Plans Systématiques Quasi-équilibrés

Lionel Qualité et Yves Tillé

Université de Neuchâtel

novembre 2012

Plans équilibrés

- Les échantillons $\mathbf{s}$ sont les vecteurs de $\mathbb{R}^N$ dont les coefficients sont les indicatrices de sélection des unités : les sommets du cube $[0,1]^N$.
- Un plan $p(\cdot)$, avec probabilités d'inclusion $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ est équilibré sur $\mathbf{X}$ ssi

$$\widehat{\mathbf{X}}_{HT} = \sum_{k \in \mathbf{s}} \frac{\mathbf{x}_k}{\pi_k} = t_{\mathbf{X}}, \ \forall \mathbf{s} \text{ t.q. } p(\mathbf{s}) > 0.$$

- Il n'en existe que si un plan affine, dont la direction et la position sont fonctions de $\mathbf{X}$ et $\boldsymbol{\pi}$ passe par des sommets du cube, et que $\boldsymbol{\pi}$ est une combinaison convexe de ces sommets.

Méthode du Cube

- Proposée par Deville et Tillé (2004) : marche aléatoire qui reste dans le plan affine des contraintes, et va se coller sur des faces du cube de dimensions de plus en plus basses.
- Si les sommets de l'intersection entre le cube et l'espace des contraintes ne sont pas tous des sommets du cube, il peut rester un problème d'arrondi (même si un plan équilibré existe).
- Celui-ci concerne au maximum m unités, où m est le nombre de variables auxiliaires dans $\mathbf{X}$.
- C'est en général le cas lorsque l'on utilise des variables auxiliaires "numériques".

Fonction de coût

- Pour un plan parfaitement équilibré, $\text{var}(\widehat{\mathbf{X}}_{HT}) = 0$.
- Lorsque l'on n'a pas de plan équilibré, on veut minimiser un critère, par exemple

$$C(p) = \sum_{\ell=1}^{m} \frac{\text{var}(\widehat{X}_\ell)}{t_{X_\ell}^2}.$$

- C'est une fonction linéaire en p.
- Le problème d'arrondi à la fin de la phase de vol de la méthode du cube est traité de cette manière.

Représentation des plans

- On liste tous les échantillons (evt. de taille fixe n) comme colonnes d'une matrice $\mathbf{S}$.
- Un plan de sondage avec probabilités $\boldsymbol{\pi}$ est un vecteur $\mathbf{p}$ de nombres positifs ou nuls dont la somme vaut 1 et tel que $\mathbf{Sp} = \boldsymbol{\pi}$.
- L'ensemble des plans de sondages $\mathcal{C}_{\boldsymbol{\pi}}$ est un polytope convexe de $\mathbb{R}^{2^N}$ (evt. $\mathbb{R}^{C_N^n}$).
- Support d'un plan $p(\cdot)$: $\mathcal{Q}_p = \{\mathbf{s} \text{ t.q. } p(\mathbf{s}) > 0\}$.

Plans à support minimal

- Les minima de $C(\cdot)$ sont atteints en des sommets de $\mathcal{C}_{\boldsymbol{\pi}}$: les plans extremaux (et si on en est capable, on utilise alors l'algorithme du simplexe pour trouver un minimum global).
- On sait caracteriser les plans extremaux : ce sont les plans à support minimal.
- Support minimal (pour $\boldsymbol{\pi}$) : ensemble d'échantillons $\mathcal{Q}$ tels que $\boldsymbol{\pi} \in \text{Conv}(\mathcal{Q})$ et $\mathcal{R} \subsetneq \mathcal{Q} \Rightarrow \boldsymbol{\pi} \notin \text{Conv}(\mathcal{R})$.
- Remarque : le cardinal de $\mathcal{Q}$ ne dépasse pas $N+1$ (evt. N), cf. Wynn (1977), et être un support minimal ne veut pas dire qu'il n'existe pas de support de cardinal plus petit.
- On trouve un algorithme pour de tels plans dans Deville et Tillé (1998).

Plans systématiques

- Le plan systématique est aussi un plan à support minimal (Pea et coll., 2007).
- Faire les sommes : $V_0 = 0$, $V_1 = \pi_1, \ldots,$ $V_k = \sum_{i=1}^{k} \pi_k$, générer une réalisation u de $\mathcal{U}([0,1])$, et sélectionner tous les k tels que $V_{k-1} < u \leq V_k$.
- Algorithme dépend de l'ordre de la population : l'ensemble des plans systématiques $\mathcal{P}_{\mathrm{SYST}}$ est l'ensemble des plans obtenus pour tous les ordres possibles.
- On a dans Pea et coll. (2007) un algorithme rapide pour calculer le support et le plan $p(\cdot)$.

Groupe symétrique

- $\mathcal{P}_{\mathrm{SYST}}$ est l'orbite du plan systématique $p_e(\cdot)$ (e désigne l'ordre initial) sous l'action du groupe de permutations Σ_N.
- Le $N-$cycle $(1, \ldots, N)$ et le retournement $(1, \ldots, N) \mapsto (N, \ldots, 1)$ ne modifient pas p_e : ils sont dans le stabilisateur $\mathcal{S}_e$.
- Sauf cas particuliers, ces deux permutations engendrent $\mathcal{S}_e$, qui est alors de cardinal 2N.
- $\mathcal{P}_{\mathrm{SYST}}$ est en bijection (notée Ψ) avec $\Sigma_N/\mathcal{S}_e$, et a un cardinal en général égal à $(N-1)!/2$.
- Trop gros pour une recherche exhaustive, et trouver le minimum de $C(\cdot)$.

Distance sur le groupe symétrique

- On cherche un minimum local : dépend d'une métrique sur $\Sigma_N/\mathcal{S}_e$.
- Si $\mathcal{G}$ est un ensemble symétrique de générateurs de Σ_N, on peut définir $d_{\mathcal{G}}(\sigma, \tau)$, le nombre minimum d'éléments de $\mathcal{G}$ pour écrire $\sigma\tau^{-1}$.
- Graphe des points adjacents : graphe de Cayley. Si $\mathcal{G}$ est l'ensemble des transpositions, distance de Cayley.
- Distance transportée sur $\Sigma_N/\mathcal{S}_e$ par

$$\delta_G(p_1, p_2) = \inf_{\sigma \in \Psi^{-1}(p_1), \tau \in \Psi^{-1}(p_2)} d_{\mathcal{G}}(\sigma, \tau).$$

Minimisation locale

- Propriété :

$$\delta_G(p_1, p_2) = 1 \Leftrightarrow \forall \sigma \in \Psi^{-1}(p_1), \exists \tau \in \Psi^{-1}(p_2) | d_{\mathcal{G}}(\sigma, \tau) = 1 \quad (1)$$

- Algorithme :
  1. On part d'un ordre initial (ex: selon π_k croissants, voir Hartley, 1966),
  2. On essaye les transpositions jusqu'à en trouver une qui fait baisser $C(\cdot)$, que l'on applique,
  3. Alternativement, on n'applique que la meilleure transposition,
  4. On recommence jusqu'à ne plus arriver à faire baisser $C(\cdot)$.

- Grâce à la propriété 1, on arrive bien à un minimum local pour $\delta_{\mathcal{G}}$.

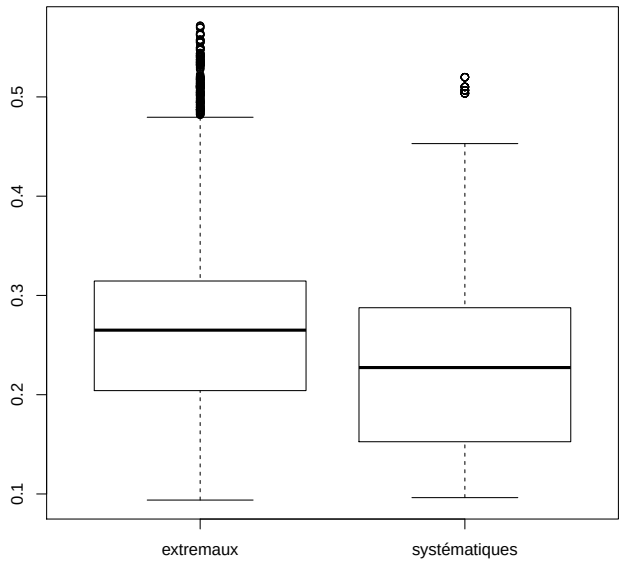
- En ne mettant dans $\mathcal{G}$ que les transpositions $(i, i+1)$ on accélère les calculs de $C(\cdot)$, mais on péjore les résultats.

Exemple 1 : faible dimension

- $N = 7,\ n = 3,\ m = 2,\ \boldsymbol{\pi} = (.11, .51, .31, .61, .32, .41, .73)$
- $\mathbf{x}_1 = (10.42, 8.31, 10.05, 8.55, 14.49, 9.46, 13.70)$,
  $\mathbf{x}_2 = (7.28, 11.72, 16.23, 5.42, 14.27, 6.86, 13.18)$
- Il y a $29'402$ plans extremaux, le coût minimum est 0.09385,
- Il y a 360 plans systématiques, les minima locaux avec la distance de Cayley sont 0.09631, 0.101353, et 0.12685
- Le coût du plan à entropie maximale est 0.21960.

Exemple 1 : suite

- Distribution de $C(\cdot)$ sur les plans extremaux et les plans systématiques.



- Minimum atteint selon ordre de départ :

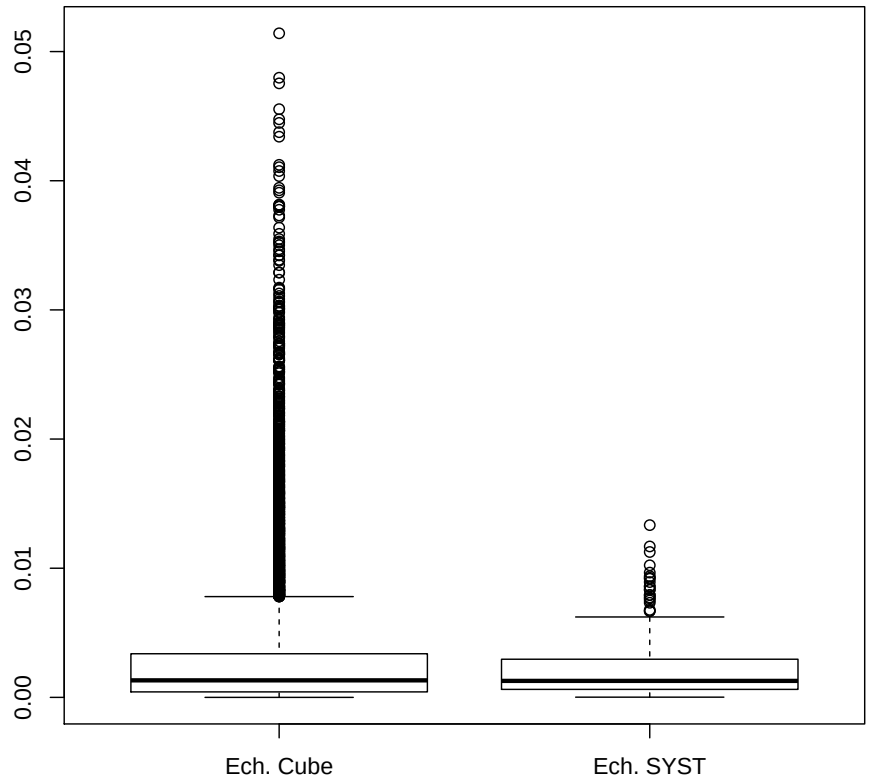
$C(\cdot)$	0.09631	0.101353	0.12685
proportion	52.4%	32.2%	15.4%

Exemple 2 : MU284

- $\pi \propto P75$, auxiliaires : P75, RMT85, SS82, ME84, $n = 20$, $N = 282$
- Cube : 10'000 échantillons, Systématique : P75 trié.
- Plan à entropie maximale : $C = 0.20497$, Cube : $\widehat{C} = 0.00304$,
- Systématique : $C < 0.0004$ en utilisant toutes les transpositions, et $C = 0.00205$ avec les transpositions $(i, i+1)$.

Exemple 2 : MU284 suite

- Coûts des échantillons sélectionnés par Cube, et dans le support du plan systématique trouvé.

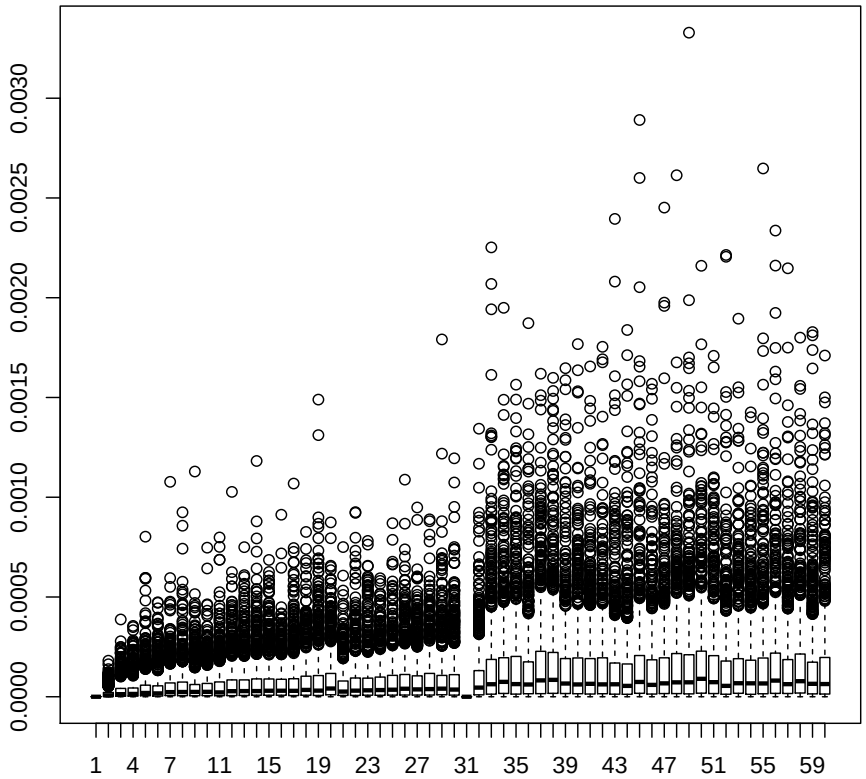


Exemple 3 : Utilisation comme phase d'atterrissage

- $N = 1000$, $n = 150$, $m = 30$ (π, 1, et 28 normales indépendantes),
- 1000 échantillons tirés avec atterrissage par élimination de variables, et 1000 autres par plans systématiques,
- Coût moyen par élimination : 0.00189, Coût moyen par systématique : 0.00434
- Variables traitées de manière inégale avec la méthode d'élimination, mais jusqu'ici meilleurs résultats qu'avec le systématique.

Exemple 3 : Atterrissage suite

- CV^2 pour chaque variable, élimination et systématique



Commentaires

- Les métriques utilisées font totalement perdre le caractère linéaire de $C(\cdot)$.
- L'ordre initial, et le système de générateurs $\mathcal{G}$ ont leur importance.
- Le minimum sur $\mathcal{P}_{\mathrm{SYST}}$ peut être loin du minimum sur $\mathcal{C}_{\boldsymbol{\pi}}$.
- Mais cela fonctionne relativement bien.

Références


Deville, J.-C. and Tillé, Y. (1998).
Unequal probability sampling without replacement through a splitting method.
Biometrika, 85:89–101.


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


Hartley, H. O. (1966).
Systematic sampling with unequal probability and without replacement.
Journal of the American Statistical Association, 61:739–748.


Pea, J., Qualité, L., and Tillé, Y. (2007).
Systematic sampling is a minimal support design.
Computational Statistics and Data Analysis, 51:5591–5602.


Wynn, H. P. (1977).
Convex sets of finite population plans.
Annals of Statistics, 5:414–418.