Determinantal sampling design

SMURF Workshop: Survey Methods and their use in Related Fields
20 - 22 August 2018, Neuchâtel, Switzerland

Loonis Vincent[1], Mary Xavier[2]

[1] *Insee, Spatial Method Unit*
[2] *Université Paris Nanterre, Modal'X*

September 2, 2018

Insee
Mesurer pour comprendre

Introduction

- A point process on a finite discrete set U is exactly a sampling design, that is to say a probability law on $\mathcal{P}(U)$
- Among the various point processes, the Determinantal Point Process has attracted a lot of interest over the last years
- Due to its repulsiveness, Determinantal Point process is a good candidate for being fruitfully used in survey sampling theory.

Outline

This presentation mainly relies on Loonis, V. and Mary, X. (2018). Determinantal sampling designs. Journal of Statistical Planning and Inference.

1. Definition and general properties of determinantal sampling designs (DSDs)
2. Estimating a total
3. Constructing fixed-size DSDs with prescribed first-order inclusion probabilities
4. Constructing fixed-size DSDs with prescribed first and second order inclusion probabilities
5. Perspectives

Notations

- U size N poputation indexed by $k = 1, \ldots, N$
- s subset of U
- $\mathbb{S}$ random sample
- t_y, t_x total of a variable of interest, and of an auxiliary variable
- Π size N vector of prescribed probabilities
- z complex number, $\overline{z}$ its conjuguate, $|z|$ its modulus (resp. $A, \overline{A}$ for matrix A).
- λ vector of eigenvalues

Definition and general properties

- Definition
- Inclusion probabilities
- Sample size
- Sampling algorithm

Definition

Definition (Determinantal sampling design (Macchi (1975), Soshnikov (2000)))

A sampling design $\mathcal{P}$ on a finite set U is a determinantal sampling design if there exists a Hermitian contracting matrix K indexed by U, called kernel, such that for all $s \in 2^U$, $\sum_{s' \supseteq s} \mathcal{P}(s') = \det(K_{|s})$. This sampling design is denoted by $DSD(K)$. A random variable $\mathbb{S}$ with values in 2^U and law $DSD(K)$ is called a determinantal random sample (with kernel K). It satisfies, for all $s \in 2^U$,

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

where $K_{|s}$ denotes the submatrix of K. whose rows and columns are indexed by s. We will also write $\mathbb{S} \sim DSD(K)$.

Remark

Determinantal sampling designs form a parametric family of sampling designs, parametrized by contracting matrices.

Définition : example

$$K = \begin{pmatrix}
\frac{1}{2} & \frac{1}{\sqrt{10}} & \frac{\sqrt{3}}{2\sqrt{14}} & \frac{\sqrt{3}}{\sqrt{70}} & \frac{1}{\sqrt{35}} & \frac{1}{\sqrt{65}} & \frac{1}{2\sqrt{26}} \\
\frac{1}{\sqrt{10}} & \frac{1}{5} & \frac{\sqrt{3}}{2\sqrt{35}} & \frac{\sqrt{3}}{5\sqrt{7}} & \frac{\sqrt{2}}{5\sqrt{7}} & \frac{\sqrt{2}}{5\sqrt{13}} & \frac{1}{2\sqrt{65}} \\
\frac{\sqrt{3}}{2\sqrt{14}} & \frac{\sqrt{3}}{2\sqrt{35}} & \frac{3}{4} & -\frac{1}{2\sqrt{5}} & -\frac{1}{\sqrt{30}} & -\frac{\sqrt{7}}{\sqrt{390}} & -\frac{\sqrt{7}}{4\sqrt{39}} \\
\frac{\sqrt{3}}{\sqrt{70}} & \frac{\sqrt{3}}{5\sqrt{7}} & -\frac{1}{2\sqrt{5}} & \frac{4}{5} & -\frac{\sqrt{2}}{5\sqrt{3}} & -\frac{\sqrt{14}}{5\sqrt{39}} & -\frac{\sqrt{7}}{2\sqrt{195}} \\
\frac{1}{\sqrt{35}} & \frac{\sqrt{2}}{5\sqrt{7}} & -\frac{1}{\sqrt{30}} & -\frac{\sqrt{2}}{5\sqrt{3}} & \frac{2}{5} & \frac{2\sqrt{7}}{5\sqrt{13}} & \frac{\sqrt{7}}{\sqrt{130}} \\
\frac{1}{\sqrt{65}} & \frac{\sqrt{2}}{5\sqrt{13}} & -\frac{\sqrt{7}}{\sqrt{390}} & -\frac{\sqrt{14}}{5\sqrt{39}} & \frac{2\sqrt{7}}{5\sqrt{13}} & \frac{3}{5} & -\frac{1}{\sqrt{10}} \\
\frac{1}{2\sqrt{26}} & \frac{1}{2\sqrt{65}} & -\frac{\sqrt{7}}{4\sqrt{39}} & -\frac{\sqrt{7}}{2\sqrt{195}} & \frac{\sqrt{7}}{\sqrt{130}} & -\frac{1}{\sqrt{10}} & \frac{3}{4}
\end{pmatrix}.$$

$$pr(s = \{1\} \subseteq \mathbb{S}) = \det(K_{|1}) = \det(\frac{1}{2}) = \frac{1}{2} = \pi_1, \pi_3 = \frac{3}{4}, \pi_5 = \frac{2}{5}$$

$$pr(s = \{3;5\} \subseteq \mathbb{S}) = \det(K_{|\{3;5\}}) = \det \begin{pmatrix} \frac{3}{4} & -\frac{1}{\sqrt{30}} \\ -\frac{1}{\sqrt{30}} & \frac{2}{5} \end{pmatrix} = \frac{6}{20} - \frac{1}{30} = \frac{4}{15} = \pi_{35}$$

$$pr(s = \{1;3;5\} \subseteq \mathbb{S}) = \det(K_{|\{1;3;5\}}) = \det \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2\sqrt{14}} & \frac{1}{\sqrt{35}} \\ \frac{\sqrt{3}}{2\sqrt{14}} & \frac{3}{4} & -\frac{1}{\sqrt{30}} \\ \frac{1}{\sqrt{35}} & -\frac{1}{\sqrt{30}} & \frac{2}{5} \end{pmatrix} = \frac{8}{105} = \pi_{135}$$

Inclusion probabilities

Let $\mathbb{S} \sim DSD(K)$.

$$\pi_k = pr(k \in \mathbb{S}) = K_{kk}, \quad (1)$$

$$\pi_{kl} = pr(k, l \in \mathbb{S}) = K_{kk}K_{ll} - \mid K_{kl} \mid^2 \; (k \neq l), \quad (2)$$

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k \pi_l = - \mid K_{kl} \mid^2 \; (k \neq l), \\ \pi_k(1 - \pi_k) = K_{kk}(1 - K_{kk}) \, (k = l). \end{cases} \quad (3)$$

it holds that

$$\Delta = \overline{(I_N - K)} * K = (I_N - K) * \overline{K},$$

where $*$ is the Schur-Hadamard (entrywise) matrix product.

Remark

From (3) a determinantal sampling design satisfies the so-called Sen-Yates-Grundy conditions:

$$\pi_{kl} \leq \pi_k \pi_l \, (k \neq l).$$

Sample size

Theorem (Sample size (Hough et al. (2006)))

Let $\mathbb{S} \sim DSD(K)$. Then the random size $\sharp\mathbb{S}$ of the random variable $\mathbb{S}$ has the law of a sum of N independent Bernoulli variables $B_1, \cdots, B_N$ of parameters $\lambda_1, \cdots, \lambda_N$, set of K's eigenvalues.

Corollary (Sample size (2))

Let $\mathbb{S} \sim DSD(K)$. Then

1. $E(\sharp\mathbb{S}) = tr(K)$.
2. $var(\sharp\mathbb{S}) = tr(K - K^2) = \sum_{i=1}^{N} \lambda_i(1-\lambda_i) = \sum_{k,l \in U} \Delta_{kl}$.
3. *$DSD(K)$ is a fixed size determinantal sampling design iff K is a projection matrix.*

Sampling algorithm

Algorithm (Lavancier et al. (2015))

Let K be a projection matrix.

1. **Find a (N, n) matrix V such that $K = V\overline{V}^T$. Let v_k^T be the k^{th} line of V.**
2. *Sample one element k_n of U with probabilities $\Pi_k^n = ||v_k||^2/n$, $k \in U$.*
3. *Set $e_1 = v_{k_n}/||v_{k_n}||$.*
4. *For i = (n-1) to 1 do:*
    1. *sample one k_i of U with probabilities $\Pi_k^i = \frac{1}{i}[||v_k||^2 - \sum_{j=1}^{j=n-i} |\overline{e_j}^T v_k|^2]$, $k \in U$,*
    2. *set $w_i = v_{k_i} - \sum_{j=1}^{j=n-i} |\overline{e_j}^T v_{k_i}| e_j$ and $e_{n-i+1} = w_i/||w_i||$.*
5. *End for.*
6. *Return $\{k_1, \cdots, k_n\}$.*

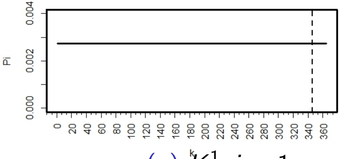
The resulting sample is a realization of $DSD(K)$.

Remark

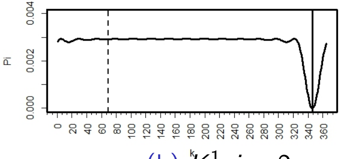
It is preferable to have a description of the matrix K directly in terms of V.

Sampling algorithm

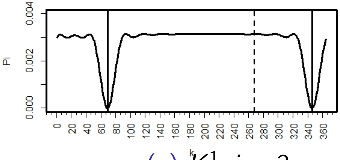
Figure 1: Example : Selecting an equal probability sample



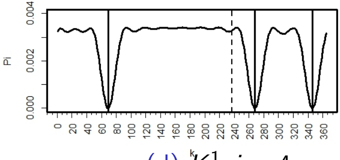
(a) $K^1, i = 1$



(b) $K^1, i = 2$



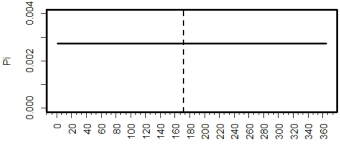
(c) $K^1, i = 3$



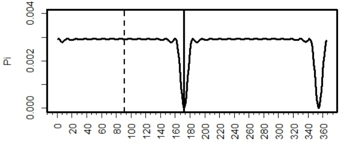
(d) $K^1, i = 4$

Sampling algorithm

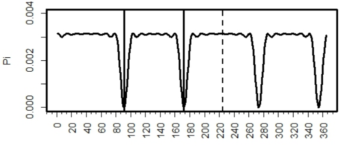
Figure 3: Example : Selecting an equal probability sample



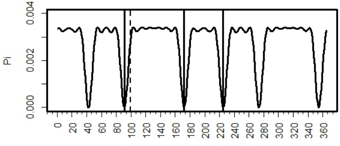
(a) $K^2, i = 1$



(b) $K^2, i = 2$



(c) $K^2, i = 3$



(d) $K^2, i = 4$

Estimating a total

1. Linear homogeneous estimators
2. Perfect estimation
3. Central limit theorem

Linear homogeneous estimators

Definition

Let $w_k, k \in U$ be N given reals, y be a variable of interest and $\mathcal{P}$ be a sampling design, then $\hat{t}_{yw} = \sum_{k \in \mathbb{S}} w_k y_k,$ with $\mathbb{S} \sim \mathcal{P}$ is a linear homogeneous estimator of t_y, whose Mean Square Error writes

$$\mathrm{MSE}(\hat{t}_{yw}) = \overbrace{\sum_{k \in U} \sum_{l \in U} w_k w_l y_k y_l \Delta_{kl}}^{\text{Variance}} + \left[\overbrace{\sum_{k \in U} (w_k \pi_k - 1) y_k}^{\text{Bias}} \right]^2$$

Example

- if $\pi_k > 0$ for all $k \in U$ and $w_k = \pi_k^{-1}$, $\hat{t}_{yHT} = \sum_{k \in \mathbb{S}} \pi_k^{-1} y_k$, is known as the Horvitz-Thompson estimator
- Let x be a strictly positive auxiliary variable, then $\hat{t}_{yw^{opt}}$, where $w_k^{opt} = (n x_k)^{-1} t_x$, will perfectly estimate t_x for a fixed size $\mathcal{P}$.

Optimal weights

Theorem (Optimal weights (Loonis and Mary (2018)))

Let $\mathcal{P}$ be a sampling design whose first and second order probabilities are π_k, π_{kl} ($\pi_{kk} = \pi_k$) and $x^1, \ldots, x^Q$ be Q vectors of auxiliary variables. The linear homogeneous estimators that minimize the sum of the Q MSEs correspond to weights w^{opt} in the affine subspace:

$$w^{opt} \in \left(\left(\sum_{q=1}^{Q} x^q {x^q}^T \right) * \Omega \right)^{\dagger} \left(\left(\sum_{q=1}^{Q} t_{x^q} x^q \right) * \pi \right) + \ker \left(\left(\sum_{q=1}^{Q} x^q {x^q}^T \right) * \Omega \right)$$

where $\Omega = (\pi_{kl})$ is the joint probability matrix of $\mathcal{P}$, π the vector of first order inclusion probabilities, and $M^{\dagger}$ the Moore-Penrose inverse of a matrix M.

Perfect estimation

Theorem (Perfect Estimation (Loonis and Mary (2018)))

Assume y takes only non-zero values and let $\mathbb{S} \sim DSD(K)$, then

$$\mathrm{MSE}(\hat{t}_{yHT}) = var(\hat{t}_{yHT}) \quad = \quad (diag(K)^{-1} * y)^{T}((I_N - K) * \overline{K})(diag(K)^{-1} * y)$$

and the total t_y is perfectly estimated by $\hat{t}_{yHT}$ ($var(\hat{t}_{yHT} = 0)$) iff $DSD(K)$ is a stratified determinantal sampling design of fixed size within each stratum, and with $\pi_k^{-1} y_k$ constant on each stratum.

Asymptotic theory

Theorem (Central Limit Theorem (Soshnikov (2002)))

Let $\mathbb{S} \sim DSD(K)$. Define for all $N \in \mathbb{N}$ the homogeneous linear estimators

$$\hat{t}_{yw} = \sum_{k \in \mathbb{S}} w_k y_k \text{ and } \hat{t}_{|y|w} = \sum_{k \in \mathbb{S}} w_k |y_k|$$

If the variance $var(\hat{t}_{yw}) \to +\infty$ as $N \to \infty$ and if

$$\sup_{k \in U_N} |w_k y_k| = o\left(var(\hat{t}_{yw})\right)^{\epsilon} \text{ and } E(\hat{t}_{|y|w}) = O\left(var(\hat{t}_{yw})\right)^{\delta}$$

for any $\epsilon > 0$ and some $\delta > 0$, then

$$\frac{\hat{t}_{yw} - E(\hat{t}_{yw})}{\sqrt{var(\hat{t}_{yw})}} \stackrel{law}{\to} \mathcal{N}(0,1).$$

Constructing a fixed size determinantal sampling design with prescribed first order inclusion probabilities

- General properties
- A closed form DSDs with any set of inclusion probabilities
  - Description and construction algorithm
  - Practical application (balancing on one variable, well spatially spread sampling).
- Going one step further with optimization routines.

General Properties

Constructing a fixed size determinantal sampling design with prescribed first order inclusion probabilities

- is equivalent to constructing a projection matrix with a prescribed diagonal,
- is a particular case of the more general issue of constructing Hermitian matrices with prescribed diagonal and spectrum.

General properties

A non-constructive proof of the existence of Hermitian matrices with prescribed diagonal and spectrum (in the context of DSDs)

Theorem (Schur (1911), Horn (1954))

Let Π and λ be two vectors of $[0,1]^N$ and $\Pi_{(k)}$ (resp. $\lambda_{(k)}$) denotes the k-th largest entry of Π (resp. λ), there exists a kernel K with diagonal Π and spectrum λ if and only if λ dominates Π

$$\begin{aligned} \sum_{k'=1}^{k'=k} \lambda_{(k')} &\geq \sum_{k'=1}^{k'=k} \Pi_{(k')} \text{ for all } k = 1, \ldots, N-1 \\ \sum_{k'=1}^{k=N} \lambda_k &= \sum_{k=1}^{k=N} \Pi_k \end{aligned}$$

General properties

A constructive proof of the existence of a real projection with prescribed diagonal, that nevertheless does not provide a closed form for the matrix.

Theorem (Kadison (2002))

Let Π a vector of $[0,1]^N$ such that $\sum_{k=1}^{N} \Pi_k = n \in \mathbb{N}^*$ there exists a fixed size DSD whose kernel is real with diagonal Π.

A closed form : description

Let Π be a vector of size N such that $0 < \Pi_k < 1$ and $\sum_{k \in U} \Pi_k = n \in \mathbb{N}^*$. For all integer r such that $1 \leq r \leq n$, let

- $1 < k_r \leq N$ be the integer such that $\sum_{k=1}^{k_r - 1} \Pi_k < r$ and $\sum_{k=1}^{k_r} \Pi_k \geq r$,
- $\alpha_{k_r} = r - \sum_{k=1}^{k_r - 1} \Pi_k$
- $\gamma_r^{r'} = \sqrt{\prod_{j=r+1}^{r'} \frac{(\Pi_{k_j} - \alpha_{k_j})\alpha_{k_j}}{(1-\alpha_{k_j})(1-(\Pi_{k_j} - \alpha_{k_j}))}}$ for $r < r'$, $\gamma_r^{r'} = 1$ otherwise.

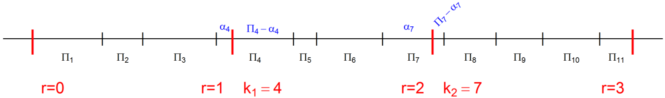


Figure 5: Example: $N = 11$ and $n = 3$.

A closed form : description

Define the real symmetric kernel P^{Π} as follows:

- for all $1 \leq k \leq N$, $P^{\Pi}_{kk} = \Pi_k$,
- for all $k > l$: P^{Π}_{kl} is computed according to formulas in table 1.

Table 1: Values of P^{Π}_{kl} : $k > l$

	Values of l	
Values of k	$l = k_r$	$k_r < l < k_{r+1}$
$k_{r'} < k < k_{r'+1}$	$-\sqrt{\Pi_k}\sqrt{\frac{(1-\Pi_l)(\Pi_l-\alpha_l)}{1-(\Pi_l-\alpha_l)}}\gamma_r^{r'}$	$\sqrt{\Pi_k \Pi_l}\gamma_r^{r'}$
$k = k_{r'+1}$	$-\sqrt{\frac{(1-\Pi_k)\alpha_k}{1-\alpha_k}}\sqrt{\frac{(1-\Pi_l)(\Pi_l-\alpha_l)}{1-(\Pi_l-\alpha_l)}}\gamma_r^{r'}$	$\sqrt{\frac{(1-\Pi_k)\alpha_k}{1-\alpha_k}}\sqrt{\Pi_l}\gamma_r^{r'}$

Theorem (Loonis and Mary (2018))

The matrix P^{Π} is a real projection matrix, and $DSD(P^{\Pi})$ is a fixed size sampling design with first order inclusion probabilities $\pi_k = \Pi_k, 1 \leq k \leq N$.

A closed form : description

Corollary

Let P^{Π} be the matrix previously constructed, and $DSD(P^{\Pi})$ the associated sampling design.

1. *If $(k, l) \in]k_r, k_{r+1}[^2$ then $\pi_{kl} = 0$.*
2. *If $j \in]k_r, k_{r+1}[$, $k = k_{r+1}$, $l \in]k_{r+1}, k_{r+2}[$ then $\pi_{jkl} = 0$.*
3. *Set $B_r = [1, k_r + 1]$. Then the random sample $\mathbb{S}$ has r or $r+1$ elements in B_r.*
4. *If $k - l$ is large then $P^{\Pi}_{kl} \approx 0$, and $\pi_{kl} \approx \Pi_k \Pi_l$.*
5. *Let $r_1, \ldots, r_H$ be the set of values of $1 \leq r \leq n$ such that $\sum_{k=1}^{k_r} \Pi_k = r$, and set $r_0 = 0$. Then $DSD(P^{\Pi})$ is stratified with H strata $]k_{r_{h-1}}, k_{r_h}]$.*

Figure 6: Examples of unfeasible samples $\mathbb{S} \sim DSD(P^{\Pi})$, $n = 3$, $N = 11$

(a) impossible according to Point 1

(b) impossible according to Point 2

A closed form : Construction algorithm

Algorithm (Rank one decomposition for P^{Π})

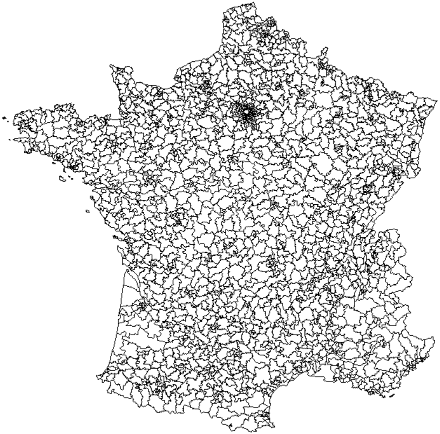
1. *For* $k \in U$, *let* s_k *and* c_k *be*
   - *if* $\exists r | k = k_r$, $s_{k_r} = \sqrt{\frac{1-\Pi_{k_r}}{1-\alpha_{k_r}}}$, $s_k = \sqrt{\frac{\Pi_k}{r+1-\sum_{i=1}^{k-1}\Pi_i}}$ *otherwise*
   - $c_k = \sqrt{1 - s_k^2}$
2. *Let* V *be a* (N, n) *matrix whose entries equal 0 apart from* $V_{k_r+1,r}$ $(r = 0, \ldots, n-1)$ *that equals* 1. *Let* V_k^T *be the* k^{th} *line of* V.
3. *For* $k = 1, \ldots, N-1$
   1. *Compute* $L_1^T = s_k V_k^T - c_k V_{k+1}^T$
   2. *Compute* $L_2^T = c_k V_k^T + s_k V_{k+1}^T$
   3. *Replace* V_k^T *(resp.* V_{k+1}^T*) by* L_1^T *(resp* L_2^T*).*

Remark

Using SAS, V *is computed in less than 9 seconds with* $N = 100\ 000$ *et* $n = 1\ 000$.

Practical application

- The French master sample for household surveys is drawn according to a two-stage sampling design.
- We consider the first stage that consists of thousands of geographical entities (PSUs) with proportional to size inclusion probabilities.
- We aim at drawing a sample that is balanced on one auxiliary variable or that is spatially well spread.



Balancing on one auxiliary variable

- Let N=4000, and n=$30, 60, \ldots, 630$
- Let $\Pi_k = n\frac{d_k}{t_d}$ where d_k is the number of dwellings in PSU K.
- Let x_k be an auxiliary variables for PSU k (total amount of wages for example).
- Let U (population of PSUs) be sorted by $\frac{x_k}{\Pi_k}$.
- We consider three different sampling designs, $DSD(P^{\Pi})$, Cube method, and systematic sampling.
- We compute for each sampling design and each sample size $CV(\hat{t}_x) = \frac{\sqrt{V(\hat{t}_x)}}{t_x}$, where $\hat{t}_x$ is the Horvitz-Thompson estimator of t_x.
- For $DSD(P^{\Pi})$, $V(\hat{t}_x)$ is known exactly whereas it is estimated with Monte Carlo methods for the other methods.

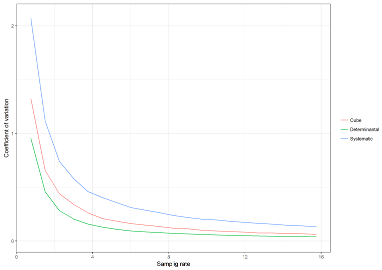
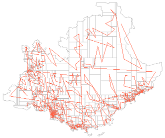


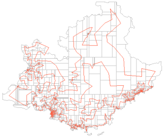
Figure 8: $DSD(P^{\Pi})$ performs better than its *opponents*.

Spatial Determinantal sampling

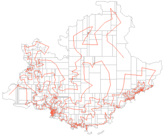
- GRTS is a well known spatial sampling method.
- It consists in drawing a path through the location of the units and selecting the units along the path according to a systematic sampling.
- There exists various ways to construct such a path (GRTS, TSP, Hamilton...)
- We suggest ordering the units according to the path and selecting the units with a $DSD(P^{\Pi})$
- We compute the variance of the HT-estimator for several auxiliary variables with a Moran-Index ranging from 0.1 to 0.8.



(a) GRTS path

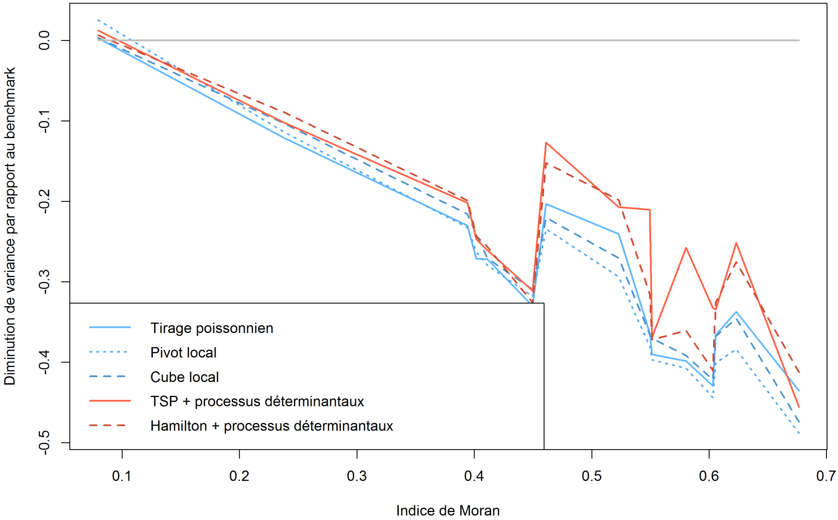


(b) TSP path



(c) Hamilton path

Spatial determinantal sampling



Going one step further

- P^{Π} proves useful for balanced sampling (one auxiliary variable) or for spatial sampling
- What about balancing on more than one auxiliary variable, or achieving other goals ?
- Let $C(K)$ be a criteria to be minimized subject to K being a contracting matrix with at least a given trace, for instance:

$$C(K) = \sum_{q=1}^{Q} V(\hat{t}_{x^q})$$

- Solving min $C(K)$ falls within the scope of non-linear semi-definite optimization that can be tough.
- We aim at finding heuristics relying on P^{Π}.

Going one step further

To do so, the following well known result proves very uselful

Proposition (Unitary transform)

Let $K \in \mathcal{M}_{N\times N}(\mathcal{C})$ be a contracting matrix and $\mathbb{S} \sim DSD(K)$. Let also $W \in \mathcal{M}_{N\times N}(\mathcal{C})$ be a unitary matrix ($W\overline{W}^T = I_N$). Then $K_W = WK\overline{W}^T$ is a Hermitian matrix with the same eigenvalues as K.

Remark

K_w has not necessarily the same diagonal entries as K.
Let $W(\rho)$ be a large enough parametrized family of unitary matrices, solving $\min C(K)$ can be approximated by solving $\min C(K_{W(\rho)})$.

Going one step Further

- We aim at minimizing $C(K) = \sum_{q=1}^{Q} V(\hat{t}_{x^q})$ subject to prescribed inclusion probabilities.
- We consider an ordered population U and the associated P^{Π}
- We use the following unitary matrix

$$W_{kl}(\theta) = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \ddots & 0 & 0 & 0 & \vdots & 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \cos(\theta) & 0 & \dots & 0 & -\sin(\theta) & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \vdots & 0 & 0 & 0 & \ddots & 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \sin(\theta) & 0 & 0 & 0 & \cos(\theta) & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \dots & 0 \\ 0 & 0 & \vdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Going one step further

- For any (k, l) and any θ in U^2, $W_{kl}(\theta)P^{\Pi}W_{kl}^T(\theta)$ is a projection matrix as well (The spectrum remains unchanged).
- If $\Pi_k \neq \Pi_l$ choosing θ_{kl} such that

$$t = \frac{2P_{kl}^{\Pi}}{K_{kk} - K_{ll}}, \cos\theta_{kl} = \frac{1}{\sqrt{1+t^2}} \textit{ and } \sin\theta_{kl} = t\cos\theta_{kl}.$$

does not change the diagonal either (Dhillon et al. (2005)).

Algorithm

- *Let* $K^0 = P^{\Pi}$
- *for* $r = 1$ *to* R *(fixed in advance) do:*
  1. *For each* (k, l) *in* U^2 *such that* $\Pi_k \neq \Pi_l$ *compute* θ_{kl}^r
  2. *Define* $(k^r, l^r) = \underset{(k,r)\in U^2}{\mathrm{argmin}}\ C(W_{kl}(\theta_{kl}^r)K_j^{r-1}W_{kl}^T(\theta_{kl}^r))$;
  3. *Set* $K^r = W_{k^r l^r}(\theta_{k^r l^r}^r)K^{r-1}W_{k^r l^r}^T(\theta_{k^r l^r}^r)$ *and* $r = r + 1$.

Going one step further

Implementing the previous algorithm with $N = 148$ PSUs and $n = 14$ for $Q = 2$ auxiliary variables x^1 and x^2 (total amount of unemployement benefits and of taxable income).

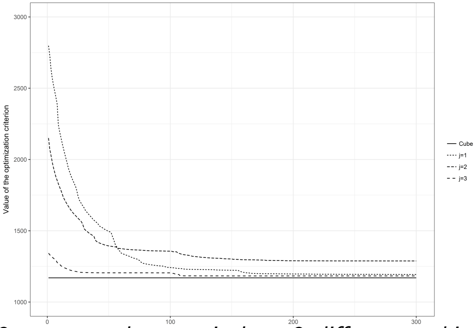


Figure 10: Evolution of the optimization criterion, DSDs perform as good as the cube.

The curves $j = 1, 2, 3$ correspond respectively to 3 different ranking methods: by x_k^2/Π_k ($j = 1$), by $(x_k^2 + x_k^3)/\Pi_k$ ($j = 2$) and by the Hamilton path ($j = 3$).

Constructing fixed-size DSDs with prescribed first and second order inclusion probabilities

- Apart from $n = 1$ or $n = N - 1$, SRS is not a determinantal sampling design.
- Does there exist K such that $DSD(K)$ has the same first and second order probabilities as a SRS ? That is to say ,
  - K is a projection matrix,
  - $\pi_k = K_{kk} = \frac{n}{N}$
  - $\pi_{kl} = K_{kk}K_{kk} - |K_{kk}|^2 = \frac{n(n-1)}{N(N-1)} \iff |K_{kk}|^2 = \frac{n(N-n)}{N^2(N-1)}$.
- Finding such a kernel is equivalent to finding an Equiangular Tight Frame (ETF).
- Many results on the existence of such frames are available in the corresponding literature.

Constructing fixed-size DSDs with prescribed first and second order inclusion probabilities

Table 2: Existence of (N, n)-simple determinantal sampling designs, depending on the kernel type (real or complex) for $n < 9$.

n	3	3	4	4	5	5	6	6	6	7	7	7	8	8	8
N	6	7	7	13	10	11	11	16	31	14	15	28	15	29	57
	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{C}$	$\mathbb{C}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{C}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{C}$	$\mathbb{C}$

1. For a given family of (non-determinantal) sampling designs, there may or may not exist a DSD with the same first and second order inclusion probabilities ;
2. There exists a $DSD(C)$, C complex kernel such that no $DSD(R)$, R real kernel has the same first and second order inclusion probabilities. This plaids in favor of using complex kernels.

Perspectives

- Implementing the selection algorithm efficiently;
- Delving deeper into the properties of P^{Π};
- Delving deeper into complex kernels;
- Delving deeper in optimization algorithm.

THANK YOU FOR YOUR ATTENTION

References I

Dhillon, I. S., Heath Jr, R. W., Sustik, M. A., and Tropp, J. A. (2005). Generalized finite algorithms for constructing hermitian matrices with prescribed diagonal and spectrum. *SIAM Journal on Matrix Analysis and Applications*, 27(1):61–71.

Horn, A. (1954). Doubly stochastic matrices and the diagonal of a rotation matrix. *American Journal of Mathematics*, 76(3):620–630.

Hough, J. B., Krishnapur, M., Peres, Y., Virág, B., et al. (2006). Determinantal processes and independence. *Probability surveys*, 3:206–229.

Kadison, R. V. (2002). The pythagorean theorem: I. the finite case. *Proceedings of the National Academy of Sciences*, 99(7):4178–4184.

Lavancier, F., Møller, J., and Rubak, E. (2015). Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):853–877.

References II

Loonis, V. and Mary, X. (2018). Determinantal sampling designs. *Journal of Statistical Planning and Inference*.

Macchi, O. (1975). The coincidence approach to stochastic point processes. *Advances in Applied Probability*, pages 83–122.

Schur, J. (1911). Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und Angewandte Mathematik*, 140:1–28.

Soshnikov, A. (2000). Determinantal random point fields. *Russian Mathematical Surveys*, 55(5):923–975.

Soshnikov, A. (2002). Gaussian limit for determinantal random point fields. *Annals of probability*, pages 171–187.