A prelude to statistics arising from optimal transport theory

Hung T. Nguyen
Department of Mathematical Sciences, New Mexico State University, Las Cruces, New Mexico, USA

Received 4 May 2023
Revised 26 May 2023
Accepted 26 May 2023

Abstract

Purpose – This paper aims to offer a tutorial/introduction to new statistics arising from the theory of optimal transport to empirical researchers in econometrics and machine learning.
Design/methodology/approach – Presenting in a tutorial/survey lecture style to help practitioners with the theoretical material.
Findings – The tutorial survey of some main statistical tools (arising from optimal transport theory) should help practitioners to understand the theoretical background in order to conduct empirical research meaningfully.
Originality/value – This study is an original presentation useful for new comers to the field.

Keywords Multivariate quantiles, Optimal transport, Partial identification, Random sets, Wasserstein metrics
Paper type General review

1. Introduction

A significant contribution to statistics in general and to econometrics and machine learning in particular from optimal transport theory has surfaced recently. As such it is about time for practitioners to be aware of it to apply it to real-world problems, especially in econometrics, to improve credibility of empirical findings. That is precisely the purpose of this prelude.

This paper is organized as follows. Although this is a prelude where detailed technical material is not spelled out, the main purpose is to call practitioners' attention to new improved statistical tools arising from optimal transport theory, and as such, optimal transport in a nutshell will be presented in Section 2. Section 3 is about the most significant new tool in statistical analysis, namely the notion of multivariate quantiles. Section 4 is devoted to the elaboration of another new tool for statistics, namely the Wasserstein metrics. In Section 5 we elaborate on the interesting connection between partial identification and random set statistics, also thanks to optimal transport.

2. Optimal transport in a nutshell

Monge (1781) was concerned with the problem of finding the cheapest way to transport, say, soil from a collection of mines to a collection of construction sites.

In mathematical language, the problem is formulated as follows. Let $\mathcal{P}(\mathcal{X}), \mathcal{P}(\mathcal{Y})$ denote the spaces of probability measures on $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ respectively. Given $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$, and a (cost) function $c(.,.): \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. A transport map is a (measurable) function $T(.)$:

JEL Classification — A23, B23, C40, C65, C83
MSC2020 Classification — 60-01, 62A01, 62H10

Asian Journal of Economics and Banking
Vol. 7 No. 2, 2023
pp. 166-179
Emerald Publishing Limited
e-ISSN: 2633-7991
p-ISSN: 2615-9821
DOI 10.1108/AJEB-05-2023-0038

$\mathcal{X} \to \mathcal{Y}$ such that $\nu(.) = \mu \circ T^{-1}$, in symbol $\nu = T\#\mu$ (T pushes μ forward to ν). The transport cost of T is

$$\int_{\mathcal{X}} c(x, T(x))d\mu(x)$$

The Monge's problem is to find an optimal transport map T^*, i.e.

$$T^* = \arg\min\left\{\int_{\mathcal{X}} c(x, T(x))d\mu(x) : T \ni T\#\mu = \nu\right\}$$

This functional optimization problem might not have a solution in general, e.g. when μ is a Dirac probability measure whereas ν is not. But more importantly, with the optimization variable being T, the objective function $T \to \int_{\mathcal{X}} c(x, T(x))d\mu(x)$ is not linear. Also, the constraint set $\{T: T\#\mu = \nu\}$ is not convex. As such, the computation of a solution is difficult.

Because of these issues, Monge's problem was unsolved until Kantorovich (1942) who reformulated Monge's problem to a setting avoiding the two main difficulties mentioned above.

It is interesting to note that Monge's difficulties have analogies in mathematics. When a quadratic equation does not have real solutions, we enlarge its solution domain (the real line $\mathbb{R}$) to the complex plane so that the equation has complex solutions. Similarly, we consider mixed (random) strategies in non-cooperative games to establish the existence of Nash equilibria for any such games.

The same methodology could be used here to "solve" Monge's problem. And that is exactly what Kantorovich has done.

If $T(.) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ is a transport map, and $I_{\mathcal{X}}(.) : \mathcal{X} \to \mathcal{X}$ is the identity map $I_{\mathcal{X}}(x) = x$, then $(I_{\mathcal{X}} \times T)(.) : \mathcal{X} \to \mathcal{X} \times \mathcal{Y}$, $(I_{\mathcal{X}} \times T)(x) = (x, T(x))$ pushes μ forward to a joint probability measure $\mu \circ (I_{\mathcal{X}} \times T)^{-1}$ on $\mathcal{X} \times \mathcal{Y}$ having μ, ν as marginal probability measures. Thus the space of joint probability measures on $\mathcal{X} \times \mathcal{Y}$ having μ, ν as marginal probability measures, denoted as $\Pi(\mu, \nu)$, contains the set of all (Monge) transport maps (by identification). Elements of $\Pi(\mu, \nu)$ are referred to as transport plans. Hence, by enlarging transport maps to transport plans, Kantorovich reformulated Monge's problem as follows. Given μ on $\mathcal{X}$, ν on $\mathcal{Y}$, and $c(.,.) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, find a transport plan $\lambda^* \in \Pi(\mu, \nu)$ such that

$$\lambda^* = \arg\min\left\{\int_{\mathcal{X} \times \mathcal{Y}} c(x,y)d\lambda(x,y) : \lambda \in \Pi(\mu, \nu)\right\}$$

The difficulties in Monge's problem are avoided: Kantorovich's problem always has solutions since $\mu \otimes \nu \in \Pi(\mu, \nu)$; with the optimization variable $\lambda \in \Pi(\mu, \nu)\}$, the objective function $\lambda \to \int_{\mathcal{X} \times \mathcal{Y}} c(x,y)d\lambda(x,y)$ is linear, and the constraint set $\{\lambda : \lambda \in \Pi(\mu, \nu)\}$ is convex, so that we are in the domain of convex optimization!

3. Multivariate quantiles

The focus on (univariate) quantile functions as a basis for statistical analysis has been advocated by Parzen (1979). In fact, in the comment to Breiman's paper (2001), Parzen even suggested that there are many possible "cultures" for statistical modeling where "quantile culture" could be one of them.

Without digging into whether Parzen's quantile culture is a culture in Breiman's sense, we could view that the use of quantile functions is part of the standard statistical analysis in which, instead of distribution functions, we focus on quantile functions. The two cultures elaborated in Breiman's paper (2001), namely the data modeling and algorithmic modeling, might be not really disjoint, i.e. they could be combined to form a new culture. That was precisely suggested in "Statistical Modeling: The Three cultures" in 2023 by Daoud and

Dubhashi (2023) as a hybrid modeling culture! Perhaps, it could be so as we witness at present the interests of Econometricians in Machine (or Statistical) Learning?

Anyway, the point we want to make is this. It is true that the use of quantile functions, such as in quantile regression, provides more information than that of mean regression. But why Parzen's "quantile culture" did not get off the ground or ring the bell, say, in multivariate analysis?

The answer could be twofold. As mean (linear) regression, in multivariate analysis, is the bread-and-butter tool in statistics, quantile regression, introduced by Koenker and Basett (1978), is also only for one dimension. The second reason is crucial: there is no counterpart of multivariate mean regression, and this is because of the lack of a "correct" notion of multivariate (vector) quantile functions, let alone its associated regression analysis.

Specifically, the mathematical problem of how to generalize the familiar notion of an univariate quantile function to higher dimensions is difficult because the explicit definition of a quantile function on the real line $\mathbb{R}$ is based on the total order relation $\leq$ of $\mathbb{R}$, whereas there is no such order relation on $\mathbb{R}^n$ with $n > 1$.

In the literature, among various attempts to "solve" the problem, e.g. Hallin *et al.* (2010) and Serfling and Zuo (2010), an attempt was to consider the partial order relation of $\mathbb{R}^n$ when $n > 1$, exemplified by Belloni and Winkler (2011), leading to the notion of "partial multivariate quantiles". This is typical of an approach to avoid the lack of the total order relation on $\mathbb{R}$, but does not really address the original problem, i.e. partial vector quantile functions are not generalizations of univariate quantile functions. They are just substitutes.

Finally, as Koenker (2017) acknowledged, the happy ending arrived in 2016 with the works of Carlier *et al.* (2016, 2017), and that was inspired from the Theory of Optimal Transport, Villani (2003).

This Section aims at elaborating a bit on the notion of vector (multivariate) quantile functions correctly generalizing the familiar notion of univariate quantile functions.

Let X be a real-valued random variable (the name of some quantity), i.e. a measurable map from a probability space $(\Omega, \mathcal{A}, P)$, its source of uncertainty, to the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, its sampling space. Its law is the probability measure P_X on $\mathcal{B}(\mathbb{R})$ obtained by pushing forward P by X, i.e. $P_X(.) = P \circ X^{-1}$, in symbol $P_X = X\#P$. By Lebesgue-Stieltjes theorem, $P_X = dF$ where $F(.) : \mathbb{R} \to [0, 1]$ is the distribution function of X. The distribution function F of X contains all information about the random evolution of X. If we know F, can we create the data from X? This is the problem known as simulations. Yes, but not directly by using F. Instead, we consider its "pseudo inverse" function $F^{[-1]}$ known as its (univariate) quantile function defined explicitly as $F^{[-1]}(.) : (0, 1) \to \mathbb{R}$,

$$F^{[-1]}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}$$

and show that $F^{[-1]}$ will push forward the uniform probability measure du on (0,1) to dF ($F^{[-1]} \# du = dF$, i.e. $dF(.) = du \circ (F^{[-1]})^{-1}$) so that $X \overset{D}{=} F^{[-1]}(U)$ (equal in distribution) where U denotes the (uniform) random variable on (0,1) with law du.

While the pseudo inverse $F^{[-1]}(.)$ provides a reasonable mathematical definition for quantiles, its explicit definition involves the total order relation $\leq$ of the real line $\mathbb{R}$ and that is the difficulty to extend it to higher dimensions, say, as a function $G(.) : (0, 1)^n \to \mathbb{R}^n$ with $n > 1$.

A short story of this extension problem seems interesting to note. Traditionally, the extension of a concept in one dimension to several dimensions could be done componentwise, such as the concept of the mean of a random vector. But defining a vector quantile function componentwise does not work since the property $G\#du = dF$, for du as uniform law on $(0,1)^n$, and dF as law of a random vector on $\mathbb{R}^n$ (for $n > 1$) is not satisfied.

Remark. Of course $F(.)$ is characterized by $F^{[-1]}(.)$, since if $Q(.) : (0,1) \to \mathbb{R}$ is monotone non decreasing and left continuous then there exists a unique distribution function $F(.)$ such that $Q(.) = F^{[-1]}(.)$. However, such a characterization of $F^{[-1]}(.)$ does not extend to higher dimensions.

Also, traditionally, if we cannot use directly an established concept in one setting to extend it to another setting, we look for a possible equivalent concept (a characterization of the established concept) that can be generalized. For example, to generalize ordinary sets to fuzzy sets, we use the indicator function of an ordinary set (as its membership function) as a characterization of the set from which to extend to the new setting. Here the question is: what is a characterization of $F^{[-1]}$, i.e. another equivalent way to define it.

Perhaps, previous attempts to generalize univariate quantile functions to vector quantile functions did not ask this question. It turns out that the answer is hidden in plain sight! Besides the property $F^{[-1]}\#du = dF$, the (explicitly defined) function $F^{[-]}(.) : (0,1) \to \mathbb{R}$ is monotone non decreasing, and these two properties provide a characterization for $F^{[-1]}$. Specifically, $F^{[-1]}(.)$: $(0,1) \to \mathbb{R}$ is the unique function that is monotone non decreasing and satisfies $G\#du = dF$:

Lemma. If $G(.) : (0,1) \to \mathbb{R}$ is monotone non decreasing and satisfies $G\#du = dF$, then $G(.) = F^{[-1]}$, i.e.

Proof. By monotonicity of G, we have

$$(-\infty, x] \subseteq G^{-1}((-\infty, G(x)])$$

so that

$$F_{du}(x) = du(-\infty, x] \leq du\{G^{-1}((-\infty, G(x)]) = dF(-\infty, G(x)] = F(G(x))$$

and $G(x) \geq F^{[-1]}(x)$

Consider the points x such that $G(x) > F^{[-1]}(x)$. This means that there exists $\varepsilon_o > 0$ such that $F(G(x) - \varepsilon) \geq F_{du}(x)$ for every $\varepsilon \in [0, \varepsilon_o]$. Also, since $G^{-1}((-\infty, G(x) - \varepsilon) \subseteq (-\infty, x)$, we have $F(G(x) - \varepsilon) < F_{du}(x)$. Thus, $F(G(x) - \varepsilon) = F_{du}(x)$ for any $\varepsilon \in [0, \varepsilon_o]$. Note that $F(G(x) - \varepsilon)$ is the value of F which F takes on an interval where it is constant. But these intervals are a countable quantity, so that the values y_j of F on these intervals are also countable. Therefore, the points x where $G(x) > F^{[-1]}(x)$ are contained in $\cup_j \{x : F_{du}(x) = y_j\}$ which is $du -$ negligible (since du is atomless). As a consequence, $G(x) = F^{[-1]}(x)$, $du -$ almost everywhere. Q.E.D.

As a consequence, the above characterization of $F^{[-1]}$ can be used to obtain its counterpart in higher dimensions since on $\mathbb{R}^n$, with $n > 1$, the property $G\#du = dF$ makes sense and the monotone non decreasing property for $G(.) : (0,1)^n \to \mathbb{R}^n$ is equivalent to

$$< u - v, G(u) - G(v) > \geq 0$$

where $< ., . >$ denotes the scalar product on $\mathbb{R}^n$.

Remark. The characterization of $F^{[-1]}$ brings out the fact that the total order relation on $\mathbb{R}$ does not play an essential role in defining it.

The point is this. If $G(.) : (0,1)^n \to \mathbb{R}^n$ (for $n > 1$) is going to be an extension of $F^{[-]}(.) : (0,1) \to \mathbb{R}$, $G(.)$ has to be monotone non decreasing and pushing forward du to dF (in dimension $n > 1$).

The upshot is that, for $n \geq 1$, these two properties are characteristic for the notion of quantiles, in the sense that there is uniquely one such function $G(.) : (0,1)^n \to \mathbb{R}^n$, so that,

for $n = 1$, it coincides with $F^{[-1]}(.)$. Thus, in dimension 1, the familiar univariate quantile function $F^{[-1]}$ can be defined without using explicitly the total order relation of $\mathbb{R}$!

This upshot was discovered in the context of Optimal Transport, see Villani (2003), Brenier (1991), McCam (1995), Carlier *et al.* (2017) and Galichon (2016), where a (n-dimensional) vector quantile function is the unique monotone noncreasing function $G(.) : (0,1)^n \to \mathbb{R}^n$ such that $G\#du = dF$.

Clearly, the upshot tells us that the familiar univariate quantile function can be generalized to higher dimensions rigorously. However, except in dimension 1, the vector quantile functions so determined are not obtained in a close form. Practitioners should consult the literature for computational works.

Remark. The following notes could give a flavor of optimal transport in getting, finally, the correct notion of multivariate quantiles.

In the setting of optimal transport, $F^{[-1]}(.)$ is characterized by a unique "transport map" $T^*(.) : (0,1) \to \mathbb{R}$, monotone non decreasing and $T^*\#du = dF$, where

$$T^* = \arg\min\left\{\int_0^1 \frac{1}{2}\left|u - T(u)\right|^2 du : Tdu = dF\right\}$$

i.e. the solution of Monge's problem with cost function $c(.,.) : (0,1) \times \mathbb{R} \to \mathbb{R}^+ : (u,x) \to \frac{1}{2}|u - T(u)|^2$. On the other hand, the function $\varphi(.) : (0,1) \to \mathbb{R}$

$$\varphi(u) = \int_0^u F^{[-1]}(v)dv$$

is convex, so that $F^{[-1]}(.)$ is the derivative of the convex function $\varphi(.)$ on (0,1).

In dimension $n > 1$, the above leads to the notion of multivariate quantile function by McCam's theorem (1995): Let $F(.) : \mathbb{R}^n \to [0,1]$ be a multivariate distribution function, then there exists a unique gradient $\nabla\varphi(.) : (0,1)^n \to \mathbb{R}^n$ of some convex function $\varphi(.) : (0,1)^n \to \mathbb{R}$ (φ is not unique, but $\nabla\varphi$ is unique) such that $\nabla\varphi\#du = dF$, where du is the uniform law on $(0,1)^n$.

4. Wasserstein metrics

We elaborate now upon a new improved type of metrics on spaces of probability measures arising from optimal transport theory. The main improvement seems to be that these new metrics, called Wasserstein metrics, do take into account of the geometry of the underlying sample space. Their construction surfaces naturally in the setting of optimal transport theory. Such metrics are useful, e.g. for machine learning.

Recall that, in applications of statistics, we often use a divergence $D(.,.)$ on a space of probability measures to "measure" of the difference between two probability measures. Such a divergence is used to compare probability measures, for example $D(\mu, \nu)$ is the difference between a model μ and a data ν.

The most well-known divergence is the Kullback-Leibler divergence on probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$:

$$KL(\mu//\nu) = \int_{\mathbb{R}} f(x)\log\left(\frac{f(x)}{g(x)}\right)d\gamma(x)$$

where $f(.)$, $g(.)$ are probability density of μ, ν respectively (with respect to some dominating measure γ on $\mathcal{B}(\mathbb{R})$). The KL divergence is not a distance since it is not symmetric, but it does have analogous properties which could be used to substitute for a metric, such as the Total Variation metric

$$TV(\mu,\nu) = \sup\{|\mu(A) - \nu(A)| : A \in \mathcal{B}(\mathbb{R})\}$$

The *KL* divergence appears in the model selection criterion *AIC*.

Metrics on spaces of probability measures are viewed as special divergences. Divergences abound. The choice of a divergence or a metric for comparing probability measures depends on its usefulness for the problem at hand. For example, the Kullback-Leibler divergence is used in *AIC* because of its relation to Maximum Likelihood Estimation.

Consider the case where $\mathcal{X} = \mathcal{Y} \subseteq \mathbb{R}^n$, we are interested in the following Wasserstein divergence (*a priori*) on the subset $\mathcal{P}_p(\mathcal{X})$ of the set $\mathcal{P}(\mathcal{X})$ of all (Borel) probability measures on $\mathcal{X}$, where

$$\mathcal{P}_p(\mathcal{X}) = \left\{\mu \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} \|x\|^p d\mu(x) < \infty\right\}$$

namely, $W_p(.,.): \mathcal{P}_p(\mathcal{X}) \times \mathcal{P}_p(\mathcal{X}) \to [0,\infty)$

$$W_p(\mu,\nu) = \inf_{\lambda \in \Pi(\mu,\nu)} \left[\int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^p d\lambda(x,y)\right]^{\frac{1}{p}}$$

Specifically, we are going to show that the Wasserstein divergence $W_p(.,.)$ is in fact a bona fide metric on $\mathcal{P}_p(\mathcal{X})$, a well-known fact in the literature.

We will carry out the complete proof that Wasserstein divergence is in fact a bona fide metric to emphasize the interesting notion of disintegration (of measures).

Disintegration is a process of extracting a conditional probability measure from a joint probability measure on a product space.

To be concrete, let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$, and $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, $(\mathcal{X}\times\mathcal{Y}, , \mathcal{B}(\mathcal{X}\times\mathcal{Y}))$, be (Borel) measurable spaces. We denote by $\mathcal{P}(\mathcal{X})$, $\mathcal{P}(\mathcal{Y})$, $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ the set of all probability measures on these spaces.

For $\lambda \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, its marginal probability measure on $\mathcal{X}$ is $\mu \in \mathcal{P}(\mathcal{X})$, defined as, for any $A \in \mathcal{B}(\mathcal{X})$, $\mu(A) = \lambda(A\times\mathcal{Y})$.

A disintegration of λ with respect to μ is a family of probability measures $\nu_x \in \mathcal{P}(\mathcal{Y})$, for any $x \in \mathcal{X}$, such that, for $A \in \mathcal{B}(\mathcal{X})$ and $B \in \mathcal{B}(\mathcal{Y})$, we have

$$\lambda(A\times B) = \int_A \nu_x(B) d\mu(x)$$

Symbolically,

$$\lambda = \int_{\mathcal{X}} (\delta_x \otimes \nu_x) d\mu(x)$$

where δ_x is the Dirac probability measure on $\mathcal{X}$, at $x \in \mathcal{X}$, and $\delta_x \otimes \nu_x$ denotes the product measure $(\delta_x \otimes \nu_x)(A\times B) = \delta_x(A)\nu_x(B)$.

The representation of λ is so written since

$$\lambda(A\times B) = \int_{\mathcal{X}} (\delta_x \otimes \nu_x)(A\times B) d\mu(x) =$$

$$\int_{\mathcal{X}} (\delta_x(A)\nu_x)(B) d\mu(x) = \int_A \nu_x(B) d\mu(x)$$

Below is a tutorial on disintegration, just enough for using it in proving the triangle inequality for Wasserstein metrics. A reference could be Dudley (2003) or Graf and Mauldin (1989).

Now, for $\mathcal{X} = \mathbb{R}^n$, with norm $\|.\|$, and $p \geq 1$, the pth- Wasserstein metric is

$$W_p^p(dF, dG) = \inf\{E_\pi \|X - Y\|^p : X \sim dF, Y \sim dG, \pi \in \Pi(dF, dG)\}$$

where F, G are $n-$ dimensional distribution functions of X, Y, respectively, and π has $2n-$ dimensional distribution function H with F, G as marginals, i.e.

$$H(x_1, \ldots x_n, \infty, \ldots, \infty) = F(x_1, \ldots x_n), \quad H(\infty, \ldots, \infty, y_1, \ldots y_n,) = G(y_1, \ldots y_n)$$

More generally, Wasserstein distance is a metric on spaces of probability measures. Let $(\mathcal{X}, \rho)$ be a metric space. Consider the situation where we are interested in probability measures governing the random evolution of random elements taking values in $\mathcal{X}$ (i.e. their "laws" operating on Borel $\sigma-$ field $\mathcal{B}(\mathcal{X})$). Comparisons of probability measures are standard concerns in applications, such as in the so-called empirical processes.

For μ, ν two probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, consider the nonnegative quantity

$$W(\mu, \nu) = \inf\left\{\int_{\mathcal{X}\times\mathcal{X}} \rho(x,y) d\pi(x,y)\right\} \leq +\infty$$

where the infimum is taken over all joint probability measure π with marginals (projections) μ, ν.

We will denote by $\Pi(\mu, \nu)$ the set of probability measures π on the product space $\mathcal{X} \times \mathcal{X}$ having μ, ν as marginal measures, i.e. $\mu(.) = \pi(. \times \mathcal{X}), \nu(.) = \pi(\mathcal{X} \times .)$.

$$ Note that the above quantity can be written as:

$$W(\mu, \nu) = \inf\{E\rho(X, Y) : X \sim \mu, Y \sim \nu\}$$

i.e. the infimum is taken over all random variables X, Y with values in $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, and X, Y are distributed as μ, ν, respectively.

On a subset of $\mathcal{P}(\mathcal{X})$ where $W(\mu, \nu) < \infty$, for μ, ν in it, $W(.,.)$ is a bona fide metric.

We come now to the main investigation of Wasserstein divergences (*a priori*) on a metric space $\mathcal{X}$ for which disintegration exists, such as $\mathbb{R}^n$ or a polish space.

Let $\mathcal{P}(\mathcal{X})$ denotes the set of all (Borel) probability measures on $\mathcal{B}(\mathcal{X})$. For $p \geq 1$, let $\mathcal{P}_p(\mathcal{X}) \subseteq \mathcal{P}(\mathcal{X})$, be the subset of probability measures with finite $p-$ moment, i.e.

$$\mathcal{P}_p(\mathcal{X}) = \left\{\mu \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} \|x\|^p d\mu(x) < \infty\right\}$$

where $\|.\|$ is the Euclidean norm of $\mathbb{R}^n$.

Consider the Wasserstein divergence on $\mathcal{P}_p(\mathcal{X})$: For $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, and $p \geq 1$, let

$$W_p(\mu, \nu) = \inf_{\lambda \in \Pi(\mu,\nu)} \left[\int_{\mathcal{X}\times\mathcal{X}} \|x - y\|^p d\lambda(x,y)\right]^{\frac{1}{p}}$$

This is just an exercise to verify that $W_p(.,.)$ does satisfy the axioms of a metric, i.e. $W_p(.,.) : \mathcal{P}_p(\mathcal{X}) \times \mathcal{P}_p(\mathcal{X}) \to \mathbb{R}^+ = [0, \infty)$ is such that

(1) $W_p(\mu, \nu) = W_p(\nu, \mu)$

(2) $W_p(\mu, \nu) = 0 \Leftrightarrow \mu = \nu$

(3) For any $\mu, \nu, \gamma \in \mathcal{P}_p(\mathcal{X})$, $W_p(\mu, \nu) \leq W_p(\mu, \gamma) + W_p(\gamma, \nu)$

First, since, for $x, y \in \mathbb{R}^n$, $\|x - y\|^p \leq c(\|x\|^p + \|y\|^p)$, so that , for $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^n)$, we have

$$W_p(\mu,\nu) \le c\left[\int_{\mathcal{X}} \|x\|^p d\mu(x) + \int_{\mathcal{X}} \|x\|^p d\nu(x)\right] < \infty$$

While (i) is obvious (since the function $(x,y) \to \|x-y\|^p$ is symmetric, and $\Pi(\mu,\nu) \simeq \Pi(\nu,\mu)$) and (ii) can be seen as follows.

For $\nu = \mu$, the optimal transport map $T : \mathcal{X} \to \mathcal{X}$ is the identity map $I(x) = x$, so that the optimal transport plan is $\lambda = (I, I)\#\mu$ concentrated on $\{(x,y) : x = y\}$ and hence

$$W_p(\mu,\mu) = \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^p d\lambda(x,y) = 0$$

Conversely, if $W_p(\mu,\nu) = 0$, then, since

$$\inf_{\lambda\in\Pi(\mu,\nu)} \left[\int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^p d\lambda(x,y)\right]^{\frac{1}{p}}$$

is attained, there exists $\lambda \in \Pi(\mu,\nu)$ such that $\int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^p d\lambda(x,y) = 0$ so that λ is concentrated on $\{(x,y) : x = y\}$ which, in turn, implies that, for any $A \in \mathcal{B}(\mathbb{R}^n)$,

$$\mu(A) = \lambda(A\times\mathbb{R}^n) = \lambda(A\times A) = \lambda(\mathbb{R}^n\times A) = \nu(A)$$

i.e. $\mu = \nu$.

Remark. For $p \le q$, we have $W_p(\mu,\nu) \le W_q(\mu,\nu)$, since, by Jensen's inequality (with respect to the convex function $t \to t^{\frac{q}{p}}$), for any $\lambda \in \Pi(\mu,\nu)$,

$$W_p^q(\mu,\nu) \le \left[\int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^p d\lambda(x,y)\right]^{\frac{q}{p}} \le$$

$$\left[\int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^q d\lambda(x,y)\right] = W_q^q(\mu,\nu)$$

However, the triangle inequality (iii) is not so obvious!

Interestingly, it is the notion of disintegration which will provide a method to verify it.

We wish to show that, for any $\mu_j, j = 1,2,3$ in $\mathcal{P}_p(\mathbb{R}^n)$, for $p \ge 1$, with support $\mathcal{X}_j \subseteq \mathbb{R}^n, j = 1, 2, 3$, respectively, we should have

$$W_p(\mu_1,\mu_2) \le W_p(\mu_1,\mu_3) + W_p(\mu_3,\mu_2)$$

For this, we follow Villani (2003).

Lemma. Let μ_j, $j = 1, 2, 3$ in $\mathcal{P}_p(\mathbb{R}^n)$, for $p \ge 1$, with support $\mathcal{X}_j \subseteq \mathbb{R}^n$, $j = 1, 2, 3$, respectively. Let $\lambda_{12} \in \Pi(\mu_1, \mu_2)$, and $\lambda_{23} \in \Pi(\mu_2, \mu_3)$.

Then there exists a probability measure λ on $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ having marginals λ_{12} and λ_{23} on $\mathcal{X}_1 \times \mathcal{X}_2$, and $\mathcal{X}_2 \times \mathcal{X}_3$, respectively.

Proof. Disintegrate both λ_{12} and λ_{23} with respect to their common marginal μ_2, and denote their disintegrations as ν_{x_1}, γ_{x_3}, respectively, so that

$$\lambda_{12} = \int_{\mathcal{X}_2} (\nu_{x_1} \otimes \delta_{x_2}) d\mu_2(x_2)$$

$$\lambda_{23} = \int_{\mathcal{X}_2} (\delta_{x_2} \otimes \gamma_{x_3}) d\mu_2(x_2)$$

Then $\lambda \in P(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3)$ constructed as

$$\lambda = \int_{\mathcal{X}_2} (\nu_{x_1} \otimes \delta_{x_2} \otimes \gamma_{x_3}) d\mu_2(x_2)$$

Then, for $A_1 \in \mathcal{B}(\mathcal{X}_1), A_2 \in \mathcal{B}(\mathcal{X}_2), A_3 \in \mathcal{B}(\mathcal{X}_3)$, we have

$$\lambda(A_1 \times A_2 \times \mathcal{X}_3) = \int_{\mathcal{X}_2} (\nu_{x_1} \otimes \delta_{x_2} \otimes \gamma_{x_3})(A_1 \times A_2 \times \mathcal{X}_3) d\mu_2(x_2) =$$

$$\int_{\mathcal{X}_2} (\nu_{x_1}(A_1)\delta_{x_2}(A_2)\gamma_{x_3}(\mathcal{X}_3) d\mu_2(x_2) = \int_{\mathcal{X}_2} (\nu_{x_1}(A_1)\delta_{x_2}(A_2) d\mu_2(x_2) =$$

$$\int_{\mathcal{X}_2} (\nu_{x_1} \otimes \delta_{x_2})(A_1 \times A_2) d\mu_2(x_2) = \lambda_{12}(A_1 \times A_2)$$

Similarly for

$$\lambda(\mathcal{X}_1 \times A_2 \times A_3) = \lambda_{23}(A_2 \times A_3)$$

QED.

Then the proof of the triangle inequality for Wasserstein metrics follows:
Let $\mu_j, j = 1, 2, 3$ in $\mathcal{P}_p(\mathbb{R}^n)$, for $p \geq 1$, with support $\mathcal{X}_j \subseteq \mathbb{R}^n, j = 1, 2, 3$, respectively.
Note that, from OT theory (existence of solutions of Kantorovich's problem),

$$W_p(\mu_i, \mu_j) = \inf_{\lambda \in \Pi(\mu_i, \mu_j)} \left[\int_{\mathcal{X}_i \times \mathcal{X}_j} \|x_i - x_j\|^p d\lambda(x_i, x_j) \right]^{\frac{1}{p}}$$

is attained with some *optimal* transport plan $\lambda_{ij} \in \Pi(\mu_i, \mu_j)$. Thus,

$$W_p(\mu_i, \mu_j) = \left[\int_{\mathcal{X}_i \times \mathcal{X}_j} \|x_i - x_j\|^p d\lambda_{ij}(x_i, x_j) \right]^{\frac{1}{p}}$$

Now, let λ, in the above Lemma corresponding to $\mu_j, j = 1, 2, 3$, be the probability measure on $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ having marginals λ_{12} and λ_{23} on $\mathcal{X}_1 \times \mathcal{X}_2$, and $\mathcal{X}_2 \times \mathcal{X}_3$, respectively.
We then have

$$W_p(\mu_1, \mu_3) = \left[\int_{\mathcal{X}_1 \times \mathcal{X}_3} \|x_1 - x_3\|^p d\lambda_{13}(x_1, x_3) \right]^{\frac{1}{p}} =$$

$$\left[\int_{\mathcal{X}_1 \times \mathcal{X}_3 \times \mathcal{X}_3} \|x_1 - x_3\|^p d\lambda(x_1, x_2, x_3) \right]^{\frac{1}{p}} \leq$$

$$\left[\int_{\mathcal{X}_1 \times \mathcal{X}_3 \times \mathcal{X}_3} (\|x_1 - x_2\| + \|x_2 - x_3\|)^p d\lambda(x_1, x_2, x_3) \right]^{\frac{1}{p}} \leq$$

$$\left[\int_{\mathcal{X}_1\times\mathcal{X}_3\times\mathcal{X}_3}\|x_1-x_2\|^p d\lambda(x_1,x_2,x_3)\right]^{\frac{1}{p}}+\left[\int_{\mathcal{X}_1\times\mathcal{X}_3\times\mathcal{X}_3}\|x_2-x_3\|^p d\lambda(x_1,x_2,x_3)\right]^{\frac{1}{p}}=$$

$$\left[\int_{\mathcal{X}_1\times\mathcal{X}_3}\|x_1-x_2\|^p d\lambda_{12}(x_1,x_2)\right]^{\frac{1}{p}}+\left[\int_{\mathcal{X}_3\times\mathcal{X}_3}\|x_2-x_3\|^p d\lambda_{23}(x_1,x_2)\right]^{\frac{1}{p}}=$$

$$W_p(\mu_1,\mu_2)+W_p(\mu_2,\mu_3)$$

Q.E.D.

5. Connection with random set statistics

One more useful statistical methodology arising from optimal transport theory was the unexpected connection between the current topic of partial identification (of statistical models) and random set statistics via optimal transport, as pointed out by Galichon (2016).

First, it seems here is a good place to spell out briefly what is statistics and how statisticians should conduct statistics! Roughly speaking, statistics is about finding the truth from data, and statistical works should be credible.

Unlike physical science, we need models to conduct statistics. Based upon observed data, statisticians propose models. A model is a subjective (stochastic) equation together with a set of assumptions supporting it. Of course each model contains unknown "parameters" which need to be estimated (from data) to specify it for, e.g. prediction and decision-making.

As we all know (since we "follow" the traditional approach to without any hesitation) that the maintained assumptions (whether they are justified or not) are there to allow us to use available data to consistently estimate the model parameters, noting that estimability of parameters in this sense is related to the notion of identification.

In order to justify our statistical estimation of our model parameter, say, in the model $\{F_\theta\ \theta\in\Theta\}$, we impose assumptions to make the true (but unknown) parameter θ_o identifiable (i.e. point identifiable) in the sense that the map $\theta\to F_\theta$ is injective. A well-known example for all is the linear supply and demand model in microeconomics. General supply and demand models are provided by economic theory, but when a text book advises us to use a linear model (for simplicity?), it puts down assumptions without justifications to make sure that the model parameter of interest is point identifiable.

If the maintained assumptions are not plausible, the map $\theta\to F_\theta$ might not be injective, i.e. there are $\theta'\neq\theta$ such that $F_\theta=F_{\theta'}$ (θ and θ' are said to be observationally equivalent) so that the model parameter is not point-identifiable, such as in games with multiple Nash equilibria. In such a situation, should we give up the analysis or the empirical attempt ? No, as Manski (e.g. 2007) put it, we could live with it and look for a new way to estimate the model parameter, not as a point but as a subset of the parameter space, called the identified set. Thus, estimating an identified set is the main goal for partially identified statistical models.

In this improved statistical setting, we are facing partially identified models where point estimation becomes set estimation. But when the estimation target is a set, the identified set (i.e. set of observationally equivalent parameters), its estimated set is a random set (a set-valued function of the data). Thus, we are facing a natural extension of classical statistics, namely statistics with random sets rather with random points.

Now, the general theory of probability supporting statistical analysis should cover the theory of random sets (as an extension of random vectors) which are well defined random elements. See Matheron (1975) or Nguyen (2006). In other words, in view of credible statistics, statistics of random sets should take a central stage in empirical research. However, the statistical theory of set-valued statistics is still young. In some contexts, e.g. estimating the level sets of an unknown probability density function, the estimation method is Hartigan's

(1987) excess mass which is the counterpart of maximum likelihood method in traditional statistics. See also Nguyen (2006).

What is "interesting" is that some partial identification problems can be formulated as an optimal transport problem which in turn provides a connection with random sets useful for computational purposes. See Galichon (2016) for details. Here we elaborate a bit on the theory of random sets since after all as the identified set is a set, its estimator will be a random set statistic, and we need to investigate its properties just like the special case of random vector statistics. The point is this. While random set statistics is the natural approach to inference about set parameters in partially identified models, the context in which these partially identified models can be formulated as optimal transport problems brings out specific ways for conducting inference.

Now, in spirit, partial identification setting is somewhat similar to statistics with coarse data where the data from the desired DGP (an unknown distribution) are not observable, but instead the data from a random set containing it are observed, i.e. the latent random variable of interest is an almost sure selector of the observed random set. As such, it is related to the estimation of the identified set from a random set viewpoint.

In Galichon's analysis (2016) the focus is the identification of an identified set of a partially identified model, and the connection with random set is based upon a result of Artstein (1983) which is generalized by Norberg (1992) as follows.

First of all, capacity functionals play the role of probability laws of random (closed) sets on $\mathbb{R}^d$ by Choquet's Theorem (the counterpart of Lebesgue- Stieltjes Theorem for random vectors), see Nguyen (2006) for an introduction. Two capacity functionals T_1, T_2 are said to form an ordered coupling if there exists a common probability space $(\Omega, \mathcal{A}, P)$ on which are defined two random closed sets S_1, S_2 such that $S_2 \subseteq S_1$ P, i.e. where S_1, S_2 have T_1, T_2 as capacity functionals, respectively. When the random set S_2 is single-valued, a special case which is identified with a random vector, it becomes an a.s. selector of S_1, i.e. $P(S_2 \in S_1) = 1$. This special case corresponds to the situation in coarse data analysis as well as in partial identification estimation (of identified sets). An useful result from random set theory for it is the following which allows us to characterize an identified set as the core of a capacity functional of a random set.

Theorem (Norberg, 1992). Let μ be a probability measure on $\mathcal{B}(\mathbb{R}^d)$ and T be a capacity functional, then the following are equivalent:

(1) $\mu \leq T$ on compact sets of $\mathbb{R}^d$

(2) There exists a common probability space $(\Omega, \mathcal{A}, P)$ on which are defined a random closed set S with capacity functional T and a random vector X with law μ, and which is an a.s. selector of S.

We elaborate a bit on the essentials of random set theory to introduce statisticians to random set statistics.

Just like traditional or standard way to start a probability theory for statistical applications, we consider the simple situation where random quantities take values in a finite set.

Let U be a finite set with n elements. The power set of U is denoted as 2^U (set of functions $U \to \{0, 1\}$). For $A \subseteq U$, $\#(A)$ denotes the number of elements of the subset A.

The source of uncertainty is a probability space $(\Omega, \mathcal{A}, P)$. A map $X(.) : \Omega \to 2^U$ is a finite random set (a set obtained at random).

The law of X is the probability measure P_X on the power set of 2^U, where $P_X(.) = P \circ X^{-1}$ (the pushforward of P by X).

As in the case of finite random variables, P_X is completely determined by the probability density of X, namely $f(.) : 2^U \to [0, 1]$ where $f(A) = P(X = A)$. Alternatively, P_X is characterized

by the distribution function $F(.): 2^U \to [0, 1]$, where $F(A) = P(X \subseteq A)$. The counterpart of the characterization of distribution functions of random variables is this.

A set-function $F(.): 2^U \to [0, 1]$ is a distribution function of a (finite) random set X if it satisfies the following conditions:

(1) $F(\varnothing) = 0, F(U) = 1$

(2) For any $k \geq 2$, and A_i, $i = 1, 2, \ldots, k$, subsets of U,

$$F\left(\cup_{i=1}^{k} A_i\right) \geq \sum_{\varnothing \neq I \subseteq \{1,2,\ldots,k\}} (-1)^{\#(I)+1} F(\cap_{i \in I} A_i)$$

Alternatively, since $T(A) = P(X \cap A \neq \varnothing) = 1 - F(A^c)$, the law of X can be also characterized by the set function $T(.): 2^U \to [0, 1]$, called the capacity functional of X.

Axiomatically, a capacity function is a function $T(.): 2^U \to [0, 1]$ satisfying the following:

(1) $T(\varnothing) = 0, T(U) = 1$

(2) For any $k \geq 2$, and A_i, $i = 1, 2, \ldots, k$, subsets of U, we have

$$T\left(\cap_{i=1}^{k} A_i\right) \leq \sum_{\varnothing \neq I \subseteq \{1,2,\ldots,k\}} (-1)^{\#(I)+1} T(\cup_{i \in I} A_i)$$

Now let X be a non-empty random set on the finite set U (i.e. $P(X = \varnothing) = 0$). The core $\mathcal{C}(T)$ of its capacity functional T is the set of probability measures μ on U such that $\mu(.) \leq T(.)$.

We extend all the above to the case where the sampling space $U = \mathbb{R}^d$.

Remember that, for random vectors, i.e. random elements taking values in $\mathbb{R}^d$, their probabilistic background was based on the theory of measures on the Borel measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ where the Borel σ – field $\mathcal{B}(\mathbb{R}^d)$ is constructed using the topology of $\mathbb{R}^d$. For random sets taking values as subsets of $\mathbb{R}^d$, we need a topology on $2^{\mathbb{R}^d}$. Now a random vector is identified as a random set taking singletons as values. But each $\{x\}$ is a closed set of $\mathbb{R}^d$. Thus, following Matheron (1975), we consider random sets taking values as closed subsets of $\mathbb{R}^d$, denoted as $\mathcal{F}(\mathbb{R}^d)$ on which a "hit-or-miss topology" is established to obtain its Borel σ – field, denoted as $\mathcal{B}(\mathcal{F})$.

A random closed set, defined on a probability space $(\Omega, \mathcal{A}, P)$ (its source of uncertainty), is a map $X(.): \Omega \to \mathcal{F}(\mathbb{R}^d)$, $\mathcal{A} - \mathcal{B}(\mathcal{F})$- measurable. Its probability law is the probability P_X on $\mathcal{B}(\mathcal{F})$ obtained as $P_X(.) = P \circ X^{-1}(.)$.

The notion of capacity functionals in the finite case is extended as follows. Let $\mathcal{K}(\mathbb{R}^d)$ denote the set of compact subsets of $\mathbb{R}^d$. Then $T(.): \mathcal{K}(\mathbb{R}^d) \to \mathbb{R}$ is called a capacity functional if it satisfies:

(1) $0 \leq T(.) \leq 1, T(\varnothing) = 0$

(2) For any $k \geq 2$, and A_i, $i = 1, 2, \ldots, k$, subsets of U, we have

$$T\left(\cap_{i=1}^{k} A_i\right) \leq \sum_{\varnothing \neq I \subseteq \{1,2,\ldots,k\}} (-1)^{\#(I)+1} T(\cup_{i \in I} A_i)$$

(3) If $K_n \in \mathcal{K}(\mathbb{R}^d)$ and $K_n \searrow K \in \mathcal{K}(\mathbb{R}^d)$ then $T(K_n) \searrow T(K)$.

The counterpart of Lebesgue-Stieltjes Theorem is the Choquet's Theorem: If $T(.): \mathcal{K}(\mathbb{R}^d) \to \mathbb{R}$ is a capacity functional, then there exists a unique probability measure Q on $\mathcal{B}(\mathcal{F})$ such that, for all $K \in \mathcal{K}(\mathbb{R}^d)$, $Q(\mathcal{F}_K) = T(K)$, where $\mathcal{F}_K = \{A \in \mathcal{F}(\mathbb{R}^d) : A \cap K \neq \varnothing\}$.

In other words, the capacity functional characterizes the probability law of a random closed set. The core of a capacity functional is the set of probability measures μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that $\mu \leq T$ on $\mathcal{K}(\mathbb{R}^d)$. Norberg's Theorem (1992) is valid for $\mathbb{R}^d$ so that the core of a capacity functional (of a random closed set on $\mathbb{R}^d$) is related to identified sets in partially identified statistical models.

References

Artstein, Z. (1983), "Distributions of random sets and random selections", *Israel Journal of Mathematics*, Vol. 46, pp. 313-324.

Belloni, A. and Winkler, R. (2011), "On multivariate quantiles under partial order", *The Annals of Statistics*, Vol. 39 No. 2, pp. 1125-1179.

Brenier, Y. (1991), "Polar factorization and monotone rearrangement of vector-valued functions", *Communications on Pure and Applied Mathematics*, Vol. 44 No. 4, pp. 375-417.

Breiman, L. (2001), "Statistical modeling: the two cultures", *Statistical Science*, Vol. 16 No. 4, pp. 199-231.

Carlier, G., Chernozukov, V. and Galichon, A. (2016), "Vector quantile regression: an optimal transport approach", *The Annals of Statistics*, Vol. 44 No. 3, pp. 1165-1192.

Carlier, G., Chernozukov, V. and Galichon, A. (2017), "Vector quantile regression beyond the specific case", *Journal of Multivariate Analysis*, Vol. 161, pp. 96-102.

Daoud, A. and Dubhashi, D. (2023), "Statistical modeling: the three cultures". doi: 10.1162/99608f92.89f6fe66.

Dudley, R.M. (2003), *Real Analysis and Probability*, Cambridge University Press, Cambridge, MA.

Galichon, A. (2016), *Optimal Transport Methods in Economics*, Princeton Univ. Press., Princeton, NJ.

Graf, S. and Mauldin, R.D. (1989), "A classification of disintegration of measures", in *Measure and Measurable Dynamics*, No 94, Contemp.Math, pp. 147-158, AMS.

Hallin, M., Paindaveine, D. and Siman, M. (2010), "Multivariate quantiles and multiple-output regression quantiles: from L1 optimization to half-space depth", *The Annals of Statistics*, Vol. 38 No. 2, pp. 635-669.

Hartigan, J.A. (1987), "Estimation of a convex density contour in two dimensions", *Journal of the American Statistical Association*, Vol. 82 No. 397, pp. 267-270.

Kantorovich, L.V. (1942), "On the translocation of masses", *C.R. Academy of Sciences of URSS*, Vol. 77, pp. 199-201.

Koenker, R. (2017), "Quantile regression 40 years on", *Annual Review of Economics*, Vol. 9 No. 1, pp. 155-176.

Koenker, R. and Basett, G. (1978), "Regression quantiles", *Econometrica*, Vol. 46, pp. 33-50.

Manski, C. (2007), *Identification for Prediction and Decision*, Harvard University Press, Cambridge, MA.

Matheron, G. (1975), *Random Sets and Integral Geometry*, J. Wiley, New York.

McCam, R. (1995), "Existence and uniqueness of monotone measure-preserving maps", *Duke Math. J.*, Vol. 80 No. 2, pp. 309-323.

Monge, G. (1781), "Memoire sur la theorie des deblais et des remblais", in *Histoire de l' Academie Royale des Sciences de Paris*, pp. 666-704.

Nguyen, H.T. (2006), *A Introduction to Random Sets*, Chapman & Hall/CRC, Boca Raton, FL.

Norberg, T. (1992), "On the existence of ordered coupling of random sets with applications", *Israel Journal of Mathematics*, Vol. 77 No. 3, pp. 241-264.

Parzen, E. (1979), "Nonparametric data modeling", *Journal of the American Statistical Association*, Vol. 74 No. 365, pp. 105-121.

Serfling, R. and Zuo, Y. (2010), "Discussion", *The Annals of Statistics*, Vol. 38 No. 2, pp. 676-684.

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

Corresponding author
Hung T. Nguyen can be contacted at: hunguyen@nmsu.edu