

Un processus ponctuel déterminantal pour la sélection d'attributs

Ayoub Belhadji ¹, Rémi Bardenet ¹, Pierre Chainais²

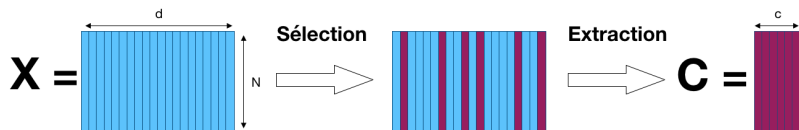
¹CNRS & CRIStAL, Univ. Lille, France

²Ecole Centrale de Lille

Gretsi, Août 2019



Sélection d'attributs pour la réduction de dimension



On cherche un sous-ensemble de colonnes de \mathbf{X} avec des garanties théoriques :

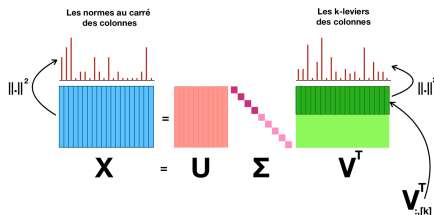
$$\|\mathbf{X} - \mathbf{\Pi}_k \mathbf{X}\|_{\kappa} \leq \|\mathbf{X} - \mathbf{\Pi}_C \mathbf{X}\|_{\kappa} \leq \gamma \|\mathbf{X} - \mathbf{\Pi}_k \mathbf{X}\|_{\kappa}, \quad (1)$$

avec $\|\cdot\|_{\kappa}$ est une norme matricielle ($\kappa \in \{\text{Fr}, 2\}$).

$$\|\mathbf{M}\|_{\text{Fr}}^2 = \sum_{j \in \text{rk}(\mathbf{M})} \sigma_j(\mathbf{M}\mathbf{M}^T), \quad \|\mathbf{M}\|_2^2 = \sigma_1(\mathbf{M}\mathbf{M}^T). \quad (2)$$

- 1 Les algorithmes de sous-échantillonnage matriciel
- 2 Un nouveau algorithme et des intuitions
- 3 Extension pour les problèmes d'interpolation et quadrature

Les garanties théoriques : les normes vs les k -leviers



Théorème (Drineas et al. (2004))

Avec les normes au carré : pour $\delta > 0$

$$\mathbb{P} \left(\|\mathbf{X} - \mathbf{\Pi}_C \mathbf{X}\|_{\text{Fr}}^2 \leq \|\mathbf{X} - \mathbf{\Pi}_k \mathbf{X}\|_{\text{Fr}}^2 + M(\delta) \|\mathbf{X}\|_{\text{Fr}}^2 \right) > 1 - \delta. \quad (3)$$

Théorème (Drineas et al. (2006))

Avec les k -leviers : pour $\delta > 0$ et $c \geq \frac{4000k^2}{\epsilon^2} \log(\frac{1}{\delta})$, on a :

$$\mathbb{P} \left(\|\mathbf{X} - \mathbf{\Pi}_C \mathbf{X}\|_{\text{Fr}}^2 \leq (1 + \epsilon) \|\mathbf{X} - \mathbf{\Pi}_k \mathbf{X}\|_{\text{Fr}}^2 \right) \geq 1 - \delta. \quad (4)$$

$$\mathbb{P}(S) \propto \text{Det}(\mathbf{C}^T \mathbf{C}) \Rightarrow \mathbb{P}(S) \ll \mathbb{P}(S)$$

Théorème (Deshpande et al. (2006))

Soit $S \subset [d]$, un sous-ensemble de k -colonnes sélectionné avec probabilité

$$\mathbb{P}(S) \propto \text{Det}(\mathbf{C}^T \mathbf{C}).$$

On a

$$\mathbb{E} \|\mathbf{X} - \Pi_{\mathbf{C}} \mathbf{X}\|_{\text{Fr}}^2 \leq (1+k) \|\mathbf{X} - \Pi_k \mathbf{X}\|_{\text{Fr}}^2, \quad (5)$$

$$\mathbb{E} \|\mathbf{X} - \Pi_{\mathbf{C}} \mathbf{X}\|_2^2 \leq (d-k)(1+k) \|\mathbf{X} - \Pi_k \mathbf{X}\|_2^2. \quad (6)$$

L'échantillonnage matriciel examiné par les PPDs

Soit $\mathbf{K} \in \mathbb{R}^{d \times d}$ une matrice semi définie positive. Soit \mathcal{S} un sous-ensemble aléatoire de $[d]$.

Définition (Kulesza et Taskar, (2012))

\mathcal{S} suit la loi d'un PPD de noyau \mathbf{K} si pour tout $S \subset [d]$

$$\mathbb{P}(S \subset \mathcal{S}) = \text{Det}(\mathbf{K}_{S,S}). \quad (7)$$

L'échantillonnage volumique est une mixture de PPDs :

$$\mathbb{P}_{\text{VS}}(S) \propto \sum_{T \subset [d], |T|=k} \prod_{i \in T} \sigma_i^2 \text{Det}(\mathbf{K}(T)_{S,S}), \quad (8)$$

avec

$$\mathbf{K}(T) = \mathbf{V}_{T,:} \mathbf{V}_{T,:}^T. \quad (9)$$

L'échantillonnage selon un PPD de projection :

$$\mathbb{P}(\mathcal{S}) = \text{Det}(\mathbf{V}_{[k],\mathcal{S}})^2. \quad (10)$$

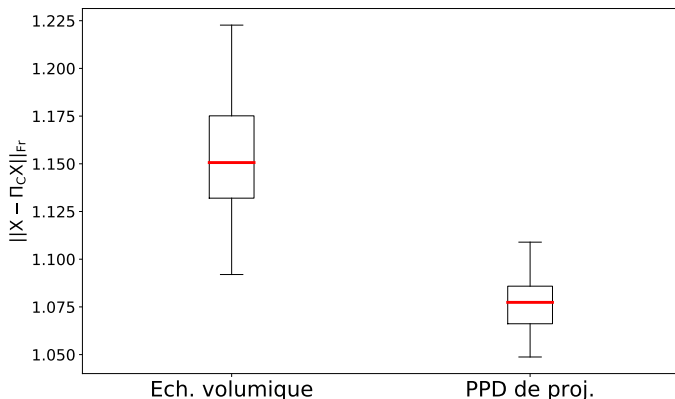
En plus, pour $j \in [d]$,

$$\mathbb{P}(j \in \mathcal{S}) = \mathbb{P}(\{j\} \subset \mathcal{S}) = \text{Det}(\mathbf{K}([k])_{[j],[j]}) = \mathbf{K}([k])_{j,j} = \ell_j^k. \quad (11)$$

La probabilité d'inclusion d'une colonne est égale au k -levier correspondant.

Un exemple réel

Les résultats de la comparaison pour la base de donnée BASEHOCK (N = 1993, d = 4862 ,k=10).



Théorème (B., Bardenet et Chainais (2018))

Soit $S \subset [d]$, un sous-ensemble de k -colonnes sélectionné avec probabilité

$$\mathbb{P}(S) = \text{Det}(\mathbf{V}_{[k],S})^2.$$

- $\mathcal{H}_0 : \text{Card}\{i, \|\mathbf{V}_{k,i}\|_2 \neq 0\} \leq p$
- $\mathcal{H}_\beta : \forall l \geq k + 1, \sigma_{k+1} \leq \beta \sigma_l$

On a

$$\mathbb{E} \|\mathbf{X} - \Pi_{\mathbf{C}} \mathbf{X}\|_{\text{Fr}}^2 \leq \left(1 + \beta^2 \frac{p-k}{d-k} k\right) \|\mathbf{X} - \Pi_k \mathbf{X}\|_{\text{Fr}}^2, \quad (12)$$

$$\mathbb{E} \|\mathbf{X} - \Pi_{\mathbf{C}} \mathbf{X}\|_2^2 \leq \left(1 + \frac{p-k}{d-k} (d-k)k\right) \|\mathbf{X} - \Pi_k \mathbf{X}\|_2^2. \quad (13)$$

La parcimonie p est une mesure de la "complexité" de \mathbf{X} .

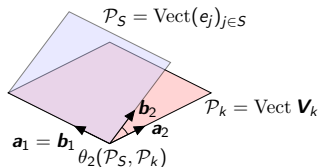
Intuition géométrique : les angles principaux

$$\cos \theta_1(\mathcal{P}_S, \mathcal{P}_k) = \max_{\substack{\|\mathbf{a}\|=\|\mathbf{b}\|=1 \\ \mathbf{a} \in \mathcal{P}_S, \mathbf{b} \in \mathcal{P}_k}} \mathbf{a}^\top \mathbf{b} := \mathbf{a}_1^\top \mathbf{b}_1$$

$$\cos \theta_2(\mathcal{P}_S, \mathcal{P}_k) = \max_{\substack{\|\mathbf{a}\|=\|\mathbf{b}\|=1 \\ \mathbf{a} \in \mathcal{P}_S \cap \text{Vect}\{\mathbf{a}_1\}^\perp \\ \mathbf{b} \in \mathcal{P}_k \cap \text{Vect}\{\mathbf{b}_1\}^\perp}} \mathbf{a}^\top \mathbf{b} := \mathbf{a}_2^\top \mathbf{b}_2$$

$$\cos \theta_k(\mathcal{P}_S, \mathcal{P}_k) = \max_{\substack{\|\mathbf{a}\|=\|\mathbf{b}\|=1 \\ \mathbf{a} \in \mathcal{P}_S \cap \text{Vect}\{\mathbf{a}_1, \dots, \mathbf{a}_{k-1}\}^\perp \\ \mathbf{b} \in \mathcal{P}_k \cap \text{Vect}\{\mathbf{b}_1, \dots, \mathbf{b}_{k-1}\}^\perp}} \mathbf{a}^\top \mathbf{b} := \mathbf{a}_k^\top \mathbf{b}_k$$

$$0 \leq \theta_1(\mathcal{P}_S, \mathcal{P}_k) \leq \dots \leq \theta_k(\mathcal{P}_S, \mathcal{P}_k) \leq \frac{\pi}{2}$$



On montre l'identité

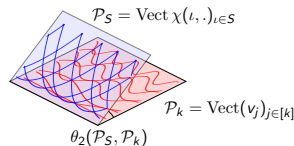
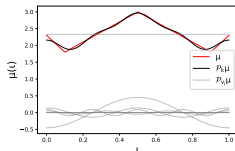
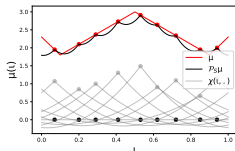
$$\cos^2(\mathcal{P}_k, \mathcal{P}_S) := \prod_{\ell=1}^k \cos^2 \theta_\ell(\mathcal{P}_k, \mathcal{P}_S) = \text{Det}(\mathbf{V}_{[k],S})^2. \quad (14)$$

Extension : interpolations et quadratures dans les EHNR

Extension à un autre problème :

soit \mathcal{D} un espace topologique et soit $\chi : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}_+$ un noyau, $d\omega$ une mesure Borélienne sur \mathcal{D} et $\mathcal{F} = \overline{\text{Vect}\{\chi(\iota, \cdot), \iota \in \mathcal{D}\}}$. On note (σ_j, v_j) les éléments propres de l'opérateur intégral :

$$(\mathcal{C}\mu)(\iota) = \int_{\mathcal{D}} \mu(\iota') \chi(\iota, \iota') d\omega(\iota'). \quad (15)$$



Analogie :

Discret	$[d]$	\mathbf{X}_i	$\mathbf{X}^T \mathbf{X}$	$\sigma_j(\mathbf{X}^T \mathbf{X})$	$\mathbf{v}_{:,j}$	$\ell_i^k = \sum_{j \in [k]} v_{i,j}^2$
Continu	\mathcal{D}	$\chi(\iota, \cdot)$	\mathcal{C}	$\sigma_j(\mathcal{C})$	v_j	$\sum_{j \in [k]} v_j(\iota)^2$

Application : les quadratures à noyau

On s'intéresse au problème d'approximation d'intégrale de $y \in \mathcal{F}$:

$$\int_{\mathcal{D}} y(\iota)g(\iota)d\omega(\iota) \approx \sum_{\iota \in S} w_{\iota}(g)y(\iota). \quad (16)$$

On définit l'élément moyen

$$\mu = \int_{\mathcal{D}} g(\iota)\chi(\iota, \cdot)d\omega(\iota). \quad (17)$$

L'erreur d'approximation

$$\begin{aligned} \left| \int_{\mathcal{D}} y(\iota)g(\iota)d\omega(\iota) - \sum_{\iota \in S} w_{\iota}(g)y(\iota) \right| &= \langle y, \mu_g - \sum_{\iota \in S} w_{\iota}(g)\chi(\iota, \cdot) \rangle_{\mathcal{F}} \\ &\leq \|y\|_{\mathcal{F}} \|\mu_g - \sum_{\iota \in S} w_{\iota}(g)\chi(\iota, \cdot)\|_{\mathcal{F}}. \end{aligned}$$

Pour les poids optimaux $w_{\iota}^*(g)$ on a

$$\|\mu_g - \sum_{\iota \in S} w_{\iota}(g)^* \chi(\iota, \cdot)\|_{\mathcal{F}} = \|\mu_g - \mathcal{P}_S \mu_g\|_{\mathcal{F}}. \quad (18)$$

Théorème (B., Bardenet et Chainais (2019))

Soit $S \subset \mathcal{D}$ soit un sous ensemble aléatoire tiré selon la distribution de densité

$$\text{Det}(v_j(\iota))_{\substack{j \in [k] \\ \iota \in S}}^2 \otimes_{\iota \in S} d\omega(\iota). \quad (19)$$

Alors

$$\mathbb{E} \sup_{\|g\|_{d\omega} \leq 1} \|\mu_g - \mathcal{P}_S \mu_g\|_{\mathcal{F}} \leq 2\sigma_{k+1} + k \sum_{m \geq k+1} \sigma_m. \quad (20)$$

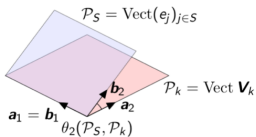
\mathcal{D}	$[0, 1]$	$[0, 1]^d$	\mathbb{R}	\mathbb{R}^d
\mathcal{F}	Sobolev	Korobov	"Gaussien"	"Gaussien"
σ_k	$\mathcal{O}(k^{-2s})$	$\mathcal{O}(\log^{2(s-1)d}(k)k^{-2s})$	$\mathcal{O}(e^{-\alpha k})$	$\mathcal{O}(e^{-\alpha dk^{1/d}})$

comparés au taux classique du TCL $\mathcal{O}(1/\sqrt{k})$.

Conclusion

- Un PPD de projection pour l'échantillonnage matricielle avec des garanties théoriques.
- Généralisation au cas continu : la décroissance des valeurs propres des opérateurs d'intégration fournissent les taux de convergence de l'interpolation et quadrature à noyau.

Réduction de dimension



Quadratures

Interpolations

Merci pour votre attention !



A. Belhadji, R. Bardenet, P. Chainais

A determinantal point process for column subset selection
arXiv :1812.09771, 2018.



A. Belhadji, R. Bardenet, P. Chainais

Kernel quadrature with DPPs
arXiv :1906.07832, 2019.



Deshpande, A. and Rademacher, L. and Vempala, S. and Wang, G.

Matrix Approximation and Projective Clustering via Volume Sampling
Proc. ACM-SIAM SODA, 2006.



Drineas, P. and Frieze, A. and Kannan, R. and Vempala, S. and Vinay, V.

Clustering Large Graphs via the Singular Value Decomposition
Mach. Learn., 2004.



Kulesza, A. and Taskar, B.

Determinantal point processes for machine learning
FTML, 2012.