An Invitation to Multivariate Quantiles Arising from Optimal Transport Theory

Hung T. Nguyen[1,2(✉)]

[1] New Mexico State University, Las Cruces, USA
hunguyen@nmsu.edu
[2] Chiang Mai University, Chiang Mai, Thailand

Abstract. While a variety of new concepts and methods arised from Optimal Transport theory recently in the literature, they are somewhat theoretical for empirical researchers, including statisticians and econometricians. This tutorial paper aims at elaborating on one of these new concepts and methods, namely multivariate quantile functions, in order to invite empirical researchers to take a closer look at this new concept to apply to their empirical works, such as multivariate quantile regression.

Keywords: Gradient of convex functions · Multivariate quantiles · Optimal transport · Quantile regression

1 Introduction

Motivated by economics issues, in 1942, Kantorovich reformulated (and solved) the unsolved "Optimal Transport" (OT) problem of Gaspard Monge (1781) and got the Nobel Prize in Economics (shared with Koopmans) in 1975, for their contributions to optimal allocation of resources.

Recently, it was "discovered" that OT provides a variety of modern methods for economics. This tutorial paper focuses only on one of these modern methods, namely multivariate quantile functions for quantile regression and related topics.

Mean linear regression models are possible (as it is obvious how to generalize the mean of a random variable to the mean of a random vector) and are useful when dealing with multivariate distribution functions. Now, over 40 years since univariate quantile regression was invented (Koenker and Bassett [5]), can we extend it to multivariate quantile regression in some acceptable way? Of course, like multivariate *mean* linear regressions, multivariate quantile regressions should be very useful in a variety of contexts.

Since there is no total order relation on $\mathbb{R}^d$ when $d > 1$, a direct extension of univariate quantile functions to higher dimensions is hopeless. Thus, in order to obtain a "correct" way to generalize univariate quantile functions, we must look for some other ways. Generalizations of mathematical concepts appear often in mathematics. When Lotfi Zadeh generalized crisp sets to fuzzy sets, he cannot do it directly, so he took

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*, SSDC 483, pp. 1–37, 2024.
https://doi.org/10.1007/978-3-031-35763-3_1

an equivalent definition of a crisp set, namely its indicator function which is a function taking only values 0 and 1, and extend it to the unit interval $[0,1]$. To view Kolmogorov probability theory as a special case of quantum probability theory, we can take an equivalent representation of finite standard probability, namely, identifying a random variable, an event, and a probability measure, as diagonal matrices, and then extend them to arbitrary self adjoint matrices.

Now, in the above spirit, to generalize a univariate quantile function $F^{[-1]}(.) : [0,1] \to \mathbb{R}$, defined as $F^{[-1]}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}$, of a real-valued random variable X with distribution function F, we look at some appropriate equivalent representation.

Note that, since we are going to derive a characterization of a univariate quantile function in the setting of Optimal Transport theory, we denote it as $F^{[-1]}$ instead of F^{-1} to avoid a possible confusion with the set-valued set-function $T^{-1}(.) : 2^{\mathbb{R}} \to 2^{[0,1]}$ of a map $T(.) : [0,1] \to \mathbb{R}$, pushing the uniform probability measure du on $[0,1]$ to a probability measure on $\mathscr{B}(\mathbb{R})$, since actually, for $T = F^{[-1]}$, we have $dF = T\#du = du \circ T^{-1}$!

The first characteristic property of $F^{[-1]}(.)$ is this: If U is a random variable, uniformly distributed on the unit interval $[0,1]$, then the random variable $F^{[-1]}(U)$ has the same distribution F as X, which is the basis of simulations. But saying that $X \overset{D}{=} F^{[-1]}(U)$ simply means that the probability "law" of X, written as $dF(-\infty, x]) = F(x)$, is the probability measure on $\mathscr{B}(\mathbb{R})$ obtained from the uniform probability du on $\mathscr{B}([0,1])$ via $du \circ (F^{[-1]})^{-1}$, written as $F^{[-1]}\#du = dF$ (a notation we will use in the context of Optimal Transport Theory), meaning "The transport map $F^{[-1]} : [0,1] \to \mathbb{R}$ pushes the probability du on $[0,1]$ forward to the probability dF on $\mathbb{R}$".

There is another property of $F^{[-1]}(.)$, kind of "hidden", since we did not use it often.

From the explicit definition of $F^{[-1]}(u)$, it is clear that the function $F^{[-1]}(.)$ is monotone non decreasing on $\mathbb{R}$, i.e., if $x \leq y$ then $F^{[-1]}(x) \leq F^{[-1]}(y)$, with is equivalent to: for any $x, y \in \mathbb{R}$, we have

$$(F^{[-1]}(x) - F^{[-1]}(y))(x-y) \quad \geq 0$$

and which can be generalized to higher dimensions (needed for our subsequent analysis), as follows. A function $g(.) : \mathbb{R}^d \to \mathbb{R}^d$ is monotone (non decreasing) if, for any $x, y \in \mathbb{R}^d$, we have

$$<g(x) - g(y), x - y> \quad \geq 0$$

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

These two properties characterize the univariate quantile function $F^{[-1]}(.) : [0,1] \to \mathbb{R}$. Thus, we should expect that a function $Q_F(.) : [0,1]^d \to \mathbb{R}^d$ is "called" the (multivariate) quantile function of the multivariate distribution function $F(.) : \mathbb{R}^d \to [0,1]$ if it possess these two "extended" properties, namely

(i) Q_F is monotone non decreasing on $\mathbb{R}^d$ (in the above equivalent sense),
(ii) Q_F pushes the uniform probability du on $[0,1]^d$ forward to the probability dF on $\mathbb{R}^d$, in symbol $Q_F\#du = dF$.

Having these requirements, let's see if we can get a candidate for Q_F in some "simple" way. To simplify the notations, consider the case where the dimension $d = 2$.

Thus, let $F(.) : \mathbb{R}^2 \to [0,1]$ a bivariate distribution of the random vector $X = (X_1, X_2)$, so that

$$F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2) = c(F_1(x_1), F_2(x_2))$$

where $c(.,.) : [0,1]^2 \to [0,1]$ is a bivariate copula capturing the dependence structure between the components of X.

The mean of the random vector $X = (X_1, X_2)'$ is defined componentwise, as a mean vector, namely $EX = (EX_1, EX_2)'$ (transpose).

Can we define bivariate quantile function componentwise?

Let $Q_F(.) : [0,1]^2 \to \mathbb{R}^2$ be defined as, for $u = (u_1, u_2) \in [0,1]^2$, $Q_F(u) = (F_1^{[-1]}(u_1), F_2^{[-1]}(u_2))'$.

a) Monotonicity is satisfied: let $v = (v_1, v_2)$, we have

$$< Q_F(u) - Q_F(v), u - v > =$$

$$[(F_1^{[-1]}(u_1) - F_1^{[-1]}(v_1))(u_1 - v_1)][(F_2^{[-1]}(u_2) - F_2^{[-1]}(v_2))(u_2 - v_2)] \geq 0$$

since both $F_1^{[-1]}, F_2^{[-1]}$ are monotone.

b) How about $Q_F \# du \stackrel{?}{=} dF$? We have

$$Q_F \# du((-\infty, a] \times (-\infty, b]) = du\{u : Q_F^{-1}((-\infty, a] \times (-\infty, b])\} =$$

$$du\{u : F_1^{[-1]}(u_1) \leq a, F_2^{[-1]}(u_2) \leq b\} = du\{u : u_1 \leq F_1(a), u_2 \leq F_2(b)\} =$$

$$F_1(a)F_2(b) \neq F(a,b)$$

unless $X = (X_1, X_2)$ has *independent components*, i.e., X_1, X_2 are independent. This is, in fact, expected since the componentwise definition $Q_F(u) = (F_1^{[-1]}(u_1), F_2^{[-1]}(u_2))'$ ignores the dependence structure of X_1 and X_2 (given by copulas).

Thus, $Q_F \# du \neq dF$, i.e., $X \stackrel{D}{\neq} dF$, in general, meaning that $Q_F(u) = (F_1^{[-1]}(u_1), F_2^{[-1]}(u_2))'$ is not a good candidate for what we could call a bivariate quantile function, a counterpart of univariate quantile function.

It turns out that a correct candidate for a multivariate quantile function came from an area of mathematics called *Optimal Transport (OT)* theory, in 2016. See Carlier et al. [2,3].

Let μ, ν be two Borel probability measures on $\mathbb{R}^d$, $d \geq 1$. A transport map sending μ to ν is a map $T(.) : \mathbb{R}^d \to \mathbb{R}^d$, such that $\nu(.) = \mu \circ T^{-1}(.)$, i.e., $T \# \mu = \nu$.

Of course, for $d=1$, and $\mu = du$, uniform on $[0,1]$ and $\nu = dF$, for arbitrary distribution function F on $\mathbb{R}$, the quantile function $F^{[-1]}(.)$ is a transport map sending du to dF.

Moreover, $F^{[-1]}$ is the *unique* monotone transport map (there are other transport maps, but $F^{[-1]}$ is the only one which is monotone).

Since monotonicity and measure-preserving # are concepts which are valid in any dimensions, the question of interest to us is: "Is there a unique monotone transport map on $\mathbb{R}^d$ for $\mu = du$, uniform on $[0,1]^d$, and arbitrary $\nu = dF$?". If the answer to it is yes, then we get our desired multivariate quantile function! The answer is in fact yes.

McCann Theorem (McCann [7]). Let $F(.) : \mathbb{R}^d \to [0,1]$ be an arbitrary multivariate distribution function, then there exists a unique gradient $\nabla\varphi(.) : [0,1]^d \to \mathbb{R}^d$ of some convex function $\varphi(.) : [0,1]^d \to \mathbb{R}$ (φ is not unique, but $\nabla\varphi$ is unique) such that $\nabla\varphi \# du = dF$, where du is the uniform probability measure on $[0,1]^d$.

Let's elaborate a bit on McCann's Theorem. In dimension 1, let $\nu = dF$ where F is the uniform distribution on the interval $[1,2]$, i.e.,

$$F(x) = \begin{cases} 0 & \text{for} \quad x < 1 \\ x & \text{for} \quad 1 \leq x \leq 2 \\ 1 & \text{for} \quad x > 2 \end{cases}$$

Then we know that $F^{[-1]}(.) : [0,1] \to \mathbb{R}$ is $F^{[-1]}(u) = 1+u$ which is monotone (non decreasing) since its derivative is positive. It is the derivative of the convex function $\varphi(u) = \frac{1}{2}(1+u)^2$. And of course, $F^{[-1]} \# du = dF$.

In dimension 1, the graph of a convex function lies above the tangent at each point x where the function is differentiable (a convex function is differentiable almost everywhere, with respect to the Lebesgue measure on $\mathbb{R}$), and as such its (a.e.) derivative is monotone non decreasing.

In dimension $d > 1$, the gradient is the vector of partial derivatives of the multivariate function. The whole graph of a convex function on $\mathbb{R}^d$ lies above each tangent hyperplane at each point where it is differentiable, as a consequence, the gradient $\nabla\varphi$ of the convex function φ is monotone non decreasing in the sense that, for any $x, y \in \mathbb{R}^d$, we have

$$<\nabla\varphi(x) - \nabla\varphi(y), x - y> \quad \geq 0$$

McCann's theorem is an existence theorem, it does not tell us how to obtain explicitly the multivariate quantile function in dimension $d > 1$. In other words, it is not a "constructive" theorem. There is much more work to do to get a "constructive" result, and we need it for applications.

It is precisely here that OT comes in.

In 1781, Gaspard Monge [8] considered the following problem. Let μ, ν be two probability measures on $\mathscr{X}, \mathscr{Y} \subseteq \mathbb{R}^3$, respectively. Let $c(.,.) : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}^+$ be a "cost" function (of moving the mass μ on $\mathscr{X}$ to the mass ν on $\mathscr{Y}$, think about "supply and demand"). A transport map $T(.) : \mathscr{X} \to \mathscr{Y}$ is a map such that $T\#\mu = \nu$. The Monge's problem (MP) is to find a transport map T^* which is optimal, with respect to the cost c, in the sense that it minimizes the total cost, i.e.,

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

This optimization problem is very difficult to study since the objective function is not linear in T, and the constraint set is not convex. This is why the problem was dormant for 200 years. Then, in 1942, Kantorovich, motivated by economic problems, solved it, earning him a Nobel Prize in Economics.

The (MP) might not even have solutions! So first of all, when a mathematician faces a such problem, she will enlarge the domain to have solutions, just like considering complex plane for solutions of equations, or extending pure (deterministic) strategies in games to mixed (random) strategies to have Nash equilibrium.

Kantorovich observed that if T is a solution of (MP), then $\gamma_T = \mu \circ (I,T)^{-1}$ is a joint probability measure on $\mathscr{X} \times \mathscr{Y}$ admitting μ and ν as its marginal measures, i.e., $\gamma_T(A \times \mathscr{Y}) = \mu(A)$, and $\gamma_T(\mathscr{Y} \times B) = \nu(B)$, for any $A \in \mathscr{B}(\mathscr{X})$, $B \in \mathscr{B}(\mathscr{Y})$. Therefore, the set of all joint probability measures with μ and ν as marginal measures, denoted as $\Pi(\mu,\nu)$, is larger than the set of transport maps in (MP). Note that, by (I,T), where I is the identity map on $\mathscr{X}$, $I(x) = x$, we mean the map $(I,T)(.) : \mathscr{X} \to \mathscr{X} \times \mathscr{Y}$, $(I,T)(x) = (x,T(x))$, so that $(I,T)^{-1}(.) : 2^{\mathscr{X}\times\mathscr{Y}} \to 2^{\mathscr{X}}$.

The Kantorovich problem (KP) is this. Find the optimal transport *plan* $\pi^* \in \Pi(\mu,\nu)$ such that

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

If a solution π^* of (KP) is of the form $\gamma_T = \mu \circ (I,T)^{-1}$, then, in it, T is a solution for (MP).

The breakthrough of Kantorovich is this. First of all, unlike (MP), the (KP) always have solutions since $\Pi(\mu,\nu) \neq \varnothing$ (the product measure $\mu \otimes \nu$ is in it).

Next, the (KP) seems "solvable" since its avoids the difficulties of (MP): The objective function is linear in π, and the constraint set $\Pi(\mu,\nu)$ is convex.

As such, the problem can be solved by duality, i.e., changing an "inf" problem to a "sup" problem in which constraints are written as (infinite) inequalities, suitable for using linear programming (invented by Kantorovich himself, 1942, of course, with the help from George B. Danzig). See Villani [11,12].

The following Sections will explain this program, at least as a gentle introduction, to obtain a constructive theory of multivariate quantiles functions.

2 A Closer Look at Quantiles

We are familiar with the notion of (univariate) quantiles when considering order statistics, say, in extreme value theory.

While in practice, we are mainly concerned with distributions of order statistics which are derived soly from the distribution of the population, you may not notice the extremely important role played by the quantile function of the population, although it is derived from the population distribution.

Since the notion of quantile function is essential in various contexts, such as risk analysis, regression models, but until recently is only available for univariate case, i.e., for real-valued random variables, it is desirable to extend it to the multivariate case, i.e., for random vectors, for applications.

The search for such an extension finally arrived (in 2016) by looking closely at the univariate quantile function, triggered by a paper of Brenier [1]. The buzz words in his paper are "polar factorization" and "Monotone rearrangement" of *vector-valued functions* In fact, this paper first triggered a return to Optimal Transport Theory (OT) since it is precisely in the setting of OT. Specifically, Brenier's paper is about extension of the above buzz words from *random variables to random vectors*.

Let $X_1, X_2, ..., X_n$ be a random sample drawn from a population X, i.e., the $X_j's$ are random variables independent and identically distributed (i.i.d.) as X. These values of X are in $\mathbb{R}$ in any possible order. Suppose we are interested in ordering these observed values of X, i.e., forming the order statistic $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$, we can just do it!

Can we do it in some more "sophisticated" way? i.e., providing a map that realizes such an ordering.

Note that the order statistic $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ is a *monotone rearrangement* of the values $X_1, X_2, ..., X_n$, i.e., arranging the unordered set $\{X_1, X_2, ..., X_n\}$ into the ordered set $\{X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}\}$. Of course, this is possible since $\mathbb{R}$ is totally ordered. Clearly, there is only one such monotone rearrangement.

It is right here that quantile function is related to order statistics. The empirical distribution of the sample is

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

Let the quantile function of F_n be $F_n^{[-1]}$. Then

$$F_n^{[-1]}(u) = X_{(j)} \quad \text{for } u \in [\frac{j-1}{n}, \frac{j}{n})$$

Thus, the quantile function $F_n^{[-1]}(.)$ (of F_n) realizes the monotone rearrangement of the observed values of X, noting that $F_n^{[-1]}(.)$ is a monotone non decreasing function.

In fact, we do get a stronger representation than $F^{[-1]} \# du = dF$, a weak representation of X, sufficient for simulation purpose, namely: there exists a random variable V distributed uniformly on $[0,1]$ such that $X \overset{a.s.}{=} F^{[-1]}(V)$, a *polar factorization* of X.

Thus, in summary, the quantile function $F^{[-1]} : [0,1] \to \mathbb{R}$ provides a polar factorization and a monotone rearrangement for the random variable X.

The next question is: What is the counterpart of $F^{[-1]}$ in higher dimensions?, i.e., for X being a random vector, taking values in $\mathbb{R}^d$, with $d > 1$.

The answer was given in Brenier's paper, and later generalized by McCann [7].

Now in the Text *Approximation Theorems of Mathematical Statistics (*Robert J. Serfling,1980), Serfling started (p. 2–3) as: Let $F(.) : \mathbb{R}^d \to [0,1]$ be the (multivariate) distribution function of a random vector $\mathbf{X} = (X_1, X_2, ..., X_d)$, defined on $(\Omega, \mathscr{A}, P)$. The mean of $\mathbf{X}$ is defined as the mean vector $E\mathbf{X} = (EX_1, EX_2, ..., EX_d)$.

How about quantiles? Well, without explaining why, he considered only the *univariate case ($d = 1$).*

We may ask: Why Serfling did not "consider", in a parallel way with distribution functions, the notion of quantiles for multivariate distribution functions (but only talked about means of random vectors)? It turns out that this definition of quantile functions for univariate distribution functions is "good" for *simulations*.

The simulation of a univariate random variable X with distribution function F is based on the fact that X and $F^{[-1]}(U)$, where U is the random variable uniformly distributed on the unit interval $[0,1]$, have the same distribution F. As such, if $U = u$, then $X = x = F^{[-1]}(u)$ is a simulated observation of X.

Remark. As far as simulation of random variables is concerned, we only need the "weak" representation of X, namely $X \overset{D}{=} F^{[-1]}(U)$, for any F, and uniformly distributed U on $[0,1]$. This representation is termed "weak" since the two random variables X and $F^{[-1]}(U)$ are "equal" only in distribution, i.e., having the same distribution, and not necessarily equal almost surely (with probability one) which is a stronger condition. We will see later that *there exists* some random variable $V \sim U$ such that $X \overset{a.s.}{=} F^{[-1]}(V)$.

How about *simulations of random vectors*? i.e., how to simulate random vectors when we *do not have* the counterpart notion of quantiles for multivariate distribution functions? In the above mentioned Text, simulation of random vectors is carried out as follows (based on univariate quantile functions only). We elaborate on it in the simple case of dimension two.

Let $\mathbf{X} = (X_1, X_2)$ with marginal distribution functions F_1, F_2, and joint distribution F. The simulation of $\mathbf{X} = (X_1, X_2)$ is based on the Rosenblatt transformation (1952). Let $F(x_1, x_2) = F_1(x_1)F(x_2|x_1)$, define $(x_1, x_2) \in \mathbb{R}^2 \to (u_1, u_2) \in [0,1]^2$ by

$$u_1 = F_1(x_1), u_2 = F(x_2|x_1)$$

The intent is to generate u_1, u_2 independently from a uniform distribution du on $[0,1]$, then solve the above system of equations (with known marginal and conditional distribution functions of course) to get $x_1 = F_1^{[-1]}(u_1)$, $x_2 = F_{X_2|X_1}^{[-1]}(u_2)$, and "view" them as simulated values for X_1, X_2. This can be justified if, e.g., $F_1^{[-1]} \circ F_1 = I$ (identity), and X_1, X_2 obtained this way is distributed as F. This requires that F is continuous, so that the Rosenblatt transformation produces (U_1, U_2) uniformly on $[0,1]^2$.

If X_1, X_2 are independent, i.e., $F(x_1, x_2) = F_1(x_1)F_2(x_2)$, then the simulation process is justified since then the vector $(F_1^{[-1]}(u_1), F_{X_2}^{[-1]}(u_2))' : [0,1]^2 \to \mathbb{R}^2$ pushes the uniform measure $d\mathbf{u}$ on $[0,1]^2$ to dF on $\mathbb{R}^2$.

Thus, $(F_1^{[-1]}(.), F_{X_2}^{[-1]}(.))'$ acts like a *multivariate quantile function* $Q_F(.) : [0,1]^2 \to \mathbb{R}^2$: monotone and $Q_F \# d\mathbf{u} = dF$.

As a final note, observe that for $d = 1$, the univariate distribution function F and its quantile function $F^{[-1]}$ (of a real-valued random variable X) are referred to as its *rank* and quantile functions, respectively. If F is continuous, then $F(X)$ is uniformly distributed on $[0,1]$. By McCann's theorem applied to $d = 1$, $F^{[-1]}$ is the (a.e.) unique

monotone map from $[0,1]$ to $\mathbb{R}$ such that $F^{[-1]}\#du = dF$, and if, in addition, F has a finite second moment, then

$$F^{[-1]} = \arg\min\{E(U - T(U))^2 : T\#du = dF\}$$

where U is the random variable uniformly distributed on $[0,1]$.

McCann's theorem for $d \geq 1$ affirms the existence and uniqueness of a "multivariate" quantile function $Q_F(.) : [0,1]^d \to \mathbb{R}^d$, monotone and $Q_F\#d\mathbf{u} = dF$, and if, in addition, F has a finite second moment, then

$$Q_F = \arg\min\{E||U - T(U)||^2 : T\#d\mathbf{u} = dF\}$$

3 Characterization of Univariate Quantile Functions

The notion of (univariate) quantiles is useful in various statistical analyses, mainly in *univariate* quantile regression (Koenker and Bassett [5]) which was developed based on characterizations of other quantities (for computation purposes).

The application of univariate quantile functions to simulations turns out to have a deeper effect.

Recall that a real-valued random variable X is a measurable map from $\Omega \to \mathbb{R}$, where its source of uncertainty is the "background" probability space $(\Omega, \mathscr{A}, P)$ and its range space is the measurable space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, and where X represents a measure-preserving map transporting the probability measure P, from its source of uncertainty, to its "law" P_X on its observation space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, i.e. $P \circ X^{-1} = P_X$. We also denote the law of X as $P_X = dF$, where $dF((-\infty, x]) = F(x)$.

Is there some other *concrete and equivalent* source of uncertainty that can replace $(\Omega, \mathscr{A}, P)$ and a measure-preserving map $T(.)$ from it to $(\mathbb{R}, \mathscr{B}(\mathbb{R}), P_X)$?

We are asking for another representation of X. The following is well known in simulations. If $F^{[-1]}(.) : [0,1] \to \mathbb{R}$ is the univariate quantile function of X (or of F), and U is a random variable uniformly distributed on $[0,1]$, then the random variable $F^{[-1]} \circ U : \Omega \to \mathbb{R}$ has the same distribution F, i.e., $X \overset{D}{=} F^{[-1]} \circ U$. Thus, if we know F, we can simulate X, i.e., obtaining simulated data from X: pick a random number u in $[0,1]$, then $F^{[-1]}(u) = x$ is an outcome from X. Note that this "concrete" specification of X (with its source on uncertainty being $([0,1], \mathscr{B}([0,1], du))$ is used for simulation purpose.

Proof of $X \overset{D}{=} F^{[-1]} \circ U$. Let's clarify first the following. If $Y = T(X)$, then

$$P(Y \in A) = P(T(X) \in A) = P(X \in T^{-1}(A))$$

so that $P_Y(A) = P_X(T^{-1}(A))$. Thus, $X \overset{D}{=} F^{[-1]} \circ U$ means, for any $A \in \mathscr{B}(\mathbb{R})$, we have $dF(A) = du((F^{[-1]})^{-1}(A)$, where du denotes the probability measure of U.

As stated earlier, for probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, it suffices to consider A of the form $A = (-\infty, x]$. We have

$$(F^{[-1]})^{-1}((-\infty,x]) = \{u \in [0,1] : F^{[-1]}(u) \in (-\infty,x]\} =$$

$$\{u \in [0,1] : F^{[-1]}(u) \leq x\} = \{u \in [0,1] : F(x) \geq u\}$$

since

$$F^{[-1]}(u) \leq x \Longleftrightarrow F(x) \geq u$$

therefore,

$$du\{(F^{[-1]})^{-1}((-\infty,x])\} = du\{\{u \in [0,1] : F(x) \geq u\}\} =$$

$$F(x) = dF((-\infty,x])$$

Thus, we have the"concrete" probability space $([0,1],\mathscr{B}([0,1]),du)$, replacing the abstract $(\Omega,\mathscr{A},P)$, and the *polar factorization* $X = F^{[-1]}(U)$ (by analogy of polar factorization of complex numbers, or of matrices) which requires the quantile function $F^{[-1]}(.) : [0,1] \to \mathbb{R}$, which is a measure-preserving map, pushing the probability measure du on $[0,1]$ to the probability measure dF on $\mathbb{R}$. (in symbol $dF = (F^{[-1]})\#du$).

Remark. Following the standard notations in Optimal Transport Theory, when a map T is a push forward for a probability μ to a probability measure ν, i.e., $\nu = \mu T^{-1}$, we write $\nu = T\#\mu$. Thus, $X \overset{D}{=} F^{[-1]}(U)$ means $dF = F^{[-1]}\#du$.

The univariate quantile function $F^{[-1]}$ satisfies two properties:

(1) $F^{[-1]}$ is monotone (non decreasing), and hence the derivative of some convex function,
(2) $dF = (F^{[-1]})\#du$: it pushs du forward to dF.

These properties are well known, but what is "new" is that they characterize univariate quantile functions, in the sense that they are obtained from an "abstract" setting, without evoking the total order of the underlying space $\mathbb{R}$ in the explicit definition of $F^{[-1]}$.

Specifically, there is only one map $T(.) : [0,1] \to \mathbb{R}$ satisfying these two conditions. In other words, if a map $T(.) : [0,1] \to \mathbb{R}$ is monotone (non decreasing) and $dF = T\#du$, then it is $F^{[-1]}(.)$.

Of course, that remains to be proved. But before that, let's announce what we are going to proceed. Once we prove this characterization of $F^{[-1]}$, we can use it to generalize to *multivariate quantile functions of random vectors*, without bother about the lack of a total order relation on $\mathbb{R}^d$ when $d > 1$, *thanks to Optimal Transport Theory* (it is precisely because of OT that Econometricians discover the above characterization of univariate quantile function for generalization to higher dimensions which is so needed in applications, but for so long, no such a generalization is available).

Specifically, the two characteristic properties of the univariate quantile function $F^{[-1]}$ can be addressed on $\mathbb{R}^d$ when $F^{[-1]}$ is replaced by a map $Q_F(.) : [0,1]^d \to \mathbb{R}^d$ which is monotone (non decreasing) as: for any $x,y \in [0,1]^d$,

$$<Q_F(x) - Q_F(y), x - y> \ \geq 0$$

where $<.,.>$ is the scalar product on $\mathbb{R}^d$. Clearly, the property (2) is meaningful on any probability spaces.

Note that the property (1) is very important! even, usually we do not emphasize it. Being a monotone non decreasing function, $F^{[-1]}$ is qualified as the derivative a some convex function. For example, if $F^{[-1]}(u) = u + 1$, then it is the derivative of the convex function $\frac{1}{2}(u+1)^2$. Note that, for a convex function $\varphi(.) : \mathbb{R}^d \to \mathbb{R}$, its gradient (vector of partial derivatives) $\nabla\varphi(.) : \mathbb{R}^d \to \mathbb{R}^d$ is monotone non decreasing in the above sense.

So what we will do next in this Section is to show the following. Consider the probability spaces $([0,1], \mathscr{B}([0,1]), du)$, and $(\mathbb{R}, \mathscr{B}(\mathbb{R}), dF)$. While we know that $F^{[-1]}(.) : [0,1] \to \mathbb{R}$ is a map having two properties (1), (2) above, we need to show two more things, namely its uniqueness, and optimality. Why? Well, our purpose is to characterize $F^{[-1]}$ in a setting (which will be Optimal Transport/OT) suitable for generalizing to higher dimensions.

Without exaggeration, it can be said that, like Copulae, OT theory will invade statistics of this 21st century!

Uniqueness of $F^{[-1]}$. Suppose T is a monotone non decreasing map and $T\#du = dF$. We are going to show that $T = F^{[-1]}$ so that $F^{[-1]}$ is unique.

Proof. By monotonicity of T, we have

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

so that

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

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

Suppose the inequality is strict. This means that there exists $\varepsilon_o > 0$ such that $F(T(x) - \varepsilon) \geq F_{du}(x)$ for every $\varepsilon \in [0, \varepsilon_o]$. Also, since $T^{-1}((-\infty, T(x) - \varepsilon) \subseteq (-\infty, x)$, we have $F(T(x) - \varepsilon) < F_{du}(x)$. Thus, $F(T(x) - \varepsilon) = F_{du}(x)$ for any $\varepsilon \in [0, \varepsilon_o]$. Note that $F(T(x) - \varepsilon)$ 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 $T(x) > F^{[-1]}(x)$ are contained in $\cup_j \{x : F_{du}(x) = y_j\}$ which is du–negligible (since du is atomless). As a consequence, $T(x) = F^{[-1]}(x)$, du–almost everywhere.

Remark. More generally, if μ, ν are Borel probability measures on $\mathbb{R}$, with supports $\mathscr{X}, \mathscr{Y} \subseteq \mathbb{R}$, respectively, with μ atomless, there is a unique, monotone non decreasing transport map, namely $x \to F_\nu^{[-1]}(F_\mu(x))$.

Optimality of $F^{[-1]}$. A transport map (monotone or not) $T(.) : [0,1] \to \mathbb{R}$ is a measure-preserving map, i.e., $T\#du = dF$. By optimality, we mean the following. Let $c(.,.)$:

$[0,1]\times\mathbb{R}\to\mathbb{R}^+$ be a "cost" function (of transporting elements of $[0,1]$ to elements of $\mathbb{R}$). A transport map T^* is optimal, with respect to c if

$$T^*=\arg\min\{\int_0^1 c(u,T(u))du : T : T\#du=dF\}$$

If the cost function is of the form $c(u,x)=h(u-x)$ with $h(.):\mathbb{R}\to\mathbb{R}$ strictly convex (such as $h(y)=y^2$), then, independent of c, $F^{[-1]}$ is optimal, i.e., we got an explicit formula for the unique monotone transport map, in this one-dimensional case. We will illustrate this via an example here. With a bit of Optimal Transport theory in the next Section, we will provide a general theorem in higher dimensions, together with a dual formulation for the computation of the optimal solution.

Let $([0,1],\mu=du)$, and $([1,2],\nu=dv)$, where dv is uniform on $[1,2]$ ($dv=dF$), and $F^{[-1]}(.)=[0,1]\to\mathbb{R}$: $F^{[-1]}(x)=x+1$. We will check that it is the unique monotone and optimal map. Clearly, it is monotone (and is the derivative of some convex function, e.g., $\frac{1}{2}(1+x)^2$). That $F^{[-1]}\#du=dv$ because $X\overset{D}{=}F^{[-1]}(U)$ where $U\simeq du$, and $X\simeq dv$. By the above proof of uniqueness, it is the only monotone map pushing du on $[0,1]$ to dv on $[1,2]$.

It remains to show that it is optimal with respect to convex cost function, such as $c(.,.):[0,1]\times[1,2]\to\mathbb{R}^+$, $c(u,v)=h(u-v)$, with $h(.)$ convex, e.g., $h(x)=x^2$.

Let $T(.):[0,1]\to[1,2]$ be a transport map, i.e., $T_{\#du}=dv$, monotone or not. The total cost of T is

$$C(T)=\int_0^1 (u-T(u))^2du$$

We have, using Jensen's inequality ($h(EX)\leq Eh(X)$):

$$C(T)=\int_0^1 h(T(x)-x)dx\geq h[\int_0^1 (T(x)-x)dx]=$$

$$h[\int_0^1 T(x)dx-\int_0^1 xdx]=$$

Note that the following are non-monotone transport maps: $T(x)=2-x$, and

$$S(x)=\begin{cases}x+\frac{3}{2} & \text{for} \quad x\in[0,\frac{1}{2}]\\ 2-x & \text{for} \quad x\in[\frac{1}{2},1]\end{cases}$$

with

$$C(T)=\int_0^1 (x-T_2(x))^2dx=\int_0^1 (2x-2)^2dx=\frac{4}{3}$$

$$C(S)=\int_0^1 (x-T_3(x))^2dx=\int_0^{\frac{1}{2}}(\frac{3}{2})^2dx+\int_{\frac{1}{2}}^1 (2x-2)^2=\frac{31}{24}$$

whereas

$$C(F^{[-1]}) = \int_0^1 (x - T_1(x))^2 dx = 1$$

which is the smallest.

Note that, in the above calculations, we only use the fact that $h(.)$ is (strictly) convex, but not its specific form. Thus, in fact, $F^{[-1]}$ is optimal with respect to any convex loss.

4 Optimal Transport and Multivariate Quantiles

We elaborate a bit on the theory of Optimal Transport from which to derive multivariate quantile functions.

As the polar factorization of a real-valued random variable is $X \overset{D}{=} F^{[-1]}(U)$, we are looking for the polar factorization of a random vector $\mathbf{X} \overset{D}{=} Q(\mathbf{U})$.

We are interested in the question: What could be the counterpart of a univariate quantile function $F^{[-1]}$ in higher dimensions, i.e., for a multivariate distribution function F on $\mathbb{R}^d$, with $d > 1$? The lack of a total order on $\mathbb{R}^d$ seems responsible for unsuccessful attempts in the past.

We are interested in quantile functions of distribution functions for a variety of reasons. We know very well what is the quantile function $F^{[-1]}$ explicitly of a univariate distribution function F of a random variable X, for arbitrary distribution function, heavy-tailed or not.

The characterization of $F^{[-1]}$ in Sect. 3 serves as a prototype for a generalization to higher dimensions. Thus, first, we call upon McCann's theorem to have the existence of a unique measure-preserving map $T : [0,1]^d \to \mathbb{R}^d$, $T\#du = dF$, where du is the uniform probability measure on $[0,1]^d$, and F is an arbitrary multivariate distribution function on $\mathbb{R}^d$, with T being monotone. Then, we rely upon Brenier's theorem to emphasize that such T in MacCann's theorem is "optimal" in Monge's problem (MP) which, in fact, also optimal in Kantorovich extended problem (KP). While (KP) is "solvable", we need a dual formulation to get the solution, via linear programming, and finally obtain a computable form of our desired transport map which will be our multivariate quantile function for the multivariate distribution function F on $\mathbb{R}^d$.

In one dimension, the quantile function $F^{[-1]}$ of the univariate distribution function F (of a real-valued random variable X) is the unique monotone map from $[0,1]$ to $\mathbb{R}$ such that $X \overset{D}{=} F^{[-1]}(U)$ (a polar factorization of X), where U is the random variable uniformly distributed on $[0,1]$, i.e., with $F_U(u) = u$, or with probability measure du on $[0,1]$, or equivalently $dF = du \circ (F^{[-1]})^{-1}$, in symbol. $F^{[-1]}\#du = dF$.

Quick question: Is there a polar factorization for $F(.) : \mathbb{R}^d \to [0,1]$, with $d > 1$? The answer is yes!

McCann's Theorem. There exists a unique (du-a.e., where du is the uniform probability measure on $[0,1]^d$) measurable map $Q_F(.) : [0,1]^d \to \mathbb{R}^d$ which is the gradient of some convex function $\varphi(.) : [0,1]^d \to \mathbb{R}$ (hence monotone) and such that $Q_F\#du = dF$ (i.e., $du \circ Q_F^{-1}(.) = dF(.)$).

Thus, if we let X be the random vector with multivariate distribution function F on $\mathbb{R}^d$, and U being the random vector uniformly distributed on the unit cube $[0,1]^d$, then we have $X \stackrel{D}{=} Q_F(U)$.

Note that the *multivariate quantile function* Q_F of F on $\mathbb{R}^d$ exists for any distribution functions F (just like in dimension 1 where $F^{[-1]}$ is defined regardless whether dF has finite moments or not). In dimension 1, $F^{[-1]}(.) : [0,1] \to \mathbb{R}$ is monotone and $F^{[-1]}\#du = dF$, it is $\nabla\varphi$ in view of the uniqueness in McCann's theorem.

When dF has finite second moment, $\nabla\varphi = F^{[-1]}$ in McCann's theorem is "optimal", with respect to square loss function, in the following Monge's problem (MP):

$$\nabla\varphi = \arg\min\{\int_0^1 (x - T(x))^2 dx : T\#du = dF\}$$

as we have seen in Sect. 3. In fact, $F^{[-1]}$ can be determined as $\nabla\varphi$ in the MP above. In fact, this situation is general (by Brenier's theorem).

If we are just interested in the existence of vector quantiles (i.e., quantile functions of random vectors) then McCann's theorem is enough. However, if we want to use vector quantiles to conduct, say, multivariate quantile regression, or to define multivariate (financial) risk measures, then their existence is not enough! We need to determine them explicitly for applications.

For dimension $d > 1$, the situation is not simple (!) as the Monge's minimization is somewhat intractable because its objective function is not linear in T, and the constraint set $\{T : T\#du = dF\}$ is not convex. We need to avoid these difficulties by embedding (MP) into the Kantorovich problem (KP) to use linear programming in its dual formulation. Such a program will help us to "compute" multivariate quantile functions. Thus, we need to evoke a bit of Optimal Transport (OT) theory.

Roughly speaking, observe that if T is in the constraint set of (MP), then $du \circ (I,T)^{-1}$ is a joint probability measure on $[0,1] \times \mathbb{R}$ having du, dF as marginal measures, we consider the (larger) convex constraint set $\Pi(du,dF)$ of all joint measures with du, dF as marginals, and the linear objective function (in $\pi \in \Pi(du,dF)$)

$$\pi \to \int_{[0,1]\times\mathbb{R}} c(x,y)d\pi(x,y)$$

and address the Kantorovich problem (KP)

$$\min\{\int_{[0,1]\times\mathbb{R}} c(x,y)d\pi(x,y) : \pi \in \Pi(du,dF)\}$$

which is "tractable", thanks to duality. If the (KP) has a solution of the form $\pi_T = du \circ (I,T)^{-1}$ then T will be a solution of the (MP).

To complete our agenda description, here is what we will proceed. The (MP) is enlarged to (KP) in view of

$$\{T : T\#du = dF) \subseteq \Pi(du,dF)$$

by the identification of T with $\pi_T = du \circ (I,T)^{-1}$. While the (KP) is linear under convex constraint set, its constraint set is not expressed as inequalities (in infinitely

dimensional form). Thus, we need to use duality, i.e., relating the "inf" problem of (KP) to a "sup" problem whose constraint set is expressed as inequalities, and then using linear programming to solve it, noting that Kantorovich is the inventor of linear programming (for solving Monge's original problem).

While we seek a candidate for a multivariate quantile function of a distribution function $F(.) : \mathbb{R}^d \to [0,1]$, generalizing the well-known univariate quantile function $F^{[-1]} : [0,1] \to \mathbb{R}$ which is characterized as the unique monotone map pushing the uniform probability measure du on $[0,1]$ to dF, namely a unique map $Q_F(.) : [0,1]^d \to \mathbb{R}^d$, monotone and pushing the uniform probability on $[0,1]^d$ to the probability measure dF on $\mathbb{R}^d$, we have McCann's theorem affirming the existence and uniqueness of a such candidate, we still need to obtain it constructively for applications.

The roads leading to them are as follows. First, we extend (MP) to (KP) to make sure that there are solutions for (KP) which came from solutions of (MP), i.e., of the form $\gamma_T = (I,T)\#du$. For the strict convex loss $c(x,y) = h(x-y)$ with $h(t) = \frac{t^2}{2}$ or t^2, it turns out that there exists uniquely (du-a.e.) an optimal γ_T for (KP), for which, the associated T is optimal for (MP). How to determine that T? (which will be our desired Q_F). We need results from duality. The unique optimal pair (φ, φ^*) of the dual problem, where φ^* is the c–transform of φ, is related to T as $T(x) = x - (\nabla h)^{-1}(\nabla \varphi(x))$ (which is the gradient of the convex function $x \to \frac{x^2}{2} - \varphi(x)$). Thus, $T(.)$ is determined once we can determine φ in the dual problem.

Example. Let $\mu = dF$ and $\nu = dG$ on $\mathbb{R}$, and $c(x,y) = (x-y)^2$. Then $\pi^* = (F^{[-1]}, G^{[-1]})\#du$, for du uniform on $[0,1]$, is optimal for (KP). Thus, for $dF = du$, we have $\pi^* = (I, G^{[-1]})\#du$, i.e., $\pi^* = \gamma_T = (I,T)\#du$, with $T = G^{[-1]}$ optimal for (MP).

As we have elaborated in Sect. 3, the concept of a "transport map" appeared already from the beginning of probability theory. Indeed, if X is a (real-valued) random variable, defined on a probability space $(\Omega, \mathscr{A}, P)$, then X acts like a map from Ω to the real line $\mathbb{R}$ (the observed values of X are "outcomes" or results from what happened in Ω), transporting the probability measure P on Ω to its law P_X on $\mathbb{R}$, in the sense that $P_X = PX^{-1}$, in symbol $X\#P = P_X$.

In fact we have a more concrete transport map which is the (univariate) quantile function $F^{[-1]}$ which is a map transporting the uniform probability measure du on $[0,1]$ to P_X (or dF) on $\mathbb{R}$, i.e., $F^{[-1]}\#du = dF$ (the polar factorization of X). In both settings, we have a map which preserves probabilities. In other words, $X : \Omega \to \mathbb{R}$, and $F^{[-1]} : [0,1] \to \mathbb{R}$ are measure-preserving maps.

What is *optimal* transport problem? In 1781 Gaspard Monge considered the following problem. Let $(\mathscr{X}, \mu), (\mathscr{Y}, \nu)$, with $\mathscr{X}, \mathscr{Y} \subseteq \mathbb{R}^d$, say, be two (Borel) probability spaces, and $c(.,.) : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}^+$ be a cost function (of transporting elements of $\mathscr{X}$ to elements of $\mathscr{Y}$). Find the best (optimal) preserving map T^* which transports μ (mass distribution) to ν, i.e., $T^*\#\mu = \nu$, in the sense of minimizing the total transport cost, i.e.,

$$T^* = \inf\{\int_{\mathscr{X}} c(x, T(x)) d\mu(x) : T : T\#\mu = \nu\}$$

In our analysis, we can take $\mathscr{X} = \mathscr{Y} = \mathbb{R}^d$, $\nu = dF$ the probability measure associated with the multivariate distribution function $F(.) : \mathbb{R}^d \to [0,1]$, and $\mu = du$, the

(non atomic) uniform probability measure on $[0,1]^d$ which is considered as on $\mathbb{R}^d$, as follows.

For $d=1$, the distribution of the variable U, uniformly distributed on $[0,1]$ has the distribution $F_U(.)$, and its associated probability measure $dF_U(-\infty,x] = F_U(x) = x$, for $x \in [0,1]$.

For $d>1$, if U is uniformly distributed on $[0,1]^d$, then its distribution function $F_U(.) : [0,1]^d \to [0,1]$ is

$$F_U(u_1,u_2,...,u_d) = \prod_{j=1}^{d} u_j \quad \text{for} \qquad (u_1,u_2,...,u_d) \in [0,1]^d$$

i.e., dF_U is the product measure with uniform marginals on $[0,1]$, a special d–copula.

We have seen an example, in Sect. 3, of this problem. Specifically, if $(\mathscr{X},\mu),(\mathscr{Y},\nu)$ are $([0,1],du)$ (a nonatomic probability measure with finite second moment), and $(\mathbb{R},dF)$ (an arbitrary probability measure), respectively, then the Monge's optimal transport map with respect to a convex loss function, e.g., $c(x,y) = (x-y)^2$, is $F^{[-1]}$.

Moreover, the optimal transport map $F^{[-1]} : [0,1] \to \mathbb{R}$ is an unique monotone non decreasing map, qualifying as the derivative of some convex function on $[0,1]$.

Let's reexamine this example again. We know in advance that the Monge's solution must be $F^{[-1]}$ since the quantile function $F^{[-1]}(x) = x+1$ (where $F(.)$ is the uniform distribution function on $[1,2]$) satisfies the two basic properties of a monotone optimal map, and in view of the uniqueness of such a map. But can we actually get that explicit optimal map without knowing the notion of univariate quantile functions? and "define" $F^{[-1]}$ as such?

Answering the above question opens the door for defining and determining multivariate quantile functions.

From Monge to Kantorovich. Since we are only interested in defining multivariate quantile functions, we will consider only two specific probability spaces $([0,1]^d,\mathscr{B}([0,1]^d),du)$ and $(\mathbb{R}^d,\mathscr{B}(\mathbb{R}^d),dF)$, where du is the uniform probability measure on $[0,1]^d$, with uniform distribution function F_U on $[0,1]^d$, and dF is the probability measure associated with the multivariate distribution function F on $\mathbb{R}^d$.

In terms of random variables, we refer to U as the uniform random vector with distribution F_U, and X as the random vector with distribution function F.

We will not need to consider the general theory of Optimal Transport (OT).

For dimension $d=1$, we have the explicit form of the univariate quantile function $F^{[-1]}$ (which is the "solution" of Monge's problem for convex loss functions), and, we will have existence and uniqueness of its counterpart in any dimension $d>1$, without evoking OT. However, for computations of Monge's solutions in higher dimensions, we need to address them in the setting of OT, in order to use linear programing in a dual formulation. Thus, we will mention a bit of OT which is beneficial in larger contexts.

The Monge's problem can be extended to a more general formulation (due to Kantorovich) as follows. Let $(\mathscr{X},\mu)$, $(\mathscr{Y},\nu)$ be Borel probability spaces with $\mathscr{X}$, $\mathscr{Y} \subseteq \mathbb{R}^d$.

Let $c(.,.) : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}^+$ be a cost function. Then the solution problem to the Monge's problem is an optimal transport map $T^*(.) : \mathscr{X} \to \mathscr{Y}$, *i.e.*,

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

In general, Monge's problem might not even have solutions. And when it does have solutions, it is not easy to compute them, because the objective function is not linear, and the constraint set is not convex.

The Kantorovich's reformulation avoids these difficulties.

Kantorovich's formulation is based on the idea of enlarging Monge's problem so that, first of all, it always has solutions. This is somewhat similar to the introduction of complex numbers, or more closely to von Neumann's mixed strategies in game theory (extending pure (determinist) strategies to random strategies).

Observe that, if T is a transport map, i.e., $T\#\mu = \nu$ ($\mu T^{-1}(.) = \nu(.)$), then, denoting by I the identity function on $\mathscr{X}$, $\gamma_T = (I,T)\#\mu$ is the probability measure on $\mathscr{X} \times \mathscr{Y}$ admitting μ, ν as marginal measures.

Remark. The map $(I,T) : \mathscr{X} \to \mathbb{R}^2$, is defined as $(I,T)(x) = (x,T(x)) \in \mathbb{R}^2$.

The joint measure $\gamma_T = (I,T)\#\mu$ is characterized as: for any $f(.) : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}^+$, we have

$$\int_{\mathscr{X}\times\mathscr{Y}} f(x,y)d\gamma_T(x,y) = \int_{\mathscr{X}} f(x,T(x))d\mu(x)$$

Proof. Use "standard argument of measure theory", starting out with f being an indicator function, i.e., $f(x,y) = 1_{A\times B}(x,y)$. Then we have

$$\int_{\mathscr{X}\times\mathscr{Y}} f(x,y)d\gamma_T(x,y) = \int_{\mathscr{X}\times\mathscr{Y}} 1_{A\times B}(x,y)d\gamma_T(x,y) = \int_{A\times B} d\gamma_T(x,y) =$$

$$d\mu(I,T)^{-1}(A\times B) = d\mu(A\cap T^{-1}(B)) = \int_{\mathscr{X}} 1_{A\times B}(x,T(x))d\mu(x)$$

Indeed,

$$(I,T)\#\mu(A\times\mathscr{Y}) = \mu(I,T)^{-1}(A\times\mathscr{Y}) = \mu\{x\in\mathscr{X} : (I,T)(x)\in A\times\mathscr{Y}\} =$$

$$\mu\{x\in\mathscr{X} : (x,,T(x))\in A\times\mathscr{Y}\} = \mu\{x\in\mathscr{X} : x\in A\} = \mu(A)$$

and

$$(I,T)\#\mu(\mathscr{X}\times B) = \mu(I,T)^{-1}(\mathscr{X}\times B) = \mu\{x\in\mathscr{X} : (I,T)(x)\in\mathscr{X}\times B\} =$$

$$\mu\{x\in\mathscr{X} : (x,,T(x))\in\mathscr{X}\times B\} = \mu\{x\in\mathscr{X} : x\in\mathscr{X}, T(x)\in B\} =$$

$$\mu\{x\in\mathscr{X} : x\in\mathscr{X}, T(x)\in B\} = \mu\{x : x\in T^{-1}(B)\} = \nu(B)$$

Thus, if $c(.,.) : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}^+$ be a given cost function, we have

$$V_c(\gamma_T) = \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)d\gamma_T(x,y) = \int_{\mathscr{X}} c(x,T(x)d\mu(x) = V_c(T)$$

where V_c denote the value of the transport plan γ_T and of the transport map T, with respect to c.

Thus, Monge's transport maps are special cases of transport "plans" (i.e., joint probability measures on $\mathscr{X} \times \mathscr{Y}$ having μ, ν as marginals).

Thus, if we denote by $\Pi(\mu,\nu)$ the set of joint probability measures on $\mathscr{X} \times \mathscr{Y}$ having μ, ν as marginals, then we enlarge the setting of Monge's problem (MP) in which $\Pi(\mu,\nu)$ is the solution set for the Kantorovich problem (KP):

$$\min\{\int_{\mathscr{X}\times\mathscr{Y}} c(x,y)d\pi(x,y) : \pi \in \Pi(\mu,\nu)\}$$

which always has solutions since $\Pi(\mu,\nu) \neq \varnothing$ (the product measure $\mu \otimes \nu \in \Pi(\mu,\nu)$).

Note that if γ_T is a solution for (KP), then T is solution for (MP).

For example, if we take $\mu = du$, and $\nu = dF$ in dimension 1 (i.e., on $\mathbb{R}$), then $\gamma = du \circ (I, F^{[-1]})^{-1} \in \Pi(\mu,\nu)$, noting that the identity I on $[0,1]$ is the quantile function of the uniform distribution.

Note that the dependence structure of the random variables U (uniform on $[0,1]$ with distribution function $F_U(u) = u$) and X (with distribution function F) in the polar factorization $X = F^{[-1]}(U)$ is that U and X are *comonotone*, i.e., they go up or down together (the subset $\{U(\omega), X(\omega)) : \omega \in \Omega\}$ is totally ordered in $\mathbb{R}^2$, which is the subset $\{(x,y) \in \mathbb{R}^2\}$ such that, for any, $(x,y),(x',y')$ in it, we have $<x-x',y-y'> \geq 0$). According the Sklar's theorem, U and X are comonotone if and only if the *copula of their dependence structure* is $\mathscr{C}(u,v) = u \wedge v$. This can be seen as follows. The joint measure of du, dF is $\gamma = du \circ (I, F^{[-1]})^{-1}$, so that the associated joint distribution function of (U,X) is

$$H(a,b) = P(U \leq a, X \leq b) = \gamma((-\infty,a] \times (-\infty,b])$$

then

$$H(a,b) = dH((-\infty,a] \times (-\infty,b]) = du(I, F^{[-1]})^{-1}((-\infty,a] \times (-\infty,b]) =$$

$$du\{u : u \leq a, F^{[-1]}(u) \leq b\} = du\{u \leq a, F(b) \geq u\} =$$

$$du\{u : u \leq a \wedge F(b)\} = a \wedge F(b) = F_U(a) \wedge F(b)$$

so that U and X are comonotone, or, by abuse of language, their joint probability measure $\gamma = du \circ (I, F^{[-1]})^{-1}$ is comonotone.

Therefore, if T is a monotone transport map then its corresponding transport plan $\gamma_T = du \circ (I,T)^{-1}$is comonotone.

Remark. In fact, the above can be extended to two variables, namely the joint measure $\gamma = (F_Y^{[-]}, F_X^{[-1]})\#du \in \Pi(dF_Y, dF_X)$ is comonotone.

Indeed, let's verify first that $\gamma = (F_Y^{[-1]}, F_X^{[-1]})\#du \in \Pi(dF_Y, dF_X)$. We have

$$\gamma(A \times \mathbb{R}) = du((F_Y^{[-1]}, F_X^{[-1]})^{-1}(A \times \mathbb{R}) =$$

$$du\{u \in [0,1] : F_Y^{[-1]}(u), \in A, F_X^{[-1]}(u) \in \mathbb{R}\} =$$

$$du\{u \in [0,1] : F_Y^{[-1]}(u), \in A\} = F_Y^{[-1]}\#du(A) = dF_Y(A)$$

Similarly,

$$\gamma(\mathbb{R} \times B) = dF_X(B)$$

Next, we have

$$H(a,b) = P(Y \leq a, X \leq b) = \gamma((-\infty, a] \times (-\infty, b]) =$$

$$du\{u \in [0,1] : F_Y^{[-1]}(u) \leq a, F_X^{[-1]}(u) \leq b\} =$$

$$du\{u \in [0,1] : F_Y(a) \geq u, F_X(b) \geq u\} =$$

$$du\{u \in [0,1] : u \leq F_Y(a) \wedge F_X(b)\} = F_Y(a) \wedge F_X(b)$$

Remark. The space $\Pi(du, dF)$ can be specified as follows. For $F(.)$ continuous, each $\gamma \in \Pi(du, dF)$ is indexed by a (binary) copula $\mathscr{C}$, say $\gamma_{\mathscr{C}}$, since the joint distribution function $H_{\mathscr{C}}$ of $\gamma_{\mathscr{C}}$, i.e., $\gamma_{\mathscr{C}} = dH_{\mathscr{C}}$, is $H_{\mathscr{C}}(u,x) = \mathscr{C}(u, F(x))$. Thus, each copula $\mathscr{C}$ determines a joint measure $\gamma_{\mathscr{C}} \in \Pi(du, dF)$. In particular, for $\mathscr{C}(u,v) = u \wedge v$, we get $\gamma_{\mathscr{C}} = du \circ (I, F^{[-1]})^{-1}$ (corresponding to the extremal copula). If $F(.)$ is not continuous (e.g., it's an empirical distribution) we must include sub-copulas.

For $\mu, \nu \in \mathscr{P}(\mathbb{R})$, set of Borel probability measures on $\mathbb{R}$, a joint measure $\gamma \in \Pi(\mu, \nu)$ is such that $\gamma(A \times \mathbb{R}) = \mu(A)$, $\gamma(\mathbb{R} \times B) = \nu(B)$. For example, let $\mathscr{C}$ be a bivariate copula, then $\gamma_{\mathscr{C}} \in \Pi(\mu, \nu)$ is determined by $\gamma_{\mathscr{C}} = dH_{\mathscr{C}}$, where $H_{\mathscr{C}}(x,y) = \mathscr{C}(F_\mu(x), F_\nu(y))$. For $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, with $d > 1$, this procedure will need the generalization of copulas to *vector copulas.*

The above characteristics of the univariate quantile function $F^{[-1]}$ is considered as its definition, i.e., let F be an arbitrary (univariate) distribution function, then its (univariate) quantile function is the unique monotone non decreasing optimal transport map between $([0,1], du)$ and $(\mathbb{R}, dF)$ with respect to a convex loss function.

What we have in mind is this. Let $(\mathscr{X}, \mu), (\mathscr{Y}, \nu)$, with $\mathscr{X}, \mathscr{Y} \subseteq \mathbb{R}^d$, $d > 1$, where $(\mathscr{X}, \mu) = ([0,1]^d, du)$, $(\mathscr{Y}, \nu) = (\mathbb{R}^d, dF)$, and $c(x,y) = ||x - y||^2$. If there exists an unique gradient $\nabla\varphi$ (of some convex function (not unique) $\varphi : [0,1]^d \to \mathbb{R}$) which is

the optimal transport, then $\nabla\varphi$ is defined as the multivariate quantile function of F, noting that the gradient $\nabla\varphi$ is monotone non decreasing as the generalization of the same concept in one dimension, i.e., for any $x, y \in \mathbb{R}^d$, we have

$$<\nabla\varphi(x) - \nabla\varphi(y), x - y> \quad \geq 0$$

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

Of course such a result is only an existence result. We need to find ways to compute it, at least for applications!

Notes on Convex Functions. A function $f(.) : \mathbb{R}^d \to \mathbb{R}$ is said to be convex it for any $x, y \in \mathbb{R}^d$ and $t \in [0,1]$ we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

it is strictly convex if the above inequality is strict.

A convex function is a..e. differentiable. In dimension 1, the graph of a convex function lies above any tangent to it, and hence its derivative is monotone non decreasing. For $d > 1$, the whole graph of $f(.)$ lies above its tangent hyperplane at any xc where it is differentiable, so as a consequence, it gradient is monotone in the above sense.

The Kantorovich's reformulation (of Monge's problem) is this. Find an optimal transport plan, i.e., a joint measure $\pi^* \in \Pi(\mu, \nu)$ such that

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

Now the problem seems solvable since the objective function is linear in π, and the constraint set is convex.

Note that, although, as far as quantile functions are concerned, we are interested in transport maps (not necessary transport plans), we still need to evoke Kantorovich's formulation in order to compute multivariate quantile functions.

Duality. In order to solve the "inf" problem of (KP) we will transform it into a "sup" problem (this procedure is referred to as duality, where the "inf" is the primal problem, and the "sup" is dual problem) where the constraints in the "sup" problem can be expressed as inequalities (In infinite dimensions). The relations between the primal and dual problems will allow us to get solution for the primal problem from the dual problem.

For $\pi \in \Pi(\mu, \nu)$, let

$$V(\pi) = \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)d\pi(x,y)$$

we are going to relate the (KP) $P = \inf\{V(\pi) : \pi \in \Pi(\mu,\nu)\}$ to a "sup" problem. For that, first observe that, for suitable function φ, ψ defined on $\mathscr{X}, \mathscr{Y}$, respectively, we have

$$\int_{\mathscr{X}\times\mathscr{Y}} [\varphi(x) + \psi(y)]d\pi(x,y) = \int_{\mathscr{X}} \varphi(x)d\mu(x) + \int_{\mathscr{Y}} \psi(y)d\nu(y)$$

Proof. Use "standard argument of measure theory"!

Thus, for φ, ψ such that $\varphi(x) + \psi(y) \leq c(x,y)$, for all x,y we have

$$\int_{\mathscr{X}} \varphi(x) d\mu(x) + \int_{\mathscr{Y}} \psi(y) d\nu(y) \leq \int_{\mathscr{X}\times\mathscr{Y}} c(x,y) d\pi(x,y)$$

Consider

$$D = \sup\{\int_{\mathscr{X}} \varphi(x) d\mu(x) + \int_{\mathscr{Y}} \psi(y) d\nu(y) : (\varphi, \psi) : \varphi(,) + \psi(,) \leq c(.,.)\}$$

then clearly $D \leq P$. In fact, $D = P$ which is our desired duality. The dual formulation has a linear objective function with inequality constraints (suitable for linear programming).

The Kantorovich duality is this (1942). Let

$$J(\varphi, \psi) = \{\int_{\mathscr{X}} \varphi(x) d\mu(x) + \int_{\mathscr{Y}} \psi(y) d\nu(y)\}$$

Then

$$P = \inf\{V(\pi) : \pi \in \Pi(\mu, \nu)\} = \sup\{J(\varphi, \psi) : \varphi(.) + \psi(.) \leq c(.,.)\} = D$$

The sup on the right hand side is attained.

We study the duality in the case of quadratic cost $c(x,y) = \frac{1}{2}||x-y||^2$, when μ and ν have finite second moments, i.e.,

$$\int_{\mathbb{R}^d} ||x||^2 d\mu(x) < \infty, \qquad \int_{\mathbb{R}^d} ||y||^2 d\nu(x) < \infty$$

so that

$$V(\pi) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{||x-y||^2}{2} d\pi(x,y) \leq \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{||x||^2 + ||y||^2}{2} d\pi(x,y) < \infty$$

From the duality

$$\inf\{V(\pi) : \pi \in \Pi(\mu, \nu)\} = \sup\{J(\varphi, \psi) : \varphi(.) + \psi(.) \leq c(.,.)\}$$

we get, for the (KP) primal problem: The left hand side admits a minimizer, i.e., there exists $\pi^* \in \Pi(\mu, \nu)$ such that

$$V(\pi^*) = \inf\{V(\pi) : \pi \in \Pi(\mu, \nu)\}$$

As for the dual problem, here $\varphi(x) + \psi(y) \leq c(x,y)$ means

$$\varphi(x) + \psi(y) \leq \frac{||x-y||^2}{2} = \frac{||x||^2}{2} + \frac{||y||^2}{2} - <x,y>$$

from it, we have

$$<x,y> \leq [\frac{||x||^2}{2} - \varphi(x)] + [\frac{||y||^2}{2} - \psi(x)]$$

Let

$$\widetilde{\varphi}(x) = \frac{||x||^2}{2} - \varphi(x), \quad \widetilde{\psi}(x) = \frac{||x||^2}{2} - \psi(x)$$

and

$$M = \int_{\mathbb{R}^d} ||x||^2/2 d\mu(x) < \infty + \int_{\mathbb{R}^d} ||y||^2/2 d\nu(x) < \infty$$

we have

$$\inf\{V(\pi) : \pi \in \Pi(\mu,\nu)\} = M - \sup\{\int_{\mathbb{R}^d \times \mathbb{R}^d} <x,y> d\pi(x,y) : \pi \in \Pi(\mu,\nu)\}$$

and

$$\sup\{J(\varphi,\psi) : \varphi(.) + \psi(.) \leq c(.,.)\} = M - \inf\{J(\varphi,\psi) : (\varphi,\psi) \in \widetilde{\Theta}\}$$

where

$$\widetilde{\Theta} = \{(\varphi,\psi) : \nabla x, y : \varphi(x) + \psi(y) \geq <x,y>\}$$

and

$$<x,y> \leq [\frac{||x||^2}{2} - \varphi(x)] + [\frac{||y||^2}{2} - \psi(x)]$$

becomes

$$\sup\{\int_{\mathbb{R}^d \times \mathbb{R}^d} <x,y> d\pi(x,y) : \pi \in \Pi(\mu,\nu)\} = \inf\{J(\varphi,\psi) : (\varphi,\psi) \in \widetilde{\Theta}\}$$

Let $\widetilde{\varphi}(x) = [\frac{||x||^2}{2} - \varphi(x)]$, and $\widetilde{\psi}(y) = [\frac{||y||^2}{2} - \psi(y)]$, we have the constraint $\widetilde{\varphi}(x) + \widetilde{\psi}(y) \geq <x,y>$.

For simplicity, we just drop the symbol $\sim$ on the functions φ, ψ from our writing (but not from our mind).

Thus, from $\varphi(x) + \psi(y) \geq <x,y>$, we have

$$\psi(y) \geq <x,y> - \varphi(x) \Longrightarrow \psi(y) \geq \sup_x [<x,y> - \varphi(x)] = \varphi^*(y)$$

so that

$$J(\varphi,\psi) \geq J(\varphi,\varphi^*)$$

We call (φ, φ^*) a potential pair. Note that, from

$$\varphi^*(y) = \sup_x [<x,y> - \varphi(x)]$$

it follows that $\varphi(x) + \varphi^*(y) \geq <x,y>$, i.e., each potential pair $(\varphi, \varphi^*) \in \widetilde{\Theta}$, the constraint set of $J(\varphi, \psi)$.

In fact, we have

Theorem. If μ, ν are (Borel) probability measures on $\mathbb{R}^d$, with finite second moments, then, with respect to the cost function $c(x,y) = \frac{1}{2}||x-y||^2$,

(i) There exists an potential pair (φ, φ^*), convex conjugate, minimizing $J(\varphi, \psi)$ on $\widetilde{\Theta}$,
(ii) If, in addition, μ ia nonatomic, there exists a unique optimal $\pi^* \in \Pi(\mu, \nu)$ of the form $\pi^* = \gamma_{T^*} = (I_{\mathscr{X}}, T^*)\#\mu$, with $T^* = \nabla\varphi$ unique.

Comments. T^* is the unique minimizer of (MP), with the strict convex loss, which is the gradient of a convex function (hence monotone non decreasing). The optimal potential pair (φ, φ^*) is obtained from the dual Kantorovich problem.

Thus, for μ being the uniform probability du on $\mathscr{X} = [0,1]^d$, the (Brenier) map T^* (pushing du to $dF = \nu$) is our *multivariate quantile function* of the multivariate distribution function $F(.)$ on $\mathbb{R}^d$.

The convex function φ is not unique, but the gradient $\nabla\varphi$ is unique (μ – a.e.).

Note also that $\nu = dF$ could be arbitrary, i.e., having finite second moment or not, in view of McCann's theorem.

Important: As you have said, we drop the symbol $\widetilde{\varphi}$ for simplicity, the φ in the theorem is really $\widetilde{\varphi(x)} = \frac{||x||^2}{2} - \varphi(x)$, i.e.,it is the pair

$$(\frac{||x||^2}{2} - \varphi(x), \frac{||y||^2}{2} - \varphi^*(y))$$

which solves the Monge-Kantorovich problem, and φ is "c–concave" in the sense that it is the function $x \to \frac{||x||^2}{2} - \varphi(x)$ that is convex. Thus, the explicit formula for T^* is

$$T^*(x) = x - \nabla\varphi$$

which is the gradient of the convex function $\frac{||x||^2}{2} - \varphi(x)$. If the cost is $c(x,y) = h(x-y)$ with $h(.) : \mathbb{R} \to \mathbb{R}$ strictly convex, then $T^*(x) = x - (h')^{-1}(\nabla\varphi(x))$.

Some Examples

(1) Let μ, ν be probability measures of $\mathbb{R}$, where μ is uniform du on $[0,1]$, and ν is uniform dv on $[1,2]$. Then μ is nonatomic, and both μ, ν have finite second moments.Let the cost function be $c(x,y) = (x-y)^2$. We are going to verify that the univariate quantile function $F_\nu^{[-1]}(.) : [0,1] \to \mathbb{R}$ of $F_\nu(.) : \mathbb{R} \to [0,1]$ is indeed the unique monotone transport map solving the Monge's problem (i.e., it is the optimal transport map).

We have $F_\nu^{[-1]}(v) = 1+v$. It's monotone non decreasing. It is a transport map pushing du on $[0,1]$ to dv on $[1,2]$. Indeed, let $F_\nu^{[-1]}(.) = T(.)$, and $a \in [1,2]$,

$$(F_\nu^{[-1]})\#du((-\infty,a]) = du \circ T^{-1}((-\infty,a]) = du\{u : T(u) \le a\} =$$

$$du\{u : 1+u \le a\} = du\{u \le a-1\} = a-1 = \nu(-\infty,a])$$

From theory, we know that such a map $F_\nu^{[-1]}(v) = 1+v$ with the above two characteristic properties is the unique solution of the Monge's problem with respect to the given quadratic loss function, i.e.,

$$T = \arg\min\{\int_0^1 (x-S(x))^2 dx : S\#du = d\nu\}$$

so let's verify it. We let $c(x-y)^2 = h(x-y)$ where $h(.) : \mathbb{R} \to \mathbb{R}$, being $h(t) = t^2$, a convex function.

For any transport map $S(.) : [0,1] \to [1,2]$, we have

$$M(S) = \int_0^1 (x-S(x))^2 dx = \int_0^1 h(x-S(x))dx$$

Since $h(.)$ is convex, we have, by Jensen's inequality ($h(EX) \le E(h(X))$,

$$M(S) = \int_0^1 h(x-S(x))dx \ge h[\int_0^1 (x-S(x))dx] =$$

$$h[\int_0^1 xdx - \int_0^1 S(x))dx] = h[\int_0^1 xdx - \int_1^2 ydy] = h(\frac{1}{2} - \frac{3}{2}) = h(-1) = 1 =$$

$$\int_0^1 (T(x)-x)^2 dx = M(T)$$

since $T(x) = x+1$. Thus, for any S such that $S\#\mu = \nu$, $M(S) \ge M(T)$.

Notes. In the above calculations, since $S\#du = dv$, we have $\int_0^1 S(u))du = \int_1^2 vdv$.

Thus, $T(x) = 1+x = F_\nu^{[-1]}(x)$ in the above example is optimal for the Monge's problem with quadratic loss: $M(T) = \min\{M(S) : S\#\mu = \nu\}$.

The optimal transport map $T(x) = 1+x = F_\nu^{[-1]}(x)$ is unique by Brenier's theorem, since it is a monotone and optimal!. It is the derivative of the convex function $g(.) : [0,1] \to \mathbb{R},\ g(x) = \frac{1}{2}(1+x)^2$.

In general, Brenier's theorem affirms that the unique monotone and optional transport map T, with respect to the strictly convex $h(.)$, is of the form

$$T(x) = x - (h')^{-1}(\nabla\varphi)$$

for an optimal potential pair (φ,ψ) of the Kantorovich dual problem, noting that $\frac{1}{2}h^2(x) - \varphi(x)$ is the convex function such that $T = \nabla(\frac{1}{2}h^2(x) - \varphi(x))$.

In our example, with $\varphi(x) = -2x$,

$$h(x) = x^2 \Longrightarrow h'(x) = 2x \Longrightarrow (h')^{-1}(x) = \tfrac{x}{2} \Longrightarrow$$

$$T(x) = x - (h')^{-1}(\nabla\varphi) = x - (-2)/2 = x + 1$$

(2) As for optimality in the Kantorovich formulation, here is a simple example.

Let $\mu = dF$, $\nu = dG$ be two probability measures on $\mathbb{R}$, then the transport map (joint measure with μ, ν as marginals) $\pi^* = dH$, where $H(x,y)$ is the bivariate distribution function $H(x,y) = H(x) \wedge G(y)$ is the optimal joint measure, i.e., with $c(x,y) = (x-y)^2$,

$$\pi^* = \arg\min\{\int_{\mathbb{R}^2} c(x,y)d\pi(x,y) : \pi \in \Pi(\mu,\nu)\}$$

First, let verify that $\pi^* = dH$ is indeed in $\Pi(\mu,\nu)$. We have

$$dH((-\infty,a] \times (-\infty,b]) = H(a,b) = F(a) \wedge G(b)$$

It follows that

$$dH((-\infty,a] \times \mathbb{R}) = F(a) \wedge G(\infty) = F(a) = dF((-\infty,a])$$

Similarly, $dH(\mathbb{R} \times (-\infty,b])) = dG((-\infty,b])$.

Note that, in fact, using copula, it is obvious that $H(x,y) = H(x) \wedge G(y)$ is a bona fide bivariate distribution function on $\mathbb{R}^2$, with marginal distribution functions F and G, since $\mathscr{C}(u,v) = u \wedge v$ is a copula!

Also, in fact, we have $\pi^* = du \circ (F^{[-1]}, G^{[-1]})^{-1}$, i.e., $\pi^* = (F^{[-1]}, G^{[-1]})\#du$, where du is uniform on $[0,1]$. Indeed,

$$du \circ (F^{[-1]}, G^{[-1]})^{-1}((-\infty,a] \times (-\infty,b]) = du\{u : F^{[-1]}(u) \leq a, G^{[-1]}(u) \leq b\} =$$

$$du\{u : u \leq F(a), u \leq G(b)\} = F(a) \wedge G(b) = H(a,b)$$

As such, we have

$$K(\pi^*) = \int_{\mathbb{R}\times\mathbb{R}} c(x,y)d\pi^*(x,y) = \int_0^1 (F^{[-1]}(u) - G^{[-1]}(u))^2 du$$

since, in general, when $\pi^* = du \circ (F^{[-1]}, G^{[-1]})^{-1}$, we have for any function $\varsigma(x,y)$,

$$\int_{\mathbb{R}^2} \varsigma(x,y)d\pi^*(x,y) = \int_0^1 \varsigma(F^{[-1]}(u), G^{[-1]}(u))du = \inf\{K(\pi) : \pi \in \Pi(\mu,\nu)\}$$

The quantity

$$W_2(\mu,\nu) = [\inf\{K(\pi) : \pi \in \Pi(\mu,\nu)\}]^{\frac{1}{2}} = [\int_0^1 (F^{[-1]}(u) - G^{[-1]}(u))^2 du]^{\frac{1}{2}}$$

is a *Wasserstein distance* between μ and ν.

(3) Let F and G be two distribution functions on $\mathbb{R}$, and let $H(x,y) = F(x)G(y)$.

What will be the bivariate quantile function $Q_H(.,.) : [0,1]^2 \to \mathbb{R}^2$ of H, i.e., monotone and $Q_H \# du = dH$, where du is uniform on $[0,1]^2$.

5 Notes on Multivariate Quantile Regression

Like a blessing, one of the inventors of univariate quantile regression, Roger Koenker wrote in his recent paper "Quantile Regression 40 years on" the following about multivariate quantiles:

"...Despite generating an extensive literature, it is fair to say that no general agreement has emerge... in contrast to the sample mean of d-dimensional vectors, there is no consensus about an appropriate notion of multivariate median. In an exciting new development, Carlier, Chernozhukov and Galichon [2] have proposed a vector quantile regression notion motivated by classical Monge-Kantorovich optimal transport theory".

Armed with the notion of multivariate quantile functions, we elaborate first on its application to regression.

Recall that the notion of unconditional (multivariate) quantile functions is this. Let Y be a random vector with values in $\mathbb{R}^d$. By Lebesgue-Stieltjes ' theorem, the law of Y is the Borel probability measure ν on $\mathscr{B}(\mathbb{R}^d)$ derived from its distribution function $F(.) : \mathbb{R}^d \to [0,1]$ as

$$\nu((-\infty, y]) = dF((-\infty, y]) = F(y)$$

Note that, only when needed that we will call upon the "background" setting: the random vector Y is "defined" on a probability space $(\Omega, \mathscr{A}, P)$, so that $\nu = PY^{-1}$, and $F(y) = P(\omega \in \Omega : Y(\omega) \leq y\}$. In our analysis, the polar factorization $Y = Q_F(U)$ is more "concrete" to use, where U is a random vector uniformly distributed on the unit cube $[0,1]^d$, with probability measure denoted as du, and Q_F denotes the (multivariate) quantile function.

The quantile function $Q_F(.) : [0,1]^d \to \mathbb{R}^d$ is the (a.e.) unique monotone (non decreasing) map, such that $Q_F \# du = dF$, or equivalently, $dF(.) = du \circ Q_F^{-1}(.)$.

In multivariate regression analysis, besides our "target" random vector Y, we have another random vector X, taking values in $\mathbb{R}^k$, and playing the role of covariates (or regressors) of Y. As "usual", we wish to establish a statistical model relating Y to its covariates X.

As far as (linear) quantile regression is concerned, the main analysis tool is conditional (multivariate) quantile functions.

It should be noted that the computational aspects in multivariate quantile regression are expected to be much more complicated than the univariate case. Not only the OT framework allowed us to generalize appropriately the univariate case to general case, it provides us with computational methods as well.

As such, to appreciate how OT can help, let's reformulate univariate conditional quantile analysis in the language of OT.

Let Y be a real-valued random variable with distribution $F_Y(.)$. We keep, in our mind, the abstract setting: Y is defined on $(\Omega, \mathscr{A}, P)$, but focus on its concrete polar factorization $Y = F_Y^{[-1]}(U)$.

First, recall that the univariate quantile function $F_Y^{[-1]}(.) : [0,1] \to \mathbb{R}$ is the *pseudo-inverse* of the distribution function F_Y (since, in general, F_Y is only monotone non decreasing, and right continuous, so that it does not have an inverse) defined as

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

This is well-defined since $\geq$ is a total order relation on $\mathbb{R}$. In fact, the infimum is attained, i.e., the infimum is a minimum.

Some useful properties of $F_Y^{[-1]}(.)$ are as follows.

a) $F_Y^{[-1]}(.)$ is *monotone non decreasing*, i.e., $u \leq v \Longrightarrow F_Y^{[-1]}(u) \leq F_Y^{[-1]}(u)$ (Note that, strictly increasing means, $u < v \Longrightarrow F_Y^{[-1]}(u) < F_Y^{[-1]}(u)$).

Proof. For $u \leq v$, we have $\{y \in \mathbb{R} : F_Y(y) \geq v\} \subseteq \{y \in \mathbb{R} : F_Y(y) \geq u\}$ and hence $\inf\{y \in \mathbb{R} : F_Y(y) \geq v\} \geq \inf\{y \in \mathbb{R} : F_Y(y) \geq u\}$.

b) $\inf\{y \in \mathbb{R} : F_Y(y) \geq u\}$ is attained.

Proof. "$\inf\{y \in \mathbb{R} : F_Y(y) \geq u\}$ is attained" means $F_Y^{[-1]}(u)$ is one of the y such that $F_Y(y) \geq u$, i.e., $F_Y(F_Y^{[-1]}(u)) \geq u$.

For each $n \geq 1$, by the definition of infimum, there exists $y_n \in \mathbb{R}$ such that $F_Y(y_n) \geq u$ and $y_n \leq F_Y^{[-1]}(u) + \frac{1}{n}$.

Since $F_Y(.)$ is nondecreasing, we then have

$$u \leq F_Y(y_n) \leq F_Y(F_Y^{[-1]}(u) + \frac{1}{n})$$

Next, by right continuity of F_Y, we have

$$\lim_{n \to \infty} F_Y(F_Y^{[-1]}(u) + \frac{1}{n}) = F_Y(F_Y^{[-1]}(u))$$

so that $F_Y(F_Y^{[-1]}(u)) \geq u$, since for each n, $u \leq F_Y(F_Y^{[-1]}(u) + \frac{1}{n})$.

c) *A weak representation of Y* is this. For any random variable U distributed uniformly on $[0,1]$, Y has the same distribution as $F_Y^{[-1]}(U)$ (written as $Y \stackrel{D}{=} F_Y^{[-1]}(U)$).

Proof. It suffices to show that

$$F_Y^{[-1]}(u) \leq y \Longleftrightarrow u \leq F_Y(y)$$

since then

$$P(F_Y^{[-1]}(U) \leq y) = P(U \leq F_Y(y)) = F_Y(y)$$

Thus, let's show the above equivalence.

If $\omega \in \{\omega \in \Omega : U(\omega) \leq F_Y(y)\}$, i.e., $U(\omega) \leq F_Y(y)$, then, by definition of $F_Y^{[-1]}(.)$, $F_Y^{[-1]}(U(\omega)) \leq y$, and hence

$$\{\omega \in \Omega : U(\omega) \leq F_Y(y)\} \subseteq \{\omega : F_Y^{[-1]}(U(\omega)) \leq y\}$$

Conversely, if $\omega \in \{\omega : F_Y^{[-1]}(U(\omega)) \leq y\}$, i.e., $F_Y^{[-1]}(U(\omega)) \leq y$, then $F_Y(y+\varepsilon) \geq U(\omega)$ for all $\varepsilon > 0$, and hence $F_Y(y) \geq U(\omega)$ by right continuity of F_Y, so that

$$\{\omega : F_Y^{[-1]}(U(\omega)) \leq y\} \subseteq \{\omega \in \Omega : U(\omega) \leq F_Y(y)\}$$

therefore equality.

Remark. (i) By taking set complement, we also have

$$F_Y^{[-1]}(u) > y \Longleftrightarrow u > F_Y(y)$$

(ii) The representation is weak since the equality between Y and $F_Y^{[-1]}(U)$ is "in distribution" which is weaker than "almost sure equality", noting that if $Y \overset{a.s.}{=} F_Y^{[-1]}(U)$ (a "strong representation", called the polar factorization of Y) then $Y \overset{D}{=} F_Y^{[-1]}(U)$.

d) *A strong representation of Y.* Every time we have a uniform random variable V on $[0,1]$, Y and $F_Y^{[-1]}(V)$ have the same distribution. If we look at the joint distribution $\pi_{(Y,V)}$ of (Y,V), then we see differences among these variables V although they all have the same uniform distribution du on $[0,1]$. Indeed, according to Sklar's theorem, the joint distribution function $H(y,v)$ of the random vector (Y,V) is of the form $c(F_Y,F_V)$ where c is a (bivariate) copula. Thus, each V is in fact determined by its own copula c, in other words, these V are indexed by copulas. While they are all in $\Pi(F_Y,du)$, the set of all joint distributions with the same marginals F_Y,F_U, there are different by their associated copulas. Thus, saying that there is a U with distribution du, such that $F_Y^{[-1]}(U) \overset{a.s.}{=} Y$, we mean a special V, or rather, a special copula c^* of (Y,V) such that we actually have $F_Y^{[-1]}(U) = Y$.

Let's elaborate a bit more on "Strong representation", i.e., equality between random variables in the "almost surely" (with probability one) sense.

The question is: is there a random variable V^* distributed as U, i.e., uniformly on $[0,1]$ such that $F_Y^{[-1]}(V^*) \overset{a.s.}{=} Y$?

The answer is affirmative. Its proof will shed light on how actually to "construct" such a random variable.

Proof. Let F_Y be the distribution function of Y.

(i) If F_Y is strictly increasing, then F_Y and $F_Y^{[-1]}$ are bijections with $F_Y = (F_Y^{[-1]})^{-1}$. Define $V^*(\omega) = F_Y(Y(\omega))$, then $F_Y^{[-1]}(V^*(\omega)) = Y(\omega)$, and V^* is uniform on $[0,1]$ since

$$P(\omega : V^*(\omega) \leq u) = P(\omega : F_Y(Y(\omega)) \leq u) =$$

$$P(\omega : Y(\omega) \leq F_Y^{[-1]}(u)) = F_Y(F_Y^{[-1]}(u)) = u$$

(ii) If F_Y is not strictly increasing, the announced V^* is constructed as follows.

Let $A_Y = \{y \in \mathbb{R} : P(\omega : Y(\omega) = y) > 0\} \neq \varnothing$.

For any $y \in A_Y$, define a uniform random variable V_y on $\{u \in [0,1] : F_Y^{[-1]}(u) = y\}$. Then define

$$V^*(\omega) = F_Y(Y(\omega))1_{(Y(\omega)\notin A_Y)} + V_{Y(\omega)}1_{(Y(\omega)\in A_Y)}$$

Then V^* is distributed uniformly on $[0,1]$, and $F_Y^{[-1]}(V^*) \overset{a.s.}{=} Y$.

Next, you may ask: What does it mean by, say, minimizing an objective function over a collection of random variables? i.e., the solution of the optimization problem is a random variable?

Well, remember how mean linear regression was originated? When predicting a random variable Y from a covariate X, using mean squared error, we seek the best random variable built from X, i.e., minimizing the objective function $E(Y - \varphi(X))^2$ over all random variables Z of the form $\varphi(X)$, i.e., a function of X. And, of course, the solution is the special random variable $E(Y|X)$.

e) For any distribution function F_Y on $\mathbb{R}$, $F_Y \circ F_Y^{[-1]}(u) \geq u$, for any $u \in [0,1]$,

If F_Y is continuous then $F_Y \circ F_Y^{[-1]}(.) =$ Identity on $[0,1]$, and $F_Y(Y)$ is distributed uniformly on $[0,1]$,

F_Y is continuous if and only if $F_Y^{[-1]}$ is strictly increasing; F_Y is strictly increasing if and only if $F_Y^{[-1]}$ is continuous,

If F_Y is continuous and strictly increasing then $F_Y^{[-1]}$ is the inverse of F_Y : $(F_Y^{[-1]})^{-1} = F_Y$.

A quick recap of univariate quantile regression.

Let X be a real-valued random variable with distribution function F. Unlike moments, quantiles exist for any distributions (heavy-tailed or not). Quantiles are used to define financial risk measures, such as Value-At-Risk which is $F^{[-1]}(\alpha)$, ($P(X > F^{[-1]}(\alpha)) = 1 - \alpha$), and in Linear Quantile regression models.

The α – quantile $q_\alpha(F)$ minimizes the objective function

$$a \to E\rho_\alpha(Y-a) = \int_{\mathbb{R}} \rho_\alpha(y-a)dF(y) = \int_{\mathbb{R}} (y-a)[\alpha - 1_{(-\infty,a)}(y)]dF(y)$$

where

$$\rho_\alpha(u) = u[\alpha - 1_{(u\leq 0)}] = \begin{cases} -(1-\alpha)u & \text{for} \quad u < 0 \\ \alpha u & \text{for} \quad u \geq 0 \end{cases}$$

i.e.,

$$q_\alpha(F) = \arg\min_a E\rho_\alpha(Y-a)$$

and hence its sample $\alpha-$ quantile $q_\alpha(F_n)$ is

$$\arg\min_a \sum_{i=1}^{n} \rho_\alpha(Y_i - a)$$

leading to the following plausible conditional quantile estimator. Since the conditional $\alpha-$quantile $q_\alpha(Y|X)$ minimizes the LAD loss, i.e., minimizing $E\rho_\alpha(Y-\varphi(X))$ *over all possible* $\varphi(X)$, if we specify $q_\alpha(Y|X)$ linearly, i.e., $q_\alpha(Y|X)=X\theta(\alpha)$, then the coefficient $\theta(\alpha)$ could be estimated by the extremum estimator $\hat{q_\alpha}(Y|X)=\hat{\theta}(\alpha)$ which is

$$\arg\min_\theta \sum_{i=1}^n \rho_\alpha(Y_i-X_i\theta)=\arg\min_\theta \sum_{i=1}^n (Y_i-X_i\theta)[\alpha-1_{(Y_i-X_i\theta<0)}]$$

for data $(X_i,Y_i), i=1,2,...,n$, drawn from (X,Y).

What is important is this. The quantile function $F^{[-1]}(.)$ (which is left continuous) satisfies the following: For $a>0$ and $b\in\mathbb{R}$,

$$q_\alpha(aX+b)=aq_\alpha(X)+b$$

i.e.,

$$F^{[-1]}_{aX+b}(\alpha)=aF^{[-1]}_X(\alpha)+b$$

That is, $F^{[-1]}(\alpha)$ is *affine equivariant* (the transformation $x\longrightarrow ax+b$ is an affine transformation): the quantile representation of a point after affine transformation agrees with its original quantile representation similarly transformed.

This invariance properly is essential to use the quantile regression

$$Y=\beta_\alpha X+\varepsilon_\alpha$$

since, given X, suppose we model $q_\alpha(Y|X)=\beta_\alpha X$, then we have

$$q_\alpha(Y|X)=q_\alpha(\beta_\alpha X+\varepsilon_\alpha)=\beta_\alpha X+q_\alpha(\varepsilon_\alpha|X)=\beta_\alpha X$$

when we impose the condition $q_\alpha(\varepsilon_\alpha|X)=0$. In other words,

$$Y=\beta_\alpha X+\varepsilon_\alpha ...\text{with } q_\alpha(\varepsilon_\alpha|X)=0$$

is equivalent to $q_\alpha(Y|X)=\beta_\alpha X$.

Note, however, that unlike the mean, in general, $q_\alpha(X+Y)\neq q_\alpha(X)+q_\alpha(Y)$

For $\alpha=\frac{1}{2}$, the median $F^{-1}(\frac{1}{2})$ minimizes $E|X-a|$ over $a\in\mathbb{R}$. How about other $\alpha\in(0,1)$?

Remark. We need to figure out an objective function for $F^{[-1]}(\alpha)$ to minimize also to suggest an *extremum estimator* for it.

Now, observe that the median minimizes also the objective function (risk) $\frac{1}{2}E|X-a|$ whose loss function is

$$\frac{1}{2}|x-a|=\begin{cases}-\frac{1}{2}(x-a) \text{ for } (x-a)<0\\ \frac{1}{2}(x-a) \text{ for } (x-a)\geq 0\end{cases}$$

or

$$\frac{1}{2}|x-a| = (x-a)[\frac{1}{2} - 1_{(x-a<0)}]$$

This observation leads to other loss functions generalizing

$$(x-a)[\frac{1}{2} - 1_{(x-a<0)}] = \rho_{\frac{1}{2}}(x-a) = \frac{1}{2}|x-a|$$

where $\rho_{\frac{1}{2}}(u) = \frac{|u|}{2}$, by replacing $\frac{1}{2}$ by an arbitrary $\alpha \in (0,1)$ in $(x-a)[\frac{1}{2} - 1_{(x-a<0)}]$, namely

$$L_\alpha(x,a) = \rho_\alpha(x-a) = (x-a)[\alpha - 1_{(x-a)<0)}]$$

Note that $\rho_\alpha(u) = u[\alpha - 1_{(u<0)}]$ is a nonnegative function.

Theorem. The α–quantile of X minimizes $E\rho_\alpha(X-a)$ over $a \in \mathbb{R}$.

Proof. As a function of a, the objective (associated risk) function

$$E\rho_\alpha(X-a) = \alpha[EX-a] - \int_{-\infty}^{a}(x-a)dF(x) =$$

$$\alpha[EX-a] - \int_{-\infty}^{a} xdF(x) + a\int_{-\infty}^{a} dF(x)$$

is differentiable with (assuming for simplicity that F is absolutely continuous)

$$\frac{d(E\rho_\alpha(X-a))}{da} = -\alpha - a\frac{dF}{dx}(a) + a\frac{dF}{dx}(a) + \int_{-\infty}^{a}\frac{dF}{dx}(x)dx = F(a) - \alpha$$

Since $F(.)$ is nondecreasing, the function $a \to F(a) - \alpha$ is increasing, so that the function $a \to E\rho_\alpha(X-a)$ is convex. As such, the first order condition

$$\frac{d(E\rho_\alpha(X-a))}{da} = F(a) - \alpha = 0$$

implies that the minimum of $E\rho_\alpha(X-a)$ over a is attained at $F(a) = \alpha$, i.e., $a = F^{-1}(\alpha)$, the $\alpha-$quantile of F. In other words, the $\alpha-$quantile $F^{-1}(\alpha)$ minimizes the risk $E\rho_\alpha(X-a)$ over a.

Remark. Thus, since $F^{[-1]}(\alpha)$ minimizes $E\rho_\alpha(X-a)$, its empirical counterpart (sample quantile), namely $\hat{a_n} = \inf\{x \in \mathbb{R} : F_n^{[-1]}(x) \geq \alpha\}$, minimizes

$$\int_{\mathbb{R}} \rho_\alpha(x-a)dF_n(x) = \frac{1}{n}\sum_{i=1}^{n}\rho_\alpha(X_i - a)$$

Note that, unlike moments, quantiles existent for any kind of distributions including heavy-tailed ones. Note also that, unlike $(x-a)^2$, the function $a \to \rho_\alpha(X-a)$ is not differentiable at any $a \in \mathbb{R}$. However, it is continuous and convex.

A similar result for conditional quantiles is this. First, the conditional distribution of Y given $X = x$ is $F_{Y|X=x}(y|x) = P(Y \leq y|X = x) = E[1_{(Y\leq y)}|X = x)$. Its α–quantile is simply

$$q_{Y|X}(\alpha) = F^{[-1]}_{Y|X}(\alpha) = \inf\{x \in \mathbb{R} : F_{Y|X}(x) \geq \alpha\}$$

where,

$$F_{Y|X}(y) = P(Y \leq y|X) = E(1_{(Y\leq y)}|X)$$

which always exists, since, for each $y \in \mathbb{R}$, the random variable $1_{(Y\leq y)}$ is bounded, and hence the conditional expectation $E(1_{(Y\leq y)}|X)$ exists (as a Radon-Nikodym derivative).

Theorem. The conditional α–quantile of Y given X minimizes $E\rho_\alpha(Y - \varphi(X))$ over all possible $\varphi(X)$.

Proof. Indeed, using the same proof for unconditional quantiles, $q_\alpha(Y|X = x)$ minimizes $E[\rho_\alpha(Y - a)|X = x]$ so that (integrating over P_X) the function $x \to q_\alpha(Y|X = x)$ minimizes $E\rho_\alpha(Y - \varphi(X))$. Q.E.D.

For applications, a linear conditional quantile model is

$$Y = \beta(\alpha)X + \varepsilon_\alpha$$

where $q_{\varepsilon_\alpha|X}(\alpha) = 0$.

Remark. Another useful application of quantiles. The one dimensional notion of quantiles plays an interesting role in connections with *copulas, OT*, with applications to *production theory in econometrics*. *What is the "rationale" of the Coob-Douglas production function?*

Recall that the Cobb-Douglas production function (in one dimensional case) is of the form

$$\Phi(.,.) : \mathbb{R} \times \mathbb{R} \to [0, \infty), \Phi(x, y) = x^a y^b$$

with $x, y, a, b \geq 0$.

In econometrics, Φ represents the technological relationship between the amount of two inputs such as labor (X) and physical capital (Y), and the amount of outputs that can be produced by these inputs. In the context of OT, it can model the situation where we wish to assign managers (characterized by scalar characteristic/ talent/ X) to firms (characterized by their market capitalization Y). Consider the case where the number of managers is the same as the number of firms. Of course, an "optimal" assignment is the one which should produce the maximum of outputs (say, surplus).

The economic value generated by a manager with talent x, when working for a firm with size y, is the production output $\Phi(x, y)$.

Let P, Q denote the distribution of X, Y, respectively on $\mathbb{R}$. An assignment of managers to firms is a transport map T such that $T(X) = Y$, in distribution. That constraint means that each manager is assigned only to one firm.

The total value created is $E[\Phi(X,T(X)] = E[XT(X)]$.

It is intuitive to view an optimal assignment should be such that most talented managers will run largest firms. in other words, the variables X,Y should be *comonotone* (varying in the same way). This desirable property could be realized when the production function $\Phi(x,y)$ possesses some appropriate condition.

If we look at the Cobb-Douglas production function, then we see that $\frac{\partial^2 \Phi(x,y)}{\partial x \partial y} \geq 0$, a property that we call *supermodularity.*

Remark. This property is similar to *affiliation* in the theory of *common value auctions*, where it is reasonable to assume that the bidders' (latent) are call*ed affiliated.*

This so since it is expected that a high value of one bidder's estimate (of the auctioned object) makes high values of the other estimates more likely.

Now, for a uniform distribution U on $[0,1]$, we have $F_P^{[-1]}(U) \sim P$. This univariate transport is generalized, say, to two dimensions as follows.

Let U,V be two uniform random variables on $[0,1]$, then for $\pi \in \mathscr{M}(P,Q)$, we have

$$(F_P^{[-1]}(U), F_Q^{[-1]}(V)) \sim \pi$$

where the joint distribution of (U,V) is a copula. And the OT problem is formulated as

$$\sup_{\lambda \in \mathscr{M}(U,V)} E_\lambda [\Phi(F_P^{[-1]}(U), F_Q^{[-1]}(V))]$$

i.e., an extremal copula problem.

Thus, X,Y are *comonotone* if there is U uniform on $[0,1]$ such that $X = F_P^{[-1]}(U), Y = F_Q^{[-1]}(U)$.

It is well known that the copula associated with comonotone variables X,Y is $C(u,v) = \min(u,v)$.

An important theoretical result is this.

Theorem. If the surplus (production) function Φ is supermodular, then the OT problem

$$\sup_{\pi \in \mathscr{M}(P,Q)} E_\pi[\Phi(X,Y)]$$

has a solution. In particular, if P has no mass points, then $F_Q^{-1} \circ F_P(x) = T(x)$ is an optimal transport map satisfied $Y = T(X)$.

Looking back at Cobb-Douglas production function, the above result indicates that it is optimal to match higher talented managers to larger firms (and less talented managers to smaller firms).

Just like a complex number $z = x + iy$ that can be written in polar coordinates as $z = re^{i\theta}$, a random variable Y with distribution F can be "factored" as $Y \overset{D}{=} F^{-1}(U)$, called a *polar factorization* of Y**.** It is this polar factorization which is the appropriate equivalent representation for univariate quantile function to be extended to higher dimensions, as $Y = \nabla\varphi(U)$, when Y is a random vector in $\mathbb{R}^d, d \geq 2$, U is uniformly distributed on $[0,1]^d$, and $\nabla\varphi$ is the gradient of a (unique) convex function $\varphi : [0,1]^d \to \mathbb{R}$.

Specifically, the vector quantile of a multivariate distribution function F is the gradient of a convex function, and its justification is within Optimal Transport Theory.

If X is a k–dimensional random vector, then the *conditional vector quantile* of Y given $X = x$ is the multivariate quantile of the random vector $Y|X = x$.

Not only for a parallel with mean linear regression, but in view of natural applications, it seems desirable to extend univariate quantile regression model to *multivariate quantile regression.*

There are many different approaches to defining the notion of multivariate (vector) quantile, but the BEST one is the (recent, 2016) approach based upon OT that we recommend, and elaborate now.

Quoting R. Koenker, with respect to multivariate extension of one dimensional quantiles: "...Despite generating an extensive literature, it is fair to say that no general agreement has emerged...." In contrast to the sample mean of $d-$ dimensional vectors, there is no consensus about an appropriate notion of multivariate median.

Remark on Orders in $\mathbb{R}^d$. The problem seems to be the lack of a natural total order on $\mathbb{R}^d$. The *Pareto* order, $(x_1,x_2,...,x_d) \leq (y_1,y_2,...,y_d)$ if and only if $x_i \leq y_i$ for all $i = 1,2,...,d$, is only a partial (but not total) order. The *lexicographic* order (used in dictionary) is a total order on $\mathbb{R}^d$ where components can be ranked as to importance. It is defined as follows. $(x_1,x_2,...,x_d) \leq (y_1,y_2,...,y_d)$ if $x_1 < y_1$, or $x_1 = y_1$ and $x_2 < y_2$, or $x_1 = y_1, x_2 = y_2$ and $x_3 < y_3$, or...or $x_i = y_i, i = 1,2,...,d-1$ and $x_d < y_d$, or $x_i = y_i, i = 1,2,...,d$.

Why the problem of extending univariate quantiles to multivariate quantiles so difficult? Well, we have just said it "there is no natural total order relation on $\mathbb{R}^d$ for $d > 1$".

The extension problem is difficult since we tried to extend the univariate quantile *directly* from its definition. In history of mathematics, often when we face an extension problem, such as fuzzy sets, quantum probability, and even "extension of transport maps to transport plans" in OT (!), while we cannot directly extend an existing notion, we look for some equivalent representation of it which can be extended. In the case of univariate quantile, perhaps mathematicians have this "extension methodology" in mind, but it was not easy to find an equivalent representation of univariate quantile which can be extended.

Finally, the extension problem was found in 2016, thanks to OT! It was R. Koenker himself to announce it.

It is impossible to generalize this one dimensional quantile function to $\mathbb{R}^d$, with $d > 1$, since there is no (natural) total order relation of $\mathbb{R}^d$, if we try to generalize this function so defined. In other words, we cannot "directly" generalize this concept. We could try to generalize it "indirectly"?

Remember, how Kantorovich generalized Monge's OT formulation? For example, how to generalize a permutation σ (a pure assignment) on $(1,2,...,n\}$ to transport plan?

We cannot do it "directly", so we search for an equivalent representation of σ, i.e., looking for some indirect way. An equivalent representation (an one-to-one map) of a permutation is a permutation matrix to be generalized.

We could do the same thing to generalize quantiles. Perhaps, the difficulty is to find a "canonical" equivalent representation for the quantile map $F^{-1}(.)$ which could be extended.

Perhaps, it was so since an equivalent representation of $F^{[-1]}(.)$ is somewhat "hidden"!

Although we all know that $F^{[-1]}(.)$ is basic for *simulations* because if U is a random variable, uniformly distributed on $[0,1]$, then the random variable $F^{[-1]}(U) = F^{[-1]} \circ U$ has $F(.)$ as its distribution.

Note again that, while the polar factorization of a random variable is used for simulations, it is somewhat hidden (latent) in quantile regression analysis (not needed).

Thus, a characteristic of $F^{[-1]}(.) : [0,1] \to \mathbb{R}$ is that it transports the uniform distribution $\mathscr{U}$ on $[0,1]$ to dF on $\mathbb{R}$, in the "language" of OT, in other words, the quantile function $F^{[-1]}(.)$ is a transport map in OT theory. Is it an equivalent representation for quantiles? Not obviously!

Any way, what seems to be missing is that the probability space $([0,1], \mathscr{U})$ is hidden in the "background": When we define $F^{[-1]}(.)$, we did not (in fact, need) mention it at all. Only it surface after, for simulations.

It is hidden, but it's there! in the language OT, we need to involve the "background" $([0,1], \mathscr{U})$ to describe $F^{[-1]}(.)$ as a transport map.

So let say this. The quantile function $F^{[-1]}(.) : [0,1] \to \mathbb{R}$ is a transport map pushing $\mathscr{U}$ forward to dF.

If this is an equivalent representation of $F^{[-1]}(.)$ in the context of OT, then we hope to be able to say this.

Let $X : \Omega \to \mathbb{R}^d$ with multivariate distribution function $F(.) : \mathbb{R}^d \to [0,1]$. Then the quantile map of F is $Q_F(.) : [0,1]^d \to \mathbb{R}^d$, defined as the transport map pushing forward the uniform probability on $[0,1]^d$ to dF on $\mathbb{R}^d$.

We proceed now to justify the above definition of multivariate (vector) quantiles, to specify it, to give meaning to it, to provide examples, to define *conditional multivariate quantiles, and multivariate quantile regression.*

If we look closerly at the notion of (univariate) quantile function $F^{[-1]}$ of a random variable X with distribution function F, then we realize something fundamental in Monte Carlo (simulation), namely $F^{[-1]}(U) \overset{D}{=} X$, for a random variable U, uniformly distributed on $[0,1]$.

The upshot is this. Rather than "look" at the very definition of $F^{[-1]}(.)$, we could "look" at $F^{[-1]}(.)$ as a map from $[0,1]$ to $\mathbb{R}$, having the property that $F^{[-1]}(U) \overset{D}{=} X$.

Specifically, consider $(\mathscr{X}, \mu) = ([0,1], u)$, where u is the uniform probability measure on $[0,1]$, and $(\mathscr{Y}, \nu) = (\mathbb{R}, dF)$. Then we realize that $F^{[-1]}(.) : \mathscr{X} \to \mathscr{Y}$ is a transport map (of Monge!).

However, in order to say that $F^{[-1]}(.)$ is characterized by such an OT map, we need to show that it is the only transport map in this OT formulation.

Next, for extending this to the multivariate case, we need to show that in the extended OT formulation, namely $(\mathscr{X}, \mu) = ([0,1]^n, u_n)$, where u_n is the uniform probability measure on the unit cube $[0,1]^n$, and $(\mathscr{Y}, \nu) = (\mathbb{R}^n, dF_n)$, where $F_n(.)$ is the multivariate distribution function on $\mathbb{R}^n$, there is a unique transport map.

If it is so, then *the unique transport map between* $([0,1]^n, u_n)$ *and* $(\mathbb{R}^n, dF_n)$ *can be used as the multivariate quantile function of the distribution* F_n**.**

It turns out that we do have a theoretical result confirming the above! Thanks to McCann [7].

Theorem. (McCann, 1995). Let μ, ν be two probability measures on $\mathbb{R}^n$, with μ being continuous (i.e., it has no mass points, or equivalently, its associate distribution function is continuous on $\mathbb{R}^n$, e.g., uniform measure on unit cube). Then there is a measurable map $T(.) : \mathbb{R}^n \to \mathbb{R}^n$ which is the gradient of some convex function φ, and such that $\nu = \mu T^{-1}$ (equivalently, $T(X) = Y$, where $X \sim \mu, Y \sim \nu$). Moreover, T is unique μ–a.s.

Remark. The gradient of a multivariate (differentiable) function ($\varphi : \mathbb{R}^n \to \mathbb{R}$) is the vector of its partial derivatives. If φ is differentiable, i.e., having first order partial derivatives, then φ is convex if and only if for any $x, y \in \mathbb{R}^n$, we have $<\nabla\varphi(x) - \nabla\varphi(y), x-y> \geq 0$ (gradient monotonicity). For $n = 1$, a differentiable convex function has nondecreasing derivative.

The Theorem says that if $X \sim \mu$, then there is a unique convex function φ such that its gradient $\nabla\varphi(X) \sim \nu$, i.e.,$\nabla\varphi(.)$ is a transport map.

Let's elaborate a bit on this fundamental theorem.

For $n = 1$, consider $(\mathscr{X}, \mu) = ([0,1], u)$, noting that the uniform measure u is continuous, and $(\mathscr{Y}, \nu) = (\mathbb{R}, dF)$. The quantile function $F^{[-1]}(.) : [0,1] \to \mathbb{R}$, which is non decreasing, and transporting u to dF, because $F^{[-1]}(U) \sim dF$. The quantile function F^{-1} is nondecreasing and hence is the derivative of a convex function. Thus, $F^{[-1]}$ fits perfectly McCann's Theorem, and hence is a (a.s.) unique transport map.

Note that if μ is an arbitrary continuous probability measure on $\mathbb{R}^d$ with associate multivariate distribution function F_μ, then the transport map is $F_\nu^{[-1]} \circ F_\mu(x) = \varphi'(x)$.

Thus, the univariate quantile function $F^{[-1]}(.) : [0,1] \to \mathbb{R}$ with its equivalent representation as a transport map pushing forward the uniform measure on $[0,1]$ to the probability measure dF on $\mathbb{R}$, can be extended to higher dimensions, as THE transport map being the gradient $\nabla\varphi$ of some convex function φ on $\mathbb{R}^n$ ($\nabla\varphi$ push forward $([0,1]^n, u_n)$ to $(\mathbb{R}^n, dF_n)$.

The $d-$quantile function of a multivariate distribution function $F(.) : \mathbb{R}^d \to [0,1]$ is the gradient $\nabla\varphi : [0,1]^d \to \mathbb{R}^d$, of some convex function $\varphi : [0,1]^d \to \mathbb{R}$, such that $\nabla\varphi(U) \sim dF$, where U is the uniform random vector on $[0,1]^d$.

The above map $\nabla\varphi$ (Brenier map) is the map between dU (uniform probability measure on the unit cube $[0,1]^d$) and dF.

In one dimension, $\nabla\varphi$ is $F^{[-1]}(.) : [0,1] \to \mathbb{R}$ (a nondecreasing function, such that $F^{-1}(U) \sim F$).

Let X, Y be random vectors on $\mathbb{R}^d$, $\mathbb{R}^k$ with distribution F, G, respectively. Then the conditional multivariate quantile function of $Y|X = x$ is the Brenier map between dU on $[0,1]^d$ and the conditional probability measure of $Y|X = x$, i.e., the multivariate quantile of the conditional distribution.

Specifically, the conditional quantile function of $Y|X = x$ is $\nabla\varphi_x$ where $\varphi_x(.)$ is a convex function on $[0,1]^d$ with $Y = \nabla\varphi_X(U)$.

Note that there are many attempts to define multivariate quantiles in the literature, but as R. Koenker said, this approach based on OT seems the best! mainly because it capture two basic properties of the univariate quantile function $F^{[-1]}(.) : [0,1] \to \mathbb{R}$ (as a kind of "inverse" of F, with a precise meaning, e.g., of median), namely $F^{[-1]}(.)$ is a monotone (nondecreasing) function, and $F^{[-1]}(U) = Y$ (where $Y \sim dF$). This is

so because, as the gradient of a convex function, $\nabla\varphi$ is the natural generalization of monotonicity in one dimension case, and $Y = \nabla\varphi_X(U)$ when X is a covariate.

Let $Q_{Y|X}(u|x)$ be the conditional (multivariate) quantile of $Y|X = x$ at level $u \in [0,1]^d$. A linear model for it is

$$Q_{Y|X}(u|x) = \beta_o(u)^T g(x)$$

so that we have the representation

$$Y = \beta_o(U)^T g(X)$$

with $U|X \sim$ uniform $[0,1]^d$, $\beta(u)$ is $k \times d$ matrix ($X \in \mathbb{R}^k$).

This formulation leads to a linear programming to computing $\beta(u)$ both for population and sample settings.

Remark. Why do we need to consider multivariate quantile regression?

Well, let's spell it out loud again. At the "beginning", Gaussian models made statisticians to center their attention only on the mean, and conditional mean of variables of interest. Then it was discovered that linear (univariate) quantile regression can address more issues in economics. However, we only have univariate quantile regression (Koenker & Bassett, 1982). As such, even we are really interested in, say, how household expenditures affect total income, we can only look at a specific component of household expenditures, e.g., food expenditure, one among 9 possible components of household expenditures: Food, Clothing, Housing, Heating and Lighting, Tools, Education, Safety, Medical care, Services. A multivariate ($d = 9$) quantile regression is desirable, and now possible!

Multivariate quantile functions are useful in a variety of fields, see e.g., Galichon [4], Matzkin [6], Panaretos and Zemel [9], Santambrogio [10].

References

1. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Commun. Pure Appl. Math. **44**(4), 375–417 (1991)
2. Carlier, G., Chernozhukov, V., Galichon, A.: Vector quantile regression: an optimal transport approach. Ann. Stat. **44**, 1165–1192 (2016)
3. Carlier, G., Chernozhukov, V., De Bie, G., Galichon, A.: Vector quantile regression and optimal transport, from theory to numerics. Empir. Econ. **62**, 35–62 (2020). https://doi.org/10.1007/s00181-020-01919-y
4. Galichon, A.: Optimal Transport Methods in Economics. Princeton University Press, Princeton (2016)
5. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**(1), 33–50 (1978)
6. Matzkin, R.L.: Nonparametric estimation of nonadditive random functions. Econometrica **71**(5), 1339–1375 (2003)
7. McCann, R.J.: Existence and uniqueness of monotone measure-preserving maps. Duke Math. J. **80**(2), 309–323 (1995)
8. Monge, G.: Memoire sur la theorie des dblais et des remblais. Histoire de l'Academie Royale des Sciences de Paris 666–704 (1781)

9. Panaretos, V.M., Zemel, Y.: An Invitation to Statistics in Wasserstein Space. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-38438-8
10. Santambrogio, V.: Optimal Transport for Applied Mathematicians. Birkhauser, Cham (2015)
11. Villani, V.: Topics in Optimal Transportation, vol. 58. American Mathematical Society, Providence (2003)
12. Villani, V.: Optimal Transport: Old and New, vol. 338. Springer, Heidelberg (2008)