Quantile Regression: Multi-Task Approaches

Delanoue Pierre
ENS Paris Saclay & ENSAE
pierre.delanoue@ensae.fr

Meunier Dimitri
ENS Paris Saclay & ENSAE
dimitri.meunier@ensae.fr

Abstract

In the report we explore two ways to simultaneously estimate multiple quantiles of a conditional distribution. The first approach is based on the article of Sangnier, Fercoq and d'Alché-Buc [19] and makes use of the Reproducing Kernel Theory. One can see that the non-crossing requirement for multiple quantiles estimation can be integrated into a vector-valued kernel while preserving satisfying quantile properties. Using the representer theorem and duality, the resulting problem can be casted as a quadratic optimisation problem. The second approach by Carlier, Chernozhukov and Galichon [3] is an extension of the quantile regression in the case were the target is a vector. This Optimal Transport approach handles simultaneous estimations of quantiles of a target in $\mathbb{R}^d$ with $d > 1$.

1 Motivations

The quantile regression was introduced by Koenker in the late 70's [11] as an alternative to the classical mean regression. It has desirable properties such as robustness and has been extensively used in the literature. To fully understand a conditional model, we would like to estimate several quantiles. Estimating them one by one often leads to the "crossing problem". The obtained estimated quantiles functions can cross each other which is inconsistent with the theory. In this report we will see two approaches to get **simultaneous** estimations of multiples quantiles that do no cross each other.

The first approach is to look at the Kernel Quantile Regression (KQR). Any algorithm that process finite-dimensional vectors that can be expressed only in terms of pairwise inner products can be applied to potentially infinite-dimensional vectors in the feature space of a positive kernel by replacing each inner product evaluation by a kernel evaluation. Kernel Methods has the advantages of being extensively used in practice with strong theoretical guarantees. Many authors have been studying simultaneous Kernel Quantile Regression from the kernel theory point a view. As an example, Liu and Wu [16], or Takeuchi et al. [15] deal with the crossing problem with multiple scalar valued kernels and add hard constraints into the optimisation problem. Alternatively, Sangnier et al. [19] incorporates the constraints inside a vector-valued kernel and derived efficient algorithms to solve the (quadratic) dual problem.

The second method studied is an Optimal Transport approach proposed by Carlier, Chernozhukov and Galichon [3] to extend the quantile regression to multidimensional targets. Indeed, the definitions of quantiles are based on the fact that $\mathbb{R}$ is ordered. This makes the extension of the quantile definition for Y belonging to $\mathbb{R}^d$ with $d \geq 2$ non-trivial. This approach retains two important properties of the classical quantile regression: (i) It avoids the "crossing problem" (ii) A uniform on the cube $[0.1]^d$ composed by the proposed quantile function has the same law as the target Y. Numerous other notions of multivariate quantiles have been proposed (see [5], [9], [13] and [21]). However, none of these proposals has both properties (i) and (ii) at the same time.

The remainder is organised as follows: in section 2 we recall the setting of the classical Quantile Regression. In section 3, we present the Kernel Quantile Regression. In section 4, we extend the idea of the Kernel Quantile Regression to the multiple-quantile setting following the ideas from Sangnier et al. [19]. In section 5, we present the Vector Quantile Regression Optimal Transport approach of

Carlier, Chernozhukov and Galichon [3]. And finally, in section 6, we illustrate both approaches with numerical applications on Engel's data on household expenditures [12].

Notations

- For all $n \in \mathbb{N}$, $\mathbf{1}_n \in \mathbb{R}^n$ a vector with only ones.
- For $x \in \mathbb{R}$, $x^+ = \max(0,x)$ and $x^- = \max(0,-x)$. For a vector $x \in \mathbb{R}^d$, x^+ is applied coordinate by coordinate.
- For a positive kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $x \in \mathbb{R}$, $K_x : \mathcal{X} \to \mathbb{R}$ is such that $K_x(x') = K(x,x')$
- $\preceq$ represents the partial ordering derived from a proper cone. If applied between vectors, it is the partial ordering induced by the nonnegative orthant (i.e. $x \preceq y \Leftrightarrow x_i \leq y_i, \quad \forall i \in [n]$). If applied between symmetric matrices, it is the partial ordering induced by the positive semi-definite cone (i.e. $A \preceq B \Leftrightarrow B - A$ is positive semi-definite matrix).
- For all $(n,m) \in \mathbb{N}^2$, $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, we note $U(a,b)$ the set of admissible couplings for the discrete Kantorovitch relaxation such that :
$$U(a,b) := \{P \in \mathbb{R}_+^{n \times m};\, P\mathbf{1}_m = a \text{ and } P^T\mathbf{1}_n = b\}$$
- Push-forward operator $T_\#$. For a measurable map $T : \mathcal{X} \mapsto \mathcal{Y}$, the push forward measure $\beta = T_\#\alpha \in \mathcal{P}(\mathcal{Y})$ of some $\alpha \in \mathcal{P}(\mathcal{X})$ satisfies $\beta(B) = \alpha(T^{-1}(B)) \quad \forall B \subset \mathcal{Y}$
- For a vector $u = (u^{(1)}, \ldots, u^{(n)}) \in \mathbb{R}^n$, $u^{-(i)}$ represents the vector u from which the i^{th} element has been removed.

2 Classical Quantile Regression

2.1 Formulation

The quantile function is easily defined for random variables with values in $\mathbb{R}$:

Definition 1 (α-th quantile on $\mathbb{R}$)**.** *For $\alpha \in (0,1)$, the α-th quantile of a random variable Y on $\mathbb{R}$ is defined by:*

$$q_Y(\alpha) = \inf\{x \in \mathbb{R}, F_Y(x) \geq \alpha\}$$

where F_Y is the distribution function of Y. If Y is continuous and strictly increasing, we have $q_{\mathbf{y}}(\alpha) = F_{\mathbf{y}}^{-1}(\alpha)$.

Conditional quantiles can then naturally be defined by:

$$q_{Y|X}(\alpha) = \inf\{z, F_{Y|X}(z) \geq \alpha\}$$

where conditional quantiles are random variables depending on the random variable X.

It is easy to observe the two important following properties for the quantile function:

- (i) $\alpha \longmapsto q_Y(\alpha)$ is non-decreasing
- (ii) If $U \sim \mathcal{U}([0,1])$, then $q_Y(U) = Y$ with probability one.

We also recall the crucial property of the classical quantile function:

Proposition 1. *In the uni-dimensional case where Y is a random variable in $\mathbb{R}$, we have*

$$q_Y(\alpha) \in \arg\min_a \mathbb{E}[\rho_\alpha(Y - a)]$$

where $\rho_\alpha(.)$ is the check function *(or* pinball loss*),*

$$\rho_\alpha(u) = (\alpha - \mathbb{1}_{\{u<0\}})u = \max(\alpha u, (\alpha - 1)u)$$

2.2 Regression

Here we are going to choose to set $X = \phi(Z)$ where $Z \in \mathbb{R}^l$ are covariates and ϕ is a known function from $\mathbb{R}^l$ to $\mathbb{R}^q$. We choose ϕ such that the first component of X is an intercept. We assume the models to be linear:

$$\forall \alpha \in (0,1),\ \exists \beta_\alpha \in \mathbb{R}^q \quad s.t. \quad q_\alpha(Y|X) = \beta_\alpha^T X \tag{1}$$

Following the linear modelling and proposition (1), we then want to solve:

$$\beta_\alpha \in \underset{\beta \in \mathbb{R}^q}{\arg\min}\, \mathbb{E}[\rho_\alpha(Y - X'\beta)] \tag{2}$$

Given a dataset $\mathcal{D} = \{(y_1, x_1), ..., (y_n, x_n)\}$, the quantile regression estimator is naturally built as:

$$\widehat{\beta}_\alpha \in \underset{\beta \in \mathbb{R}^q}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(\mathrm{y}_i - x_i'\beta) \tag{3}$$

This estimator possesses several interesting properties (see Koenker [10] for in depth proofs or D'Haultfeuille [6] for insights of the proofs) such as:

- Identification
- Consistency
- Asymptotic normality
- Enables the building of confidence intervals and statistical tests.

These last two properties are two major advantages of the classical quantile regression. The constitution of statistical tests and confidence intervals provide a lot of information on the data studied. In the next section, we generalise the linear setting to potentially non linear infinite-dimensional estimators with the help of Kernel theory. This will strongly improve the power of the quantile estimators while scarifying interpretability.

3 Kernelised Quantile Regression

3.1 Main idea

Given a positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and the associated (unique) *Reproducing Kernel Hilbert Space* $\mathcal{H} \subset (\mathbb{R})^{\mathcal{X}}$. We can *kernelized* problem 3 with the following problem:

$$(f_\alpha, b_\alpha) \in \underset{f \in \mathcal{H}, b \in \mathbb{R}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(\mathrm{y}_i - f(x_i) - b) + \frac{\lambda}{2} ||f||_{\mathcal{H}}^2 \qquad \lambda > 0 \tag{4}$$

Adding the regularisation term will enforce the norm of the solution $||f||_{\mathcal{H}}$ to be "small", which can be beneficial to ensure a sufficient level of smoothness for the solution (the RKHS norm has a regularization effect). As always with Kernel Methods, the representer theorem is extremely helpful to solve the problem.

Theorem 1 (Representer Theorem)**.** *Le $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ be a function of $n+1$ variables, strictly increasing with respect to the last variable. Then any solution to the optimisation problem:*

$$\inf_{f \in \mathcal{H}} \Psi(f(x_1), \cdots, f(x_n), ||f||_{\mathcal{H}})$$

admits a representation of the form:

$$f(x) = \sum_{i=1}^{n} w_j K(x_i, x), \quad \forall x \in \mathcal{X}$$

In other words, the solution f lives in a finite-dimensional subspace: $f \in Span(K_{x_1}, \cdots, K_{x_n})$

Let's forget the intercept in a first step. We fix $b \in \mathbb{R}$ and solve the following problem,

$$f_\alpha^b \in \underset{f\in\mathcal{H}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(\mathrm{y}_i - f(x_i) - b) + \frac{\lambda}{2}||f||_{\mathcal{H}}^2 \tag{5}$$

It is clear that we can apply the representer theorem to this problem, the solution will have the form $f_\alpha^b = \sum_{i=1}^n w_j^b K_{x_i}$. The following properties are helpful to rewrite the problem:

$$f_\alpha^b(x_j) = \sum_{i=1}^{n} w_j^b K(x_i, x_j) = [K^n w^b]_j, \quad \forall j \in [n]$$

Where K^n is the $n \times n$ matrix, that contains the pairwise distances according to the kernel K, i.e.,

$$\mathrm{K}^n = \begin{pmatrix} K(x_1,x_1) & \dots & K(x_1,x_n) \\ \vdots & \ddots & \vdots \\ K(x_n,x_1) & \dots & K(x_n,x_n) \end{pmatrix}$$

Secondly, we recall the reproducing property in a RKHS,

$$\forall f \in \mathcal{H}, \quad \forall x \in \mathcal{X}, \quad f(x) = \langle f, K_x \rangle_{\mathcal{H}}$$

Applied to $f = K_x$, it gives $\forall (x,x') \in \mathcal{X}^2, \quad K(x,x') = K_x(x') = \langle K_x, K_{x'} \rangle_{\mathcal{H}}$. Using this equality, we get the following,

$$||f_\alpha^b||_{\mathcal{H}}^2 = ||\sum_{i=1}^{n} w_j^b K_{x_i}||_{\mathcal{H}}^2 \underbrace{=}_{\langle .,. \rangle_{\mathcal{H}} \text{bilinear}} \sum_{i,j=1}^{n} w_j^b w_i^b \langle K_{x_i}, K_{x_j} \rangle_{\mathcal{H}} = w^T K^n w$$

Plugging those transformations into problem 4, it can be rewritten as,

$$(w_\alpha, b^\alpha) \in \underset{w\in\mathbb{R}^n, b\in\mathbb{R}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(y_i - [K^n w]_i - b) + \frac{\lambda}{2} w^T K^n w \tag{6}$$

The final estimator is then,

$$h^\alpha(x) = \sum_{i=1}^{n} w_{\alpha,j}^{b^\alpha} K(x_i, x) + b^\alpha, \quad \forall \in \mathcal{X}$$

This form is standard with kernel methods (eg. SVM), the solution is expressed as a weighted sum of the similarity with the other observations measured by the Kernel. We will see in later sections that problem 6 can be solved efficiently with quadratic programming. Let's note that a standard kernel used in practice is the Gaussian kernel: $K(x,x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \forall (x,x') \in (\mathbb{R}^q)^2$

3.2 Getting back to the linear case

The linear Quantile Regression introduced in equation (3) can be penalised with the Tikhonov regularisation, as follow[1]:

$$(\beta_\alpha, b^\alpha) \in \underset{\beta\in\mathbb{R}^q, b\in\mathbb{R}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(\mathrm{y}_i - x_i'\beta - b) + \frac{\lambda}{2}\beta^T\beta \tag{7}$$

On the other hand, we can apply the Linear Kernel $K(x,y) = x^T y$ (which is a well-defined positive-definite kernel) to problem 6 to get,

$$(w^\alpha, b^\alpha) \in \underset{w\in\mathbb{R}^n . b\in\mathbb{R}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(y_i - \sum_{j=1}^{n} w_j x_j^T x_i - b) + \frac{\lambda}{2}\left(\sum_{i=1}^{n} x_i w_i\right)^2 \tag{8}$$

[1] We take the intercept out of the features for the clarity of the comparison.

Both problems are linked through the relation $\beta_\alpha = \sum_{i=1}^n w_{\alpha,i} x_i$. Their solutions then gives:

$$h_\alpha(x) = \sum_{i=1}^n w_{\alpha,j}^{b^\alpha} K(x_i, x) + b^\alpha = \beta_\alpha^T x + b^\alpha, \quad \forall \in \mathcal{X}$$

It's important to note that while being related, problem 7 and 8 are not solved in the same way. For low dimensional problems with $q < n$, problem 7 can be easier to solve while problem 8 can be preferable for high dimensional problems with $q > n$.

3.3 Computation

Let see how we could solve the problem 6. Let $\mathcal{D} = \{(y_1, x_1), ..., (y_n, x_n)\}$ be a dataset of n observations where $y_1, ..., y_n$ are i.i.d samples from F_Y and $x_1, ..., x_n$ are i.i.d. samples from F_X.

Without any explicit solution to 6, we need to solve the program numerically. Specifically, the check function ρ_α is non-differentiable, thus, we cannot use Newton-Raphson algorithms directly. However, we can reformulate 6 as a quadratic programming problem. Let us introduce the slack variables
$\xi_i = y_i - [K^n w]_i - b, \forall i \in [n]$, and let us note that we have,

$$\sum_{i=1}^n \rho_\alpha(\xi_i) = \sum_{i=1}^n \alpha |\xi_i| \mathbb{1}(\xi_i \geq 0) + (1-\alpha)|\xi_i| \mathbb{1}(\xi_i < 0) = \alpha \mathbf{1}_n^T \xi^+ + (1-\alpha)\mathbf{1}_n^T \xi^-$$

We can treat ξ^+ and ξ^- as two slack variables with the constraint $\xi^+ \succeq 0$, $\xi^- \succeq 0$ and $\xi = \xi^+ - \xi^-$, the problem can therefore be re-written as,

$$\begin{aligned} \underset{w\in\mathbb{R}^n, b\in\mathbb{R}, \xi^+\in\mathbb{R}^n, \xi^-\in\mathbb{R}^n}{\text{minimize}} \quad & \alpha \mathbf{1}_n^T \xi^+ + (1-\alpha)\mathbf{1}_n^T \xi^- + \frac{\lambda}{2} w^T K^n w \\ \text{subject to} \quad & \xi^- \succeq 0, \xi^+ \succeq 0 \\ & \xi^+ - \xi^- = y - [K^n w] - b\mathbf{1}_n \end{aligned} \tag{9}$$

This is a well defined convex problem with quadratic objective and linear constraints. We can directly solve 9 with interior points methods (see [2]). However, we will see in a later section that the dual programm is easier to solve.

4 Simultaneous Kernelised Quantile Regression

4.1 Core Intuition

In this section we study how one can extend the estimation of a single quantile to a simultaneous estimation of several quantiles. We want to simultaneously estimate p quantile functions with $0 < \alpha_1 < ... < \alpha_p < 1$. One theoretical property of quantile functions is that they do not cross each other. We therefore want to make sure that we have $h_{\alpha_i}(x) \leq h_{\alpha_j}(x), \forall x \in \mathcal{X}$ and for all (i, j) in $[\![1, p]\!]^2$ if $i \leq j$ where the $(h_{\alpha_i})_{i\in[n]}$ are our estimations of the quantile functions. If we estimate the quantile one by one, there is a possibility that it violates the non-crossing property. The main challenge is therefore to enforce the non-crossing constraint into the optimisation scheme.

4.1.1 Enforcing non-crossing with hard constraints

Several authors introduced algorithms that enforce the non-crossing property with hard constraints (Liu and Wu [16], or Takeuchi et al. [15]).

Liu and Wu [16] introduce the following hypothesis on their kernel: $K(x, y) \geq 0, \quad \forall (x, y) \in \mathcal{X}^2$. We call it the *positivity-bis* hypothesis. It should not be confused with the assumption that a kernel is positive definite, which means that K^n is a non-negative matrix for any n and any cloud of n points. In this framework, we always assume that the kernels are positive definite, and we, in addition, (in this section only) assume that they satisfy *positivity-bis*[2]. Under the assumption *positivity-bis*, they observe that for all (k,k') in $[\![1,p]\!]^2$ such that $k<k'$ and for all i in $[\![1,n]\!]$, if $w_{\alpha_k,i} \leq w_{\alpha'_k,i}$ and $b_{\alpha_k,i} \leq b_{\alpha'_k,i}$, then $h_{\alpha_k}(x) \leq h_{\alpha'_k}(x), \forall x \in \mathcal{X}$. This naturally leads us to the following problem to estimate simultaneously multiple quantiles that do not cross.

$$\begin{array}{ll} \underset{(w_{\alpha_1},\cdots,w_{\alpha_p})\in\mathbb{R}^{n\times p}, b\in\mathbb{R}^p}{\text{minimize}} & \sum_{k=1}^{p}\sum_{i=1}^{n}\rho_{\alpha_k}\left(y_i-\sum_{j=1}^{n}w_{\alpha_k,j}K(x_j,x_i)-b_k\right)+\frac{\lambda}{2}\sum_{k=1}^{p}w_{\alpha_k}^T K^n w_{\alpha_k} \\ \text{subject to} & w_{\alpha_k,i}\leq w_{\alpha_{k+1},i}, \quad \forall i\in[\![1,n]\!], \quad \forall k\in[\![1,p-1]\!] \\ & b_k\leq b_{k+1}, \quad \forall k\in[\![1,p-1]\!] \end{array} \tag{10}$$

λ is the regularisation strength: the stronger λ, the stronger we penalise a large RKHS norm and the smoother the resulting optimal function will be. λ is treated as an hyperparrameter that can be choosen by cross validation for example. The set of constraints ensure that the estimated quantiles functions will not cross. Exactly like in section 3.3, we can rewritte this problem as a Quandratic Program.

4.1.2 Non-crossing with vector-valued kernels

Instead of learning p different kernels simultaneously, Sangnier et al. [19] chose to learn a single vector-valued kernel that takes values in $\mathbb{R}^{p\times p}$. A vector-valued kernel can potentially learn p functions and how their are linked together and is therefore used for multi-task learning [1]. The theory behind vector-valued kernels is essentially the same as for scalar-valued kernels.

Let $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}^{p\times p}$, we say that K is a vector-valued positive definite kernel (v.p.d kernel) if, $\forall(x,x')\in\mathcal{X}^2, K(x,x')$ is symmetric, positive, $K(x,x')=K(x',x)^T$ and $\forall n\in\mathbb{N},\forall(x_i,a_i)_{i\in[n]}\in(\mathcal{X}\times\mathbb{R}^p)^n$

$$\sum_{i,j=1}^{n}\langle a_i, K(x_i,x_j)a_j\rangle \geq 0$$

Similarly to the scalar case, there exists a unique RKHS $\mathcal{H}\subset(\mathbb{R}^p)^{\mathcal{X}}$ associated to K. Any calculus that has been derived in section 3 can be extended to the multidimensional case (see [1] for the details). In particular, problem 4 is simply adapted to the multivariate case,

$$(f_\alpha,b_\alpha)\in \underset{f\in\mathcal{H},b\in\mathbb{R}^p}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{p}\rho_{\alpha_k}(\mathrm{y}_i-f_k(x_i)-b_k)+\frac{\lambda}{2}||f||_{\mathcal{H}}^2 \qquad \lambda>0 \tag{11}$$

Applying the representer theorem and plugging back f similarly to section 3 leads to the following quadratic problem:

$$\begin{array}{ll} \underset{f\in\mathcal{H},b\in\mathbb{R}^p,\xi^+\in(\mathbb{R}^n)^p,\xi^-\in(\mathbb{R}^n)^p}{\text{minimize}} & \sum_{k=1}^{p}\alpha_k\mathbf{1}_n^T\xi_k^+ +(1-\alpha_k)\mathbf{1}_n^T\xi_k^- +\frac{\lambda}{2}||f||_{\mathcal{H}}^2 \\ \text{subject to} & \xi_k^-\succeq 0, \xi_k^+\succeq 0 \quad \forall k\in[p] \\ & \xi_{i,k}^+-\xi_{i,k}^- = y_i-f_k(x_i)-b_k \quad \forall k\in[p] \quad \forall i\in[n] \end{array} \tag{12}$$

Problem 12 is a Quadratic Program and we could apply standard interior point methods. However using a general kernel could lead to memory issues as we have to store a matrix with $np\times np$ entries. The main difficulty is therefore to select an appropriate kernel. Sangnier et al. [19] chose the simplest case for optimisation: a *separable kernel*,

$$[K(x,x')]_{l,m}=k(x,x')\bar{k}(l,m) \quad \forall(l,m)\in\{1,\cdots,p\}^2, \quad \forall(x,x')\in\mathcal{X}^2$$

where k is a p.d. scalar kernel on $\mathcal{X}\times\mathcal{X}$ and $\bar{k}$ is a p.d. scalar kernel on $\{1,\cdots,p\}\times\{1,\cdots,p\}$. Therefore the relationships between the components of the inputs vectors are specified by a kernel

[2] As explain in the article, this is not a strong assumption if we assume that $\mathcal{X}$ is compact as we can always translate our kernel by the minimum value.

that only looks at their coordinate positions. Since $\bar{k}$ does not depend on the input we can rewrite $K(x,x') = k(x,x')B$ where $B_{l,m} = \bar{k}(l,m)$.

Sangnier et al. selected a Gaussian Kernel for $\bar{k}$, i.e. $\bar{k}(l,m) = e^{-\gamma(l-m)^2}, \quad \forall(l,m) \in \{1,\cdots,p\}^2$, $\gamma \geq 0$. The two extremal cases for γ have interesting properties. If $\gamma \to +\infty$, then $B \to I_p$ and therefore K takes values into the space of diagonal matrices; there is no relationship between the components of the input. In that case, by expanding the RKHS norm we get $||f_\alpha||^2_{\mathcal{H}} = \sum_{k=1}^{p} ||f_{\alpha,k}||^2_{\mathcal{H}'}$ where $\mathcal{H}'$ is the RKHS associated to k. Therefore, one can see that problem 12 is separable and equivalent to solving for the quantiles separately. On the other hand, if $\gamma = 0$, B is a matrix full of 1's and the coordinates of f_α will always be equal, since our quantile predictions will only differ by the intercept they will be totally parallel. This ensures that the quantiles will not cross but it is not a desirable property as we would like the relative distances between the quantiles to vary with the values of our covariates. Therefore, as a trade-off, a γ should exist between those two extreme cases that ensures non-crossing and good simultaneous estimations.

4.2 Computation

Problems 10 and 12 are Quadratic programs that can be solved by efficient solvers. However, as pointed out in [19], it is much more efficient to derive and solve the dual of those problems. Taking advantage of the form of the separable kernel, Sangnier et al. derive different algorithms: one is simply to apply interior point methods to the dual of 12, another is derived by noting that with the structure of the separable kernel we can apply a variant of the coordinate descent on the primal and the dual (it is a Primal-Dual Coordinate Descent), and a third algorithms uses a Lagrangian formulation. In is not clear which approach between 10 and 12 is the best to solve the simultaneous estimations of quantiles as they are both supported with theoretical guarantees. However, the Primal-Dual Coordinate Descent algorithm derived by Sangnier et al. seems the most promising as it is computationally efficient. In the numerical sections, we will compare the different algorithms to solve 10 and 12.

5 Optimal Transport Approach

5.1 Vectorized Formulation

We now study the method invented by Carlier, Chernozhukov and Galichon [3] and keep optimal transport notations of Peyré and Cuturi [17]. We are in the context where Y is a random variable taking values in $\mathbb{R}^d$. We admit that the second moment of Y is finite.

The idea is to built a deterministic function $(u,z) \longmapsto Q_{Y|Z}(u,z)$ from $[0,1]^d \times \mathbb{R}^q$ to $\mathbb{R}^d$ where we find some equivalent properties as (i) and (ii) in the univariate case 2.1. We want to have the following properties:

- (i) $(u,z) \longmapsto Q_{Y|Z}(u,z)$ being monotone with respect to u, in the sense of being a gradient of a convex function :

$$(Q_{Y|Z}(u,z) - Q_{Y|Z}(u',z))^T(u-u') \geq 0 \qquad \forall (u,u') \in [0,1]^d \times [0,1]^d, z \in \mathbb{R}^q \tag{13}$$

- (ii) Having with probability one :

$$Y = Q_{Y|Z}(U,Z), \qquad U|Z \sim \mathcal{U}([0,1]^d) \tag{14}$$

In other words, for every $z \in \mathbb{R}^q$ we want to look for a transport T from U to Y:

$$\inf_{T:\mathbb{R}^d \longrightarrow \mathbb{R}^d} \int c(x,T(x))\, F_U(\mathrm{d}x) \qquad T_\# F_U = F_Y,\ F_{U|Z=z} = F_{\mathcal{U}([0,1]^d)} \tag{15}$$

where we will choose the cost c to be $c(x,y) = ||x-y||^2$. Then, we can apply the Brenier's theorem for each $z \in \mathbb{R}^d$ to prove (i).

Theorem 2 (Brenier's theorem, Theorem 2.32 from Villani [4]). *Let μ, ν be two probability measures on $\mathbb{R}^d$. Then with probability one*

$$\exists!\ T\ measurable,\ T\#\mu = \nu\ and\ T = \nabla\phi$$

for some convex function ϕ.

For each $z \in \mathbb{R}^d$ we then find a unique transport T_z. We then define the map $(u,z) \longmapsto Q_{Y|Z}(u,z) = T_z(u)$ which respects property (i).

Property (ii) is proven in Carlier et al. [3] showing that the probability law of $(Q_{Y|Z}(U,Z),Z)$ is the same than the law of (Y,Z). It should be noted that Y is not necessarily continuous.

We can reformulate (15) from probabilistic point of view:

$$\min_U \{\mathbb{E}[||Y-U||^2] :\ U|Z \sim \mathcal{U}([0,1]^d)\} \tag{16}$$

which is equivalent to :

$$\max_U \{\mathbb{E}[U^T Y] :\ U|Z \sim \mathcal{U}([0,1]^d)\} \tag{17}$$

According to Carlier et al. [3], the dual of problem 17 is,

$$\min_{(\psi,\phi)} \mathbb{E}(\phi(U,Z)+\psi(Y,Z)) : \phi(u,z)+\psi(y,z) \geq u^T y \qquad \forall (z,y,u) \in \mathbb{R}^p \times \mathbb{R}^d \times \mathbb{R}^d \tag{18}$$

so that ϕ, the solution of 18 gives:

$$(u,z) \longmapsto Q_{Y|Z}(u,z) = \nabla_u \phi(u,z)$$

We can see here, that one of the big differences with the univariate case is that we will not be able to compute the quantile function at a single point without computing $Q_{Y|Z}(.,.)$ over its entire definition domain.

5.2 Model

Here, like in 2.2, we choose to set $X = f(Z)$ where $Z \in \mathbb{R}^l$ are covariates and f is a known function from $\mathbb{R}^l$ to $\mathbb{R}^q$. We choose f such that the first component of X is an intercept.

We also suppose the model to be linear:

$$Q_{Y|X}(U,X) = \beta(U)^T X, \qquad U|X \sim \mathcal{U}([0,1]^d) \tag{19}$$

where $u \longmapsto \beta(u)$ is a function from $[0,1]^d$ to $\mathbb{R}^{q\times d}$.

This condition will actually be relaxed to have the following condition:

$$Q_{Y|X}(U,X) = \beta(U)^T X, \qquad U \sim \mathcal{U}([0,1]^d)\ and\ \mathbb{E}[X|U] = \mathbb{E}[X] \tag{20}$$

However, 19 and 20 are equivalent if $\forall g \in L^2(F_Z),\ \exists \delta_g,\ g(Z) = X^T \delta_g$.

With the condition of linearity 19, we recall that in this framework, where we wish to solve an optimal transport problem, we try to estimate the function $u \longmapsto \beta_0(u)$ from $\mathbb{R}^d$ to the matrix space $\mathbb{R}^{q\times d}$ such that $\forall x \in \mathbb{R}^q$, $u \longmapsto \beta_0(u)^T x$ is a monotonous smooth map, being the gradient of a certain convex function Φ:

$$\forall (u,x) \in \mathbb{R}^d \times \mathbb{R}^q, \qquad \beta_0(u)^T x = \nabla_u \Phi_x(u) \qquad \Phi_x(u) = B_0(u)^T x$$

where $u \longmapsto B_0(u)$ is continuously differentiable form $\mathbb{R}^d$ to $\mathbb{R}^q$. Therefore, finding B_0 would allow us to approximate β_0.

To construct a vector quantile regression model, Carlier et al.[3] is based on 17 with a relaxed form of the condition $U|X \sim \mathcal{U}([0,1]^d)$:

$$\max_U \{\mathbb{E}[U^T Y] :\ U \sim \mathcal{U}([0,1]^d)\ and\ \mathbb{E}[X|U] = \mathbb{E}[X]\} \tag{21}$$

Then, under 20, the dual program of 21 is proven by Carlier et al. [3] to be the following:

$$\inf_{(\psi,b)} \mathbb{E}[\psi(X,Y)] + \mathbb{E}[b(U)]^T \mathbb{E}[X] : \psi(x,y) + b(u)^T x \geq u^T y \qquad \forall (y,x,u) \in \mathbb{R}^d \times \mathbb{R}^q \times \mathbb{R}^d \tag{22}$$

Carlier, Chernozhukov and Galichon then prove the following result which allows to implement a vector quantile estimation method:

Theorem 3 (Dual solutions, Theorem 3.2 in Carlier et al. [3])**.** *Under the linearity condition, if the second moment of Y and the second moment of U are finite then the solution (ψ^*, b^*) of the dual problem 22 meet the following equalities :*

$$b^*(u) = B_0(u) \qquad \psi^*(x,y) = \sup_u \{u^T y - B_0(u)^T x\}$$

More explicitly, thanks to the theorem 3 we have the following equality:

$$\forall (u,x) \in \mathbb{R}^d \times \mathbb{R}^q, \qquad \beta_0(u)^T x = \nabla_u \Phi_x(u) \qquad \Phi_x(u) = B_0(u)^T x = b^*(u)^T x \tag{23}$$

thus,

$$\forall (u,x) \in \mathbb{R}^d \times \mathbb{R}^q, \qquad \beta_0(u)^T x = \nabla_u (b^*(u)^T x) \tag{24}$$

This is where one of the big differences between the classical version and the optimal transport approach of quantile regression emerges. There does not seem to be a clear framework within which to create statistical tests of significance of the β parameters. Similarly, it does not seem obvious at first glance to build confidence intervals for β.

5.3 Computation

We are given a dataset $D_n = \{(Y_1, Z_1), ..., (Y_n, Z_n)\}$ of n observations where $Y_1, ..., Y_n$ are iid sample following F_Y with value in $\mathbb{R}^d$ and $Z_1, ..., Z_n$ are iid sample following F_Z. Here we are going to choose to set $X_i = f(Z_i)$ for all i in $[\![1,n]\!]$. We also set X and Y such that:

$$\mathsf{X} = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} \in \mathbb{R}^{n\times q} \qquad \mathsf{Y} = \begin{pmatrix} Y_1^T \\ \vdots \\ Y_n^T \end{pmatrix} \in \mathbb{R}^{n\times d} \tag{25}$$

We will estimate the uniform distribution over $[0.1]^d$ by a finite grid of m points $(U_i)_{i\in[\![1,m]\!]}$ of $[0.1]^d$ spaced evenly. We then set $\nu \in \mathbb{R}^n$ and $\mu \in \mathbb{R}^m$ such that:

$$\nu = \begin{pmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix} \qquad \mu = \begin{pmatrix} \frac{1}{m} \\ \vdots \\ \frac{1}{m} \end{pmatrix} \tag{26}$$

The discrete form of our transportation problem is therefore:

$$\max_{P\succeq 0} \sum_{i,j} P_{i,j} Y_j^T U_i \qquad s.t. \qquad P^T \mathbf{1}_m = \nu[\psi],\ PX = \mu\nu^T X[b] \tag{27}$$

where the square brackets indicate the associated Lagrange multiplier. We solve the problem 27 to get $\widehat{b^*}$, the Lagrange multiplier associated with the second equality constraint. By calculation, we can estimate that $\widehat{b^*}$ satisfies the following equality:

$$\widehat{b^*} = \begin{pmatrix} b^*(U_1) \\ \vdots \\ b^*(U_m) \end{pmatrix} = \begin{pmatrix} b_1^*(U_1) \dots b_q^*(U_1) \\ \vdots \\ b_1^*(U_m) \dots b_q^*(U_m) \end{pmatrix} \tag{28}$$

For convenience we write $\widehat{b^*}_i^{(j)} = b_j^*(U_i)$. We wish to calculate an estimator of $u \longmapsto \beta_0(u)$. We recall that:

$$\beta_0(u) = \nabla b^*(u) \approx (\frac{b_j^*(u^{(i)} + \varepsilon, u^{-(i)}) - b_j^*(u^{(i)}, u^{-(i)})}{\varepsilon})_{i\in[\![1,d]\!], j\in[\![1,q]\!]} \tag{29}$$

where $u = (u^{(1)}, ..., u^{(d)})$ and $\varepsilon > 0$ as small as possible.

Remember that we constructed $(U_i)_{i\in[\![1,m]\!]} = ((u_i^{(1)}, ..., u_i^{(d)}))_{i\in[\![1,m]\!]}$ so as to form a grid of $[0,1]^d$ with steps equal to ε.

We introduce the following notation to express the neighbour of a sample U_i on its dimension k:

$$U_i^{(n:k)} = \begin{cases} (u_i^{(1)}, ..., u_i^{(k)} + \varepsilon, ..., u_i^{(d)}) & \text{if } 0 \leq u_i^{(k)} < 1 \\ (u_i^{(1)}, ..., u_i^{(k)} - \varepsilon, ..., u_i^{(d)}) & \text{if } u_i^{(k)} = 1 \end{cases}$$

We have the following property on our dataset:

$$\forall (i,k) \in [\![1,m]\!] \times [\![1,d]\!],\ \exists j \in [\![1,m]\!], \qquad U_j = U_i^{(n:k)}$$

Thus we can create an estimator $\widehat{\beta}$ of the function $u \longmapsto \beta_0(u)$ at each point U_i at our disposal while calculating:

$$\forall i \in [\![1,m]\!], \widehat{\beta}(U_i) := (\frac{b_j^*(U_i^{(n:k)}) - b_j^*(U_i)}{\varepsilon})_{k\in[\![1,d]\!], j\in[\![1,q]\!]} \tag{30}$$

A linear estimator of the vectorized version of quantile regression was thus highlighted.

It is important to note that several points of this implementation will suffer from the curse of dimentionality. Indeed, the number m of samples of the Uniform Act $(U_i)_{i\in[\![1,m]\!]}$ increases exponentially with d the dimension of the definition space of Y. It will therefore be important to discuss the computational limitations of such a method.

6 Implementations

We now implement the methods presented above. These implementations are then used on the dataset classically used to study quantile regression since Koenker and Bassett [12]: Engel's data on household expenditures. Yet, the use of these data has been mostly done in the univariate framework but the target variable is multivariate.

Altogether we ran four approaches. The first one, as a baseline, uses the statsmodel package [20] to see the results in the classical framework of the part 2. The second is an attempt to solve the kernelized problem in part 3 when quantiles are predicted one by one. For this approach we ran directly CVXPY on the primal formulation of 10 but the results were not convincing. Going through the dual problem instead of the primal should improve the results. In a third, step we use the code proposed by Sangnier and Fercoq [19] to solve the kernel problem of part 4 by predicting several quantiles simultaneously. Finally we proposed our own implementation of the Optimal Transport approach method in part 5.

Our implementations of the Kernel and Optimal Transport approaches use CVXPY ([7]) to solve the optimisation problems. All the codes can be found on the GitHub page https://github.com/DimSum2k20/Multi-Task-Quantile-Regression

Figures 1, 2 and 3 are the usual representations of the results of a classical regression. We thus represent the evolution of 5 key quantiles (for α equal to 0.1, 0.3, 0.5, 0.7 and 0.9) as a function of the income for food expenses. This representation testifies notably the power of quantile regression against the ordinary last square to study the heterodestacidity of the data.

Our optimal transport implementation also allows us to use quantile regression for vectors of dimension 2. We propose to illustrate it with figure 4. This plot consists in displaying the values of the quantiles of food and housing expenditures as a function of U1 and U2. As explained in the original paper of Carlier, Chernozhukov and Galichon [3], U1 and U2 can be interpreted as a propensity to consume the good with which they are associated. The interest of figure 4 is to see how Y1 evolves with U2. Indeed we know by construction that Y1 is increasing with U1, but we learn that Y1 covaries strongly with U2. It is the same for Y2 with U1. This shows us that for a median income, these two goods are locally substitutable goods.

7 Conclusion

We studied and implemented four different methods of quantile regression. Beyond the classical method, we were able to address three exciting research topics around quantile regression.

On one hand, the non-crossing problem. This problem has been the subject of numerous publications, of which we have only addressed a few proposals among many. On the other hand, we introduced the Kernel and Optimal Transport approaches to quantile regression. Both approaches represent rich and exciting research areas for quantile regression and machine learning in general. With the Optimal Transport approach, we got a glimpse of the multidimensional quantile regression problem.

With the perspective to continue this work, several other aspects can be investigated to improve the methods. First, kernel methods have a known issue of scalability. One should check if the Random Kernel Features methods [18] or the Nyström method [8] can be applied to the vector-Kernel Quantile regression. Secondly, a trade-off can be addressed between the mean regression and the median regression through the Huber loss [14]. One could check if other similar links exist outside the median as it could lead to more hybrid algorithms. Finally, it would also be interesting to study a method using a kernelized approach to propose an answer to the problem of simultaneous multidimensional quantile regression.

References

[1] Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. “Kernels for vector-valued functions: A review”. In: *Foundations and Trends® in Machine Learning* 4.3 (2012), pp. 195–266.

[2] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[3] Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. “Vector quantile regression: An optimal transport approach”. In: *The Annals of Statistics* 44.3 (2016), pp. 1165–1192. DOI: 10.1214/15-aos1401.

[4] Villani Cédric. *Topics in optimal transportation*. American mathematical society, 2016.

[5] Probal Chaudhuri. “On a Geometric Notion of Quantiles for Multivariate Data”. In: *Journal of the American Statistical Association* 91.434 (1996), pp. 862–872. DOI: 10.1080/01621459.1996.10476954.

[6] Xavier D’Haultfœuille. *Semi and Nonparametric Econometrics, Lectures Note*. 2017.

[7] Steven Diamond and Stephen Boyd. “CVXPY: A Python-Embedded Modeling Language for Convex Optimization”. In: *Journal of Machine Learning Research* 17.83 (2016), pp. 1–5.

[8] Petros Drineas and Michael W Mahoney. “On the Nyström method for approximating a Gram matrix for improved kernel-based learning”. In: *journal of machine learning research* 6.Dec (2005), pp. 2153–2175.

[9] Marc Hallin, Davy Paindaveine, and Miroslav Šiman. “Multivariate quantiles and multiple-output regression quantiles: From L 1 optimization to halfspace depth”. In: *The Annals of Statistics* 38.2 (2010), pp. 635–669. DOI: 10.1214/09-aos723.

[10] Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[11] Roger Koenker and Gilbert Bassett. “Regression Quantiles”. In: *Econometrica* 46.1 (1978), p. 33. DOI: 10.2307/1913643.

[12] Roger Koenker and Gilbert Bassett. “Robust Tests for Heteroscedasticity Based on Regression Quantiles”. In: *Econometrica* 50.1 (1982), p. 43. DOI: 10.2307/1912528.

[13] V. I. Koltchinskii. “M -estimation, convexity and quantiles”. In: *The Annals of Statistics* 25.2 (1997), pp. 435–477. DOI: 10.1214/aos/1031833659.

[14] Sophie Lambert-Lacroix, Laurent Zwald, et al. “Robust regression through the Huber’s criterion and adaptive lasso penalty”. In: *Electronic Journal of Statistics* 5 (2011), pp. 1015–1053.

[15] Quoc V Le, Tim Sears, and Alexander J Smola. *Nonparametric quantile regression*. Tech. rep. Technical report, National ICT Australia, June 2005. Available at http://sml . . ., 2005.

[16] Yufeng Liu and Yichao Wu. “Simultaneous multiple non-crossing quantile regression estimation using kernel constraints”. In: *Journal of nonparametric statistics* 23.2 (2011), pp. 415–437.

[17] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. 2018. arXiv: 1803.00567 [stat.ML].

[18] Ali Rahimi and Benjamin Recht. “Random features for large-scale kernel machines”. In: *Advances in neural information processing systems*. 2008, pp. 1177–1184.

[19] Maxime Sangnier, Olivier Fercoq, and Florence d’Alché-Buc. “Joint quantile regression in vector-valued RKHSs”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 3693–3701.

[20] Skipper Seabold and Josef Perktold. “Statsmodels: Econometric and statistical modeling with python”. In: *9th Python in Science Conference*. 2010.

[21] Ying Wei. “An Approach to Multivariate Covariate-Dependent Quantile Contours With Application to Bivariate Conditional Growth Charts”. In: *Journal of the American Statistical Association* 103.481 (2008), pp. 397–409. DOI: 10.1198/016214507000001472.

A Plots

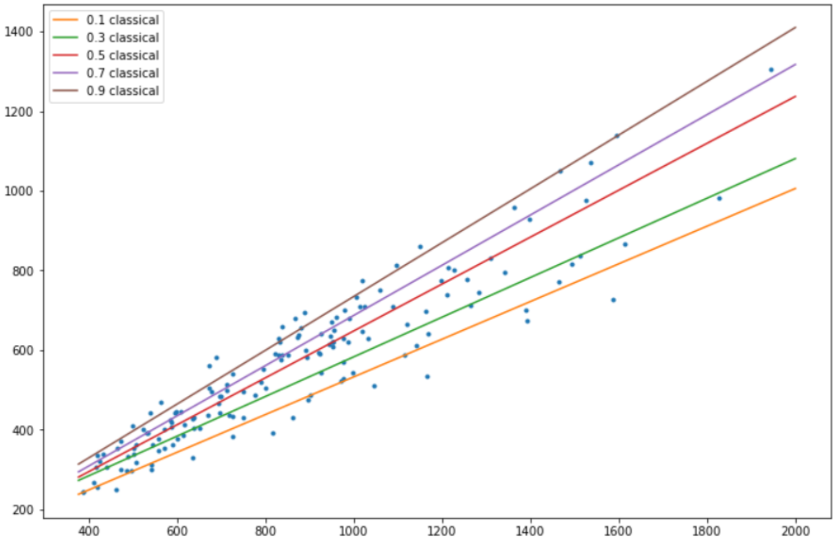

Figure 1: Results of the one by one estimation of quantile regression with statsmodel (classical method) for food expenses as a function of salary for 5 values of quantiles: 0.1, 0.3, 0.5, 0.7 and 0.9

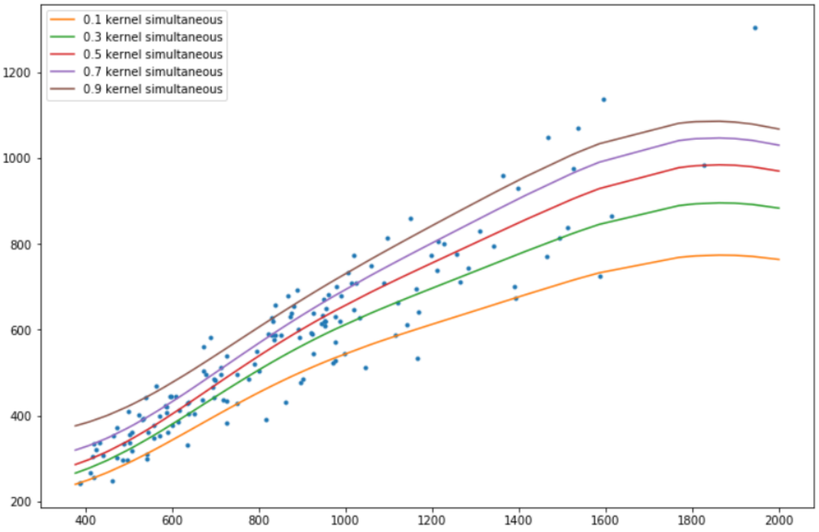

Figure 2: Results ot the simultaneous estimation of quantile regression with Sangnier and Fercoq implementation [19] for food expenses as a function of salary for 5 values of quantiles: 0.1, 0.3, 0.5, 0.7 and 0.9
.

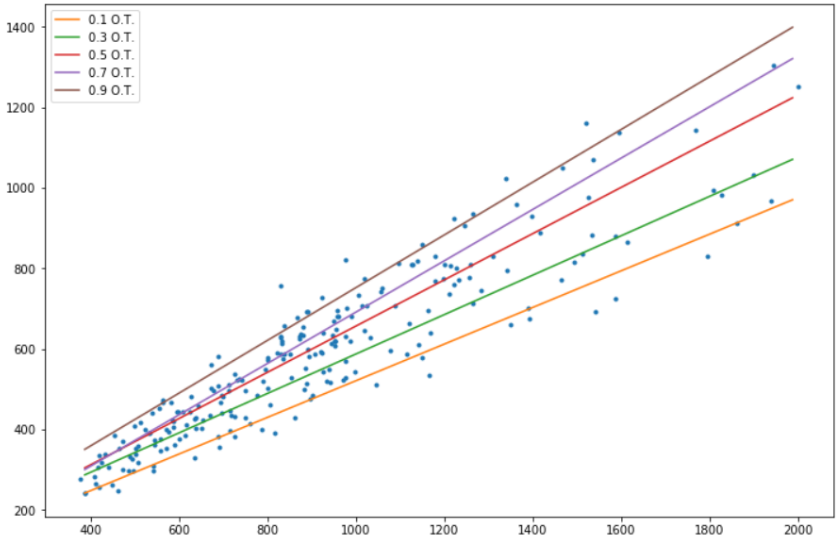

Figure 3: Results of our optimal transport implementation of quantile regression for food expenses as a function of salary for 5 values of quantiles: 0.1, 0.3, 0.5, 0.7 and 0.9

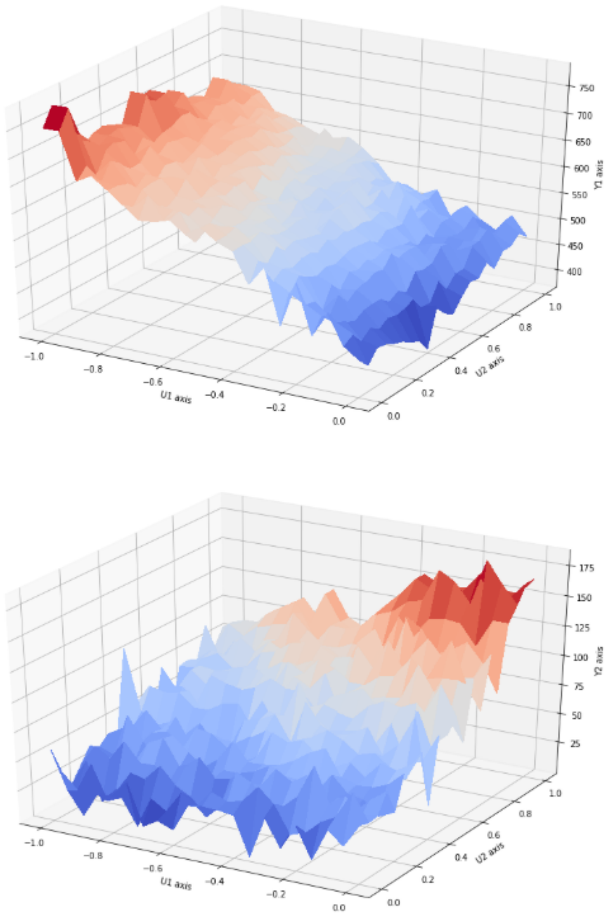

Figure 4: Values predicted by the vectorized quantile regression for X = 883.99 (median value of the Engel's data). The top graph shows food and the bottom graph shows house expenses.