Kernel Conditional Quantile Estimation via Reduction Revisited

Novi Quadrianto*, Kristian Kersting†, Mark D. Reid*, Tibério S. Caetano* and Wray L. Buntine*
**SML, NICTA & RSISE, ANU*
Canberra ACT, Australia
Email: {firstname.lastname}@nicta.com.au
†*Fraunhofer IAIS*
Sankt Augustin, Germany
Email: {firstname.lastname}@iais.fraunhofer.de

Abstract—**Quantile regression refers to the process of estimating the quantiles of a conditional distribution and has many important applications within econometrics and data mining, among other domains. In this paper, we show how to estimate these conditional quantile functions within a Bayes risk minimization framework using a Gaussian process prior. The resulting non-parametric probabilistic model is easy to implement and allows non-crossing quantile functions to be enforced. Moreover, it can directly be used in combination with tools and extensions of standard Gaussian Processes such as principled hyperparameter estimation, sparsification, and quantile regression with input-dependent noise rates. No existing approach enjoys all of these desirable properties. Experiments on benchmark datasets show that our method is competitive with state-of-the-art approaches.**

Keywords-**Regression; Quantile Regression; Gaussian Processes;**

I. Introduction

In most regression studies, we are typically interested in inferring a real-valued function whose values correspond to the mean of response variables conditioned on the explanatory variables. The application of this conditional mean regression is ubiquitous. There are, however, many important applications where we are interested in estimating either the median or other quantiles such as estimating the potential amount of money a customer can spend on a product rather than his/her expected spending [1]. This is called quantile regression and was introduced by Koenker and Bassett [2].

The unobservable nature of quantiles means that their prediction is a challenging task. If we had a model $p(y|x)$ of the conditional distribution of response variables y conditioned on the explanatory variables x, however, their prediction would be much simpler as quantile estimation essentially involves slicing this distribution at a certain quantile level. This slicing operation is a convex optimization problem. Although we are *reducing* a hard quantile estimation problem to yet another hard problem, i.e. distribution modeling, the latter is a well-studied subject in machine learning in particular Gaussian processes. At first glance, the usage of Gaussian process distribution modeling for learning problems such as classification or regression might violate Vapnik's paradigm of estimating only the relevant parameters directly [3]. This paradigm is in favor of estimating latent functions while sidestepping distribution modeling. However, there have been several studies that show superiority of Gaussian processes based methods to infer flexible latent functions [4], [5]. Our reduction approach is similar in spirit to the Langford's et al. [6] method in reducing quantile estimation problem to series of classification problems.

Therefore, we propose in this paper to estimate conditional quantile functions within a Gaussian process model [7]. The well-known advantage of using such type of model over non-Bayesian models is that of having an explicit probabilistic formulation. This allows us to have a principled way of performing model selection, as well as a predictive posterior probability distribution over response variables. In terms of quantile estimation, the latter is particularly useful when we have censored or missing response variables [8]. From a practical point of view, our estimator can be easily sparsified, therefore being able to handle large datasets, and can take input dependent noise into account. More importantly, our derived quantile estimator has the desirable property that the estimated conditional functions at different quantiles can never cross or overlap each other. Quantile crossing occurs because each conditional quantile function is independently estimated, and it has traditionally been one of the challenging problems in the field [8], [9]. To our knowledge, our quantile estimator is the first that enjoys both sparsifiable and non-crossing properties while being competitive with state-of-the-art alternatives as we will show in our experiments.

Our contributions:

– A quantile estimator within a Bayes risk minimization framework using a Gaussian process prior.
– A specific example of Bayes quantile estimator which enjoys non-parametric, probabilistic model, principled learning of free parameters, sparse approximation, heteroscedasticity and enforced non-crossing constraint.
– A novel Gaussian processes treatment of input-dependent noise which allows jointly learning the free parameters of the latent and observed processes.
– A theoretical analysis of our proposed estimator in term of regret transform bound.

II. Related Work

Most of existing work focuses on estimating each conditional quantile function separately. A standard technique for conditional quantile estimation is based on a linear model [2]. In this model, the τ-th conditional quantile function of y given x is assumed to be a linear function of the vector of regressors, i.e. $q_\tau(x) := \langle x, \beta(\tau) \rangle$, where $\beta(\tau)$ is a vector of coefficients dependent on τ. Estimation of coefficients is done by minimizing the pinball loss function. It is shown that the minimization can be reformulated as a linear programming problem and can be solved efficiently with interior point techniques [8].

The assumption of a linear relationship between the regressors and the conditional quantile function is quite restrictive. Takeuchi et al. [10] propose a nonparametric approach to quantile regression based on kernel methods. The dual of a regularized version of pinball loss minimization is solved via standard quadratic programming techniques. Several extensions to incorporate commonly desired constraints such as non-crossing constraints and a monotonicity constraint are also discussed. This method provides state-of-the-art performance.

Langford et al. [6] show that the quantile regression problem can be reduced to a series of classification problems such that a small average error rate on the classification problems leads to a provably accurate estimate of the conditional quantile. The estimation of τ-th conditional quantile function is first reduced to a set of importance weighted binary classification problems. This problem is further reduced to ubiquitous unweighted binary classification problem via rejection sampling. This method is computationally efficient thus is able to handle large datasets.

The closest work to Quantile Gaussian processes is the work of Yu and Moyeed [11]. They introduce a Bayesian approach for quantile estimation based on the linear model. An asymmetric Laplace distribution is used as a likelihood function, $p(y|x, \beta(\tau))$, and the prior on the coefficients is chosen to be improper uniform, $(p(\beta(\tau)) \propto 1)$. Although the prior is improper, they proved that the posterior distribution will be proper. A Markov chain Monte Carlo (MCMC) method is used to infer this posterior distribution, $p(\beta(\tau)|x, y)$. Finally, the posterior mean is used for quantile estimation, i.e. $q_\tau(x) := \langle x, \beta_{\text{MAP}}(\tau) \rangle$. We focus on a different notion of prior distribution, usage of any likelihood function, and a different procedure for quantile estimation. Precisely these differences allow us to derive the desirable properties mentioned in the introduction.

III. Quantile Estimation as an Optimization Problem

Given m observed data points $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$, where $y_i \in \mathcal{Y}$ (the set of outputs) and $x_i \in \mathcal{X}$ (the set of regressors or inputs), the goal of quantile regression is to infer a conditional quantile function $q_\tau(x)$ from observed data points.

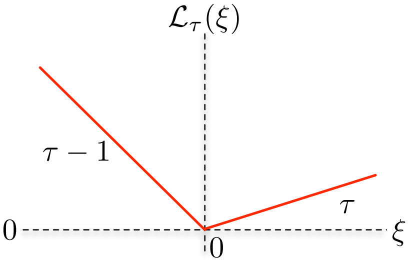


Figure 1. The pinball loss function.

Definition 1 (Conditional Quantile): Let $\tau \in (0, 1)$. The conditional quantile $q_\tau(x)$ for a pair of random variables $(x, y) \in \mathcal{X} \times \mathbb{R}$ is defined as the function $q_\tau : \mathcal{X} \to \mathbb{R}$ for which pointwise $q_\tau(x)$ is the infimum over q for which $\Pr(y \leq q_\tau | x) = \tau$.

The idea behind quantile regression arises from the observation that minimizing the ℓ_1-loss function yields the median. The symmetry of the ℓ_1-loss function implies that minimization of $\sum_{i=1}^m |y_i - q|$ must give an equal number of $y_i - q$ terms lying on either side of zero. Koenker and Bassett [2] generalize this idea to obtain a quantile regression estimator by tilting the loss function. This loss function is given in Figure 1 and is known as a pinball loss function,

$$\mathcal{L}_\tau(\xi) := \begin{cases} \tau\xi, & \text{if } \xi \geq 0 \\ (\tau - 1)\xi, & \text{if } \xi < 0. \end{cases}$$

In this paper, our goal is to estimate the latent quantile function $q_\tau(x)$ in a Bayesian framework with a Gaussian Process prior, which we will develop in the next section.

IV. Bayesian Framework

Assuming we can estimate the conditional distribution $p(y|x)$, the Bayes quantile estimator is found by minimizing expected value of the pinball loss function, i.e.

$$q_\tau^{(\text{opt})} = \underset{q_\tau}{\operatorname{argmin}} \int \mathcal{L}_\tau(y - q_\tau) p(y|x) dy = \underset{q_\tau}{\operatorname{argmin}} \mathcal{R}_\tau(q_\tau) \tag{1}$$

where $\mathcal{R}_\tau(q_\tau) := \mathbf{E}_{p(y|x)}[\mathcal{L}_\tau(y - q_\tau)]$ is the Bayes risk.

Lemma 2: The Bayes risk in Equation (1) is convex in q_τ.

Proof: The Bayes risk is a convex combination of convex loss functions, which must itself be convex. ■

The subsequent sections will deal with modeling the conditional distribution $p(y|x)$. We will describe a model based on Gaussian processes framework. Before introducing the model, let us briefly review Gaussian processes.

Gaussian Process Prior: In the Gaussian process framework, the output y_i at input location x_i is assumed to be a corrupted version of a latent function $q(x_i)$, i.e. $y_i = q(x_i) + \epsilon_i$ where ϵ_i is the noise term. A Gaussian process can be used to define a prior distribution over these latent functions [7], $q \sim \mathcal{GP}(m(x), k(x, x'))$, where $m(x)$ is the mean function (assumed to be zero) and the covariance $k(x, x')$ between functions at input x and x' is defined by Mercer kernel functions [7].

Likelihood for Quantile Regression: In the Bayesian setting, there is a distinction between the likelihood function and the loss function. The likelihood defines the probability of observing the noisy outputs given the latent functions, whereas the loss function measures the regret of making a specific decision. We can in fact define any likelihood function to model the data. For the purpose of this paper, we give a specific example for the likelihood function where we choose to *believe* that the noise term is independent and normally distributed, $\epsilon_i \sim \mathcal{N}(0, \sigma_n^2)$ where σ_n^2 is the noise variance.

Predictive Distribution: Choosing a Gaussian likelihood leads to tractable Bayesian inference, i.e. the standard Gaussian process conditional mean regression [7]. Thus, the predictive distribution over latent functions is given as $q^*|x_*, X, Y \sim \mathcal{N}(\mu_*, \sigma_*^{'2})$ with the moments as follows

$$\begin{aligned} \mu_* &= k^{*T}(\sigma_n^2 I + K)^{-1} Y && (2) \\ \sigma_*^{'2} &= k(x_*, x_*) - k^{*T}(\sigma_n^2 I + K)^{-1} k^*. && (3) \end{aligned}$$

In these equations, we have $K \in \mathbb{R}^{m \times m}$, $K_{ij} = k(x_i, x_j)$ and $k^* \in \mathbb{R}^{m \times 1}$, $k_i^* = k(x^*, x_i)$. Here k denotes covariance function. The predictive distribution over output y_* is also normally distributed, i.e. $y_*|x_*, X, Y \sim \mathcal{N}(\mu_*, \sigma_*^{'2} + \sigma_n^2 := \sigma_*^2)$.

Quantile Estimator: Under the assumption that the true conditional distribution p over y is Gaussian with mean μ and variance σ^2, we can evaluate the risk in (1) for a given quantile estimate q[1]:

$$\mathcal{R}_\tau(q) = (\mu - q)\left[\tau - \Phi_{\mu,\sigma^2}(q)\right] + \sigma\phi_{\mu,\sigma^2}(q) \qquad (4)$$

Proposition 3 (Quantile Estimator): The empirical solution q_τ^* to (1) using the predictive distribution $y_*|x_*, X, Y \sim \mathcal{N}(\mu_*, \sigma_*^2)$ is given by the zero of the following function: $f_\tau(q) = \Phi_{\mu_*,\sigma_*^2}(q) - \tau$.

Proof: Since the objective function is convex, the (global) minimizer of $\mathcal{R}_\tau(.)$ with $p(y_*|x_*, X, Y) = \mathcal{N}(\mu_*, \sigma_*^2)$ is given by $\partial_{q_\tau}\{\int \mathcal{L}_\tau(y_* - q_\tau)p(y_*|x_*, X, Y)dy_*\} = 0$. ■

Thus, the τ-th quantile estimate is given by

$$q_\tau^* = \mu_* + \sigma_*\Phi^{-1}(\tau). \qquad (5)$$

Remark: In literatures, (5) is known as a *location-scale* model. Several methods have been proposed to estimate both location and scale functions simultaneously (c.f. [8]). This is a special case of our framework with a specific choice of likelihood function.

Our estimator carries several advantages. The first is that our estimator inherently enforces a *non-crossing constraint*. Estimation of several conditional quantile functions can cause two or more estimated functions to cross or overlap. This is due to each conditional quantile function being independently estimated. This phenomenon should not happen as the true quantile functions are defined to be non-crossing.

Corollary 4 (Non-Crossing Estimator): For $p(y_*|x_*, X, Y)$ is independent of τ, the Bayes quantile estimator is a monotone increasing function of τ.

Proof: Provided $p(y_*|x_*, X, Y)$ is independent of τ and has finite density this is immediate from the fact that the inverse CDF is monotonically increasing. ■

There have been several approaches addressing the non-crossing constraint. *He* [9] transformed the non-crossing constraint into a positivity constraint, however, this might not be desirable from the non-parametric point of view. *Takeuchi et al.* [10] imposed the non-crossing constraint as linear constraints, however, this means that every adjacent pair of conditional quantile functions should be computed when multiple quantiles are needed. Recently, *Shim et al.* [12] used a location–scale model and estimated both location and scale functions simultaneously via SVM. It is shown that the proposed method works slightly better than the method of [10] but offers conceptual simplicity since it estimates the location and scale functions simultaneously.

Secondly, by approximating the predictive distribution over quantile functions with conditional mean Gaussian process regression, we have a large pool of *sparse approximation* methods at our disposal. Several approximative models, such as subset of regressors, subset of datapoints, projected process, and Bayesian committee machine have been proposed for Gaussian process regression in order to deal with the high time and storage requirements for large training datasets. Many of the approximations use a subset $\mathcal{I}$, $|\mathcal{I}| = n$, of datapoints (the support set) from the full training set $\mathcal{D}$, $|\mathcal{D}| = m$, see e.g. [7] for more details. Finally, we can elegantly deal with input-dependent noise as shown below.

V. Heteroscedastic Quantile Estimation

In many real-world problems, the local noise rates are important features of data distributions and hence of conditional quantiles that have to be modeled accurately. Our Bayesian approach allows a simple but elegant solution to handle this locally varying noise: use heteroscedastic Gaussian processes instead of standard ones.

In contrast to the standard Gaussian process approaches discussed so far, we now do not assume a constant noise level $n(x)$ at location x but place a prior over it. More precisely, an independent Gaussian process is used to model the logarithms of the noise levels, denoted as $z(x) = \log(n(x))$. This noise process is governed by a different

[1] $\phi_{\mu,\sigma^2}(x)$ denotes the density at x of the Gaussian random variable with mean μ and variance σ^2 and $\Phi_{\mu,\sigma^2}(z) = \int_{-\infty}^{z} \phi_{\mu,\sigma^2}(x)dx$ denotes the CDF. $\Phi(z) := \Phi_{0,1}(z)$ is the standard Gaussian CDF.

covariance function k_z, parameterized by θ_z. The locations $\bar{X} = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_l\}$ for the "training" noise levels $\bar{Z} = \{\bar{z}_1, \bar{z}_2, \dots, \bar{z}_l\}$ can be chosen differently from the ones used for the noise-free process.

Since the noise rates z_i at the original locations $x_1, x_2, \dots, x_m$ are now independent latent variables in the combined regression model, the predictive distribution for y_* changes to $p(y_*|x_*, \mathcal{D}, \theta) = \iint p\left(y_*|x_*, z_*, \mathcal{D}, Z, \theta\right) \cdot p\left(z_*, Z|x_*, X, \bar{X}, \bar{Z}, \theta_z\right) dz_* dZ$, where Z denotes the predicted (logarithmized) noise levels at the original locations X. Given (z_*, Z), the prediction $p\left(y_*|x_*, z_*, \mathcal{D}, Z\right)$ is Gaussian with mean and variance as defined by (2) and (3) replacing the constant noise level $\sigma_n^2 \times I$ with the diagonal matrix $\operatorname{diag}(\exp(Z))$.

The problematic term is $p(z_*, Z|x_*, X, \bar{X}, \bar{Z})$. It makes the integral difficult to handle analytically. Instead, we seek for the solution using the most probable noise estimates, i.e., $p(y_*|x_*, \mathcal{D}, \theta) \approx p\left(y_*|x_*, z_*, X, Y, Z\right)$ where (z_*, Z) are the mean predictions of the latent noise process. To jointly estimate θ, θ_z, and $\bar{Z}$ from data, we seek an MAP solution that maximizes $\log p(Z|y, X, \bar{X}, \theta) = \log p(y|X, Z, \theta) + \log p(Z|X, \bar{X}, \theta) + \text{const.}$, where (overloading notation) θ now also includes θ_z and $\bar{Z}$. One may now find the gradient of this objective function with respect to the hyperparameters θ and employ it within a gradient-based optimization to find the corresponding solution.

VI. Regret Transform Bound

Our approach to solving the quantile estimation problem can be thought of as a reduction. The problem we really wish to solve is quantile estimation (Problem A) but the problem we actually solve is a Gaussian Process regression (Problem B) and then we use this to get a quantile estimation. A natural question in this sort of reductions is: given a measure of how well we solve Problem B, what can we say about how well will we solve Problem A?

Questions like this are typically answered in terms of *regrets*. The regret of a τ-th quantile point estimate q under a true point distribution $p(y|x)$ is $\Delta\mathcal{R}_\tau(q) := \mathcal{R}_\tau(q) - \mathcal{R}_\tau(q^{opt})$, where q^{opt} is the best τ-th quantile estimate in (1) under p.

Theorem 5: Suppose $p_* = \mathcal{N}(\mu_*, \sigma_*^2)$ is a predictive distribution at the point x_* and the true point distribution is $p = \mathcal{N}(\mu, \sigma^2)$. Then, if $KL(p_*||p) \leq \epsilon$, the regret of the corresponding τ-th quantile estimator q_τ^* satisfies

$$\Delta\mathcal{R}_\tau(q) \leq \sqrt{2\epsilon}(\tau\sigma + 1)(|\Phi^{-1}(\tau)| + 1). \qquad (6)$$

This bound depends not only on how well the true normal distribution can be estimated (the $\sqrt{\epsilon}$ term) but also on the quantile being estimated (τ) and the variance of the true distribution (σ). These dependencies are quite natural. If the true distribution is spread out a small error in estimating it can lead to large errors in its quantiles. Also, since there is little mass in the highest and lowest quantiles, an error in estimating the true distribution can potentially make large changes to a quantile location.

The proof of this theorem relies on the Lipschitz continuity of several functions related to Gaussian densities which we establish in the following lemma.

Lemma 6: Upper bounds on the Lipschitz constants for the functions $q \mapsto \phi_{\mu,\sigma^2}(q)$ and $q \mapsto (\mu - q)\Phi_{\mu,\sigma^2}(q)$ are σ^{-2} and σ^{-1}, respectively.

Proof: The first and second derivatives of $\phi_{\mu,\sigma^2}(q)$ are $\phi'_{\mu,\sigma^2}(q) = \frac{\mu-q}{\sigma^2}\phi_{\mu,\sigma^2}(q)$ and $\phi''_{\mu,\sigma^2}(q) = \frac{(\mu-q)^2-\sigma^2}{\sigma^4}\phi_{\mu,\sigma^2}(q)$. Thus, the maximal/minimal values for $\phi'_{\mu,\sigma^2}(q)$ occur when $\phi''_{\mu,\sigma^2}(q) = 0$. That is, when $q = \mu \pm \sigma$. Thus, for all $q \in \mathbb{R}$, we have $|\phi'_{\mu,\sigma^2}(q)| \leq |\phi'_{\mu,\sigma^2}(\mu \pm \sigma)| = \frac{\sigma}{\sigma^2}\frac{1}{\sqrt{2\pi\sigma^2}} = \frac{1}{\sigma^2\sqrt{2\pi}} < \sigma^{-2}$.

Similarly, the first and second derivatives of $(\mu - q)\Phi_{\mu,\sigma^2}$ are $\frac{(\mu-q)^2-\sigma^2}{\sigma^2}\phi_{\mu,\sigma^2}(q)$ and $\frac{(\mu-q)^3-3\sigma^2(\mu-q)}{\sigma^4}\phi_{\mu,\sigma^2}(q)$. Thus, its first derivative is maximal/minimal at either $q = \mu$ or $\mu - q = \pm\sqrt{3}\sigma$. Substituting these solutions back into the first derivative gives $\left|\frac{d}{dq}(\mu - q)\Phi_{\mu,\sigma^2}(q)\right| \leq \frac{1}{\sqrt{2\pi\sigma^2}}\max\left\{1, 2e^{-3/2}\right\} < \sigma^{-1}$ and proves the lemma. ■

Proof: [Theorem 5] The regret under the assumption that p is the true point distribution is given by (4) for q and q^{opt} as $\Delta\mathcal{R}_\tau(q) = \tau(q^{opt} - q) + \sigma[\phi_{\mu,\sigma^2}(q) - \phi_{\mu,\sigma^2}(q^{opt})] + (\mu - q^{opt})\Phi_{\mu,\sigma^2}(q^{opt}) - (\mu - q)\Phi_{\mu,\sigma^2}(q)$. Letting $\Gamma_{\mu,\sigma^2}(q) := (\mu - q)\Phi_{\mu^2_\sigma}(q)$ and by the Lipschitz conditions of Lemma 6 we can write, $|\Delta\mathcal{R}_\tau(q)| \leq \tau|q^{opt} - q| + \sigma|\phi_{\mu,\sigma^2}(q) - \phi_{\mu,\sigma^2}(q^{opt})| + |\Gamma_{\mu,\sigma^2}(q^{opt}) - \Gamma_{\mu,\sigma^2}(q)| = (\tau + \sigma^{-1})|q^{opt} - q|$. We now note that, by equation (5) that $|q^{opt} - q| \leq |\mu - \mu_*| + |\Phi^{-1}(\tau)||\sigma - \sigma_*|$, where (μ, σ) and (μ_*, σ_*) are the moments for the true and predictive distribution, respectively. Thus, it is sufficient to bound the difference in means and variances between these distributions.

We now make use of our assumption that $KL(p_*||p) < \epsilon$. The KL-divergence for two Gaussians is the well-known expression, $KL(p_*||p) = \ln\left(\frac{\sigma}{\sigma_*}\right) + \frac{\sigma_*^2}{2\sigma^2} + \frac{(\mu-\mu_*)^2}{2\sigma^2} - \frac{1}{2}$. Since σ_* and μ_* can be chosen independently, we note that the upper bound implies both $|\mu - \mu_*| < \sigma\sqrt{2\epsilon}$ and $\ln(\sigma/\sigma_*) + \frac{1}{2}(\sigma_*^2/2\sigma^2 - 1) < \epsilon$. The well-known bound $\ln(x) \leq x - 1$ for all $x > 0$ can be rearranged to give $\ln(x) \geq 1 - \frac{1}{x}$ and so $\epsilon > \ln(\frac{\sigma}{\sigma_*}) + \frac{1}{2}(\frac{\sigma_*^2}{\sigma^2} - 1) \geq 1 - \frac{\sigma_*}{\sigma} + \frac{1}{2}(\frac{\sigma_*^2}{\sigma^2} - 1) = \frac{(\sigma-\sigma_*)^2}{2\sigma^2}$. Thus, it is also the case that $|\sigma - \sigma_*| < \sigma\sqrt{2\epsilon}$. Combining all these bounds proves the result. ■

VII. Experimental Evaluation

Our intention here is to investigate to which extent the performance of the Quantile GP is comparable to state-of-the-art quantile estimation approaches.

A. Synthetic Data

In this experiment, we are interested to analyze the quality of our estimator under the condition of *known* noise

	τ				
	0.1	0.25	0.5	0.75	0.9
Example 1					
QSVM	0.0822	0.0641	0.0274	0.0238	0.0937
HQGP	0.0621	0.0410	0.0306	0.0379	0.0563
Example 2					
QSVM	0.0987	0.2465	0.5090	0.8044	0.9393
HQGP	0.5286	0.2938	0.2398	0.4793	0.9927

Table I

Absolute loss comparison on Example 1 and Example 2. QSVM: Quantile SVM, [10]; HQGP: Heteroscedastic Quantile GP, i.e. our approach.

distribution. We focus on two cases, namely the Gaussian noise case and the Chi-squared noise case:

Example 1 (Heteroscedastic Gaussian Noise) We generate 100 samples from the following stochastic process: $x \sim U(-1, 1)$ and $y = \mu(x) + \sigma(x)\xi$ with $\mu(x) = \mathrm{sinc}(x), \sigma(x) = 0.1 \exp(1 - x)$, and noise, $\xi \sim \mathcal{N}(0, 1)$.
Example 2 (Heteroscedastic Chi-squared Noise) We generate 200 samples from the following stochastic process: $x \sim U(0, 2)$ and $y = \mu(x) + \sigma(x)\xi$ with $\mu(x) = \sin(2\pi x), \sigma(x) = \sqrt{\frac{2.1-x}{4}}$, and noise, $\xi \sim \chi^2_{(1)} - 2$.

For the case of known noise distribution, the true quantile values can be computed simply via inverse cumulative distribution function of noise density, i.e. $q^{\mathrm{true}}_\tau = \mu(x) + \sigma(x)\Phi^{-1}_\xi(\tau)$ with $\Phi_\xi(.)$ is given as $\Phi_{\mathcal{N}(0,1)}(.)$ or $\Phi_{\chi^2_{(1)}}(.) - 2$ for Example 1 or 2, respectively. The absolute errors for each estimated quantile regression functions are given in Table I. For comparison, we contrast the performance of our method with SVM based quantile estimator [10]. It is of no surprise that our estimator shows superior performance in Gaussian noise corrupted data while falls short in Chi-squared noise model case. In the latter case, our estimator tries to approximate a single right-tail Chi-squared density with a double tail Gaussian density and thus the produced estimates suffer badly in the lower quantile regime. However, in the real data, the noise model miss-specification is less apparent and as it is shown in the next section our quantile estimator delivers competitive performance with state-of-the-art approaches. This is partly due to the competitive advantage of Gaussian processes based estimator to be a superior mean predictor.

B. Real Data

We are interested to assess the effectiveness of our quantile estimator in comparison to the linear model by [2], the SVM based approach by [10], and the learning reduction based approach by [6] for several real datasets. We implemented our approach in `Matlab` using [7] GPML Toolbox.

		τ		
Dataset		0.1	0.5	0.9
Antigen	A	0.293±0.105	0.264±0.050	0.292±0.087
	B	0.123±0.033	**0.249±0.033**	0.128±0.018
	C	0.122±0.031	0.266±0.021	0.131±0.015
	D	**0.116±0.021**	0.255±0.028	**0.126±0.015**
Weather	A	0.291±0.034	0.293±0.024	0.301±0.045
	B	0.067±0.015	0.218±0.034	0.118±0.013
	C	0.075±0.011	0.176±0.028	0.123±0.011
	D	**0.057±0.010**	**0.097±0.015**	**0.068±0.017**
Mcycle	A	0.396±0.080	0.389±0.019	0.387±0.056
	B	0.090±0.012	0.202±0.019	0.085±0.008
	C	0.094±0.011	0.190±0.015	0.083±0.010
	D	0.092±0.025	**0.186±0.018**	0.089±0.010
	E	**0.079±0.019**	0.187±0.021	**0.070±0.016**
BMD	A	0.328±0.028	0.325±0.0340	0.324±0.073
	B	0.122±0.017	**0.306±0.039**	**0.152±0.025**
	C	**0.121±0.020**	0.311±0.041	0.154±0.027
	D	0.135±0.014	0.310±0.045	0.168±0.030
	E	0.123±0.017	0.309±0.045	0.153±0.027
Calif. Housing	A	0.283±0.038	**0.225±0.009**	0.254±0.068
	B	†	†	†
	C	0.108±0.014	0.263±0.023	**0.167±0.006**
	F	**0.104±0.016**	0.272±0.018	0.175±0.024

Table II

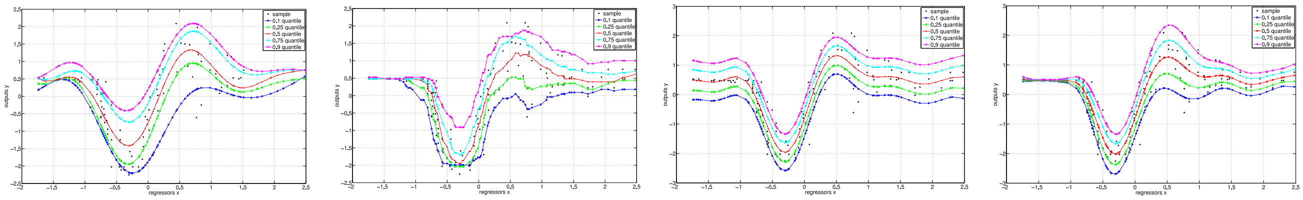
Pinball loss comparison: 5-fold cross validation errors ± STD. The best result is in boldface. A: Linear, [2]; B: Quantile SVM, [10]; C: Reduction to classifications, [6]; D: Quantile GP; E: Heteroscedastic Quantile GP; F: Sparse Quantile GP. †: Program fails (large dataset).

Datasets: We used three regression datasets from the UCI repository (Antigen-97[2], Weather-238 and Motorcycle-133); one dataset from the Elements of Statistical Learning Book (BMD-485); and one dataset from StatLib repository (California Housing-20640). We normalized all datasets to have zero mean and unit standard deviation for each coordinate [10].

Model Selection: In our approach, we use squared exponential covariance function. Gaussian RBF Kernel is used for Quantile SVM with the kernel width and regularization parameters fitted with the trick described in [10]. For learning the reduction method, Expectation Propagation (EP) approximation of Gaussian process classification [7] with squared exponential covariance function is used as base classifier learners. We fix the number of classifiers at 100 for all datasets [6]. For the linear model, there is no parameter to be tuned.

Sparse Approximation: As stated in Section IV, our estimator can be easily sparsified. This relies on advances of sparse approximation methods for conditional mean Gaussian process regression. In our experiments, we use the *projected process* (PP) [7]. We assess the performance of

[2] Number of observations.



(a) Takeuchi *et al.* (b) Langford *et al.* (c) Quantile GP (d) Heteroscedastic QGP

Figure 2. Illustration of conditional quantile analysis for Silverman's motorcycle dataset via Quantile SVM (Takeuchi *et al.*), Reduction (Langford *et al.*), Quantile GP, and Heteroscedastic QGP. The dataset exhibits heteroscedasticity. Non-crossing constraint is not enforced in Reduction and enforceable in Quantile SVM via additional linear constraints and is an inherent property of Quantile GP.

this sparse approximation in the California Housing dataset.

Input-dependent noise: We optimized the hyperparameters of both the noise-free and the noise process jointly using a scaled conjugate gradient approach. As locations $\bar{X}$ of the latent noise process, we selected 10 points linearly spaced in the bounding interval of the given input location.

Results: We run the linear model, SVM, learning reduction, and Gaussian process methods for 3 different quantile values for each dataset, i.e. at 0.1, 0.5 and 0.9. For Quantile SVM, we perform *nested* 5-fold cross-validation. There is no need to perform nested cross-validation for our approach, learning reduction approach, and the linear model as the free parameters are selected via log evidence, EP approximation of log evidence or there is no free parameter, respectively. The 5-fold cross validation results are summarized in Table II. Arguably, our estimator performs on par with (if not exceeding) state-of-the-art SVM and learning reduction based method. Noticeably, our approach has a competitive advantage over Quantile SVM for large datasets where the later approach might fail due to high memory and computational time requirements.

An illustration of the estimated quantile regression functions via SVM, learning reduction, and Gaussian process methods on the Motorcycle dataset is given in Figure 2.

VIII. Conclusions and Future Research

We tackled the quantile estimation problem by modeling the conditional distribution and subsequently slicing the distribution at the respective quantile level to get the estimate of the latent quantile function. This approach is preferable when multiple quantile regression functions are needed and captures rather well the characteristics of the datasets.

In this paper, we have focussed on the specific example of Bayes quantile estimator, i.e. with Gaussian likelihood function. While this offers several appealing properties, the framework is by no means restricted to this. Our proposed Bayes quantile estimator offers two design parameters: likelihood function (for robustness, a heavier tail distribution function might be more preferable) and non-parametric CDF (estimation of CDF *directly* from the data via residuals / errors).

Acknowledgment

The authors would like to thank Alex Smola, Marcus Hutter, and Marconi Barbosa for useful comments and suggestions. NICTA is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the ARC. The research was supported by the Pascal Network and the Fraunhofer ATTRACT fellowship STREAM.

References

[1] C. Perlich, S. Rosset, R. D. Lawrence, and B. Zadrozny, "High-quantile modeling for customer wallet estimation and other applications," in *KDD '07*. ACM, 2007, pp. 977–985.

[2] R. Koenker and G. Bassett, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[3] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. Springer, 1982.

[4] C. Rasmussen, "Evaluation of Gaussian processes and other methods for non-linear regression," Ph.D. dissertation, Department of Computer Science, University of Toronto, 1996, http://www.kyb.mpg.de/publications/pss/ps2304.ps.

[5] H. Nickisch and C. E. Rasmussen, "Approximations for binary gaussian process clasification," *JMLR*, vol. 9, pp. 2035–2078, 2008.

[6] J. Langford, R. Oliveira, and B. Zadrozny, "Predicting conditional quantiles via reduction to classification," in *UAI*, 2006.

[7] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[8] R. Koenker, *Quantile Regression*. Cambridge University Press, 2005.

[9] X. He, "Quantile curves without crossing," *The American Statistician*, vol. 51, no. 2, pp. 186–192, may 1997.

[10] I. Takeuchi, Q. V. Le, T. Sears, and A. J. Smola, "Nonparametric quantile estimation," *J. Mach. Learn. Res.*, vol. 7, 2006.

[11] K. Yu and R. A. Moyeed, "Bayesian quantile regression," *Statistics & Probability Letters*, vol. 54, pp. 437–447, 2001.

[12] J. Shim, C. Hwang, and K. H. Seok, "Non-crossing quantile regression via doubly penalized kernel machine," in *Computational Statistics*, vol. 24, 2009, pp. 83–94.