Problèmes d'adéquations entre distributions : une approche par un modèle de déformations et la distance de Wasserstein.

Hélène Lescornel[1], Eustasio del Barrio[2] & Jean-Michel Loubes[3]

[1] *INRIA Saclay, 1 rue Honoré d'Estienne d'Orves, 91 120 Palaiseau.*
helene.lescornel@inria.fr
[2] *Universitad de Valladolid, Facultad de Sciencas, C/ Prado de la Magdalena s/n, 47005 Valladolid, ESPAGNE.*
tasio@eio.uva.es
[3] *Institut de Mathématiques de Toulouse, 118, route de Narbonne F-31062 Toulouse Cedex 9.*
loubes@math.univ-toulouse.fr

Résumé. Nous proposons des procédures de tests d'adéquation de données à un modèle de déformations de distributions.

Le problème considéré est le suivant.

On dispose de J échantillons indépendants constitués chacun de n variables aléatoires identiquement distribuées :

$$\left\{(X_{i,1})_{1\leqslant i\leqslant n}\,,\ \text{où } X_{i,1}\sim\mu_1;\quad \ldots\quad ;(X_{i,J})_{1\leqslant i\leqslant n}\,,\ \text{où } X_{i,J}\sim\mu_J.\right.$$

Le but est de savoir si, pour un ensemble de fonctions de déformations inversibles $\mathcal{G}=\mathcal{G}_1\times\cdots\times\mathcal{G}_J$, il existe $\varphi^\star=(\varphi_1^\star,\ldots,\varphi_J^\star)\in\mathcal{G}$ et des variables aléatoires indépendantes et de même loi $(\varepsilon_{i,j})_{\substack{1\leqslant i\leqslant n\\ 1\leqslant j\leqslant J}}$ telles que

$$X_{i,j}=\left(\varphi_j^\star\right)^{-1}(\varepsilon_{i,j})\,,\quad \forall\ 1\leqslant i\leqslant n,\ 1\leqslant j\leqslant J. \tag{1}$$

Pour cela, nous introduisons les variables aléatoires

$$\{Z_{i,1}(\varphi)=\varphi_j(X_{i,1})\sim\mu_1(\varphi)\,;\quad \ldots\quad ;Z_{i,J}(\varphi)=\varphi_j(X_{i,J})\sim\mu_J(\varphi)\,;\ 1\leqslant i\leqslant n.$$

En faisant varier $\varphi=(\varphi_1,\ldots,\varphi_J)$ dans l'ensemble $\mathcal{G}$, nous allons chercher à aligner les distributions $\mu_j(\varphi)$ pour $1\leqslant j\leqslant J$.

Pour quantifier l'alignement de ces mesures nous utilisons la distance de Wasserstein sur l'ensemble des probabilités sur $\mathbb{R}$ admettant un moment d'ordre 2, ensemble noté $\mathcal{W}_2(\mathbb{R})$. Si $\mu,\nu\in\mathcal{W}_2(\mathbb{R})$, cette distance a pour expression

$$W_2^2(\mu,\nu)=\int_0^1\left(F_\mu^{-1}(t)-F_\nu^{-1}(t)\right)^2dt,$$

où F_η^{-1} désigne la fonction quantile associée à la loi η, (l'inverse généralisé de sa fonction de répartition).

Cependant, nous considérons ici J distributions, nous devons donc introduire une mesure "globale" de séparation entre les $\mu_j(\varphi)$. Nous proposons pour cela d'utiliser leur barycentre :

$$\mu_B(\varphi) \in \arg\min_{\mu \in \mathcal{W}_2(\mathbb{R})} \left\{ \mu \mapsto \frac{1}{J} \sum_{j=1}^{J} W_2^2(\mu_j(\varphi), \mu) \right\}.$$

Dans le cas où les distributions sont réelles, la loi du barycentre est connue : sa fonction quantile est la moyenne des fonctions quantiles des $\mu_j(\varphi)$.

Ainsi, notre critère d'alignement théorique est $M : \varphi \mapsto \frac{1}{J} \sum_{j=1}^{J} W_2^2(\mu_j(\varphi), \mu_B(\varphi))$.

Cependant, cette version est inaccesssible. Nous remplaçons donc les lois $(\mu_j(\varphi))_{1 \leqslant j \leqslant J}$ par leurs versions empiriques obtenues à l'aide des observations, et obtenons le critère suivant :

$$M^n : \varphi \mapsto \frac{1}{J} \sum_{j=1}^{J} W_2^2\left(\mu_j^n(\varphi), \mu_B^n(\varphi)\right) = \frac{1}{J} \sum_{j=1}^{J} \int_0^1 \left(\varphi_j\left(\left(F_j^n\right)^{-1}(t)\right) - \left(F_B^n\right)^{-1}(\varphi)(t)\right)^2 dt,$$

où $\left(F_B^n\right)^{-1}(\varphi)(t) = \frac{1}{J} \sum_{k=1}^{J} \varphi_k \circ \left(F_k^n\right)^{-1}(t)$, et $\varphi_k \circ \left(F_k^n\right)^{-1}$ est la fonction quantile empirique associée à l'échantillon $(Z_{i,k}(\varphi))_{1 \leqslant i \leqslant n} = (\varphi_k(X_{i,k}))_{1 \leqslant i \leqslant n}$.

La statistique de test proposée est $\inf_{\mathcal{G}} M^n$, et le but est de déterminer sa distribution asymptotique.

Sous des conditions de régularité sur les fonctions de déformation et sur les distributions $(\mu_j)_{1 \leqslant j \leqslant J}$, nous obtenons un premier résultat de convergence en distribution de $\sqrt{n} \inf_{\mathcal{G}} M^n$ vers une loi Gaussienne.

Cependant, lorsque le modèle (1) est valide, ce résultat donne une convergence en probabilité vers 0.

Dans ce cas, nous devons changer de point de vue, tout d'abord en considérant un modèle paramétrique pour les déformations. Dans ce cas, on suppose connue la forme de la déformation, mais pas son importance représentée par un paramètre $\theta_j^\star \in \Theta_j \subset \mathbb{R}^d$. Ainsi, $\varphi_j^\star = \varphi_{\theta_j^\star}$, et $\mathcal{G}$ est représenté par $\Theta \subset \mathbb{R}^{dJ}$. Pour un résultat plus précis nous devons également spécifier la loi de ε, ce qui peut se faire dans le cadre général en supposant par exemple $\theta_1^\star$ connu.

Nous obtenons alors, en renforçant les hypothèses, un résultat de convergence plus fort nous donnant la loi asymptotique de $n \inf_{\Theta} M^n$. Mais cette fois la loi limite n'est pas gaussienne. Pour estimer ses quantiles, nous utilisons une propriété de convergence d'une version bootstrappée de notre critère empirique.

Ainsi, avec ces deux résultats, nous pouvons envisager différentes procédures de tests d'un niveau asymptotique fixé.

Mots-clés. Modèle de déformations, Distance de Wasserstein, Test.

Abstract. We propose procedures for Goodness-of-fit tests to a model of deformations between distributions.

The problem is the following.

We have at hand J independent samples made of n independent random variables having the same distribution:

$$\left\{(X_{i,1})_{1\leqslant i\leqslant n}\text{, where } X_{i,1}\sim\mu_1;\quad \ldots\quad ;(X_{i,J})_{1\leqslant i\leqslant n}\text{, where } X_{i,J}\sim\mu_J.\right.$$

Given a set of invertible deformation functions $\mathcal{G}=\mathcal{G}_1\times\cdots\times\mathcal{G}_J$, our aim is to know if there exist $\varphi^\star=(\varphi_1^\star,\ldots,\varphi_J^\star)\in\mathcal{G}$ and independent, identically distributed random variables $(\varepsilon_{i,j})_{\substack{1\leqslant i\leqslant n\\1\leqslant j\leqslant J}}$ such that

$$X_{i,j}=\left(\varphi_j^\star\right)^{-1}(\varepsilon_{i,j}),\quad \forall\ 1\leqslant i\leqslant n,\ 1\leqslant j\leqslant J. \tag{2}$$

To answer this question, we introduce the following quantities

$$\{Z_{i,1}(\varphi)=\varphi_j(X_{i,1})\sim\mu_1(\varphi);\quad \ldots\quad ;Z_{i,J}(\varphi)=\varphi_j(X_{i,J})\sim\mu_J(\varphi);\ 1\leqslant i\leqslant n.$$

By varying $\varphi=(\varphi_1,\ldots,\varphi_J)$ in $\mathcal{G}$, we will try to align the distributions $\mu_j(\varphi)$ for $1\leqslant j\leqslant J$.

To measure the alignment of the distributions, we will use the Wasserstein distance on the set of probabilities on $\mathbb{R}$ with a moment of order 2, denoted by $\mathcal{W}_2(\mathbb{R})$. If $\mu,\nu\in\mathcal{W}_2(\mathbb{R})$, their Wasserstein distance is given by

$$W_2^2(\mu,\nu)=\int_0^1\left(F_\mu^{-1}(t)-F_\nu^{-1}(t)\right)^2dt,$$

where F_η^{-1} is the quantile function associated with η, (the generalized inverse of the distribution function).

However, we have at hand J distributions, then we have to consider a global measure of separation between the $\mu_j(\varphi)$'s. We will use their barycenter, defined as:

$$\mu_B(\varphi)\in\arg\min_{\mu\in\mathcal{W}_2(\mathbb{R})}\left\{\mu\mapsto\frac{1}{J}\sum_{j=1}^J W_2^2(\mu_j(\varphi),\mu)\right\}.$$

In the case of distributions on $\mathbb{R}$, the distribution of the barycenter is known: it has for quantile function the mean average of the quantile functions of the $\mu_j(\varphi)$'s.

Thus we have an alignment criterion $M:\varphi\mapsto\frac{1}{J}\sum_{j=1}^J W_2^2(\mu_j(\varphi),\mu_B(\varphi))$.

However we can not compute this theoretical version. We have to replace the distributions $(\mu_j(\varphi))_{1\leqslant j\leqslant J}$ by their empirical versions. Then we get the following criterion

$$M^n:\varphi\mapsto\frac{1}{J}\sum_{j=1}^J W_2^2\left(\mu_j^n(\varphi),\mu_B^n(\varphi)\right)=\frac{1}{J}\sum_{j=1}^J\int_0^1\left(\varphi_j\left(\left(F_j^n\right)^{-1}(t)\right)-\left(F_B^n\right)^{-1}(\varphi)(t)\right)^2dt,$$

where $\left(F_B^n\right)^{-1}(\varphi)(t) = \frac{1}{J}\sum_{k=1}^{J} \varphi_k \circ \left(F_k^n\right)^{-1}(t)$ and $\varphi_k \circ \left(F_k^n\right)^{-1}$ is the empirical quantile function associated with $\left(Z_{i,k}(\varphi)\right)_{1\leqslant i\leqslant n} = \left(\varphi_k(X_{i,k})\right)_{1\leqslant i\leqslant n}$.

The test statistic that we consider is then $\inf_{\mathcal{G}} M^n$, and we have to determinate its asymptotic distribution.

Under regularity conditions on deformation functions and on the μ_j's, we obtain the convergence in distribution of $\sqrt{n}\inf_{\mathcal{G}} M^n$ to a random Gaussian variable.

However, when the model (2) holds, this result gives a convergence in probability to 0.

In this case, we have to change our point of view, first by considering a parametric model for the deformations. More precisely, we assume that we know the shape of the deformation but not the amount of deformation which is represented by a parameter $\theta_j^\star \in \Theta_j \subset \mathbb{R}^d$. Then $\varphi_j^\star = \varphi_{\theta_j^\star}$, and $\mathcal{G}$ is represented by $\Theta \subset \mathbb{R}^{dJ}$.

To get sharper results we have also to specify the distribution of ε. This can be done in the general case by assuming for instance that $\theta_1^\star$ is known.

Then in this case, with stronger assumptions on the distributions μ_j's, we get the asymptotic distribution of $n\inf_{\Theta} M^n$. However here the limit distribution is no more Gaussian. Then, to estimate its quantiles, we use a convergence property for a bootstrap version of our empirical criterion. Then, with these two results, we are able to propose different test procedures with a given asymptotic level.

Keywords. Deformation model, Wasserstein distance, Tests.

1 Description de la communication

Nous proposons des procédures de tests d'adéquation de données à un modèle de déformations de distributions.

Dans un premier temps, nous détaillons le problème considéré et le point de vue que nous avons adopté pour en donner une solution.

On dispose de J échantillons indépendants constitués chacun de n variables aléatoires identiquement distribuées :

$$\left\{(X_{i,1})_{1\leqslant i\leqslant n},\ \text{où } X_{i,1}\sim\mu_1;\quad \ldots\quad ;(X_{i,J})_{1\leqslant i\leqslant n},\ \text{où } X_{i,J}\sim\mu_J. \right. \tag{3}$$

Le but est de savoir si, pour un ensemble de fonctions de déformations inversibles $\mathcal{G} = \mathcal{G}_1\times\cdots\times\mathcal{G}_J$, il existe $\varphi^\star = (\varphi_1^\star,\ldots,\varphi_J^\star) \in \mathcal{G}$ et des variables aléatoires indépendantes et de même loi $(\varepsilon_{i,j})_{\substack{1\leqslant i\leqslant n\\ 1\leqslant j\leqslant J}}$ telles que

$$X_{i,j} = \left(\varphi_j^\star\right)^{-1}(\varepsilon_{i,j}),\quad \forall\ 1\leqslant i\leqslant n,\ 1\leqslant j\leqslant J. \tag{4}$$

Pour cela, nous introduisons les variables aléatoires

$$\left\{Z_{i,1}(\varphi) = \varphi_j(X_{i,1})\sim\mu_1(\varphi);\quad \ldots\quad ;Z_{i,J}(\varphi) = \varphi_j(X_{i,J})\sim\mu_J(\varphi);\ 1\leqslant i\leqslant n.\right. \tag{5}$$

En faisant varier φ dans l'ensemble $\mathcal{G}$, nous allons chercher à aligner les distributions $\mu_j(\varphi)$ pour $1 \leqslant j \leqslant J$. En effet, le modèle (4) est valide s'il existe φ_0 ($\varphi^\star$, par exemple) et une loi μ tels que $\mu_j(\varphi_0) = \mu$ pour tout j, $1 \leqslant j \leqslant j$.

Pour quantifier l'alignement de ces mesures nous utilisons la distance de Wasserstein sur l'ensemble des probabilités sur $\mathbb{R}$ admettant un moment d'ordre 2, ensemble noté $\mathcal{W}_2(\mathbb{R})$. Si $\mu, \nu \in \mathcal{W}_2(\mathbb{R})$, cette distance a pour expression

$$W_2^2(\mu,\nu) = \int_0^1 \left(F_\mu^{-1}(t) - F_\nu^{-1}(t)\right)^2 dt, \tag{6}$$

où F_η^{-1} désigne la fonction quantile associée à la loi η, (l'inverse généralisé de sa fonction de répartition).

Cependant, nous considérons ici J distributions, nous devons donc introduire une mesure "globale" de séparation entre les $\mu_j(\varphi)$. Nous proposons pour cela d'utiliser leur barycentre :

$$\mu_B(\varphi) \in \arg\min_{\mu\in\mathcal{W}_2(\mathbb{R})} \left\{ \mu \mapsto \frac{1}{J}\sum_{j=1}^{J} W_2^2(\mu_j(\varphi), \mu) \right\}.$$

Cette notion de barycentre pour des mesures dans $\mathbb{R}^k$, $k \geqslant 1$, est étudiée dans [1].

Dans le cas où les distributions sont réelles, la loi du barycentre est connue : sa fonction quantile est la moyenne des fonctions quantiles des $\mu_j(\varphi)$.

Ainsi, notre critère d'alignement théorique est $M : \varphi \mapsto \frac{1}{J}\sum_{j=1}^{J} W_2^2(\mu_j(\varphi), \mu_B(\varphi))$.

Cependant, cette version théorique est inaccesssible. Nous remplaçons donc les lois $\mu_j(\varphi)$ par leurs versions empiriques obtenues à l'aide des observations, et obtenons le critère suivant :

$$M^n : \varphi \mapsto \frac{1}{J}\sum_{j=1}^{J} W_2^2\left(\mu_j^n(\varphi), \mu_B^n(\varphi)\right) = \frac{1}{J}\sum_{j=1}^{J}\int_0^1 \left(\varphi_j\left(\left(F_j^n\right)^{-1}(t)\right) - \left(F_B^n\right)^{-1}(\varphi)(t)\right)^2 dt,$$

où $\left(F_B^n\right)^{-1}(\varphi)(t) = \frac{1}{J}\sum_{k=1}^{J} \varphi_k \circ \left(F_k^n\right)^{-1}(t)$, et $\varphi_k \circ \left(F_k^n\right)^{-1}$ est la fonction quantile empirique associée à l'échantillon $\left(Z_{i,k}(\varphi)\right)_{1\leqslant i\leqslant n} = \left(\varphi_k(X_{i,k})\right)_{1\leqslant i\leqslant n}$.

La statistique de test proposée est $\inf_{\mathcal{G}} M^n$, et le but est de déterminer sa distribution asymptotique.

Dans la seconde partie de l'exposé, nous présentons le résultat suivant obtenu sous des conditions de régularité sur les fonctions de déformation et sur les distributions $(\mu_j)_{1\leqslant j\leqslant J}$.

Théorème 1.

$$\sqrt{n}\left\{\inf_{\mathcal{G}} M^n - \inf_{\mathcal{G}} M\right\} \rightharpoonup$$
$$\inf_{\varphi\in\mathcal{G}} \frac{2}{J}\sum_{j=1}^{J}\int_0^1 \varphi_j'\left(F_j^{-1}(t)\right)\frac{B_j(t)}{f_j\left(F_j^{-1}(t)\right)}\left(\varphi_j\left(F_j^{-1}(t)\right) - F_B^{-1}(\varphi)(t)\right) dt,$$

où F_j est la fonction de répartition de la loi μ_j, $F_B^{-1}(\varphi)(t) = \frac{1}{J}\sum_{k=1}^{J} \varphi_k \circ F_k^{-1}(t)$, et $(B_j)_{1\leqslant j\leqslant J}$ sont des ponts browniens standards indépendants.

Ce théorème est obtenu via un résultat d'approximation forte du processus quantile par une suite de ponts browniens suivant la procédure utlisée dans [3]. Cependant, lorsque le modèle (4) est valide, ce résultat donne une convergence en probabilité vers 0, ne nous renseignant pas sur la distribution asymptotique de notre statistique de test.

Dans un troisième temps, nous allons donc changer de point de vue, tout d'abord en considérant un modèle paramétrique pour les déformations. Dans ce cas, on suppose connue la forme de la déformation, mais pas son importance représentée par un paramètre $\theta_j^\star \in \Theta_j \subset \mathbb{R}^d$. Ainsi, $\varphi_j^\star = \varphi_{\theta_j^\star}$, et $\mathcal{G}$ est représenté par $\Theta \subset \mathbb{R}^{dJ}$.

Le théorème 1 reste vrai dans ce cadre, sous des hypothèses plus faibles en ce qui concerne la régularité des distributions. Cependant pour un résultat plus précis nous devons également spécifier la loi de ε, ce qui peut se faire dans le cadre général en supposant par exemple $\theta_1^\star$ connu.

Nous obtenons alors, en renforçant les hypothèses, un résultat de convergence plus fort nous donnant la loi asymptotique de $n \inf_\Theta M^n$. Mais cette fois la loi limite n'est pas gaussienne ; pour estimer ses quantiles, nous utilisons une propriété de convergence d'une version bootstrappée de notre critère empirique.

Ainsi, avec ces deux résultats, nous pouvons envisager différentes procédures de tests d'un niveau asymptotique fixé que nous exposons en dernier lieu.

Il est également possible de considérer le cas où les données sont d-dimensionnelles : par exemple, dans le cas où les données proviennent du modèle (4) dans sa version paramétrique, c'est à dire $\varphi_j^\star = \varphi_{\theta_j^\star}$, [2] propose une procédure d'estimation du paramètre de déformation $\theta^\star$ lorsque les observations ne sont plus à valeurs réelles.

Bibliographie

[1] Agueh, M. et Carlier, G. (2011), Barycenters in the Wasserstein space, *SIAM Journal on Mathematical Analysis*, 43(2):904924.

[2] Agullò, M., Cuesta-Albertos, J. A., Lescornel, H. et Loubes, J.-M. (2015), A Parametric Registration model for warped distributions with Wasserstein's distance, *Journal of Multivariate Analysis*, à paraître.

[3] Alvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J. A. et Matrán, C. (2008), Trimmed comparison of distributions, *Journal of the American Statistical Association*, 103(482):697704.