Joint quantile regression in vector-valued RKHSs

Maxime Sangnier, Olivier Fercoq, Florence d'Alché-Buc

Joint quantile regression in vector-valued RKHSs

Maxime Sangnier, Olivier Fercoq and Florence d'Alché-Buc
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France

February 10, 2016

Abstract

Building upon kernel-based multi-task learning, a novel methodology for estimating and predicting simultaneously several conditional quantiles is proposed. We particularly focus on curbing the embarrassing phenomenon of quantile crossing. Moreover, this framework comes along with a uniform convergence bound and an efficient coordinate descent learning algorithm. Numerical experiments on benchmark datasets highlight the enhancements of our approach regarding the prediction error, the crossing occurrences and the training time.

1 Introduction

Given a couple (X, Y) of random variables, where Y takes scalar continuous values, a common aim in statistics and machine learning is to estimate the conditional expectation $\mathbb{E}[Y \mid X = x]$ as a function of x. In the previous setting, called regression, one assumes that the main information in Y is a scalar value corrupted by a centered noise. However, in some applications such as econometrics, social sciences and ecology, Y may carry a *structural* information, represented by its conditional distribution. Such a scenario raises the will to know more than the expectation of the distribution and for instance, expectiles and quantiles are different quantities able to achieve this goal.

This paper deals with this last setting, called (conditional) quantile regression. This topic has been championed by Koenker & Bassett (1978) as the minimization of the pinball loss (see (Koenker, 2005) for an extensive presentation) and brought to the attention of the machine learning community by Takeuchi et al. (2006); Rosset (2009). Ever since then, several studies have built upon this framework and the most recent ones include a definition of multivariate quantiles (when Y is a random vector) and the corresponding framework for multiple-output quantile regression (where we are interested in a single quantile level) Hallin et al. (2010, 2015); Hallin & Šiman (2016). On the contrary, we are interested in estimating and predicting simultaneously several quantiles of a scalar-valued random variable $Y|X$ (see Figure 1), what is called joint quantile regression. For this purpose, we focus on non-parametric hypotheses from a vector-valued Reproducing Kernel Hilbert Space (RKHS).

Since quantiles of a distribution are closely related, joint quantile regression is subsumed under the field of multi-task learning Jebara (2004); Evgeniou et al. (2005); Argyriou et al. (2008); Ciliberto et al. (2015). As a consequence, vector-valued kernel methods Micchelli & Pontil (2005b) are appropriate for such a task. They have already been used for various applications, such as image colorization Minh et al. (2010), classification Dinuzzo et al. (2011); Mroueh et al. (2012), manifold regularization Minh & Sindhwani (2011); Brouard et al. (2011), vector autoregression Lim et al. (2014), functional regression Kadri et al. (2010, 2015) and structured regression Brouard et al. (2015). Quantile regression is a new opportunity for vector-valued RKHSs to perform in a multi-task problem, along with a loss that is different from the ℓ_2 cost predominantly used in the previous references.

In addition, such a framework offers a novel way to deal with an embarrassing phenomenon: often, estimated quantiles cross, thus violating the basic principle that the cumulative distribution function should be monotonically non-decreasing. The method proposed in this paper can curb that phenomenon while preserving the so called *quantile property*. This one guarantees that the ratio of observations lying below a predicted quantile is bounded by the quantile level of interest. The quantile property may not be satisfied if, for instance, hard non-crossing constraints are enforced during the estimation (Takeuchi et al., 2006).

In a nutshell, this work provides the following contributions (reflecting the outline of the paper): **i)** a novel methodology for joint quantile regression, that is based on vector-valued RKHSs; **ii)** enhanced predictions thanks to a multi-task approach along with limited appearance of crossing curves; **iii)** a uniform bound regarding the generalization of the model, which is, as far as we know, the first such result based on the Rademacher average for kernelized hypothesis

spaces; **iv)** an efficient coordinate descent algorithm, that is able to handle the intercept of the model in a manner that is simple and different from Sequential Minimal Optimization (SMO). Besides these novelties, the enhancements of the proposed method and the efficiency of our learning algorithm are supported by numerical experiments on benchmark datasets.

2 Related work

Since the introduction of quantile regression Koenker & Bassett (1978); Koenker (2005) research spread in two noteworthy directions. First, estimators opened onto scalar-valued RKHSs (Takeuchi & Furuhashi, 2004; Takeuchi et al., 2006; Li et al., 2007). Second, different losses have been used, such as the ϵ-insensitive loss (Takeuchi & Furuhashi, 2004; Takeuchi et al., 2006; Steinwart & Christmann, 2008) and a re-weighted least squares penalty (Schnabel & Eilers, 2012).

Regarding the embarrassing phenomenon of quantile crossing, it has been first tackled with a location-scale model $Y = \mu(X) + \sigma(X)\epsilon$, where ϵ is a noise with zero mean and prescribed variance. A multi-step strategy to estimate sequentially (linear or spline-based) mappings μ and σ by enforcing σ to be positive enables to get non-crossing τ-quantile regressors of the form $\mu(X) + \sigma(X)c_\tau$ (He, 1997). Thereafter, an extension based on penalized kernel machines has been proposed (Shim et al., 2009).

Even though the quantile property may not be satisfied in this case, a common way to prevent curve crossing is to enforce hard non-crossing constraints during the M-estimation process. For instance, for linear estimators, all it takes to make sure that quantile functions do not cross on the vertices of the hypercube including the data (Wu & Liu, 2009). On the other hand, for non-linear estimators, hard non-crossing constraints are generally enforced on training points (Takeuchi & Furuhashi, 2004; Takeuchi et al., 2006; Wu & Liu, 2009; Bondell et al., 2010). This solution cannot guarantee that curve crossing does not occur elsewhere. To fight this remark, an option is to work with RKHSs based on a non-negative-valued kernel and to enforce constraints on the weights of the kernel expansion (Liu & Wu, 2011).

Other techniques exist, such as computing a simultaneous inversion and monotonization of an initial estimate of the conditional distribution function (Dette & Volgushev, 2008) or rearranging the quantile estimations to make them monotonely increasing with respect to the quantile level (Chernozhukov et al., 2010). In another study, Schnabel & Eilers (2012) consider a quantile as a bivariate functions (of the input variable and the quantile level) and estimate it as a tensor product spline surface, smooth enough to prevent crossing.

Last but not least, Takeuchi et al. (2013) showed that for linear regressors, estimators are piecewise linear functions of the quantile level. Thus, estimating simultaneously all conditional quantiles is theoretically feasible.

In comparison to the literature, we propose a novel methodology, based on vector-valued RKHSs, with a one-step estimation, no post-processing, and keeping the quantile property while dealing with curve crossing.

3 Framework

3.1 Problem definition

Let $\mathcal{Y} \subset \mathbb{R}$ be a compact set, $\mathcal{X}$ be an arbitrary input space and $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ a pair of random variables following an unknown joint distribution. For a given probability $\tau \in (0, 1)$, the τ-quantile of Y, denoted μ_τ, is the minimal scalar value μ for which $\mathbb{P}(Y \leq \mu) = \tau$. Likewise, the conditional τ-quantile of (X, Y) is the function $\mu_\tau \colon \mathcal{X} \to \mathbb{R}$ such that $\mu_\tau(x) = \min\{\mu \in \mathbb{R} : \mathbb{P}(Y \leq \mu \mid X = x) = \tau\}$.

Given a training set $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ and a vector $\boldsymbol{\tau} \in (0, 1)^p$ of quantile levels, the paradigm is to estimate the vector-valued function $\mathbf{x} \in \mathcal{X} \mapsto (\mu_{\tau_1}(\mathbf{x}), \ldots, \mu_{\tau_p}(\mathbf{x})) \in \mathbb{R}^p$ of conditional quantiles (hence the name *joint* regression).

3.2 Loss function

Following Koenker (2005), we estimate quantiles by minimization of the pinball loss: for a residual $\boldsymbol{r} \in \mathbb{R}^p$,

$$\ell_{\boldsymbol{\tau}}(\boldsymbol{r}) = \sum_{j=1}^{p} \begin{cases} \tau_j r_j & \text{if } r_j \geq 0, \\ (\tau_j - 1) r_j & \text{if } r_j < 0. \end{cases}$$

Using such a loss arose from the observation that the location parameter μ that minimizes the ℓ_1-loss $\sum_{i=1}^n |y_i - \mu|$ is an estimator of the median. This idea has been extended by Koenker & Bassett (1978) to the estimation of other quantiles. Let $\mathbb{N}_n$ be the range of integers between 1 and n and suppose for a while that we are interested in a single quantile τ ($p = 1$). Let μ_τ be a minimizer of the function $\mu \in \mathbb{R} \mapsto \sum_{i=1}^n \ell_\tau(y_i - \mu)$ and denote $n_- = \operatorname{card}(\{i \in \mathbb{N}_n, y_i < \mu_\tau\})$, $n_+ = \operatorname{card}(\{i \in \mathbb{N}_n, \mu_\tau < y_i\})$ and $n_0 = \operatorname{card}(\{i \in \mathbb{N}_n, y_i = \mu_\tau\})$, respectively the number of observations y_i that lie below, above and on the location parameter μ_τ. As a consequence, we know that $(\tau - \frac{n_0}{n}) \leq \frac{n_-}{n} \leq \tau$ and $(1 - \tau - \frac{n_0}{n}) \leq \frac{n_+}{n} \leq (1 - \tau)$ (Koenker & Bassett, 1978) [1]. Moreover, if Y has a continuous distribution, $\frac{n_-}{n} \underset{n\to\infty}{\longrightarrow} \tau$ (Takeuchi et al., 2006). The previous statements indicate that the expected *quantile property* is empirically satisfied: the ratio of observations lying below μ_τ is bounded by above by τ and converges to τ when the sample grows (and respectively for the ratio of observations lying above μ_τ).

The quantile property comes with another interesting fact: when one estimates jointly two quantiles with the pinball loss, the *natural order* is respected. Let $\mathbb{1}$ be the all-ones vector (the size depends on the context). Then, for two quantile levels $\boldsymbol{\tau} = (\tau, \tau') \in (0, 1)^2$, such that $\tau > \tau'$, any minimizer $(\mu_\tau, \mu_{\tau'})$ of the function $\mu \in \mathbb{R}^2 \mapsto \sum_{i=1}^n \ell_{\boldsymbol{\tau}}(y_i \mathbb{1} - \mu)$ is such that $\mu_\tau \geq \mu_{\tau'}$. An original proof is given in Appendix A.

The next sections will focus on the estimation of conditional quantiles, instead of unconditional ones. In particular, we will see that the *natural order* of conditional quantiles is not easily satisfied.

3.3 Estimation of conditional quantiles

Spinning the estimation process introduced in the previous section, let us define the joint quantile risk:

$$R \colon h \in (\mathbb{R}^p)^{\mathcal{X}} \mapsto \mathbb{E}\left[\ell_{\boldsymbol{\tau}}(Y\mathbb{1} - h(X))\right],$$

where $(\mathbb{R}^p)^{\mathcal{X}}$ is the set of functions from $\mathcal{X}$ to $\mathbb{R}^p$. In the scalar case ($p = 1$ and $\boldsymbol{\tau}$ reduces to a scalar τ), Li et al. (2007) have shown that the conditional τ-quantile of (X, Y) is a minimizer of R (see Appendix A for a reminder of the proof). This is also true for several quantiles ($p \geq 1$). To see this, all it takes to remark that the objective function is separable: $R(h) = \sum_{j=1}^p \mathbb{E}\left[\ell_{\tau_j}(Y - h_j(X))\right]$, where h_j is the j^{th} component of h. Thus minimizing R boils down to minimizing each contribution $\mathbb{E}\left[\ell_{\tau_j}(Y - h_j(X))\right]$ independently, for which we know that the conditional τ_j-quantile of (X, Y) is a minimizer (Li et al., 2007).

Since the joint probability of (X, Y) is unknown but we are provided with an independent and identically distributed (*iid*) sample of observations $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$, we resort to minimizing the empirical risk within a class $\mathcal{H} \subset (\mathbb{R}^p)^{\mathcal{X}}$ of functions, calibrated in order to overcome the shift from the true risk to the empirical one. Thus the estimation procedure is to solve the following optimization problem:

$$\underset{h \in \mathcal{H}}{\text{minimize}} \; R_{\text{emp}}(h) = \frac{1}{n} \sum_{i=1}^n \ell_{\boldsymbol{\tau}}(y_i \mathbb{1} - h(\mathbf{x}_i)). \tag{1}$$

Let $\psi \colon (\mathbb{R}^p)^{\mathcal{X}} \to \mathbb{R}$ be a convex function and a constant $c > 0$. In the following, we assume that $\mathcal{H}$ is a convex set of bounded functions with an unconstrained intercept: $\mathcal{H} = \{h = f + \boldsymbol{b} : \boldsymbol{b} \in \mathbb{R}^p, f \in (\mathbb{R}^p)^{\mathcal{X}}, \psi(f) \leq c\}$. Then the estimator $\hat{h}$, obtained by minimizing R_{emp}, comes with the expected quantile property for each level τ_j ($j \in \mathbb{N}_p$), as for unconditional quantile estimation [2]: $(\tau_j - \frac{n_0^{(j)}}{n}) \leq \frac{n_-^{(j)}}{n} \leq \tau_j$ and $(1 - \tau_j - \frac{n_0^{(j)}}{n}) \leq \frac{n_+^{(j)}}{n} \leq (1 - \tau_j)$, where $n_-^{(j)} = \operatorname{card}\left(\left\{i \in \mathbb{N}_n, y_i < \hat{h}(\mathbf{x}_i)\right\}\right)$ $n_+^{(j)} = \operatorname{card}\left(\left\{i \in \mathbb{N}_n, \hat{h}(\mathbf{x}_i) < y_i\right\}\right)$ and $n_0^{(j)} = \operatorname{card}\left(\left\{i \in \mathbb{N}_n, y_i = \hat{h}(\mathbf{x}_i)\right\}\right)$. Moreover, under some regularity assumptions, $\frac{n_-^{(j)}}{n} \underset{n\to\infty}{\longrightarrow} \tau_j$ almost surely. These statements were proven by Takeuchi et al. (2006) in the scalar case. For several probabilities τ_j, it is enough to remark that the objective function is separable.

The proof relies on the fact that the intercept $\boldsymbol{b}$ is unconstrained. Without this assumption (for instance when hard non-crossing constraints are enforced during the estimation process (Takeuchi et al., 2006)), the previous quantile property for conditional quantiles may not be satisfied.

[1] Formal statement and proof are given in Appendix A.

[2] Formal statement and proof are given in Appendix A.

3.4 The crossing problem

A nice feature of quantile regression is to enable us to look at slices of the conditional distribution of $Y|X$. However, when quantiles are estimated independently, an embarrassing phenomenon often appears: quantile functions cross, thus violating the basic principle that the cumulative distribution function should be monotonically non-decreasing.

The crossing problem is often handled empirically (like in this paper) and few theoretical insights exist. Yet, it is comforting to know that if the regularizer ψ is separable (that is $\psi(f)$ does not exhibit interaction terms between components f_j), the conditional quantile estimator is monotonically non-decreasing in the probability τ_j *on average* [3]: $\forall j \in \mathbb{N}_{p-1}, \frac{1}{n}\sum_{i=1}^{n} h_j^*(\mathbf{x}_i) \geq \frac{1}{n}\sum_{i=1}^{n} h_{j+1}^*(\mathbf{x}_i)$, as soon as $\tau_j > \tau_{j+1}$.

Let us remark that, on the one hand, when the different components h_j are estimated independently from each other, ψ is separable, so quantile functions effectively do not cross *on average*, even though crossing may occur. On the other hand, ψ can be chosen in order to reflect the *similarity* between the components h_j. In this paper, we propose to use this mean to prevent curve crossing.

In the forthcoming section, we focus on a particular kind of subspace $\mathcal{H}$, built upon an RKHS. Then, it will be shown that, in this space $\mathcal{H}$, the crossing problem can be avoided.

3.5 Reproducing kernel Hilbert space

Let us denote $\cdot^T$ the transpose operator and $\mathcal{L}(\mathbb{R}^p)$ the set of linear and bounded operators from $\mathbb{R}^p$ to itself. In our (finite) case, $\mathcal{L}(\mathbb{R}^p)$ comes down to the set of $p \times p$ real-valued matrices. A matrix-valued kernel is a function $K\colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathbb{R}^p)$, that is symmetric and positive (Senkene & Tempel'man, 1973; Micchelli & Pontil, 2005b): $\forall (\mathbf{x}, \mathbf{x}') \in \mathcal{X}\times\mathcal{X}, K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})^T$ and $\forall m \in \mathbb{N}, \forall \{(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)\}_{1\leq i\leq m} \in (\mathcal{X}\times\mathbb{R}^p)^m, \sum_{1\leq i,j\leq m} \left\langle \boldsymbol{\beta}_i \mid K(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j)\boldsymbol{\beta}_j \right\rangle_{\ell_2} \geq 0$.

Let K be such a kernel and for any $\mathbf{x} \in \mathcal{X}$, let $K_{\mathbf{x}}\colon \boldsymbol{y} \in \mathbb{R}^p \mapsto K_{\mathbf{x}}\boldsymbol{y} \in (\mathbb{R}^p)^{\mathcal{X}}$ be the linear operator such that: $\forall \mathbf{x}' \in \mathcal{X}, (K_{\mathbf{x}}\boldsymbol{y})(\mathbf{x}') = K(\mathbf{x}', \mathbf{x})\boldsymbol{y}$. Then, there exists a unique Hilbert space of functions $\mathcal{K}_K \subset (\mathbb{R}^p)^{\mathcal{X}}$ (with an inner product and a norm respectively denoted $\langle \cdot \mid \cdot \rangle_{\mathcal{K}}$ and $\|\cdot\|_{\mathcal{K}}$), called the RKHS associated to K, such that $\forall \mathbf{x} \in \mathcal{X}$ (Senkene & Tempel'man, 1973; Carmeli et al., 2010): $K_{\mathbf{x}}$ spans the space $\mathcal{K}_K$ ($\forall \boldsymbol{y} \in \mathbb{R}^p\colon K_{\mathbf{x}}\boldsymbol{y} \in \mathcal{K}$), $K_{\mathbf{x}}$ is bounded for the uniform norm ($\sup_{\boldsymbol{y}\in\mathbb{R}^p} \|K_{\mathbf{x}}\boldsymbol{y}\|_{\mathcal{K}} < \infty$) and $\forall f \in \mathcal{K}\colon f(\mathbf{x}) = K_{\mathbf{x}}^* f$ (reproducing property), where $\cdot^*$ is the adjoint operator.

From now on, we assume that we are provided with a kernel K and we limit the hypothesis space to: $\mathcal{H} = \{f + \boldsymbol{b} : f \in \mathcal{K}_K, \|f\|_{\mathcal{K}} \leq c, \boldsymbol{b} \in \mathbb{R}^p\}$ (we have chosen the regularizer $\psi = \|\cdot\|_{\mathcal{K}}$). To this point, the choice of the kernel K is critical, since it controls both the data-dependent part of the hypothesis $f \in \mathcal{K}_K$ and the way the estimation procedure is regularized ($\|f\|_{\mathcal{K}} \leq c$).

Though several candidates are available Micchelli & Pontil (2005a); Alvarez et al. (2012); Baldassarre et al. (2012), we focus on one of the simplest and most efficiently computable kernels, called *decomposable kernel*: $K\colon (\mathbf{x}, \mathbf{x}') \mapsto k(\mathbf{x}, \mathbf{x}')\boldsymbol{B}$, where $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar-valued kernel and $\boldsymbol{B}$ is a $p \times p$ symmetric positive semi-definite matrix. In this particular case, the matrix $\boldsymbol{B}$ encodes the relationship between the components f_j and thus, the link between the different conditional quantile estimators. A rational choice is to consider $\boldsymbol{B} = \left(\exp(-\gamma(\tau_i - \tau_j)^2)\right)_{1\leq i,j\leq p}$. To explain it, let us consider two extreme cases.

First, when $\gamma = 0$, $\boldsymbol{B}$ is the all-ones matrix. Since $\mathcal{K}_K$ is the closure of $\operatorname{span}\{K_{\mathbf{x}}\boldsymbol{y} : (\mathbf{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathbb{R}^p\}$ (Senkene & Tempel'man, 1973), then any $f \in \mathcal{K}_K$ has all its components equal. This means that quantile estimations $h_j = f_j + b_j$ only differ thanks to the intercept b_j. Moreover, a straightforward application of the natural order property stated in Section 3.2 leads to $b_j \geq b_{j+1}$ when $\tau_j > \tau_{j+1}$. Consequently, the quantile estimators are parallel (and non-crossing) curves. In this case, we talk about a *homoscedastic* regressor (see Figure 1).

Second, when $\gamma \to +\infty$, then $\boldsymbol{B} \to \boldsymbol{I}$ (identity matrix). In this situation, it is easy to show that the components of $f \in \mathcal{K}_K$ are independent from each other and that $\|f\|_{\mathcal{K}}^2 = \sum_{j=1}^{p} \|f_j\|_{\mathcal{K}'}^2$, where $\|\cdot\|_{\mathcal{K}'}$ is the norm coming with the RKHS associated to k. Thus, each quantile function is learned independently from the others during the estimation procedure. We are then talking about *heteroscedastic* regressors (see Figure 1).

It appears clearly that between these two extreme cases, there is a room for learning a non-homescedastic and non-crossing quantile regressor by tuning γ. Before studying this fact empirically in Section 6.1, the next section will highlight a uniform convergence bound.

[3] A formal definition of a separable regularizer along with a proof of the previous statement is provided in Appendix A. This is an extension of the argument given by Koenker (2005) for a linear and unregularized estimator.

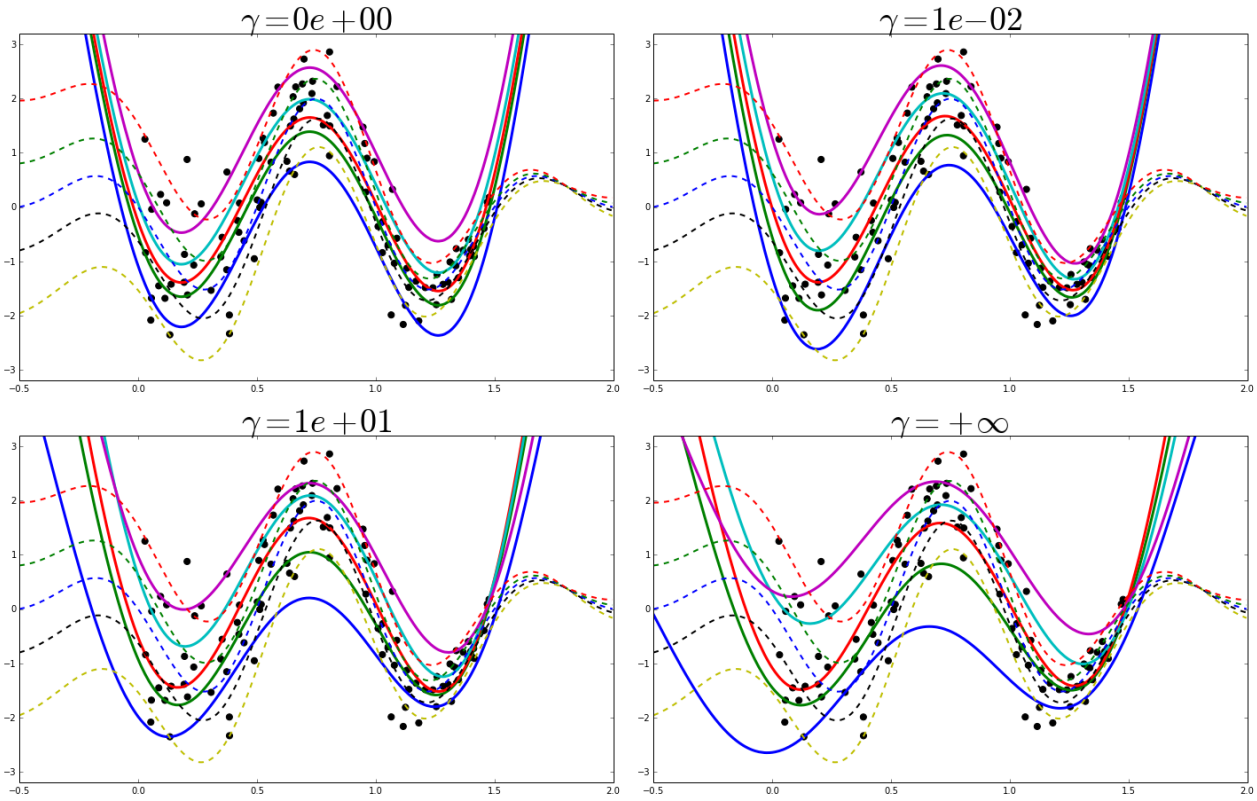

Figure 1: Estimated (plain lines) and true (dashed lines) conditional quantiles of $Y|X$ (synthetic dataset) from homoscedastic regressors ($\gamma = 0$) to heteroscedastic ones ($\gamma \to +\infty$).

4 Theoretical analysis

This section provides a theoretical analysis based on the Rademacher complexity of the hypothesis space, which is a standard technique to obtain uniform bounds for scalar-valued functions (Bartlett & Mendelson, 2002). Here, we do assume working in an RKHS but not specifically with a decomposable kernel.

The Rademacher complexity, previously defined for scalar-valued functions, is readily generalizable to vector-valued hypothesis spaces (Maurer, 2006). Following this trend, bounds for multiple-output RKHSs recently appeared (Sindhwani et al., 2013). In this last paper, however, the authors do not explicitly discuss how to use such a result in order to get a generalization bound, yet it is not trivial. For this reason, we propose to use a slightly different definition of the Rademacher complexity, which uses the maximum on each component instead of the summation of them. Given an *iid* sample $(X_i)_{1 \leq i \leq n} \in \mathcal{X}^n$ and independent Rademacher variables $(\epsilon_i)_{1 \leq i \leq n} \in \{-1, 1\}^n$ (*i.e.* uniformly distributed on $\{-1, 1\}$), the Rademacher complexity (or average) of a class $\mathcal{F} \subset (\mathbb{R}^p)^{\mathcal{X}}$ of functions is:

$$\mathcal{R}_n(\mathcal{F}) = \max_{1 \leq j \leq p} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f_j(X_i)\right],$$

where the expectation is computed jointly on $(X_i, \epsilon_i)_{1 \leq i \leq n}$.

The choice of the last definition is motivated by two facts. First, for scalar-valued hypotheses ($p = 1$), the last quantity boils down to the usual Rademacher average (Kakade et al., 2009). Second, it naturally appears in the extension to vector-valued hypotheses of Bartlett & Mendelson's generalization bound (for which we derive a full proof in Appendix A). Indeed, this definition makes easier getting the so called *composition lemma*, which is the key element to bridge the gap between bounding a risk and bounding the Rademacher average of a hypothesis space. This lemma is true for a mapping $\phi \colon \mathbb{R}^p \to \mathbb{R}$, that is separable ($\forall \boldsymbol{z} \in \mathbb{R}^p \colon \phi(\boldsymbol{z}) = \sum_{j=1}^{p} \phi_j(z_j)$ for some ϕ_j) and with each

component ϕ_j Lipschitz continuous.

Lemma 4.1 (Composition lemma)**.** *Let $\phi\colon \mathbb{R}^p \to \mathbb{R}$ be a separable Lipschitz continuous mapping with Lipschitz constant L_ϕ and denote $\Phi = \{\phi \circ f, f \in \mathcal{F}\}$. Then:*

$$\mathcal{R}_n(\Phi) \leq pL_\phi \mathcal{R}_n(\mathcal{F}).$$

Proof. In Appendix A. □

Given the composition lemma, the next step is to bound the Rademacher complexity of our hypothesis space $\mathcal{F} = \{f \in \mathcal{K}_K, \|f\|_{\mathcal{K}} \leq c\}$. It appears that this quantity is bounded the same way as in (Sindhwani et al., 2013). Let us denote $\mathrm{tr}(\cdot)$ the trace operator.

Lemma 4.2. *Assume that there exists $\kappa \geq 0$ such that: $\sup_{\mathbf{x}\in\mathcal{X}} \mathrm{tr}(K(\mathbf{x},\mathbf{x})) \leq \kappa$. Then:*

$$\mathcal{R}_n(\mathcal{F}) \leq c\sqrt{\frac{\kappa}{n}}.$$

Proof. Our proof (given in Appendix A) is an extension of Sindhwani et al.'s one for their own definition of the Rademacher average. □

The last step of the derivation is to combine these two lemmas in order to get a uniform generalization bound (Theorem 4.3). For this purpose, let $((X_i, Y_i))_{1\leq i\leq n} \in (\mathcal{X}\times\mathcal{Y})^n$ be an *iid* sample and denote $\hat{R}_n(h) = \frac{1}{n}\sum_{i=1}^n \ell_{\boldsymbol{\tau}}(Y_i\mathbb{1} - h(X_i))$, the random variable associated to the empirical risk of an hypothesis h.

Theorem 4.3 (Generalization)**.** *Let $a \in \mathbb{R}_+$ such that $\sup_{y\in\mathcal{Y}} |y| \leq a$, $\boldsymbol{b} \in \mathcal{Y}^p$ and $\mathcal{H} = \{f + \boldsymbol{b} : f \in \mathcal{F}\}$ be the class of hypotheses. Then under the assumption of Lemma 4.2 and with probability at least $1-\delta$ (for $\delta \in (0,1]$):* $\quad \forall h \in \mathcal{H}$,

$$R(h) \leq \hat{R}_n(h) + 2pc\sqrt{\frac{\kappa}{n}} + p(2a + c\sqrt{\kappa})\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Sketch of proof (full derivation in Appendix A). Following the technique developed by Bartlett & Mendelson (2002); Kakade et al. (2009) for real-valued functions, the proof uses McDiarmid's inequality. We prove that: $\forall (f,\mathbf{x}) \in \mathcal{F}\times\mathcal{X}, \|f(\mathbf{x})\|_{\ell_2} \leq c\sqrt{\kappa}$. Finally, $\ell_{\boldsymbol{\tau}}$ being separable 1-Lipschitz, Lemmas 4.1 and 4.2 give the middle term of the right hand side. □

The uniform bound in Theorem 4.3 states that, with high probability, all the hypotheses of interest have a true risk which is less that an empirical risk to an additive bias in $O(1/\sqrt{n})$. Let us remark that it makes use of the output dimension p. However, there exist non-uniform generalization bounds for operator-valued kernel-based hypotheses, which do not depend on the output dimension p (Audiffren & Kadri, 2013; Kadri et al., 2015), being thus well-suited for infinite-dimensional output spaces. Yet those results, only hold for optimal solutions $\hat{h}$ of the learning problem, which we never obtain in practice.

Finally, following the example of Takeuchi et al. (2006) and using the same technique as for Theorem 4.3, a bound on the quantile property can also be derived (see Theorem A.12 in Appendix A). This one states that $\mathbb{E}\left[\mathbb{P}\left(Y \leq h_j(X) \mid X\right)\right]$ does not deviate to much from τ_j.

5 Optimization algorithm

In order to finalize the M-estimation of a non-parametric function, we need a way to jointly solve the optimization problem of interest and compute the estimator. For ridge regression in vector-valued RKHSs, representer theorems enable to reformulate the estimator (Micchelli & Pontil, 2005b; Brouard et al., 2011) and to derive algorithms based on matrix inversion (Micchelli & Pontil, 2005b), Sylvester equation (Dinuzzo et al., 2011) or proximal gradient Lim et al. (2014). Since the optimization problem we are tackling is quite different, those methods can not be applied. Yet, we will shortly see that Karush-Kuhn-Tucker (KKT) conditions (see Appendix B) lead straightforwardly to a representer property and that an efficient coordinate descent can be devised.

Quantile estimation, as presented in this paper, comes down to minimizing a regularized empirical risk, defined by the pinball loss $\ell_{\boldsymbol{\tau}}$. Since this loss function is non-differentiable, we introduce slack variables $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$ to get the following primal formulation. We also consider a regularization parameter C to be tuned:

$$\begin{array}{ll} \underset{\substack{f\in\mathcal{K}_K, \boldsymbol{b}\in\mathbb{R}^p,\\ \boldsymbol{\xi},\boldsymbol{\xi}^*\in(\mathbb{R}^p)^n}}{\text{minimize}} & \frac{1}{2}\|f\|_{\mathcal{K}}^2 + C\sum_{i=1}^{n}\left(\langle\boldsymbol{\tau} \mid \boldsymbol{\xi}_i\rangle_{\ell_2} + \langle\mathbb{1}-\boldsymbol{\tau} \mid \boldsymbol{\xi}_i^*\rangle_{\ell_2}\right) \\ \text{s. t.} & \begin{cases} \forall i\in\mathbb{N}_n\colon \boldsymbol{\xi}_i \succcurlyeq 0, \boldsymbol{\xi}_i^* \succcurlyeq 0 \\ y_i - f(\mathbf{x}_i) - \boldsymbol{b} = \boldsymbol{\xi}_i - \boldsymbol{\xi}_i^* \quad : \boldsymbol{\alpha}_i \in \mathbb{R}^p, \end{cases} \end{array} \tag{2}$$

where $\succcurlyeq$ is a pointwise inequality. Problem (2) also introduces the dual vectors $\boldsymbol{\alpha}_i$ ($i\in\mathbb{N}_n$) corresponding to the linear constraint. In order to make the numerical approximation easier, we focus on a dual formulation of Problem (2) (see Appendix B for a full derivation):

$$\begin{array}{ll} \underset{\boldsymbol{\alpha}\in(\mathbb{R}^p)^n}{\text{minimize}} & \frac{1}{2}\sum_{i,j=1}^{n}\langle\boldsymbol{\alpha}_i \mid K(\mathbf{x}_i,\mathbf{x}_j)\boldsymbol{\alpha}_j\rangle_{\ell_2} - \sum_{i=1}^{n} y_i\langle\boldsymbol{\alpha}_i \mid \mathbb{1}\rangle_{\ell_2} \\ \text{s. t.} & \begin{cases} \forall i\in\mathbb{N}_n\colon C(\boldsymbol{\tau}-\mathbb{1}) \preccurlyeq \boldsymbol{\alpha}_i \preccurlyeq C\boldsymbol{\tau} \\ \displaystyle\sum_{i=1}^{n}\boldsymbol{\alpha}_i = 0_{\mathbb{R}^p}. \end{cases} \end{array} \tag{3}$$

The KKT conditions of Problem (2) indicate that a minimizer $\hat{f}$ of (2) can be recovered from a solution $\hat{\boldsymbol{\alpha}}$ of (3) with the formula $\hat{f} = \sum_{i=1}^{n} K_{\mathbf{x}_i}\hat{\boldsymbol{\alpha}}_i$. Moreover, $\hat{\boldsymbol{b}}$ can also be obtained thanks to KKT conditions. However, as we deal with a numerical approximate solution $\boldsymbol{\alpha}$, in practice $\boldsymbol{b}$ is computed by solving primal Optimization Problem (2) with all other variables fixed. This boils down to taking b_j as the τ_j-quantile of $(y_i - f_j(\mathbf{x}_i))_{1\leq i\leq n}$.

As soon as the $\boldsymbol{\alpha}_i$'s are stacked in a big vector, Optimization Problem (3) becomes a common quadratic program. However, since we are essentially interested in decomposable kernels $K(\cdot,\cdot) = k(\cdot,\cdot)\boldsymbol{B}$, it appears that the quadratic part of the objective function would be defined by the $np\times np$ matrix $\boldsymbol{K}\otimes\boldsymbol{B}$, where $\otimes$ is the Kronecker product and $\boldsymbol{K} = (k(\mathbf{x}_i,\mathbf{x}_j))_{1\leq i,j\leq n}$. Storing this matrix explicitly is likely to be time and memory expensive.

To overcome this issue, Dinuzzo et al. (2011) showed that ridge regression with a decomposable kernel boils down to solving a Sylvester equation (which can be done efficiently) and Minh et al. (2015) recently proposed a coordinate descent for vector-valued Support Vector Machine (SVM) without intercept. However, these methods can not be used in our setting since the loss function is different and considering the intercept is necessary for the quantile property. Theoretically, coordinate descent could be extended in an SMO technique, able to handle the linear constraint introduced by the intercept in Optimization Problem 3 (Platt, 1999). However, SMO works usually with a single linear constraint and needs heuristics to pick the pair of points of interest at each iteration. Even though it has already been implemented for two linear constraints (Takeuchi & Furuhashi, 2004), those heuristics are quite difficult to find.

Therefore, for the sake of efficiency, we propose to use a Primal-Dual Coordinate Descent (PDCD) technique, recently introduced by Fercoq & Bianchi (2015). This algorithm (which is proved to converge) is able to deal with the linear constraint coming from the intercept and is thus utterly workable for the problem at hand. Moreover, PDCD has been proved favorably competitive with SMO for SVMs.

From now on, for $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_i)_{1\leq i\leq n} \in (\mathbb{R}^p)^n$, let us denote $\boldsymbol{\alpha}^j$ the *row* vector $\left((\boldsymbol{\alpha}_i)_j\right)_{i=1}^{n} \in \mathbb{R}^n$ and α_i^j its i^{th} component. Moreover, let diag be the operator mapping a vector to a diagonal matrix and $\mathrm{proj}_{\mathbb{1}}$ and $\mathrm{proj}_{[C(\tau_l-1),C\tau_l]}$ be respectively the projectors onto the vector $\mathbb{1}$ and the compact set $[C(\tau_l-1), C\tau_l]$. PDCD is described in Algorithm 1. It uses dual variables $\boldsymbol{\theta}\in(\mathbb{R}^p)^n$ (which are updated during the descent) and has two sets of parameters $\boldsymbol{\nu}\in(\mathbb{R}^p)^n$ and $\boldsymbol{\mu}\in(\mathbb{R}^p)^n$, that verify ($\forall(i,l)\in\mathbb{N}_n\times\mathbb{N}_p$):

$$\mu_i^l < \frac{1}{(K(\mathbf{x}_i,\mathbf{x}_i))_{l,l} + \nu_i^l}.$$

In practice, we kept the same parameters as in (Fercoq & Bianchi, 2015): $\nu_i^l = 10(K(\mathbf{x}_i,\mathbf{x}_i))_{l,l}$ and μ_i^l equal to 0.95 times the bound. Moreover, as it is standard for coordinate descent methods, our implementation makes use of efficient updates for the computation of both $\sum_{j=1}^{n} K(\mathbf{x}_i,\mathbf{x}_j)\boldsymbol{\alpha}_j$ and of $\overline{\boldsymbol{\theta}}^l$.

Algorithm 1 Primal-Dual Coordinate Descent

Initialize $\boldsymbol{\alpha}_i, \boldsymbol{\theta}_i \in \mathbb{R}^p$ $(\forall i \in \mathbb{N}_n)$.
repeat
  Choose $(i, l) \in \mathbb{N}_n \times \mathbb{N}_p$ uniformly at random.
  Set $\overline{\boldsymbol{\theta}}^l \leftarrow \mathrm{proj}_{\mathbb{1}}\left(\boldsymbol{\theta}^l + \mathrm{diag}(\boldsymbol{\nu}^l)\boldsymbol{\alpha}^l\right)$.
  Set $d_i^l \leftarrow \sum_{j=1}^n (K(\mathbf{x}_i, \mathbf{x}_j)\boldsymbol{\alpha}_j)^l - y_i + 2\overline{\theta}_i^l - \theta_i^l$.
  Set $\overline{\alpha}_i^l \leftarrow \mathrm{proj}_{[C(\tau_l - 1), C\tau_l]}\left(\alpha_i^l - \mu_i^l d_i^l\right)$.
  Update coordinate (i, l): $\alpha_i^l \leftarrow \overline{\alpha}_i^l$, $\theta_i^l \leftarrow \overline{\theta}_i^l$,
  and keep other coordinates unchanged.
until duality gap (2)-(3) is small enough

6 Numerical experiments

Two sets of experiments are presented [4]. The first one is aimed at assessing the ability of the methodology introduced in this paper to predict quantiles. The second one compares an implementation of Algorithm 1 with an off-the-shelf solver and an augmented Lagrangian scheme.

Following the previous sections, a decomposable kernel $K(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}')\boldsymbol{B}$ is used, with $\boldsymbol{B} = (\exp(-\gamma(\tau_i - \tau_j)^2))_{1 \leq i,j \leq p}$. The parameter γ is either fixed or chosen by cross-validation. Moreover, $k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{\|\mathbf{x}-\mathbf{x}'\|_{\ell_2}^2}{2\sigma^2})$, with σ being the 0.7-quantile of the pairwise distances of the training data $\{\mathbf{x}_i\}_{1 \leq i \leq n}$. Eventually, regressors are build in order to estimate quantile functions for levels $\boldsymbol{\tau} = (0.1, 0.3, 0.5, 0.7, 0.9)$.

6.1 Quantile regression

Quantile regression is assessed with three criteria. First, the pinball loss $\frac{1}{n}\sum_{i=1}^n \ell_{\boldsymbol{\tau}}(y_i - h(\mathbf{x}_i))$ is the one minimized to build the proposed estimator. Second, the quantile loss $\sum_{j=1}^p \left[\left[\frac{1}{n}\sum_{i=1}^n \mathrm{I}_{\mathbb{R}_-}(y_i - h_j(\mathbf{x}_i))\right] - \tau_j\right]$, where $\mathrm{I}_{\mathbb{R}_-}$ is the indicator function of the set $\mathbb{R}_-$, measures the deviation of the estimators h_j to the prescribed quantile levels τ_j. Third, the crossing loss $\sum_{j=1}^{p-1}\left[\frac{1}{n}\sum_{i=1}^n \max(0, h_{j+1}(\mathbf{x}_i) - h_j(\mathbf{x}_i))\right]$, assuming that $\tau_j > \tau_{j+1}$, quantifies how far h_j goes below h_{j+1}, while h_j is expected to stay always above h_{j+1}.

This study is restricted to three non-parametric models based on the RKHS theory. Other linear and spline-based models have been dismissed since Takeuchi et al. (2006) have already provided a comparison of these ones with kernel methods. First, we considered an independent estimation of quantile regressors (IND.), which boils down to set $\boldsymbol{B} = \boldsymbol{I}$. This can be done out of the vector-valued RKHS theory, considering only scalar-valued kernels. Second, hard non-crossing constraints on the training data have been imposed (IND. (NC)), as proposed by Takeuchi et al. (2006). Third, the proposed joint estimator (JOINT) uses the Gaussian matrix $\boldsymbol{B}$ presented above.

These three methods are compared based on 20 regression datasets, which are the ones used in (Takeuchi et al., 2006). These datasets come from the UCI repository and three R packages: quantreg, alr3 and MASS. Their names are indicated in Table 1. The sample sizes vary from 38 (CobarOre) to 1375 (heights) and the numbers of explanatory variables vary from 1 (5 sets) to 12 (BostonHousing) [5]. The datasets were standardized coordinate-wise to have zero mean and unit variance. Results are given in Tables 1, 2 and 3 thanks to the mean and the standard deviation of the test losses recorded on 10 random splits train-test with ratio 0.7-0.3 (except for the dataset heights: 0.5-0.5). The best result of each line is boldfaced and the bullet indicate that it is significantly different from JOINT or from both IND. and IND. (NC). All these statements are based on a Wilcoxon signed-rank test with significance level 0.05.

The parameter C is chosen by cross-validation (minimizing the pinball loss) inside a logarithmic grid $(10^{-5}, 10^{-4}, \ldots, 10^5)$ for all methods and datasets. For our approach (JOINT), the parameter γ is chosen the same way as C but we also added 0 and $+\infty$ in the grid. Finally, the dual optimization problems corresponding to each approach are solved with CVXOPT (Anderson et al., 2012).

Regarding the pinball loss, (Table 1), joint quantile regression compares favorably to independent and hard non-crossing constraint estimations for 13 datasets (5 significantly different). These results bear out the assumption con-

[4] Numerical experiments ran on an Intel® Core™ i7-4600U CPU, operating at 2.10 GHz with 4 cores and 8 Gb of RAM.

[5] Categorical and indexing variables were dropped (see (Takeuchi et al., 2006)).

Table 1: Empirical pinball loss $\times 100$ (the less, the better).

DATA SET	IND.	IND. (NC)	JOINT
CAUTION	**99.01** $\pm$ 20.72	100.33 $\pm$ 20.54	99.46 $\pm$ 21.82
FTCOLLINSSNOW	152.13 $\pm$ 8.99	151.78 $\pm$ 8.84	**151.55** $\pm$ 8.43
HIGHWAY	107.14 $\pm$ 40.97	**107.08** $\pm$ 40.97	109.23 $\pm$ 35.24
HEIGHTS	127.93 $\pm$ 2.09	127.93 $\pm$ 2.09	• **127.47** $\pm$ 2.20
SNIFFER	45.29 $\pm$ 5.84	45.17 $\pm$ 5.87	**44.92** $\pm$ 5.22
SNOWGEESE	71.27 $\pm$ 32.52	**71.19** $\pm$ 32.54	80.25 $\pm$ 26.97
UFC	81.96 $\pm$ 3.76	82.08 $\pm$ 3.71	• **80.54** $\pm$ 3.90
BIRTHWT	139.93 $\pm$ 10.56	139.92 $\pm$ 10.55	**139.21** $\pm$ 12.91
CRABS	12.48 $\pm$ 0.83	12.46 $\pm$ 0.85	**12.19** $\pm$ 0.68
GAGURINE	62.61 $\pm$ 8.99	62.61 $\pm$ 8.98	**62.37** $\pm$ 8.58
GEYSER	108.07 $\pm$ 8.34	**108.06** $\pm$ 8.33	108.65 $\pm$ 8.46
GILGAIS	46.42 $\pm$ 4.76	46.25 $\pm$ 4.83	**45.67** $\pm$ 5.52
TOPO	67.65 $\pm$ 8.18	**66.63** $\pm$ 9.56	70.52 $\pm$ 8.93
BOSTONHOUSING	50.12 $\pm$ 6.14	50.05 $\pm$ 6.13	• **48.97** $\pm$ 5.52
COBARORE	• **0.54** $\pm$ 0.62	0.54 $\pm$ 0.62	0.63 $\pm$ 0.62
ENGEL	59.28 $\pm$ 7.18	**58.77** $\pm$ 6.32	64.96 $\pm$ 17.62
MCYCLE	83.48 $\pm$ 7.77	83.15 $\pm$ 7.64	• **78.92** $\pm$ 8.43
BIGMAC2003	70.25 $\pm$ 21.11	69.90 $\pm$ 21.59	• **66.24** $\pm$ 19.62
UN3	101.95 $\pm$ 8.26	101.86 $\pm$ 8.21	**100.31** $\pm$ 6.97
CPUS	18.83 $\pm$ 15.55	18.81 $\pm$ 15.58	**18.73** $\pm$ 15.57

Table 2: Empirical quantile loss $\times 100$ (the closer to 0, the better).

DATA SET	IND.	IND. (NC)	JOINT
CAUTION	12.00 $\pm$ 38.79	**9.33** $\pm$ 35.80	13.00 $\pm$ 35.54
FTCOLLINSSNOW	6.79 $\pm$ 47.85	6.79 $\pm$ 47.85	**5.36** $\pm$ 43.89
HIGHWAY	-9.17 $\pm$ 64.38	-8.33 $\pm$ 65.30	**-5.00** $\pm$ 61.26
HEIGHTS	-0.16 $\pm$ 6.03	**-0.13** $\pm$ 6.12	-0.65 $\pm$ 6.13
SNIFFER	-10.00 $\pm$ 22.39	-10.26 $\pm$ 25.17	**-5.79** $\pm$ 26.05
SNOWGEESE	• **3.57** $\pm$ 57.17	3.57 $\pm$ 55.54	-8.57 $\pm$ 58.19
UFC	6.07 $\pm$ 11.06	6.34 $\pm$ 10.91	**3.93** $\pm$ 12.65
BIRTHWT	16.49 $\pm$ 24.00	16.49 $\pm$ 24.00	**15.61** $\pm$ 28.45
CRABS	**3.67** $\pm$ 23.16	5.33 $\pm$ 24.66	5.00 $\pm$ 21.90
GAGURINE	-2.53 $\pm$ 20.42	**-2.53** $\pm$ 20.48	-3.47 $\pm$ 18.53
GEYSER	**4.44** $\pm$ 19.26	5.33 $\pm$ 19.29	7.00 $\pm$ 18.16
GILGAIS	4.36 $\pm$ 22.19	3.91 $\pm$ 22.22	**3.09** $\pm$ 23.90
TOPO	-16.25 $\pm$ 63.57	**-12.50** $\pm$ 63.92	-20.00 $\pm$ 62.80
BOSTONHOUSING	10.39 $\pm$ 19.14	10.46 $\pm$ 18.85	**7.76** $\pm$ 14.68
COBARORE	**29.17** $\pm$ 80.99	29.17 $\pm$ 80.99	44.17 $\pm$ 70.12
ENGEL	-8.45 $\pm$ 20.52	-8.73 $\pm$ 20.60	**-7.46** $\pm$ 20.20
MCYCLE	11.25 $\pm$ 31.71	**10.25** $\pm$ 35.42	13.75 $\pm$ 36.88
BIGMAC2003	-4.76 $\pm$ 38.47	**-2.38** $\pm$ 38.69	7.14 $\pm$ 40.01
UN3	**1.90** $\pm$ 24.03	2.22 $\pm$ 23.93	4.29 $\pm$ 24.17
CPUS	3.33 $\pm$ 20.29	4.60 $\pm$ 20.80	**3.02** $\pm$ 23.85

cerning the relationship between conditional quantiles and the usefulness of multiple-output methods for quantile regression.

The quantile loss is quite an equivocal criterion, since it measures *how much* the *unconditional* quantile property is satisfied. This unconditional indicator is indeed the only way to get a piece of information concerning the *conditional* quantile property. For instance, Takeuchi et al. (2006) empirically showed (with the same datasets) that the constant function based on the unconditional quantile estimator performs best under this criterion, even though it is expected to be a poor conditional quantile regressor. The numerical results in Table 2 follow this remark and the results previously obtained (Takeuchi et al., 2006). No significant ranking comes out.

Last but not least, the results for the crossing loss (Table 3) clearly show that joint regression enables to weaken the crossing problem, in comparison to independent estimation and hard non-crossing constraints (13 favorable datasets and 6 significantly different). Note that for the estimation with hard non-crossing constraints (IND. (NC)), the crossing loss is null on the training data but is not guaranteed to be null on the test data. In addition, let us remark that model selection (and particularly for the parameter γ, which tunes the trade-off between hetero and homoscedastic regressors) has been performed based on the pinball loss only. It seems that, in a way, the pinball loss embraces the crossing loss as a subcriterion.

6.2 Training algorithms

This section is aimed at comparing three implementations of algorithms for estimating joint quantile regressors, following their running (CPU) time. First, the off-the-shelf solver included in CVXOPT (Anderson et al., 2012) (QP) is applied to Optimization Problem (3) turned into a standard form of linearly constrained quadratic program. This solver is based on an interior-point method. Second, an augmented Lagrangian scheme (AUG. LAG) is used in order to get

Table 3: Empirical crossing loss $\times 100$ (the less, the better).

DATA SET	IND.	IND. (NC)	JOINT
CAUTION	0.46 ± 0.74	0.38 ± 0.95	**0.07** ± 0.10
FTCOLLINSSNOW	**0.00** ± 0.00	0.00 ± 0.00	0.00 ± 0.00
HIGHWAY	10.01 ± 7.88	9.90 ± 7.93	**9.52** ± 8.10
HEIGHTS	0.03 ± 0.05	0.01 ± 0.02	**0.00** ± 0.00
SNIFFER	0.93 ± 0.67	0.48 ± 0.63	**0.10** ± 0.17
SNOWGEESE	2.92 ± 2.66	2.17 ± 2.32	**1.68** ± 4.77
UFC	0.22 ± 0.22	0.33 ± 0.58	• **0.02** ± 0.07
BIRTHWT	0.00 ± 0.00	0.00 ± 0.00	**0.00** ± 0.00
CRABS	0.47 ± 0.28	0.40 ± 0.25	• **0.13** ± 0.27
GAGURINE	0.06 ± 0.08	**0.05** ± 0.07	0.05 ± 0.10
GEYSER	0.60 ± 1.41	**0.60** ± 1.41	0.82 ± 1.49
GILGAIS	0.95 ± 0.27	• **0.69** ± 0.23	0.89 ± 0.42
TOPO	1.83 ± 1.25	**0.67** ± 0.90	1.79 ± 2.53
BOSTONHOUSING	0.64 ± 0.20	• **0.47** ± 0.18	0.62 ± 0.26
COBARORE	0.10 ± 0.15	0.10 ± 0.15	• **0.02** ± 0.03
ENGEL	0.33 ± 0.62	**0.03** ± 0.04	0.09 ± 0.18
MCYCLE	2.77 ± 2.23	1.30 ± 1.45	• **0.07** ± 0.14
BIGMAC2003	2.24 ± 2.30	1.63 ± 1.60	**1.05** ± 1.26
UN3	0.85 ± 0.52	0.67 ± 0.43	• **0.14** ± 0.41
CPUS	0.91 ± 0.34	0.85 ± 0.33	• **0.15** ± 0.15

Table 4: Comparison of CPU time (s) for training a model.

SIZE	QP	AUG. LAG.	PDCD
100	**0.85** ± 0.07	49.21 ± 12.28	3.76 ± 0.80
250	**8.73** ± 0.34	261.11 ± 46.69	18.69 ± 3.54
500	75.53 ± 2.98	865.86 ± 92.26	**61.30** ± 7.05
1000	621.60 ± 30.37	–	**266.50** ± 41.16

rid of the linear constraint that defines the intercept [6]. In this scheme, the inner solver is the algorithm proposed in (Shalev-Shwartz & Zhang, 2013), which boils down to be the same as Algorithm 1 when the intercept is dismissed. The last approach (PDCD) is Algorithm 1.

We use a synthetic dataset (the same as in Figure 1), for which $X \in [0, 1.5]$. The target Y is computed as a sine curve at 1 Hz modulated by a sine envelope at 1/3 Hz and mean 1. Moreover, this pattern is distorted with a random Gaussian noise with mean 0 and a linearly decreasing standard deviation from 1.2 at $X = 0$ to 0.2 at $X = 1.5$.

Parameters for the models are: $(C, \gamma) = (10^2, 10^{-2})$. To compare the implementations of the three algorithms, we first run CVXOPT (QP), with a relative tolerance set to 10^{-2}, and store the *optimal* objective value. Then, the two other methods (AUG. LAG and PDCD) are launched and stopped when they pass the objective value reached by QP [7]. Table 4 gives the mean and standard deviation of the CPU time required by each method for 10 random datasets and several sample sizes. Some statistics are missing because AUG. LAG. ran out of time.

As expected, it appears that for a not too tight tolerance and big datasets, implementation of Algorithm 1 outperforms the two other competitors. Let us remark that CVXOPT is also more expensive in memory than the coordinate-based algorithms like ours. To conclude, training time may seem high in comparison to usual SVMs. However, let us first remind that we jointly learn p regressors. Thus, a fair comparison should be done with an SVM applied to an $np \times np$ matrix, instead of $n \times n$. Moreover, there is no *natural* sparsity in quantile regression (while there is one in SVM), which slows down computation.

7 Conclusion

This paper introduces a novel methodology for joint quantile regression, which is based on vector-valued RKHSs. It comes along with theoretical guarantees and an efficient learning algorithm. Moreover, this framework enjoys enhanced performances and few occurrences of curve crossing, compared to independent estimations and hard non-crossing constraints.

[6] The procedure AUG. LAG is detailed in Appendix C.

[7] During the descent, we used an efficient accumulated objective value, which is not exact since the iterate $\boldsymbol{\alpha}$ is not feasible to the linear constraint. Appendix D gives the average objective values reached by each algorithm after projection of the best candidate onto the set of constraints. We can check that PDCD always reach a smaller objective value than the target QP.

As a future work, we envisage: **i)** an empirical comparison with the several approaches from the literature; **ii)** a theoretical analysis of the crossing probability; **iii)** a tighter and dimensional-free generalization bound (Boucheron et al., 2005; Kadri et al., 2015); **iv)** predicting all quantile curves (Takeuchi et al., 2013; Kadri et al., 2015); **v)** an extension to the multivariate setting (Hallin et al., 2010).

References

Alvarez, M.A., Rosasco, L., and Lawrence, N.D. Kernels for Vector-Valued Functions: a Review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012. arXiv: 1106.6251.

Anderson, M.S., Dahl, J., and Vandenberghe, L. CVXOPT: A Python package for convex optimization, version 1.1.5., 2012.

Argyriou, A., Evgeniou, T., and Pontil, M. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Audiffren, J. and Kadri, H. Stability of Multi-Task Kernel Regression Algorithms. In *Proceedings of 5th Asian Conference on Machine Learning*, 2013.

Baldassarre, L., Rosasco, L., Barla, A., and Verri, A. Multi-output learning via spectral filtering. *Machine Learning*, 87(3):259–301, 2012.

Bartlett, P.L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Bondell, H.D., Reich, B.J., and Wang, H. Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838, 2010.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of Classification: a Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Brouard, C., d'Alché Buc, F., and Szafranski, M. Semi-supervised Penalized Output Kernel Regression for Link Prediction. In *Proceedings of The 28th International Conference on Machine Learning*, 2011.

Brouard, C., d'Alché Buc, F., and Szafranski, M. Input Output Kernel Regression. *hal-01216708 [cs]*, 2015.

Carmeli, C., De Vito, E., Toigo, A., and Umanità, V. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 08(01):19–61, 2010.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. Quantile and Probability Curves Without Crossing. *Econometrica*, 78(3):1093–1125, 2010.

Ciliberto, C., Mroueh, Y., Poggio, T., and Rosasco, L. Convex Learning of Multiple Tasks and their Structure. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Dette, H. and Volgushev, S. Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):609–627, 2008.

Dinuzzo, F., Ong, C.S., Gehler, P., and Pillonetto, G. Learning Output Kernels with Block Coordinate Descent. In *Proceedings of the 28th International Conference of Machine Learning*, 2011.

Evgeniou, T., Micchelli, C.A., and Pontil, M. Learning Multiple Tasks with Kernel Methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

Fercoq, O. and Bianchi, P. A Coordinate Descent Primal-Dual Algorithm with Large Step Size and Possibly Non Separable Functions. *arXiv:1508.04625 [math]*, 2015.

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

Hallin, M., Lu, Z., Paindaveine, D., and Šiman, M. Local bilinear multiple-output quantile/depth regression. *Bernoulli*, 21(3):1435–1466, 2015.

Hallin, M. and Šiman, M. Elliptical multiple-output quantile regression and convex optimization. *Statistics & Probability Letters*, 109:232–237, 2016.

He, X. Quantile Curves without Crossing. *The American Statistician*, 51(2):186–192, 1997.

Jebara, T. Multi-task Feature and Kernel Selection for SVMs. In *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004.

Kadri, H., Duflos, E., Preux, P., Canu, S., and Davy, M. Nonlinear functional regression: a functional RKHS approach. In *Proceedings og the 13th International Conference on Artificial Intelligence and Statistics (AISTATS'10)*, 2010.

Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A., and Audiffren, J. Operator-valued Kernels for Learning from Functional Response Data. *Journal of Machine Learning Research*, 16:1–54, 2015.

Kakade, S.M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*, 2009.

Koenker, R. *Quantile Regression*. Cambridge University Press, Cambridge, New York, 2005.

Koenker, Roger and Bassett, Jr., Gilbert. Regression Quantiles. *Econometrica*, 46(1):33–50, 1978.

Li, Y., Liu, Y., and Zhu, J. Quantile Regression in Reproducing Kernel Hilbert Spaces. *Journal of the American Statistical Association*, 102(477):255–268, 2007.

Lim, N., d'Alché Buc, F., Auliac, C., and Michailidis, G. Operator-valued kernel-based vector autoregressive models for network inference. *Machine Learning*, 99(3):489–513, 2014.

Liu, Y. and Wu, Y. Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of nonparametric statistics*, 23(2):415–437, 2011.

Maurer, A. The Rademacher Complexity of Linear Transformation Classes. In *Proceedings of the 19th Annual Conference on Learning Theory*, 2006.

Micchelli, C.A. and Pontil, M. Kernels for Multi–task Learning. In *Advances in Neural Information Processing Systems 17*, 2005a.

Micchelli, C.A. and Pontil, M. Learning the Kernel Function via Regularization. *Journal of Machine Learning Research*, 6:1099–1125, July 2005b.

Minh, H.Q. and Sindhwani, V. Vector-valued Manifold Regularization. In *Proceedings of The 28th International Conference on Machine Learning*, 2011.

Minh, H.Q., Kang, S.H., and Le, T.M. Image and Video Colorization Using Vector-Valued Reproducing Kernel Hilbert Spaces. *Journal of Mathematical Imaging and Vision*, 37(1):49–65, 2010.

Minh, H.Q., Bazzani, L., and Murino, V. A Unifying Framework in Vector-valued Reproducing Kernel Hilbert Spaces for Manifold Regularization and Co-Regularized Multi-view Learning. *Journal of Machine Learning Research*, 2015. To appear.

Mroueh, Y., Poggio, T., Rosasco, L., and Slotine, J.-J. Multiclass Learning with Simplex Coding. In *Advances in Neural Information Processing Systems 25*, pp. 2789–2797. Curran Associates, Inc., 2012.

Platt, J.C. *Fast training of support vector machines using sequential minimal optimization*. Advances in Kernel Methods. MIT Press, Cambridge, MA, USA, 1999.

Rosset, S. Bi-Level Path Following for Cross Validated Solution of Kernel Quantile Regression. *Journal of Machine Learning Research*, 10:2473–2505, 2009.

Schnabel, S.K. and Eilers, P.H.C. Simultaneous estimation of quantile curves using quantile sheets. *AStA Advances in Statistical Analysis*, 97(1):77–87, 2012.

Senkene, E. and Tempel'man, A. Hilbert spaces of operator-valued functions. *Lithuanian Mathematical Journal*, 13 (4):665–670, 1973.

Shalev-Shwartz, S. and Zhang, T. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. *Journal of Machine Learning Research*, 14:567–599, 2013.

Shim, J., Hwang, C., and Seok, K.H. Non-crossing quantile regression via doubly penalized kernel machine. *Computational Statistics*, 24(1):83–94, 2009.

Sindhwani, V., Quang, M.H., and Lozano, A.C. Scalable Matrix-valued Kernel Learning for High-dimensional Nonlinear Multivariate Regression and Granger Causality. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013.

Steinwart, I. and Christmann, A. How SVMs can estimate quantiles and the median. In *Advances in Neural Information Processing Systems 20*, 2008.

Takeuchi, I., Le, Q.V., Sears, T.D., and Smola, A.J. Nonparametric Quantile Estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.

Takeuchi, I., Hongo, T., Sugiyama, M., and Nakajima, S. Parametric Task Learning. In *Advances in Neural Information Processing Systems 26*, pp. 1358–1366. Curran Associates, Inc., 2013.

Takeuchi, Ichiro and Furuhashi, T. Non-crossing quantile regressions by SVM. In *2004 IEEE International Joint Conference on Neural Networks, 2004. Proceedings*, volume 1, pp. –406, July 2004. doi: 10.1109/IJCNN.2004.1379939.

Wu, Y. and Liu, Y. Stepwise multiple quantile regression estimation using non-crossing constraints. *Statistics and Its Interface*, 2:299–310, 2009.