Vector Quantile Regression: An Optimal Transport Approach

Guillaume Carlier, Victor Chernozhukov, Alfred Galichon

Submitted to the Annals of Statistics
arXiv: arXiv:1406.4643

VECTOR QUANTILE REGRESSION: AN OPTIMAL TRANSPORT APPROACH*

By Guillaume Carlier†, Victor Chernozhukov‡ and Alfred Galichon§

†*Université Paris IX Dauphine,* ‡*MIT, and* §*NYU and Sciences Po*

We propose a notion of conditional vector quantile function and a vector quantile regression. A *conditional vector quantile function* (CVQF) of a random vector Y, taking values in $\mathbb{R}^d$ given covariates $Z = z$, taking values in $\mathbb{R}^k$, is a map $u \longmapsto Q_{Y|Z}(u, z)$, which is monotone, in the sense of being a gradient of a convex function, and such that given that vector U follows a reference non-atomic distribution F_U, for instance uniform distribution on a unit cube in $\mathbb{R}^d$, the random vector $Q_{Y|Z}(U, z)$ has the distribution of Y conditional on $Z = z$. Moreover, we have a strong representation, $Y = Q_{Y|Z}(U, Z)$ almost surely, for some version of U. The *vector quantile regression* (VQR) is a linear model for CVQF of Y given Z. Under correct specification, the notion produces strong representation, $Y = \beta(U)^\top f(Z)$, for $f(Z)$ denoting a known set of transformations of Z, where $u \longmapsto \beta(u)^\top f(Z)$ is a monotone map, the gradient of a convex function, and the quantile regression coefficients $u \longmapsto \beta(u)$ have the interpretations analogous to that of the standard scalar quantile regression. As $f(Z)$ becomes a richer class of transformations of Z, the model becomes nonparametric, as in series modelling. A key property of VQR is the embedding of the classical Monge-Kantorovich's optimal transportation problem at its core as a special case. In the classical case, where Y is scalar, VQR reduces to a version of the classical QR, and CVQF reduces to the scalar conditional quantile function. An application to multiple Engel curve estimation is considered.

1. Introduction. Quantile regression provides a very convenient and powerful tool for studying dependence between random variables. The main object of modelling is the conditional quantile function (CQF) $(u, z) \longmapsto Q_{Y|Z}(u, z)$, which describes the u-quantile of the random scalar Y conditional on a k-dimensional vector of regressors Z taking a value z. Conditional

*First ArXiv submission: June, 2014. Carlier is partially supported by INRIA and the ANR for partial support through the Projects ISOTACE (ANR-12-MONU-0013) and OPTIFORM (ANR-12-BS01-0007). Chernozhukov is partially supported by an NSF grant. Galichon is partially supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement 313699.

MSC 2010 subject classifications: Primary 62J99; secondary 62H05, 62G05

Keywords and phrases: Vector quantile regression, vector conditional quantile function, Monge-Kantorovich-Brenier, LaTeX 2_ε

quantile function naturally leads to a strong representation via relation:

$$Y = Q_{Y|Z}(U, Z), \quad U \mid Z \sim U(0,1),$$

where U is the latent unobservable variable, normalized to have a uniform reference distribution, and is independent of regressors Z. The mapping $u \longmapsto Q_{Y|Z}(u, Z)$ is monotone, namely non-decreasing, almost surely.

Quantile regression (QR) is a means of modelling the conditional quantile function. A leading approach is linear in parameters, namely, it assumes that there exists a known $\mathbb{R}^p$-valued vector $f(Z)$, containing transformations of Z, and a ($p \times 1$ vector-valued) map of regression coefficients $u \longmapsto \beta(u)$ such that

$$Q_{Y|Z}(u \mid z) = \beta(u)^\top f(z),$$

for all z in the support of Z and for all quantile indices u in $(0,1)$. This representation highlights the vital ability of QR to capture differentiated effects of the explanatory variable Z on various conditional quantiles of the dependent variable Y (e.g., impact of prenatal smoking on infant birthweights). QR has found a large number of applications; see references in Koenker ([18])'s monograph. The model is flexible in the sense that, even if the model is not correctly specified, by using more and more suitable terms $f(Z)$ we can approximate the true CQF arbitrarily well. Moreover, coefficients $u \longmapsto \beta(u)$ can be estimated via tractable linear programming method ([20]).

The principal contribution of this paper is to extend these ideas to the cases of vector-valued Y, taking values in $\mathbb{R}^d$. Specifically, a vector conditional quantile function (CVQF) of a random vector Y, taking values in $\mathbb{R}^d$ given the covariates Z, taking values in $\mathbb{R}^k$, is a map $(u, z) \longmapsto Q_{Y|Z}(u, z)$, which is monotone with respect to u, in the sense of being a gradient of a convex function, which implies that

$$(Q_{Y|Z}(u,z) - Q_{Y|Z}(\overline{u},z))^\top (u - \overline{u}) \geq 0 \quad \text{for all } u, \overline{u} \in (0,1)^d, z \in \mathcal{Z}, \tag{1.1}$$

and such that the following strong representation holds with probability 1:

$$Y = Q_{Y|Z}(U, Z), \quad U \mid Z \sim U(0,1)^d, \tag{1.2}$$

where U is latent random vector uniformly distributed on $(0,1)^d$. We can also use other non-atomic reference distributions F_U on $\mathbb{R}^d$, for example, the standard normal distribution instead of uniform distribution (as we can in the canonical, scalar quantile regression case). We show that this map exists and is unique under mild conditions, as a consequence of Brenier's polar factorization theorem. This notion relies on a particular, yet very important, notion of monotonicity (1.1) for maps $\mathbb{R}^d \to \mathbb{R}^d$, which we adopt here.

We define vector quantile regression (VQR) as a model of CVQF, particularly a linear model. Specifically, under correct specification, our linear model takes the form:

$$Q_{Y|X}(u \mid z) = \beta(u)^\top f(z),$$

where $u \longmapsto \beta(u)^\top f(z)$ is a monotone map, in the sense of being a gradient of convex function; and $u \longmapsto \beta(u)$ is a map of regression coefficients from $(0,1)^d$ to the set of $p \times d$ matrices with real entries. This model is a natural analog of the classical QR for the scalar case. In particular, under correct specification, we have the strong representation

$$Y = \beta(U)^\top f(Z), \quad U \mid Z \sim U(0,1)^d, \tag{1.3}$$

where U is uniformly distributed on $(0,1)^d$ conditional on Z. (Other reference distributions could also be easily permitted.)

We provide a linear program for computing $u \longmapsto \beta(u)$ in population and finite samples. We shall stress that this formulation offers a number of useful properties. In particular, the linear programming problem admits a general formulation that embeds the optimal transportation problem of Monge-Kantorovich-Brenier, establishing a useful conceptual link to an important area of optimization and functional analysis (see, e.g. [34], [35]).

Our paper also connects to a number of interesting proposals for performing multivariate quantile regressions, which focus on inheriting certain (though not all) features of univariate quantile regression– for example, minimizing an asymmetric loss, ordering ideas, monotonicity, equivariance or other related properties, see, for example, some key proposals (including some for the non-regression case) in [6], [23], [32], [16], [24], [2], which are contrasted to our proposal in more details below. Note that it is not possible to reproduce all "desirable properties" of scalar quantile regression in higher dimensions, so various proposals focus on achieving different sets of properties. Our proposal is quite different from all of the excellent aforementioned proposals in that it targets to simultaneously reproduce two fundamentally different properties of quantile regression in higher dimensions – namely the deterministic coupling property (1.3) and the monotonicity property (1.1). This is the reason we deliberately don't use adjective "multivariate" in naming our method. By using a different name we emphasize the major differences of our method's goals from those of the other proposals. This also makes it clear that our work is complementary to other works in this direction. We discuss other connections as we present our main results.

1.1. *Plan of the paper.* We organize the rest of the paper as follows. In Section 2, we introduce and develop the properties of CVQF. In Section

3, we introduce and develop the properties of VQR as well its linear programming implementation. In Section 4, we provide computational details of the discretized form of the linear programming formulation, which is useful for practice and computation of VQR with finite samples. In Section 5, we implement VQR in an empirical example, providing the testing ground for these new concepts. We provide proofs of all formal results of the paper in the Appendix, and in the Supplementary material [4].

2. Conditional Vector Quantile Function.

2.1. *Conditional Vector Quantiles as Gradients of Convex Functions.* We consider a random vector (Y, Z) defined on a complete probability space $(\Omega_1, \mathcal{A}_1, \mathrm{P}_1)$. The random vector Y takes values in $\mathbb{R}^d$. The random vector Z is a vector covariate, taking values in $\mathbb{R}^k$. Denote by F_{YZ} the joint distribution function of (Y, Z), by $F_{Y|Z}$ the (regular) conditional distribution function of Y given Z, and by F_Z the distribution function Z. We also consider random vectors V defined on a complete probability space $(\Omega_0, \mathcal{A}_0, \mathrm{P}_0)$, which are required to have a fixed reference distribution function F_U. Let $(\Omega, \mathcal{A}, \mathrm{P})$ be the a *suitably enriched* complete probability space that can carry all vectors (Y, Z) and V with distributions F_{YZ} and F_U, respectively, as well as the independent (from all other variables) standard uniform random variable on the unit interval. Formally, this product space takes the form $(\Omega, \mathcal{A}, \mathrm{P}) = (\Omega_0, \mathcal{A}_0, \mathrm{P}_0) \times (S_1, \mathcal{A}_1, \mathrm{P}_1) \times ((0,1), B(0,1), \mathrm{Leb})$, where $((0,1), B(0,1), \mathrm{Leb})$ is the canonical probability space, consisting of the unit segment of the real line equipped with Borel sets and the Lebesgue measure. The symbols $\mathcal{Y}$, $\mathcal{Z}$, $\mathcal{U}$, $\mathcal{YZ}$, $\mathcal{UZ}$ denote the support of F_Y, F_Z, F_U, F_{YZ}, F_{UZ}, and $\mathcal{Y}_z$ denotes the support of $F_{Y|Z}(\cdot|z)$. We denote by $\|.\|$ the Euclidian norm of $\mathbb{R}^d$.

We assume that the following condition holds:

(N) *F_U has a density f_U with respect to the Lebesgue measure on $\mathbb{R}^d$ with a convex support set $\mathcal{U}$.*

The distribution F_U describes a reference distribution for a vector of latent variables U, taking values in $\mathbb{R}^d$, that we would like to link to Y via a strong representation of the form mentioned in the introduction. This vector will be one of many random vectors V having a distribution function F_U, but there will only be one $V = U$, in the sense specified below, that will provide the required strong representation. The leading cases for the reference distribution F_U include:

- the standard uniform distribution on the unit d-dimensional cube, $U(0,1)^d$,
- the standard normal distribution $N(0, I_d)$ over $\mathbb{R}^d$, or

- any other reference distribution on $\mathbb{R}^d$, e.g., uniform on a ball.

Our goal here is to create a deterministic mapping that transforms a random vector U with distribution F_U into Y such that Y conditional on Z has the conditional distribution $F_{Y|Z}$. Such a map that pushes forward a probability distribution of interest onto another one is called a *transport* between these distributions. That is, we want to have a strong representation property like (1.2) that we stated in the introduction. Moreover, we would like this transform to have *a monotonicity property*, as in the scalar case. Specifically, in the vector case we require this transform to be a *gradient of a convex function*, which is a plausible generalization of monotonicity from the scalar case. Indeed, in the scalar case the requirement that the transform is the gradient of a convex map reduces to the requirement that the transform is non-decreasing. We shall refer to the resulting transform as the *conditional vector quantile function* (CVQF). The following theorem shows that such map *exists* and is *uniquely determined* by the stated requirements.

THEOREM 2.1 (**CVQF as Conditional Brenier Maps**). *Suppose condition (N) holds.*

(i) There exists a measurable map $(u, z) \longmapsto Q_{Y|Z}(u, z)$ *from* $\mathcal{UZ}$ *to* $\mathbb{R}^d$*, such that for each* z *in* $\mathcal{Z}$*, the map* $u \longmapsto Q_{Y|Z}(u, z)$ *is the unique (*F_U*-almost everywhere) gradient of convex function such that, whenever* $V \sim F_U$*, the random vector* $Q_{Y|Z}(V, z)$ *has the distribution function* $F_{Y|Z}(\cdot, z)$*, that is,*

$$F_{Y|Z}(y, z) = \int 1\{Q_{Y|Z}(u, z) \leq y\} F_U(du), \quad \text{for all } y \in \mathbb{R}^d. \tag{2.1}$$

(ii) Moreover, there exists a random variable $V = U$ *such that* P*-almost surely*

$$Y = Q_{Y|Z}(U, Z), \text{ and } U \mid Z \sim F_U. \tag{2.2}$$

The theorem is our *first main result* that we announced in the introduction. It should be noted that the theorem does not require Y to have an absolutely continuous distribution, it holds for discrete and mixed outcome variables; only the reference distribution for the latent variable U is assumed to be absolutely continuous. It is also noteworthy that in the classical case of Y and U being *scalars* we recover the classical conditional quantile function as well as the strong representation formula based on this function ([29], [18]). Regarding the proof, the first assertion of the theorem is a consequence of fundamental results due to McCann ([25]) (as, e.g, stated in [34], Theorem 2.32) who in turn refined the fundamental results of [3]. These results were obtained in the case without conditioning. The second assertion is a consequence of Dudley-Philipp ([12]) result on abstract couplings in Polish spaces.

REMARK 2.1 (**Monotonicity**). The transform $(u,z) \longmapsto (Q_{Y|Z}(u,z),z)$ has the following monotonicity property:

$$(Q_{Y|Z}(u,z) - Q_{Y|Z}(\overline{u},z))^\top (u-\overline{u}) \geq 0 \quad \forall u,\overline{u} \in \mathcal{U}, \forall z \in \mathcal{Z}. \tag{2.3}$$

∎

REMARK 2.2 (**Uniqueness**). In part (i) of the theorem, $u \longmapsto Q_{Y|Z}(u,z)$ is equal to a gradient of some convex function $u \longmapsto \varphi(u,z)$ for F_U-almost every value of $u \in \mathcal{U}$ and it is unique in the sense that any other map with the same properties will agree with it F_U-almost everywhere. In general, the gradient $u \longmapsto \nabla_u \varphi(u,z)$ exists F_U-almost everywhere, and the set of points $\mathcal{U}_e$ where it does not is negligible. Hence the map $u \longmapsto Q_{Y|Z}(u,z)$ is still definable at each $u_e \in \mathcal{U}_e$ from the gradient values $\varphi(u,z)$ on $u \in \mathcal{U} \setminus \mathcal{U}_e$, by defining it at each u_e as a smallest-norm element of $\{v \in \mathbb{R}^d : \exists u_k \in \mathcal{U} \setminus \mathcal{U}_e : u_k \to u_e, \nabla_u \varphi(u_k,z) \to v\}$. ∎

Let us assume further that the following condition holds:

(C) *For each $z \in \mathcal{Z}$, the distribution $F_{Y|Z}(\cdot,z)$ admits a density $f_{Y|Z}(\cdot,z)$ with respect to the Lebesgue measure on $\mathbb{R}^d$.*

Under this condition we can recover U uniquely in the following sense:

THEOREM 2.2 (**Conditional Inverse Vector Quantiles or Conditional Vector Ranks**). *Suppose conditions (N) and (C) holds.*

Then there exists a measurable map $(y,z) \longmapsto Q^{-1}_{Y|Z}(y,z)$, mapping $\mathcal{Y}\mathcal{Z}$ to $\mathbb{R}^d$, such that for each z in $\mathcal{Z}$, the map $y \longmapsto Q^{-1}_{Y|Z}(y,z)$ is the inverse of $u \longmapsto Q_{Y|Z}(u,z)$ in the sense that:

$$Q^{-1}_{Y|Z}(Q_{Y|Z}(u,z),z) = u,$$

for almost all u under F_U. Furthermore, we can construct U in (2.2) as follows,

$$U = Q^{-1}_{Y|Z}(Y,Z), \text{ and } U \mid Z \sim F_U. \tag{2.4}$$

REMARK 2.3 (Conditional Vector Rank Function). The mapping $y \longmapsto Q^{-1}_{Y|Z}(y,z)$, which maps $\mathcal{Y} \subset \mathbb{R}^d$ to $\mathbb{R}^d$, is the conditional rank function. When $d = 1$, it coincides with the conditional distribution function, but when $d > 1$ it does not. The ranking interpretation stems from the fact that when we set $F_U = U(0,1)^d$, vector $Q^{-1}_{Y|Z}(Y,Z) \in [0,1]^d$ measures the centrality of observation Y for each of the dimensions, conditional on Z. ∎

It is also of interest to state a further implication, which occurs under (N) and (C), on the link between the transportation map $Q_{Y|Z}$ and its

derivatives on one side, and the densities f_U and $f_{Y|Z}$ on the other side. This link is a nonlinear second order partial differential equation called a (conditional) *Monge-Ampère equation*.

COROLLARY 2.1 (**Conditional Monge-Ampère Equations**). *Assume that conditions (N) and (C) hold and, further, that the map $u \longmapsto Q_{Y|Z}(u,z)$ is continuously differentiable and injective for each $z \in \mathcal{Z}$. Under this condition, the following conditional forward Monge-Ampère equation holds for all $(u,z) \in \mathcal{UZ}$:*

(2.5)
$$f_U(u) = f_{Y|Z}(Q_{Y|Z}(u,z),z)det[\mathrm{D}_u Q_{Y|Z}(u,z)] = \int \delta(u - Q^{-1}_{Y|Z}(y,z)) f_{Y|Z}(y,z)dy,$$

where δ is the Dirac delta function in $\mathbb{R}^d$ and $\mathrm{D}_u = \partial/\partial u^\top$. Reversing the roles of U and Y, we also have the following conditional backward Monge-Ampère equation holds for all $(y,z) \in \mathcal{YZ}$:

(2.6)
$$f_{Y|Z}(y,z) = f_U(Q^{-1}_{Y|Z}(y,z))det[\mathrm{D}_y Q^{-1}_{Y|Z}(y,z)] = \int \delta(y - Q_{Y|Z}(u,z)) f_U(u)du.$$

The latter expression is useful for linking the conditional density function to the conditional vector quantile function. Equations (2.5) and (2.6) are *partial differential equations* of the Monge-Ampère type, carrying an additional index $z \in \mathcal{Z}$. These equations could be used directly to solve for conditional vector quantiles given conditional densities. We can also use them to set up maximum likelihood method for recovering conditional vector quantiles. In the next section we describe a variational approach to recovering conditional vector quantiles.

2.2. *Conditional Vector Quantiles as Optimal Transport.* Under additional moment assumptions, the CVQF can be characterized and even defined as solutions to a regression version of the Monge-Kantorovich-Brenier's optimal transportation problem or, equivalently, a conditional correlation maximization problem.

We assume that the following conditions hold:

(M) The second moment of Y and the second moment of U are finite:

$$\int\int \|y\|^2 F_{YZ}(dy,dz) < \infty \text{ and } \int \|u\|^2 F_U(du) < \infty.$$

We consider the following optimal transportation problem with conditional independence constraints:

$$\min_V \{\mathrm{E}\|Y - V\|^2 : V \mid Z \sim F_U\}, \tag{2.7}$$

where the minimum is taken over all random vectors V defined on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Note that the value of objective is the Wasserstein distance between Y and V subject to $V \mid Z \sim F_U$. Under condition (M) we will see that a *solution* exists and is given by $V = U$ constructed in the previous section.

The problem (2.7) is the conditional version of the classical Monge - Kantorovich problem with Brenier's quadratic costs, which was solved by Brenier in considerable generality in the unconditional case. In the unconditional case, the canonical Monge problem is to transport a pile of coal with mass distributed across production locations from F_U into a pile of coal with mass distributed across consumption locations from F_Y, and it can be rewritten in terms of random variables V and Y. We are seeking to match Y with a version of V that is closest in mean squared sense subject to V having a prescribed distribution. Our conditional version above (2.7) imposes the additional conditional independence constraint $V \mid Z \sim F_U$.

The problem above is equivalent to covariance maximization problem subject to the prescribed conditional independence and distribution constraints:

$$\max_V \{\mathrm{E}(V^\top Y) : \quad V \mid Z \sim F_U\}, \tag{2.8}$$

where the maximum is taken over all random vectors V defined on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. This type of problem will be convenient for us, as it most directly connects to convex analysis and leads to a convenient dual program. This form also connects to unconditional multivariate quantile maps defined in [13], who employed them for purposes of risk analysis; our definition given in the previous section is more satisfactory, because it does not require any moment conditions, as follows from the results of [25].

The dual program to (2.8) can be stated as:

$$\begin{aligned} \min_{(\psi,\varphi)} \mathrm{E}(\varphi(V,Z) + \psi(Y,Z)) \quad &: \ \varphi(u,z) + \psi(y,z) \geq u^\top y \\ &\text{for all } (z,y,u) \in \mathcal{Z} \times \mathbb{R}^{2d}, \end{aligned} \tag{2.9}$$

where V is any vector such that $V \mid Z \sim F_U$, and minimization is performed over Borel maps $(y,z) \longmapsto \psi(y,z)$ from $\mathcal{Z} \times \mathbb{R}^d$ to $\mathbb{R} \cup \{+\infty\}$ and $(u,z) \longmapsto \varphi(z,u)$ from $\mathcal{Z} \times \mathbb{R}^d$ to $\mathbb{R} \cup \{+\infty\}$, where $y \longmapsto \psi(y,z)$ and $u \longmapsto \varphi(u,z)$ are lower-semicontinuous for each value $z \in \mathcal{Z}$.

THEOREM 2.3 (**Conditional Vector Quantiles as Optimal Transport**). *Suppose conditions (N), (C), and (M) hold.*

(i) There exists a pair of maps $(u,z) \longmapsto \varphi(u,z)$ and $(y,z) \longmapsto \psi(y,z) = \varphi^*(y,z)$, each mapping from $\mathbb{R}^d \times \mathcal{Z}$ to $\mathbb{R}$, that solve the problem (2.9). For each $z \in \mathcal{Z}$, the maps $u \longmapsto \varphi(u,z)$ and $y \longmapsto \varphi^*(y,z)$ are convex and are

Legendre transforms of each other:

$$\varphi(u,z) = \sup_{y\in\mathbb{R}^d}\{u^\top y - \varphi^*(y,z)\}, \quad \varphi^*(y,z) = \sup_{u\in\mathbb{R}^d}\{u^\top y - \varphi(u,z)\},$$

for all $(u,z) \in \mathcal{UZ}$ *and* $(y,z) \in \mathcal{YZ}$.

(iii) We can take the gradient $(u,z) \longmapsto \nabla_u\varphi(u,z)$ *of* $(u,z) \longmapsto \varphi(u,z)$ *as the conditional vector quantile function, namely, for each* $z \in \mathcal{Z}$, $Q_{Y|Z}(u,z) = \nabla_u\varphi(u,z)$ *for almost every value* u *under* F_U.

(iv) We can take the gradient $(y,z) \longmapsto \nabla_y\varphi^*(y,z)$ *of* $(y,z) \longmapsto \varphi^*(y,z)$ *as the conditional inverse vector quantile function or conditional vector rank function, namely, for each* $z \in \mathcal{Z}$, $Q^{-1}_{Y|Z}(y,z) = \nabla_y\varphi^*(z,y)$ *for almost every value* y *under* $F_{Y|Z}(\cdot,z)$.

(v) The vector $U = Q^{-1}_{Y|Z}(Y,Z)$ *is a solution to the primal problem (2.8) and is unique in the sense that any other solution* U^* *obeys* $U^* = U$ *almost surely under* P. *The primal (2.8) and dual (2.9) have the same value.*

(vi) The maps $u \longmapsto \nabla_u\varphi(u,z)$ *and* $y \longmapsto \nabla\varphi^*_y(y,z)$ *are inverses of each other: for each* $z \in \mathcal{Z}$, *and for almost every* u *under* F_U *and almost every* y *under* $F_{Y|Z}(\cdot,z)$

$$\nabla_y\varphi^*(\nabla_u\varphi(u,z),z) = u, \quad \nabla_u\varphi(\nabla_y\varphi^*(y,z),z) = y.$$

REMARK 2.4. There are many maps $Q : \mathcal{UZ} \to \mathcal{Y}$ such that if $V \sim F_U$, then $Q(V,z) \sim F_{Y|Z=z}$. Any of these maps define a *transport* from F_U to $F_{Y|Z=z}$. Our choice is to take the *optimal transport*, in the sense that it minimizes the Wasserstein distance $\mathrm{E}\|Q(V,Z) - V\|^2$ among such maps. This has several benefits: (i) the optimal transport is unique as soon as F_U is absolutely continuous, as noted in Remark 2.2 and (ii) this object is easily computable through a linear programming problem. Note that the classical, scalar quantile map is the optimal transport from F_U to F_Y in this sense, so oue notion indeed extends the classical notion of a quantile. ∎

REMARK 2.5. Unlike in the scalar case, we cannot compute $Q_{Y|Z}(u,z)$ at a given point u without computing the whole map $u \to Q_{Y|Z}(u,z)$. This highlights the fact that CVQF is not a local concept with respect to values of the rank u. ∎

Theorem 2.3 provides a number of analytical properties, formalizing the variational interpretation of conditional vector quantiles, providing the potential functions $(u,z) \longmapsto \varphi(u,z)$ and $(y,z) \longmapsto \varphi^*(y,z)$, which are mutual Legendre transforms, and whose gradients are the conditional vector quantile

functions and its inverse, the conditional vector rank function. This problem is a conditional generalization of the fundamental results by Brenier as presented in [34], Theorem 2.12.

EXAMPLE 2.1 (**Conditional Normal Vector Quantiles**). Here we consider the normal conditional vector quantiles. Consider the case where

$$Y \mid Z \sim N(\mu(Z), \Omega(Z)).$$

Here $z \longmapsto \mu(z)$ is the conditional mean function and $z \longmapsto \Omega(z)$ is a conditional variance function such that $\Omega(z) > 0$ (in the sense of positive definite matrices) for each $z \in \mathcal{Z}$ with $\mathrm{E}\|\Omega(Z)\| + \mathrm{E}\|\mu(Z)\|^2 < \infty$. The reference distribution is given by $U \mid Z \sim N(0, I)$. Then we have the following conditional vector quantile model:

$$\begin{aligned} Y &= \mu(Z) + \Omega^{1/2}(Z)U, \\ U &= \Omega^{-1/2}(Z)(Y - \mu(Z)). \end{aligned}$$

Here we have the following conditional potential functions

$$\begin{aligned} \varphi(u, z) &= \mu(z)^\top u + \frac{1}{2} u^\top \Omega^{1/2}(z) u, \\ \psi(y, z) &= \frac{1}{2}(y - \mu(z))^\top \Omega^{-1/2}(z)(y - \mu(z)), \end{aligned}$$

and the following conditional vector quantile and rank functions:

$$\begin{aligned} Q_{Y|Z}(u, z) &= \nabla_u \varphi(u, z) = \mu(z) + \Omega^{1/2}(z) u, \\ Q^{-1}_{Y|Z}(y, z) &= \nabla_y \psi(y, z) = \Omega^{-1/2}(z)(y - \mu(z)). \end{aligned}$$

It follows from Theorem 2.3 that $V = U$ solves the covariance maximization problem (2.8). This example is special in the sense that the conditional vector quantile and rank functions are *linear in* u and y, respectively. ∎

2.3. *Interpretations of vector rank* U. We can provide the following interpretations of U:

1) As multivariate rank. An interesting interpretation of U is as a multivariate rank. In the univariate case, [18], Ch. 1.3 and 3.5, interprets U as a continuous notion of rank in the setting of quantile regression. The rank has a reference distribution F_U, which is typically chosen to be uniform on $(0, 1)$, but other reference distributions could be used as well. The concept of vector quantile allows us to assign a continuous rank to each of the dimensions, and the vector quantile mapping is monotone with respect to the

rank in the sense of being the gradient of a convex function. As a result, U can be interpreted as a *multivariate rank* for Y, as we are trying to map the distribution of U to a prescribed distribution F_Y at minimal distortion, as seen in (2.7).

2) As a reference outcome for defining quantile treatment effects. Another motivation is related to the classical definition of quantile treatment effects introduced by [27], and further developed by [10], [18], and others. Suppose we define U as an outcome for an untreated population; for this we simply set the reference distribution F_U to the distribution of outcome in the untreated population. Suppose Z is the indicator of the receiving a treatment ($Z = 0$ means no treatment). Then we can represent outcome $Y = Q_{Y|Z}(U, Z)$ as the multivariate health outcome conditional on Z. If $Z = 0$, then the outcome is distributed as $Q_{Y|Z}(U, 0) = U$. If $Z = 1$, then the outcome is distributed as $Q_{Y|Z}(U, 1)$. The corresponding notion of vector quantile treatment effects is $Q_{Y|Z}(u, 1) - Q_{Y|Z}(u, 0)$.

3) As nonlinear latent factors. As it is apparent in the variational formulation (2.7), the entries of U can also be thought as latent factors, independent of each other and explanatory variables Z and having a prescribed marginal distribution F_U, and that best explain the variation in Y. Therefore, the conditional vector quantile model (2.2) provides a non-linear latent factor model for Y with factors U solving the matching problem (2.7). This interpretation suggests that this model may be useful in applications which require measurement of multidimensional unobserved factors, for example, cognitive ability, persistence, and various other latent propensities; see, for example, [8].

2.4. *Overview of Other Notions of Multivariate Quantile.* We briefly review other notions of multivariate quantiles in the statistical literature. We highlight the main contrasts with the notion we are using, based on optimal transport. For the sake of clarity of exposition, we discuss the unconditional case; albeit the comparisons extend naturally to the regression case.

In [6], the following definition of multivariate quantile function is suggested: for $u \in \mathbb{R}^d$, let

$$Q_Y^C(u) = \arg\max_{y \in \mathbb{R}^d} \mathrm{E}\left[y^\top u - \|y - Y\|\right]$$

which coincides with the classical notion when $d = 1$. See also [32]. More generally, [23] offers the following definition based on M-estimators, still for $u \in \mathbb{R}^d$,

$$Q_Y^K(u) = \arg\max_{y \in \mathbb{R}^d} \mathrm{E}\left[y^\top u - K(y, Y)\right]$$

for a choice of kernel K assumed to be convex with respect to its first argument. Like our proposal, these notions of quantile maps are gradients of convex potentials. However, unlike our proposal, these notions do not provide a transport from a fixed distribution over values of u to the distribution F_Y of Y as soon as $d > 1$.

In [36], a notion of quantile based on the Rosenblatt map is investigated. In the case $d = 2$, this quantile is defined for $u \in [0,1]^2$ as

$$Q_Y^R(u_1, u_2) = \left(Q_{Y_1}(u_1), Q_{Y_2|Y_1}(u_2 \mid Q_{Y_1}(u_1))\right)$$

where Q_{Y_1} and $Q_{Y_1|Y_2}$ are the univariate and the conditional univariate quantile map. This map is a transport of the distribution of $\mathcal{U}(0,1)^2$; however, in this definition, Y_1 and Y_2 play sharply assymetric roles, as the second dimension is defined conditional on the first one. Unlike ours, this quantile map is not a gradient of convex function.

In [16], the authors specify a vector of latent indices $u \in B^d$ the unit ball of $\mathbb{R}^d$. For $u \in B^d$, they define multivariate quantiles as

$$Q_Y^{HPS}(u) = \left\{y \in \mathbb{R}^d : c^\top y = a\right\},$$

where $a \in \mathbb{R}$ and $c \in \mathbb{R}^d$ minimize $\mathrm{E}\rho_{\|u\|}\left(c^\top Y - a\right)$ subject to constraint $c^\top u = \|u\|$. In contrast to ours, their notion of quantile is a set-valued. A closely related construction is provided by [24] who define the directional quantile associated to the index $u \in \mathbf{B}^d$ via:

$$Q_Y^{KM}(u) = Q_{u^\top Y/\|u\|}(\|u\|)\, u/\|u\|,$$

where $Q_{u^\top Y/\|u\|}$ is the univariate quantile function of the random variable $u^\top Y/\|u\|$. We can provide a transport interpretation to this notion of quantiles, but unlike our proposal this map is not a gradient of convex function.

A notion of quantile based on a partial order $\preceq$ on $\mathbb{R}^d$ is proposed in [2]. For an index $u \in (0,1)$, these authors define

$$Q_Y^{BW}(u) = \left\{y \in \mathbb{R}^d : \Pr(Y \succeq y \mid C(y)) \geq 1-u, \Pr(Y \preceq y \mid C(y)) \geq u\right\}$$

where $C(y) = \left\{y' \in \mathbb{R}^d : y \succeq y' \text{ or } y' \succeq y\right\}$ is the set of elements that can be ordered by $\succeq$ relative to the point y. Unlike our proposal, the index u is scalar and the quantile is set-valued.

3. Vector Quantile Regression.

3.1. *Linear Formulation.* Here we let $X = f(Z)$ denote a vector of regressors formed as transformations of Z, such that the first component of X is 1 (intercept term in the model) and such that conditioning on X is equivalent to conditioning on Z. The dimension of X is denoted by p and we shall denote $X = (1, X_{-1}^\top)^\top$ with $X_{-1} \in \mathbb{R}^{p-1}$.

In practice, X would often consist of a constant and some polynomial or spline transformations of Z as well as their interactions. Note that conditioning on X is equivalent to conditioning on Z if, for example, a component of X contains a one-to-one transform of Z.

Denote by F_X the distribution function of X and $F_{UX} = F_U F_X$. Let $\mathcal{X}$ denote the support of F_X and $\mathcal{UX}$ the support of F_{UX}. We define *linear vector quantile regression model* (VQRM) as the following linear model of CVQF.

(L) The following linearity condition holds:

$$Y = Q_{Y|X}(U, X) = \beta_0(U)^\top X, \quad U \mid X \sim F_U,$$

where $u \longmapsto \beta_0(u)$ is a map from $\mathcal{U}$ to the set $\mathcal{M}_{p\times d}$ of $p \times d$ matrices such that $u \longmapsto \beta_0(u)^\top x$ is a monotone, smooth map, in the sense of being a gradient of a convex function:

$$\beta_0(u)^\top x = \nabla_u \Phi_x(u), \quad \Phi_x(u) := B_0(u)^\top x, \text{ for all } (u, x) \in \mathcal{UX},$$

where $u \longmapsto B_0(u)$ is C^1 map from $\mathcal{U}$ to $\mathbb{R}^d$, and $u \longmapsto B_0(u)^\top x$ is a strictly convex map from $\mathcal{U}$ to $\mathbb{R}$.

The parameter $\beta(u)$ is indexed by the quantile index $u \in \mathcal{U}$ and is a $d \times p$ *matrix* of quantile regression coefficients. Of course in the scalar case, when $d = 1$, this matrix reduces to a vector of quantile regression coefficients. This model is a natural analog of the classical QR for scalar Y where the similar regression representation holds. One example where condition (L) holds is Example 2.1, describing the conditional normal vector regression. It is of interest to specify other examples where condition (L) holds or provides a plausible approximation.

Example 3.1 (Saturated Specification). The regressors $X = f(Z)$ with $\mathrm{E}\|f(Z)\|^2 < \infty$ are *saturated* with respect to Z, if, for any $g \in L^2(F_Z)$, we have $g(Z) = X^\top \alpha_g$. In this case the linear functional form (L) is not a restriction. For $p < \infty$ this can occur if and only if Z takes on a finite set of values $\mathcal{Z} = \{z_1, \ldots, z_p\}$, in which case we can write:

$$Q_{Y|X}(u, X) = \sum_{j=1}^{p} Q_{Y|Z}(u, z_j) 1(Z = z_j) =: B_0(u)^\top X,$$

$$B_0(u) := \begin{pmatrix} Q_{Y|Z}(u,z_1)^\top \\ \vdots \\ Q_{Y|Z}(u,z_p)^\top \end{pmatrix}, \quad X := \begin{pmatrix} 1(Z=z_1) \\ \vdots \\ 1(Z=z_p) \end{pmatrix}.$$

Here the problem is equivalent to considering p unconditional vector quantiles in populations corresponding to $Z = z_1, \ldots, Z = z_p$. ∎

The rationale for using linear forms is two-fold – one is convenience of estimation and representation of functions and another one is approximation property. We can approximate a smooth convex potential by a smooth linear potential, as the following example illustrates for a particular approximation method.

EXAMPLE 3.2 (Linear Approximation). Let $(u,z) \longmapsto \varphi(u,z)$ be of class C^a with $a > 1$ on the support $(u,z) \in \mathcal{UZ} = [0,1]^{d+k}$. Consider a trigonometric tensor product basis of functions $\{(u,z) \longmapsto q_j(u)f_l(z), j \in \mathbb{N}, l \in \mathbb{N}\}$ in $L^2[0,1]^{d+k}$. Then there exists a JL vector $(\gamma_{jl} : j \in \{1,...,J\}, l \in \{1,...,L\})$ such that the linear map:

$$(u,z) \longmapsto \Phi^{JL}(u,z) := \sum_{j=1}^{J}\sum_{l=1}^{L} \gamma_{jl} q_j(u) f_l(z) =: B_0^L(u)^\top f^L(z),$$

where $B_0^L(u) = (\sum_{j=1}^J \gamma_{jl} q_j(u), l \in \{1,...,L\})$ and $f^L(z) = (f_l(z), l \in \{1,...,L\})$, provides uniformly consistent approximation of the potential and its derivative:

$$\lim_{J,L\to\infty} \sup_{(u,z)\in\mathcal{UZ}} \Big(|\varphi(u,z) - \Phi^{JL}(u,z)| + \|\nabla_u \varphi(u,z) - \nabla_u \Phi^{JL}(u,z)\| \Big) = 0. \quad ∎$$

The approximation property via the sieve-type approach provides a rationale for the linear (in parameters) specification (1.3). Another approach, based on local polynomial approximations over a collection of (increasingly smaller) neighborhoods, also provides a useful rationale for the linear (in parameters) specification, e.g., similarly in spirit to [37]. If the linear specification does not hold *exactly* we say that the model is *misspecified*. If the model is flexible enough, by using a suitable basis or localization, then the approximation error is small, and we effectively ignore the error when assuming (1.3). However, when constructing a sensible estimator we must allow the possibility that the model is misspecified, which means we can't really force (1.3) onto data. Our proposal for estimation presented next does not force (1.3) onto data, but if (1.3) is true in population, then as a result, the true conditional vector quantile function would be recovered perfectly in population.

3.2. *Linear Program for VQR.* Our approach to multivariate quantile regression is based on the multivariate extension of the covariance maximization problem with a mean independence constraint:

$$\max_{V}\{\mathrm{E}(V^\top Y):\ V\sim F_U,\ \mathrm{E}(X\mid V)=\mathrm{E}(X)\}. \tag{3.1}$$

Note that the constraint condition is a relaxed form of the previous independence condition.

REMARK 3.1. The new condition $V\sim F_U,\ \mathrm{E}(X\mid V)=\mathrm{E}(X)$ is weaker than $V\mid X\sim F_U$, but the two conditions coincide if X is saturated relative to Z, as in Example 3.1, in which case $\mathrm{E}(g(Z)V)=\mathrm{E}X^\top\alpha_g V=\mathrm{E}(X^\top\alpha_g)\mathrm{E}(V)=\mathrm{E}g(Z)\mathrm{E}V$ for every $g\in L^2(F_Z)$. More generally, this example suggests that the richer X is, the closer the mean independence condition becomes to the conditional independence. ∎

The relaxed condition is sufficient to guarantee that the solution exists not only when (L) holds, but more generally when the following quasi-linear assumption holds.

(QL) We have a quasi-linear representation a.s.

$$Y=\beta(\widetilde{U})^\top X,\quad \widetilde{U}\sim F_U,\quad \mathrm{E}(X\mid\widetilde{U})=\mathrm{E}(X),$$

where $u\longmapsto\beta(u)$ is a map from $\mathcal{U}$ to the set $\mathcal{M}_{p\times d}$ of $p\times d$ matrices such that $u\longmapsto\beta(u)^\top x$ is a gradient of convex function for each $x\in\mathcal{X}$ and a.e. $u\in\mathcal{U}$:

$$\beta(u)^\top x=\nabla_u\Phi_x(u),\quad \Phi_x(u):=B(u)^\top x,$$

where $u\longmapsto B(u)$ is C^1 map from $\mathcal{U}$ to $\mathbb{R}^d$, and $u\longmapsto B(u)^\top x$ is a strictly convex map from $\mathcal{U}$ to $\mathbb{R}$.

This condition allows for a degree of misspecification, which allows for a latent factor representation where the latent factor obeys the relaxed independence constraints.

THEOREM 3.1. *Suppose conditions (M), (N) , (C), and (QL) hold.*

(i) The random vector $\widetilde{U}$ entering the quasi-linear representation (QL) solves (3.1).

(ii) The quasi-linear representation is unique a.s. that is if we also have $Y=\overline{\beta}(\overline{U})^\top X$ with $\overline{U}\sim F_U,\mathrm{E}(X\mid\overline{U})=\mathrm{E}X$, $u\longmapsto X^\top\overline{\beta}(u)$ is a gradient of a strictly convex function in $u\in\mathcal{U}$ a.s., then $\overline{U}=\widetilde{U}$ and $X^\top\beta(\widetilde{U})=X^\top\overline{\beta}(\widetilde{U})$ a.s.

(iii) Under condition (L) and assuming that $\mathrm{E}(XX^\top)$ has full rank, $\widetilde{U}=U$ a.s. and U solves (3.1). Moreover, $\beta_0(U)=\beta(U)$ a.s.

The last assertion is important – it says that if (L) holds, then the linear program (3.1), where the independence constraint has been relaxed into a *mean independence* constraint, will find the true linear vector quantile regression in the population.

3.3. *Dual Program for Linear VQR.* As explained in details in the appendix, Program (3.1) is an infinite-dimensional linear programming problem whose dual program is:

$$\inf_{(\psi,b)} \ \mathrm{E}\psi(X,Y) + \mathrm{E}b(V)^\top \mathrm{E}(X) \ : \quad \begin{array}{l} \psi(x,y) + b(u)^\top x \geq u^\top y, \\ \forall\ (y,x,u) \in \mathcal{Y}\mathcal{X}\mathcal{U}, \end{array} \tag{3.2}$$

where $V \sim F_U$, where the infimum is taken over all continuous functions $(y,x) \longmapsto \psi(y,x)$, mapping $\mathcal{Y}\mathcal{X}$ to $\mathbb{R}$ and $u \longmapsto b(u)$ mapping $\mathcal{U}$ to $\mathbb{R}$, such that $\mathrm{E}\psi(X,Y)$ and $\mathrm{E}b(V)$ are finite.

Since for fixed b, the smallest ψ which satisfies the pointwise constraint in (3.2) is given by

$$\psi(x,y) := \sup_{u\in\mathcal{U}} \{u^\top y - b(u)^\top x\},$$

one may equivalently rewrite (3.2) as the minimization over continuous b of

$$\int \sup_{u\in\mathcal{U}} \{u^\top y - b(u)^\top x\} F_{Y|X}(dx,dy) + \int b(u)^\top \mathrm{E}(X) F_U(du).$$

By standard arguments ([34], section 1.1.7), the infimum over continuous functions coincides with the one over smooth or simply integrable functions.

THEOREM 3.2. *Under (M) and (QL), we have that the optimal solution to the dual is given by functions:*

$$\psi(x,y) = \sup_{u\in\mathcal{U}} \{u^\top y - B(u)^\top x\}, \quad b(u) = B(u).$$

This result can be recognized as a consequence of strong duality of the linear programming (e.g. [34]).

3.4. *Connecting to Scalar Quantile Regression.* We now consider the connection to the canonical, scalar quantile regression primal problem, where Y is scalar and for each probability index $t \in (0,1)$, the linear functional form $x \longmapsto x^\top \beta(t)$ is used. [20] define linear quantile regression as $X^\top \beta(t)$ with $\beta(t)$ solving the minimization problem

$$\beta(t) \in \arg\min_{\beta\in\mathbb{R}^p} \mathrm{E}\rho_t(Y - X^\top \beta), \tag{3.3}$$

where $\rho_t(z) := tz_- + (1-t)z_+$, with z_- and z_+ denoting the negative and positive parts of z. The above formulation makes sense and $\beta(t)$ is unique under the following simplified conditions:

(QR) $\mathrm{E}|Y| < \infty$ and $\mathrm{E}\|X\|^2 < \infty$, $(y,x) \longmapsto f_{Y|X}(y,x)$ is bounded and uniformly continuous, and $\mathrm{E}(f_{Y|X}(X^\top\beta(t),X)XX^\top)$ is of full rank.

We note that (3.3) can be conveniently rewritten as

$$\min_{\beta\in\mathbb{R}^p}\{\mathrm{E}(Y - X^\top\beta)_+ + (1-t)\mathrm{E}X^\top\beta\}. \tag{3.4}$$

[20] showed that this convex program admits as *dual formulation*:

$$\max\{\mathrm{E}(A_tY) \ : \ A_t \in [0,1],\ \mathrm{E}(A_tX) = (1-t)\mathrm{E}X\}. \tag{3.5}$$

An optimal $\beta = \beta(t)$ for (3.4) and an optimal rank-score variable A_t in (3.5) may be taken to be

$$A_t = 1(Y > X^\top\beta(t)), \tag{3.6}$$

and thus the constraint $\mathrm{E}(A_tX) = (1-t)\mathrm{E}X$ reads:

$$\mathrm{E}(1(Y > X^\top\beta(t))X) = (1-t)\mathrm{E}X, \tag{3.7}$$

which simply is the first-order optimality condition for (3.4).

We say that the specification of quantile regression is quasi-linear if

$$t \longmapsto x^\top\beta(t) \text{ is increasing on } (0,1). \tag{3.8}$$

Define the rank variable $\widetilde{U} = \int_0^1 A_t dt$, then under (3.8) we have that

$$A_t = 1(\widetilde{U} > t),$$

and the first-order condition (3.7) implies that for each $t \in (0,1)$

$$\mathrm{E}1(\widetilde{U} \geq t) = (1-t), \quad \mathrm{E}1(\widetilde{U} \geq t)X = (1-t)\mathrm{E}X.$$

The first property implies that $\widetilde{U} \sim U(0,1)$ and the second property can be easily shown to imply the *mean-independence* condition:

$$\mathrm{E}(X \mid \widetilde{U}) = \mathrm{E}X.$$

Thus quantile regression *naturally leads* to the mean-independence condition and the quasi-linear latent factor model. This is the reason we used mean-independence condition as a starting point in formulating the vector quantile regression. Moreover, in both vector and scalar cases, we have that, when the conditional quantile function is linear (not just quasi-linear), the quasi-linear representation coincides with the linear representation and $\widetilde{U}$ becomes fully independent of X.

The following result summarizes the connection more formally.

THEOREM 3.3 (**Connection to Scalar QR**). *Suppose that (QR) holds.*
(i) If (3.8) holds, then for $\widetilde{U} = \int_0^1 A_t dt$ *we have the quasi-linear model holding*

$$Y = X^\top \beta(\widetilde{U}) \text{ a.s., } \widetilde{U} \sim U(0,1) \text{ and } \mathrm{E}(X \mid \widetilde{U}) = \mathrm{E}(X).$$

Moreover, $\widetilde{U}$ *solves the problem of correlation maximization problem with a mean independence constraint:*

$$\max\{\mathrm{E}(VY) : V \sim U(0,1),\ \mathrm{E}(X \mid V) = \mathrm{E}(X)\}. \tag{3.9}$$

(ii) The quasi-linear representation above is unique almost surely. That is, if we also have $Y = \overline{\beta}(\overline{U})^\top X$ *with* $\overline{U} \sim U(0,1), \mathrm{E}(X \mid \overline{U}) = \mathrm{E}X$, $u \longmapsto X^\top \overline{\beta}(u)$ *is increasing in* $u \in (0,1)$ *a.s., then* $\widetilde{U} = \overline{U}$ *and* $X^\top \beta(\widetilde{U}) = X^\top \overline{\beta}(\overline{U})$ *a.s.*
(iii) Consequently, if the conditional quantile function is linear, namely $Q_{Y|X}(u) = X^\top \beta_0(u)$, *so that* $Y = X^\top \beta_0(U)$, *then the latent factors in the quasi-linear and linear specifications coincide, namely* $U = \widetilde{U}$, *and so do the model coefficients, namely* $\beta_0(U) = \beta(U)$.

4. Implementation of Vector Quantile Regression. In order to implement VQR in practice, we employ discretization of the problem, namely we approximate the distribution F_{YX} of the outcome-regressor vector (Y, X) and F_U of the vector rank U by discrete distributions ν and μ, respectively. For example, for estimation purposes we can approximate F_{YX} by an empirical distribution of the sample, and the distribution F_U of U by a finite grid.

Let $y_i \in \mathbb{R}^d$ denote values of outcomes and $x_i \in \mathbb{R}^p$ of regressors for $1 \le i \le n$; we assume the first component of x_i is 1. For estimation purposes, we assume these values are obtained as a random sample from distribution F_{YX}, and so each observation receives a point mass $\nu_i = 1/n$. When we perform computation for theoretical purposes, we can think of these values as grid points, which are not necessarily obtained as a random sample, and so each observation receives a point mass ν_i which does not have to be $1/n$. We also set up a collection of grid points u_k, for $k = 1, ..., m$, for values of the vector rank U, and assign the probability mass μ_k to each of the point. For example, if $U \sim U(0,1)^d$ and we generate values u_k as a random sample or via a uniformly spaced grid of points, then $\mu_k = 1/m$.

Thus, let Y be the $n \times d$ matrix with row vectors $y_j^\top$ and X the $n \times r$ matrix of row vectors $x_j^\top$; the first column of this matrix is a vector of ones. Let ν be a $n \times 1$ matrix such that ν_i is the probability attached to a value (x_i, y_i), so that $\nu_i \ge 0$ and $\sum_{i=1}^n \nu_i = 1$. Let m be the number of points in the support of μ. Let U be a $m \times d$ matrix, where the ith row denoted by

$u_i^\top$. Let μ be a $m \times 1$ matrix such that μ_k is the probability weight of u_k (hence $\mu_i \geq 0$ and $\sum_k \mu_k = 1$).

We are looking to find an $m \times n$ matrix π such that π_{ij} is the probability mass attached to (u_i, x_j, y_j) which maximizes

$$\sum_{ij} \pi_{ij} y_j^\top u_i = \mathrm{Tr}(\mathsf{U}^\top \pi \mathsf{Y})$$

subject to constraint $\pi^\top 1_m = \nu$, where 1_m is a $m \times 1$ vector of ones, and subject to constraints $\pi 1_n = \mu$ and $\pi \mathsf{X} = \mu \nu^\top \mathsf{X}$.

Hence, the discretized VQR program is given in its primal form by

$$\max_{\pi \geq 0} \mathrm{Tr}\left(\mathsf{U}^\top \pi \mathsf{Y}\right) : \quad \pi^\top 1_m = \nu \ [\psi] \quad \pi \mathsf{X} = \mu \nu^\top \mathsf{X} \ [b]\,, \tag{4.1}$$

where the square brackets show the associated Lagrange multipliers, and in its dual form by

$$\min_{\psi, b} \psi^\top \nu + \nu^\top \mathsf{X} b^\top \mu : \quad \psi 1_m^\top + \mathsf{X} b^\top \geq \mathsf{Y} \mathsf{U}^\top \ \left[\pi^\top\right], \tag{4.2}$$

where ψ is a $n \times 1$ vector, and b is a $m \times r$ matrix.

Problems (4.1) and (4.2) are two linear programming problems dual to each other. However, in order to implement them on standard numerical analysis software such as R or Matlab coupled with a linear programming software such as Gurobi, we need to convert matrices into vectors. This is done using the vec operation, which is such that if A is a $p \times q$ matrix, $\mathrm{vec}(A)$ is a column vector of size pq such that $\mathrm{vec}\,(A)_{i+p(j-1)} = A_{ij}$. The use of the Kronecker product is also helpful. Recall that if A is a $p \times q$ matrix and B is a $p' \times q'$ matrix, then the Kronecker product $A \otimes B$ is the $pp' \times qq'$ matrix such that for all relevant choices of indices i, j, k, l, $(A \otimes B)_{i+p(k-1), j+q(l-1)} = A_{ij} B_{kl}$. The fundamental property linking Kronecker products and the vec operator is $\mathrm{vec}\left(BXA^T\right) = (A \otimes B)\,\mathrm{vec}\,(X)$.

Introduce $\mathrm{vec}\pi = \mathrm{vec}\,(\pi)$, the optimization variable of the "vectorized problem". Note that the variable $\mathrm{vec}\pi$ is a mn-vector. Then we rewrite the objective function, $\mathrm{Tr}\left(\mathsf{U}^\top \pi \mathsf{Y}\right) = \mathrm{vec}\pi^\top \mathrm{vec}\left(\mathsf{U}\mathsf{Y}^\top\right)$; as for the constraints, $\mathrm{vec}\left(1_m^\top \pi\right) = \left(I_n \otimes 1_m^\top\right) \mathrm{vec}\pi$ is a $n \times 1$ vector; and $\mathrm{vec}\,(\pi \mathsf{X}) = \left(\mathsf{X}^\top \otimes I_m\right) \mathrm{vec}\pi$ is a mr-vector. Thus we can rewrite the program (4.1) as:

$$\begin{aligned} &\max_{\mathrm{vec}\pi \geq 0} \mathrm{vec}\left(\mathsf{U}\mathsf{Y}^\top\right)^\top \mathrm{vec}\pi : \\ &\left(I_n \otimes 1_m^\top\right) \mathrm{vec}\pi = \mathrm{vec}\,(\nu) \\ &\left(\mathsf{X}^\top \otimes I_m\right) \mathrm{vec}\pi = \mathrm{vec}\left(\mu \nu^\top \mathsf{X}\right) \end{aligned} \tag{4.3}$$

which is a LP problem with mn variables and $mr + n$ constraints. The constraints $\left(I_n \otimes 1_m^\top\right)$ and $\left(\mathsf{X}^\top \otimes I_m\right)$ are very sparse, which can be taken advantage of from a computational point of view.

5. Empirical Illustration. We demonstrate the use of the approach on a classical application of Quantile Regression since [21]: Engel's ([15]) data on household expenditures, including 199 Belgian working-class households surveyed by Ducpetiaux ([11]), and 36 observations from all over Europe surveyed by Le Play ([28]). Due to the univariate nature of classical QR, [21] limited their focus on the regression of food expenditure over total income. But in fact, Engel's dataset is richer and classifies household expenses in nine broad categories: 1. Food; 2. Clothing; 3. Housing; 4. Heating and lighting; 5. Tools; 6. Education; 7. Safety; 8. Medical care; and 9. Services. This allows us to have a multivariate dependent variable. While we could in principle have $d = 9$, we focus for illustrative purposes on a two-dimensional dependent variable ($d = 2$), and we choose to take Y_1 as food expenditure (category #1) and Y_2 as housing and domestic fuel expenditure (category #2 plus category #4). We take $X = (X_1, X_2)$ with $X_1 = 1$ and X_2= the total expenditure (income) as an explanatory variable.

5.1. *One-dimensional VQR.* To begin with, we run a pair of one dimensional VQRs, where we regress Y_1 on X, and Y_2 on X. We plot the results in Figure 1; the curves drawn here are $u \to \beta(u)^\top x$ for five percentiles of the income x_{-1} (0%, 25%, 50%, 75%, 100%), and the corresponding probabilistic representations are

$$Y_1 = \beta_1 (U_1)^\top X \text{ and } Y_2 = \beta_2 (U_2)^\top X \tag{5.1}$$

with $U_1 \mid X \sim \mathcal{U}([0,1])$ and $U_2 \mid X \sim \mathcal{U}([0,1])$. Here, U_1 is interpreted as a propensity to consume food, while U_2 is interpreted as a propensity to consume the housing good. Note that in general, U_1 and U_2 are not independent; in other words, the distribution of (U_1, U_2) differs from $\mathcal{U}([0,1]^2)$. In fact, the distribution of (U_1, U_2) is called the *copula* associated to the conditional distribution of (Y_1, Y_2) conditional on X.

As explained above, when $d = 1$, VQR is very closely connected to classical quantile regression. Hence, in Figure 1, we also draw the classical quantile regression (in red). In each case, the curves exhibit very little difference between classical quantile regression and vector quantile regression. Small differences occur, since vector quantile regression in the scalar case can be shown to impose the fact that map $t \to A_t$ in (3.5) is nonincreasing, which is not necessarily the case with classical quantile regression under misspecification in population, or even under specification in sample. As can be seen in Figure 1, the difference, however, is minimal.

From the plots in Figure 1, it is also apparent that one-dimensional VQR can also suffer from the "crossing problem," namely the fact that $\beta(t)^\top x$ may not be monotone with respect to t. Indeed, the fact that $t \to A_t$ is nonincreasing fails to imply the fact that $t \to \beta(t)^\top x$ is nondecreasing. There exist procedures to repair the crossing problem, see [7]. However, we see that the crossing problem is modest in the current example.

Running a pair of one-dimensional Quantile Regressions is interesting, but it does not immediately convey the information about the joint conditional dependence in Y_1 and Y_2 (given X). In other words, representations (5.1) are not informative about the joint propensity to consume food and income. One could also wonder whether food and housing are locally complements (respectively locally substitute), in the sense that, conditional on income, an increase in the food consumption is likely to be associated with an increase (respectively a decrease) in the consumption of the housing good. All these questions can be immediately answered with higher-dimensional VQR.

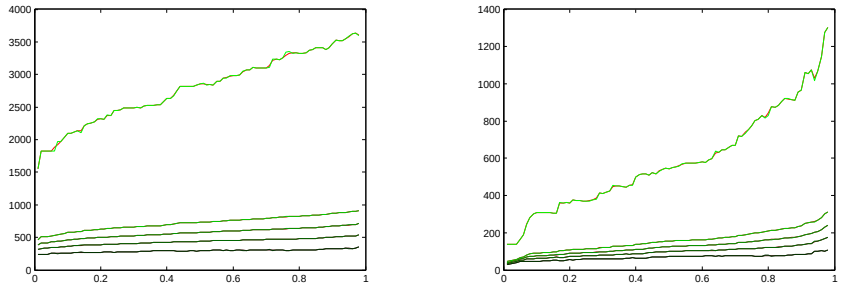

Fig 1: Classical quantile regression (red) and one-dimensional vector quantile regression (green) with income as explanatory variable and with: (i) Food expenditure as dependent variable (Left) and (ii) Housing expenditure as dependent variable (Right). The maps $t \to \beta(t)^\top x$ are plotted for five values of income x_{-1} (quartiles).

5.2. *Two dimensional VQR.* In contrast, the two-dimensional vector quantile regression with $Y = (Y_1, Y_2)$ as a dependent variable yields a representation

$$Y_1 = \beta_1(U_1, U_2)^\top X = \frac{\partial B}{\partial u_1}(U_1, U_2)^\top X,$$
$$Y_2 = \beta_2(U_1, U_2)^\top X = \frac{\partial B}{\partial u_2}(U_1, U_2)^\top X,$$

where $(U_1, U_2) \mid X \sim \mathcal{U}([0,1]^2)$. Let us make a series of remarks.

First, U_1 and U_2 have an interesting interpretation: U_1 is a *propensity for food expenditure*, while U_2 is a *propensity for domestic (housing and heating) expenditure*. Let us explain this denomination. If VQR is correctly specified, then $\Phi_x(u) = B(u)^\top x$ is convex with respect to u, and $Y = \nabla_u \Phi_X(U)$, which implies in particular that

$$\partial/\partial u_1 \left(\partial \Phi_x(u_1, u_2)/\partial u_1\right) = \partial^2 \Phi_x(u_1, u_2)/\partial u_1^2 \geq 0.$$

Hence an increase in u_1 keeping u_2 constant leads to an increase in y_1. Similarly, an increase in u_2 keeping u_1 constant leads to an increase in y_2.

Second, the quantity $U(x, y) = Q^{-1}_{Y|X}(y, x)$ is a measure of joint propensity of expenditure $Y = y$ conditional on $X = x$. This is a way of rescaling the conditional distribution of Y conditional on $X = x$ into the uniform distribution on $[0,1]^2$. If VQR is correctly specified, then (U_1, U_2) is independent from X, so that $U(X, Y) \sim F_U = \mathcal{U}([0,1]^2)$. In this case, $\Pr(U(X, Y) \geq u_1, U(X, Y) \geq u_2) = (1 - u_1)(1 - u_2)$ can be used to detect "nontypical" values of (y_1, y_2).

Third, representation (5.2) may also be used to determine if Y_1 and Y_2 are local complements or substitutes. Indeed, if VQR is correctly specified and (Y_1, Y_2) are independent conditional on X, then $B(u_1, u_2) = B_1(u_1) + B_2(u_2)$, so that the cross derivative $\partial^2 B(u_1, u_2)/\partial u_1 \partial u_2 = 0$. In this case, (5.2) becomes $Y_1 = \frac{\partial B_1}{\partial u_1}(U_1)^\top X$ and $Y_2 = \frac{\partial B_2}{\partial u_2}(U_2)^\top X$, which is equivalent to two single-dimensional quantile regressions. In this case, conditional on X, an increase in Y_1 is not associated to an increase or a decrease in Y_2. On the contrary, when (Y_1, Y_2) are no longer independent conditional on X, then the term $\partial^2 B(u_1, u_2)/\partial u_1 \partial u_2$ is no longer zero. Assume it is positive. In this case, an increase in the propensity to consume food u_1 not only increases the food consumption y_1, but also the housing consumption y_2, which we interpret by saying that food and housing are local complements.

Going back to Engel's data, in Figure 2, we set $x = (1,\ 883.99)$, where $x_2 = 883.99$ is the median value of the total expenditure X_2, and we are able to draw the two-dimensional representations.

The top pane expresses Y_1 as a function of U_1 and U_2, while the bottom pane expresses Y_2 as a function of U_1 and U_2. The insights of the two-dimensional representation become apparent. One sees that while Y_1 covaries strongly with U_1 and Y_2 covaries strongly with U_2, there is a significant and negative cross-covariation: Y_1 covaries negatively with respect to U_2, while Y_2 covaries negatively with U_1. The interpretation is that, for a median level of income, the food and housing goods are local substitutes. This makes intuitive sense, given that food and housing goods account for a large share of the surveyed households' expenditures.

APPENDIX A: PROOFS FOR SECTION 2

A.1. Proof of Theorem 2.1. The first assertion of the theorem is a consequence of the refined version of Brenier's theorem given by [25] (as, e.g, stated in [34], Theorem 2.32), which we apply for each $z \in \mathcal{Z}$. In particular, this implies that for each $z \in \mathcal{Z}$, the map $u \longmapsto Q_{Y|Z}(u,z)$ is Borel-measurable. The Borel measurability of $(u,z) \longmapsto (Q_{Y|Z}(u,z),z)$ follows from the measurability of conditional probabilities and standard measurable selection arguments [14]; for details, we refer to the supplement to this paper [4].

Next we are going to show that the probability law of $(Q_{Y|Z}(V,Z),Z)$ is the same as that of (Y,Z). Indeed, for any rectangle $A \times B \subset \mathbb{R}^{d+k}$,

$$\begin{aligned}
&\mathrm{P}((Y,Z) \in A \times B) = \int_B \left[\int_A F_{Y|Z}(dy,z) \right] F_Z(dz) \\
&= \int_B \left[\int 1\{(Q_{Y|Z}(u,z) \in A\} F_U(du) \right] dF_Z(dz) \\
&= \mathrm{P}((Q_{Y|Z}(V,Z),Z) \in A \times B),
\end{aligned}$$

where the first equality relies on the regularity of the conditional distribution function, and the penultimate equality follows from the previous paragraph. Since probability measure defined over rectangles uniquely pins down the probability measure on all Borel sets via Caratheodory's extension theorem, the claim follows.

To show the second assertion we invoke Dudley-Phillip's ([12]) coupling result given in their Lemma 2.11.

LEMMA A.1 (Dudley-Phillip's coupling). *Let S and T be Polish spaces and Q a law on $S \times T$, with marginal law μ on S. Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space and J a random variable on Ω with values in S and $J \sim \mu$. Assume there is a random variable W on Ω, independent of J, with values in a Polish space R and law ν on R having no atoms. Then there exists a random variable I on Ω with values in T such that $(J,I) \sim Q$.*

First we recall that our probability space has the form:

$$(\Omega, \mathcal{A}, \mathrm{P}) = (\Omega_0, \mathcal{A}_0, \mathrm{P}_0) \times (\Omega_1, \mathcal{A}_1, \mathrm{P}_1) \times ((0,1), B(0,1), \mathrm{Leb}),$$

where $(0,1), B(0,1), \mathrm{Leb})$ is the canonical probability space, consisting of the unit segment of the real line equipped with Borel sets and the Lebesgue measure. We use this canonical space to carry W, which is independent of any other random variables appearing below, and which has the uniform distribution on $R = [0,1]$. The space $R = [0,1]$ is Polish and the distribution of W has no atoms.

Next we apply the lemma to $J = (Y, Z)$ to show existence of $I = U$, where both J and I live on the probability space $(\Omega, \mathcal{A}, \mathrm{P})$ and that obeys the second assertion of the theorem. The variable J takes values in the Polish space $S = \mathbb{R}^d \times \mathbb{R}^k$, and the variable I takes values in the Polish space $T = \mathbb{R}^d$.

Next we describe a law Q on $S \times T$ by defining a triple (Y^*, Z^*, U^*) that lives on a suitable probability space. We consider a random vector Z^* with distribution F_Z, a random vector $U^* \sim F_U$, independently distributed of Z^*, and $Y^* = Q_{Y|Z}(U^*, Z^*)$ uniquely determined by the pair (U^*, Z^*), which completely characterizes the law Q of (Y^*, Z^*, U^*). In particular, the triple obeys $Z^* \sim F_Z$, $U^*|Z^* \sim F_U$ and $Y^* \mid Z^* = z \sim F_{Y|Z}(\cdot, z)$. Moreover, the set $\{(y^*, z^*, u^*) : \|y^* - Q_{Y|Z}(u, z^*)\| = 0\} \subset S \times T$ is assigned probability mass 1 under Q.

By the lemma quoted above, given J, there exists an $I = U$, such that $(J, I) \sim Q$, but this implies that $U|Z \sim F_U$ and that $\|Y - Q_{Y|Z}(U, Z)\| = 0$ with probability 1 under P. ∎

A.2. Proof of Theorem 2.2. We condition on $Z = z$. By reversing the roles of V and Y, we can apply Theorem 2.1 to claim that there exists a map $y \longmapsto Q^{-1}_{Y|Z}(y, z)$ with the properties stated in the theorem such that $Q^{-1}_{Y|Z}(Y, z)$ has distribution function F_U, conditional on $Z = z$. Hence for any test function $\xi : \mathbb{R}^d \to \mathbb{R}$ such that $\xi \in C_b(\mathbb{R}^d)$ we have

$$\int \xi(Q^{-1}_{Y|Z}(Q_{Y|Z}(u, z), z))F_U(du) = \int \xi(u)F_U(du).$$

This implies that for F_U-almost every u, we have $Q^{-1}_{Y|Z}(Q_{Y|Z}(u, z), z) = u$. Hence P-almost surely, $Q^{-1}_{Y|Z}(Y, Z) = Q^{-1}_{Y|Z}(Q_{Y|Z}(U, Z), Z) = U$. Thus we can set $U = Q^{-1}_{Y|Z}(Y, Z)$ P-almost surely in Theorem 2.1. ∎

A.3. Proof of Theorem 2.3. The result follows from [34], Theorem 2.12. ∎

APPENDIX B: PROOFS FOR SECTION 3

B.1. Proof of Theorem 3.1. We first establish part(i). We have a.s.

$$Y = \nabla\Phi_X(\widetilde{U}), \text{ with } \Phi_X(u) = B(u)^\top X.$$

For any $V \sim F_U$ such that $\mathrm{E}(X|V) = \mathrm{E}(X)$, and $\Phi^*_x(y) := \sup_{v\in\mathcal{U}}\{v^\top y - \Phi_x(v)\}$, we have by the mean independence

$$\mathrm{E}[\Phi_X(V) + \Phi^*_X(Y)] = \mathrm{E}B(V)^\top\mathrm{E}(X) + \mathrm{E}\Phi^*_X(Y) := M,$$

where M depends only on F_U. We have by Young's inequality

$$V^\top Y \leq \Phi_X(V) + \Phi_X^*(Y),$$

but $Y = \nabla\Phi_X(\widetilde{U})$ a.s. implies that a.s.

$$\widetilde{U}^\top Y = \Phi_X(\widetilde{U}) + \Phi_X^*(Y),$$

so taking expectations gives

$$\mathrm{E}V^\top Y \leq M = \mathrm{E}\widetilde{U}^\top Y, \quad \forall V \sim F_U : \mathrm{E}(X|V) = \mathrm{E}(X),$$

which yields the desired conclusion.

We next establish part(ii). We can argue similarly as above to show that $Y = \overline{\beta}(\overline{U})^\top X = \nabla\overline{\Phi}_X(\overline{U})$, for some convex potential $u \longmapsto \overline{\Phi}_x(u) = \overline{B}(u)^\top x$, and that for $\overline{\Phi}_x^*(y) := \sup_{v\in\mathcal{U}}\{v^\top y - \Phi_x(v)\}$ we have a.s.

$$\overline{U}^\top Y = \overline{\Phi}_X(\overline{U}) + \overline{\Phi}_X^*(Y).$$

Using the fact that $\widetilde{U} \sim \overline{U}$ and the fact that mean-independence gives $\mathrm{E}(B(\widetilde{U})^\top X) = \mathrm{E}B(\widetilde{U})^\top \mathrm{E}(X) = \mathrm{E}B(\overline{U})^\top \mathrm{E}(X) = \mathrm{E}(B(\overline{U})^\top X)$, we have

$$\mathrm{E}(\widetilde{U}^\top Y) = \mathrm{E}(\Phi_X^*(Y) + B(\widetilde{U})^\top X) = \mathrm{E}(\Phi_X^*(Y) + B(\overline{U})^\top X) \geq \mathrm{E}(\overline{U}^\top Y),$$

where we used Young's inequality again. Reversing the role of $\widetilde{U}$ and $\overline{U}$ and using the same reasoning, we also conclude that $\mathrm{E}(\widetilde{U}^\top Y) \leq \mathrm{E}(\overline{U}^\top Y)$, and hence then

$$\mathrm{E}(\overline{U}^\top Y) = \mathrm{E}(\Phi_X^*(Y) + B(\overline{U})^\top X)$$

so that, thanks to the Young's inequality again,

$$\Phi_X^*(y) + B(u)^\top x \geq u^\top y, \ \forall (u,x,y) \in \mathcal{U}\mathcal{X}\mathcal{Y},$$

we have

$$\Phi_X^*(Y) + B(\overline{U})^\top X = \overline{U}^\top Y, \ \text{ a.s. },$$

which means that $\overline{U} = \arg\max_{u\in\mathcal{U}}\{u^\top Y - B(u)^\top X\}$ a.s., which by strict concavity admits $\widetilde{U}$ as the unique solution a.s. This proves that $\widetilde{U} = \overline{U}$ a.s. and thus we have $(\overline{\beta}(\widetilde{U}) - \beta(\widetilde{U}))^\top X = 0$ a.s.

The part (iii) is a consequence of part (i). Note that by part (ii) we have that $\widetilde{U} = U$ a.s. and $(\beta(U) - \beta_0(U))^\top X = 0$ a.s. Since U and X are independent, we have that, for $e_1, ..., e_p$ denoting vectors of the canonical basis in $\mathbb{R}^p$, and each $j \in \{1, \ldots, p\}$:

$$\begin{aligned}
0 &= \mathrm{E}\left(e_j^\top(\beta(U) - \beta_0(U))^\top XX^\top(\beta(U) - \beta_0(U))e_j\right) \\
&= \mathrm{E}\left(e_j^\top(\beta(U) - \beta_0(U))^\top \mathrm{E}XX^\top(\beta(U) - \beta_0(U))e_j\right) \\
&\geq \mathrm{mineg}(\mathrm{E}XX^\top)\mathrm{E}\left(\|(\beta(U) - \beta_0(U))e_j\|^2\right).
\end{aligned}$$

Since $\mathrm{E}XX^\top$ has full rank this implies that $\mathrm{E}\|(\beta(U) - \beta_0(U))e_j\|^2 = 0$ for each $j \in \{1, \ldots, p\}$, which implies the rest of the claim. ∎

B.2. Proof of Theorem 3.2. We have that any feasible pair (ψ, b) obeys the constraint

$$\psi(x,y) + b(u)^\top x \geq u^\top y, \quad \forall (y,x,u) \in \mathcal{Y}\mathcal{X}\mathcal{U}.$$

Let $\widetilde{U} \sim F_U : \mathrm{E}(X \mid U) = \mathrm{E}(X)$ be the solution to the primal program. Then for any feasible pair (ψ, b) we have:

$$\mathrm{E}\psi(X,Y) + \mathrm{E}b(\widetilde{U})^\top \mathrm{E}X = \mathrm{E}\psi(X,Y) + \mathrm{E}b(\widetilde{U})^\top X \geq \mathrm{E}Y^\top \widetilde{U}.$$

The last inequality holds as equality if

$$\psi(x,y) = \sup_{u \in \mathcal{U}} \{u^\top y - b(u)^\top x\}, \quad b(u) = B(u), \tag{B.1}$$

since this is a feasible pair by (QL) and since

$$\psi(X,Y) + B(\widetilde{U})^\top X = Y^\top \widetilde{U},$$

as shown the proof of the previous theorem. It follows that $\mathrm{E}Y^\top \widetilde{U}$ is the optimal value and it is attained by the pair (B.1). ∎

B.3. Proof of Theorem 3.3. Obviously $A_t = 1 \Rightarrow \widetilde{U} \geq t$, and $\widetilde{U} > t \Rightarrow A_t = 1$. Hence $\mathrm{P}(\widetilde{U} \geq t) \geq \mathrm{P}(A_t = 1) = \mathrm{P}(Y \geq \beta(t)^\top X) = (1-t)$ and $\mathrm{P}(\widetilde{U} > t) \leq \mathrm{P}(A_t = 1) = (1-t)$ which proves that $\widetilde{U}$ is uniformly distributed and $1\{\widetilde{U} > t\}$ coincides with $1\{\widetilde{U}_t = 1\}$ a.s. We thus have $\mathrm{E}(X1\{\widetilde{U} > t\}) = \mathrm{E}(XA_t) = \mathrm{E}X(1-t) = \mathrm{E}X\mathrm{E}A_t$, with standard approximation argument we deduce that $\mathrm{E}(Xf(\widetilde{U})) = \mathrm{E}X\mathrm{E}f(\widetilde{U})$ for every $f \in C_b([0,1], \mathbb{R})$, which means that $\mathrm{E}(X \mid \widetilde{U}) = \mathrm{E}(X)$.

As already observed $\widetilde{U} > t$ implies that $Y > \beta(t)^\top X$ in particular $Y \geq \beta(\widetilde{U} - \delta)^\top X$ for $\delta > 0$, letting $\delta \to 0^+$ and using the a.e. continuity of $u \longmapsto \beta(u)$ we get $Y \geqq \beta(\widetilde{U})^\top X$. The converse inequality is obtained similarly by remaking that $\widetilde{U} < t$ implies that $Y \leq \beta(t)^\top X$.

Let us now prove that $\widetilde{U}$ solves (3.9). Take V uniformly distributed and mean-independent from X and set $V_t := 1\{V > t\}$, we then have $\mathrm{E}(XV_t) = 0$, $\mathrm{E}(V_t) = (1-t)$ but since A_t solves (3.5) we have $\mathrm{E}(V_t Y) \leq \mathrm{E}(A_t Y)$. Observing that $V = \int_0^1 V_t dt$ and integrating the previous inequality with respect to t gives $\mathrm{E}(VY) \leq \mathrm{E}(\widetilde{U}Y)$ so that $\widetilde{U}$ solves (3.9).

Next we show part(ii). Let us define for every $t \in [0,1]$ $B(t) := \int_0^t \beta(s)ds$. Let us also define for (x,y) in $\mathbb{R}^{N+1}$:

$$\psi(x,y) := \max_{t \in [0,1]} \{ty - B(t)^\top x\};$$

thanks to monotonicity condition, the maximization program in the display above is strictly concave in t for every y and each $x \in X$. We then note that

$$Y = \beta(\widetilde{U})^\top X = \nabla B(\widetilde{U})^\top X \text{ a.s.}$$

exactly is the first-order condition for the above maximization problem when $(x, y) = (X, Y)$. In other words, we have

$$\psi(x,y) + B(t)^\top x \geq ty, \ \forall (t,x,y) \in [0,1] \times \mathcal{X} \times \mathbb{R} \tag{B.2}$$

with an equality holding a.s. for $(x, y, t) = (X, Y, \widetilde{U})$, i.e.

$$\psi(X,Y) + B(\widetilde{U})^\top X = \widetilde{U}Y, \quad \text{a.s.} \tag{B.3}$$

Using the fact that $\widetilde{U} \sim \overline{U}$ and the fact that the mean independence gives $\mathrm{E}(B(\widetilde{U})^\top X) = \mathrm{E}(b(\overline{U})^\top X) = \mathrm{E}(X)$, we have

$$\mathrm{E}(UY) = \mathrm{E}(\psi(X,Y) + B(\widetilde{U})^\top X) = \mathrm{E}(\psi(X,Y) + B(\overline{U})^\top X) \geq \mathrm{E}(\overline{U}Y)$$

but reversing the role of $\widetilde{U}$ and $\overline{U}$, we also have $\mathrm{E}(UY) \leq \mathrm{E}(\overline{U}Y)$ and then

$$\mathrm{E}(\overline{U}Y) = \mathrm{E}(\psi(X,Y) + B(\overline{U})^\top X)$$

so that, thanks to inequality (B.2)

$$\psi(X,Y) + B(\overline{U})^\top X = \overline{U}Y, \quad \text{a.s.}$$

which means that $\overline{U}$ solves $\max_{t\in[0,1]}\{tY - \varphi(t) - B(t)^\top X\}$ which, by strict concavity admits $\widetilde{U}$ as unique solution.

Part (iii) is a consequence of Part (ii) and independence of $\widetilde{U}$ and X. Note that by part (ii) we have that $\widetilde{U} = U$ a.s. and that $(\beta(U) - \beta_0(U))^\top X = 0$ a.s. Since U and X are independent, we have that

$$\begin{aligned}
0 &= \mathrm{E}\left((\beta(\widetilde{U}) - \beta_0(U))^\top XX^\top(\beta(U) - \beta_0(U))\right) \\
&= \mathrm{E}\left((\beta(U) - \beta_0(U))^\top \mathrm{E}XX^\top(\beta(U) - \beta_0(U))\right) \\
&\geq \mathrm{mineg}(\mathrm{E}XX^\top)\mathrm{E}\left(\|(\beta(U) - \beta_0(U))\|^2\right).
\end{aligned}$$

Since $\mathrm{E}XX^\top$ has full rank this implies that $\mathrm{E}\|(\beta(U) - \beta_0(U))\|^2 = 0$, which implies the rest of the claim. ∎

ACKNOWLEDGEMENTS

We thank the editor Rungzhe Li, an associate editor, two anonymous referees, Denis Chetverikov, Xuming He, Roger Koenker, Steve Portnoy, and workshop participants at Oberwolfach 2013, Université Libre de Bruxelles, UCL, Columbia University, University of Iowa, MIT, and CORE for helpful comments. Jinxin He, Jérôme Santoul, and Simon Weber provided excellent research assistance.

SUPPLEMENTARY MATERIAL

Supplement to "Vector Quantile Regression"
(doi: XXX). In the online supplement [4], we provide additional results for Sections 2 and 3, including a proof of duality for CVQF and Linear VQR, and the measurability claims for Theorem 2.1.

REFERENCES

[1] Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression*. John Wiley & Sons.

[2] Belloni, A. and R.L. Winkler (2011). "On Multivariate Quantiles Under Partial Orders". *The Annals of Statistics*, Vol. 39 No. 2, pp. 1125–1179

[3] Brenier, Y (1991). "Polar factorization and monotone rearrangement of vector-valued functions ". *Comm. Pure Appl. Math.* 44 **4**, pp. 375–417.

[4] Carlier, G., Chernozhulov, V., and Galichon, A. (2014). "Supplementary Material to "Vector Quantile Regression".

[5] Carlier, G., Galichon, A., Santambrogio, F. (2010). "From Knothe's transport to Brenier's map". *SIAM Journal on Mathematical Analysis* 41, Issue 6, pp. 2554-2576.

[6] Chaudhuri P. (1996). "On a geometric notion of quantiles for multivariate data,". *Journal of the American Statistical Association* 91, pp. 862–872.

[7] Chernozhukov, V., Fernandez-Val, I., and A. Galichon (2010). "Quantile and Probability Curves without Crossing". *Econometrica* 78(3), pp. 1093–1125.

[8] Cunha, F., Heckman, J. and Schennach, S. (2010). "Estimating the technology of cognitive and noncognitive skill formation". *Econometrica* 78(3), pp. 883–931.

[9] Dette, H., Neumeyer, N., and K. Pilz (2006). "A simple Nonparametric Estimator of a Strictly Monotone Regression Function". *Bernoulli*, 12, no. 3, pp. 469–490.

[10] Doksum, K. (1974). "Empirical probability plots and statistical inference for nonlinear models in the twosample case". *Annals of Statistics* 2, pp. 267-277.

[11] Ducpetiaux, E. (1855). *Budgets économiques des classes ouvrières en Belgique*, Bruxelles.

[12] Dudley, R. M., and Philipp, W. (1983). "Invariance principles for sums of Banach space valued random elements and empirical processes". *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 62 (4), pp. 509–552.

[13] Ekeland, I., Galichon, A., and Henry, M. (2012). "Comonotonic measures of multivariate risks". *Mathematical Finance*, 22 (1), pp. 109–132.

[14] I. Ekeland, R. Temam, *Convex Analysis and Variational Problems*, Classics in Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, (1999).

[15] Engel, E. (1857). "Die Produktions und Konsumptionsverhältnisse des Königreichs Sachsen". *Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Misisteriums des Innerm,* 8, pp. 1–54.

[16] Hallin, M., Paindaveine, D., and Šiman, M. (2010). "Multivariate quantiles and multiple-output regression quantiles: From L1 optimization to halfspace depth". *Annals of Statistics* 38 (2), pp. 635–669.

[17] He, X. (1997). "Quantile Curves Without Crossing". *American Statistician*, 51, pp. 186–192.

[18] Koenker, R. (2005), *Quantile Regression*. Econometric Society Monograph Series 38, Cambridge University Press.

[19] Koenker, R. (2007). *quantreg: Quantile Regression.* R package version 4.10. `http://www.r-project.org`.

[20] Koenker, R. and G. Bassett (1978). "Regression quantiles," *Econometrica* 46, 33–50.

[21] Koenker, R. and Bassett, G. (1982). "Robust Tests of Heteroscedasticity based on Regression Quantiles," *Econometrica* 50, pp. 43–61.

[22] Koenker, R. and P. Ng (2005), "Inequality constrained quantile regression." *Sankhyā* 67, no. 2, pp. 418–440.

[23] Koltchinskii, V. (1997). "M-Estimation, Convexity and Quantiles". *Annals of Statistics* 25, No. 2, pp. 435–477.

[24] Kong, L. and Mizera, I. (2012). "Quantile tomography: using quantiles with multivariate data." *Statistica Sinica* 22, pp. 1589–1610.

[25] McCann, R.J. (1995). "Existence and uniqueness of monotone measure-preserving maps". *Duke Mathematical Journal* 80 (2), pp. 309–324.

[26] Laine, B. (2001). "Depth contours as multivariate quantiles: a directional approach." Master's thesis, Université Libre de Bruxelles.

[27] Lehmann, E. (1974). "Nonparametrics: statistical methods based on ranks. "Holden-Day Inc., San Francisco, Calif.

[28] Le Play, F. (1855). *Les Ouvriers Européens. Etudes sur les travaux, la vie domestique et la condition morale des populations ouvrières de l'Europe.* Paris.

[29] Matzkin, R. (2003). "Nonparametric estimation of nonadditive random functions," Econometrica 71, pp. 1339–1375.

[30] R Development Core Team (2007): *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

[31] Ryff, J.V. (1970). "Measure preserving Transformations and Rearrangements." *J. Math. Anal. and Applications* **31**, pp. 449–458.

[32] Serfling, R. (2004). "Nonparametric multivariate descriptive measures based on spatial quantiles". *Journal of Statistical Planning and Inference* 123, pp. 259–278.

[33] Tukey, J. (1975). *Mathematics and the picturing of data.* In Proc. 1975 International Congress of Mathematicians, Vancouver pp. 523–531. Montreal: Canadian Mathematical Congress.

[34] Villani, C. (2003). *Topics in Optimal Transportation.* Providence: American Mathematical Society.

[35] Villani, C. (2009). *Optimal transport: Old and New.* Grundlehren der mathematischen Wissenschaften, Springer-Verlag, Heidelberg, 2009.

[36] Wei, Y. (2008). "An Approach to Multivariate Covariate-Dependent Quantile Contours With Application to Bivariate Conditional Growth Charts". *Journal of the American Statistical Association*, 103 (481), pp. 397–409.

[37] Yu, K., and Jones, M.C. (1998), "Local Linear Quantile Regression". *Journal of the American Statistical Association*, 93 (441), pp. 228–237.

CEREMADE, UMR CNRS 7534,
Université Paris IX Dauphine
Pl. de Lattre de Tassigny,
75775 Paris Cedex 16, FRANCE
E-mail: carlier@ceremade.dauphine.fr

Department of Economics
and Center for Statistics, MIT
50 Memorial Drive, E52-361B
Cambridge, MA 02142, USA
E-mail: vchern@MIT.edu

Economics Department
and Courant Institute of Mathematical Sciences, NYU
70 Washington Square South
New York, NY 10013, USA
E-mail: ag133@nyu.edu

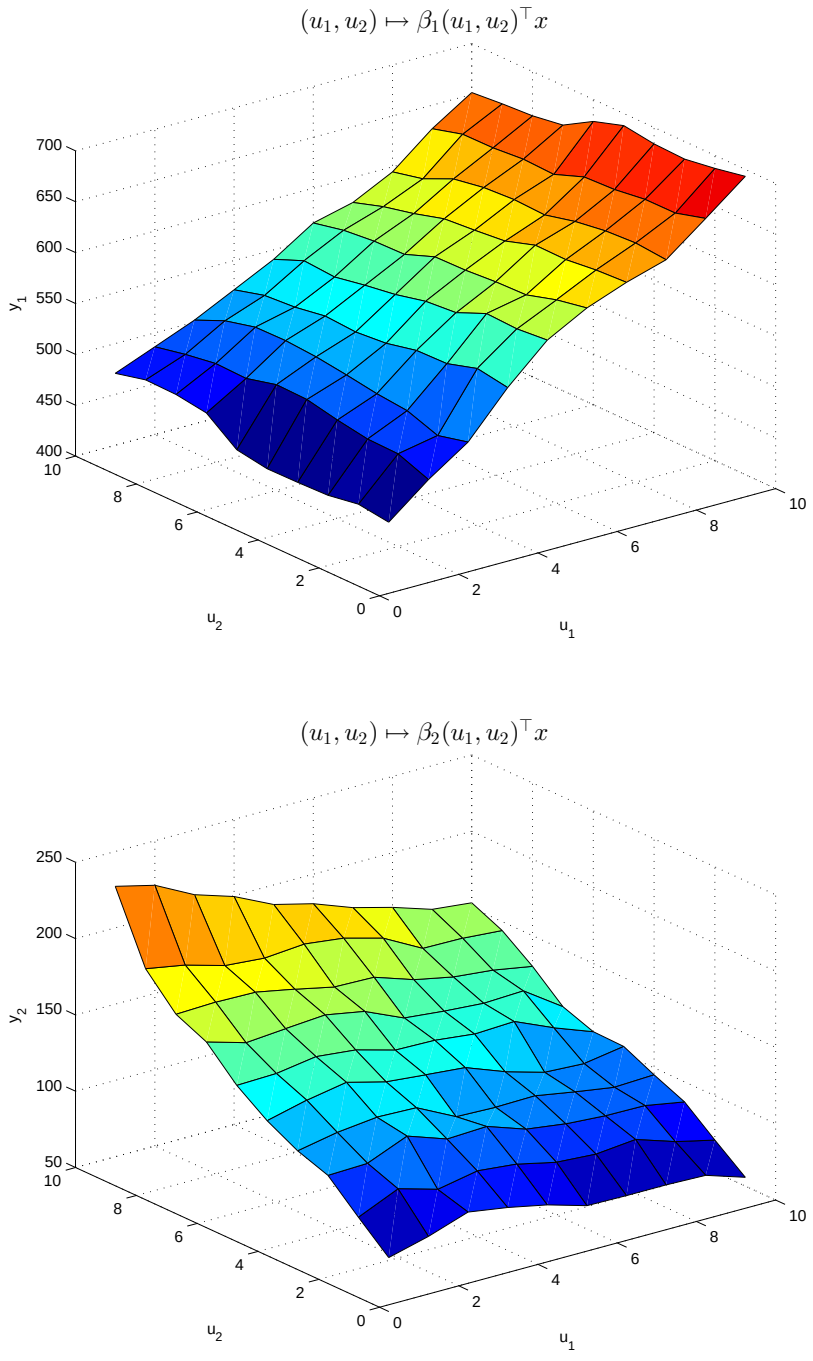

Fig 2: Predicted outcome conditional on total expenditure equal to median value, that is $X_2 = 883.99$. Top: food expenditure, Bottom: housing expenditure.