RÉGRESSION QUANTILE CONJOINTE ET RKHS

Maxime Sangnier [1], Olivier Fercoq [1] & Florence d'Alché-Buc [1]

[1] *LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France,*
prenom.nom@telecom-paristech.fr

Résumé. Tirant profit de l'apprentissage multi-tâche à noyaux, nous proposons une nouvelle méthodologie non-paramétrique pour estimer et prédire simultanément plusieurs quantiles conditionnels. L'une des particularités de celle-ci est de restreindre considérablement le phénomène disgracieux de croisement des courbes estimées. De plus, le cadre méthodologique proposé est accompagné d'une borne de généralisation uniforme et d'un algorithme efficace d'estimation. Des résultats numériques sur des données réelles de référence garantissent empiriquement les améliorations de notre approche quant à l'erreur de prédiction et à l'apparition de croisements de courbes.

Mots-clés. Noyau à valeurs opérateurs, estimation non-paramétrique, apprentissage multi-tâche, apprentissage statistique, descente par coordonnées.

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 and the crossing occurrences.

Keywords. Operator-valued kernel, non-parametric estimation, multi-task learning, statistical learning, coordinate descent.

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 and Bassett [18] as the minimization of the pinball loss (see

[17] for an extensive presentation) and brought to the attention of the machine learning community by Takeuchi et al. [27], Rosset [25]. 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) [11, 10, 12]. On the contrary, we are interested in estimating and predicting simultaneously several quantiles of a scalar-valued random variable $Y|X$, 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 [13, 9, 2, 7]. As a consequence, vector-valued kernel methods [21] are appropriate for such a task. They have already been used for various applications, such as image colorization [22], classification [8, 24], manifold regularization [23, 5], vector autoregression [19], functional regression [14, 15] and structured regression [6]. 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 [27].

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 (the description of which has been omitted due to a lack of space). 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 Quantile estimation

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. Given a vector $\boldsymbol{\tau} \in (0,1)^p$ of quantile levels, the paradigm is to estimate the vector-valued function of conditional quantiles $\mathbf{x} \in \mathcal{X} \mapsto (\mu_{\tau_1}(\mathbf{x}), \ldots, \mu_{\tau_p}(\mathbf{x})) \in \mathbb{R}^p$, where $\mu_{\tau_j}(\mathbf{x}) = \min\{\mu \in \mathbb{R} : \mathbb{P}(Y \leq \mu \mid X = \mathbf{x}) = \tau_j\}$.

Suppose we are provided with an independent and identically distributed (*iid*) sample of

observations $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$ and a matrix-valued kernel $K\colon \mathcal{X}\times\mathcal{X} \to \mathcal{L}(\mathbb{R}^p)$, where $\mathcal{L}(\mathbb{R}^p)$ is the set of linear and bounded operators from $\mathbb{R}^p$ to itself [26, 21], and let $\mathcal{K}_K \subset (\mathbb{R}^p)^{\mathcal{X}}$ be the RKHS associated to K (with a norm denoted $\|\cdot\|_{\mathcal{K}}$). Extending the work [17, 27] (regarding single quantile estimation), conditional quantiles can be estimated by minimization of the empirical risk within a class $\mathcal{H} = \{f + \boldsymbol{b} : f \in \mathcal{K}_K, \|f\|_{\mathcal{K}} \leq c, \boldsymbol{b} \in \mathbb{R}^p\}$ (with $c > 0$) of functions:

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

where $\mathbb{1}$ stands for the all-ones vector and the pinball loss $\ell_{\boldsymbol{\tau}}$ is defined for all $\boldsymbol{r} \in \mathbb{R}^p$ by:

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

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 [18]. In addition, one can show that joint conditional quantiles are minimizers of the true risk: $R\colon h \in (\mathbb{R}^p)^{\mathcal{X}} \mapsto \mathbb{E}\left[\ell_{\boldsymbol{\tau}}(Y\mathbb{1} - h(X))\right]$. To this point, let us remark that 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 ($\|f\|_{\mathcal{K}} \leq c$). The forthcoming section illustrates the room for learning a non-homescedastic and non-crossing quantile regressor by tuning the kernel K.

Now, we state a uniform generalization bound for the model at hand. This result is based on an extension of the Rademacher complexity to vector-valued hypotheses, which is a standard technique to obtain uniform bounds for scalar-valued functions [4]. 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 2.1 (Generalization)**.** *Let $a \in \mathbb{R}_+$ such that $\sup_{y\in\mathcal{Y}} |y| \leq a$, $\boldsymbol{b} \in \mathcal{Y}^p$ and $\mathcal{H} = \{f + \boldsymbol{b} : f \in \mathcal{K}_K, \|f\|_{\mathcal{K}} \leq c\}$ be the class of hypotheses, Assume that there exists $\kappa \geq 0$ such that: $\sup_{\mathbf{x}\in\mathcal{X}} \operatorname{tr}(K(\mathbf{x},\mathbf{x})) \leq \kappa$ and let $\delta \in (0, 1]$. Then:*

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left(R(h) - \hat{R}_n(h)\right) > 2pc\sqrt{\frac{\kappa}{n}} + p(2a + c\sqrt{\kappa})\sqrt{\frac{\log(1/\delta)}{2n}}\right) \leq \delta.$$

Sketch of proof. Following the technique developed in [4, 16] for real-valued functions, we start with an extension of the Rademacher complexity to vector-valued functions and prove the so called *composition lemma*. Then, we bound the Rademacher average and the pinball loss using the fact that $\forall (f, \mathbf{x}) \in \mathcal{F}\times\mathcal{X}, \|f(\mathbf{x})\|_{\ell_2} \leq c\sqrt{\kappa}$. Finally, we use McDiarmid's inequality. □

In practice, a quantile regressor $\hat{h} = \hat{f} + \hat{\boldsymbol{b}}$ is obtained by maximization of an optimization problem dual to (1). Then, Karush-Kuhn-Tucker (KKT) conditions indicate that $\hat{f}(\cdot) = \sum_{i=1}^n K(\cdot, \mathbf{x}_i)\hat{\boldsymbol{\alpha}}_i$, where $\hat{\boldsymbol{\alpha}} \in (\mathbb{R}^p)^n$ is a solution of the dual problem. Moreover, $\hat{\boldsymbol{b}}$ can also be obtained thanks to KKT conditions.

3 Numerical experiments

Though several candidates are available [20, 1, 3] for the kernel K, 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. Since the matrix $\boldsymbol{B}$ encodes the relationship between the components f_j, we set $\boldsymbol{B} = (\exp(-\gamma(\tau_i - \tau_j)^2))_{1 \leq i,j \leq p}$, where $\gamma \geq 0$. Ranging from 0 to $+\infty$, the parameter γ offers versatile hypotheses from homoscedastic to heteroscedastic ones. In practice, γ, as well as c, are chosen by cross-validation (minimizing the pinball loss). Moreover, we chose $k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{\|\mathbf{x}-\mathbf{x}'\|_{\ell_2}^2}{2\sigma^2})$, with σ being the 0.7-quantile of the pairwise distances of the training data $\{\mathbf{x}_i\}_{1 \leq i \leq n}$. Eventually, quantile levels considered are $\boldsymbol{\tau} = (0.9, 0.7, 0.5, 0.3, 0.1)$.

Quantile regression is assessed with two 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 crossing loss $\sum_{j=1}^{p-1}\left[\frac{1}{n}\sum_{i=1}^{n}\max(0, h_{j+1}(\mathbf{x}_i) - h_j(\mathbf{x}_i))\right]$ quantifies how far h_j goes below h_{j+1}, while h_j is expected to stay always above h_{j+1}. Moreover, this study is restricted to three non-parametric models based on the RKHS theory. Other linear and spline-based models have been dismissed since [27] 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 setting $\boldsymbol{B} = \boldsymbol{I}$. 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 in [27]. 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 [27]. These datasets come from the UCI repository and three R packages: quantreg, alr3 and MASS. Results are given in Table 1 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. The best result of each line is boldfaced and the bullet indicates 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.

Regarding the pinball loss, 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 concerning the relationship between conditional quantiles and the usefulness of multiple-output methods for quantile regression.

In addition, the results for the crossing loss 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.

Table 1: Empirical pinball loss (left table) and crossing loss (right table). The less, the better.

Data set	Ind.	Ind. (nc)	Joint
caution	**99.01** ± 20.72	100.33 ± 20.54	99.46 ± 21.82
ftcollinssnow	152.13 ± 8.99	151.78 ± 8.84	**151.55** ± 8.43
highway	107.14 ± 40.97	**107.08** ± 40.97	109.23 ± 35.24
heights	127.93 ± 2.09	127.93 ± 2.09	• **127.47** ± 2.20
sniffer	45.29 ± 5.84	45.17 ± 5.87	**44.92** ± 5.22
snowgeese	71.27 ± 32.52	**71.19** ± 32.54	80.25 ± 26.97
ufc	81.96 ± 3.76	82.08 ± 3.71	• **80.54** ± 3.90
birthwt	139.93 ± 10.56	139.92 ± 10.55	**139.21** ± 12.91
crabs	12.48 ± 0.83	12.46 ± 0.85	**12.19** ± 0.68
GAGurine	62.61 ± 8.99	62.61 ± 8.98	**62.37** ± 8.58
geyser	108.07 ± 8.34	**108.06** ± 8.33	108.65 ± 8.46
gilgais	46.42 ± 4.76	46.25 ± 4.83	**45.67** ± 5.52
topo	67.65 ± 8.18	**66.63** ± 9.56	70.52 ± 8.93
BostonHousing	50.12 ± 6.14	50.05 ± 6.13	• **48.97** ± 5.52
CobarOre	• **0.54** ± 0.62	0.54 ± 0.62	0.63 ± 0.62
engel	59.28 ± 7.18	**58.77** ± 6.32	64.96 ± 17.62
mcycle	83.48 ± 7.77	83.15 ± 7.64	• **78.92** ± 8.43
BigMac2003	70.25 ± 21.11	69.90 ± 21.59	• **66.24** ± 19.62
UN3	101.95 ± 8.26	101.86 ± 8.21	**100.31** ± 6.97
cpus	18.83 ± 15.55	18.81 ± 15.58	**18.73** ± 15.57

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[17] Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, Cambridge, New York.

[18] Koenker, R. and Bassett, Jr., G. (1978). Regression Quantiles. *Econometrica*, 46(1):33–50.

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

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

[21] Micchelli, C. and Pontil, M. (2005b). Learning the Kernel Function via Regularization. *Journal of Machine Learning Research*, 6:1099–1125.

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

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

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

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

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

[27] Takeuchi, I., Le, Q., Sears, T., and Smola, A. (2006). Nonparametric Quantile Estimation. *Journal of Machine Learning Research*, 7:1231–1264.