Sous-échantillonnage aléatoire par processus déterminantaux pour l'estimation

Pierre-Olivier AMBLARD[1] * et Jean-François CŒURJOLLY[2]

[1]GIPSAlab, UMR CNRS 5216
Saint Martin d'Hères, France

[2]Dépt. de Statistiques, Université de Montréal, Canada

pierre-olivier.amblard@gipsa-lab.grenoble-inp.fr, coeurjolly.jean-francois@uqam.ca

Résumé – Ce papier s'intéresse à des stratégies de sous-échantillonnage aléatoire pour estimer des grandeurs statistiques de grandes masses de données. Le sous-échantillonnage déterminantal est utilisé pour apporter de la diversité par rapport à des échantillonnage de type Poissonien. La diversité est un corollaire du caractère répulsif des processus ponctuels déterminantaux. Nous détaillons particulièrement l'estimation de grandeurs statistiques comme des moments ou des fonctions de corrélation. Nous montrons dans quelles situations un sous-échantillonnage déterminantal peut être bénéfique. Dans le cadre du sous-échantillonnage indépendant des observations, nous illustrons sur un exemple simple les gains potentiels qui peuvent être obtenus.

Abstract – We investigate in this paper random subsampling to estimate statistics in very large data sets. Determinantal subsampling is used for the diversity offered compared to Poissonian subsampling. We particularly study the estimation of moments or correlation functions. We exhibit the conditions under which determinantal subsampling is a better strategy than Poissonian subsampling. If the sampling is independent from the data, we illustrate on a very simple example the potential benefits of determinantal subsampling.

1 Introduction

Le problème étudié ici est proche du problème standard en théorie des sondages, à savoir sous-échantillonner une population et estimer à partir du sous-échantillon une grandeur caractéristique. Nous nous intéressons à ce problème pour des situations dans lesquelles le nombre d'échantillons est gigantesque et ne peut pas être traité en utilisant l'ensemble de la population initiale.

Ce papier s'intéresse à la compression de l'information contenue dans un très grand échantillon, non pas à des fins de reconstruction de l'échantillon global, mais plutôt à des fins d'inférence. Si l'échantillon initial sert à l'inférence d'un paramètre θ, un but est de compresser l'échantillon le plus possible tout en gardant un maximum d'information sur θ. Ce point de vue est celui adopté dans la méthode de l'$\ll$ information bottleneck $\gg$[10] , mais également par exemple dans les techniques de sketching [12], ou encore dans les théories de type PAC (Provably Approximately Correct).

Proches du sketching, les méthodes de sous-échantillonnage aléatoire ont démontré dans les deux dernières décennies leur puissance pour résoudre des problèmes rendus difficiles pas la taille des données [3, 9], comme par exemple dans l'implantation du produit de matrices de grande taille ou la recherche d'approximation de rang faible. Dans ces techniques, l'échantillonnage aléatoire consiste à représenter une matrice par un sous-ensemble aléatoire de ses colonnes, en utilisant une probabilité discrète. Les approches développées utilisent des tirages *indépendants* de colonnes. Les probabilités utilisées sont souvent optimisées, et reposent essentiellement sur la notion de $\ll$ statistical leverage scores $\gg$, liés à la géométrie de l'espace image des matrices. Lorsque certains de ces scores sont élevés, les colonnes ou lignes associées ont tendance a être sur-échantillonnées. Il s'agit ici d'un manque typique de diversité qui apparaît dans les techniques d'échantillonnage avec remise. Pour échantillonner avec plus de diversité, l'idée est d'introduire de la dépendance négative entre les échantillons. Des techniques heuristiques sont connues comme par exemple l'échantillonnage antithétique ou le $\ll$ herding $\gg$[2].

Le regain d'intérêt récent pour les processus déterminantaux en statistique et apprentissage repose précisément sur leur caractère répulsif qui en font des candidats de choix pour sonder aléatoirement des espaces avec de la diversité [1, 5, 7]. Nous les utilisons ici pour extraire aléatoirement d'un ensemble de données un échantillon qui se veut aussi représentatif que possible de l'ensemble, au sens de l'estimation de grandeurs statistique.

Précisément, on considère l'observation de N variables aléatoires x_i, $i = 1, \ldots, N$ prenant leurs valeurs dans $\mathbb{R}^d$, *identiquement distribuées*. On souhaite estimer des statistiques du type $\mathrm{E}[h(x)]$ où h est une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$. Typiquement, ce formalisme nous permet d'estimer des fonctions de corrélation à tout ordre de signaux aléatoires. Par exemple, pour un signal s_t, considérer $x_i = (s_i, s_{i+k})$ et h définie sur $\mathbb{R}^2$ par $h(y_1, y_2) = y_1 y_2$ permet d'estimer une fonction de corrélation. Notons de suite que l'ensemble des résultats qui suivent se transcrivent sans difficulté à des fonctions vectorielles $h : \mathbb{R}^d \to \mathbb{R}^p$, mais que les interprétations simples que nous donnerons seraient un peu dissimulées dans ce cas général. Pour cette raison, on se restreint ici à $p = 1$. Dans la section 2, nous donnons quelques éléments sur les processus déterminantaux sur un espace discret. L'estimation utilisant un sous-échantillon est développé en 3.

*Travail soutenu par le LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) et le LIA CNRS/MelbourneUni Geodesic.

Le cas général d'un sous-échantillonnage dépendant des observations étant trop difficile théoriquement, nous présentons le cas particulier du sous-échantillonnage indépendant dans le paragraphe 4. L'avantage d'utiliser des processus déterminantaux par rapport à des processus de Poisson est mis en lumière par un exemple simple.

2 Processus ponctuels sur un espace discret

L'étape de sous-échantillonnage peut être modélisée comme composition du signal x_i par une réalisation S d'un processus ponctuel, soit x_S. On donne dans le suite quelques éléments sur les processus ponctuels et les idées sous-tendant les processus déterminantaux.

Généralités. Un processus ponctuel défini sur l'ensemble discret $\mathcal{U} = \{1, \dots, N\}$ est une mesure de probabilité sur $2^{\mathcal{U}}$, l'ensemble des sous-ensembles de $\mathcal{U}$. Soit P une telle mesure. Alors tout sous-ensemble s de $\mathcal{U}$ a la probabilité $P(s)$ d'être observée. On définit un échantillon S de $\mathcal{U}$ comme étant une variable aléatoire à valeur dans $2^{\mathcal{U}}$ avec comme mesure de probabilité P, c'est-à-dire $\Pr(S = s) = P(s)$.

Une description équivalente est donnée par les probabilités d'inclusion (équivalents discrets des fonctions intensités pour les processus ponctuels à espaces d'états continus). Considérons un échantillon S de $\mathcal{U}$. La probabilité d'inclusion d'ordre 1 est la probabilité qu'un élément i de $\mathcal{U}$ appartienne à S, soit $\pi_i = \Pr(i \in S) = \sum_{s/i \in s} P(s)$ De même, la probabilité d'inclusion d'ordre n est la probabilité qu'un sous-ensemble s de taille n appartienne à S, c'est-à-dire $\pi_s = \Pr(s \in S) = \sum_{s'/s \in s'} P(s')$ Si les éléments s sont notés $s_1, \dots, s_n$, on note $\pi_{s_1 \dots s_n}$ la probabilité d'inclusion d'ordre n.

On peut décrire le processus par une suite de variables aléatoires de Bernoulli ε_i de moyenne π_i. La covariance à deux indices s'écrit $\text{Cov}\,[\varepsilon_i, \varepsilon_j] = \pi_{ij} - \pi_i \pi_j$. Pour les processus *négativement associés* tels que $\pi_{ij} < \pi_i \pi_j$ pour tout (i, j), la covariance est négative et traduit un caractère répulsif. Si $i \in S$, j a tendance à ne pas être inclus. A l'opposé, la covariance est positive pour des processus *positivement associé* si $\pi_{ij} > \pi_i \pi_j$ pour tout (i, j). Dans ce cas, le processus présente des aggrégats puisque si $i \in S$, j a tendance à être inclus également. La frontière entre ces classes de processus est donnée par le processus de Poisson, pour lequel les inclusions de deux éléments distincts sont des évènements indépendants, et donc $\pi_{ij} = \pi_i \pi_j$ pour tout (i, j), conduisant à une covariance nulle.

Processus déterminantaux, un archétype de processus négativement associés. Considérons une matrice K de dimension N symétrique, définie positive et contractante ($K \leq I$). On note K_s la sous matrice de K correspondant aux lignes et colonnes $s_1, \dots, s_n$ des éléments de s. De plus, avec abus de notation, on note également K_{ij} l'élément (i, j) de K.

Un processus ponctuel sur $\mathcal{U}$ est déterminantal s'il existe une telle matrice K de sorte que $\forall n \leq N, \forall s = (s_1, \dots, s_n), \Pr(s \in S) = \text{Det}\, K_s$.

En particulier, la probabilité d'inclusion d'ordre 1 est $\pi_i = K_{ii} = \pi_{ii}$; la probabilité d'inclusion d'ordre 2 est $\pi_{ij} = K_{ii}K_{jj} - |K_{ij}|^2$. Comme K est définie positive, $\pi_{ij} \leq \pi_{ii}\pi_{jj}$ qui signe le caractère répulsif du processus. Plus de détails se trouvent dans [8, 4, 5, 6]. On s'intéresse ici au cas où le processus ponctuel est doublement stochastique, c'est-à-dire que le processus ponctuel ne dépend des données qu'à travers les probabilités d'inclusion, soit $\Pr(i \in s |\, x) = \pi_i(x_i)$ et $\Pr((i, j) \in s |\, x) = \pi_{ij}(x_i, x_j)$.

3 Estimateurs

Considérons maintenant divers estimateurs de $\text{E}[h(x)]$. Le plus naturel est l'estimateur empirique que l'on notera C_N. Nous étudions l'influence d'un sous échantillonnage d'un facteur α. On notera C_π l'estimateur utilisant un processus ponctuel de probabilité d'inclusion π. La notation sera alors déclinée en C_P pour un sous échantillonnage de Poisson, C_D pour un échantillonnage déterminantal. Dans le cas de signaux, *i.e.* les x_i correspondent aux échantillons successifs d'un signal temporel à temps discret, on notera C_α l'estimateur utilisant un sous-échantillonnage périodique de facteur α.

Les performances sont examinées dans un régime asymptotique pour lequel le nombres d'échantillons initiaux N et sous-échantillonnés M tendent vers l'infini mais à *ratio* $\alpha = N/M$ constant. Ce régime n'est en général pas considéré dans la littérature statistique sur les sondages (≪ survey sampling ≫), voir par exemple [7], qui considère en général que N et M tendent vers l'infini avec le *ratio* $\alpha = N/M$ tendant vers l'infini.

Nous envisageons des estimateurs qui s'écrivent sous la forme

$$C_\pi = \frac{1}{N} \sum_{i=1}^{N} \frac{h(x_i)\varepsilon_i}{\pi_i}$$

L'estimateur empirique usuel C_N s'obtient avec $\pi_i = 1, \forall i = 1, \dots, N$; l'estimateur de Poisson avec $\pi_{ij} = \pi_i \pi_j$; l'estimateur empirique sur M points périodiques avec $\pi_{kN/M+1} = 1, \forall k = 0, \dots, M-1$ et 0 sinon.

L'estimateur C_π est non biaisé puisque

$$\text{E}\big[C_\pi\big] = \frac{1}{N} \sum_{i=1}^{N} \text{E}\Big[\frac{h(x_i)}{\pi_i} \text{E}\big[\varepsilon_i \big| x_i\big]\Big] = \frac{1}{N} \sum_{i=1}^{N} \text{E}\Big[\frac{h(x_i)}{\pi_i}\pi_i\Big] = \text{E}[h(x)]$$

La variance de C_π se calcule aisément en utilisant $\text{Cov}\,[X, Y] = \text{E}[\text{Cov}\,[X, Y|Z]] + \text{Cov}\,[\text{E}[X|Z], \text{E}[Y|Z]]$. On obtient alors

$$\begin{aligned} \text{Var}[C_\pi] &= \frac{1}{N^2} \sum_{i,j} \text{Cov}\,\Big[\frac{h(x_i)\varepsilon_i}{\pi_i}, \frac{h(x_j)\varepsilon_j}{\pi_j}\Big] &(1)\\ &= \text{Var}[C_N] + \frac{1}{N^2} \sum_{i,j} \text{E}\Big[h(x_i)h(x_j)\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\Big] &(2)\end{aligned}$$

où $\pi_{ii} = \pi_i$ et $\text{Var}[C_N] = N^{-2} \sum_{i,j} \text{Cov}\,\Big[h(x_i), h(x_j)\Big]$. L'équation (2) exhibe la perte de performance d'un sous-échantillonnage par rapport à l'utilisation de toutes les variables observées. L'intérêt du résultat (2) est toutefois mineur si on ne compare pas différentes méthodes de sous-échantillonnage. En particulier, la méthode de sous-échantillonnage aléatoire la plus simple à mettre en œuvre est celle pour laquelle les points du processus ponctuel (ou les ε_i) sont indépendants. Dans ce cas on a affaire à un processus de Poisson discret, et puisqu'alors $\pi_{ij} = \pi_i \pi_j$ on obtient

$$\text{Var}[C_P] = \text{Var}[C_N] + \frac{1}{N^2} \sum_{i} \text{E}\Big[h(x_i)^2 \frac{1 - \pi_i}{\pi_i}\Big] \qquad (3)$$

de sorte que l'on peut décomposer pour des sous-échantillonnages non Poissonien la variance d'estimation en

$$\text{Var}[C_\pi] = \text{Var}[C_P] + \frac{1}{N^2} \sum_{i \neq j} \text{E}\Big[h(x_i)h(x_j)\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\Big] \qquad (4)$$

Dans cette forme par contre, le terme complémentaire peut être négatif ou positif selon les propriétés des processus stochastiques x et ε et de la fonction h considérée.

Un premier résultat évident est le suivant. Si h est à valeur dans $\mathbb{R}^+$ (resp. $\mathbb{R}^-$), un processus d'échantillonnage à association négative

(resp. association positive) est meilleur (au sens de la variance d'estimation de $\mathrm{E}[h(x)]$) que le sous-échantillonnage de Poisson. Ce résultat n'est pas complètement inutile puisqu'il s'applique directement si l'on estime un moment d'ordre pair de x par exemple. Par contre, il est difficile d'obtenir des résultats généraux dans des cas aussi simples que l'estimation d'une fonction de corrélation, cas pour lequel $h : \mathbb{R}^2 \to \mathbb{R}$. Le cas de l'échantillonnage indépendant (pour lequel π ne dépend pas de x) permet toutefois d'obtenir des résultats intéressants, et parfois surprenants.

4 Sous-échantillonnage indépendant des observations

Le cas général d'un sous-échantillonnage dépendant des observations est très difficile à étudier dans un cadre général. On réduit ici la difficulté en autorisant le sous-échantillonnage à ne pas dépendre des variables. La variance pour une loi d'échantillonnage π s'écrit dans ce cas

$$\mathrm{Var}[C_\pi] = \mathrm{Var}[C_P] + \frac{1}{N^2}\sum_{i\neq j}\mathrm{E}\Big[h(x_i)h(x_j)\Big]\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} \quad (5)$$

$$\mathrm{Var}[C_P] = \mathrm{Var}[C_N] + \frac{1}{N^2}\sum_{i}\mathrm{E}\Big[h(x_i)^2\Big]\frac{1-\pi_i}{\pi_i} \quad (6)$$

Probabilité d'inclusion optimale. L'excès de variance par rapport à C_N en sous-échantillonnage Poissonien est donné au facteur N^{-1} près par

$$\sum_i \mathrm{E}\Big[h(x_i)^2\Big]\frac{1-\pi_i}{\pi_i} = \sum_i \frac{\mathrm{E}\Big[h(x_i)^2\Big]}{\pi_i} - \sum_i \mathrm{E}\Big[h(x_i)^2\Big]$$

terme qui peut être minimisé sous la contrainte d'un nombre moyen d'échantillons $M = \sum_i \pi_i$ donné. On montre alors aisément que l'excès de variance est minimisé pour

$$\pi_i^\star = M\mathrm{E}\Big[h(x_i)^2\Big]^{1/2}\Big/\sum_i E\Big[h(x_i)^2\Big]^{1/2}$$

Comme nous avons supposé les x_i identiquement distribués, on obtient $\pi_i^\star = M/N = 1/\alpha$ et $\mathrm{Var}[C_P] = \mathrm{Var}[C_N] + (\alpha - 1)/N\mathrm{E}[h(x)^2]$. Notons que $\pi_i^\star$ est constant, conclusion naturelle puisque les variables sont supposées identiquement distribuées.

Cas de données issues d'un signal stationnaire. Supposons que les x_i soient les échantillons d'une séquence stationnaire. Alors $\mathrm{E}\Big[h(x_i)h(x_j)\Big]$ est une fonction de $i-j$ seulement. Appelons-la f^h_{i-j}. Soit $\widetilde{f}^h_{i-j}$ la covariance associée. D'après le paragraphe précédent, la probabilité d'inclusion optimale dans le cas Poissonien est donnée par $\pi_i^\star = M/N$, et puisque nous souhaitons mesurer l'intérêt des corrélations entre ε_i, nous conservons pour les autres échantillonnages cette probabilité d'inclusion d'ordre 1. Il est également sensé d'imposer aux probabilités d'inclusion d'ordre 2 π_{ij} de respecter la stationnarité et de ne dépendre que de $i-j$, $\pi_{ij} = \bar{\pi}_{i-j}$. Si on impose M éléments en moyenne, on notera $\pi_i^\star = \bar{\pi}_0 = M/N$. On obtient alors pour la variance d'estimation

$$\mathrm{Var}[C_\pi] = \mathrm{Var}[C_P] + \frac{2}{N^2}\sum_{k=1}^{N}(N-|k|)(\frac{\bar{\pi}_k}{\bar{\pi}_0^2}-1)f_k^h$$

et la variance de C_P est dans ce cas égale à $\mathrm{Var}[C_P] = \mathrm{Var}[C_N] + (N/M-1)f_0^h/N$.

Dans le cas d'un échantillonnage déterminantal, on a vu que $\pi_{ij} = K_{ii}K_{jj} - |K_{ij}|^2$. Pour que π_{ij} ne dépendent que de $i-j$, la matrice K de taille N doit dont être de Toeplitz et doit vérifier $K_{ii} = M/N = 1/\alpha$. K dépend donc explicitement de M et N , et on a alors $\pi_{ij} = K_{N,M}(0)^2 - |K_{N,M}(i-j)|^2$ où $K_{N,M}$ est une fonction de covariance avec $K_M(0) = M/N$. On suppose de plus que cette fonction possède une limite K lorsque N et M tendent vers l'infini à rapport constant.

Un exemple simple est fourni par la somme des M premiers vecteurs de la base de Fourier discrète de dimension N, soit

$$K_{N,M}(k) = \frac{1}{N}\sum_{l=0}^{M-1} e^{2\imath\pi\frac{kl}{N}} = \frac{e^{\imath\pi\frac{k(M-1)}{N}}}{N}\frac{\sin\left(\frac{\pi kM}{N}\right)}{\sin\left(\frac{\pi k}{N}\right)}$$

qui tend vers $e^{\imath\pi k/\alpha}\sin(\pi k/\alpha)/\pi k$, où $\imath = \sqrt{-1}$.

Les performances de l'estimateur utilisant l'échantillonnage déterminantal sont données par

$$\mathrm{Var}[C_D] = \mathrm{Var}[C_P] - \frac{2}{N^2}\sum_{k=1}^{N}(N-|k|)\frac{|K_{N,M}(k)|^2}{K_{N,M}(0)^2}f_k^h$$

On peut alors obtenir le résultat asymptotique suivant pour l'excès de variance par rapport à C_N (pour $N, M \to +\infty$, $N/M = \alpha$)

$$\begin{aligned} N\big(\mathrm{Var}[C_\alpha] - \mathrm{Var}[C_N]\big) &\longrightarrow \alpha\sum_{k\in\mathbb{Z}}\widetilde{f}^h_{\alpha k} - \sum_{k\in\mathbb{Z}}\widetilde{f}^h_k \\ N\big(\mathrm{Var}[C_P] - \mathrm{Var}[C_N]\big) &\longrightarrow (\alpha-1)f_0^h \\ N\big(\mathrm{Var}[C_D] - \mathrm{Var}[C_N]\big) &\longrightarrow \alpha f_0^h - \alpha^2\sum_{k\in\mathbb{Z}}\big|K(k)\big|^2 f_k^h \end{aligned}$$

Ce résultat présuppose des propriétés de sommabilité des fonctions en jeu. Il est intéressant de noter à partir du dernier résultat que la fonction K qui minimise la variance de C_D est à module carré proportionnel à f_k^h. En effet le minimum est obtenu en maximisant $\sum_k \big|K(k)\big|^2 f_k^h$, produit scalaire entre $\big|K(k)\big|^2$ et f_k^h. On vérifie l'assertion en appliquant l'inégalité de Cauchy-Schwartz.

Dans l'exemple où $K_{N,M}$ est la somme des M premiers vecteurs de Fourier, on obtient explicitement

$$N\Big(\mathrm{Var}[C_D] - \mathrm{Var}[C_N]\Big) \longrightarrow \alpha f_0^h - \alpha^2\sum_{k\in\mathbb{Z}}\frac{\sin^2\left(\frac{\pi k}{\alpha}\right)}{\pi^2k^2}f_k^h$$

Deux arguments forts plaident en faveur de ce choix particulier. Le premier est que les polynômes trigonométriques sont les polynômes orthogonaux associés à la mesure uniforme sur un intervalle compact. Le deuxième est que les vecteurs de Fourier sont proches des fonctions propres d'opérateurs de covariance de processus stationnaires.

Illustration. On considère un modèle AR(1) $x_n = ax_{n-1} + \sqrt{1-a^2}w_n$ où $|a| < 1$, w_n étant une suite de variables gaussiennes centrées, normalisées et indépendantes. On souhaite estimer $\mathrm{E}[x]$ and $\mathrm{E}[x^2]$, c'est-à-dire que l'on considère les fonctions $h(x) = x^1$ et $h(x) = x^2$. Pour ce choix on a $f_k^1 = a^{|k|}$ et $f_k^2 = 2a^{2|k|} + 1$. On s'intéresse ici à l'excès de variance apporté par le sous-échantillonnage par rapports à l'estimateur empirique C_N. Les limites calculées précédemment sont représentées en fonction de a dans les figures 1 pour $\alpha = 10$ et 2 pour $\alpha = 11$, et dans les deux cas pour $h(x) = x^1$ et $h(x) = x^2$. Cet exemple simple confirme quelques faits :

– Si f_k^h est toujours positive, l'échantillonnage déterminantal est toujours supérieur à l'échantillonnage de Poisson. C'est le cas par exemple dans le cas $AR(1)$ pour $h(x) = x^2$ et tout a (figures 1,2 bas), et pour $h(x) = x$ pour $a > 0$ (figures 1,2 haut).

– Ce n'est plus vrai si f_k^hpeut prendre des valeurs négatives, comme illustré dans le cas $AR(1)$ pour $a < 0$ et $h(x) = x$, (figures 1,2 haut), cas pour lequel l'échantillonnage de Poisson a une variance plus faible que l'échantillonnage déterminantal.

Lorsque l'on compare les résultats de C_D au sous échantillonnage périodique, on obtient le résultat surprenant que contrairement à l'intuition, les performances de C_α peuvent être plus faibles que celle de l'échantillonnage déterminantal ! On peut prouver cela dans l'exemple $AR(1)$ pour $h(x) = x$. Pour les trois types de sous-échantillonnage envisagés, l'excès de variance en $a = 0$ est égal à $\alpha - 1$ (ce qui en passant est logique puisque $a = 0$ correspond au bruit blanc). Si on évalue la dérivée de l'excès par rapport à a en $a = 0$ on obtiendra un ordonnancement local des trois estimateurs. Pour C_α cette dérivée vaut -2 ; pour C_D elle vaut $-2|K(1)|^2/|K(0)|^2$. Comme K est une covariance, $-2|K(1)|^2/|K(0)|^2 \geq -2$. Ainsi pour des valeurs de a négatives et suffisamment proches de 0, l'excès de variance de l'échantillonnage périodique est plus grand que celui de l'échantillonnage déterminantal. Notons malgré tout que la zone en a pour lequel ce résultat est vrai est très limitée, que le gain est très limité dans le cas α impair, et que dans cette situation étonnante, C_D et C_α sont moins bons que C_P.

Les deux figures représentent l'excès de variance pour deux valeurs de α, l'une paire l'autre impaire. Le comportement le l'estimateur à échantillonnage périodique de $\mathrm{E}[x]$ est radicalement différent en fonction de la parité de α. Pour α pair, l'excès de variance diverge quand $a \to -1$. Lorque a est positif, le sous échantillonnage poduit des échantillons positivement corrélés quelque soit la parité de α et le comportement asymptotique des variances de C_α et C_N sont du même ordre en a lorsque a approche 1. L'excès de variance tend alors vers 0. A l'opposé, pour $a < 0$, les échantillons du modèle AR(1) sont alternativement positivement puis négativement corrélés. Un échantillonnage d'un facteur pair ne considère alors que des échantillons positivement corrélés quand un échantillonnage de facteur impair conserve des échantillons négativement corrélés. Les ordres en a des variances de C_α et C_N sont alors identiques dans le cas α impair (l'excès de variance tend vers 0) mais divergent dans le cas pair (l'excès de variance diverge).

Pour conclure. L'échantillonnage déterminantal peut être très largement supérieur à l'échantillonnage de Poisson, comme en témoigne l'exemple de l'estimation du moment d'ordre 2. Toutes ces remarques qui sont tirées dans un cas très simple laissent présager des applications très intéressantes dans des situations de signaux non stationnaires, et/ou de mesures non structurées temporellement, pour lesquelles la notion de sous-échantillonnage n'est pas aussi aisée que dans le cas temporel, comme par exemple de le cas de graphes [11]. Toutefois, la complexité de l'échantillonnage devra être réduite pour que le passage à l'échelle puisse se faire.

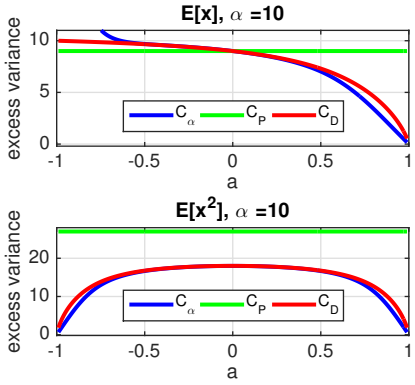

FIGURE 1 – Excès de variance (par rapport à C_N) asymptotique pour les échantillonnages de Poisson (vert), déterminantaux (rouge) et périodique (bleu) pour un taux de compression $\alpha = 10$ dans le cas de l'estimation de la moyenne (haut) et du moment d'ordre 2 (bas).

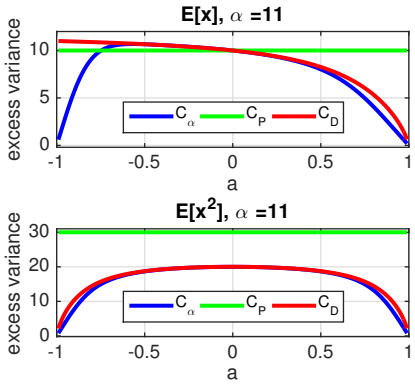

FIGURE 2 – Excès de variance (par rapport à C_N) asymptotique pour les échantillonnages de Poisson (vert), déterminantaux (rouge) et périodique (bleu) pour un taux de compression $\alpha = 11$ dans le cas de l'estimation de la moyenne (haut) et du moment d'ordre 2 (bas).

Références

[1] R. Bardenet and A. Hardy. Monte carlo with determinantal processes. *submitted, arxiv :1605.00361v1*, 2016.

[2] Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *proc. UAI*, 2010.

[3] N. Halko, P. Martinsson, and J. A. Tropp. Finding structure with randomness : Probabilistic algortihms for constructiong approximate matrix decomposition. *SIAM Rev.*, 53(2) :217–288, 2011.

[4] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. *Zeros of gaussian analytic functions and determinantal point processes*. AMS, 2010.

[5] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.*, 5(2-3) :123–286, 2012.

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

[7] V. Loonis and X. Mary. Determinantal sampling designs. *submitted, arxiv :1510.06618*, 2015.

[8] O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1) :83–122, 1975.

[9] M. W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(113–2242), 2011.

[10] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *47th Allerton conf. comm., Contr., and Comp.*, 1999.

[11] N. Tremblay, P. O. Amblard, and S. Barthelmé. Graph sampling with determinantal processes. In *arXiv preprint, 1703.01594*, 2017.

[12] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Th. Comp. Sci.*, 10(1–2) :1–157, 2014.