : $(xA)_{i,j} = x_i A_{i,j}$
- en colonnes : $(A * x)_{i,j} = A_{i,j} x_j$

– Il faut faire attention lorsqu'une expression combine des produits matriciels usuels avec des produits termes à termes. Par exemple, il n'est pas vrai en général que $(AB) * C = A(B * C)$.

– A^{*2} désigne la matrice dont les coordonnées sont les carrés de celles de la matrice A.

– Pour une variable réelle, $y : s \longmapsto \mathbb{R}$, soit $y \in \mathbb{R}_s$, ou vectorielle : $\sum y = Y = y(+)$ est le total sur l'univers, $\sum_s y = \sum_{i \in s} y_i$ est le total sur l'échantillon, y_d est le total sur une unité de d-ième phase

– $\|x\|_M$ est la norme euclidienne d'un vecteur x pour la métrique M (ie une matrice symétrique, définie positive), soit $\sqrt{x'Mx}$. $\langle , \rangle_M$ est son produit scalaire (La mention de la métrique est omise si c'est l'identité.) En particulier, pour un vecteur $c(>0)$, $\|x\|_c$ est la norme pour la métrique diagonale définie par c, ce qui correspond simplement à la somme pondérée : $\|x\|_c^2 = \sum cx^2$. Cette dernière formule permet d'étendre la définition de $\|x\|_c^2$ au cas où $c \in \mathbb{R}_s$.

– proj_F^M est la projection orthogonale pour la métrique M sur le sous-espace-vectoriel F. (La mention de $M(=\mathrm{id})$ est encore omise pour la métrique euclidienne canonique.) $\mathrm{proj}_F^{\perp M} = \mathrm{proj}_{F^{\perp(M)}}^M$ est la projection sur l'orthogonal de F, soit : $\mathrm{proj}_F^{\perp M} = \mathrm{id} - \mathrm{proj}_F^M$.

– m est la mesure de comptage réduite sur l'univers $\mathscr{P}$: $m(y) = \dfrac{Y}{N}$, où $N = |\mathscr{P}|$