La DTM-signature pour un test d'isomorphisme entre espaces métriques mesurés

Claire Brécheteau [1]

[1] *Inria Saclay, Palaiseau & Université Paris-Sud, Orsay – Université Paris-Saclay – claire.brecheteau@inria.fr*

Résumé. Nous introduisons la notion de DTM-signature. Il s'agit d'une mesure de probabilité sur $\mathbb{R}_+$ qui peut être associée à n'importe quel espace métrique. Cette signature est issue de la fonction distance à la mesure (DTM) introduite par Chazal, Cohen-Steiner et Mérigot (2011). On l'utilise pour construire une pseudo-métrique entre espaces métriques mesurés. Une borne supérieure pour une telle métrique est donnée par la distance de Gromov-Wasserstein. Des bornes inférieures peuvent être obtenues sous des hypothèses géométriques particulières.

Étant donnés deux N-échantillons, on se propose de construire un test statistique, à l'aide de la DTM-signature, permettant de rejeter l'hypothèse d'égalité des espaces métriques mesurés sous-jacents, à isométrie préservant la mesure près. Afin de justifier la validité du test, on utilisera les métriques de Wasserstein.

Mots-clés. Statistique computationnelle, sous-échantillonnage, tests statistiques, processus, espaces métriques mesurés, métriques de Wasserstein

Abstract. We introduce the notion of DTM-signature, a measure on $\mathbb{R}_+$ that can be associated to any metric-measure space. This signature is based on the distance to a measure (DTM) introduced by Chazal, Cohen-Steiner and Mérigot (2011). It leads to a pseudo-metric between metric-measure spaces, upper-bounded by the Gromov-Wasserstein distance. Under some geometric assumptions, we derive lower bounds for this pseudo-metric.

Given two N-samples, we also build an asymptotic statistical test based on the DTM-signature, to reject the hypothesis of equality of the two underlying metric-measure spaces, up to a measure-preserving isometry. We give strong theoretical justifications for this test involving Wasserstein metrics.

Keywords. Computational statistics, subsampling, statistical hypothesis testing, processes, metric-measure spaces, Wasserstein metrics

Parmi les nombreuses données disponibles, beaucoup peuvent être représentées par un ensemble de points dans un espace métrique. Une question naturelle qui se pose alors est de décider si deux ensembles de telles données sont similaires, i.e. si ces dernières sont issues de deux distributions semblables, si leurs formes sont proches, etc. Depuis les tests de Kolmogorov-Smirnov, de rangs-signés de Wilcoxon ou encore les t-tests, de nombreuses méthodes permettant de tester l'égalité des distributions ou des moyennes des distributions, etc. ont été étudiés. Des travaux récents sur le sujet comprennent les tests de Gretton et al. (2012) utilisant la divergence moyenne maximale et les espaces de Hilbert à noyaux reproduisants. Cependant, une telle comparaison peut se révéler compliquée dès lors que les données de sont pas issues du même espace métrique, ou simplement si les systèmes de coordonnées dans lesquels elles ont été mesurées sont différents. Afin de palier à ce problème, on peut faire abstraction de ces espaces métriques et ne conserver que les ensembles de points ainsi que les distances associées à chaque paire de points.

1 Comparaison d'espaces métriques mesurés

Le cadre statistique que nous adoptons est le suivant. Les données sont des réalisations de variables aléatoires indépendantes issues d'une même mesure de probabilité, à valeurs dans un espace métrique. On décide de ne faire aucune distinction entre deux espaces métriques mesurés lorsqu'ils sont égaux à isomorphisme près, comme défini ci-dessous.

Définition 1 *Un* ***espace métrique mesuré*** *est un triplet $(\mathcal{X}, \delta, \mu)$, avec $(\mathcal{X}, \delta)$ un espace métrique compact et μ une mesure de probabilité définie sur la tribu borélienne associée à $(\mathcal{X}, \delta)$.*

Définition 2 *Deux espaces métriques mesurés $(\mathcal{X}, \delta, \mu)$ et $(\mathcal{Y}, \gamma, \nu)$ sont dit isomorphes s'il existe deux boréliens $\mathcal{X}_0 \subset \mathcal{X}$ et $\mathcal{Y}_0 \subset \mathcal{Y}$ tels que $\mu(\mathcal{X} \backslash \mathcal{X}_0) = 0$ et $\nu(\mathcal{Y} \backslash \mathcal{Y}_0) = 0$, et une isométrie bijective $\phi : \mathcal{X}_0 \to \mathcal{Y}_0$ préservant la mesure, i.e. satisfaisant $\nu(\phi(A \cap \mathcal{X}_0)) = \mu(A \cap \mathcal{X}_0)$ pour tout borélien A de $\mathcal{X}$. Une telle application ϕ est appelée* ***isomorphisme*** *entre les espaces métriques mesurés $(\mathcal{X}, \delta, \mu)$ et $(\mathcal{Y}, \gamma, \nu)$.*

Afin de construire des tests de comparaison de deux espaces métriques mesurés à partir de deux échantillons, nous prenons le parti de définir une métrique ou pseudo-métrique entre de tels espaces, en gardant à l'esprit qu'une telle application se doit d'être stable relativement à certaines perturbations, stable par échantillonage, discriminante et facile à implémenter lorsque les espaces métriques considérés sont discrets.

Certaines distances entre espaces métriques mesurés ont été largement étudiées, on peut penser par exemple aux distances de Gromov-Wasserstein, voir Facundo Mémoli (2011). L'inconvénient de telles distances est que leur coût algorithmique est très élevé, quand bien même les espaces métriques seraient discrets. Voilà pourquoi nous n'utilisons pas cette distance, mais plutôt une pseudo-distance construite à partir de signatures, i.e.

d'objets associés à des espaces métriques mesurés, invariants par isomorphisme. Dans le papier de Mémoli (2011), il est proposé un panorama de signatures fréquemment utilisées.

La signature que nous considérons est une mesure de probabilité sur $\mathbb{R}^+$, et nous proposons de comparer deux signatures à l'aide des distances de Wasserstein.

Définition 3 *La **distance de Wasserstein** de paramètre $p \in [1, \infty)$ entre deux mesures boréliennes de probabilité μ et ν sur le même espace métrique $(\mathcal{X}, \delta)$ est définie par:*

$$W_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^2} \delta^p(x, y) \mathrm{d}\pi(x, y) \right)^{\frac{1}{p}},$$

*où $\Pi(\mu, \nu)$ représente l'ensemble des **plans de transport** entre μ et ν, c'est-à-dire l'ensemble des mesures de probabilité π sur $\mathcal{X} \times \mathcal{Y}$ telles que $\pi(A \times \mathcal{Y}) = \mu(A)$ et $\pi(\mathcal{X} \times B) = \nu(B)$ pour tous boréliens A dans $\mathcal{X}$ et B dans $\mathcal{Y}$.*

Les signatures ont largement été utilisées pour la classification de formes. On peut citer le papier de Osada, Funkhouser, Chazelle et Dobkin (2002). Pour autant, il semblerait qu'elles n'aient pas encore été utilisées dans le cadre des tests statistiques.

2 La DTM-signature

La signature que nous proposons ici est issue de la fonction distance à la mesure, introduite par Chazal, Cohen-Steiner et Mérigot (2011).

Soit $(\mathcal{X}, \delta)$ un espace métrique muni d'une mesure borélienne de probabilité μ. Soit m dans $[0, 1]$, la **fonction pseudo-distance** est définie en chaque point x de $\mathcal{X}$, par:

$$\delta_{\mu,m}(x) = \inf\{r > 0 \,|\, \mu(\overline{\mathrm{B}}(x, r)) > m\}.$$

La fonction **distance à la mesure** μ de paramètre de masse m notée $\mathrm{d}_{\mu,m}$ est alors définie pour tout x dans $\mathcal{X}$ par:

$$\mathrm{d}_{\mu,m}(x) = \frac{1}{m} \int_{l=0}^{m} \delta_{\mu,l}(x) \, \mathrm{d}l.$$

La distance à la mesure est une généralisation de la fonction distance au compact; voir le papier de Chazal, Cohen-Steiner et Mérigot (2011). Cette fonction est continue par rapport au paramètre de masse m, et lipschitzienne par rapport à μ.

Proposition 1 (cas $\mathbb{R}^d$: voir Chazal, Cohen-Steiner et Mérigot (2011)) *Soit $(\mathcal{X}, \delta, \mu)$ et $(\mathcal{Y}, \delta, \nu)$ deux espaces métriques mesurés inclus dans le même espace métrique, l'inégalité suivante est satisfaite :*

$$\|\mathrm{d}_{\mu,m} - \mathrm{d}_{\nu,m}\|_{\infty,\mathcal{X}\cup\mathcal{Y}} \leq \frac{1}{m} W_1(\mu, \nu).$$

De plus, pour une mesure empirique $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}$ sur un espace métrique $(\mathcal{X}, \delta)$, la distance à la mesure $\hat{\mu}_N$ de paramètre $\frac{k}{N}$ pour un k dans $[\![0, N]\!]$ en un point x de $\mathcal{X}$ satisfait:

$$\mathrm{d}_{\hat{\mu}_N, \frac{k}{N}}(x) = \frac{1}{k}\sum_{i=1}^{k}\delta(X^{(i)}, x),$$

où $X^{(1)}$, $X^{(2)}$, $\ldots X^{(k)}$ sont des k plus proches voisins de x parmi les N points X_1, X_2, $\ldots X_N$.

La distance à la mesure $\hat{\mu}_N$ est ainsi égale à la moyenne des distances aux k-plus proches voisins. En particulier, dans ce cas, le calcul de la distance à la mesure revient simplement à une recherche des k premiers plus proches voisins.

Définition 4 (DTM-signature) *La **DTM-signature** associée à un espace métrique mesuré $(\mathcal{X}, \delta, \mu)$, et notée $\mathrm{d}_{\mu,m}(\mu)$, est définie comme la loi de la variable aléatoire réelle positive $\mathrm{d}_{\mu,m}(X)$ où X est une variable aléatoire de loi μ.*

Remarquons qu'à deux espaces métriques mesurés isomorphes $(\mathcal{X}, \delta, \mu)$ et $(\mathcal{Y}, \gamma, \nu)$ seront associées les mêmes DTM-signatures. Aussi, la quantité $W_1(\mathrm{d}_{\mu,m}(\mu), \mathrm{d}_{\nu,m}(\nu))$ sera nulle. Grâce aux propriétés de régularité et de stabilité de la fonction distance à la mesure, la quantité précédente peut-être bien approximée par $W_1(\mathrm{d}_{\hat{\mu}_N,m}(\hat{\mu}_N), \mathrm{d}_{\hat{\nu}_N,m}(\hat{\nu}_N))$, voire par $\frac{T_{N,n,m}(\mu,\nu)}{\sqrt{n}} = W_1(\mathrm{d}_{\hat{\mu}_N,m}(\hat{\mu}_n), \mathrm{d}_{\hat{\nu}_N,m}(\hat{\nu}_n))$ pour $n \leq N$ avec $\hat{\mu}_N$ la mesure uniforme sur un N-échantillon de loi μ et $\hat{\mu}_n$ la mesure uniforme sur un sous-ensemble de n points du N-échantillon.

Pour construire le test, nous nous appuyons donc sur le fait simple que si deux espaces métriques mesurés sont isomorphes, alors la statistique $T_{N,n,m}(\mu, \nu)$ sera petite.

3 Le test

Étant donnés deux espaces métriques mesurés $(\mathcal{X}, \delta, \mu)$ et $(\mathcal{Y}, \gamma, \nu)$, on construit le test de l'hypothèse nulle

$$H_0 : \textit{Les espaces } (\mathcal{X}, \delta, \mu) \textit{ et } (\mathcal{Y}, \gamma, \nu) \textit{ sont isomorphes},$$

contre l'alternative :

$$H_1 : \textit{Les espaces } (\mathcal{X}, \delta, \mu) \textit{ et } (\mathcal{Y}, \gamma, \nu) \textit{ ne sont pas isomorphes}.$$

Pour $\alpha \in (0, 1)$, le test est le suivant :

$$\phi_N = \mathbb{1}_{T_{N,n,m}(\mu,\nu) \geq \hat{\mathrm{q}}_{\alpha,N,n}},$$

avec $\hat{\mathrm{q}}_{\alpha,N,n}$ un estimateur du quantile de la statistique $T_{N,n,m}(\mu,\nu)$ sous l'hypothèse H_0 défini comme le α-quantile de la loi

$$\mathcal{L}^* = \frac{1}{2}\sqrt{n}W_1(\mathrm{d}_{\hat{\mu}_N,m}(\mu_n^*), \mathrm{d}_{\hat{\mu}_N,m}(\mu'^*_n)) + \frac{1}{2}\sqrt{n}\, W_1(\mathrm{d}_{\hat{\nu}_N,m}(\nu_n^*), \mathrm{d}_{\hat{\nu}_N,m}(\nu'^*_n)).$$

Ici, μ_n^* et μ'^*_n sont deux mesures empiriques à partir de deux n-échantillons de loi $\hat{\mu}_N$, de même pour ν_n^* et ν'^*_n, à partir de ν.

Dans la suite, on choisira N de l'ordre de n^ρ. Alors, si la métrique est euclidienne, on montre que ce test est de niveau asymptotique α sous la condition $\rho > \frac{\max\{d,2\}}{2}$. De plus, sous cette hypothèse, on obtient une borne inférieure pour l'espérance de la distance de Wasserstein entre la loi de $T_{N,n,m}(\mu,\nu)$ sous H_0 et la loi $\mathcal{L}^*$ de l'ordre de $N^{\frac{1}{2\rho}-\frac{1}{\max\{d,2\}}}$.

Si l'on ajoute l'hypothèse que les mesures considérées soient (a,b)-standard pour a et b dans $\mathbb{R}^+$, c'est-à-dire satisfaisant : $\mu(\mathrm{B}(x,r)) \geq ar^b$ pour tout x dans le support de μ et tour rayon $r > 0$, alors on a de meilleures vitesses de convergence. Le test est de niveau asymptotique α dès lors que ρ est strictement supérieur à 1. De plus, l'espérance de la distance de Wasserstein est de l'ordre de $N^{\frac{1}{2\rho}-\frac{1}{2}}$.

Pour ce second cas, on utilisera les résultats de Chazal, Massart et Michel (2016) fournissant des vitesses de convergence de l'ordre de $\frac{1}{\sqrt{N}}$ à m fixé pour $\|\mathrm{d}_{\mu,m} - \mathrm{d}_{\hat{\mu}_N}\|_\infty$ dans le cas (a,b)-standard. Dans le cas général, on n'a pas mieux qu'une vitesse en $\frac{1}{N^{\frac{1}{\max\{d,2\}}}}$. Il s'agit de la vitesse de convergence de $W_1(\mu,\hat{\mu}_N)$ pour des mesures de probabilités générales dans $\mathbb{R}^d$ euclidien, établies dans le papier de Fournier et Guillin (2015); voir Bobkov et Ledoux (2014) pour le cas de la dimension 1.

Enfin, on montre que l'erreur de deuxième espèce est majorée par

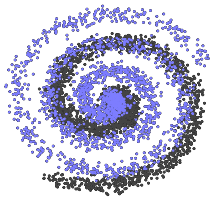
$$4\exp\left(-c\frac{W_1^2\left(\mathrm{d}_{\mu,m}(\mu), \mathrm{d}_{\nu,m}(\nu)\right)}{\max\left\{\mathcal{D}_{\mu,m}^2, \mathcal{D}_{\nu,m}^2\right\}}n\right),$$

si n est assez grand, avec c une constante positive et $\mathcal{D}_{\mu,m}$ le diamètre du support de $\mathrm{d}_{\mu,m}(\mu)$.

Le test est de plus facilement implémentable. Dans la suite on a généré deux 2000-échantillons de variables aléatoires de type $(R\sin(vR) + 0.03N, R\cos(vR) + 0.03N')$ avec R, N et N' indépendantes; N et N' de loi normale centrée réduite et R uniforme sur $]0,1[$; pour des valeurs de v différentes. Pour la loi μ, on choisira $v = 10$. On choisit $\alpha = 0.05$ et réalise les tests 1000 fois afin d'approcher la puissance et l'erreur de première espèce. On aura utilisé les paramètres $m = 0.05$ et $n = 20$, et approximé le quantile de la loi $\mathcal{L}^*$ par des méthodes de Monte-Carlo, en simulant 1000 réalisations de cette loi. On compare notre test (**DTM**) à un test de Kolmogorov-Smirnov (**KS**) sur deux $\frac{N}{2}$-échantillons de loi $\delta(X,X')$ (resp. $\gamma(X,X')$) avec X et X' indépendantes de loi μ (resp. ν).

v	15	20	30	40	100
erreur de type I **DTM**	0.050	0.049	0.051	0.044	0.051
puissance **DTM**	0.525	0.884	0.987	0.977	0.985
puissance **KS**	0.768	0.402	0.465	0.414	0.422

Figure 1: Erreur de Type I et puissance



Les méthodes présentées ici sont facilement implémentables et viennent avec des justifications théoriques. Elles pourraient être généralisées à d'autres signatures. Enfin, le choix des meilleurs paramètres à utiliser en pratique, m et n, reste une question ouverte qu'il pourrait être intéressant de traîter dans des travaux futurs.

Bibliographie

[1] Frédéric Chazal, David Cohen-Steiner et Quentin Mérigot (2011), Geometric inference for probability measures, *Foundations of Computational Mathematics* 11(6), 733–751.
[2] Facundo Mémoli (2011), Gromov-Wasserstein distances and the metric approach to object matching, *Foundations of Computational Mathematics* 11(4), 417–487.
[3] Frédéric Chazal, Pascal Massart et Bertrand Michel (2016), Rates of convergence for robust geometric inference, *Electronic Journal of Statistics* 10(2), 2243–2286.
[4] Nicolas Fournier et Arnaud Guillin (2015), On the rate of convergence in Wasserstein distance on the empirical measure, *Probability Theory & Related Fields* 162, 707–738.
[5] Sergey Bobkov et Michel Ledoux (2014), One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *unpublished*.
[6] Arthur Gretton et al. (2012), A Kernel Two-Sample Test, *Journal of Machine Learning Research* 13, 723-773.
[7] Robert Osada, Thomas Funkhouser, Bernard Chazelle et David Dobkin (2002), Shape distributions, *ACM Transactions on Graphics* 21, 807–832.