Approximation de la variance d'un sondage équilibré : évaluation sur des simulations de l'enquête Emploi

1. propriétés théoriques
2. comment mettre en oeuvre
3. évaluation empirique

rappel du principe de l'approximation de variance

- Pour le tirage sans remise d'un échantillon s, 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 $\frac{y}{\pi}(\mathscr{P})$:

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

$\rightarrow$ si $\forall i,j \in \mathscr{P}, \pi(i,j) > 0$ (et si $|s|$ fixe pour $\widehat{V}_{YG}$) alors deux formes quadratiques estiment sans biais $Var\left(\widehat{Y}\right)$ $(\forall y(\mathscr{P}))$:

$\triangleright$ $\widehat{Var}^{HT}\left(\widehat{Y}\right) = \widehat{V}_{HT} = \sum_{i,j\in s} \frac{y}{\pi}(i)\,\frac{y}{\pi}(j)\,\frac{\pi(i,j) - \pi(i)\,\pi(j)}{\pi(i,j)}$

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

$\rightarrow$ $\widehat{V}_{YG}$ positive si $\forall i \neq j, \pi(i,j) \leqslant \pi(i)\,\pi(j)$ (Yates-Grundy)

$\rightarrow$ $\mathrm{diag}\left(\widehat{V}_{HT}\right) = 1 - \pi$; $\mathrm{diag}\left(\widehat{V}_{YG}\right) \geqslant 1 - \pi$ (si (YG))

principe général de l'approximation de variance

- Pour un sondage équilibré d'entropie maximale ($p = p^*\left[\left|\widehat{X} = x(+)\right.\right]$ et p^* poissonnien), la variance asymptotique approximative :

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

$$\cong \sum_{\mathscr{P}} \left(\frac{y}{\pi} - \frac{x'}{\pi}\beta^*\right)^2 \pi^*(1-\pi^*) = M^*\left[\left(\frac{y}{\pi} - \frac{x'}{\pi}\beta^*\right)^2\right]$$

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

$\rightarrow$ estimée sans biais avec la mesure $\widehat{M^*}(z) = \sum_s z\frac{\pi^*(1-\pi^*)}{\pi}$ par :

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

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

→ Le principe général de l'estimation sur un échantillon est de remplacer la mesure inconnue $\widehat{M^*}(z)$ par une approximation de la même forme $\widetilde{M^*}(z) = \sum_s cz$ (c_i dépend en général de s) :

$$\widehat{Var}\left(\widehat{Y}\right) = \widehat{V}_{DT}(y) = \widetilde{M^*}\left[\left(\frac{y}{\pi} - \frac{x}{\pi}'\widehat{\beta}\right)^2\right] = Q\left(\frac{y}{\pi}(s)\right)$$

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

⟹ En pratique, la variance est estimée par la somme pondérée des carrés des résidus de la régression de la variable d'intérêt par le vecteur d'équilibrage sur l'échantillon pondéré par $\frac{c}{\pi^2}$.

→ sous le modèle de l'approximation, $\exists c^* / \widetilde{M^*_{c^*}}(z)$ soit une approximation de $Var\left(\widehat{Y}\right)$ convergente sous certaines hypothèses

→ si $y(\mathscr{P}) \in \text{Im}(x)$ alors $\widehat{V}_{DT}(y) = 0 =$ variance de $\widehat{Y}$ si équilibrage exact pour tous les échantillons tirables

expressions de la forme quadratique approximante

Avec les notations $u = \frac{x}{\pi}$ et $z = \frac{y}{\pi}$:

- comme le carré d'une norme :
$$\widehat{V}_{DT} = \min_{\beta} \left\| z - u'\beta \right\|_c^2 = \left\| \left(\mathrm{id} - \mathrm{proj}^c_{\mathrm{Im}(u_s)} \right)(z) \right\|_c^2$$
- avec la matrice de la forme quadratique :
$\widehat{V}_{DT} = z_s' \Delta(c) z_s$ où :
$$\Delta(c) = \mathrm{diag}(c) \left(\mathrm{id} - \mathrm{proj}^c_{\mathrm{Im}(u_s)} \right)$$
$$= \mathrm{diag}(c) - (cu)_s \left(\sum cuu' \right)^- (cu)^s$$
$$\left(\text{formule du résidu : } \widehat{\epsilon} = \left(\mathrm{id} - \mathrm{proj}^c_{\mathrm{Im}(u_s)} \right)(z_s) = \mathrm{diag}(c)^{-1} \Delta(c) z_s \right)$$

$\rightarrow$ formule de ses coefficients diagonaux :
$$\delta_i(c) = c \left(1 - cu' \left(\sum cuu' \right)^- u \right)(i) = \min_{\beta} \left\| \mathbb{1}_i - u'\beta \right\|_c^2$$

propriétés comme fonction de la pondération

- L'approximation de variance et les coefficients diagonaux se formulent par la fonction :

$$\varphi(c) = \varphi_z(c) = \min_{\beta} \left\| z - u'\beta \right\|_c^2 = \sum_s c \left(z - u'\beta \right)^2$$

$\rightarrow$ positive et croissante en c

$\rightarrow$ homogène de degré 1

$\rightarrow$ concave

$\rightarrow$ continue sur $\mathbb{R}_s^+$

$\rightarrow$ dérivable en $c > 0$ et de dérivée :

$$\dot{\varphi}(c) = \left\| z - u'\widehat{\beta}_c \right\|_{dc}^2 = \sum_{i \in s} \left[z_i - u_i'\widehat{\beta}_c \right]^2 dc_i$$

$$\text{où } \widehat{\beta}_c \in \operatorname*{argmin}_{\beta} \left\| z - u'\beta \right\|_c$$

dérivée des coefficients diagonaux

- La matrice de la dérivée de la diagonale est calculable ainsi :

$\rightarrow \dot{\delta}(c) = \left(\mathrm{id} - (cu)_s \left[\sum_s cuu'\right]^- u^s\right)^{\circledast 2}$ (matrice formée des carrés des coefficients)

  - $\frac{\partial \delta_i(c)}{\partial c_i} = \left[1 - u_i' \left[\sum_s cuu'\right]^- c_i u_i\right]^2$
  - pour $i \neq j$: $\frac{\partial \delta_i(c)}{\partial c_j} = \left[u_j' \left[\sum_s cuu'\right]^- c_i u_i\right]^2$

(Le signe positif de ces dérivées correspond à la croissance de $\varphi(c)$.)

optimalité 'a priori' de la pondération

Si les probabilités d'inclusion doubles sont indéterminées, l'information connue sur l'estimateur sans biais de la variance est sa diagonale.

- La pondération optimale selon les tests de l'article de Deville et Tillé identifie les diagonales de l'ESB et de la FQA :

$$\delta(c) = 1 - \pi \quad \text{(équation de pondération)}$$

→ pour une variable z, l'écart entre les deux formes quadratiques tronqué aux termes diagonaux est $\sum_s (\delta(c) - (1-\pi)) z^2$

→ Lorsque l'identité stricte n'est pas réalisable, il semble raisonnable de chercher à minimiser la valeur absolue de cet écart sur un ensemble de variables selon un critère à préciser.

bornes de la pondération

- la croissance de $\delta(c) \Longrightarrow \min(c)\,\delta(1) \leqslant \delta(c) \leqslant \max(c)\,\delta(1)$
- d'autre part : $0 \leqslant \delta(c) = \min_{\beta} \left\| \mathbb{1}_s - u'\beta \right\|_c^2 \leqslant \left\| \mathbb{1}_s \right\|_c = c$

$\Longrightarrow$ si c est une solution exacte alors $c \geqslant 1 - \pi$

$\Longrightarrow$ si $\exists i \in s \,/\, \delta_i(1) = 0$ et $\pi_i < 1$ alors pas de solution exacte (et $\forall c \geqslant 0, \delta_i(c) = 0$)

$\Longrightarrow$ si $\delta(c) = 1 - \pi$ alors $\max(c) \geqslant \max\left[\dfrac{1-\pi}{\delta(1)}\right]$

- Comme $\lim_{c_i \longrightarrow \infty} \|\delta(c)\| < \infty$, il ne semble pas exister de majorant imposable à la recherche d'une solution minimisante.

analyse de la condition d'exclusion d'une solution exacte

$$\delta_i(1) = 0 \iff \min_{\beta} \|\mathbb{1}_i - u_s\beta\|_1^2 = 0$$

$$\iff \mathbb{1}_i \in \mathrm{Im}(u_s)$$

$$\iff \exists\beta \,/\, \begin{cases} \beta' u^{\neq i} = 0 \\ \beta' u_i \neq 0 \end{cases}$$

$$\iff u_i \notin \mathrm{Im}(u^{\neq i}) \iff x_i \notin \mathrm{Im}(x^{\neq i}) :$$

$\leftrightarrow$ Le vecteur d'équilibrage en i n'est pas dans l'espace généré par les autres unités (i est atypique pour les variables d'équilibrage).

- moins vraisemblable si $|s|$ grand $/\mathrm{rang}(u_s) = \mathrm{rang}(x_s)$
- vérifiée si $|s| = \mathrm{rang}(u_s)$
- lien avec l'équilibrage : si $u_i \notin \mathrm{Im}\left(u^{\mathscr{P}\setminus i}\right)$ et $\pi_i < 1$ alors l'équilibrage n'est pas toujours (exactement) possible

critères d'optimalité

- Pour la qualité de l'estimation de variance, la proximité entre $\delta(c)$ et $1-\pi$ peut être considérée comme bonne si l'écart : $\sum (\delta(c) - (1-\pi)) z^2$ est 'petit en moyenne' sur $z \in]-1,1[_s$

$\rightarrow$ La pondération $c_2 = (1-\pi)\left(\sum_s 1-\pi\right) \Big/ \sum_s \delta(1-\pi)$:

  - ▷ annule l'écart sur les fonctions constantes
  - ▷ s'annule pour $\pi = 1$, comme la cible $1-\pi$
  - ▷ $c_2 \geqslant 1-\pi$, condition nécessaire pour une solution

$\rightarrow$ Chercher à minimiser $\omega(c) = \|\delta(c) - (1-\pi)\|^2$ présente l'avantage théorique d'exploiter toute l'information sur la diagonale de l'ESB.

un algorithme de minimisation récursif

- Une solution exacte de l'équation de pondération est un point fixe de la fonction $\xi(c) = c - \delta(c) + 1 - \pi$.

(corrige le point de l'itération précédente par l'erreur correspondante)

→ si $\left\|\mathrm{id} - \dot{\delta}\right\| \leqslant 1$ sur $[c, \xi(c)]$ alors la fonction objectif baisse entre c et $\xi(c)$:

$$\|\delta[\xi(c)] - (1-\pi)\| = \|(\mathrm{id} - \delta)(c + 1 - \pi - \delta(c)) - (\mathrm{id} - \delta)(c)\|$$
$$\leqslant \sup_{[c,\xi(c)]} \left\|\mathrm{id} - \dot{\delta}\right\| \|\delta(c) - (1-\pi)\| \leqslant \|\delta(c) - (1-\pi)\|$$

→ Cet algorithme présente l'avantage informatique de ne pas nécessiter de calcul du gradient $\dot{\delta}$.

→ si initialisation à $1 - \pi$, deux propriétés bénéfiques ;

- ▷ $\xi^{\mathbb{N}}(1-\pi) \geqslant 1 - \pi$
- ▷ si $\pi_i = 0$ alors $[\xi^{\mathbb{N}}(1-\pi)]_i = 0$

efficacité pour la minimisation

L'algorithme récursif :

$\rightarrow$ est plus simple à programmer et nettement plus rapide d'exécution que les algorithmes à gradient testés

Sur les données testées :

$\rightarrow$ Il fournit la distance $\|\delta(c) - (1-\pi)\|$ minimale dans la grande majorité des simulations et des régions.

$\rightarrow$ Il minimise largement la somme des coefficients diagonaux supérieurs à 1.

$\Longrightarrow$ Il parait clairement optimal pour l'objectif 'technique' de rapprochement des deux diagonales, du moins selon le critère de proximité retenu.

les différences de minimisation obtenues sont-elles importantes ?

- La baisse de la fonction objectif entre les pondérations v_2 et récursive est entre 32% et 71% selon la région, en médiane sur les simulations.

$\rightarrow$ Relativement à la norme (médiane) de $1 - \pi$, ça représente une baisse encore sensible : entre 3% et 18%.

$\rightarrow$ Toutefois l'importance de cette amélioration de la minimisation pour l'objectif final de l'estimation de variance est indéterminée a priori, en l'absence d'une formulation du lien entre la distance et la précision de l'estimation de variance.

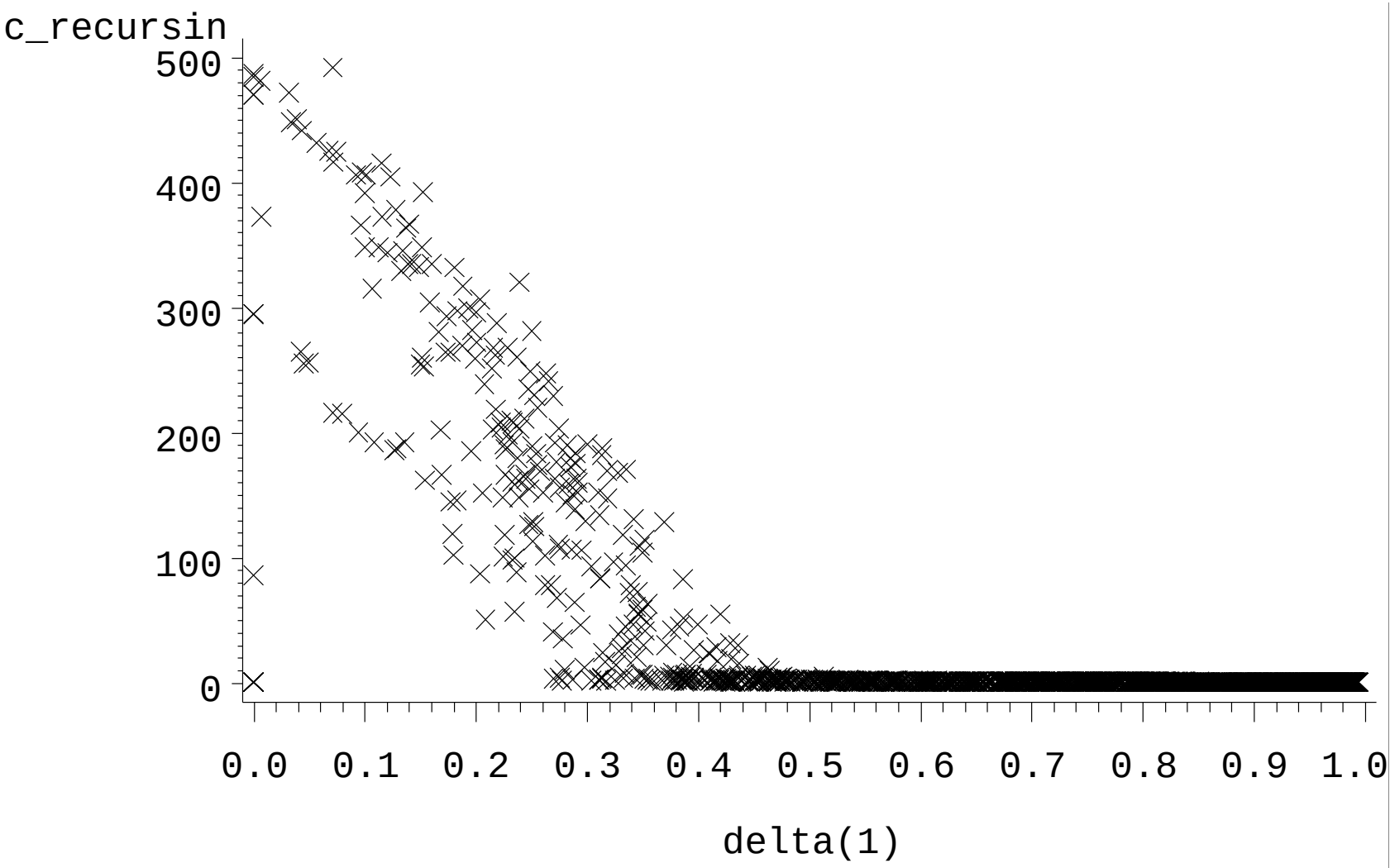
pondérations élevées et incertitude numérique

- La pondération maximale produite par cet algorithme est entre 200 et 500, en médiane régionale sur les simulations, alors qu'elle est de l'ordre de 1 à 3 pour la pondération simple calée sur la trace.

$\rightarrow$ Or les poids élevés accroissent la sensibilité de l'approximation aux gros résidus.

- De plus, la qualité de l'inverse généralisée de $\sum cuu'$, selon une mesure (absolue) du type $\|A - AA^{-}A\|$, parait décroître en fonction de la pondération maximale. Ceci brouille la comparaison de l'efficacité des méthodes.

les poids élevés sont causés par les petits $\delta(1)$



(sur le premier échantillon simulé)

contrôle des poids maximaux v/s contrôle des coefficients diagonaux

Pour nuancer l'importance des poids élevés :

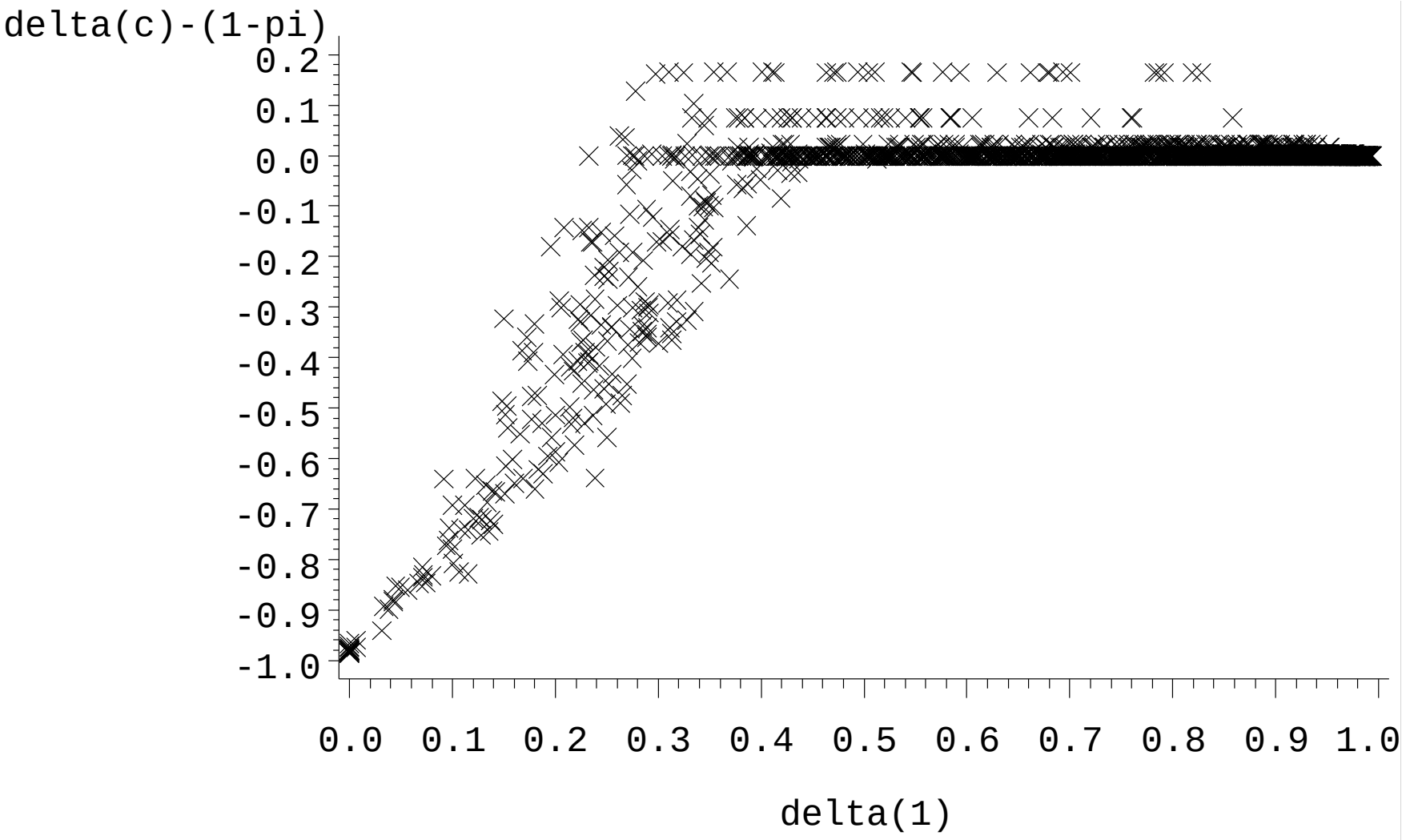
- Les coefficients diagonaux donnent un certain contrôle sur l'impact des valeurs individuelles dans l'approximation de variance. En effet, si $\widehat{\epsilon}$ désigne le résidu de la régression d'équilibrage :

$$c_i^2 \left[\widehat{\epsilon_i}\right]^2 = \left[\Delta_i(c)\, z_s\right]^2 \leqslant \left[\sum_{j \in s} \left|\Delta_{i,j}(c)\right| |z_j|\right]^2 \leqslant \delta_i(c) \left[\sum |z_s| \sqrt{\delta_s(c)}\right]^2$$

  - ▷ Les coefficients diagonaux et les valeurs prises par la variable d'intérêt sur l'échantillon déterminent ce majorant.

  → De plus, contrôler le poids maximal peut augmenter en compensation le poids d'autres unités, avec un effet pervers sur les coefficients diagonaux de ceux-ci. Globalement, ceci se traduit par un éloignement entre les deux diagonales. Les tests effectués ne sont pas favorables à cette option.

la minimisation est inachevée pour les petits $\delta(1)$ avec une compensation à l'excès pour d'autres unités



la pondération optimisée est un peu plus précise

variable		moy	moyvar	min	p5	p10	q1	médian	q3	p90	p95	max
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
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
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
nbpi2	v1	-7,8	-7,8	-16,6	-11,4	-10,6	-9,3	-7,8	-6,3	-4,9	-4,1	0,7
nbpi2	v2	3,0	3,1	-6,5	-0,9	-0,1	1,4	3,0	4,6	6,1	7,1	13,5
nbpi2	ren	2,9	3,0	-6,6	-1,1	-0,3	1,2	2,9	4,6	6,2	7,2	12,3
mar	v1	-18,1	-18,1	-24,8	-20,9	-20,3	-19,3	-18,1	-16,9	-15,9	-15,3	-9,7
mar	v2	-6,0	-6,0	-13,4	-9,2	-8,6	-7,4	-6,0	-4,6	-3,4	-2,7	3,5
mar	ren	-5,6	-5,6	-13,8	-9,1	-8,4	-7,1	-5,6	-4,1	-2,7	-1,9	4,2
isf	v1	13,6	13,7	-9,7	3,5	5,6	9,3	13,3	17,6	21,8	24,3	39,9
isf	v2	20,2	20,4	-3,4	10,0	12,2	15,9	20,0	24,4	28,5	31,0	47,0
isf	ren	20,0	20,2	-4,2	9,6	11,8	15,6	19,7	24,3	28,4	31,0	48,0

v1 : $c = 1 - \pi$

v2 : $c = (1 - \pi) \sum_s 1 - \pi \Big/ \sum_s \delta\,(1 - \pi)$

ren : minimisation par récursion sur $c - \delta\,(c) + 1 - \pi$

conclusion

- apports de l'étude :
  - La portée de l'optimisation de la pondération de l'estimateur de la variance équilibrée est limitée par la présence d'unités spécifiques selon les variables de calage. Cette condition dépend du rapport entre la taille de l'échantillon et le nombre de variables d'équilibrage.
  - L'algorithme récursif semble être une bonne méthode d'optimisation de la pondération.
  - Sur les données simulées, la pondération optimisée fournit un estimateur plus précis que la pondération simple calée sur la trace, mais l'avantage est plus réduit qu'attendu.
- perspectives d'approfondissement :
  - conditions d'existence d'une solution minimisante
  - propriétés théoriques de la pondération optimisée
  - conditions suffisantes de convergence de l'algorithme récursif
  - raffinement du critère de proximité entre les diagonales