Variational Inference for Nonparametric Bayesian Quantile Regression

Sachinthaka Abeywardana
School of Information Technologies
University of Sydney
NSW 2006, Australia
sachinra@it.usyd.edu.au

Fabio Ramos
School of Information Technologies
University of Sydney
NSW 2006, Australia
fabio.ramos@sydney.edu.au

Abstract

Quantile regression deals with the problem of computing robust estimators when the conditional mean and standard deviation of the predicted function are inadequate to capture its variability. The technique has an extensive list of applications, including health sciences, ecology and finance. In this work we present a nonparametric method of inferring quantiles and derive a novel Variational Bayesian (VB) approximation to the marginal likelihood, leading to an elegant Expectation Maximisation algorithm for learning the model. Our method is nonparametric, has strong convergence guarantees, and can deal with nonsymmetric quantiles seamlessly. We compare the method to other parametric and non-parametric Bayesian techniques, and alternative approximations based on expectation propagation demonstrating the benefits of our framework in toy problems and real datasets.

1 Introduction

Most regression techniques revolve around predicting an average value for a query point given a training set and, in certain cases, the predicted variance around this mean. Quantile regression was introduced as a method of modelling the variation in functions, where the mean along with standard deviation are not adequate. In this sense quantile regression provides a better statistical view of the predicted function. Quantiles are important tools in medical data, for instance in measuring a normal weight range for a particular age group or, in modelling train arrival times where (for arguments sake) 90% of trains would arrive before the allocated time and 10% late. Other areas of application are in financial data where it is important to measure what the daily worst case scenarios would be so that analysts could hedge their risks.

There are two main approaches used in inferring quantiles. The first is building a Cumulative Distribution Function (CDF) over the set of observations. Taddy and Kottas; Chen and Müller employ this approach to model the quantiles. However, the drawback of this approach is that it requires MCMC methods for inference which can be computationally intensive and prohibitive for large datasets.

The second approach uses a loss function that penalises predictive quantiles at wrong locations. Koenker and Bassett Jr introduced the tilt (pinball) loss function over the errors ξ_i for a specified quantile $\alpha \in (0, 1)$ (equation 1). The errors mentioned in this context are the errors between the observation $\mathbf{y}_i$ and the inferred quantile $\mathbf{f}_i$;

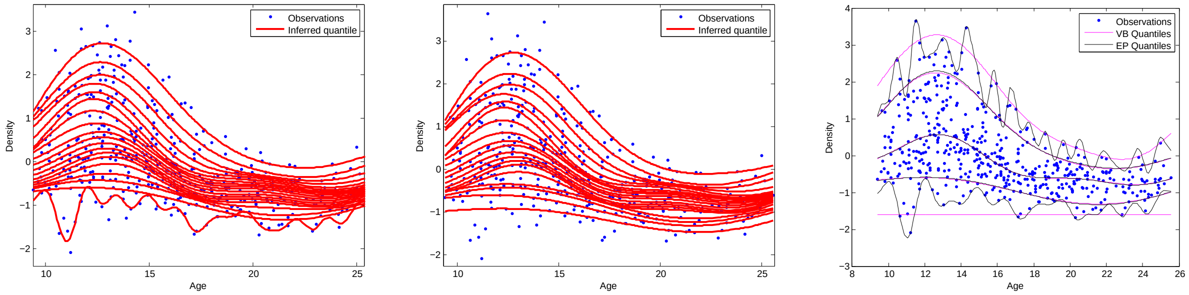
$$\mathcal{L}(\xi_i, \alpha) = \begin{cases} \alpha\xi_i & \text{if } \xi_i \geq 0, \\ (\alpha - 1)\xi_i & \text{if } \xi_i < 0. \end{cases} \tag{1}$$

However, as with many other regression techniques, regularisation is necessary to prevent overfitting. Thus, the problem can be transformed to minimising over $\mathbf{f}$ (the quantile function) for $\mathcal{L}(\alpha, \mathbf{y}, \mathbf{f}) + \lambda||\mathbf{f}||$ for some specified norm $||\cdot||$ where, $\mathcal{L}(\alpha, \mathbf{y}, \mathbf{f}) = \sum_{i=1}^{N} \mathcal{L}(\mathbf{y}_i - \mathbf{f}_i, \alpha)$. This could be solved as an optimisation problem using quadratic programming as shown in (Takeuchi et al. 2006). However, it requires finding an appropriate regularisation term λ.

In this work, we adopt the second approach where a loss is minimised but within a Bayesian framework. In addition to naturally encoding the Occam's razor principle (simpler models are preferable) therefore avoiding the manual specification of the regularisation term, the Bayesian formulation also provides posterior estimates for the predictions and the associated uncertainty.

Inspired by the ability of the l_1 norm to consistently enforce sparsity, Koenker and Bassett Jr modified this loss function to create the pinball loss function (equation 1) where, $\xi_i = \mathbf{y}_i - \mathbf{f}_i$. The l_1 norm can be thought of as a proxy to cardinality, which is exploited in Lasso regression, (Tibshirani 1996). As stated in (Takeuchi et al. 2006) the minimiser $\mathbf{f}$ of this loss has the property of having at most αN and $(1-\alpha)N$ observations for $\xi < 0$ and $\xi > 0$ respectively. Finally, for large number of observations, the proportion $|\xi < 0|/|\xi > 0|$ converges to α. In a probabilistic setting, instead of minimising this loss the goal is to maximise the exponential of the negative loss.

In this work we derive a nonparametric approach to modelling the quantile function. Similarly, (Quadrianto et al. 2009), (Takeuchi et al. 2006) and (Boukouvalas, Barillec, and Cornford 2012) use kernels as a nonparametric method of inferring quantile functions. (Quadrianto et al. 2009) minimises the expected loss function under a Gaussian Process (GP) (Rasmussen 2006) prior which is placed over the data. (Boukouvalas, Barillec, and Cornford 2012) takes a more



(a) Expectation propagation

(b) Variational Bayesian

(c) Both methods superimposed

Figure 1: Comparison of bone density quantiles as a function of age. The first two images show the quantiles 0.05 to 0.95 with increments of 0.05 for EP and VB methods. The last image shows quantiles 0.01, 0.1, 0.5, 0.9 and 0.99 with both EP and VB inferences superimposed.

direct Bayesian approach by having an Asymmetric Laplace likelihood over the data and a Gaussian Process prior over the space of quantile functions. The same approach is taken in this work however we derive a Variational Bayesian (VB) inference method which possesses theoretical advantages over the Expectations Propagation (EP) approximation.

The above mentioned methods have a series of weaknesses which we overcome with the VB formulation. Firstly, the quantiles inferred in (Takeuchi et al. 2006) are point estimates and do not have uncertainty estimates associated with it. Conversely, if the data is modelled as a GP (or its heteroskedastic extensions), it is possible to infer quantiles using the inverse Cumulative Distribution Function (CDF) of a Gaussian. The method of construction of quantiles taken by (Quadrianto et al. 2009) which strongly resembles a heteroskedastic GP, implies that the median is the mean and the quantiles are symmetric about the median (mean). The symmetric assumption of quantiles is a weakness when inspecting datasets as those in figure 1. In fact, the authors report that this heteroskedastic GP framework performs poorly in conditions of non-Gaussian errors. (Boukouvalas, Barillec, and Cornford 2012) use Expectation Propagation (EP) as a tool to approximate Bayesian inference, overcoming some of these limitations. Our VB formulation has the same properties but with the following additional advantages over EP: 1. A guaranteed lower bound on the marginal log likelihood is provided. 2. An explicit formulation of the family of functions used in the approximation do not need to be specified. 3. It is guaranteed to converge (Bishop and others 2006, p. 510).

In other works, Yu and Moyeed; Kozumi and Kobayashi use Bayesian formulations for quantile regression but, in a parametric setting. Both settings use asymmetric likelihoods of which the log likelihood is the pinball loss function. (Yu and Moyeed 2001) uses a uniform prior over the parameters whereas (Kozumi and Kobayashi 2011) uses a Gaussian prior with MCMC inference to learn the model. Also, the asymmetric Laplacian distribution can be shown to be a scalar mixture of Gaussians as pointed out in (Kotz, Kozubowski, and Podgorski 2001) and (Kozumi and Kobayashi 2011) with interesting properties for quantile regression.

One of the defining features of our framework is that there are no assumptions on the type of the distribution used for the generative function. Instead, the prior lies over the quantile in question. The advantage of this is that the required quantile can be inferred over non-symmetric and even multimodal functions. The advantages of this are summarised in table 1.

	VB	EP	MCMC	GP
Nonparametric	✓	✓		✓
Fast inference	✓	✓		✓
Convergence guarantees	✓		✓	✓
Non-symmetric quantiles	✓	✓	✓	

Table 1: Main properties of different approaches for quantile regression.

The remainder of the paper is structured as follows. We define the hierarchical Bayesian model in section 2 and show how to find the posterior using approximate Bayesian inference in section 3. In order to learn the model over kernel hyper-parameters, we present and analyse the data likelihood term in section 4. We devise the inference equations in section 5 and present experiments and comparisons in section 6.

2 Bayesian Quantile Regression

In a Bayesian setting the aim is to derive the posterior $p(\mathbf{f}_\star|\mathbf{y},\mathbf{x}_\star,\mathbf{x})$ where $\mathbf{f}_\star$ is a prediction for some input $\mathbf{x}_\star$ and $\mathbf{y},\mathbf{x}$ is the set of observations. This is done by marginalising out all latent variables. We assume that the function is locally smooth which leads to Gaussian Process prior (which employs a stationary kernel) on the space of functions, and use an Inverse Gamma prior ($\text{IG}(10^{-6},10^{-6})$) for the uncertainty estimate σ (equation 4). Finally, the data likelihood is an exponentiation of the Pinball loss (equation 1) function.

$$p(\mathbf{y}_i|\mathbf{f}_i,\alpha,\sigma,\mathbf{x}_i) = \frac{\alpha(1-\alpha)}{\sigma}\exp\left(-\frac{\xi_i(\alpha - I(\xi_i<0))}{\sigma}\right) \tag{2}$$

$$p(\mathbf{f}|\mathbf{x}) = \mathcal{N}(\mathbf{m}(\mathbf{x}),\mathbf{K}(\mathbf{x})) \tag{3}$$

$$p(\sigma) = \text{IG}(10^{-6},10^{-6}) \tag{4}$$

where, $\xi_i = \mathbf{y}_i - \mathbf{f}_i$[1], I is the indicator function and $\mathbf{K}$ is the covariance matrix whose elements are $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ for some kernel function $k(\cdot,\cdot)$ and mean function $\mathbf{m}(\cdot)$ which is assumed to be zero without loss of generality. This likelihood function is an Asymmetric Laplace distribution (Kotz, Kozubowski, and Podgorski 2001). The σ parameter is a dispersion measurement of the observations about the latent quantile function $\mathbf{f}$. An important property of the likelihood function is that $p(\mathbf{y}_i < \mathbf{f}_i) = \alpha$. Specifically, $100\alpha\%$ of the observations are below the quantile function.

Alternatively, the likelihood $p(\mathbf{y}_i|\mathbf{f}_i, \alpha)$ can be written as a scalar mixture of Gaussians (Kotz, Kozubowski, and Podgorski 2001; Kozumi and Kobayashi 2011) such that,

$$p(\mathbf{y}_i|\mathbf{f}_i, \mathbf{x}_i, \sigma, \alpha) = \int \mathcal{N}(\mathbf{y}_i|\mu_{\mathbf{y}_i}, \sigma_{\mathbf{y}_i}) \exp(-\mathbf{w}_i)\, d\mathbf{w} \quad (5)$$

where, $\mu_{\mathbf{y}_i} = \mathbf{f}_i(\mathbf{x}_i) + \frac{1-2\alpha}{\alpha(1-\alpha)}\sigma\mathbf{w}_i$ and $\sigma_{\mathbf{y}_i} = \frac{2}{\alpha(1-\alpha)}\sigma^2\mathbf{w}_i$. Thus the likelihood can be represented as a joint distribution with $\mathbf{w}$ (which will be marginalised out) where, the prior on $\mathbf{w}$ is $\prod_{i=1}^{N}\exp(-\mathbf{w}_i)$. This extra latent variable $\mathbf{w}$ will be useful in a Variational Bayesian setting which is shown in section 3.

3 Variational Bayesian Inference

The marginal likelihood $p(\mathbf{y}|\mathbf{x}, \theta, \alpha)$ as well as the posterior on the latent variables $p(\mathbf{f}, \mathbf{w}, \sigma|\mathbf{y}, \theta, \alpha)$ are not analytically tractable (where, θ are the hyper-parameters and are discussed in section 4). VB aims to approximate this intractable posterior distribution with an approximate posterior $q(\mathbf{f}, \mathbf{w}, \sigma)$.

The data likelihood, $\log p(\mathbf{y}|\mathbf{x}, \alpha, \theta)$ can alternatively be expressed as: $\mathcal{L}(q(\mathbf{f}, \mathbf{w}, \sigma), \theta|\alpha) + KL(q(\mathbf{f}, \mathbf{w}, \sigma)||p(\mathbf{f}, \mathbf{w}, \sigma|\mathbf{y}, \theta, \alpha))$ where, $\mathcal{L} = \int\int q(\mathbf{f}, \mathbf{w}, \sigma) \log \frac{p(\mathbf{f}, \mathbf{w}, \sigma, \mathbf{y}|\theta, \alpha)}{q(\mathbf{f}, \mathbf{w}, \sigma)} d\mathbf{f} d\mathbf{w} d\sigma$ and, KL is the Kullback-Leibler divergence between the proposal distribution on the latent variables and the posterior distribution of the latent variables. The Expectation Maximisation (EM) algorithm maximises the likelihood by initially minimizing the KL divergence for a given set of hyper parameters (i.e. finding an appropriate $q(\cdot)$). Ideally, this is usually done by setting $p(\mathbf{f}, \mathbf{w}, \sigma|\mathbf{y}) = q(\mathbf{f}, \mathbf{w}, \sigma)$ in which case $\log p(\mathbf{y}|\theta) = \mathcal{L}(q(\mathbf{f}, \mathbf{w}, \sigma), \theta)$. However, in this case an analytic distribution for $p(\mathbf{f}, \mathbf{w}, \sigma|\mathbf{y})$ cannot be found. Instead, the approximation, $q(\mathbf{f}, \mathbf{w}, \sigma) = q(\mathbf{f})q(\mathbf{w})q(\sigma) \approx p(\mathbf{f}, \mathbf{w}, \sigma|\mathbf{y})$ is used (Tzikas, Likas, and Galatsanos 2008). Under this assumption the closed form solution for the approximate distribution $q(z_i) = \exp(E(\log p(\mathbf{z}, \mathbf{y}))/Z$ where, $\{\mathbf{z}_i\}$ is the set of latent variables, Z is the normalising constant and the expectation, E is taken w.r.t. to approximate distributions $q(\mathbf{z})$ with the exception of $\mathbf{z}_i$ itself. In the approximate distributions that follow, $\langle\cdot\rangle$ indicates the expectation with respect to all the latent variables except, the variable being investigated.

The approximate posterior on the function space is $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ [2] where,

$$\boldsymbol{\Sigma} = \left(\left\langle \mathbf{D}^{-1} \right\rangle + \mathbf{K}^{-1}\right)^{-1} \quad (6)$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\left\langle \mathbf{D}^{-1} \right\rangle \mathbf{y} - \frac{1-2\alpha}{2}\left\langle \frac{1}{\sigma} \right\rangle \mathbf{1}\right) \quad (7)$$

where, $\mathbf{D} = \frac{2}{\alpha(1-\alpha)}\sigma^2 \text{diag}(\mathbf{w})$. The expectations, $\langle \mathbf{f} \rangle = \boldsymbol{\mu}$ and $\left\langle \mathbf{f}\mathbf{f}^T \right\rangle = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T$ will be required for the computation of subsequent approximate distributions.

The approximate posterior on $\mathbf{w}_i$ is a Generalised Inverse Gaussian $\mathbf{GIG}(\frac{1}{2}, \alpha_i, \beta_i)$ where,

$$\alpha_i = \left(\frac{(1-2\alpha)^2}{2\alpha(1-\alpha)} + 2\right) \quad (8)$$

$$\beta_i = \frac{\alpha(1-\alpha)}{2}\left\langle \frac{1}{\sigma^2} \right\rangle \left(\mathbf{y}_i^2 - 2\mathbf{y}_i \langle \mathbf{f}_i \rangle + \left\langle \mathbf{f}_i^2 \right\rangle\right) \quad (9)$$

The expectations, $\left\langle \frac{1}{\mathbf{w}_i} \right\rangle = \sqrt{\frac{\alpha_i}{\beta_i}}$ and $\langle \mathbf{w}_i \rangle = \sqrt{\frac{\beta_i}{\alpha_i}} + \frac{1}{\alpha_i}$ are used in the computation of other approximate distributions.

The VB approximate posterior on $q(\sigma)$ suffers from numerical problems due to calculations of the parabolic cylindrical function (Abramowitz and Stegun 1972, p. 687). Hence, we shall restrict $q(\sigma) = IG(a, b)$, an Inverse Gamma distribution with parameters a, b. VB maximises the lower bound $\mathcal{L}_\sigma$ which can be expressed as $-KL(q_j||\tilde{p}) - \sum_{i\neq j}\int q_i \log q_i dz$ where $\log\tilde{p} = \int \log p(\mathbf{y}, \mathbf{z}) \prod_{i\neq j}(q_i dz_i)$. Thus we are required to maximise,

$$\begin{aligned}
\mathcal{L}_\sigma =& -\left(N + 1 + 10^{-6}\right)\langle \log\sigma \rangle - \gamma\left\langle \frac{1}{\sigma} \right\rangle - \delta\left\langle \frac{1}{\sigma^2} \right\rangle \\
& - \int q(\sigma) \log q(\sigma)\, d\sigma \\
\therefore \mathcal{L}_\sigma =& (a - N - 10^{-6})(\log b - \psi(a)) + (b - \gamma)\frac{a}{b} \\
& - \delta\frac{a(a+1)}{b^2} - a\log b + \log\Gamma(a) \quad (10) \\
\frac{\partial\mathcal{L}_\sigma}{\partial a} =& (N - a + 10^{-6})\psi^{(1)}(a) - \frac{\gamma}{b} - \frac{\delta(2a+1)}{b^2} + 1 \\
& \quad (11) \\
\frac{\partial\mathcal{L}_\sigma}{\partial b} =& -\frac{N}{b} + \frac{\gamma a}{b^2} + \frac{2\delta a(a+1)}{b^3} \quad (12)
\end{aligned}$$

where, $\Gamma(\cdot)$ is the gamma function, $\gamma = -\frac{1-2\alpha}{2}\sum_{i=1}^{N}(\mathbf{y}_i - \langle \mathbf{f}_i \rangle) + 10^{-6}$, $\delta = \frac{\alpha(1-\alpha)}{4}\sum_{i=1}^{N}\left\langle \frac{1}{\mathbf{w}_i} \right\rangle \left(\mathbf{y}_i^2 - 2y_i \langle \mathbf{f}_i \rangle + \left\langle \mathbf{f}_i^2 \right\rangle\right)$ and as before the expectations, $\left\langle \frac{1}{\sigma} \right\rangle = \frac{a}{b}$, $\left\langle \frac{1}{\sigma^2} \right\rangle = \frac{a(a+1)}{b^2}$ and $\langle \log\sigma \rangle = \log b - \psi(a)$ (where $\psi(\cdot)$ is the digamma function) are required. $\mathcal{L}_\sigma$ is maximised using a numerical optimiser which employs the given derivatives.

[1] Notation: Bold lower case letters represent vectors, and subscripts indicate the i-th element. Bold upper case represent matrices.

[2] Derivation shown in section A.

4 Hyper-parameter Optimisation

The only hyper-parameters in this formulation are the kernel hyper-parameters $\theta_{\mathbf{K}}$. In this framework the lower bound, $\mathcal{L}(q(\mathbf{f}, \mathbf{w}, \sigma), \theta_{\mathbf{K}})$ is maximised. In the formulations that follow, $\langle \cdot \rangle$ indicates the expectation with respect to all the latent variables, unlike what was used in the VB approximate distributions.

In order to use the lower bound it is convenient to represent $p(\mathbf{y}|\mathbf{f}, \mathbf{w}, \sigma, \mathbf{x}) = \prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{f}_i, \mathbf{w}_i, \sigma, \mathbf{x}_i)$ from equation 5 as $\mathcal{N}\left(\mathbf{y}|\mathbf{f} + \frac{1-2\alpha}{\alpha(1-\alpha)}\sigma\mathbf{w}, \frac{2}{\alpha(1-\alpha)}\sigma^2\mathrm{diag}(\mathbf{w})\right)$, its multivariate format. Due to the symmetricity of the Normal distribution with respect to its mean we may depict this distribution as, $\mathcal{N}\left(\mathbf{f}|\mathbf{y} - \frac{1-2\alpha}{\alpha(1-\alpha)}\sigma\mathbf{w}, \frac{2}{\alpha(1-\alpha)}\sigma^2\mathrm{diag}(\mathbf{w})\right)$. Hence, substituting $\mathbf{u} = \mathbf{f} - \left(\mathbf{y} - \frac{1-2\alpha}{\alpha(1-\alpha)}\sigma\mathbf{w}\right)$, $\mathbf{v} = \left\langle \mathbf{D}^{-1}(\mathbf{y} - \frac{1-2\alpha}{\alpha(1-\alpha)}\sigma\mathbf{w})\right\rangle = \left\langle \mathbf{D}^{-1}\right\rangle \mathbf{y} - \frac{1-2\alpha}{2}\left\langle \frac{1}{\sigma}\right\rangle \mathbf{1}$ and ignoring terms that do not contain $\theta_{\mathbf{K}}$ we obtain the lower bound,

$$\begin{aligned}\mathcal{L} =& \int q(\mathbf{f}|\theta_{\mathbf{K}})q(\mathbf{w})q(\sigma)\log p(\mathbf{y}|\mathbf{f}, \mathbf{w}, \sigma)p(\mathbf{f}|\theta_{\mathbf{K}})\, d\sigma d\mathbf{w} d\mathbf{f}\\ &- \int q(\mathbf{f}|\theta_{\mathbf{K}})\log q(\mathbf{f}|\theta_{\mathbf{K}})\, d\mathbf{f}\\ =& -\frac{1}{2}\left\langle \mathbf{u}^T\mathbf{D}^{-1}\mathbf{u} + \mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} + \log|\mathbf{K}|\right\rangle\\ &+ \frac{1}{2}\left\langle (\mathbf{f}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{f}-\boldsymbol{\mu}) + \log|\boldsymbol{\Sigma}|\right\rangle\\ =& -\frac{1}{2}\Big\langle \mathbf{f}^T(\mathbf{D}^{-1}+\mathbf{K}^{-1})\mathbf{f} - 2\mathbf{f}^T\mathbf{v}\\ &- \mathbf{f}^T\boldsymbol{\Sigma}^{-1}\mathbf{f} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big\rangle + \frac{1}{2}\left(\log|\boldsymbol{\Sigma}| - \log|\mathbf{K}|\right)\end{aligned}$$

Noting the three identities, $\boldsymbol{\Sigma} = \left\langle \mathbf{D}^{-1} + \mathbf{K}^{-1}\right\rangle^{-1} = \left\langle \mathbf{D}^{-1}\right\rangle^{-1}\left(\left\langle \mathbf{D}^{-1}\right\rangle^{-1} + \mathbf{K}\right)^{-1}\mathbf{K}$, $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \mathbf{v}$ and finally $\left\langle \mathbf{f}^T\mathbf{A}\mathbf{f}\right\rangle = Tr(\boldsymbol{\Sigma}\mathbf{A}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$ and ignoring terms without $\theta_{\mathbf{K}}$ the following expression is obtained,

$$\begin{aligned}\mathcal{L} =& \frac{1}{2}\left(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \log\left|\left\langle \mathbf{D}^{-1}\right\rangle^{-1} + \mathbf{K}\right|\right)\\ =& \frac{1}{2}\left(\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} - \log\left|\left\langle \mathbf{D}^{-1}\right\rangle^{-1} + \mathbf{K}\right|\right)\end{aligned} \tag{13}$$

In this setting $\mathbf{K}$ and thus $\boldsymbol{\Sigma}$ are the only terms that depends on the hyper-parameters $\theta_{\mathbf{K}}$. Equation 13 was optimised using a numerical optimiser.

5 Prediction

For a query point $\mathbf{x}_\star$, the output $\mathbf{y}_\star$ that minimises equation 1 is $\mathbf{f}_\star$. Thus unlike most Bayesian formulations where the objective is to learn $p(\mathbf{y}_\star|\mathbf{x}_\star, \mathbf{y}, \mathbf{x})$ in this particular formulation the objective is to learn the latent function $p(\mathbf{f}_\star|\mathbf{x}_\star, \mathbf{y}, \mathbf{x})$. To obtain the posterior, $p(\mathbf{f}_\star|\mathbf{x}_\star, \mathbf{y}, \mathbf{x})$ we are required to marginalise out all latent variables, $\int p(\mathbf{f}_\star|\mathbf{f}, \sigma, \mathbf{w}, \mathbf{x}_\star, \mathbf{y}, \mathbf{x}, \alpha)p(\mathbf{f}, \sigma, \mathbf{w}|\mathbf{x}, \alpha)\, d\mathbf{f}\, d\mathbf{w}\, d\sigma$. This marginalisation can be approximated to $\int p(\mathbf{f}_\star|\mathbf{f}, \mathbf{x}_\star, \mathbf{y}, \mathbf{x})q(\mathbf{f})q(\sigma)q(\mathbf{w})\, d\mathbf{f}\, d\mathbf{w}\, d\sigma$. Thus we obtain a Gaussian distribution for $p(\mathbf{f}_\star|\mathbf{x}_\star, \mathbf{y}, \mathbf{x}) \approx \mathcal{N}(\boldsymbol{\mu}_\star, \boldsymbol{\Sigma}_\star)$ for the approximate posterior where,

$$\boldsymbol{\mu}_\star = \mathbf{K}_{\mathbf{x}_\star,\mathbf{x}}\mathbf{K}_{\mathbf{x},\mathbf{x}}^{-1}\boldsymbol{\mu} \tag{14}$$

$$\boldsymbol{\Sigma}_\star = \sigma_{GP}^2 + \mathbf{K}_{\mathbf{x}_\star,\mathbf{x}}\mathbf{K}_{\mathbf{x},\mathbf{x}}^{-1}\boldsymbol{\Sigma}\mathbf{K}_{\mathbf{x},\mathbf{x}}^{-1}\mathbf{K}_{\mathbf{x}_\star,\mathbf{x}}^T \tag{15}$$

and, $\sigma_{GP}^2 = \mathbf{K}_{\mathbf{x}_\star,\mathbf{x}_\star} - \mathbf{K}_{\mathbf{x}_\star,\mathbf{x}}\mathbf{K}_{\mathbf{x},\mathbf{x}}^{-1}\mathbf{K}_{\mathbf{x}_\star,\mathbf{x}}^T$. Note in equation 15 that the variance is slightly different to that of a usual GP. This follows from using the result that $E(\mathbf{f}_\star\mathbf{f}_\star^T) = \int\int \mathbf{f}_\star\mathbf{f}_\star^T p(\mathbf{f}_\star|\mathbf{f})q(\mathbf{f})\, d\mathbf{f}_\star d\mathbf{f}$ and $Var(\mathbf{f}_\star) = E(\mathbf{f}_\star\mathbf{f}_\star^T) - E(\mathbf{f}_\star)E(\mathbf{f}_\star)^T$.

6 Experiments

Following the examples set out in (Quadrianto et al. 2009) two toy problems are conducted which are constructed as follows:

Toy Problem 1 (Heteroscedastic Gaussian Noise): 100 samples are generated from the following process. $x \sim U(-1, 1)$ and $y = \mu(x) + \sigma(x)\xi$ where $\mu = \mathrm{sinc}(x), \sigma(x) = 0.1\exp(1-x)$ and $\xi \sim \mathcal{N}(0, 1)$.

Toy Problem 2 (Heteroscedastic Chi-squared noise): 200 samples are generated from $x \sim U(0, 2)$ and $y = \mu(x) + \sigma(x)\xi$ where $\mu = \sin(2\pi x), \sigma(x) = \sqrt{\frac{2.1-x}{4}}$ and $\xi \sim \chi_{(1)}^2 - 2$.

Our algorithm is also tested in four real world examples. In the motorcycle dataset, acceleration experienced by a helmet in a crash is measured over time with the goal of interpolating between existing measurements. This is a popular dataset to assess heteroscedastic inference methods. In the bone density dataset, the goal is to predict the bone density of individuals as a function of age. The birth weight dataset aims to predict infants weight as a function of the mothers age and weight. Finally, the snow fall dataset, attempts to predict snow fall at Fort Collins in January, as a function of snow fall in September-December. We have used 80% of the data as training and the rest as testing and iterated over 20 times for each experiment. The cases were randomly permuted in each iteration.

The proposed method is compared against its nearest competitor, the EP approach, Heteroscedastic Quantile Gaussian Processes (HQGP) as well as, against a linear method (Lin) which attempts to find the quantile as a polynomial function of the inputs (polynomial basis function, in this case having $f_\alpha = \beta_0 + \beta_1 x + \beta_1 x^2 + ... + \beta_7 x^7$). The square exponential kernel was used in evaluating the VB, EP and HQGP methods. In the case of the real world datasets, the output is standardised to have zero mean and unit variance so that comparisons could be made across datasets. Note that this standardisation has not been applied to the toy data sets. Since the exact quantiles can be found for the toy datasets the Mean Absolute Deviation (MAD) and Root Mean Squared Error (RMSE) metrics have been used and are presented in table 2. The true quantiles for the real world datasets are not known *a priori*. Therefore, the average pinball loss is used as a proxy for a function that penalises incorrect quantile inference. These results are presented in table 3. Finally an empirical observed quantile error (OQE) defined as $\left|\frac{\sum_{i=1}^{N} I(\mathbf{y}_i<\boldsymbol{\mu}_{\star(i)})}{N}-\alpha\right|$ is used where I is the indicator function and the results are shown in table 3. This metric gives an estimate as to what proportion of observations are below the inferred quantile and how far this is from the intended quantile, α. This metric was provided in order to illustrate that a bias was not introduced by using the pinball loss as a metric. Different metrics were used for toy and real world problems as the true quantiles were not known for real world examples. Note that there was no code freely available for HQGP inference. Thus, the results portrayed in (Quadrianto et al. 2009) was used. [3].

The toy problem 1 was specifically designed for HQGP and therefore is not surprising that it outperforms the VB method. However, as shown in problem 2 for non-Gaussian problems the HQGP is not able to model the underlying quantiles. The HQGP inherently assumes that the quantiles lie symmetrically about the inferred mean on the dataset. This weakness is highlighted in toy problem 2.

One of the strengths of using the VB framework is its ability to infer quantiles even where observations are sparse. This is evident in its ability to infer the quantiles more accurately for the extreme quantile of 0.99 in toy problem 2 as well quantiles 0.01 and 0.99 in the real world examples. This strength is also evident when inspecting the tails of the motor cycle dataset in figure 2. The variations in accelerations experienced at the start and end of the experiment are expected to be low. This detail is better captured using VB than the EP framework as is evident in the plot. The difference in the inferred quantiles could be attributed to the fact that the posterior is better approximated by exploiting the scalar mixture of Gaussians than forcefully applying a Gaussian to the posterior (which is done in the EP method).

One of the biggest weaknesses of the HGQP is that it implies that the mean is the median, and that the quantiles are symmetrical about the mean (median). These two requirements are seemingly satisfied in the motor cycle dataset. However, in the bone density dataset there is a clear deviation from the symmetric assumption when inspecting figure 1.

The linear method, despite giving competitive error estimates, is a parametric method. This suggests that in order to get good estimates the user must manually tune the inputs and generate features. In fact, for the Fort Collins Snow dataset, instead of having a polynomial of 7th power, a cubic polynomial provided much better results. This was due to the fact that non-sensible errors (probably due to overfitting) were observed when using a polynomial of 7th power as the basis function.

7 Discussion and Future Work

In this work we have presented a Variational Bayesian approach to estimating quantiles exploiting the Gaussian scale mixture properties of Laplacian distributions. Results show that our method is able to outperform other frameworks.

Figure 2: Comparison of the quantiles obtained with (a) Variational Bayesian and (b) Expectation Propagation approaches for the motorcycle dataset. The quantiles 0.01, 0.1, 0.5, 0.9 and 0.99 are shown.

The methodology presented here can be trivially extended to parametric models by setting $f=\Phi(\mathbf{x})\mathbf{w}$ where, $\Phi(\mathbf{x})$ is a suitable basis for the problem, resulting in $p(\mathbf{f})=\mathcal{N}(\mathbf{0},\Phi(\mathbf{x})^T\Phi(\mathbf{x}))$ instead. The computational cost of inference is $\mathcal{O}(n^3)$, that of a GP. The underlying GP prior allows other GP frameworks such as those for large datasets exploiting low rank approximations and sparsity of the kernel matrices to be employed here.

One of the weaknesses of our particular setting is that quantiles are not non-crossing. Future area of research would be to impose this restriction when certain quantiles are found in previous iterations of the given algorithm. It should however be noted that in the presence of enough data, this constraint seems to be self imposing as seen in figure 1b.

A Approximate Distribution Calculations

This section will render the detailed calculations used in obtaining the approximate distributions in section 3. Recall that $\log q(z_i) \propto \langle \log p(\mathbf{y}|\mathbf{z})p(\mathbf{z})\rangle_{\prod_{j\neq i} q(z_j)}$. In fact any term that does not contain z_i can be omitted from this expression as it will form part of the normalising constant.

In order to calculate $q(\mathbf{f})$ let, $\mathbf{u}=\mathbf{f}-\left(\mathbf{y}-\frac{1-2\alpha}{\alpha(1-\alpha)}\sigma\mathbf{w}\right)$, $\mathbf{v}=\left\langle\mathbf{D}^{-1}(\mathbf{y}-\frac{1-2\alpha}{\alpha(1-\alpha)}\sigma\mathbf{w})\right\rangle$ and $\mathbf{D}=\frac{2}{\alpha(1-\alpha)}\sigma^2\mathrm{diag}(\mathbf{w})$. As shown in section 4,

[3] Code and data are available at http://www.bitbucket.org/sachinruk/gpquantile

		MAD				RMSE		
Dataset	α	VB	EP	Lin	HQGP	VB	EP	Lin
(1)	0.01	0.808±0.173	**0.233±0.145**	0.246±0.054		0.883±0.188	**0.303±0.161**	0.331±0.077
	0.10	0.109±0.089	0.110±0.088	0.121±0.035	**0.062**	**0.142±0.105**	0.146±0.104	0.177±0.074
	0.50	0.077±0.057	0.077±0.057	0.092±0.021	**0.031**	**0.100±0.069**	0.100±0.069	0.135±0.037
	0.90	0.096±0.063	0.093±0.059	0.125±0.035	**0.056**	0.128±0.094	**0.124±0.087**	0.184±0.061
	0.99	0.364±0.066	**0.199±0.093**	0.241±0.068		0.514±0.090	**0.257±0.132**	0.337±0.102
(2)	0.01	1.114±0.055	**0.016±0.003**	0.042±0.004		1.281±0.051	**0.018±0.003**	0.066±0.011
	0.10	**0.010±0.003**	0.012±0.004	0.035±0.003	0.099	**0.016±0.008**	0.018±0.007	0.053±0.010
	0.50	0.101±0.104	0.102±0.104	**0.080±0.021**	0.509	0.137±0.129	0.138±0.128	**0.115±0.045**
	0.90	0.400±0.143	0.511±0.154	**0.363±0.109**	0.804	0.526±0.210	0.663±0.209	**0.478±0.167**
	0.99	1.120±0.208	1.938±0.629	**1.027±0.261**		1.356±0.253	2.164±0.641	**1.295±0.303**

Table 2: MAD and RMSE metric for the toy problems. (1) and (2) represents the respective toy problem.

		Pin-Ball				OQE		
Dataset	α	VB	EP	Lin	HQGP	VB	EP	Lin
(1)	0.01	0.025±0.018	0.030±0.020	**0.020± 0.016**		**0.042±0.042**	0.066±0.046	0.044±0.035
	0.10	**0.076±0.020**	0.082±0.020	0.099± 0.025	0.079±0.019	0.051±0.047	0.050±0.038	**0.046±0.036**
	0.50	**0.168±0.031**	0.171±0.030	0.255± 0.046	0.187±0.020	**0.078±0.052**	0.080±0.054	0.091±0.042
	0.90	0.070±0.016	0.073±0.014	0.115± 0.061	**0.070±0.016**	0.062±0.049	0.067±0.067	**0.050±0.045**
	0.99	**0.015±0.012**	0.016±0.013	0.050± 0.080		**0.055±0.049**	0.055±0.057	0.072±0.045
(2)	0.01	**0.017±0.002**	0.025±0.007	0.021± 0.006		**0.009±0.008**	0.043±0.023	0.013±0.016
	0.10	**0.119±0.010**	0.119±0.009	0.120± 0.010	0.123±0.017	**0.031±0.019**	0.031±0.018	0.036±0.022
	0.50	0.303±0.025	**0.303±0.025**	0.304± 0.025	0.309±0.045	0.051±0.045	0.055±0.044	**0.048±0.045**
	0.90	**0.153±0.014**	0.153±0.014	0.153± 0.014	0.153±0.027	0.026±0.024	**0.025±0.020**	0.033±0.022
	0.99	**0.024±0.004**	0.038±0.021	0.025± 0.004		**0.011±0.006**	0.042±0.035	0.014±0.008
(3)	0.01	**0.063±0.039**	0.370±0.078	0.246± 0.475		**0.057±0.048**	0.420±0.085	0.077±0.050
	0.10	**0.210±0.032**	0.382±0.060	0.319± 0.274		0.061±0.048	0.323±0.098	**0.050±0.050**
	0.50	**0.404±0.024**	0.411±0.024	0.590± 0.322		0.039±0.043	**0.033±0.023**	0.060±0.055
	0.90	**0.177±0.029**	0.369±0.062	0.272± 0.178		**0.053±0.050**	0.333±0.080	0.060±0.039
	0.99	**0.040±0.018**	0.355±0.078	0.145± 0.226		**0.049±0.036**	0.428±0.078	0.080±0.036
(4)	0.01	**0.029±0.011**	0.148±0.106	0.136±0.165		**0.033±0.035**	0.216±0.134	0.061±0.040
	0.10	0.214±0.053	0.235±0.075	**0.187±0.023**		0.094±0.099	0.116±0.121	**0.041±0.035**
	0.50	**0.421±0.026**	0.437±0.020	0.483±0.075		0.060±0.042	**0.059±0.042**	0.066±0.048
	0.90	**0.237±0.041**	0.279±0.072	0.370±0.248		0.086±0.058	0.133±0.115	**0.074±0.076**
	0.99	**0.049±0.052**	0.229±0.136	0.220±0.334		**0.059±0.067**	0.255±0.155	0.096±0.089

Table 3: Pin-Ball loss and Observed Quantile Error (OQE) for real world datasets. (1): Motor Cylce, (2): Bone Density, (3): Birth Weight, (4): ftCollins Snowfall. The numbers represent the average loss for the 20 iterations and the standard deviation associated with them.

$p(\mathbf{y}|\mathbf{f},\mathbf{w},\sigma) = \mathcal{N}\left(\mathbf{f}|\mathbf{y} - \frac{1-2\alpha}{\alpha(1-\alpha)}\sigma\mathbf{w}, \mathbf{D}\right)$

$$\begin{aligned}
\log q(\mathbf{f}) &= \langle \log p(\mathbf{y}|\mathbf{f},\mathbf{w},\sigma)\rangle_{q(\mathbf{w})q(\sigma)} + \log p(\mathbf{f}) + const \\
\log q(\mathbf{f}) &\propto -\frac{1}{2}\left(\left\langle \mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}\right\rangle_{q(\mathbf{w})q(\sigma)} + \mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}\right) \\
&\propto -\frac{1}{2}\left[\mathbf{f}^T\left(\left\langle \mathbf{D}^{-1}\right\rangle + \mathbf{K}^{-1}\right)\mathbf{f} - 2\mathbf{v}^T\mathbf{f}\right] \qquad (16)
\end{aligned}$$

Simplifying $\mathbf{v}$ such that $\mathbf{v} = \left\langle \mathbf{D}^{-1}\right\rangle \mathbf{y} - \frac{1-2\alpha}{2}\left\langle \frac{1}{\sigma}\right\rangle \mathbf{1}$ and comparing equation 16 with the log of a normal distribution, $-\frac{1}{2}(\mathbf{f}^T\mathbf{\Sigma}^{-1}\mathbf{f} - \boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{f}) + const$ we obtain equations 6 and 7.

Similarly, in order to obtain $q(\mathbf{w}_i)$,

$$\begin{aligned}
\log q(\mathbf{w}_i) &= \langle \log p(\mathbf{y}|\mathbf{f},\mathbf{w},\sigma)\rangle_{q(\mathbf{f})q(\sigma)\prod_{j\neq i} q(\mathbf{w}_j)} \\
&\quad + \log p(\mathbf{w}_i) + const \\
\log(q(\mathbf{w}_i)) &= -\mathbf{w}_i - \frac{1}{2}\log(\mathbf{w}_i) - \frac{1}{2}\left\langle \frac{\alpha(1-\alpha)}{2\sigma^2\mathbf{w}_i}\mathbf{u}_i^2\right\rangle_{q(\mathbf{f})q(\sigma)}
\end{aligned}$$

For the term $\left\langle \frac{\alpha(1-\alpha)}{2\sigma^2\mathbf{w}_i}\mathbf{u}_i^2\right\rangle_{q(\mathbf{f})q(\sigma)}$ ignoring the terms that do not contain $\mathbf{w}_i$ we obtain the expression $\frac{(1-2\alpha)^2}{2\alpha(1-\alpha)}\mathbf{w}_i + \frac{\alpha(1-\alpha)}{2}\left\langle \frac{1}{\sigma^2}\right\rangle \left(\mathbf{y}_i^2 - 2y_i\langle \mathbf{f}_i\rangle + \langle \mathbf{f}_i^2\rangle\right)\frac{1}{\mathbf{w}_i}$. Thus,

$$\begin{aligned}
\log q(\mathbf{w}_i) = -\frac{1}{2}\Bigg(&\log(\mathbf{w}_i) + \left(\frac{(1-2\alpha)^2}{2\alpha(1-\alpha)} + 2\right)\mathbf{w}_i + \\
&\frac{\alpha(1-\alpha)}{2}\left\langle \frac{1}{\sigma^2}\right\rangle \left(\mathbf{y}_i^2 - 2\mathbf{y}_i\langle \mathbf{f}_i\rangle + \langle \mathbf{f}_i^2\rangle\right)\frac{1}{\mathbf{w}_i}\Bigg)
\end{aligned} \qquad (17)$$

Comparing the above to the log of a GIG distribution, $(p-1)\log \mathbf{w}_i - \frac{1}{2}\left(\alpha\mathbf{w}_i + \frac{\beta}{\mathbf{w}_i}\right) + const$ we obtain equations 8 and 9 where $p = 1/2$.

References

Abramowitz, M., and Stegun, I. A. 1972. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Number 55. Courier Dover Publications.

Bishop, C. M., et al. 2006. *Pattern recognition and machine learning*, volume 1. springer New York.

Boukouvalas, A.; Barillec, R.; and Cornford, D. 2012. Gaussian process quantile regression using expectation propagation. *arXiv preprint arXiv:1206.6391*.

Chen, K., and Müller, H.-G. 2012. Conditional quantile analysis when covariates are functions, with application to growth data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74(1):67–89.

Koenker, R., and Bassett Jr, G. 1978. Regression quantiles. *Econometrica: journal of the Econometric Society* 33–50.

Kotz, S.; Kozubowski, T.; and Podgorski, K. 2001. *The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Exonomics, Engineering, and Finance*. Number 183. Springer.

Kozumi, H., and Kobayashi, G. 2011. Gibbs sampling methods for bayesian quantile regression. *Journal of statistical computation and simulation* 81(11):1565–1578.

Quadrianto, N.; Kersting, K.; Reid, M. D.; Caetano, T. S.; and Buntine, W. L. 2009. Kernel conditional quantile estimation via reduction revisited. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, 938–943. IEEE.

Rasmussen, C. E. 2006. Gaussian processes for machine learning.

Taddy, M. A., and Kottas, A. 2010. A bayesian nonparametric approach to inference for quantile regression. *Journal of Business & Economic Statistics* 28(3).

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

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Tzikas, D. G.; Likas, C.; and Galatsanos, N. P. 2008. The variational approximation for bayesian inference. *Signal Processing Magazine, IEEE* 25(6):131–146.

Yu, K., and Moyeed, R. A. 2001. Bayesian quantile regression. *Statistics & Probability Letters* 54(4):437–447.