Plan de sondage déterminantal

équilibrage et calage a priori

Vincent Loonis [a], Xavier Mary [b]

(a) INSEE Division des Méthodes et Référentiels Géographiques (DMRG),
(b) laboratoire MODAL'X : Université Paris-Ouest Nanterre-La Défense

31 mars 2015

Rappel : le plan de sondage

- U de taille $N = 3$: $\{1, 2, 3\}$.
- $2^3 = 8$ sous-ensembles (échantillons) différents : $\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$
- On attribue à chaque échantillon, une probabilité d'être sélectionné :

$$p(\emptyset) = p(\{1\}) = p(\{2\}) = p(\{3\}) = 0$$

$$p(\{1, 2\}) = p(\{1, 3\}) = \frac{4}{9},\ p(\{2, 3\}) = \frac{1}{9}$$

$$p(\{1, 2, 3\}) = 0$$

- $\pi_1 = \frac{4}{9} + \frac{4}{9} = \frac{8}{9}, \pi_2 = \pi_3 = \frac{5}{9}$ inclusion simple,
- $\pi_{12} = \frac{4}{9}, \pi_{13} = \frac{4}{9}, \pi_{23} = \frac{1}{9}$ inclusion double,
- $\pi_{123} = 0$,inclusion triple.

Plans de sondage et processus ponctuels

- Un plan de sondage $\mathcal{P}$ non ordonné sans remise sur une population U est une loi de probabilité sur $\mathcal{S}=2^U$ ensemble des parties de U.

Cette définition correspond exactement, au vocabulaire près, à celle de certains processus ponctuels.

- Un processus ponctuel simple $\mathcal{P}$ sur $\mathcal{X}$ est une loi de probabilité sur l'ensemble $2^{\mathcal{X}}$ des parties de $\mathcal{X}$.

On peut importer dans la théorie des sondages les résultats de la théorie des processus ponctuels et réciproquement.

Plans de sondage et processus ponctuels

Les définitions sont trop générales pour être utiles dans la théorie et dans la pratique. On se limite en général à certains plans ou processus ayant des propriétés particulières :

- **plans de sondage particuliers :** Sondage aléatoire simple, stratifié, de taille fixe, poissonnien, équilibré, à plusieurs degrés, indirects...
- **processus :** déterminantal.

Plans déterminantaux

- Définitions,
- Propriétés,
- Plans déterminantaux particuliers.

Plan de sondage déterminantal définition et existence

- **Définition** : Un plan de sondage $\mathcal{P}$ est déterminantal s'il existe une matrice hermitienne K de taille (N, N), appelée noyau, indéxée par les individus de U, telle que pour tout $s \in \mathcal{S}$:

$$P(s \subseteq \mathbb{S}) = \det(K_{|s})$$

où $\mathbb{S} \sim \mathcal{P}$ et $K_{|s}$ est la sous matrice de K indéxée par les individus de s. Un tel plan est noté $\mathcal{P}_K$.

- **Existence** : Il existe un plan de sondage associé au noyau K, si et seulement si K est contractante.
- **Rappel** : K est hermitienne contractante si elle est telle que $K = {}^t\overline{K}$ et ses valeurs propres sont dans $[0; 1]$

Exemple

L'exemple précédent est en fait déterminantal de noyau :

$$K = \frac{1}{9}\begin{pmatrix} 8 & -2 & -2 \\ -2 & 5 & -4 \\ -2 & -4 & 5 \end{pmatrix}.$$

- K est hermitienne, car réelle et symétrique, et contractante,
- $pr(1 \subset \mathbb{S}) = \det(\frac{1}{9}(8)) = \frac{8}{9}$,
- $pr(\{1,2\} \subset \mathbb{S}) = \det(\frac{1}{9}\begin{pmatrix} 8 & -2 \\ -2 & 5 \end{pmatrix}) = \frac{4}{9}$,
- $pr(\{1,2,3\} \subset \mathbb{S}) = \det(\frac{1}{9}\begin{pmatrix} 8 & -2 & -2 \\ -2 & 5 & -4 \\ -2 & -4 & 5 \end{pmatrix}) = 0$,

Exemples

Quelques exemples de matrices hermitiennes contractantes

- les matrices de projection orthogonale,
- les matrices diagonales de termes diagonaux dans $[0; 1]$, qui correspondent aux plans poissonniens.

Ces matrices s'écrivent plus généralement $P\Gamma^t\overline{P}$ où P est orthogonale et Γ diagonale de termes diagonaux compris dans $[0\,;1]$,

Probabilités d'inclusion simple, double

- $\pi_k = pr(\{k\} \subseteq \mathbb{S}) = \det(K_{kk}) = K_{kk}$
- $\pi_{kl} = pr(\{k,l\} \subseteq \mathbb{S}) = \det(\begin{pmatrix} K_{kk} & K_{kl} \\ K_{lk} & K_{ll} \end{pmatrix}) =$

  $\det(\begin{pmatrix} K_{kk} & K_{kl} \\ \overline{K_{kl}} & K_{ll} \end{pmatrix}) = K_{kk}K_{ll} - \mid K_{kl} \mid^2 \quad (k \neq l)$
- $\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k\pi_l = - \mid K_{kl} \mid^2 (k \neq l) \\ \pi_k(1-\pi_k) = K_{kk}(1-K_{kk})(k=l) \end{cases}$;
- $\Delta = \overline{(I-K)} * K = (I-K) * \overline{K}$

où $*$ désigne le produit de Schur-Hadamard.

Probabilités d'inclusion simple, double

Conséquences sur les plans de sondage déterminantaux

- Ils vérifient les conditions de Sen-Yates-Grundy :

$$\Delta_{kl} = - \mid K_{kl} \mid^2 \leq 0 \text{ si } k \neq l$$

$$\Leftrightarrow$$

$$P(\{k,l\} \subseteq \mathbb{S}) \leq P(\{k\} \subseteq \mathbb{S}) P(\{l\} \subseteq \mathbb{S}),$$

- De manière plus générale, on montre que les plans déterminantaux
  - sont à association négative :
    $P(A \cup B \subseteq \mathbb{S}) \leq P(A \subseteq \mathbb{S}) P(B \subseteq \mathbb{S})$
  - vérifient la condition forte de Rayleigh.

Taille des échantillons d'un plan déterminantal $\mathcal{P}_K$

- **Théorème : Taille des échantillons**
  La taille de l'échantillon aléatoire : $\sharp\mathbb{S}$, a la même loi qu'une somme de N variables indépendantes de Bernouilli de paramètre λ_i, où les λ_i sont les valeurs propres de K.
- **Conséquence(s)** :
  - $var(\sharp\mathbb{S}) = \sum_{k,l \in U} \Delta_{kl} = \sum_{i=1}^{i=N} \lambda_i(1-\lambda_i) = tr(K - K^2)$
  - Un plan de sondage déterminantal est de taille fixe si et seulement si la matrice K est une matrice de projection orthogonale.

Autres propriétés d'un plan déterminantal $\mathcal{P}_K$

- $pr(\mathbb{S} = s) = \det(B(s))$ où $B(s)$ est la matrice carrée de taille N définie par $B_{kl} = K_{kl}$ si $k \in s$ et $B_{kl} = (I_N - K)_{kl}$ si $k \notin s$,
- Il existe un algorithme de tirage en n étapes.

Autres propriétés d'un plan déterminantal $\mathcal{P}_K$

- l'échantillon aléatoire $\mathbb{S}^c$, complémentaire de $\mathbb{S}$ dans U, a pour loi un plan de sondage déterminantal de matrice $I - K$.
- l'échantillon aléatoire $\mathbb{S}_A = \mathbb{S} \cap A$ pour loi un plan de sondage determinantal sur de matrice $K_{|A}$, restriction de K aux individus de A (Loi sur un domaine).

Plans de sondage particuliers

- **plans de taille fixe** : matrices de projection,
- **plans de probabilités d'inclusion fixées** : matrices diagonales (plan poissonnien),

Existe-t-il des plans déterminantaux de taille et de probabilités d'inclusion fixées ?

Existence et Construction

- **Existence :** Soit Π un vecteur de taille N tel que $0 \leq \Pi_k \leq 1 (k \in 1, \ldots, N)$ et $\sum_{k=1}^{N} \Pi_k = n \in \mathbb{N}$. Alors il existe un plan de sondage déterminantal de taille fixe n et de probabilités d'inclusions simples $\pi_k = \Pi_k (k \in 1, \ldots, N)$.
- **Construction :** Il existe des algorithmes permettant de construire des matrices de projection de diagonale fixée quelconque.
- Il existe des formules explicites lorsque les probabilités d'inclusion sont constantes.
- Si $n \notin \{0, 1, N, N-1\}$, le plan de sondage aléatoire simple n'est pas déterminantal.

Estimation d'un total

- Propriétés de l'estimateur,
- équilibrage et calage par optimisation.

Estimation d'un total : cas général

Un estimateur du total t_y est dit linéaire homogène s'il sécrit sous la forme :

$$\boxed{\hat{t}_{yw} = \sum_{k \in \mathbb{S}} w_k(\mathbb{S}) y_k}$$

Quand les poids $w_k(\mathbb{S})$ ne dépendent pas de $\mathbb{S}$, l'erreur quadratique moyenne de $\hat{t}_{yw}$ est :

$$\boxed{EQM(\hat{t}_{yw}) = \overbrace{y^T \delta_w \Delta \delta_w y}^{Variance} + \overbrace{[e^T(\delta_w \delta_\pi - I_N)y]^2}^{Biais}}$$

L'estimateur d'Horvitz-Thompson est un cas particulier avec $w_k = \frac{1}{\pi_k}$.

Estimation d'un total : cas déterminantal

$$EQM_K(\hat{t}_{yw}) = y^T \delta_w((I-K) * \overline{K})\delta_w y + [e^T(\delta_w(I * K) - I_N)y]^2$$

- **Théorème : Equilibrage parfait**
  A $\pi_k = \Pi_k$, $EQM_K(\hat{t}_{yw}) = 0$ si et seulement si K est une matrice de projection, $w_k = \pi_k^{-1}$ et π_k proportionnel à y_k par strate,
- $EQM_K(\hat{t}_{yw}) = g_y(K, w)$
- à $\pi_k = \Pi_k$ fixés et $w_k = \pi_k^{-1}$, $EQM_K(\hat{t}_{yw})$ est maximale pour le plan poissonnien ($K = \delta_\Pi$).

Estimation d'un total : cas déterminantal

- **propriétés asymtotiques :** sous des conditions classiques on montre que

$$\frac{\hat{t}_{yw}^N - E(\hat{t}_{yw}^N)}{\sqrt{Var(\hat{t}_{yw}^N)}} \xrightarrow{loi} \mathcal{N}(0,1).$$

- **Propriétés à distance finie :**

$$pr(|\hat{t}_{yw} - E(\hat{t}_{yw})| > a) \leq 5 \exp\left(-\frac{a^2}{16^2\,(aC + 2\mu C^2)}\right).$$

avec $C = \max_{k \in U} |w_k y_k|$ et $a > 0$.

Equilibrage et calage a priori par optimisation

Soient X matrice de Q variables auxiliaires : x^1 à x^Q, et g_X telle que $g(K, w) = 0$ si et seulement si $EQM_K(\hat{t}_{x^q w}) = 0$ pour tout q :

- **Equilibrage** $\leftrightarrow \underset{K \in \Theta_K}{\mathcal{M}in}\ g_X(K, w)$ (w : fixé)
- **Calage *a priori*** $\leftrightarrow \underset{w \in \Theta_w}{\mathcal{M}in}\ g_X(K, w)$ (K : fixée)
- **Equilibrage et calage *a priori*** $\leftrightarrow \underset{(K,w) \in \Theta}{\mathcal{M}in}\ g_X(K, w)$.

Par exemple,

$$g_X(K, w) = \sum_{q=1}^{q=Q} \alpha_q \frac{\sqrt{var(\hat{t}_{x^q w})}}{t_{x^q}} + \beta_q \frac{|biais(\hat{t}_{x^q w})|}{t_{x^q}}$$

Application

- Sélection de 9 Unités primaires parmi 19 dans la strate des agglomérations de 20 000 à 100 000 habitants de la région Rhône Alpes,
- trois jeux de probabilités d'inclusion
  - constantes $= \frac{n}{N}$,
  - proportionnelles au nombre de résidences principales Π_1,
  - optimales (seulement pour le déterminantal)
- stratégies d'équilibrage,
  - non équilibré,
  - équilibré sur la population, le nombre de Hlm, le nombre logements individuels.

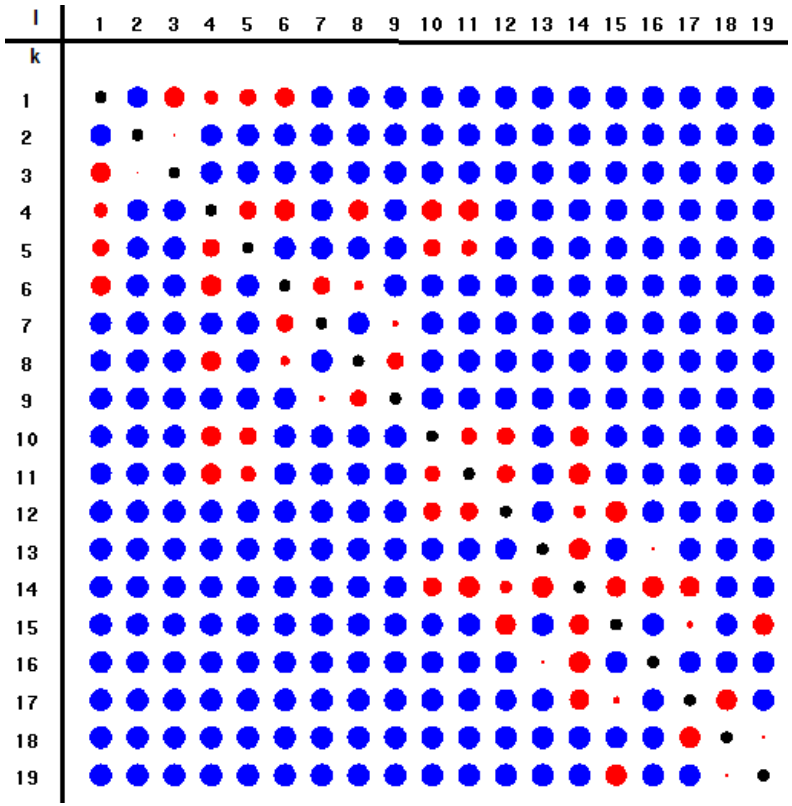
Application

TABLE: Précision moyenne des estimations selon la stratégie

méthode	π_k	Equilibrage	$g(X)$
SAS	$\frac{n}{N}$		0.6013106
Déterminant	$\frac{n}{N}$	π	0.5839211
Deville-Tillé	$\frac{n}{N}$	X,π	0.2667057
Déterminant	$\frac{n}{N}$	X	0.2523981
Sampford	Π_k^1		0.2241620
Déterminant	Π_k^1	π	0.2240632
Deville-Tillé	Π_k^1	X,π	0.1156012
Déterminant	Π_k^1	X	0.1130944
Déterminant	opt	X	0.1071314

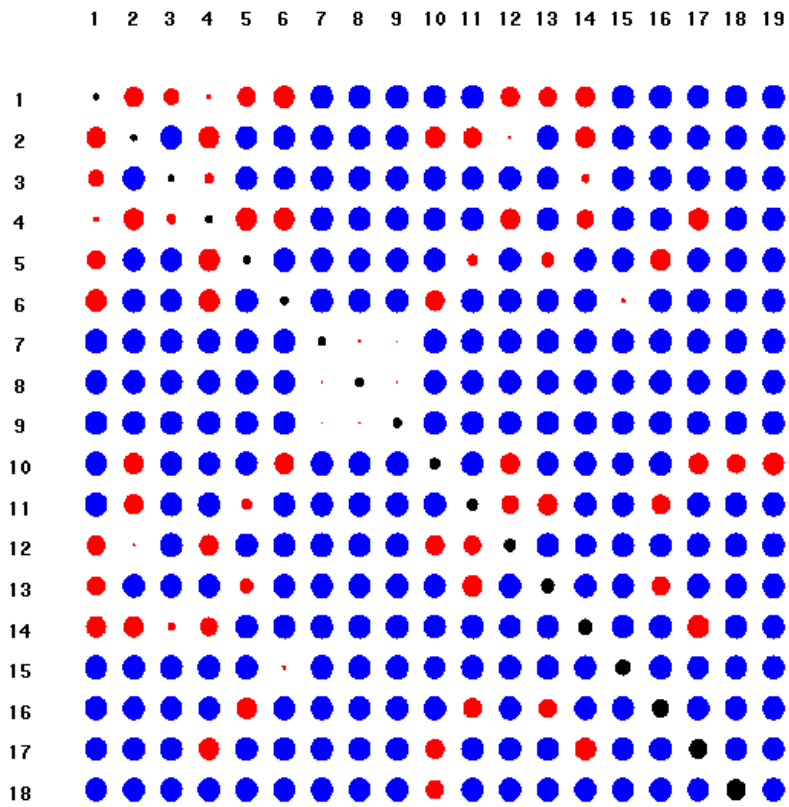
Plan déterminantal probabilités d'inclusion constantes, équilibré sur X

FIGURE: Disques proportionnels à π_k sur la diagonale (noir), à $\frac{\pi_{kl}}{\pi_k \pi_l}$ autrement : bleu $0.95 \leq \frac{\pi_{kl}}{\pi_k \pi_l} \leq 1$, rouge $0 \leq \frac{\pi_{kl}}{\pi_k \pi_l} \leq 0.95$. UP triées par taille.



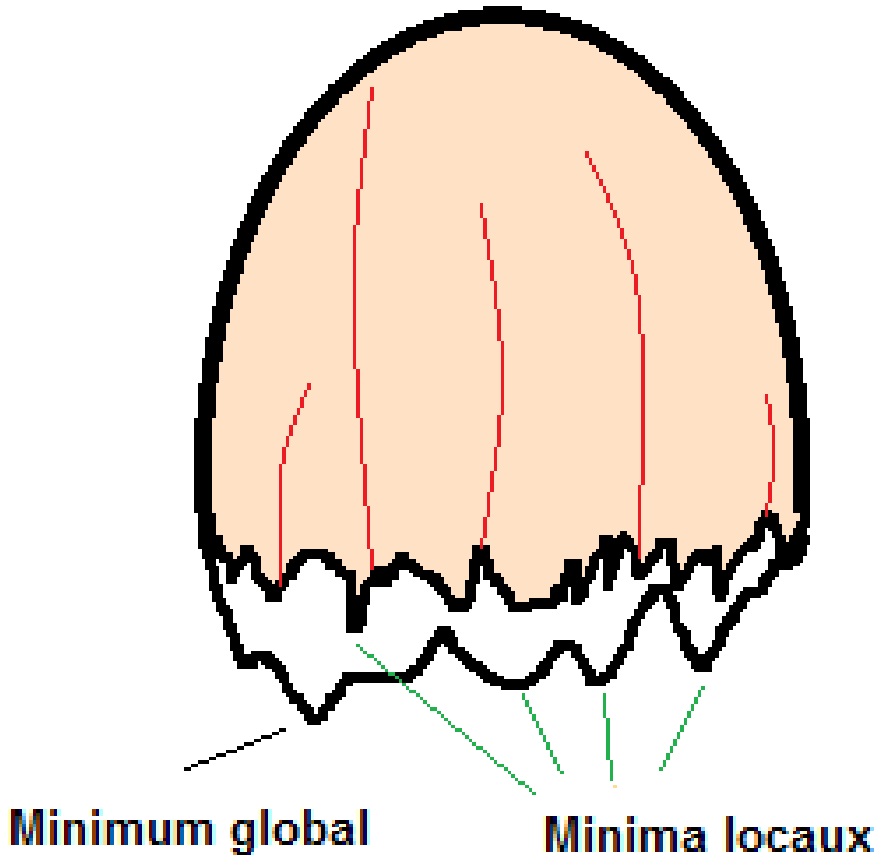
Plan déterminantal probabilités égale à Π_1, équilibré sur X

FIGURE: Disques proportionnels à π_k sur la diagonale (noir), à $\frac{\pi_{kl}}{\pi_k \pi_l}$ autrement : bleu $0.95 \leq \frac{\pi_{kl}}{\pi_k \pi_l} \leq 1$, rouge $0 \leq \frac{\pi_{kl}}{\pi_k \pi_l} \leq 0.95$. UP triées par taille.



Difficultés algorithmiques

Minimiser une fonction concave de très grande dimension ($\approx N^2$) avec des algorithmes de recherche de la plus grande pente sur un espace dont les points extrêmaux ont une structure complexe.



Merci pour votre attention ! !