

École nationale
de la statistique
et de l'analyse
de l'information

Gaussian Processes on Distributions based on Regularized Optimal Transport

Julien Heurtin, Léonard Gousset, Louis Allain

Tuteurs:
Brian Staber
Sébastien Da Veiga

2024

27 Mars

Introduction

- Distributions comme variable d'un problème de machine learning
- Comparer des distributions
- Transport Optimal Régularisé

$$\min_{\mathbf{P} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon \sum_{i,j} P_{ij} \left(\log(P_{ij}) - 1 \right)$$

- Un algorithme itératif pour résoudre ce problème: l'algorithme de Sinkhorn
- Cet algorithme fait apparaître des *potentiels de Sinkhorn*
- Le papier de Bachoc et al. propose de s'appuyer sur ces potentiels pour comparer des distributions.

Sommaire

1. Le noyau Sinkhorn
2. Les processus Gaussien
3. Le dataset Rotor37
4. Un modèle de référence
5. Le noyau Sinkhorn en pratique
6. Conclusion

Le noyau Sinkhorn

Le noyau Sinkhorn

Definition (Noyau Sinkhorn)

Pour P, Q deux distributions et $\mathcal{U}$ une mesure de réference, le noyau de Sinkhorn se définit:

$$F\left(\|g_{\mathcal{U}}^{P} - g_{\mathcal{U}}^{Q}\|_{L^2(\mathcal{U})}\right)$$

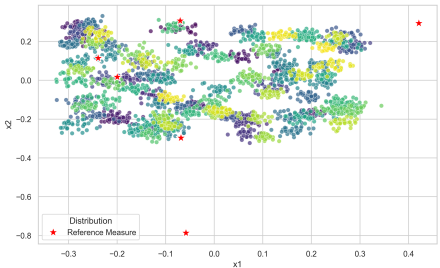
Si F est une fonction continue sur $[0, +\infty[$ telle que $F \circ \sqrt{\cdot}$ soit complétement monotone sur $[0, +\infty[$, alors ce noyau est bien défini positif.

Le papier montre trois propriétés de ce noyaux:

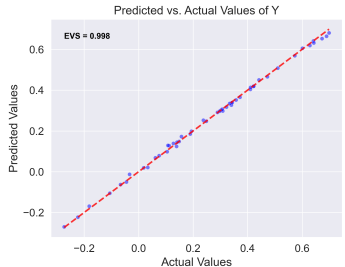
1. La distance entre les potentiels est dominée par celle entre les distributions.
2. Les potentiels caractérisent effectivement les distributions.
3. En considérant des échantillons des distributions, le noyau empirique converge vers le noyau des distributions.

Méthodes à noyaux

Dans tous les cas où le noyau est défini positif, ce noyau peut être utilisé dans les méthodes à noyaux, comme la Kernel Ridge Regression.



(a) Les distributions



(b) Performance de la KRR avec noyau Sinkhorn.

Figure: Un exemple sur des données simulées.

Les processus Gaussien

ENSAI & SAFRAN

Les processus Gaussien

Une autre méthode de régression qui utilise les noyaux définis positifs:

Definition (Processus Gaussien)

On dit qu'une fonction est un processsus Gaussien de moyenne m et de covariance k si pour tout ensemble fini $X = (x_1, \cdots, x_n)$, $f_X = (f(x_1), \ldots, f(x_n))^\top$ suit une loi normale multivariée de moyenne $m(X)$ et de matrice de covariance $k(X, X)$. On écrit

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$$

Avantages:

- Paradigme bayésien
- Intervalles de confiance
- Petits échantillons
- Fonction de covariance - noyau

Entraînement du modèle

On considère le modèle $y = f(\mathbf{x}) + \varepsilon$, avec ε des bruits iid Gaussien de variance σ_n^2.
Sous la prior on peut écrire:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) + \sigma_n^2 I & K(\mathbf{X},\mathbf{X}_*) \\ K(\mathbf{X}_*,\mathbf{X}) & K(\mathbf{X}_*,\mathbf{X}_*) \end{bmatrix}\right)$$

On remarque alors que:

$$\mathbf{y} \sim \mathcal{N}\left(0, K + \sigma_n^2 I\right)$$

Et la log-vraissemblance marginale est alors donnée par:

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K} + \sigma_n^2\mathbf{I}| - \frac{n}{2}\log(2\pi)$$

Inférence

En cherchant la loi de la fonction évaluée en les observations tests, sachant les données d'entrainements et les observations tests, on obtient:

$$\mathbf{f}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}\left(\bar{\mathbf{f}_*}, \text{cov}(\mathbf{f}_*)\right)$$

Avec

$$\bar{\mathbf{f}_*} = K(X_*, X)\left[K(X, X) + \sigma_n^2 I\right]^{-1}\mathbf{y}$$

et

$$\text{cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)\left[K(X, X) + \sigma_n^2 I\right]^{-1}K(X, X_*)$$

Des noyaux usuels (1/4)

Definition (Radial Basis Function)

Soit $\mathcal{X} \subset \mathcal{H}$, où $\mathcal{H}$ est un espace de Hilbert. Soit $l > 0$ la longueur caractéristique. Le noyau RBF $k_l \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est défini par

$$k_l(x, x') = \exp\left(-\frac{\|x - x'\|_{\mathcal{H}}^2}{2l^2}\right), \quad \forall x, x' \in \mathcal{X}$$

où $\|.\|_{\mathcal{H}}$ est la norme associée au produit scalaire de l'espace de Hilbert $\mathcal{H}$.

Des noyaux usuels (2/4)

Definition (Noyau de Matérn)

Soit $\mathcal{X} \subset \mathcal{H}$, où $\mathcal{H}$ est un espace de Hilbert. Soit $l > 0$ la longueur caractéristique et $\nu > 0$. Le noyau de Matérn $k_{l,\nu} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est défini par

$$k_{l,\nu}(x, x') = \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{l} \|x - x'\|_{\mathcal{H}} \right)^{\nu} B_{\nu} \left(\frac{\sqrt{2\nu}}{l} \|x - x'\|_{\mathcal{H}} \right)$$

où $\|.\|_{\mathcal{H}}$ est la norme associée au produit scalaire de l'espace de Hilbert $\mathcal{H}$ et B_{ν} est la fonction de Bessel modifiée.

Des noyaux usuels (3/4)

Definition (Matérn $3/2$)

Pour $\nu = 3/2$, le noyau de Matérn, composé de fonctions une fois différentiable, est donné par :

$$k_{l,\frac{3}{2}}(x, x') = \left(1 + \frac{\sqrt{3}\|x - x'\|_{\mathcal{H}}}{l}\right) \exp\left(-\frac{\sqrt{3}\|x - x'\|_{\mathcal{H}}}{l}\right)$$

Des noyaux usuels (4/4)

Definition (Matérn $5/2$)

Pour $\nu = 5/2$, le noyau de Matérn, composé de fonctions deux fois différentiables, est donné par :

$$k_{l,\frac{5}{2}}(x, x') = \left(1 + \frac{\sqrt{5}\|x - x'\|_{\mathcal{H}}}{l} + \frac{5\|x - x'\|_{\mathcal{H}}^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}\|x - x'\|_{\mathcal{H}}}{l}\right)$$

Le dataset Rotor37

ENSAI & SAFRAN

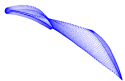
Le problème de machine learning

Variables d'entrées

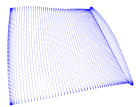
1. Les aubes
2. La pression moyenne du fluide
3. La velocité moyenne du fluide

Variables de sorites

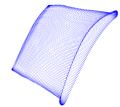
1. Efficacité des aubes
2. Débit massique
3. Le ratio de compression



(a) Première perspective.



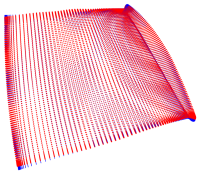
(b) Deuxième perspective



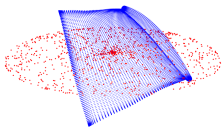
(c) Troisième perspective.

Figure: Différentes perspectives du nuage de points d'une aube.

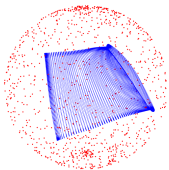
Le transport optimal pour des aubes



(a) Une autre aube.



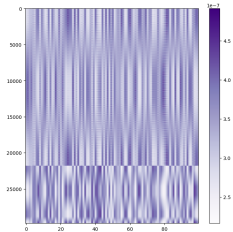
(b) Un disque.



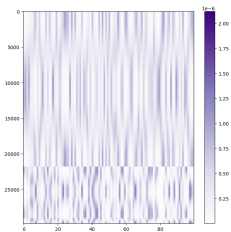
(c) Une sphere.

Figure: Le nuage de points d'une aube (bleu) et différentes mesures de références (rouge).

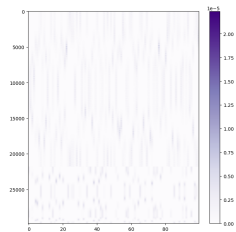
Le transport optimal pour des aubes



(a) $\varepsilon = 0.01$



(b) $\varepsilon = 0.001$



(c) $\varepsilon = 0.0001$

Figure: La matrice de coût de transport entre une aube et la mesure de référence disque pour différentes valeurs de ε.

Le sous-échantillonnage des aubes

La complexité de l'algorithme de Sinkhorn est donnée par:

$$\mathcal{O}\left(\frac{n|\mathbf{u}|log(n|\mathbf{u}|)}{\varepsilon^2}\right)$$

Pour réduire le coût de calcul du transport optimal nous avons considéré trois méthodes de sous-échantillonnage:

1. Sous-échantillonnage optimisé
2. Sous-échantillonnage aléatoire et indépendant
3. Sous-échantillonnage aléatoire unique

Un modèle de référence

Construction du modèle

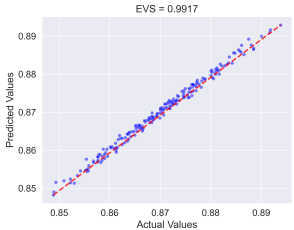
Un modèle très simple sur ce dataset consiste à utiliser les coordonnées des points comme variables de la manière suivante.

Une aube sous la forme $(29773, 3)$:

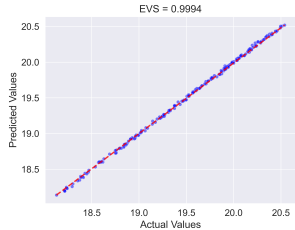
$$\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_{29773} & y_{29773} & z_{29773} \end{bmatrix} \longrightarrow \text{Transformée en la forme } (1, 89319): \begin{bmatrix} x_1 & x_2 & \dots & x_{29773} & y_1 & \dots & z_{29772} & z_{29773} \end{bmatrix} \quad (1)$$

Le modèle possède alors 89319 variables. On réduit la dimension grâce à une Analyse en Composante Principale de dimensions 32.

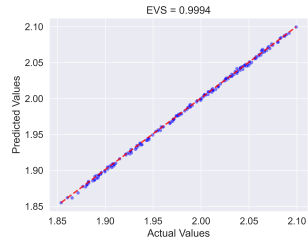
Performance de ce modèle



(a) Efficacité, EVS = 0.9917



(b) Débit massique, EVS = 0.9994



(c) Ratio de compress., EVS = 0.9994

Figure: Performance du modèle de référence sur l'ensemble des données, évalué sur le dataset de test.

Évaluation du sous-échantillonnage

	Optimisé	Aléatoire indépendant		Aléatoire unique		
Taille	2000	29772	2000	2000	200	50
EVS	0.5263	0.0002	0.001	0.9993	0.9992	0.9988

Table: Performance pour toutes les méthodes d'échantillonnage selon la taille.

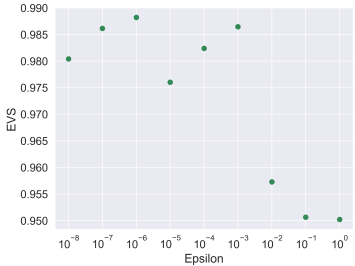
Ces derniers résultats montre les limites du dataset Rotor37.
Sa construction particulière lie les nuages de points des aubes entre eux.

Le noyau Sinkhorn en pratique

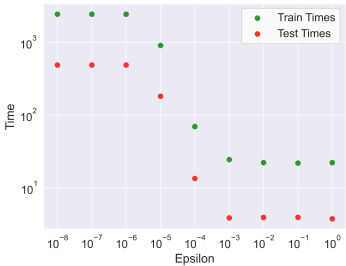
ENSAI & SAFRAN

Hyperparamètres: ε

On rappelle la complexité de l'algorithme de Sinkhorn: $\mathcal{O}\left(\frac{n|\mathbf{u}|log(n|\mathbf{u}|)}{\varepsilon^2}\right)$



(a) Performance (EVS).



(b) Temps (secondes).

Figure: Évolution de la performance et du temps de calcul en fonction de ε.

Hyperparamètres: méthodes de sous-échantillonnages

Optimisé
Performance similaire au modèle de référence: $EVS = 0.5858$ $(vs.\ 0.526)$.

Aléatoire indépendant
Le modèle n'est pas utilisable.

Aléatoire unique
Avec 2000 points, on obtient $EVS = 0.978$. Proche du modèle de référence.

Le sous-échantillonnage pour le noyau de Sinkhorn est toujours sujet à la dépendance entre les aubes. Il est donc difficile de juger cet hyperparamètre.

Hyperparamètres: les mesures de références

On rappelle que l'on a considéré trois mesures de références différentes: **i)** une autre aube, **ii)** une sphere, **iii)** un disque.

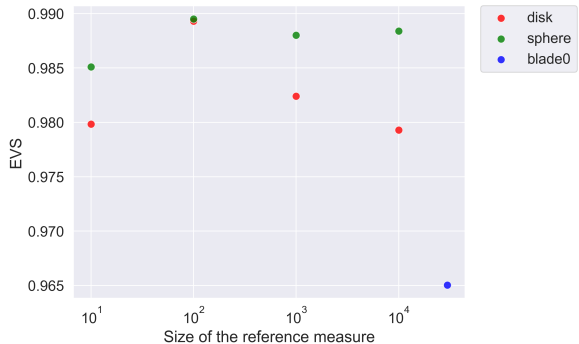
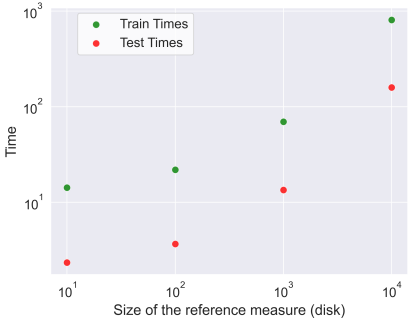
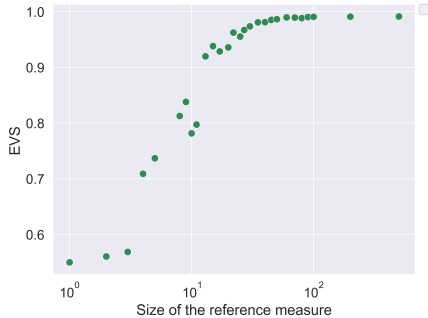


Figure: EVS pour différenets mesures de références.

Hyperparamètres: les mesures de références



(a) Temps (secondes).



(b) EVS.

Figure: Évolution de la performance et du temps de calcul en fonction de de la taille de la mesure de référence.

Hyperparamètres: Le nombre de dimensions dans l'ACP

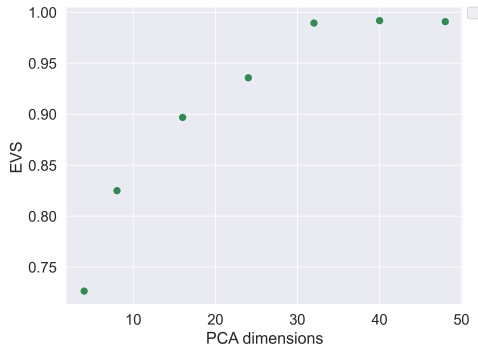


Figure: EVS pour différentes mesures de références.

Hyperparamètres: différentes tailles de problèmes

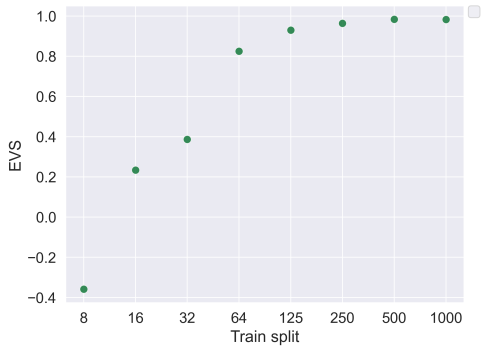


Figure: EVS pour différentes tailles de problème.

Hyperparamètres: Processus Gaussien vs. Kernel Ridge Regression

	RBF	Matérn $3/2$	Matérn $5/2$
Processus Gaussien	0.9882	0.9885	0.9895
Kernel Ridge Regression	0.9807	0.9812	0.9834

Table: Le score de variance expliqué en fonction de modèle et du noyau utlisé.

Le modèle le plus performant

Les hyperparamètres retenus:

- $\varepsilon = 10^{-4}$
- Aucun sous-échantillonnage.
- Sphère avec 100 points.
- ACP de dimensions 32.
- 1000 obs. d'entrainement.
- Processus Gaussien, Matérn $5/2$.
- `GPy`, 24 init. aléatoire et 10000 itérations maximum.

Résultats:

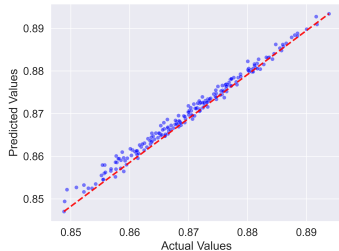


Figure: **Efficiency:** EVS = 0.9895, $\bar{CI} = 1.7^{-6}$
Massflow: EVS = 0.9983, $\bar{CI} = 1.1^{-3}$
Compression Rate: EVS = 0.9983, $\bar{CI} = 1.05^{-5}$

Conclusion

Conclusion

Le papier

- Un noyau pour les distributions
- Potentiels de Sikhorn
- Des garanties théoriques fortes

Nos résultats

- OT sur des aubes
- Hyperparamètres
- Mesure de référence réduite

Approfondir

- Sliced-Wasserstein
- Modèle à vecteur de paramètres
- Optimisation de la mesure de référence



École nationale
de la statistique
et de l'analyse
de l'information

Merci pour votre attention

Julien Heurtin, Léonard Gousset, Louis Allain

Tuteurs:
Brian Staber
Sébastien Da Veiga
2024

27 Mars

Maximum Mean Discrepancy

On considère une feature map $\phi : \mathcal{X} \to \mathcal{H}$ telle que $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ soit un noyau défnit positif.

Pour une probabilité P sur $\mathcal{X}$, on définit le *mean embedding* $\mu_P(X) = [\mathbb{E}(\phi(X_1)), \cdots, \mathbb{E}(\phi(X_m))]^\top$.

La *Maximum Mean Discrepancy* entre deux distributions P et Q est alors la distance entre les mean embeddings des deux distributions:

$$MMD^2(P, Q) = ||\mu_P - \mu_Q||^2_{\mathcal{H}}$$

Le noyau de Sinkhorn est défini positif (1/3)

ENSAI

Proposition 1

Soit $F : \mathbb{R}_+ \to \mathbb{R}$. Les deux propositions suivantes sont équivalentes:

- Il existe $d \in \mathbb{N}$, tel que $K_d : (\mathbb{R}^d)^2 \longrightarrow \mathbb{R}$ soit défini positif, où $(x, y) \longmapsto F(||x - y||)$.
- Pour tout espace de Hilbert $\mathcal{H}$, la fonction $K_{\mathcal{H}}(u, v) = F\left(||u - v||_{\mathcal{H}}\right)$ est définie positive.

Le noyau de Sinkhorn est défini positif (2/3)

ENSAI

Proposition 2

Soit $F : \mathbb{R}_+ \to \mathbb{R}$. Les trois propositions suivantes sont équivalentes:

- $\forall\ \mathcal{H}$, espace de Hilbert, $K_{\mathcal{H}}(u, v) = F(||u - v||_{\mathcal{H}})$ est défini positif.
- $F \circ \sqrt{\cdot}$ soit complétement monotone sur $[0, +\infty[$
- Il existe une mesure de Borel non nulle ν sur $[0, +\infty[$ telle que pour tous $t > 0$, $F(t) = \int_0^{+\infty} e^{-ut^2} d\nu(u)$

Le noyau de Sinkhorn est défini positif (3/3)

$L^2(\mathcal{U})$

$$L^2(\mathcal{U}) = \{f : \mathcal{U} \to \mathbb{R} \ / \ ||f||_2 = \left(\int_{\mathcal{U}} |f|^2 du \right)^{1/2} < +\infty \}$$

est un espace de Hilbert avec produit scalaire: $\langle f, g \rangle = \int_{\mathcal{U}} f \bar{g} du$