Vector quantile regression beyond the specified case

Journal of Multivariate Analysis 161 (2017) 96–102

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Vector quantile regression beyond the specified case

Guillaume Carlier [a,b], Victor Chernozhukov [c], Alfred Galichon [d,*]

[a] CEREMADE, UMR CNRS 7534, Université Paris IX Dauphine, Pl. de Lattre de Tassigny, 75775 Paris Cedex 16, France
[b] MOKAPLAN Inria Paris, France
[c] Department of Economics, MIT, 50 Memorial Drive, E52-361B, Cambridge, MA 02142, USA
[d] Economics Department and Courant Institute of Mathematical Sciences, NYU, 70 Washington Square South, New York, NY 10013, USA

ARTICLE INFO

Article history:
Received 21 October 2016
Available online 27 July 2017

Keywords:
Duality
Optimal transport
Vector quantile regression

ABSTRACT

This paper studies vector quantile regression (VQR), which models the dependence of a random vector with respect to a vector of explanatory variables with enough flexibility to capture the whole conditional distribution, and not only the conditional mean. The problem of vector quantile regression is formulated as an optimal transport problem subject to an additional mean-independence condition. This paper provides results on VQR beyond the specified case which had been the focus of previous work. We show that even beyond the specified case, the VQR problem still has a solution which provides a general representation of the conditional dependence between random vectors.

1. Introduction

Vector quantile regression was recently introduced in [4] in order to generalize the technique of quantile regression when the dependent random variable is multivariate. Quantile regression, pioneered by Koenker and Bassett [12], provides a powerful way to study dependence between random variables assuming a linear form for the quantile of the endogenous variable Y given the explanatory variables X. It has therefore become a very popular tool in many areas of economics, program evaluation, biometrics, etc. However, a well-known limitation of the approach is that Y should be scalar so that its quantile map is defined. When Y is multivariate, there is no canonical notion of quantile, and the picture is less clear than in the univariate case. There is actually an important literature that aims at generalizing the notion of quantile to a multidimensional setting and various different approaches have been proposed; see in particular [1,10,15] and the references therein.

The approach proposed in [4] is based on optimal transport ideas and can be described as follows. For a random vector Y taking values in $\mathbb{R}^d$, we look for a random vector U uniformly distributed on the unit cube $[0,1]^d$ and which is maximally correlated to Y; finding such a U is an optimal transport problem. A celebrated result called Brenier's theorem [2,3,17] implies that such an optimal U is characterized by the existence of a convex function φ such that $Y = \nabla\varphi(U)$. When $d = 1$, of course, the optimal transport map of Brenier $\nabla\varphi = Q$ is the quantile map of Y. In higher dimensions it still retains one of the main properties of univariate quantiles, namely monotonicity. Thus Brenier's map $\nabla\varphi$ is a natural candidate to be considered as the vector quantile of Y, and one advantage of such an approach is the point-wise relation $Y = \nabla\varphi(U)$, where U is a uniformly distributed random vector which best approximates Y in L^2.

If, in addition, we are given another random vector X capturing a set of observable explanatory variables, we may wish to have a tractable method to estimate the conditional quantile of Y given $X = x$, i.e., the map $u \in [0,1]^d \mapsto Q(x,u) \in \mathbb{R}^d$.

* Corresponding author.
E-mail addresses: carlier@ceremade.dauphine.fr (G. Carlier), vchern@mit.edu (V. Chernozhukov), ag133@nyu.edu (A. Galichon).

http://dx.doi.org/10.1016/j.jmva.2017.07.003

In the univariate case $d = 1$, and if the conditional quantile is affine in x, i.e., $Q(x, u) = \alpha(u) + \beta(u)x$, the quantile regression method of Koenker and Bassett gives a constructive and powerful linear programming approach to compute the coefficients $\alpha(t)$ and $\beta(t)$ for any fixed $t \in [0, 1]$. When quantile regression is specified, i.e., when the true conditional quantile is affine in x, this variational approach estimates the true coefficients $\alpha(t)$ and $\beta(t)$. In [4], we have shown that in the multivariate case as well, when the true vector quantile is affine in x, one may estimate it by a variational problem which consists in finding the uniformly distributed random variable U such that $\mathrm{E}(X \mid U) = \mathrm{E}(X)$ (mean independence) and maximally correlated with Y.

The purpose of the present paper is to convey what these variational approaches tell about the dependence between Y and X in the general case, i.e., without assuming any particular form for the conditional quantile. We will characterize the solution of the optimal transport problem with a mean-independence constraint from [4] and relate it to a relaxed form of specified quantile regression. To be more precise, our Theorem 3 below will provide the following general representation of the distribution of (X, Y):

$$Y \in \partial\Phi_X^{**}(U) \quad \text{with } X \mapsto \Phi_X(U) \text{ affine}, \qquad \Phi_X(U) = \Phi_X^{**}(U) \quad \text{almost surely}, \qquad U \sim \mathcal{U}[0,1]^d, \quad \mathrm{E}(X \mid U) = \mathrm{E}(X),$$

where Φ_x^{**} denotes the convex envelope of $u \mapsto \Phi_x(u)$ for a fixed x, and ∂ denotes the subdifferential. The main ingredients are convex duality and an existence theorem for optimal dual variables. The latter is a non-trivial extension of Kantorovich duality: indeed, the existence of a Lagrange multiplier associated to the mean-independence constraint is not straightforward and we shall prove it thanks to Komlos' theorem (Theorem 2). Vector quantile regression is *specified* if $u \mapsto \Phi_x(u)$ is convex and differentiable for every x, in which case one can write

$$Y = \nabla\Phi_X(U) \quad \text{with } \Phi_X(\cdot) \text{ convex}, \qquad X \mapsto \Phi_X(U) \text{ affine}, \qquad U \sim \mathcal{U}[0,1]^d, \quad \mathrm{E}(X \mid U) = \mathrm{E}(X).$$

While our previous paper [4] focused on the specified case, the results we obtain in the present paper are general.

In the paper, we will use the following notations. For two vectors x and y of $\mathbb{R}^d$, the scalar product of x and y is denoted $x^\top y$. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, f^* shall denote the Legendre–Fenchel transform, defined as

$$f^*(y) = \max_{x \in \mathbb{R}^d} \{x^\top y - f(x)\}.$$

The subdifferential of a convex function f, denoted $\partial f(x)$ is defined as $\arg\max_{y \in \mathbb{R}^d} \left\{x^\top y - f^*(y)\right\}$. Given a subset Ω of $\mathbb{R}^d$, $\overline{\Omega}$ denotes the closure of Ω, and $|\Omega|$ denotes the Lebesgue measure of Ω. Given a random variable X, Law (X) denotes the probability distribution of X. Given a measure ν, spt (ν) denotes the support of ν. Given a smooth map $Z : \mathbb{R}^n \to \mathbb{R}^m$, DZ denotes the Jacobian of Z, which is the $m \times n$ matrix with entries $\partial Z^i(x)/\partial x^j$, we also denote its transpose by $DZ^\top$.

The rest of the paper is organized as follows. Section 2 gives reminders on optimal transport in relation to multivariate quantiles. Section 3 provides the main results on vector quantile regression beyond the specified case. Section 4 discusses two possible applications of the method, one for biometric purposes, and the other for economic purposes.

2. Vector quantiles and optimal transport

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some nonatomic probability space, and let (X, Y) be a random vector, where the vector of explanatory variables X is valued in $\mathbb{R}^N$ and the vector of dependent variables Y is valued in $\mathbb{R}^d$.

The notion of vector quantile was recently introduced by Ekeland et al. [7] and Galichon and Henry [9]. It was used in the framework of quantile regression in our companion paper [4]. The starting point for this approach is the correlation maximization problem

$$\max_V \{\mathrm{E}(VY), \ \mathrm{Law}(V) = \mu\} \tag{1}$$

where $\mu = \mathcal{U}[0,1]^d$ is the uniform measure on the unit d-dimensional cube $[0,1]^d$. This problem is equivalent to the optimal transport problem which consists in minimizing $\mathrm{E}(|Y - V|^2)$ among uniformly distributed random vectors V. This quadratic optimal transport problem has received a lot of attention since the 1980s; see [6,11,20]. An important result in this field is Brenier's polar factorization theorem [2,3,17] ensuring that (1) has a solution U which is characterized by the condition

$$Y = \nabla\varphi(U)$$

for some (essentially uniquely defined) convex function φ obtained by solving a dual formulation of (1). Arguing that gradients of convex functions are the natural multivariate extension of monotone nondecreasing functions, the authors of [7,9] considered the function $Q = \nabla\varphi$ as the vector quantile of Y. We therefore shall define the quantile of Y as the optimal transport map (for the quadratic cost) $Q = \nabla\varphi$ between the uniform measure on $[0,1]^d$ and Law(Y). We refer to the textbooks [18,19,21] for a presentation of optimal transport theory, and to [8] for a survey of applications to economics.

Let us now assume that in addition, there is an N-dimensional random vector X of regressors, $\nu = \mathrm{Law}(X, Y)$, $m = \mathrm{Law}(X)$, $\nu = \nu^x \otimes m$, where m is the law of X and ν^x is the law of Y given $X = x$. One can consider $Q(x, u) = \nabla\varphi(x, u)$ as the optimal transport between μ and ν^x, viz.

$$Y = Q(X, U) = \nabla_u\varphi(X, U), \quad U \sim \mathcal{U}[0,1]^d.$$

By definition, $Q(X,\cdot)$ is the conditional vector quantile of Y given X. Note that in the specified case, i.e., when the conditional quantile function is affine in X and $Y = Q(X,U) = \alpha(U) + \beta(U)X$, where U is uniform and independent from X, the function $u \mapsto \alpha(u) + \beta(u)x$ should be the gradient of some function of u which requires

$$\alpha = \nabla\varphi, \qquad \beta = Db^{\top}$$

for some potential φ and some vector-valued function b, in which case $Q(x,\cdot)$ is the gradient of $u \mapsto \varphi(u) + b(u)x$. Moreover, since quantiles are gradients of convex potentials, one should also have that $u \in [0,1]^d \mapsto \varphi(u) + b(u)x$ is convex.

3. Vector quantile regression

3.1. Correlation maximization

Without loss of generality we normalize X so that it is centered, i.e., $\mathrm{E}(X) = 0$. Our approach to vector quantile regression is based on the following correlation maximization problem, subject to a mean-independence constraint:

$$\max_{V}\{\mathrm{E}(V\cdot Y),\ \mathrm{Law}(V) = \mu,\ \mathrm{E}(X \mid V) = 0\}, \tag{2}$$

where $\mu = \mathcal{U}[0,1]^d$ is the uniform measure on the unit d-dimensional cube; the existence of a solution to (2) is standard: in terms of the joint law of (X,Y,V) it is a linear maximization on a weakly-$\star$ compact set. We now make the following assumption on the dependence structure of (U,X,Y).

Assumption 1. Assume that (U,X,Y) is a random vector such that:

(i) U is a random vector of $\mathbb{R}^d$ with distribution μ;
(ii) (X,Y) is a random vector of $\mathbb{R}^N \times \mathbb{R}^d$ with distribution ν;
(iii) X is mean-independent from U, i.e., $\mathrm{E}(X \mid U) = 0$.

The connection with the specification of vector quantile regression is then given by the following result from [4].

Proposition 1. *Let (U,X,Y) satisfy Assumption 1. Suppose that there exists a smooth function $\varphi : \mathbb{R}^d \to \mathbb{R}$ and a smooth function $b : \mathbb{R}^d \to \mathbb{R}^N$ such that $u \mapsto \varphi(u) + b(u)^{\top}x$ is convex for m-almost every value of x, so that (U,X,Y) satisfy $Y = \nabla\varphi(U) + Db(U)^{\top}X$. Then U solves (2).*

3.2. Duality

We now aim at emphasizing the kind of information provided by relation (2) regarding the dependence of X and Y. A good starting point is convex duality. As explained in [4], the dual of (2) takes the form

$$\inf_{(\psi,\varphi,b)} \mathrm{E}\{\psi(X,Y) + \varphi(U)\} : \psi(x,y) + \varphi(t) + b(t)x \geq ty, \tag{3}$$

where $\mathrm{Law}(U) = \mu = \mathcal{U}[0,1]^d$ and the infimum is taken over continuous functions $\psi \in \mathcal{C}(\mathrm{spt}(\nu),\mathbb{R})$, $\varphi \in \mathcal{C}([0,1]^d,\mathbb{R})$ and $b \in \mathcal{C}([0,1]^d,\mathbb{R}^N)$ satisfying the pointwise constraint

$$\forall_{(x,y,t)\in \mathrm{spt}(\nu)\times[0,1]^d} \quad \psi(x,y) + \varphi(t) + b(t)x \geq ty.$$

Since for fixed (φ,b), the smallest ψ that satisfies the pointwise constraint in (3) is given by the convex function

$$\psi(x,y) = \max_{t\in[0,1]^d}\{ty - \varphi(t) - b(t)x\},$$

one may equivalently rewrite (3) as the minimization over continuous functions φ and b of

$$\int \max_{t\in[0,1]^d}\{ty - \varphi(t) - b(t)x\}\nu(dx,dy) + \int_{[0,1]^d}\varphi(t)\mu(dt). \tag{4}$$

By standard approximation techniques, one can show that the infimum in (4) over continuous functions (φ,b) coincides with the infimum over smooth or simply integrable functions. Let us make the following assumption.

Assumption 2. Assume that ν is an absolutely continuous probability measure over $\mathbb{R}^N \times \mathbb{R}^d$ with density g such that:

(i) the support of ν is $\overline{\Omega}$, where Ω is an open bounded convex subset of $\mathbb{R}^N \times \mathbb{R}^d$;
(ii) g is bounded on Ω, and bounded away from zero on compact subsets of Ω.

The existence of optimal (L^1) functions ψ, φ and b is our first main result.

Theorem 2. *Let ν satisfy Assumption 2. Then, the dual problem (3) admits at least one solution.*

Proof. Let us denote by $(0, \bar{y})$ the mean of ν, viz.

$$\int_\Omega x\ \nu(dx, dy) = 0, \quad \int_\Omega y\ \nu(dx, dy) = \bar{y},$$

and observe that $(0, \bar{y}) \in \Omega$; otherwise, by convexity, ν would be supported on $\partial\Omega$ which would contradict our assumption that $\nu \in L^\infty(\Omega)$.

We wish to prove the existence of optimal potentials for the problem

$$\inf_{\psi,\varphi,b} \int_\Omega \psi(x, y)d\nu(x, y) + \int_{[0,1]^d} \varphi(u)d\mu(u) \tag{5}$$

subject to the pointwise constraint that

$$\psi(x, y) + \varphi(u) \geq uy - b(u)x, \quad (x, y) \in \overline{\Omega},\ u \in [0, 1]^d. \tag{6}$$

Of course, we can take ψ that satisfies

$$\psi(x, y) = \sup_{u\in[0,1]^d} \{uy - b(u)x - \varphi(u)\}$$

so that ψ can be chosen convex and 1-Lipschitz with respect to y, hence

$$\psi(x, \bar{y}) - |y - \bar{y}| \leq \psi(x, y) \leq \psi(x, \bar{y}) + |y - \bar{y}|. \tag{7}$$

The problem being invariant by the transform $(\psi, \varphi) \mapsto (\psi + C, \psi - C)$ – C being an arbitrary constant – we can add as a normalization the condition that $\psi(0, \bar{y}) = 0$. This normalization and the constraint (6) imply that

$$\varphi(t) \geq t\bar{y} - \psi(0, \bar{y}) \geq -|\bar{y}|.$$

We note that there is one extra invariance of the problem: if one adds an affine term qx to ψ, this does not change the cost and neither does it affect the constraint, provided one modifies b accordingly by subtracting the constant vector q from it. Take then q in the subdifferential of $x \mapsto \psi(x, \bar{y})$ at 0 and change ψ into $\psi - qx$, we obtain a new potential with the same properties as above and with the additional property that $\psi(\cdot, \bar{y})$ is minimal at $x = 0$, and thus $\psi(x, \bar{y}) \geq 0$, together with (7) this gives the lower bound

$$\psi(x, y) \geq -|y - \bar{y}| \geq -C,$$

where the bound comes from the boundedness of Ω. From now on, C will denote a generic constant which may change from one line to the next.

Now take a minimizing sequence $(\psi_n, \varphi_n, b_n) \in \mathcal{C}(\overline{\Omega}, \mathbb{R}) \times \mathcal{C}([0, 1]^d, \mathbb{R}) \times \mathcal{C}([0, 1]^d, \mathbb{R}^N)$, where for each n, ψ_n has been chosen with the same properties as above. Since φ_n and ψ_n are bounded from below ($\varphi_n \geq -|\bar{y}|$ and $\psi_n \geq C$) and since the sequence is minimizing, we deduce immediately that ψ_n and φ_n are bounded sequences in L^1. Let $z = (x, y) \in \Omega$ and $r > 0$ be such that the distance between z and the complement of Ω is at least $2r$ – so that $B_r(z)$ is in the set of points that are at least at distance r from $\partial\Omega$ – by assumption there is an $\alpha > 0$ such that $g \geq \alpha$ on $B_r(z)$. We then deduce from the convexity of ψ_n that

$$C \leq \psi_n(z) \leq \frac{1}{|B_r(z)|} \int_{B_r(z)} \psi_n \leq \frac{1}{|B_r(z)|\alpha} \int_{B_r(z)} |\psi_n| g \leq \frac{1}{|B_r(z)|\alpha} \|\psi_n\|_{L^1(\nu)}$$

so that ψ_n is actually locally bounded and by convexity, we also have

$$\|\nabla\psi_n\|_{L^\infty(B_r(z))} \leq \frac{2}{R - r} \|\psi_n\|_{L^\infty(B_R(z))}$$

whenever $R > r$ and $B_R(z) \subset \Omega$; see, e.g., the proof of Lemma 5.1 in [5] for the derivation of such a bound. We can thus conclude that ψ_n is also locally uniformly Lipschitz. Therefore, thanks to Ascoli's theorem, we can assume, taking a subsequence if necessary, that ψ_n converges locally uniformly to some potential ψ.

Let us now prove that b_n is bounded in L^1. To this end, take $r > 0$ such that $B_{2r}(0, \bar{y})$ is included in Ω. For every $x \in B_r(0)$, any $t \in [0, 1]^d$ and any n, we then have

$$-b_n(t)x \leq \varphi_n(t) - t\bar{y} + \|\psi_n\|_{L^\infty(B_r(0,\bar{y}))} \leq C + \varphi_n(t),$$

and maximizing in $x \in B_r(0)$ immediately gives $|b_n(t)|r \leq C + \varphi_n(t)$, from which we deduce that b_n is bounded in L^1 since φ_n is. From Komlos' theorem [14], we may find a subsequence such that the Cesàro means

$$\frac{1}{n}\sum_{k=1}^n \varphi_k, \qquad \frac{1}{n}\sum_{k=1}^n b_k$$

converge a.e. to some φ and b, respectively. Clearly ψ, φ and b satisfy the linear constraint (6), and since the sequence of Cesàro means $(\psi_n', \phi_n', b_n') = \sum_{k=1}^n (\psi_k, \phi_k, b_k)/n$ is also minimizing, we deduce from Fatou's Lemma that

$$\int_\Omega \psi(x,y)d\nu(x,y) + \int_{[0,1]^d} \varphi(u)d\mu(u) \le \liminf_n \int_\Omega \psi_n'(x,y)d\nu(x,y) + \int_{[0,1]^d} \varphi_n'(u)d\mu(u) = \inf\{\text{Eq. (5)}\}.$$

This concludes the proof of Theorem 2. □

3.3. Vector quantile regression as optimality conditions

Let U solve (2) and (ψ, φ, b) solve its dual (3). Recall that, without loss of generality, we can take ψ convex and given by

$$\psi(x,y) = \sup_{t\in[0,1]^d} \{ty - \varphi(t) - b(t)x\}. \tag{8}$$

The constraint of the dual is

$$\forall_{(x,y,t)\in\Omega\times[0,1]^d} \qquad \psi(x,y) + \varphi(t) + b(t)x \ge ty \tag{9}$$

and the primal–dual relations give that, almost surely

$$\psi(X,Y) + \varphi(U) + b(U)X = UY, \tag{10}$$

which, since ψ given by (8) is convex, yields

$$(-b(U), U) \in \partial\psi(X,Y), \text{ or, equivalently } (X,Y) \in \partial\psi^*(-b(U), U).$$

Problems (2) and (3) have thus enabled us to find:

- ✓ U uniformly distributed with X mean-independent from U,
- ✓ $\phi : [0,1]^d \to \mathbb{R}$, $b : [0,1]^d \to \mathbb{R}^N$ and $\psi : \Omega \to \mathbb{R}$ convex,

such that $(X,Y) \in \partial\psi^*(-b(U), U)$. Specification of vector quantile regression rather asks whether one can write $Y = \nabla\varphi(U) + Db(U)^\top X = \nabla\Phi_X(U)$ with $u \mapsto \Phi_x(u) = \varphi(u) + b(u)x$ convex in u for fixed x. The smoothness of φ and b is actually related to this specification issue. Indeed, if φ and b were smooth, then (by the envelope theorem) we would have

$$Y = \nabla\varphi(U) + Db(U)^\top X = \nabla\Phi_X(U).$$

But the smoothness of φ and b is not enough to guarantee that the conditional quantile is affine in x, which would also require $u \mapsto \Phi_x(u)$ to be convex. Note also that if ψ was smooth, we would then have

$$U = \nabla_y\psi(X,Y), \qquad -b(U) = \nabla_x\psi(X,Y).$$

In general (without assuming any smoothness), define $\psi_x(y) = \psi(x,y)$. We then have, thanks to (9)–(10), $U \in \partial\psi_X(Y)$, i.e., $Y \in \partial\psi_X^*(U)$. The constraint of (3) also gives $\psi_x(y) + \Phi_x(t) \ge ty$ since the Legendre Transform is order-reversing; this implies

$$\psi_x \ge \Phi_x^* \tag{11}$$

hence $\psi_x^* \le (\Phi_x)^{**} \le \Phi_x$, where Φ_x^{**} denotes the convex envelope of Φ_x. Duality between (2) and (3) thus yields the following result.

Theorem 3. *Let U be a random variable over $\mathbb{R}^d$ solution to (2), and let $\psi : \mathbb{R}^N \times \mathbb{R}^d \to \mathbb{R}$, $\varphi : \mathbb{R}^d \to \mathbb{R}$, $b : \mathbb{R}^d \to \mathbb{R}^N$ be functions such that (ψ, φ, b) solve the corresponding dual problem (3). For every $(t,x) \in [0,1]^d \times \mathrm{spt}(m)$, define $\Phi_x(t) = \varphi(t) + b(t)x$. Then*

$$\Phi_X(U) = \Phi_X^{**}(U) \text{ and } U \in \partial\Phi_X^*(Y), \quad \text{i.e., } Y \in \partial\Phi_X^{**}(U) \text{ a.s.} \tag{12}$$

Proof. From the duality relation (10) and (11), we have $UY = \psi_X(Y) + \Phi_X(U) \ge \Phi_X^*(Y) + \Phi_X(U)$, so that $UY = \Phi_X^*(Y) + \Phi_X(U)$ and then $\Phi_X^{**}(U) \ge UY - \Phi_X^*(Y) = \Phi_X(U)$. Hence, $\Phi_X(U) = \Phi_X^{**}(U)$ and $UY = \Phi_X^*(Y) + \Phi_X^{**}(U)$, i.e., $U \in \partial\Phi_X^*(Y)$ almost surely, and the latter is equivalent to the requirement that $Y \in \partial\Phi_X^{**}(U)$. □

The previous theorem thus gives the following interpretation of the correlation maximization with a mean independence constraint (2) and its dual (3). These two variational problems in duality lead to the pointwise relations (12) which can be seen as best approximations of a specification assumption:

$$Y = \nabla\Phi_X(U), \ (X,U) \mapsto \Phi_X(U) \text{ affine in } X, \text{ convex in } U.$$

Indeed in (12), Φ_X is replaced by its convex envelope, the uniform random variable U solving (2) is shown to lie a.s. in the contact set $\Phi_X = \Phi_X^{**}$ and differentiability is replaced by a subdifferential condition.

4. Discussion

To conclude the paper, let us make some remarks on computations and highlight two possible uses of VQR, one pertaining to biometrics, and one to economics.

4.1. Computation

In this paper we have not discussed the implementation issues (discretization and computation), which are discussed in Section 4 of [4]. Let us simply mention that when μ and ν are discrete probability measures with respective supports $\{(u_i) : 1 \leq i \leq I\}$ and $\{(x_j, y_j) : 1 \leq j \leq J\}$ with associated weights $\mu_i > 0$ and $\nu_j > 0$, so that $\mu = \sum_{i=1}^{I} \mu_i \delta_{u_i}$ and $\nu = \sum_{j=1}^{J} \nu_j \delta_{(x_j,y_j)}$, problem (3) becomes a finite-dimensional linear programming problem

$$\min_{\psi_j, \varphi_i, b_i^n} \sum_{j=1}^{J} \nu_i \varphi_i + \sum_{j=1}^{J} \mu_j \psi_j, \quad \text{subject to } \psi_j + \varphi_i + \sum_{n=1}^{N} b_i^n x_j^n \geq \sum_{k=1}^{d} u_i^k y_j^k \tag{13}$$

where x_i^n is the nth dimension of $x_i \in \mathbb{R}^N$, b_i^n stands for the nth dimension of $b(x_i) \in \mathbb{R}^N$, and u_i^k and y_j^k are the kth dimension of respectively u_i and y_j, which are two vectors of $\mathbb{R}^d$. Problem (13) can be computed using standard large-scale linear programming solvers. When μ and ν are continuous probability measures, they will be replaced by a sampled version; the study of the stability of the VQR parameters is left for future work, but it is possible to investigate version of (13) with entropic regularization: letting $T > 0$ be a temperature parameter, consider

$$\min_{\psi_j, \varphi_i, b_i^n} \sum_{j=1}^{J} \nu_i \varphi_i + \sum_{i=1}^{J} \mu_j \psi_j + T \sum_{\substack{1 \leq i \leq I \\ 1 \leq j \leq J}} \exp\left(\frac{\sum_{k=1}^{d} u_i^k y_j^k - \psi_j - \varphi_i - \sum_{n=1}^{N} b_i^n x_j^n}{T} \right),$$

which can be efficiently computed via coordinate descent, in the spirit of the Iterative Proportional Fitting Procedure (IPFP) also known as the Sinkhorn algorithm; see [16]. The detailed study of this numerical procedure is again left for future work.

4.2. Measurements of newborn babies

One first possible application of VQR will help understand how the height and weight of newborn babies is affected by the characteristics of their mother, such as education. This relates to pioneering work done by Koenker and Hallock [13], using scalar quantile regression. Assume Y is a vector of biometric measures of a newborn baby, where Y_1 is the height and Y_2 is the weight; and assume that X measures the mother's education, one would like to understand the effect of education (X) on the joint distribution of height (Y_1) and weight (Y_2). Note that scalar quantile regression does not allow to study the impact on the joint distribution, but only the impact on the marginal distributions. Nevertheless, it is of interest to understand whether conditional on X, the rank of the weight in the conditional distribution is or not correlated with the rank of the height.

Letting $\Phi_X(U) = \varphi(U) + b(U)^\top X$, one has $Y \in \partial \Phi_X^{**}(U)$, i.e., $U \in \partial \Phi_X^{*}(Y)$ and thus the barycenter of $\partial \Phi_x^{*}(y)$ gives the "multivariate rank" $(u_1, u_2) \in [0, 1]^2$ associated to observation (x, y). Similarly, this construction allows to define the "multivariate median" of Y conditional on $X = x$ as the barycenter of $\partial \Phi_x^{**}(1/2, 1/2)$. The plot of the multivariate median of Y conditional on X is informative as it represents the status of a "typical" individual in the population. Similarly, we may define four "extremal" individuals associated with respectively $u^{TH} = (0.9, 0.9)$ (tall and heavy), $u^{TL} = (0.9, 0.1)$ (tall and light), $u^{SH} = (0.1, 0.9)$ (small and heavy), and $u^{SL} = (0.1, 0.1)$ (small and light). For $u \in \{u^{TH}, u^{TL}, u^{SH}, u^{SL}\}$, one may be interested in plotting $\partial \Phi_x^{**}(u)$ as the evolution of these various profiles of individuals when the level of prematurity increases.

4.3. Willingness to pay for real estate amenities

We now consider an economic application of VQR to the real estate market. Assume $y \in \mathbb{R}^d$ is a vector of house characteristics (price, square footage, and amenities), and $x \in \mathbb{R}^N$ is a vector of observable characteristics of the buyer (income, size of the household, age). On top of the observable characteristics x, it is assumed that the consumer is represented by a vector of unobservable characteristics u, so that the valuation of good y by buyer x is

$$u^\top y + V(x, y),$$

where it is assumed that $u \in \mathbb{R}^d$ is uniformly distributed and independent from x and V is concave with respect to y. The goal of this exercise is to identify $V(x, y)$, which is the systematic part of the valuation of y by consumers of observable characteristics x, based on the observation of sales data, which specify the characteristics of the good sold y, jointly with the characteristics of the characteristics x of the corresponding buyers.

A consumer of type (x, u) chooses quality y in order to maximize utility, i.e., the consumer solves

$$\Phi_x(u) = \max_{y \in \mathbb{R}^d} \{u^\top y + V(x, y)\}. \tag{14}$$

Letting $\psi(x, y) = -V(x, y)$, it follows that $\Phi_x(\cdot)$ and $\psi(x, \cdot)$ form a pair of convex conjugate functions. VQR assumes a parameterization of $\Phi_x(u)$ of the form $\Phi_x(u) = \varphi(u) + b(u)^\top x$, and therefore, it follows from the envelope theorem in (14) that $Y \in \Phi_X^{**}(U)$.

Assume now a number of real estate transactions are observed. For each transaction, one observes a vector Y of characteristics of the house sold, and a vector X of characteristics of the buyer. It follows from the considerations above that ψ is identified by

$$\psi(x, y) = \Phi_x^*(y),$$

where $\Phi_x(u) = \varphi(u) + b(u)^\top x$ has been obtained by Vector Quantile Regression of Y on X.

Acknowledgments

The authors would like to thank two anonymous reviewers and the Editor-in-Chief, Pr. Christian Genest, for very valuable comments on previous versions of the manuscript that have improved the paper. Chernozhukov gratefully acknowledges support from NSF. Galichon's work has received support from NSF grant DMS-1716489, and ERC grants FP7-295298, FP7-312503, FP7-337665, and ANR grant Famineq.

References

[1] A. Belloni, R.L. Winkler, On multivariate quantiles under partial orders, Ann. Statist. 39 (2011) 1125–1179.
[2] Y. Brenier, Décomposition polaire et réarrangement monotone des champs de vecteurs, C. R. Acad. Sci. Paris I 305 (1987) 805–808.
[3] Y. Brenier, Polar factorization and monotone rearrangement of vector-valued functions, Comm. Pure Appl. Math. 44 (1991) 375–417.
[4] G. Carlier, V. Chernozhukov, A. Galichon, Vector quantile regression: an optimal transport approach, Ann. Statist. 44 (2016) 1165–1192.
[5] G. Carlier, A. Galichon, Exponential convergence for a convexifying equation, ESAIM Control Optim. Calc. Var. 18 (2012) 611–620.
[6] J.A. Cuesta-Albertos, C. Matrán, Notes on the Wasserstein metric in Hilbert spaces, Ann. Probab. 17 (1989) 1264–1276.
[7] I. Ekeland, A. Galichon, M. Henry, Comonotonic measures of multivariate risks, Math. Finance 22 (2012) 109–132.
[8] A. Galichon, Optimal Transport Methods in Economics, Princeton University Press, 2016.
[9] A. Galichon, M. Henry, Dual theory of choice with multivariate risks, J. Econom. Theory 47 (2012) 1501–1516.
[10] M. Hallin, D. Paindaveine, M. Šiman, Multivariate quantiles and multiple-output regression quantiles: From L^1 optimization to halfspace depth, Ann. Statist. 38 (2010) 635–669.
[11] M. Knott, C.S. Smith, On the optimal mapping of distributions, J. Optim. Theory Appl. 43 (1984) 39–49.
[12] R. Koenker, G. Bassett, Regression quantiles, Econometrica 46 (1978) 33–50.
[13] R. Koenker, K. Hallock, Quantile regression, J. Econ. Perspect. 15 (2001) 143–156.
[14] J. Komlos, A generalization of a problem of Steinhaus, Acta Math. Acad. Sci. Hungar. 18 (1967) 217–229.
[15] G. Puccetti, M. Scarsini, Multivariate comonotonicity, J. Multivariate Anal. 101 (2010) 291–304.
[16] L. Rüschendorf, Convergence of the iterative proportional fitting procedure, Ann. Statist. 23 (1995) 1160–1174.
[17] L. Rüschendorf, S.T. Rachev, A characterization of random variables with minimum L^2-distance, J. Multivariate Anal. 32 (1990) 48–54.
[18] L. Rüschendorf, S.T. Rachev, Mass Transportation Problems. Vol. I. Theory, Vol. II. Applications, Probability and its Applications, Springer, New York, 1998.
[19] F. Santambrogio, Optimal transport for applied mathematicians, in: Progress in Nonlinear Differential Equations and Their Applications 87, Brezis, H., Birkhäuser, Basel, 2015.
[20] C.S. Smith, M. Knott, Note on the optimal transportation of distributions, J. Optim. Theory Appl. 52 (1987) 323–329.
[21] C. Villani, Topics in Optimal Transportation, American Mathematical Society, Providence, RI, 2003.