A prelude to statistics in Wasserstein metric spaces

Chon Van Le
International University, Vietnam National University – Ho Chi Minh City, Ho Chi Minh City, Vietnam, and
Uyen Hoang Pham
University of Economics and Law, Vietnam National University – Ho Chi Minh City, Ho Chi Minh City, Vietnam

Received 8 October 2023
Revised 17 October 2023
19 October 2023
31 October 2023
Accepted 6 November 2023

Abstract

Purpose – This paper aims mainly at introducing applied statisticians and econometricians to the current research methodology with non-Euclidean data sets. Specifically, it provides the basis and rationale for statistics in Wasserstein space, where the metric on probability measures is taken as a Wasserstein metric arising from optimal transport theory.

Design/methodology/approach – The authors spell out the basis and rationale for using Wasserstein metrics on the data space of (random) probability measures.

Findings – In elaborating the new statistical analysis of non-Euclidean data sets, the paper illustrates the generalization of traditional aspects of statistical inference following Frechet's program.

Originality/value – Besides the elaboration of research methodology for a new data analysis, the paper discusses the applications of Wasserstein metrics to the robustness of financial risk measures.

Keywords Frechet mean sets, Histogram data sets, Optimal transport, Random probability measures, Robustness of financial risk measures, Wasserstein metrics, Wasserstein sampling spaces, WGAN

Paper type Research paper

1. Introduction

As we are witnessing the current extension of statistical analysis to more general data sets in data science, it is about time to let applied statisticians and econometricians be aware of this useful and important phenomenon. The cornerstone of statistical theory for applications is data. Traditionally, data are elements of Euclidean spaces which are naturally equipped with Euclidean distances which are essential for analysis. Modern applications call for more general data sets, such as histograms or non-Euclidean data. To use statistics to make predictions and decisions with this new type of data, we need to extend traditional statistical theory. The first basic ingredient to generalize is metrics on new data sets. This short note aims simply at elaborating a bit on a popular new metric which applied econometricians can learn to apply to their empirical applications from current research literature. This popular new metric is called Wasserstein metric (distance) which is shown to be suitable for a variety of non-Euclidean data space, such as Wasserstein space which is a space of probability distributions equipped with a Wasserstein metric.

Simple and elementary examples will serve as illustrating the usefulness and rationale of modern statistics with non-Euclidean data. The note elaborates theoretical aspects in simple

JEL Classification — C10

Asian Journal of Economics and Banking
Vol. 8 No. 1, 2024
pp. 54-66
Emerald Publishing Limited
e-ISSN: 2633-7991
p-ISSN: 2615-9821
DOI 10.1108/AJEB-10-2023-0099

settings, as well as mentioning some concrete applications. Our purpose is simply introducing applied statisticians and econometricians to modern data analysis based upon statistical theory.

The paper is organized as follows. In Section 2, we elaborate on Wasserstein metrics in a concrete data set consisting of (random) histograms which are probability measures, together with the notion of Wasserstein metrics. In Section 3, we touch upon the starting point to generalize traditional statistics in Euclidean spaces to Wasserstein spaces. In Section 4, we mention an application of Wasserstein metrics to the robustness issue of financial risk management. Section 5 provides the conclusions.

2. Wasserstein metrics on histogram data sets

We can take it as self-evidence that statistics is based on data. While we do have a general theory of statistics to guide us each time we need statistics, there is something hidden in the practices of statistics that we start looking at nowadays.

Traditionally, most of our data are Euclidean elements and in practicing statistics on $\mathbb{R}^k$, we take for granted their Euclidean distances $||.||_k$, without bothering spelling out that our data set is a metric space (which is, in fact, essential for all statistical investigations, such as comparing data points, summarizing observed data sample).

Before our times, i.e. before we actually run into modern applications where our data could be non-Euclidean, Maurice Frechet has forseen the future (i.e. nowadays) for us. Indeed, recognizing that our traditional data space is the metric space $(\mathbb{R}^d, ||.||_d)$, Frechet (1906) first axiomatized the notion of a metric on arbitrary spaces, to have rigorous metric spaces, not only for mathematical functional calculus, but specifically for probability and statistics.

A well-known situation for all statisticians where "data points" are non-Euclidean is this. Let $X_1, X_2, \ldots, X_n$ be an observed (IID) random sample drawn from a real-valued random variable (population) X whose distribution function F is unknown. To improve the classical practices (e.g. estimating some population parameters of interest), and to take into account the advantages of computer science, the method of bootstrap was invented to improve the accuracy of estimators and their confidence intervals. The method consists of creating new "data points" via simulations.

Specifically, given the observed sample $X_1, X_2, \ldots, X_n$, we obtain the known empirical distribution function (but, ex ante, it is a random distribution function):

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{(-\infty,x]}(X_j)$$

whose corresponding probability measure (law) is $dF_n(.) = \frac{1}{n}\sum_{j=1}^{n}\partial_{X_j}(.)$ (by Lebesgue-Stieltjes Theorem) where $\partial_{X_j}(.) = 1_{(.)}(X_j)$ is the (random) Dirac probability measure at X_j on $\mathcal{B}(\mathbb{R})$.

Having the known probability measure dF_n, we can create simulated data from it via $F_n^{-1}(U)$, where

$$F_n^{-1}(.) : [0,1] \to \mathbb{R}, \quad F_n^{-1}(u) = \inf\{x \in \mathbb{R} : F_n(x) \geq u\}$$

is the (univariate) quantile function of F_n and U is the random variable uniformly distributed on [0,1].

Roughly speaking, a simulated sample (a new "data point") is obtained as a result of drawing with replacement n points from the set $\{X_1, X_2, \ldots, X_n\}$, say, m times, resulting in m sets $B_k = \{b_{1,k}, b_{2,k}, \ldots, b_{n,k}\}$, $k = 1, 2, \ldots, m$.

Because of the drawings with replacement, the elements $b_{j,k}$ in each B_k could be equal, i.e. appearing more than once in it, so that each new "data point" B_k is not really a subset of n elements of $\mathbb{R}$ as in set theory. Instead, each B_k is a multiset, i.e. a collection of points distinct or not (multiplicities of occurences are allowed).

As a remark, such a collection of n points $b_{j,k}, j = 1, 2, \ldots, n$, can be viewed as a fuzzy subset of $\mathbb{R}$, consisting of distinct points whose degrees of membership are equal to the ratios of their multiplicity of occurence and the size n.

But, in the setting of statistics, it is more representative if we view the new "data points" B_k as a histogram (a random probability measure on $\mathcal{B}(\mathbb{R})$), so that our new data set is a space of (random) probability measures denoted as $\mathcal{P}(\mathbb{R})$ where each "data point" is not an element of the Euclidean space $\mathbb{R}$, but is a probability measure on the metric space $(\mathbb{R}, |.|)$.

Data sets which are (random) probability measures on a metric space $(\mathcal{X}, \rho)$ abound in applications. As such, we need a suitable metric between probability measures.

Remark. But we know well that a large part of probability theory was about precisely the metrization of weak convergence of probability measures on metric spaces, i.e. producing metrics on the space of probability measures, see, e.g. Billingsley (1995), Parthasarathy (1967). Can we just pick some known metric among, say, Levy, Prokhorov, Total Variation metrics to use? Well, it depends on what we want our chosen metric to "behave!" So far, metrics on probability measures are invented to study asymptotic sampling distributions, such as in the Central Limit Theorem. They were not invented to handle data analysis, in which we need, for example, to use a suitable metric to compare probability measures (as data points in our new data set of an application). For example, if we observe three data "points" as three probability densities f, g, h which are uniformly distributed on $[-3, -2]$, $[-2, -1]$ (Bernton *et al.*, 2019; Bhat and Prashanth, 2019), respectively, (and denoting F as the distribution function with density f and dF its associated probability measure) then

$$TV(dF, dG) = \frac{1}{2}\int_{-\infty}^{\infty} |f(x) - g(x)|dx = 1 = TV(dF, dH)$$

i.e. the total variation metric cannot capture the locations of these histogram data.

This is similar to the recognition that Hausdorff distance on subsets of a space cannot be used when data are curved in the space, although curves are subsets. The reason is clear: Hausdorff distance does not capture the structure of curves which is needed in data analysis when curves are data "points".

So, what are other metrics (on space of probability measures) which can be used for data analysis/statistics with data sets as histograms?

We need to compare histograms (as data points) in applications when each histogram represents the observed information about an "object", or the return of a stock in financial econometrics. Then it is obvious that we must take into account of their locations! On the other hand, if data points are elements of an Euclidean space, e.g. $x, y \in (\mathbb{R}, |.|)$, the suitable metric W we wish to have should be a natural extension of the Euclidean metric $|.|$ on $\mathbb{R}$, in the sense that $W(\partial_x, \partial_y) = |x - y|$, i.e. when we identify a number $x \in \mathbb{R}$ with the Dirac probability measure ∂_x.

We are going to "mention" a suitable and popular metric $W(.,.)$ on histogram data. It seems important for applied statisticians and econometricians to have a good understanding of that metric to feel comfortable to use it in real-world applications, rather than just take it for granted!.

The following elaboration is for this purpose.

As far as history is concerned, it is fair to start with Maurice Frechet, the pioneer of modern statistics.

In 1937, Levy (1937) defined several metrics on probability measures on $\mathcal{B}(\mathbb{R})$. One is

$$L(F,G)=\inf\{\varepsilon>0: G(x-\varepsilon)-\varepsilon\leq F(x)\leq G(x+\varepsilon)+\varepsilon,\ \nabla x\in\mathbb{R}\}$$

which metrized the convergence in distribution (or weak convergence of probability measures, i.e. $F_n\overset{w}{\rightarrow}F$ if $F_n(x)\rightarrow F(x)$, as $n\rightarrow\infty$, for any $x\in C(F)$, the continuity set of $F(.)$) i.e. $F_n\overset{w}{\rightarrow}F\Leftrightarrow L(F_n,F)\rightarrow 0$.

Pursuing Levy's work, in 1957, Frechet (1957) observed that Levy's distance $L(F, G)$ of the distribution functions of two random variables X and Y involved F and G alone. He suggested that a "global" distance $W_H(X, Y)$ should involve the joint distribution function $H(x, y)$ of the random vector (X, Y), say, $W_H(F, G)$ where $H(.,.):\mathbb{R}^2\rightarrow[0,1]$ is the joint distribution with marginals F, G, i.e. $H(x,\infty)=F(x)$, $H(\infty,y)=G(y)$.

Another definition of Levy's distance on distribution functions on $\mathbb{R}$ is of the form

$$W(F,G)=\inf\{W_H(F,G): H\in C(F,G)\}$$

where $C(F, G)$ is the set of joint distributions with marginals F, G (later in 1959, Abe Sklar specified it as copulas).

But for $W(F, G)$ to be a bona fide "metric" (in particular, $W(F, G)\Leftrightarrow F=G$), the above infimum must be attained at some special H^*.

Let's see whether it is the case or not for the example given in Frechet (1957)

$$W_H(F,G)=\sqrt{E_H(X-Y)^2}=\sqrt{\int_{\mathbb{R}^2}(x-y)^2dH(x,y)}$$

Remark. In 1969, Vassershtein (Wasserstein) (1969) proposed exactly

$$W_1(\mu,\nu)=\inf\left\{\int_{\mathbb{R}^2}\left|x-y\right|d\lambda(x,y):\lambda\in\Pi(\mu,\nu)\right\}$$

where $\Pi(\mu,\nu)$ is the set of probability measures on $\mathcal{B}(\mathbb{R}^2)$ with projections (marginals) μ, ν.

Upfront: $\inf\left\{\sqrt{E_H(X-Y)^2}\ : H\in C(F,G)\right\}$ is attained at $H^*(x,y)=F(x)\wedge G(y)$ because, for U uniformly distributed on $[0, 1]$, $X\overset{D}{=}F^{-1}(U)$, $Y\overset{D}{=}G^{-1}(U)$, the joint distribution function of $(F^{-1}(U), G^{-1}(U))$ is H^* and

$$W_{H^*}(F,G)=W_{H^*}\left(F^{-1}(U),G^{-1}(U)\right)=\sqrt{\int_0^1\left(F^{-1}(u)-G^{-1}(u)\right)^2du}$$

which is the minimum of $\left\{\sqrt{E_H(X-Y)^2}\ :H\in C(F,G)\right\}$.

It suffices to show that $W_1(F,G)=\inf\{\int_{\mathbb{R}^2}|x-y|dH(x,y):H\in C(F,G)\}$ is attained at $H^*(x,y)=F(x)\wedge G(y)$. The same result holds for W_p, $p\geq 1$, where

$$W_p(F,G)=\left[\inf\left\{\int_{\mathbb{R}^2}\left|x-y\right|^p dH(x,y):H\in C(F,G)\right\}\right]^{\frac{1}{p}}$$

Here are the details, see Vallender (1973), that the infimum of $\int_{\mathbb{R}^2}|x-y|dH(x,y)$ over $H\in C(F, G)$ is indeed attained (at $H(x,y)=F(x)\wedge G(y)$).

Let $X,Y:(\Omega,\mathcal{A},P)\rightarrow(\mathbb{R},\mathcal{B}(\mathbb{R}))$ be random variables with distributions F, G, respectively.

Since

$$|X - Y| = (X - Y)1_{(X \geq Y)} + (Y - X)1_{(X<Y)}$$

we let $\alpha = \max(X - Y, 0)$ and $\beta = \max(Y - X), 0)$, so that

$$E|X - Y| = E\alpha + E\beta$$

Since $\alpha \geq 0$, we have

$$E(\alpha|Y = y) = \int_0^\infty P(\alpha > z|Y = y)dz$$

Now, $E(\alpha) = EE(\alpha|Y)$, so that

$$E\alpha = \int_{-\infty}^{\infty} dG(y) \int_0^\infty P(X - Y \geq z|Y = y) =$$

$$\int_{-\infty}^{\infty} dG(y) \int_0^\infty P(X \geq y + z|Y = y)dz =$$

$$\int_{-\infty}^{\infty} dG(y) \int_y^\infty P(X \geq x|Y = y)dx =$$

$$\iint_{(x,y);x>y} P(X \geq y, Y < y)dx = \int_{-\infty}^{\infty} P(X \geq y, Y < y)dy$$

Similarly,

$$E\beta = \int_{-\infty}^{\infty} P(Y \geq y, X < y)dy$$

Thus,

$$E|X - Y| = \int_{-\infty}^{\infty} P(X \geq y, Y < y)dy + \int_{-\infty}^{\infty} P(Y \geq y, X < y)dy =$$

$$\int_{-\infty}^{\infty} [P(X < y, Y \geq y) + P(Y < y, X \geq y)]dy$$

Now, look at the event $(X < y, Y \geq y)$.

Let $A = (X < y)$ and $B = (Y < y)$, then $(X < y, Y \geq y) = A \cap B^c$. But $A = (B^c \cap A) \cup (A \cap B)$, so that

$$P(X < y, Y \geq y) = P(A) - P(A \cap B) = P(X < y) - P(X < y, Y < y)$$

Thus,

$$E|X - Y| = \int_{-\infty}^{\infty} [P(X < y) + P(Y < y) - 2P(X < y, Y < y)]dy$$

Note that $P(X < y, Y < y)$ is the value of the joint distribution $H(y, y)$ of the vector (X, Y), and it is well known that $H(x, y) \leq F(x) \wedge G(y)$ which is a joint distribution with marginals F, G (from Frechet's work (1956) on correlation analysis with given marginals or from copula theory), so that

$$E|X - Y| \geq \int_{-\infty}^{\infty} [F(y) + G(y) - 2\min(F(y), G(y)]dy = \int_{-\infty}^{\infty} |F(y) - G(y)|dy$$

by noting that $|x - y| = x + y - 2(x \wedge y)$.

Therefore,

$$W_1(dF, dG) = \inf\{W_H(F, G) : H \in C(F, G)\} \geq \int_{-\infty}^{\infty} |F(y) - G(y)|dy$$

But

$$\int_{-\infty}^{\infty} |F(y) - G(y)|dy = \int_0^1 |F^{-1}(u) - G^{-1}(u)|du$$

(by an "analytic" proof below) so that the infimum of $\int_{\mathbb{R}^2} |x - y| dH(x, y)$ over $H \in C(F, G)$ is $\int_0^1 |F^{-1}(u) - G^{-1}(u)|du$ which turns out to be a minimum since

$$\int_0^1 |F^{-1}(u) - G^{-1}(u)|du = E|F^{-1}(U) - G^{-1}(U)| = E_{H^*}|X - Y|$$

where $H^*(x, y) = F(x) \wedge G(y)$ is the joint distribution function of

$(F^{-1}(U), G^{-1}(U))$. Q.E.D.

Remarks.

(1) Let $X \overset{D}{=} F^{[-1]}(U)$ and $Y \overset{D}{=} G^{[-1]}(U)$, we have $dH^* = du \circ (F^{-1}, G^{-1})^{-1}$, so that

$$H^*(x, y) = dH^*((-\infty, x] \times (-\infty, y]) = du\left\{u : F^{-1}(u) \leq x, G^{-1}(u) \leq y\right\} =$$
$$du\{u : u \leq F(x), u \leq G(y)\} = du\{u : u \leq F(x) \wedge G(y)\} = F(x) \wedge G(y)$$

(2) Proof of

$$\int_{-\infty}^{\infty} |F(y) - G(y)|dy = \int_0^1 |F^{-1}(u) - G^{-1}(u)|du$$

is as follows. The following is justified by Fubini's theorem, namely if $\int_{A\times B} |f(x, y)| d(x, y) < \infty$, then

$$\int_{A\times B} |f(x, y)| d(x, y) = \int_A \left[\int_B f(x, y)dy\right] dx = \int_B \left[\int_A f(x, y)dx\right] dy$$

Now, for $u \in (0, 1)$, we have

$$|F^{[-1]}(u) - G^{[-1]}(u)| = \left[F^{[-1]}(u) - G^{[-1]}(u)\right] 1_{\{u: F^{[-1]}(u) > G^{[-1]}(u)\}}(u) +$$
$$\left[G^{[-1]}(u) - F^{[-1]}(u)\right] 1_{\{u: F^{[-1]}(u) \leq G^{[-1]}(u)\}}(u)$$

So let

$$A = \left\{u \in (0, 1) : F^{[-1]}(u) > G^{[-1]}(u)\right\}$$

$$A^c = \left\{u \in (0,1) : F^{[-1]}(u) \leq G^{[-1]}(u)\right\}$$

We have

$$\int_0^1 |F^{[-1]}(u) - G^{[-1]}(u)|du =$$

$$\int_A |F^{[-1]}(u) - G^{[-1]}|(u)|du + \int_{A^c} |F^{[-1]}(u) - G^{[-1]}(u)|du$$

where we can write

$$\int_A |F^{[-1]}(u) - G^{[-1]}|(u)|du = \int_A \left[\int_{G^{[-1]}(u)}^{F^{[-1]}(u)} dx\right] du$$

Now, observe that, by definition of the quantile functions, we have $G^{[-1]}(u) \leq x \Leftrightarrow u \leq G(x)$ (and of course, $x < F^{[-1]}(u) \Leftrightarrow u > F(x)$), so that

$$\int_A \left[\int_{G^{[-1]}(u)}^{F^{[-1]}(u)} dx\right] du = \int_{\mathbb{R}} \left[\int_{F(x)}^{G(x)} 1_A(u) 1_{\{F(x) \leq G(x)\}}(x) du\right] dx$$

Similarly,

$$\int_{A^c} |F^{[-1]}(u) - G^{[-1]}(u)|du = \int_{\mathbb{R}} \left[\int_{G(x)}^{F(x)} 1_{A^c}(u) 1_{\{F(x) > G(x)\}}(x) du\right] dx$$

Hence,

$$\int_{\mathbb{R}} \left[\int_{F(x)}^{G(x)} 1_A(u) 1_{\{F(x) \leq G(x)\}}(x) du\right] dx + \int_{\mathbb{R}} \left[\int_{G(x)}^{F(x)} 1_{A^c}(u) 1_{\{F(x) > G(x)\}}(x) du\right] dx =$$

$$\int_{\mathbb{R}} |F(x) - G(x)|dx$$

Q.E.D.

Now, the distance $W_1(F, G)$ or $W_1(dF, dG) = \int_0^1 |F^{-1}(u) - G^{-1}(u)|du$ does take into account the locations of the histogram data "points". Indeed, for the histograms f, g, h in the previous example (with associated distributions F, G, H, respectively), we have $W_1(F, G) = 1$ and $W_1(F, H) = 5$, showing that the histogram f is closer to g than h.

On the other hand, W_1 is a natural extension from Euclidean data points to histogram data points. Indeed, for $x, y \in \mathbb{R}$, we identify them as

$$\partial_x(A) = dF_x(A) = 1_A(x), \qquad \partial_y(B) == dG_y(B) = 1_B(y)$$

so that

$$F_x(t) = \partial_x((-\infty, t]) = 1_{[x,\infty)}(t)$$

Since we consider real-valued random variable, i.e. with values in $\mathbb{R} = (-\infty, \infty)$, their quantile functions, e.g. $F_x^{-1}(.) : (0,1) \to \mathbb{R}$:

$$F_x^{-1}(u) = \inf\{t \in \mathbb{R} : F_x(t) \geq u\} = x1_{(0,1)}(u)$$

and hence

$$W_1(\partial_x, \partial_y) = \int_0^1 |F_x^{-1}(u) - F_y^{-1}(u)|du = \int_0^1 |x - y|1_{(0,1)}(u)du = |x - y|$$

Remarks.

(1) In 1969, Wasserstein Vassershtein (1969) considered W_2 to investigate the uniqueness of the stationary distribution of a Markov process. And in 1970, Dobrushin (1970) used Wasserstein metric to investigate stochastic processes by conditional distributions.

(2) In 1972, Mallows (1972) considered the same $W_2 -$ metric, without referring to its existence years ago!

(3) Shorack and Wellner (1986) used Wasserstein metrics to investigate the convergence of empirical processes in their book in 1986.

(4) For general Wasserstein metrics in Optimal Transport Theory, see Villani (2003)

3. Typical positions in Frechet's program

From a historical perspective, the pioneering work of Frechet (1948) can be viewed as the first attempt to generalize probability background for statistics, such as general random elements in arbitrary metric spaces, their typical positions (e.g. mean), general parameters, general statistics and their convergences (for asymptotics, e.g. consistency of estimators).

Nowadays, we are witnessing efforts of theoretical statisticians to specify Frechet's vision while applied statisticians in various fields, such as economics and machine learning (ML), started by implementing it in real-world applications, see, e.g. Bernton *et al.* (2019), Bhat and Prashanth (2019), Bigot (2020), Chartier (2013) and Kiesel *et al.* (2016).

We will elaborate on these current efforts in the context of Wasserstein metric spaces as data sets. For an invitation to the theoretical aspects of statistics in Wasserstein space, see Panaretos and Zemel (2020).

Remark. As Breiman (2001) spelled out the useful marriage between statistics and ML, see, e.g. Morizet (2020), Shalev-Shwartz and Ben-David (2014), Wasserstein metrics are used also in ML, e.g. in WGAN.

As a starting point, let's discuss the notion of "typical positions" of a random element X with values in an arbitrary metric space $(\mathcal{X}, \rho)$.

According to Frechet (1948), generalized typical positions such as median and mean could be defined via appropriate characterizations of classical notions on Euclidean spaces. For simplicity, consider $(\mathbb{R}, |.|)$.

Let X be a real-valued random variable with distribution function F (and law dF). In classical probability theory, the median $m(X)$ of X is a value on $\mathbb{R}$ which is "equiprobable" (always existed). The mean of X is the quantity $EX = \int_{\mathbb{R}} xdF(x)$ which exists when this integral is finite.

To generalize these typical positions to arbitrary metric spaces, we need to "characterize" them.

First, a characterization of $m(X)$ is obtained when statisticians use LAD (Least Absolute Deviation) $E|X - a|$ as error, say, in quantile regression.

Specifically, when $E|X| < \infty$, we have

$$m(X) = \arg\min_{a\in\mathbb{R}} E|X - a|$$

Thus, for X being a random element with values in a metric space $(\mathcal{X}, \rho)$, it median could be taken as $\operatorname{argmin}_{a\in\mathcal{X}} E\rho(X, a)$.

Next, on $(\mathbb{R}, |, |)$, when $E|X|^2 < \infty$, using mean squared error (MSE), we have

$$EX = \arg\min_{a\in} E(X - a)^2$$

However, as mentioned by Frechet (1948), a better characterization of EX., when EX exists (i.e. $EX < \infty$) with EX^2 finite or not, is this

$$EX = \arg\min_{a\in} E\left[(X - a)^2 - X^2\right]$$

noting that the minimizers of $a \to E(X - a)^2$ and of $a \to E[(X - a)^2 - X^2]$ are the same, since $(X - a)^2 - X^2$ differs from $(X - a)^2$ only by a constant X^2 (not depending on $a \in \mathbb{R}$).

But the advantage is that, since

$$|(X - a)^2 - X^2| = |a^2 - 2aX| \leq a^2 + 2|a||X|$$

we have $E|(X - a)^2 - X^2| < \infty$ when $E|X| < \infty$.

Consider now the situation in statistics where our random element of interest is $X(.) : (\Omega, \mathcal{A}, P) \to (\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{B}(\mathcal{X})$ is the Borel $\sigma -$ field generated by the topology of the metric ρ on $\mathcal{X}$. To be specific, let $\mathcal{X}$ be the space $\mathcal{P}(\mathbb{R})$ of probability measures dF on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ having second finite moment, i.e. $\int_{\mathbb{R}} x^2 dF(x) < \infty$ and let $\rho = W$ be the Wasserstein metric on $\mathcal{X} = \mathcal{P}(\mathbb{R})$, i.e. $W_2^2(dF, dG) = \int_0^1 |F^{-1}(u) - G^{-1}(u)|^2 du$.

This corresponds to the situation where our *data space* is $\mathcal{P}(\mathbb{R})$ (histogram data) equipped with a Wasserstein metric.

Remark. The metric space $(\mathcal{P}(\mathbb{R}), W)$ is not a linear space, let alone a Banach space !

How the notion of "mean of X", a typical position of X, is generalized? Well, let's follow Frechet!

A Frechet mean of X (or "barycenter"), for $EX^2 < \infty$, is an element of $\mathcal{X}$ that is a minimizer of the map $a \in \mathcal{X} \to E[W_2^2(X, a)]$.

Remark. A priori, the set of Frechet means of X is not a singleton.

The counter part of the "sample mean" of a random sample is the empirical Frechet mean, i.e. the Frechet mean set of the empirical probability measure $\frac{1}{n}\sum_{j=1}^{n} \partial_{X_j}$ of the IID sample X_1, $X_2, \ldots, X_n$ from X.

The convergence of the sample mean to the population mean is called the Law of Large Numbers (LLN) and is essential for statistics. Now, in the context of data space as a Wasserstein space of probability measures (on a metric space) equipped with a Wasserstein metric, the means of the "population" form a subset of the data space and hence the counterpart of LLN would involve estimation (from empirical Frechet mean set) of that subset, a problem of random set estimation. For a "flavor" of the strong law of large numbers (SLLN) in similar contexts, see Artstein and Wets (1995).

4. An improvement in financial risk management

Just to illustrate an application of statistics in Wasserstein space, we elaborate a bit here the robustness of risk measures in financial econometrics.

Here is why Wasserstein metrics appear as an improvement in risk measure modeling.

Consider the simplest case of modeling risk of a loss variable X. The loss function is intrinsically a random variable. As such, it risk is a function of its distribution function F, e.g. its value-at-risk is $R(X) = F^{-1}(\alpha)$, so that the risk $R(X) = \varphi(F)$, a functional on distribution functions of random variables. By robustness of a risk measure $\varphi(F)$, we mean the continuity of $\varphi(.)$, i.e. $F_n \rightarrow F \Rightarrow \varphi(F_n) \rightarrow \varphi(F)$, as $n \rightarrow \infty$. Now $F_n \overset{D}{\rightarrow} F$ is the convergence in distribution which is equivalent to the weak convergence of their associated probability measures $dF_n \overset{w}{\rightarrow} dF$, which, in turn, is metrized by, e.g. a Wasserstein $W_2 -$ metric.

Thus, traditionally, the desirable robustness property of (financial) risk measures is investigated under the (Euclidean metric-based) of some metrics on the set of probability measures (e.g. on $\mathbb{R}$).

Now Wasserstein metrics are also metrics on the set of probability measures (as models for building risk measures) with an apparent possible advantage (as compared to other such metrics), namely they do take into account of the geometry of the underlying sample spaces. As such, we are actually witnessing such a revolution in financial risk management with the use of Wasserstein metrics replacing previous metrics on probability measures.

Specifically, a new robust risk measure should be a functional $R(.) : \mathcal{P}(W) \rightarrow \mathbb{R}$ that is continuous with respect to a $W -$ metric, i.e.

$$\mu_n \overset{W}{\rightarrow} \mu \Rightarrow R(\mu_n) \rightarrow R(\mu)$$

What is new in risk models ? Answer: Replacing Euclidean metrics by Wasserstein metrics! Why that's a good thing to do? Let's find out!

If X is the (combined) loss function of, say, an investment portfolio of the form $\sum_{j=1}^{m} \lambda_j X_j$, we can consider various risk measures for X, such as Value-at-Risk, conditional Value-at-risk, etc . . .

Each individual loss function X_j as its probability measures μ_j (on $\mathbb{R}$, for example) with corresponding distribution function F_j. A risk model X could be proposed by using a copula approach, resulting in a probability measure μ with corresponding distribution function F for X.

The approach to new risk measures consists of viewing the probability measures μ_j as points of a Wasserstein space and using the Wasserstein barycenter b of $\mu_j' s$ with weights $\lambda_j' s$. A risk measure for X is a functional of distribution functions (such as the quantile function F^{-1}).

Now in the setting of Wasserstein risk measures, as mentioned before, the Wasserstein value-at-risk is taken to be based on the quantile function $F_{\mathbf{b}}^{-1}(.)$ of the Wasserstein barycenter b of $\mu_j' s$ with weights $\lambda_j' s$.

The new types of risk measures are Wasserstein barycenter risk measures.

Remark.

ML or statistical learning should be viewed as complementary to "standard" statistics, rather than "adversaries", especially in the actual situation of big data.

The paper "Statistical modeling: The two cultures" of Breiman (2001) while distinguished two different ways of doing statistics, could give the impression that standard statistics and ML are separate ways. In fact, if we look closely at them, not only they aim to solve (and apply) to same problems, but also they complement each other, both in goals and techniques (theories). As such, they are rather complementary. Thus, researchers who are familiar with statistics should also look at available ML algorithms. It is just about the interesting and useful phenomenon of the appearance of Wasserstein metrics in it.

It is so since essentially any statistical problems (treating in ML) involve the measure of a distance between probability measures.

Without going into details of unsupervised learning using Generative Adversary Networks/algorithms (GAN), it suffices to mention the following.

In GAN, Kullback–Leibler (KL)-Divergence (or more general divergence measures) is used to compare probability measures. But these divergence measures (inspired from statistics) are applicable only for probability density functions (i.e. for absolutely continuous probability measures), in one hand, and, on the other hand, exhibit some undesirable properties with regard to computations and interpretations.

It turns out that another metric can replace divergences to avoid these undesirable properties, and it is precisely a Wasserstein metric, leading to the WGAN!

This is a significant new application of Wasserstein metrics in unsupervised ML, since the training algorithm involves a distance between probability measures, for comparision. It illustrates again the need to consider, not only a metric between probability measures, but a good one!.

Now let turn to Wasserstein distance in risk analysis

From the setting of law invariance, we see that (financial) risk measures are based on (model) distributions of loss random elements (in the Euclidian case) and probability measures (in the general Polish space case). Such risk measures are estimated from historical data involving empirical probability measures. The problem of robustness of risk measures just involves comparisons of probability measures (e.g. deviation between true but unknown law of the loss variable and its estimate empirical version) and hence requires some appropriate probability metric, especially, robustness is about the continuity of risk measures as functions of their underlying probability measures.

How Wasserstein metrics appear naturally in risk analysis?

Let X be a (nonnegative) loss random variable, say, in actuarial science, with distribution F. According to the distortion principle, the risk premium calculation is based upon the functional

$$C_g(X) = \int_0^\infty (g \circ P)(X > x)dx = \int_0^\infty g(1 - F(x))dx$$

where $g(.)$ is a *concave* distortion function.

Now, as a routine, expressing $\int_0^\infty g(1 - F(x))dx$ as a double integral, and using Fubini's theorem to exchange the order of integration and the equivalence $P(X < x) \leq u \Leftrightarrow x \leq F^{[-1]}(u)$, we get the spectral representation of the distorted risk measure

$$C_g(F) = \int_0^1 F^{[-1]}(u)g'(1 - u)du$$

The robustness of $C_g(.)$ can be carried out as the continuity of this functional with respect to some metric of the appropriate set of probability measures dF.

Now, observe that for distributions F, G such that $\int_{\mathbb{R}} |x|^p dF(x) < \infty$, $\int_{\mathbb{R}} |x|^p dG(x) < \infty$ and $g' \in \mathcal{L}^q([0, 1], du)$, with $p > 1, \frac{1}{p} + \frac{1}{q} = 1$, we have, in view of Holder's inequality:

$$|C_g(F) - C_g(G)| = |\int_0^1 F^{[-1]}(u)g'(1 - u)du - \int_0^1 G^{[-1]}(u)g'(1 - u)du| =$$

$$|\int_0^1 \left[F^{[-1]}(u) - G^{[-1]}(u)\right]g'(1 - u)du| \leq$$

$$\left[\int_0^1 \left|g'(1 - u)\right|^q du\right]^{\frac{1}{q}}\Bigg] \left[\int_0^1 \left|F^{[-1]}(u) - G^{[-1]}(u)\right|^p du\right]^{\frac{1}{p}} = ||g'||_q W_p(dF, dG)$$

The above inequality suggests the use of Wasserstein metric in the study of robustness of risk measures.

Remark. Moreover, the same suggestion can be applied to model ambiguity where, according to decision theory, ambiguity refers to the uncertainty in considering a baseline model F for the loss variable in consideration (based upon historical data, of course). Model ambiguity is investigated by forming some ambiguity set of probability measures "around dF", i.e. by specifying a ball centered at dF, which requires a metric on probability measures. Various reasons in the literature point to the appropriateness of using Wasserstein metrics for robustness and model ambiguity. As such, it is about time to look closely at Wasserstein metrics (from optimal transport) as an update statistical tool for empirical works.

5. Conclusions

Wasserstein spaces as data sets are of current interests in applications of a variety of fields such as financial econometrics and ML. Since Wasserstein data are non-Euclidean, traditional data analysis and statistical methods cannot be directly applicable. Therefore, current research efforts aim at developing new methodologies for data analysis and associated statistical inference with such new data. In this paper, besides elaborating on Wasserstein metrics, we outline some basic ingredients for statistics with Wasserstein data, focusing on realizing the Frechet's program.

References

Artstein, Z. and Wets, R.J.B. (1995), "Consistency of minimizers and the SLLN for stochastic programs", *Journal Convex Analysis*, Vol. 2 No. 1, pp. 1-17.

Bernton, E., Jacob, P.E., Gerber, M. and Robert, C.P. (2019), "On parameter estimation with the Wasserstein distance", *Information and Inference: A Journal of the IMA*, Vol. 8 No. 4, pp. 657-676.

Bhat, S.P. and Prashanth, L.A. (2019), "Concentration of risk measures: a Wasserstein distance approach", *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Bigot, J. (2020), "Statistical data analysis in the Wasserstein space", *ESAIM: Proceedings and Surveys*, Vol. 68, pp. 1-19, doi: 10.1051/proc/202068001.

Billingsley, P. (1995), *Probability and Measure*, John Wiley & Sons, New York.

Breiman, L. (2001), "Statistical modeling: the two cultures", *Statistical Science*, Vol. 16 No. 3, pp. 199-215, doi: 10.1214/ss/1009213726.

Chartier, B. (2013), "Necessary and sufficient condition for the existence of a Frechet mean on the circle", *ESAIM: Probability and Statistics*, Vol. 17, pp. 635-649, doi: 10.1051/ps/2012015.

Dobrushin, R.L. (1970), "Prescribing a system of random variables by conditional distributions", *Theory of Probability and Its Applications*, Vol. XV No. 3, pp. 458-486, doi: 10.1137/1115049.

Frechet, M. (1906), "Sur quelque poins du calcul fonctionel", Thesis, 1906, Rend. Circ. Matem. Palermo (XXII), Paris.

Frechet, M. (1948), "Les elements aleatoires de nature quelconque dans un espace distancé", in *Annales de l'institut Henri Poincaré*, Vol. 10, pp. 215-310.

Frechet, M. (1956), "Sur les tableaux de correlation dont les marges sont donne'es", in *Comptes Rendus de l'Académie des Sciences*, Paris, (242), pp. 2426-2428.

Frechet, M. (1957), "Sur la distance de deux lois de probabilite", in *Comptes Rendus de l'Académie des Sciences*, Paris, (244), pp. 689-692.

Kiesel, R., Ruhlicke, R., Stahl, G. and Zheng, J. (2016), “The Wasserstein metric and robustness in risk management”, *Risk*, Vol. 4 No. 4, p. 32, doi: 10.3390/risks4030032.

Levy, P. (1937), *Theorie de 'l"Addition des Variables Aleatoires*, Gauthier-Villars, Paris.

Mallows, C. (1972), “A note on asymptotic joint normality”, *The Annals of Mathematical Statistics*, Vol. 43 No. 2, pp. 508-515, doi: 10.1214/aoms/1177692631.

Morizet, N. (2020), “Introduction to generative adversarial Networks”, Technical report, Advestis, hal-02899937.

Panaretos, V.M. and Zemel, Y. (2020), *An Invitation to Statistics in Wasserstein Space*, Springer.

Parthasarathy, K.R. (1967), *Probability Measures on Metric Spaces*, Academic Press, New York.

Shalev-Shwartz, S. and Ben-David, S. (2014), *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, New York.

Shorack, G.R. and Wellner, J.A. (1986), *Empirical Processes with Applications to Statistics*, John Wiley & Sons, New York.

Vallender, S.S. (1973), “Calculation of the Wasserstein distance between distributions on the line”, *Theory of Probability and Its Applications* Vol. 18, pp. 784-786.

Vassershtein, L.N. (1969), “Markov process over denumbrable product of spaces describing large systems of automata”, *Problems of Information Transmission*, Vol. 5 No. 3, pp. 47-52.

Villani, C. (2003), *Topics in Optimal Transportation*, American Mathematical Society.

Corresponding author
Chon Van Le can be contacted at: lvchon@hcmiu.edu.vn