Journées de Méthodologie Statistique de l'Insee

À la recherche du plan déterminantal optimal

Kim Antunez (Insee), Vincent Loonis (Insee)

Session 15 : Échantillonnage
mercredi 30 mars 2022

Sommaire

Les données Meuse

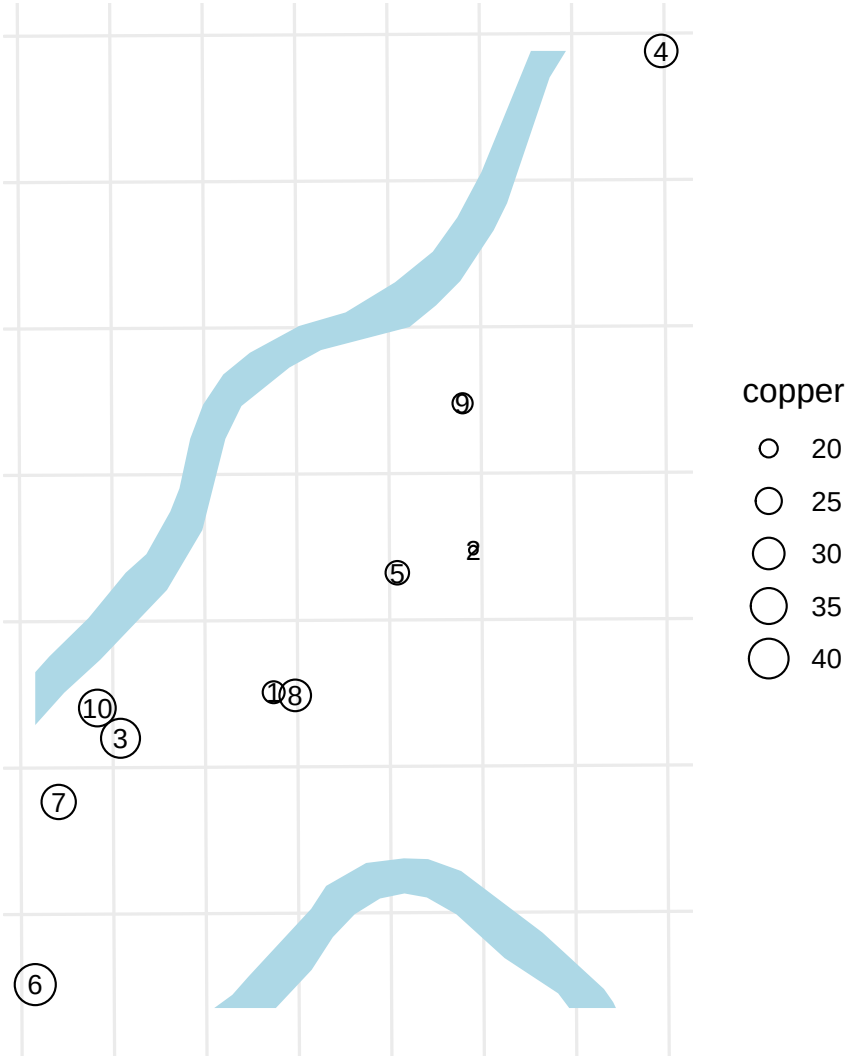

Figure 1 – N = 10 emplacements de métaux lourds dont on veut tirer un échantillon de taille n = 4.

Sommaire

Les plans déterminantaux (Loonis et Mary (2019))

Les plans de sondage déterminantaux sont paramétrés par les matrices hermitiennes contractantes :

- **hermitienne** : $K = \overline{K}^T$
- **contractante** : les valeurs propres de K sont comprises entre 0 et 1.

➔ Si les valeurs propres valent 0 ou 1, le plan est de **taille fixe**.

Les plans déterminantaux (Loonis et Mary (2019))

Les plans de sondage déterminantaux sont paramétrés par les matrices hermitiennes contractantes :

- **hermitienne** : $K = \overline{K}^T$
- **contractante** : les valeurs propres de K sont comprises entre 0 et 1.

➔ Si les valeurs propres valent 0 ou 1, le plan est de **taille fixe**.

Le plan associé $DSD(K)$ est tel que :

- la diagonale de K donne les **probabilités d'inclusion simple** ;
- les termes hors diagonale donnent les **probabilités d'inclusion double**.

➔ Toute caractéristique du plan est une fonction de K : $V(\hat{t}_y) = f(K)$; indicateur d'étalement spatial $= f(K)$...

Sommaire

Les plans déterminantaux optimaux

La recherche d'un plan de sondage peut souvent être mise sous la forme d'un problème d'optimisation :

- **Équilibrage** : Minimiser $\sum_{q=1}^{Q} V(\hat{t}_{x^q})$, avec $Q << N$, et les x^q, $q = 1, \cdots, Q$ des variables auxiliaires ;
- **Échantillonnage spatial** : Minimiser $\sum_{q=1}^{Q} V(\hat{t}_{x^q})$, avec $Q = N$, et les x^q, $q = 1, \cdots, N$ des variables auxiliaires géographiques particulières (Jauslin et Tillé (2019))

Les plans déterminantaux optimaux

La recherche d'un plan de sondage peut souvent être mise sous la forme d'un problème d'optimisation :

- **Équilibrage** : Minimiser $\sum_{q=1}^{Q} V(\hat{t}_{x^q})$, avec $Q << N$, et les x^q, $q = 1, \cdots, Q$ des variables auxiliaires ;
- **Échantillonnage spatial** : Minimiser $\sum_{q=1}^{Q} V(\hat{t}_{x^q})$, avec $Q = N$, et les x^q, $q = 1, \cdots, N$ des variables auxiliaires géographiques particulières (Jauslin et Tillé (2019))

➔ Dans le monde déterminantal, cela revient à résoudre un programme **semi-défini non linéaire sous contrainte** de la forme :

$$\underset{K}{Min}\ f(K) \text{ s.c } K = \overline{K}^T \text{ et } Sp(K) \in [0,1]^N$$

$$\underset{K}{Min}\ f(K) \text{ s.c } K = \overline{K}^T \text{ et } Sp(K) \in \{0;1\}^N \text{ (taille fixe)}$$

Équilibrage sur Q variables auxiliaires

Minimiser la somme des variances des estimateurs d'Horvitz-Thompson $\hat{t}_{x^q}$ dans le cas d'un plan de sondage de taille fixe.

$$
\begin{aligned}
C(\mathscr{P}) &= 2\sum_{q=1}^{Q} Var(\hat{t}_{x^q}) = \sum_{k=1}^{N}\sum_{l=1}^{N} \underbrace{\underbrace{\sum_{q=1}^{Q}\left(\frac{x_k^q}{\pi_k} - \frac{x_l^q}{\pi_l}\right)^2}_{Q_{kl}(x^1,\cdots,x^Q,\pi)}(\pi_k\pi_l - \pi_{kl})}_{C_{kl}(x^1,\cdots,x^Q,\pi)} \qquad (1)\\
&= e_N^t \underbrace{C(Q,\pi)}_{=\mathrm{C(Q,K)}=Q*K*\overline{K}} e_N
\end{aligned}
$$

➔ individus qui ne se ressemblent pas au sens de $\frac{x_k^q}{\pi_k} \Longrightarrow \pi_{kl}$ grandes

➔ **privilégier la sélection dans notre échantillon d'individus qui ne se ressemblent pas !**

Échantillonnage spatial

– Tirer des individus qui ne se ressemblent pas $\Longrightarrow$ privilégier l'**étalement spatial** de l'échantillon (première loi de la géographie de Tobler)

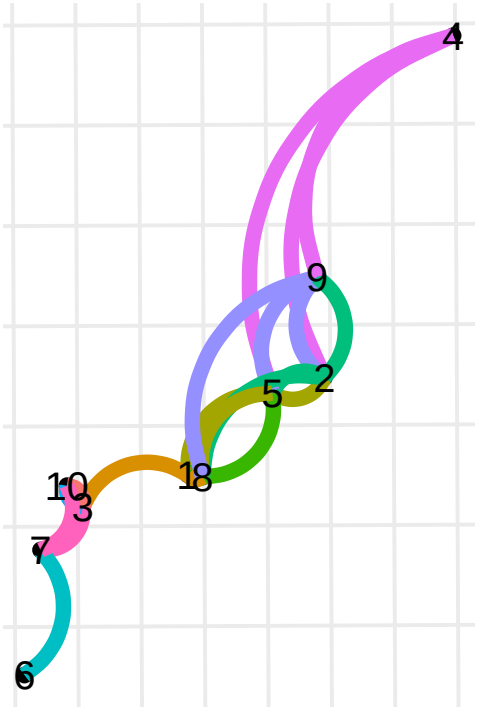

Figure 2 – Domaines D_q.

Équilibrage sur $\mathbf{Q = N}$ domaines géographiques (inspiré de Jauslin et Tillé (2019))

➔ On prend comme variables auxiliaires $x_k^q = \pi_k 1(k \in D_q)$. Équilibrer sur x^q, revient à souhaiter échantillonner un seul individu dans chaque domaine D_q.

$$\begin{aligned} C(\mathscr{P}) &= \underbrace{\sum_{k=1}^{N}\sum_{l=1}^{N}\underbrace{\sum_{q=1}^{Q}(1(k\in D_q)-1(l\in D_q))^2}_{Q_{kl}(x^1,\cdots,x^Q,\pi)}(\pi_k\pi_l-\pi_{kl})}_{C_{kl}(x^1,\cdots,x^Q,\pi)} \\ &= e_N^t\, C(Q,\pi)\, e_N \end{aligned} \tag{2}$$

Lien avec les plans déterminantaux

$$
\begin{aligned}
C(DSD(K)) &= \sum_{k=1}^{N}\sum_{l=1}^{N} Q_{kl}(x^1,\cdots,x^Q, diag(K))\,|K_{kl}|^2 \\
&= e_N^t \underbrace{Q * K * \overline{K}}_{\mathrm{C(Q,K)}} e_N \\
&= e_N^t\, C(Q,K)\, e_N
\end{aligned}
\tag{3}
$$

où $*$ désigne le produit matriciel de Hadamard (multiplication terme à terme).

➔ problème d'optimisation semi-définie non linéaire $\Longrightarrow$ La **recherche de solution est complexe** !

Sommaire

3 méthodes de résolution du problème d'optimisation

1. Sortir une solution de son chapeau : la matrice $K = P^{\Pi}$
2. Définir des heuristiques pour améliorer localement P^{Π}
3. Paramétriser les matrices hermitiennes contractantes sous la forme $K(\theta)$ et transformer le problème SDP contraint en un problème classique non contraint en θ

3 méthodes de résolution du problème d'optimisation

1. Sortir une solution de son chapeau : la matrice $K = P^{\Pi}$
2. Définir des heuristiques pour améliorer localement P^{Π}
3. Paramétriser les matrices hermitiennes contractantes sous la forme $K(\theta)$ et transformer le problème SDP contraint en un problème classique non contraint en θ

Quelles sont les **performances empiriques** des méthodes de recherche de **plans déterminantaux équilibrés** optimaux exhibées dans Loonis et Mary (2019) et Loonis (2021)

?

En fonction :

- de la taille de la population
- du nombre de variables auxiliaires considérées

Méthode 1 (Loonis et Mary (2019))

1 Les bonnes propriétés de la matrice P^{Π} triée astucieusement

Une matrice de projection particulière présente des propriétés intéressantes de répulsion (exclure au maximum de tirer des unités qui se ressemblent) quand les individus sont triés de manière à ce que deux individus dont l'ordre d'apparition est proche soient deux individus qui se ressemblent.

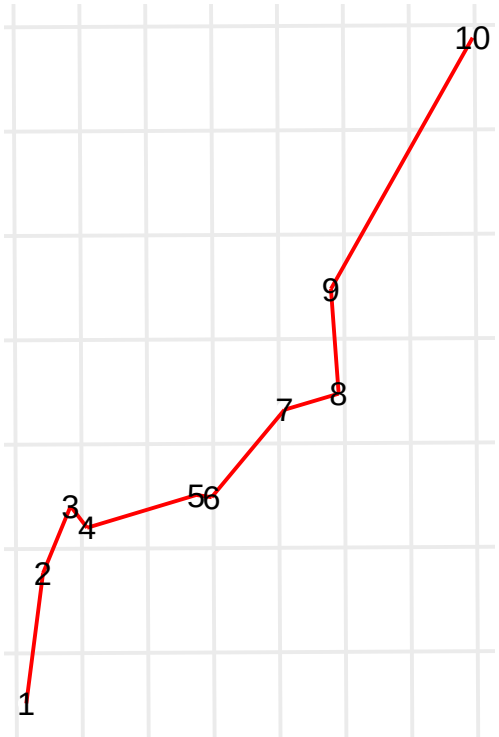

Figure 3 – Chemin de Hamilton sur données `Meuse`.

Méthode 2 (Loonis et Mary (2019))

2 Améliorer P^{Π} par rotations successives

Une modification itérative d'une matrice P^{Π} est rendue possible par des heuristiques. Elles exploitent le fait que la composition d'une matrice K par des matrices de rotation astucieusement choisies permet d'obtenir une nouvelle matrice hermitienne ayant la même diagonale et le même spectre que la matrice d'origine.

Algorithme (Minimisation de $C(DSD(K))$ par rotations)

1. *Initialiser l'algorithme à* $K = K^0 = P^{\Pi}$.
2. *For couple* $= 1$ *à* $\frac{N(N-1)}{2}$ *:*
   - *Tirer au hasard un couple* (k, l) *tel que* $l < k$ *non tiré jusqu'à présent.*
   - *Calculer* $K' = R_{kl}(\theta_{kl}) K R_{kl}^T(\theta_{kl})$.
   - *Si* $C(DSD(K')) < C(DSD(K))$ *alors* $K \leftarrow K'$.
3. *Fin de for.*
4. *Réitérer l'étape 2. jusqu'à ne plus trouver de rotation qui diminuent le critère.*
5. *Retourner* K.

Méthode 3 (Loonis (2021))

3 La paramétrisation des matrices K avec un grand nombre de paramètres

Paramétrisation par des matrices hermitiennes contractantes par un ensemble de $2Nn$ paramètres indépendants $\in [0,1]$:

- $\Omega[N,n]$ dont la $k^{ème}$ colonne conditionne la loi de la taille de l'échantillon dans le domaine des points de 1 à k.
- $\rho[N,n]$ qui, à variance de la taille de l'échantillon dans chacun des domaines 1 à k fixée par Ω, conditionne les termes hors diagonal de K, et donc les probabilités d'inclusion double.

➔ Transformation du problème d'optimisation semi-définie en un **problème non contraint** (mais comportant de nombreux paramètres !)

➔ Application $(\Omega,\rho) \longmapsto K(\Omega,\rho)$ continue mais non différentiable pour certaines composantes. $\Longrightarrow$ algorithmes d'optimisation stochastiques (type **recuit-simulé**)

Recuit simulé (Kirkpatrick *et al.* (1983))

Algorithme (Recuit simulé simplifié : pseudo-code)

Fonction recuit(s0, C, niter)
    $s := s0$
    $c := C(s)$
    $k := 0$
    pour k in $1 :$ *niter*
        $sn := voisin(s, pas)$
        $cn := C(sn)$
        si $en < e$ *alors*
            $s := sn \quad ; \quad e := en$
        $k := k + 1$
    retourne s
Fin Fonction

1. Tirage de **100 000** jeux de paramètres Ω et ρ
2. Conservation des **10** meilleurs en terme de minimisation du critère C
3. Recuit simulé sur chacun de ces 10 meilleurs (**400 000** itérations de recuit simulé)

Sommaire

Comparaison des 3 méthodes d'optimisation de plans déterminantaux

	Variance d'1 variable	Variances de 2 variables	Variances de 3 variables	Critère géographique (Q=N)
Matrice P^{Π}	1.757	3.604	5.154	4.515
Matrice P^{Π} puis rotations	1.757	3.604	5.097	4.515
Matrice par recuit simulé	1.757	3.591	5.073	4.249

TABLEAU 1 – Comparaison des 3 méthodes d'optimisation des plans déterminantaux en fonction du nombre de variables auxiliaires considérées.

Comparaison avec d'autres méthodes d'échantillonnage

	var1	var2	var3
Systematique aléatoire	10.242	14.792	18.956
Maximum d'entropie	10.443	15.053	19.258
Poisson équilibré	4.125	5.607	6.801
Pivot équilibré 1	2.995	4.299	5.665
Cube équilibré	2.811	4.733	6.572
Déterminantal (P^{Π})	1.757	3.604	5.154
Déterminantal (P^{Π} puis rotations)	1.757	3.604	5.097
Déterminantal (Recuit simulé)	1.757	3.591	5.073

TABLEAU 2 – Critères de minimisation des variances.

	IB1	B	geo
Systematique aléatoire	-0.175	0.299	10.514
Maximum d'entropie	-0.162	0.301	10.671
Poisson local	-0.432	0.189	5.718
Pivot local 1	-0.491	0.168	4.633
Cube local	-0.441	0.184	5.490
Wave Sampling	-0.569	0.164	3.354
Déterminantal (P^{Π})	-0.264	0.278	4.515
Déterminantal (P^{Π} puis rotations)	-0.526	0.147	4.515
Déterminantal (Recuit simulé)	-0.541	0.149	4.249

TABLEAU 3 – Critères géographique et d'étalement spatial.

Conclusion (1/2)

➔ Étude empirique des propriétés d'optimalité des plans déterminantaux

✚ P^{Π} semble être une borne inférieure pour les plans déterminantaux pour 1 variable auxiliaire

✚ Apport de la paramétrisation par Ω et ρ (Loonis (2021)) croissant quand le nombre de variables auxiliaires augmente

✚ Plans déterminantaux efficaces en comparaison d'autres méthodes d'échantillonnage

Conclusion (1/2)

➔ Étude empirique des propriétés d'optimalité des plans déterminantaux

✚ P^{Π} semble être une borne inférieure pour les plans déterminantaux pour 1 variable auxiliaire

✚ Apport de la paramétrisation par Ω et ρ (Loonis (2021)) croissant quand le nombre de variables auxiliaires augmente

✚ Plans déterminantaux efficaces en comparaison d'autres méthodes d'échantillonnage

▬ Malédiction de la dimension (Bellman et Kalaba (1959))

▬ Les plans déterminantaux vérifient nécessairement les conditions de Sen-Yates-Grundy ($\pi_{kl} \leq \pi_k \pi_l \quad (k \neq l)$).

Conclusion (2/2)

Prolongements

- Améliorer la parallélisation des calculs
- Identifier les paramètres les plus influents par l'analyse de sensibilité et les indices de Sobol (Sobol (2001)).
- Utiliser des algorithmes d'optimisation semi-définis adaptés : lagrangien augmenté, optimisation sur les variétés (Absil *et al.* (2009), Boumal *et al.* (2014), Fiala *et al.* (2013))

Merci pour votre attention !

Kim Antunez (Insee, Direction de la diffusion et de l'action régionale)

Vincent Loonis (Insee, Direction de la méthodologie et de la coordination statistique et internationale)

Références

Absil, P.-A., Mahony, R. et Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Bellman, R. et Kalaba, R. (1959). On adaptive control processes. *IRE Transactions on Automatic Control.*, 4(2):1–9.

Boumal, N., Mishra, B., Absil, P.-A. et Sepulchre, R. (2014). Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459.

Fiala, J., Kočvara, M. et Stingl, M. (2013). Penlab : A matlab solver for nonlinear semidefinite optimization. *arXiv preprint arXiv :1311.5240*.

Jauslin, R. et Tillé, Y. (2019). Spatial spread sampling using weakly associated vectors. *arXiv preprint arXiv :1910.13152*.

Kirkpatrick, S., Gelatt, C. D. et Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Loonis, V. (2021). Construire tous les plans de sondage déterminantaux. *Colloque francophone sur les sondages, Bruxelles, article soumis, DOI : 10.13140/RG.2.2.21214.92483.*

Loonis, V. et Mary, X. (2019). Determinantal sampling designs. *Journal of Statistical Planning and Inference*, 199:60–88.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280.