# A review on quantile regression for stochastic computer experiments

Léonard Torossian
*INRA (MIAT unit) - Institute of Mathematics, University of Toulouse*

**Supervisor(s):** DR. Robert Faivre (INRA, MIAT), Prof. Aurélien Garivier (IMT) and CR. Victor Picheny (INRA, MIAT)

**Ph.D. expected duration:** Nov. 2016 - Oct. 2019

**Adress**: INRA, 24 chemin de Borde Rouge, 31320 Auzeville-Tolosane

**Email**: leonard.torossian@inra.fr

**Abstract:** We consider a stochastic computer code of the form:

$$f : \mathcal{X} \times \Omega \to \mathbb{R}$$

with $\mathcal{X} \subset \mathbb{R}^D$ a design space and $\Omega$ a stochastic space. Contrary to deterministic black boxes, at a fixed $x$, the output is a random variable $Y_x = f(x, \omega)$ that follows an unknow distribution $\mathbb{P}(Y|X = x)$. We assume the classical constraints of computer experiments, that are: the function is only accessible through pointwise evaluations $f(x, \omega)$ ; no structural information is available regarding $f$ ; evaluations may be expensive, which limits drastically the number of calls of $f$. We assume further that

- the variance of $\mathbb{P}(Y|X = x)$ may vary with respect to $x$ (heteroscedasticity),
- the distribution has a non-parametric form and its shape may vary with respect to $x$.

Our objective is to explicit the link between $x$ and $Y$ in order to choose the $x^* \in \mathcal{X}$ that optimizes an indicator based on $\mathbb{P}(Y|X = x)$, typically the conditional expectation. However, the expectation is risk-neutral, as it does not account for the variability of $\mathbb{P}(Y|X = x)$. Here, we choose to focus on a number $s$ of conditional quantiles of order $\tau_1, .., \tau_s$ preliminarily fixed, with the conditional quantile defined as $q_\tau(x) = \{\inf q : F(q|X = x) \geq \tau\}$, where $F(.|X = x)$ is the cumulative distribution of $\mathbb{P}(Y|X = x)$.

In this work, we propose a review of the metamodeling approaches dedicated to the approximation of one or several conditional quantiles. Since the literature on quantile regression is very large, we restrict our review to the approaches that are best-suited for the framework defined above, while ensuring a good diversity of the metamodels. In particular, we selected

- two methods based on neighbourhood: $K$-Nearest Neighbours regression [2] and Random Forest regression [6]
- two methods based on functional analysis: Neural Networks regression [3] and RKHS regression [9]
- two methods based on stochastic processes: Quantile Kriging [8] and Bayesian Quantile regression [1]

Based on a set of observations $\mathcal{D}_n = ((x_1, y_1), ..., (x_n, y_n))$, each approach extracts the quantile estimator in a different way. Neighborhood-based approaches select a subset of observations that are close to each other in order to use their order statistic as a local quantile estimator. Grouping

observations could be made thanks to the Euclidian distance on $\mathcal{X}$ (K-Nearest Neighbours) or by a tree-based method (Random Forest). Functional approches consist in choosing a functional space $\mathcal{H}$ for the quantile estimator, then in optimizing the empirical risk (on $\mathcal{D}_n$) associated to a well-chosen loss function (the so-called *pinball loss* function). The main difference between the two functional approaches lies in the black box aspect of the neural network. Studying the neural network doesn't give any insights on the structure of the function being approximated while the RKHS regression defines $\mathcal{H}$ using a kernel function and several properties can be linked to each kernel. Finally, the random processes approaches we have selected are based on the assumption that the quantile is a realization of a Gaussian process. Since it is a latent (unobservable) process, the Quantile Kriging forces $\mathcal{D}_n$ to have repeated values of the $x_i$'s in order to extract order statistics, while the Bayesian Quantile regression uses a specific assumption on the distribution of the observations $y_i$ and a variational approach.

The performance of the six metamodels is analyzed through a benchmark composed of two toy functions and an agronomical model [4]. The dimensions of the problems vary from 1 to 8 and the number of observations from 80 to 2000.The relative cost and robustness of the approaches is also discussed, in particular with respect to hyperparameter tuning

## References

[1] Sachinthaka Abeywardana and Fabio Ramos. Variational inference for nonparametric bayesian quantile regression. In *AAAI*, pages 1686–1692, 2015.

[2] Pallab K Bhattacharya and Ashis K Gangopadhyay. Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics*, pages 1400–1415, 1990.

[3] Alex J Cannon. Quantile regression neural networks: Implementation in r and application to precipitation downscaling. *Computers & geosciences*, 37(9):1277–1284, 2011.

[4] Pierre Casadebaig, Lydie Guilioni, Jérémie Lecoeur, Angélique Christophe, Luc Champolivier, and Philippe Debaeke. Sunflo, a model to simulate genotype-specific performance of the sunflower crop in contrasting environments. *Agricultural and forest meteorology*, 151(2):163–178, 2011.

[5] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[6] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

[7] Rémi Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in neural information processing systems*, pages 783–791, 2011.

[8] Matthew Plumlee and Rui Tuo. Building accurate emulators for stochastic simulations via quantile kriging. *Technometrics*, 56(4):466–473, 2014.

[9] Ichiro Takeuchi, Quoc V Le, Timothy D Sears, and Alexander J Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(Jul):1231–1264, 2006.

**Short biography** – Léonard Torossian graduated from Université Pierre et Marie Curie in 2016 with a MSc in Modelling and Optimization. He started is PhD at INRA and IMT in november 2016 co-directed by R. Faivre and V. Picheny from INRA (MIAT) and A. Garivier from IMT. His work is funded by MIAT and the Occitanie region. His PhD subject is about metamodeling and robust optimization of stochastic black box, he aims to built a tool able to take optimal decisions under risk aversion.