Construction de noyau fondé sur la distance de Wasserstein et utilisation pour la prédiction d'un code à entrée fonctionnelle

Nil Venet

CEA Tech Occitanie, IMT

25 avril 2017

Plan de la présentation

Données dont les entrées sont des distributions

Dans la suite on s'intéresse à des données

$$(\mu_i, y_i)_{i=1}^n,$$

où les μ_i sont des distributions de probabilités sur $\mathbb{R}$.

Données dont les entrées sont des distributions

Dans la suite on s'intéresse à des données

$$(\mu_i, y_i)_{i=1}^{n},$$

où les μ_i sont des distributions de probabilités sur $\mathbb{R}$.

Motivations

- Entrées fonctionnelles.
- Entrées aléatoires pour un code.

Données dont les entrées sont des distributions

Dans la suite on s'intéresse à des données

$$(\mu_i, y_i)_{i=1}^n,$$

où les μ_i sont des distributions de probabilités sur $\mathbb{R}$.

Motivations

- Entrées fonctionnelles.
- Entrées aléatoires pour un code.

- Il reste à choisir une distance sur l'espace des entrées.

Distance de Wasserstein

- On suppose que les mesures considérées possèdent un moment d'ordre deux.

Distance de Wasserstein

- On suppose que les mesures considérées possèdent un moment d'ordre deux.
- La distance de Wasserstein quadratique entre μ et ν est définie par

$$W_2(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int |x - y|^2 \, d\pi(x, y) \right)^{1/2}, \qquad (1)$$

avec $\Pi(\mu, \nu)$ l'ensemble des mesures de probabilités sur $\mathbb{R}^2$ dont les lois marginales sont μ et ν.

Distance de Wasserstein

- On suppose que les mesures considérées possèdent un moment d'ordre deux.
- La distance de Wasserstein quadratique entre μ et ν est définie par

$$W_2(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int |x-y|^2 \, d\pi(x,y) \right)^{1/2}, \qquad (1)$$

  avec $\Pi(\mu,\nu)$ l'ensemble des mesures de probabilités sur $\mathbb{R}^2$ dont les lois marginales sont μ et ν.
- On cherche des *noyaux définis-positifs stationnaires* sur l'espace $\mathcal{W}_2(\mathbb{R})$ obtenu.

Noyaux défini-négatifs

Théorème

Pour tout $H \in [0,1]$,

$$(\mu,\nu) \mapsto W_2(\mu,\nu))^{2H} \tag{2}$$

est un noyau défini-négatif : $\forall \mu_1, \cdots, \mu_n \in \mathcal{W}_2(\mathbb{R})$, $\forall c_1, \cdots, c_n \in \mathbb{R}$ *t.q.* $\sum_{i=1}^n c_i = 0$,

$$\sum_{i,j=1}^n c_i c_j W_2(\mu_i,\mu_j))^{2H} \leq 0. \tag{3}$$

Conséquences, 1

Théorème (Champs browniens fractionnaires)

Pour tout $0 \leq H \leq 1$ *et* $\sigma \in \mathcal{W}_2(\mathbb{R})$,

$$K^{H,\sigma}(\mu,\nu) = \frac{1}{2}\left(W_2^{2H}(\sigma,\mu) + W_2^{2H}(\sigma,\nu) - W_2^{2H}(\mu,\nu)\right) \quad (4)$$

est une fonction de covariance sur $\mathcal{W}_2(\mathbb{R})$. *De plus cette fonction est non-dégénérée si et seulement si* $0 < H < 1$.

- On a un champ brownien fractionnaire indexé par $\mathcal{W}_2(\mathbb{R})$. Il est à accroissements stationnaires.

Conséquences, 2

Théorème (Processus stationnaires)

Pour $F : \mathbb{R}^+ \to \mathbb{R}^+$ complètement monotone et $0 < H \leq 1$,

$$(\mu, \nu) \mapsto F\left(W_2^{2H}(\mu, \nu)\right) \tag{5}$$

est la fonction de covariance d'un processus gaussien stationnaire indexé par $\mathcal{W}_2(\mathbb{R})$.

- Rappelons qu'une fonction $F : \mathbb{R}^+ \to \mathbb{R}^+$ complètement monotone est une fonction infiniment dérivable telle que $(-1)^n F^{(n)}$ est à valeurs positives pour tout $n \in \mathbb{N}$.

Modèle paramétrique Gaussien stationnaire

En particulier

$$K_{\sigma^2,\ell,H}(\nu_1,\nu_2) = \sigma^2 \exp\left(-\frac{W_2(\nu_1,\nu_2)^{2H}}{\ell}\right), \tag{6}$$

$$H \in [0,1], \sigma > 0, l > 0,$$

est un modèle paramétrique de processus gaussiens stationnaires indexés par $\mathcal{W}_2(\mathbb{R})$.

Plan de la présentation

Hypothèses pour nos résultats

Condition (1)

On considère une matrice triangulaire de points d'observations de $\mathcal{W}_2(\mathbb{R})$ $\{\mu_1, ..., \mu_n\} = \{\mu_1^{(n)}, ..., \mu_n^{(n)}\}$ *tels que pour tout* $n \in \mathbb{N}$ *et* $1 \leq i \leq n$, μ_i *est de support inclus dans* $[i, i+K]$, *où* $K < \infty$ *est fixe.*

Condition (2)

Le modèle de fonctions de covariance $\{K_\theta, \theta \in \Theta \subset \mathbb{R}^p\}$ *est tel que*

$$\forall \theta \in \Theta,\; K_\theta(\mu, \nu) = F_\theta\left(W_2(\mu, \nu)\right) \;\; et \quad \sup_{\theta \in \Theta} |F_\theta(t)| \leq \frac{A}{1 + |t|^{1+\tau}}$$

avec $A < \infty$ *et* $\tau > 1$ *des constantes.*

Condition (3)

Nous disposons d'observations $y_i = Y(\mu_i)$, $i = 1, \cdots, n$ *du processus aléatoire gaussien* Y, *centré et de covariance* K_{θ_0} *pour un* $\theta_0 \in \Theta$.

Condition (4)

La suite de matrices $R_\theta = (K_\theta(\mu_i, \mu_j))_{1 \le i,j \le n}$ *est telle que* $\lambda_{\inf}(R_\theta) \ge c$ *pour une constante* $c > 0$, *où* $\lambda_{\inf}(R_\theta)$ *désigne la plus petite valeur propre de* R_θ.

Condition (5)

$$\forall \alpha > 0, \liminf_{n \to \infty} \inf_{||\theta - \theta_0|| \ge \alpha} \frac{1}{n} \sum_{i,j=1}^{n} \left[K_\theta(\mu_i, \mu_j) - K_{\theta_0}(\mu_i, \mu_j)\right]^2 > 0.$$

Condition (6)

$\forall t \geq 0$, $F_\theta(t)$ est $\mathcal{C}^1$ en θ et vérifie $\sup_{\theta \in \Theta} \max_{i=1,\cdots,p} \left| \frac{\partial}{\partial \theta_i} F_\theta(t) \right| \leq \frac{A}{1+t^{1+\tau}}$, *où A, τ sont définis dans la Condition 2.*

Condition (7)

Pour tout $t \geq 0$, $F_\theta(t)$ est $\mathcal{C}^3$ en θ et$\forall q \in \{2,3\}$,
$\forall i_1 \cdots i_q \in \{1, \cdots p\}$,

$$\sup_{\theta \in \Theta} \max_{i=1,\cdots,p} \left| \frac{\partial}{\partial \theta_{i_1}} \cdots \frac{\partial}{\partial \theta_{i_q}} F_\theta(t) \right| \leq \frac{A}{1+|t|^{1+\tau}}.$$

Condition (8)

$\forall(\lambda_1 \cdots, \lambda_p) \neq (0, \cdots, 0)$,

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i,j=1}^{n} \left(\sum_{k=1}^{p} \lambda_k \frac{\partial}{\partial_{\theta_k}} K_{\theta_0}(\mu_i, \mu_j) \right)^2 > 0.$$

Consistance de l'EMV

Théorème

Sous les conditions 1 à 5 l'estimateur par maximum de vraisemblance est consistant, c'est-à dire :

$$\hat{\theta}_{ML} \xrightarrow[n\to\infty]{\mathbb{P}} \theta_0.$$

Normalité asymptotique de l'EMV

Théorème

Soit M_{ML} la matrice de taille $p \times p$ définie par

$$(M_{ML})_{i,j} = \frac{1}{2n} Tr \left(K_{\theta_0}^{-1} \frac{\partial K_{\theta_0}}{\partial \theta_i} K_{\theta_0}^{-1} \frac{\partial K_{\theta_0}}{\partial \theta_j} \right).$$

Sous les conditions 1 à 8, l'estimateur par maximum de vraisemblance est asympotiquement normal. Plus précisément :

$$\sqrt{n} M_{ML}^{1/2} \left(\hat{\theta}_{ML} - \theta_0 \right) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0, I_p).$$

De plus

$0 < \liminf_{n \to \infty} \lambda_{min}(M_{ML}) \leq \limsup_{n \to \infty} \lambda_{max}(M_{ML}) < +\infty.$

Krigeage sous le modèle estimé par EMV

Théorème

Sous les conditions 1 à 8, l'estimateur par Krigeage sous $\hat{\theta}_{ML}$ est asymptotiquement optimal :

$$\forall \mu \in \mathcal{W}_2(\mathbb{R}), \ \left| \hat{Y}_{\hat{\theta}_{ML}}(\mu) - \hat{Y}_{\theta_0}(\mu) \right| = o_{\mathbb{P}}(1).$$

Plan de la présentation

Performances sur des données simulées

- On note $m_k(\nu)$ le k-ième moment de ν et considère la fonction $F : \mathcal{W}_2(\mathbb{R}) \to \mathbb{R}$ telle que

$$F(\nu) = \frac{m_1(\nu)}{0.05 + \sqrt{m_2(\nu) - m_1(\nu)^2}}, \tag{7}$$

qu'on va chercher à interpoler.

Performances sur des données simulées

- On note $m_k(\nu)$ le k-ième moment de ν et considère la fonction $F : \mathcal{W}_2(\mathbb{R}) \to \mathbb{R}$ telle que

$$F(\nu) = \frac{m_1(\nu)}{0.05 + \sqrt{m_2(\nu) - m_1(\nu)^2}}, \tag{7}$$

qu'on va chercher à interpoler.

- On génère $\nu_1, \cdots, \nu_{100}$ qui sont des gaussiennes de moyennes et de variances tirées uniformément, perturbées aléatoirement afin d'exhiber des irrégularités

Performances sur des données simulées

- On note $m_k(\nu)$ le k-ième moment de ν et considère la fonction $F : \mathcal{W}_2(\mathbb{R}) \to \mathbb{R}$ telle que

$$F(\nu) = \frac{m_1(\nu)}{0.05 + \sqrt{m_2(\nu) - m_1(\nu)^2}}, \tag{7}$$

  qu'on va chercher à interpoler.
- On génère $\nu_1, \cdots, \nu_{100}$ qui sont des gaussiennes de moyennes et de variances tirées uniformément, perturbées aléatoirement afin d'exhiber des irrégularités
- Maximum de vraisemblance sur $\hat{\sigma}^2, \hat{\ell}, \hat{H}$ pour le modèle gaussien paramétrique

$$K_{\sigma^2,\ell,H}(\nu_1, \nu_2) = \sigma^2 \exp\left(-\frac{W_2(\nu_1, \nu_2)^{2H}}{\ell}\right). \tag{8}$$

Performances sur données simulées

- On évalue la méthode sur un dataset de test $(\nu_{t,i})_{i=500}$ généré de la même manière que les ν_i,

Performances sur données simulées

- On évalue la méthode sur un dataset de test $(\nu_{t,i})_{i=500}$ généré de la même manière que les ν_i, pour les critères :

$$RMSE^2 = \frac{1}{500}\sum_{i=1}^{500}\left(F(\nu_{t,i}) - \hat{F}(\nu_{t,i})\right)^2,$$

$$CIR_\alpha = \frac{1}{500}\sum_{i=1}^{n_t}\mathbf{1}\left\{\left|F(\nu_{t,i}) - \hat{F}(\nu_{t,i})\right| \leq q_\alpha \hat{\sigma}(\nu_{t,i})\right\},$$

Performances sur données simulées

- On évalue la méthode sur un dataset de test $(\nu_{t,i})_{i=500}$ généré de la même manière que les ν_i, pour les critères :

$$RMSE^2 = \frac{1}{500}\sum_{i=1}^{500}\left(F(\nu_{t,i}) - \hat{F}(\nu_{t,i})\right)^2,$$

$$CIR_\alpha = \frac{1}{500}\sum_{i=1}^{n_t}\mathbf{1}\left\{\left|F(\nu_{t,i}) - \hat{F}(\nu_{t,i})\right| \leq q_\alpha \hat{\sigma}(\nu_{t,i})\right\},$$

modèle	RMSE	$CIR_{0.9}$
"distribution"	0.094	0.92
"Legendre" ordre 5	0.49	0.92
"Legendre" ordre 10	0.34	0.89
"Legendre" ordre 15	0.29	0.91
"PCA" ordre 5	0.63	0.82
"PCA" ordre 10	0.52	0.87
"PCA" ordre 15	0.47	0.93

Travail sur des données météo avec un noyau sur le cylindre