A Spatially Adapted Score-Based Likelihood Approach to Bayesian Quantile Regression

Marshall Honaker & John Schwenck

10 December 2020

Abstract

We propose a spatial adaptation of the score-based working likelihood function for quantile regression by deriving a non-parametric spline-based weight matrix capable of performing inference for multiple conditional quantiles simultaneously. We outline the methodology for Bayesian quantile regression for a single quantile and then generalize to multiple, making effective use of the Importance Sampling algorithm to compute posterior summaries. We then demonstrate our method's effectiveness through both a simulation study and a real-world example of NYC crime data. While the most common approaches to Bayesian spatial quantile regression tend to perform poorly in the high dimensional setting, we provide an alternative to the methods outlined in Reich et al. (2011) and Smith et al. (2015) that addresses the incorrect posterior variance commonly encountered with other Bayesian quantile regression techniques.

Introduction

Quantile regression, introduced by Koenker & Bassett (1978), has become a widely studied topic for its usefulness in characterizing an entire distribution, particularly when tail risk and extreme behavior is of interest and/or the error terms exhibit heteroskedasticity. Whereas the more common conditional mean regression tends to be sensitive to outliers, quantile regression by contrast allows for a much more general relationship between Y and X as the regression coefficients change with different levels of τ.

More formally, for a given continuous response, Y, and p-dimensional design matrix, $\boldsymbol{X}$, let $\tau \in (0,1)$ represent a given quantile, and let $Q_Y(\tau|\boldsymbol{X}=\boldsymbol{x}) = \boldsymbol{x}^T\,\boldsymbol{\beta}(\tau)$ denote the τth conditional quantile of Y given $\boldsymbol{X}=\boldsymbol{x}$ where $\boldsymbol{\beta}(\tau) \in \mathbb{R}^p$. The standard quantile regression estimator is therefore given by

$$\hat{\beta}(\tau) = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\boldsymbol{\beta})$$

where $\rho_\tau(u) = u(\tau - I(u<0))$ is the asymmetric loss function, whose asymptotically normal results from Koenker & Bassett (1978) lead to valid inference on the quantile regression parameters. From a frequentist perspective, inference for multiple quantiles follows directly. In hypothesis testing scenarios, we would test whether the effect of a given covariate has a different effect on the response at multiple separate quantiles, i.e. for two quantile levels: at τ_1 and τ_2.

There has been considerable work in this area since its inception, but it has been slower to gain traction within the Bayesian community due to the need for a specified parametric likelihood. The common approach for the Bayesian framework is to base the likelihood model on the Asymmetric Laplace Distribution (ALD) proposed by Yu & Moyeed (2001). Despite the consistency of the resulting posterior distribution demonstrated by Sriram et al. (2013), and the fairly straightforward implementation of standard MCMC procedures, Sriram

(2015) and Yang et al. (2016) have shown that the posterior variance is incorrect, leading to invalid posterior inference.

While these studies propose methods to correct the posterior variance obtained using the ALD based likelihood, Wu and Narisetty (2020) recently proposed a score-based working likelihood approach that yields valid posterior results, thus circumventing the the need for any such corrections. Their proposed methodology revolves around Bayesian multiple quantile regression for linear models, where the parameters of the resulting posterior distribution were approximated by a variant of the Importance Sampling algorithm.

Our proposed method extends their approach to a spatial setting by developing an augmented feature matrix to incorporate spatial dependencies within observed data and smooth the spatial variability, based on an approach outlined by Banerjee et al. (2008). There have been previous studies that explored the use of Bayesian quantile regression for spatially-dependent data, but to the authors' knowledge, there have been no attempts to implement the score-based framework at the time of writing. Reich et al. (2011) proposed a semi-parametic approach using Bernstein basis polynomials to model the distribution of the covariates at a particular quantile level which was later expanded by Smith et al. (2015), who used cubic splines to jointly model multiple quantiles simultaneously. However, because these revolve around a model based entirely on basis functions (i.e; not a likelihood function), they are more sensitive and less efficient in the high dimensional setting. The approach by Reich et al. (2011) is typically the most popular approach in practice and has been implemented in several applications, such as Ramsey (2019)who used this approach to model the effects of adverse weather conditions and technological advances on crop yields.

Methodology

We first define a score function and working likelihood following from the Wu and Narisetty (2020) paper. The score function seeks to optimize the regression estimator given by ***(1.1)*** through the following:

$$s_\tau(\beta) = \sum_{i=1}^{n} \boldsymbol{x_i} \psi_\tau(y_i - \boldsymbol{x_i}^T \boldsymbol{\beta})$$

where $\psi_\tau(u) = \tau - I(u < 0)$ is the check-loss function for the τth quantile. The proposed working likelihood is then given by:

$$L(Y|X, \boldsymbol{\beta}) = C \exp\left(-\frac{1}{2n} s_\tau(\boldsymbol{\beta})^T W s_\tau(\boldsymbol{\beta})\right)$$

where W is a p x p positive definite weight matrix given by:

$$W = \frac{n}{\tau(1-\tau)} \left(\sum_{i=1}^{n} x_i x_i^T\right)^{-1}$$

and C is a constant that does not depend on $\boldsymbol{\beta}$.

The above working likelihood $L(Y|X, \boldsymbol{\beta})$, with this specification of the $p \times p$ positive definite weight matrix W as the sampling model, leads to a posterior distribution that is numerically asymptotically normal. Because this working likelihood is not derived from any distributional specification of $Y|(X, \beta)$, the adaptive Importance Sampling algorithm proposed by Wu and Narisetty (2020) can then be implemented to approximate the mean vector and covariance matrix of the resulting posterior distribution.

In the case of modeling multiple m quantiles jointly, the score function will then incorporate these m quantiles jointly and the weight matrix will be reconfigured into an $mp \times mp$ block matrix $W = (Q \otimes G)^{-}1$ where $Q = (\min(\tau_i, \tau_j) - \tau_i \tau_j)_{ij}$ and $G = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$

As in Wu and Narisetty (2020), we explore two choices of prior distribution for the model coefficients: a uniform prior and a multivariate normal prior. By nature, the uniform prior is non-informative and can be used when there is little or no reliable information regarding the population under study before the experiment is conducted or the data is analyzed. However, when such relevant information is available, it

can be encoded in a multivariate normal prior distribution. Though not discussed in Wu and Narisetty (2020), it may be useful to explore the performance of invariant priors such as a g-prior. In any case, we will first explore the performance of this approach when using a multivariate uniform prior.

As a way to borrow information from the multivariate Normal proposal distribution, we make use of the importance sampling algorithm to calculate the importance weights needed for posterior estimates. This method is particularly relevant in a Bayesian framework in order to harness all information from then working likelihood, prior, and proposal distribution. The idea is to approximate quantities of interest from a given distribution (in our case, the proposal) indirectly as $\mathbb{E}_g[X] = \sum_x x \frac{g(x)}{f(x)} f(x) = \mathbb{E}_f\left[X \frac{g(x)}{f(x)}\right]$. So, $\mathbb{E}_g(X) \approx \frac{1}{n} \sum_{x=1}^{n} x_i \frac{g(x_i)}{f(x_i)}$ where $w_i = \frac{g(x_i)}{f(x_i)}$ are the importance sampling weights. For our implementation, the importance weights are given by $w^{(r)} = \frac{L(b^{(r)})\pi(x^{(r)})}{q(b^{(r)})}$

When working with spatial data, it is imperative that the model consider and account for the spatial variability and dependencies within the data. Failure to do so can lead to inaccurate predictions, loss of power, and invalid inferential procedures. To introduce these spatial dependencies to the score-based working likelihood approach outlined above, we will generate an augmented feature matrix by appending a matrix of spatial basis covariate functions to the standard n x p feature matrix X. Then, implementing the framework outlined by Wu and Narisetty (2020) will yield a Bayesian quantile regression model that takes into account the locations at which various observations were made as part of its input.

To account for the spatial dependence of observed data, we introduce an augmented feature matrix $\tilde{X}$ by appending an n x b matrix of spatial basis covariate functions to the standard n x p feature matrix. Using this feature matrix to implement the framework outlined above yields a Bayesian quantile regression that incorporates the locations at which observations were made as part of its input. The spatial basis covariate functions were obtained using an approach similar to that used by Banerjee et al. (2008) to derive the spatial predictive process $\tilde{w}(\boldsymbol{s})$.

Assuming a spatial Gaussian process $w(\boldsymbol{s}) \sim MVN(\mathbf{0}, C_s)$ where C_s is the spatial covariance matrix of the observed locations, we then place m knots at equally spaced intervals over the relevant spatial domain. Since we have assumed a spatial Gaussian process, these knots can be described by $w^*(\boldsymbol{s_k}) \sim MVN_m(\mathbf{0}, C^*)$. As stated in Banerjee et al. (2008), the spatial effect at a particular location s_0 can be interpolated as:

$$\begin{aligned} \tilde{w}(s_0) &= \mathbf{E}[w(s_0)|w^*] \\ &= \boldsymbol{c}_s^T(s_0) C^{*-1} w* \end{aligned}$$

However, since w^* follows a Gaussian process with mean vector $\mathbf{0}$ and covariance matrix C^*, the above can equivalently expressed as:

$$\tilde{w}(s_0) = \boldsymbol{c}_s^T(s_0) \left(C^{*-1}\right)^{\frac{1}{2}} N(\mathbf{0}, 1)$$

As stated in Banerjee et al. (2008), the interpolated value at a particular location s_0 as given above defines another, predictive spatial process $\tilde{w}(\boldsymbol{s})$. Using $\tilde{w}(\boldsymbol{s})$, we are able to interpolate the spatial effect at the knots for each observation, yielding an n x m^2 matrix of interpolated spatial effects. We then append this matrix of interpolated values to the standard n x p feature matrix to obtain an n x $(p + m^2)$ augmented feature matrix of the form:

$$\tilde{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} & b_{11} & b_{12} & \dots & b_{1m^2} \\ x_{21} & x_{22} & \dots & x_{2p} & b_{21} & b_{22} & \dots & b_{2m^2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} & b_{n1} & b_{n2} & \dots & b_{nm^2} \end{bmatrix}$$

To implement the adaptive importance sampling approach for a single quantile, we used the steps given by Wu and Narisetty (2020) but incorporated the spatial adaptation as mentioned above. After first initializing

the proposal distribution with ordinary linear regression with $q(b) = N(a, \hat{\Sigma})$ where $\hat{\Sigma}$ is given by

$$\hat{\Sigma} = c\tau(1-\tau)\left(\sum_{i=1}^{n} \frac{x_i x_i^T}{n}\right)^{-1}$$

, we generate samples $b^{(1)}, b^{(2)}, \ldots, b^{(m)}$ from $q(b)$ from some large M and calculate the importance weights. Using the importance weights, we can then estimate the posterior mean and covariance as given by:

$$\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, ..., \hat{\mu}_m) = \frac{\sum_{i=1}^{M} w^{(r)} \mathbf{b}^{(r)}}{\sum_{i=1}^{M} w^{(r)}}$$

and

$$\hat{\boldsymbol{\sigma}}_{\boldsymbol{j,k}} = \frac{\sum_{r=1}^{M} w^{(r)} \mathbf{b}_j^{(r)} \mathbf{b}_k^{(r)}}{\sum_{r=1}^{M} w^{(r)}} - \hat{\mu}_j \hat{\mu}_k$$

for the jk^{th} entry of the posterior covariance matrix where $b_j^{(r)}$ is the j^{th} coordinate of $b^{(r)}$. These estimates will then be used to form new proposal distributions which will be repeated a pre-specifed number of iterations. The posterior mean and covariance can then be estimated from the final proposal distribution.

In the context of multiple quantiles, the process is similar but requires transforming the data structure into block-diagonal form. For each quantile of interest, perform the algorithm for a single quantile as described above. For computing the posterior and covariance estimates, the resulting mean will now become an $m \times p$ matrix and the resulting covariance will become an mp-dimensional block-diagonal matrix. The off-diagonal covariances will then need to be imputed for quantile levels τ_i, τ_j using the following from Wu and Narisetty (2020):

$$\hat{\Sigma}_{i,j} = \gamma(\min(\tau_i, \tau_j) - \tau_i \tau_j)\left(\frac{(\tau_i(1-\tau_i)\hat{\Sigma}_i^{-1} + \tau_j(1-\tau_j)\hat{\Sigma}_j^{-1})}{2}\right)^{-1}$$

From these, generate samples and estimate the final posterior mean and covariance as described for the single quantile case.

Simulation

To assess the performance of our approach, we will evaluate our model using simulated data. 10,000 observations were simulated from random locations on a standardized 10 x 10 grid using a spatial Gaussian process. The simulated data was then split into a training and a test data set. 75% of the simulated values were allocated to the training data set and the remaining 25% was allocated to the test data set. We then specify a set of quantiles $\vec{\tau} = \{\tau_1, \tau_2, ..., \tau_t\}$ and fit a model for each individual quantile and another to model all of the quantiles simultaneously, using the Importance Sampling algorithms outlined by Wu & Narisetty (2020).

Once these models have been fit, we will predict the value at the locations in the test data set. However, to make the desired predictions, we will need to derive or approximate the posterior predictive distribution, given by:

$$\begin{aligned} P(\tilde{y}|\vec{y}) &= \int p(\tilde{y}|\vec{\beta}, \vec{y}) p(\vec{\beta}|\vec{y}) d\vec{\beta} \\ &= \int p(\tilde{y}|\vec{\beta}) p(\vec{\beta}|\vec{y}) d\vec{\beta} \\ &= \int L(\tilde{y}|X, \vec{\beta}) p(\vec{\beta}|\vec{y}) d\vec{\beta} \end{aligned}$$

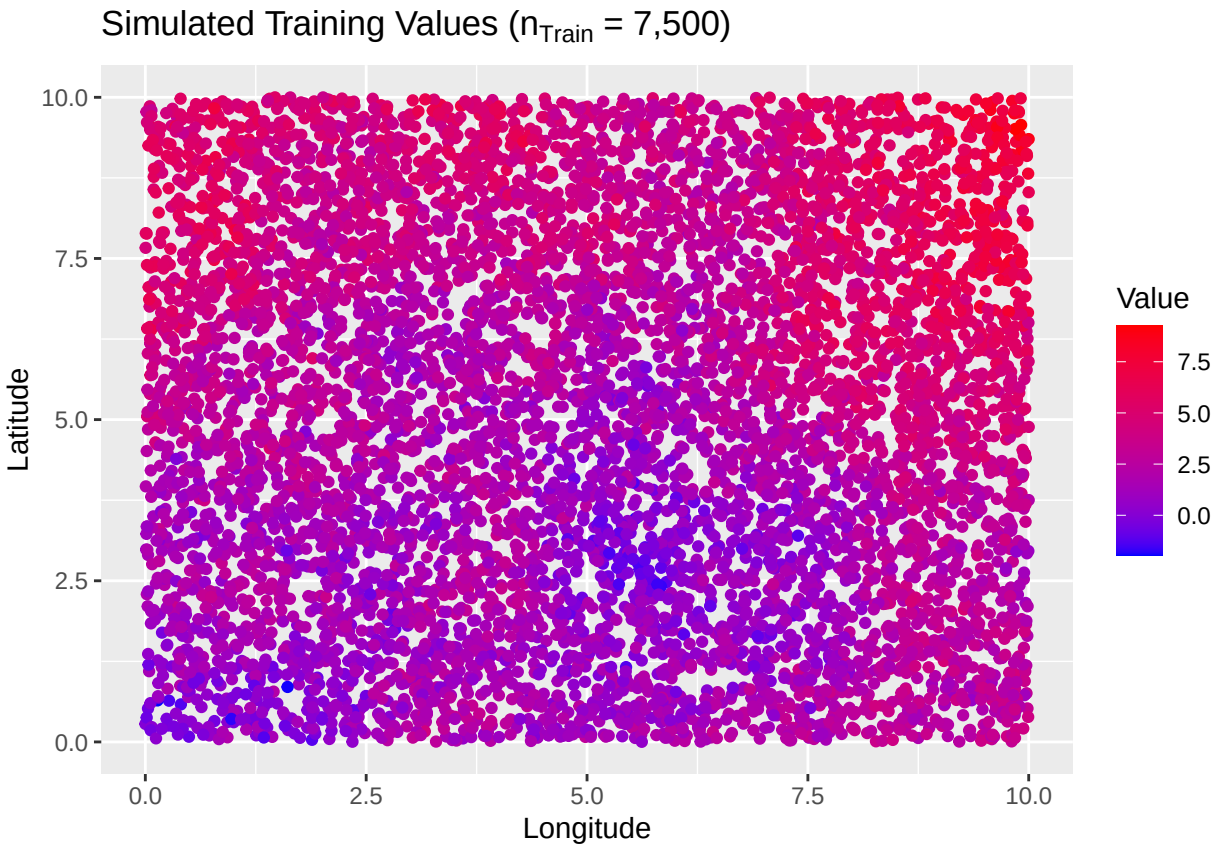
where $\vec{y}$ are the observed (simulated) observations and $\tilde{y}$ is an as of yet unobserved value from the same population as $\vec{y}$. The Metropolis-Hastings Algorithm can be used to make this approximation and obtain

the desired predictions. Once we have obtained these predicted values, we will be able to calculate and compare the training and test error rates; various information criteria (the Bayesian Information Criterion, the Deviance Information Criterion, etc.); the Mean Squared Error(s); and any other metrics we may be interested in. Using these metrics, we will be able to compare the performance of our approach to the models developed by Reich et al. (2011) and Smith et al. (2015).

Lastly, we will use our simulated data to assess the validity of inference procedures based on the posterior distribution(s) for the models for a single quantile and the model for multiple quantiles. This is a central goal of the paper by Wu & Narisetty and it is important that models developed using our approach retain this property. To assess the validity of posterior inference procedures, we will examine the coverage probabilities of confidence intervals for the parameters in each model and test for the equality of coefficients on the same variable at different quantile levels.

As mentioned previously, we have not been able to develop a fully functional model using the approach outlined above. Therefore, we are not able to present the results of the simulation study in full. However, the data that will be used to train and test the model once it is fully developed and properly functioning has been simulated and a sample is given below. (The code used to generate this and all other simulated data sets is provided in the accompanying file.)

```
## Warning: package 'GpGp' was built under R version 4.0.3
```



Real-World Application: NYC Crime

(until code is finalized, there is not much to say here. . .)

Discussion

Limitations

We have encountered several issues throughout this process, most notably involving the values within the weight matrix W. The values within the weight matrix W tend to be quite large (of the order 10^4). Because the weight matrix appears in the quadratic exponent of the likelihood function, this causes the likelihood to be extremely small. In preliminary simulations, $L(Y|X,\boldsymbol{\beta}) = 0$. This causes further problems within the importance sampling algorithm, because the importance weights (given by $w^{(r)} = \frac{L(b^{(r)}\pi^{(r)})}{q(b^{(r)})}$) also depend on the likelihood function.

Several attempts were made to address and remedy these issues, but so far none have consistently yielded valid results. The most promising of results obtained so far involved altering the expression used to compute the importance weights within the importance sampling algorithm. Instead of computing the importance weights as $w^{(r)} = \frac{L(b^{(r)}\pi^{(r)})}{q(b^{(r)})}$ where $L(Y|X,\boldsymbol{\beta}) = C\exp\left(-\frac{1}{2n}s_\tau(\boldsymbol{\beta})^T W s_\tau(\boldsymbol{\beta})\right)$ and $q(b^{(r)}) = (2\pi)^{-\frac{n}{2}}(det(\Sigma))^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$, the importance weights were computed as $w^{(r)} = C(2\pi)^{\frac{n}{2}}(det(\Sigma))^{\frac{1}{2}}\exp\left(\frac{-\frac{1}{2n}s_\tau(\boldsymbol{\beta})^T W s_\tau(\boldsymbol{\beta})}{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)}\right)$ While this yields minimally better results, the posterior covariance matrix obtained from the importance sampling algorithm is clearly incorrect.

Future Directions

While previous studies have discussed and even implemented Bayesian spatial quantile regression models, this paper presents a framework for a general model that provides valid posterior inference. However, this framework could be extended to include a regularization term that would allow for model selection at various quantile levels. By comparing the variables included in the model at various quantile levels, one would be able to examine which variables play an active role in modeling the response at various points in the distribution. It is worth noting that this extension need not be specific to a spatial context, but the regularization literature is much less established in such a domain.

From a reproducibility and optimization standpoint, we will alleviate bottlenecks within the necessary algorithms by translating R code to C++ code, and will restructure it to incorporate compatibility checks and other fail-safe measures.

Supplementary Material

All code is hosted on GitHub through the repository: https://github.com/johnschwenck/crimeBQR

Publication Note

If at all possible, we would like to submit this report for publication in a journal once it has been completed.

References

1. Banerjee, S., Gelfand, A., Finley, A., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *J R Stat Soc Series B Stat Methodol*, 70(4), 825–848. doi:10.1111/j.1467-9868.2008.00663.x.

2. Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46(1), 33-50. doi:10.2307/1913643

3. Ramsey, A.F. (2020), Probability Distributions of Crop Yields: A Bayesian Spatial Quantile Regression Approach. *Amer. J. Agr. Econ.*, 102(1), 220-239. doi:10.1093/ajae/aaz029

4. Reich, B. J., Fuentes, M., & Dunson, D. B. (2011). Bayesian Spatial Quantile Regression. *Journal of the American Statistical Association*, 106(493), 6–20. doi:10.1198/jasa.2010.ap09237

5. Smith, L. B., Reich, B. J., Herring, A. H., Langlois, P. H., & Fuentes, M. (2015). Multilevel quantile function modeling with application to birth outcomes. *Biometrics*, 71(2), 508–519. doi:10.1111/biom.12294

6. Sriram K. (2015). A sandwich likelihood correction for Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Statistics & Probability Letters*, 107, 18-26. doi:10.1016/j.spl.2015.07.035

7. Sriram, K., Ramamoorthi, R.V., Ghosh, P. (2013). Posterior Consistency of Bayesian Quantile Regression Based on the Misspecified Asymmetric Laplace Density. *Bayesian Analysis*. 8(2), 479-504. doi:10.1214/13-BA817.

8. Wu, T., & Narisetty, N. (2020). Bayesian Multiple Quantile Regression for Linear Models Using a Score Likelihood. *Bayesian Analysis*, adv issue. doi:10.1214/20-BA1217

9. Yang, Y., Wang, H., & He, X. (2016). Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *International Statistical Review*, 84(3), 327-344. doi:10.1111/insr.12114

10. Yu, K., & Moyeed, R. (2001). Bayesian Quantile Regression. *Statistics & Probability Letters*, 54(4), 437-447. doi:10.1016/S0167-7152(01)00124-9