A Bayesian Approach to Multiple-Output Quantile Regression

Michael Guggisberg*
Institute for Defense Analyses

May 4, 2022

Abstract

This paper presents a Bayesian approach to multiple-output quantile regression. The prior can be elicited as ex-ante knowledge of the distance of the τ-Tukey depth contour to the Tukey median, the first prior of its kind. The parametric model is proven to be consistent and a procedure to obtain confidence intervals is proposed. A proposal for nonparametric multiple-output regression is also presented. These results add to the literature of misspecified Bayesian modeling, consistency, and prior elicitation of nonparametric multivariate modeling. The model is applied to the Tennessee Project Steps to Achieving Resilience (STAR) experiment and finds a joint increase in *τ-quantile subpopulations* for mathematics and reading scores given a decrease in the number of students per teacher.

Keywords: Bayesian Methods, Quantile Estimation, Multivariate Methods

[*] The author gratefully acknowledges *the School of Social Sciences at the University of California, Irvine, and the Institute for Defense Analyses for funding this research. The author would also like to thank Dale Poirier, Ivan Jeliazkov, David Brownstone, Daniel Gillen, Karthik Sriram, Brian Bucks, and two anonymous reviewers for their helpful comments.*

1 Introduction

Single-output (i.e., univariate) quantile regression, originally proposed by Koenker and Bassett (1978), is a popular method of inference among empirical researchers (see Yu et al. (2003) for a survey). Yu and Moyeed (2001) formulated a Bayesian framework for quantile regression. This advance opened the doors for Bayesian inference of quantiles and generated a series of applied and methodological research.[1]

A multiple-output (i.e., multivariate) quantile can be defined in many different ways and there has been little consensus on which is the most appropriate (Small, 1990; Chaudhuri, 1996; Serfling, 2002; Wei, 2008; Serfling and Zuo, 2010; Hallin et al., 2010; Kong and Mizera, 2012; Carlier et al., 2016). Advancements for Bayesian multiple-output quantiles are sparse. A recent paper uses a geometric definition for a multiple-output quantile location (Bhattacharya and Ghosal, 2020). Two other previous approaches exist, but neither used a commonly accepted definition for a multiple-output quantile (Drovandi and Pettitt, 2011; Waldmann and Kneib, 2014).

This paper presents a Bayesian framework for multiple-output quantiles defined parametrically in Laine (2001) and Hallin et al. (2010) and nonparametrically in Hallin et al. (2015). Their directional quantiles coincide with Tukey halfspace depth contours and can be computed with standard single-output quantile regression techniques (Hallin et al., 2010). See McKeague et al. (2011) for a frequentist application to growth trajectories and Santos and Kneib (2020) for a Bayesian extension and application to student scores on the Brazilian High School National Exam.

Both the parametric and nonparametric models can be extended to include regression

[1] For example, see Taddy and Kottas (2010); Thompson et al. (2010); Alhamzawi et al. (2012); Kozumi and Kobayashi (2011); Benoit and Van den Poel (2012); Benoit and Van den Poel (2017); Feng et al. (2015); Kottas and Krnjajić (2009); Lancaster and Jae Jun (2010); Rahman (2016); Sriram et al. (2016).

with covariates. However, the τ-quantile contours in the parametric regression model require severe restrictions to be depth contours of the conditional distribution due to an issue analogous to quantile crossing in the single-output case (Hallin et al., 2015; Koenker et al., 2018). In which case the nonparametric model can provide the correct depth contours. However, the nonparametric model has a curse of dimensionality and can have large parameter spaces with long computation times.

Validity of the proposed approach leverages an idea similar to Chernozhukov and Hong (2003) using a likelihood that is not necessarily representative of the Data Generating Process (DGP). Despite misspecification, the posterior of the parametric model is proven to converge almost surely to the population parameters. The posterior of the nonparametric model is shown via simulation to converge to the population parameters as well. Frequentist confidence intervals can be obtained for the location case of the parametric model and are shown through simulation to have proper coverage. These results further motivate the use of misspecified Bayesian models for real world data analysis.

By performing inference in this framework, one gains many advantages of a Bayesian analysis. The Bayesian machinery provides a principled way of combining prior knowledge with data to arrive at conclusions. This machinery can be used in a data-rich world, where data is continuously collected, to make inferences and update them in real time. The proposed approach can take more computational time than the frequentist approach, since the proposed posterior sampling algorithm recommends initializing the Markov Chain Monte Carlo (MCMC) sequence at the frequentist estimate. Thus, if the researcher does not desire to provide prior information or perform online learning, the frequentist approach may be more desirable than the proposed approach. An anonymous reviewer pointed out that directional quantiles can be used to infer boundaries of arbitrarily bounded multivariate distributions.

The prior is a required component in Bayesian analysis where the researcher elicits their pre-analysis beliefs for the population parameters. The prior for the parametric model is closely related to the Tukey depth of a distribution, a notion of multiple-output centrality of a data point (Tukey, 1975). The prior can be elicited as the Euclidean distance of the Tukey median from a (spherical) τ-Tukey depth contour. This is the first Bayesian prior for Tukey depth and motivates further research for nonparametric prior elicitation.

Once a prior is chosen, estimates can be computed using MCMC draws from the posterior. A Gibbs MCMC sampler can be used if the researcher is willing to accept prior joint normality of the model parameters. Gibbs samplers have many computational advantages over other MCMC algorithms such as easy implementation, efficient convergence to the stationary distribution, and little to no parameter tuning. Consistency of the posterior and a Bernstein-Von Mises result are verified via a simulation study.

The models are applied to the Tennessee Project Steps to Achieving Resilience (STAR) experiment (Finn and Achilles, 1990). The goal of the experiment was to determine if classroom size has an effect on learning outcomes. The effect of decreasing classroom size is shown to improve test scores by comparing τ-quantile contours for mathematics and reading test scores of first grade students. Further it is found that τ-quantile subpopulations of mathematics and reading scores improve for both central and outlying students in smaller classrooms compared to larger classrooms. This result is consistent with, and much stronger than the result one would find with multiple-output linear regression. An analysis by multiple-output linear regression finds mathematics and reading scores improve *on average*; however, there could still be subpopulations where the score declines. The multiple-output quantile regression approach confirms there are no quantile subpopulations where the score declines (of the inspected subpopulations).

2 Bayesian multiple-output quantile regression

This section presents the parametric Bayesian approach to quantile regression. Notation common to both parametric and nonparametric approaches is presented, followed by the definition of the parametric model and a theorem of consistency for the Bayesian estimator (section 2.1). Parameter interpretations related to τ-Tukey depth contours are shown (subsection 2.1.1). Then a method to construct asymptotic frequentist confidence intervals is shown (section 2.2). The prior is then discussed (section 2.3). Expectations and probabilities in sections 2.1 and 2.2 are conditional on parameters. Expectations in section 2.3 are with respect to prior parameters. Appendix E presents the nonparametric approach and Appendix A reviews additional details on multiple-output quantiles.

Let $[Y_1, Y_2, ..., Y_k]' = \mathbf{Y}$ be a k-dimension random vector. The direction and magnitude of the directional quantile is defined by $\boldsymbol{\tau} \in \mathcal{B}^k = \{\mathbf{v} \in \Re^k : 0 < ||\mathbf{v}||_2 < 1\}$. Where $\mathcal{B}^k$ is a k-dimension unit ball centered at $\mathbf{0}$ (with center removed). Define $||\cdot||_2$ to be the l_2 norm. The vector $\boldsymbol{\tau}= \tau\mathbf{u}$ can be broken down into direction, $[u_1, u_2, ..., u_k]' = \mathbf{u} \in \mathcal{S}^{k-1} = \{\mathbf{v} \in \Re^k : ||\mathbf{v}||_2 = 1\}$ and magnitude, $\tau \in (0, 1)$.

Let $\boldsymbol{\Gamma}_{\mathbf{u}}$ be a $k \times (k-1)$ matrix such that $[\mathbf{u} \vdots \boldsymbol{\Gamma}_{\mathbf{u}}]$ is an orthonormal basis of $\Re^k$. Define $\mathbf{Y}_{\mathbf{u}} = \mathbf{u}'\mathbf{Y}$ and $\mathbf{Y}_{\mathbf{u}}^{\perp} = \boldsymbol{\Gamma}_{\mathbf{u}}'\mathbf{Y}$. Let $\mathbf{X} \in \Re^p$ to be random covariates. Define the ith observation of the jth component of $\mathbf{Y}$ to be $\mathbf{Y}_{ij}$ and the ith observation of the lth covariate of $\mathbf{X}$ to be $\mathbf{X}_{il}$ where $i \in \{1, 2, ..., n\}$ and $l \in \{1, 2, ..., p\}$.

2.1 Parametric model

Define $\Psi^u(a, \mathbf{b}) = E[\rho_\tau(\mathbf{Y}_{\mathbf{u}} - \mathbf{b}_{\mathbf{y}}'\mathbf{Y}_{\mathbf{u}}^{\perp} - \mathbf{b}_{\mathbf{x}}'\mathbf{X} - a)]$ to be the objective function of interest. Hallin et al. (2015) refers to this model as the “unconditional” model because the expectation in the objective function does not condition on the covariates. In this paper it is referred to as the parametric model. The $\boldsymbol{\tau}$th quantile regression of $\mathbf{Y}$ on $\mathbf{X}$ (and an intercept) is $\lambda_{\boldsymbol{\tau}} = \{\mathbf{y} \in \Re^k : \mathbf{u}'\mathbf{y} = \beta'_{\boldsymbol{\tau}\mathbf{y}}\boldsymbol{\Gamma}'_{\mathbf{u}}\mathbf{y} + \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X} + \alpha_{\boldsymbol{\tau}}\}$ where

$$(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}\mathbf{y}}, \beta_{\boldsymbol{\tau}\mathbf{x}}) \in \underset{a,\mathbf{b_y},\mathbf{b_x}}{argmin}\, \Psi^u(a, \mathbf{b}). \tag{1}$$

The definition of the location case is embedded in definition (1), where $\mathbf{b_x}$ and $\mathbf{X}$ are of null dimension. A τ-quantile contour is produced by obtaining the boundary of the intersection of the closed upper halfspaces of $\lambda_{\boldsymbol{\tau}}$ for all $\mathbf{u}$ and a fixed τ (rigorously defined in Appendix A equation (9)). The location specific τ-quantile contours can be interpreted as τ-Tukey depth contours. However, without severe restrictions, the interpretation is lost if covariates are included (Hallin et al., 2015; Koenker et al., 2018). In which case more flexible parametric modeling (e.g., polynomial) or a nonparametric approach can be used to obtain proper contours.

Note that $\beta_{\boldsymbol{\tau}\mathbf{y}}$ is a function of $\boldsymbol{\Gamma}_{\mathbf{u}}$. This relationship is of little importance; the uniqueness of $\beta'_{\boldsymbol{\tau}\mathbf{y}}\boldsymbol{\Gamma}'_{\mathbf{u}}$, which is guaranteed under Assumption 2 (presented in the next section), is of greater interest. Thus, the choice of $\boldsymbol{\Gamma}_{\mathbf{u}}$ is unimportant as long as $[\mathbf{u} \vdots \boldsymbol{\Gamma}_{\mathbf{u}}]$ is orthonormal.[2]

The population parameters satisfy two subgradient conditions

$$\left.\frac{\partial \Psi^u(a, \mathbf{b})}{\partial a}\right|_{\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}} = Pr(\mathbf{Y_u} - \beta'_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y}^{\perp}_{\mathbf{u}} - \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X} - \alpha_{\boldsymbol{\tau}} \leq 0) - \tau = 0 \tag{2}$$

and

$$\left.\frac{\partial \Psi^u(a, \mathbf{b})}{\partial \mathbf{b}}\right|_{\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}} = E[[\mathbf{Y}^{\perp'}_{\mathbf{u}}, \mathbf{X}']' 1_{(\mathbf{Y_u} - \beta'_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y}^{\perp}_{\mathbf{u}} - \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X} - \alpha_{\boldsymbol{\tau}} \leq 0)}] - \tau E[[\mathbf{Y}^{\perp'}_{\mathbf{u}}, \mathbf{X}']'] = \mathbf{0}_{k+p-1}. \tag{3}$$

The expectations need not exist if observations are in general position (Hallin et al., 2010).

[2] The choice of $\boldsymbol{\Gamma}_{\mathbf{u}}$ could possibly affect the efficiency of MCMC sampling and convergence speed of the MCMC algorithm to the stationary distribution.

Interpretations of the subgradient conditions are presented in Appendix A, one of which is new to the literature and will be restated here. The second subgradient condition can be rewritten as

$$E[\mathbf{Y}_{\mathbf{u}i}^{\perp}|\mathbf{Y}_{\mathbf{u}} - \beta_{\boldsymbol{\tau}\mathbf{y}}'\mathbf{Y}_{\mathbf{u}}^{\perp} - \beta_{\boldsymbol{\tau}\mathbf{x}}'\mathbf{X} - \alpha_{\boldsymbol{\tau}} \leq 0] = E[\mathbf{Y}_{\mathbf{u}i}^{\perp}] \text{ for all } i \in \{1, ..., k-1\}$$
$$E[\mathbf{X}_i|\mathbf{Y}_{\mathbf{u}} - \beta_{\boldsymbol{\tau}\mathbf{y}}'\mathbf{Y}_{\mathbf{u}}^{\perp} - \beta_{\boldsymbol{\tau}\mathbf{x}}'\mathbf{X} - \alpha_{\boldsymbol{\tau}} \leq 0] = E[\mathbf{X}_i] \text{ for all } i \in \{1, ..., p\}$$

This shows the probability mass center in the lower halfspace for the orthogonal response is equal to that of the probability mass center in the entire orthogonal response space. Likewise for the covariates, the probability mass center of being in the lower halfspace is equal to the probability mass center in the entire covariate space.

The Bayesian approach assumes

$$\mathbf{Y}_{\mathbf{u}}|\mathbf{Y}_{\mathbf{u}}^{\perp}, \mathbf{X}, \alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}} \sim ALD(\beta_{\boldsymbol{\tau}\mathbf{y}}'\mathbf{Y}_{\mathbf{u}}^{\perp} + \beta_{\boldsymbol{\tau}\mathbf{x}}'\mathbf{X} + \alpha_{\boldsymbol{\tau}}, \sigma_{\boldsymbol{\tau}}, \tau)$$

whose density is

$$f_{\boldsymbol{\tau}}(\mathbf{Y}|\mathbf{X}, \alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}, \sigma_{\boldsymbol{\tau}}) = \frac{\tau(1-\tau)}{\sigma_{\boldsymbol{\tau}}} exp(-\frac{1}{\sigma_{\boldsymbol{\tau}}}\rho_{\tau}(\mathbf{Y} - \beta_{\boldsymbol{\tau}\mathbf{y}}'\mathbf{Y}_{\mathbf{u}}^{\perp} - \beta_{\boldsymbol{\tau}\mathbf{x}}'\mathbf{X} - \alpha_{\boldsymbol{\tau}})).$$

The nuisance scale parameter, $\sigma_{\boldsymbol{\tau}}$, is fixed at 1.[3] The likelihood is

$$L_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) = \prod_{i=1}^{n} f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i, \alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}, 1). \tag{4}$$

The ALD distributional assumption likely does not represent the DGP, and is thus a misspecified distribution. However, as more observations are obtained, the posterior

[3] The nuisance parameter is sometimes taken to be a free parameter in single-output Bayesian quantile regression (Kozumi and Kobayashi, 2011). The posterior has been shown to still be consistent with a free nuisance scale parameter in the single-output model (Sriram et al., 2013). This paper does not attempt to prove consistency with a free nuisance scale parameter.

probability mass concentrates around neighborhoods of $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$, where $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$ satisfies (2) and (3). Theorem 1 shows this posterior consistency.

The assumptions for Theorem 1 are below.

Assumption 1. *The observations $(\mathbf{Y}_i, \mathbf{X}_i)$ are independent and identically distributed (i.i.d.) with true measure $\mathbf{P}_0$ for $i \in \{1, 2, ..., n, ...\}$.*

The density of $\mathbf{P}_0$ is denoted p_0. Assumption 1 states the observations are independent. This still allows for dependence among the components within a given observation (e.g., heteroskedasticity that is a function of $\mathbf{X}_i$). The i.i.d. assumption is required for the subgradient conditions to be well defined.

The next assumption causes the subgradient conditions to exist and be unique, ensuring the population parameters,$(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$, are well defined.

Assumption 2. *The measure of $(\mathbf{Y}_i, \mathbf{X}_i)$ is continuous with respect to Lebesgue measure, has connected support, and admits finite first moments for all $i \in \{1, 2, ..., n, ...\}$.*

Serfling and Zuo (2010) show this assumption can be weakened to not require that moments exist. The next assumption describes the prior.

Assumption 3. *The prior, $\Pi_{\boldsymbol{\tau}}(\cdot)$, has positive measure for every open neighborhood of $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$ and is*

a) proper, or

b) improper but admits a proper posterior.

Case b includes the Lebesgue measure on $\Re^{k+p}$ (i.e., flat prior) as a special case (Yu and Moyeed, 2001, Theorem 1). Assumption 3 is satisfied using the joint normal prior suggested in section 2.3.

The next assumption bounds the covariates and response variables to ensure the expectation of the likelihood is finite.

Assumption 4. *There exists a* $c_x > 0$ *such that* $|\mathbf{X}_{i,l}| < c_x$ *for all* $l \in \{1, 2, ..., p\}$ *and all* $i \in \{1, 2, ..., n, ...\}$. *There exists a* $c_y > 0$ *such that* $|\mathbf{Y}_{i,j}| < c_y$ *for all* $j \in \{1, 2, ..., k\}$ *and all* $i \in \{1, 2, ..., n, ...\}$. *There exists a* $c_\Gamma > 0$ *such that* $\sup\limits_{i,j} |[\mathbf{\Gamma_u}]_{i,j}| < c_\Gamma$.

The restriction on $\mathbf{X}$ is fairly mild in application; any given dataset will satisfy these restrictions. Further, $\mathbf{X}$ can be controlled by the researcher in some situations (e.g., experimental environments). The restriction on $\mathbf{Y}$ is more contentious since it is less common. However, like $\mathbf{X}$, any given dataset will satisfy this restriction (and is satisfied in this paper's application). A simulation in section 4 shows that this assumption can be violated. The assumption on $\mathbf{\Gamma_u}$ is innocuous since $\mathbf{\Gamma_u}$ is chosen by the researcher.

The next assumption ensures the Kullback-Leibler minimizer is well defined.

Assumption 5. $E \log \left(\frac{p_0(\mathbf{Y}_i, \mathbf{X}_i)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i | X_i, \alpha, \beta, 1)} \right) < \infty$ *for all* $i \in \{1, 2, ..., n, ...\}$.

The next assumption is to ensure the orthogonal response and covariate vectors are not degenerate.

Assumption 6. *There exist vectors* $\epsilon_Y > \mathbf{0}_{k-1}$ *and* $\epsilon_X > \mathbf{0}_p$ *such that*

$$Pr(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Yj}, \mathbf{X}_{il} > \epsilon_{Xl}, \forall j \in \{1, ..., k-1\}, \forall l \in \{1, ..., p\}) = c_p \notin \{0, 1\}.$$

This assumption can always be satisfied with a simple location shift as long as each variable takes on at least two different values with positive joint probability. Let $U \subseteq \Theta$, define the posterior probability of U to be

$$\Pi_{\boldsymbol{\tau}}(U | (\mathbf{Y}_1, \mathbf{X}_1), (\mathbf{Y}_2, \mathbf{X}_2), ..., (\mathbf{Y}_n, \mathbf{X}_n)) = \frac{\int_U \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i | \mathbf{X}_i, \alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}, \sigma_{\boldsymbol{\tau}})}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i | \mathbf{X}_i, \alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0}, \sigma_{\boldsymbol{\tau}0})} d\Pi_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})}{\int_\Theta \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i | \mathbf{X}_i \alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}, \sigma_{\boldsymbol{\tau}})}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i | \mathbf{X}_i, \alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0}, \sigma_{\boldsymbol{\tau}0})} d\Pi_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})}.$$

The main theorem of the paper can now be stated.

Theorem 1. *Suppose assumptions 1, 2, 3a, 4 and 6 hold or assumptions 1, 2, 3b, 4, 5 and 6. Let* $U = \{(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) : |\alpha_{\boldsymbol{\tau}} - \alpha_{\boldsymbol{\tau}0}| < \Delta, |\beta_{\boldsymbol{\tau}} - \beta_{\boldsymbol{\tau}0}| < \Delta \mathbf{1}_{k-1}\}$*. Then* $\lim_{n\to\infty} \Pi_{\boldsymbol{\tau}}(U^c|(\mathbf{Y}_1, \mathbf{X}_1), ..., (\mathbf{Y}_n, \mathbf{X}_n)) = 0$ *a.s.* $[\mathbf{P}_0]$.

The proof is presented in Appendix B. The strategy of the proof follows very closely to the strategy used in the single-output model (Sriram et al., 2013). First, construct an open set U_n containing $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$ for all n that converges to $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$, the population parameters. Define $B_n = \Pi_{\boldsymbol{\tau}}(U_n^c|(\mathbf{Y}_1, \mathbf{X}_1), ..., (\mathbf{Y}_n, \mathbf{X}_n))$. The Markov inequality and Borel-Cantelli lemma are used to prove convergence of B_n to $B = 0$ almost surely by showing that $\lim_{n\to\infty} \sum_{i=1}^n E[|B_n - B|^d] < \infty$ for some $d > 0$. The Markov inequality states if $B_n - B \geq 0$ then for any $d > 0$

$$Pr(|B_n - B| > \epsilon) \leq \frac{E[|B_n - B|^d]}{\epsilon^d}$$

for any $\epsilon > 0$. The Borel-Cantelli lemma states

$$\text{if } \lim_{n\to\infty} \sum_{i=1}^n Pr(|B_n - B| > \epsilon) < \infty \text{ then } Pr(\limsup_{n\to\infty} |B_n - B| > \epsilon) = 0.$$

Thus by Markov inequality

$$\sum_{i=1}^n Pr(|B_n - B| > \epsilon) \leq \sum_{i=1}^n \frac{E[|B_n - B|^d]}{\epsilon^d}.$$

Since $\lim_{n\to\infty} \sum_{i=1}^n E[|B_n - B|^d] < \infty$ then $\lim_{n\to\infty} \sum_{i=1}^n Pr(|B_n - B| > \epsilon) < \infty$. By Borel-Cantelli

$$Pr(\limsup_{n\to\infty} |B_n - B| > \epsilon) = 0.$$

To show $\lim_{n\to\infty} \sum_{i=1}^n E[|B_n - B|^d] < \infty$, a set G_n is created where $(\alpha_{\tau 0}, \beta_{\tau 0}) \notin G_n$. Within this set the expectation of the posterior numerator is less than $e^{-2n\delta}$ and the expectation of the posterior denominator is greater than $e^{-n\delta}$ for some $\delta > 0$. Thus the expected value of the posterior is less than $e^{-n\delta}$, which is summable.

2.1.1 Relation to τ-Tukey depth contours

Let $\boldsymbol{\mu}$ be the Tukey median of $\mathbf{Y}$, where the Tukey median is the point with maximal Tukey depth. See Appendix A for a discussion of Tukey depth and Tukey median. Define $\mathbf{Z} = \mathbf{Y} - \boldsymbol{\mu}$ to be the Tukey median centered transformation of $\mathbf{Y}$. Let $\alpha_{\boldsymbol{\tau}}$ and $\beta_{\boldsymbol{\tau}} = (\beta_{\boldsymbol{\tau}\mathbf{z}}, \beta_{\boldsymbol{\tau}\mathbf{x}})$ be the parameters of the $\lambda_{\boldsymbol{\tau}} = \{\mathbf{z} \in \Re^k : \mathbf{u}'\mathbf{z} = \beta'_{\boldsymbol{\tau}\mathbf{z}}\boldsymbol{\Gamma}'_{\mathbf{u}}\mathbf{z} + \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X} + \alpha_{\boldsymbol{\tau}}\}$ hyperplane for $\mathbf{Z}$. Under the condition that

$$\alpha_{\boldsymbol{\tau}} = \alpha_{\tau},\ \beta_{\boldsymbol{\tau}\mathbf{z}} = \mathbf{0}_{k-1} \text{ and } \beta_{\boldsymbol{\tau}\mathbf{x}} = \beta_{\tau\mathbf{x}} \text{ for all } \boldsymbol{\tau}, \tag{5}$$

$\mathbf{Y}$ has spherical Tukey contours with a Euclidean distance of $|\alpha_{\boldsymbol{\tau}} + \beta_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}|$ to the Tukey median. This result is obtained using Theorem 2 (presented below) and the fact that τ-quantile contours correspond to τ-Tukey depth contours (see equation (9) and the following text in Appendix A).

Theorem 2. *Suppose i) $\alpha_{\boldsymbol{\tau}} = \alpha_{\tau}$, $\beta_{\boldsymbol{\tau}\mathbf{z}} = \mathbf{0}_{k-1}$ and $\beta_{\boldsymbol{\tau}\mathbf{x}} = \beta_{\tau\mathbf{x}}$ for all $\boldsymbol{\tau}$ with τ fixed and ii) $\mathbf{Z}$ has spherical τ-Tukey depth contours (possibly traveling through $\mathbf{X}$) denoted by T_{τ} with Tukey median at $\mathbf{0}_k$. Then 1) the radius of the τ-Tukey depth contour is $d_{\tau} = |\alpha_{\tau} + \beta_{\tau\mathbf{x}}\mathbf{X}|$, 2) for any point $\tilde{\mathbf{Z}}$ on the τ-Tukey depth contour the hyperplane $\lambda_{\tilde{\boldsymbol{\tau}}}$ with $\tilde{\mathbf{u}} = \tilde{\mathbf{Z}}/\sqrt{\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}}$ and $\tilde{\boldsymbol{\tau}} = \tau\tilde{\mathbf{u}}$ is tangent to the contour at $\tilde{\mathbf{Z}}$ and 3) the hyperplane $\lambda_{\boldsymbol{\tau}}$ for any $\mathbf{u}$ is tangent to the τ-Tukey depth contour.*

The proof for Theorem 2 is presented in Appendix C. A corollary follows if $\beta_{\boldsymbol{\tau}\mathbf{x}} = \mathbf{0}_p$ or $\mathbf{X}$ has null dimension, then the radius of the spherical τ-Tukey depth contour is $|\alpha_{\tau}|$.

2.2 Confidence Intervals

Asymptotic frequentist confidence intervals for the location case can be obtained using Theorem 4 from Chernozhukov and Hong (2003) and asymptotic results from Hallin et al.

(2010). Frequentist confidence intervals based off a sandwich estimator are a common form of uncertainty quantification in the Bayesian single-output quantile model and for Bayesian misspecified models in general (Kleijn and van der Vaart, 2012; Müller, 2013; Yang et al., 2015; Sriram, 2015). In which case the sandwich estimator can be interpreted as an adjusted posterior estimate of frequentist uncertainty. An alternative approach not pursued in this paper is to use a score based approach (Wu and Narisetty, 2021).

Let $V_{\boldsymbol{\tau}} = V_{\boldsymbol{\tau}}^{mcmc} J'_{\mathbf{u}} V_{\boldsymbol{\tau}}^{c} J_{\mathbf{u}} V_{\boldsymbol{\tau}}^{mcmc}$ where $J_{\mathbf{u}}$ is a k by $k+1$ block diagonal matrix with blocks 1 and $\Gamma_{\mathbf{u}}$,

$$V_{\boldsymbol{\tau}}^{c} = \begin{bmatrix} \tau(1-\tau) & \tau(1-\tau)E[\mathbf{Y}'] \\ \tau(1-\tau)E[\mathbf{Y}] & Var[(\tau - 1_{(\mathbf{Y}\in H_{\boldsymbol{\tau}}^{-})})\mathbf{Y}] \end{bmatrix},$$

and $V_{\boldsymbol{\tau}}^{mcmc}$ be the covariance matrix of MCMC draws times n. The values of $E[\mathbf{Y}]$ and $Var[(\tau - 1_{(\mathbf{Y}\in H_{\boldsymbol{\tau}}^{-})})\mathbf{Y}]$ are estimated with standard moment estimators where the parameters of $H_{\boldsymbol{\tau}}^{-}$ are estimated with the Bayesian estimate plugged in. Then $\hat{\theta}_{\boldsymbol{\tau}i} \pm \Phi^{-1}(1-\alpha/2)\sqrt{V_{\boldsymbol{\tau}ii}/n}$ has a $1-\alpha$ coverage probability where Φ^{-1} is the inverse standard normal CDF. Section 4 verifies this in simulation. The sandwich estimator in one dimension is equivalent to estimators derived for the single-output location case (Sriram, 2015; Yang et al., 2015).

2.3 Choice of prior

A distinct set of population parameters are associated with each unique $\boldsymbol{\tau}$. Generally one is not interested in one unique $\boldsymbol{\tau}$ but some collection $\{\boldsymbol{\tau}_1, ..., \boldsymbol{\tau}_m\}$. This results in $m \times p \times k$ parameters to be estimated. Eliciting informative priors for such a large number of parameters can be a daunting task.

One can elicit prior beliefs for the parameters of each $\lambda_{\boldsymbol{\tau}}$ hyperplane individually. This approach is an onerous task and is thus discussed in Appendix D. However, the elicitation can be simplified if the researcher is simply interested in the τ-quantile (regression) contours.

If the prior is centered over (5) (e.g., $E[\alpha_{\boldsymbol{\tau}}] = \alpha_{\tau}$, $E[\beta_{\boldsymbol{\tau}\mathbf{z}}] = \mathbf{0}_{k-1}$ and $E[\beta_{\boldsymbol{\tau}\mathbf{x}}] = \beta_{\tau\mathbf{x}}$) then the implied ex-ante belief is $\mathbf{Z}$ has spherical Tukey contours with a Euclidean distance of $|\alpha_{\boldsymbol{\tau}} + \beta_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}|$ to the Tukey median. In which case there are only $1 + k$ parameters to be elicited for each τ-quantile (regression) contour. If there are no regressors then there is only one parameter per τ-quantile contour, $|\alpha_{\tau}|$, representing the distance of the spherical τ-Tukey depth contour from the Tukey median. Additionally, if $p = 2$ then for any $\mathbf{u} \in \mathcal{S}^{k-1}$ the population parameter $\alpha_{\boldsymbol{\tau}0}$ is negative when $\tau < 0.5$ and positive when $\tau > 0.5$.

Many basic statistical distributions have spherical Tukey contours such as the standard normal and t distributions. Thus one can elicit prior beliefs for τ-contour distances by adopting a prior belief that the data is distributed multivariate normal and standardizing the data to be mean zero with an identity covariance matrix. In which case the τ-contour distances are simply the quantiles of the univariate standard normal distribution (i.e., $E[\alpha_{\boldsymbol{\tau}}] = \Phi^{-1}(\tau)$). The prior variance then represents the strength of the researcher's belief. See Dutta et al. (2011) and Rousseeuw and Ruts (1999) for a more detailed discussion of distributions with and without spherical Tukey contours.

If one is willing to accept joint normality of $(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})$ then a Gibbs sampler can be used for estimation. The sampler is presented in Section 3. Further, if data is being collected and analyzed in real time, then the prior of the current analysis can be centered over the estimates from the previous analysis and the variance of the prior is how willing the researcher is to allow for departures from the previous estimates.

3 MCMC estimation

This section presents a Gibbs sampler to obtain draws from the posterior distribution of the parametric model. The MCMC sampler for the nonparametric model is presented in Appendix E.2.

Assuming joint normality of the prior distribution, estimation can be performed using Gibbs sampler draws from the posterior distribution developed in Kozumi and Kobayashi (2011). The approach assumes $\mathbf{Y}_{\mathbf{u}i} = \beta'_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y}^{\perp}_{\mathbf{u}i} + \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}_i + \alpha_{\boldsymbol{\tau}} + \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} ALD(0,1)$. The random component, ϵ_i, can be written as a mixture of a normal distribution and an exponential distribution, $\epsilon_i = \eta W_i + \gamma\sqrt{W_i}U_i$ where $\eta = \frac{1-2\tau}{\tau(1-\tau)}$, $\gamma = \sqrt{\frac{2}{\tau(1-\tau)}}$, $W_i \overset{iid}{\sim} exp(1)$ and $U_i \overset{iid}{\sim} N(0,1)$ are mutually independent (Kotz et al., 2001). This mixture representation allows for efficient simulation using data augmentation (Tanner and Wong, 1987). It follows that $\mathbf{Y}_{\mathbf{u}i}|\mathbf{Y}^{\perp}_{\mathbf{u}i}, \mathbf{X}_i, W_i, \beta_{\boldsymbol{\tau}}, \alpha_{\boldsymbol{\tau}}$ is normally distributed. Further, if the prior is $\theta_{\boldsymbol{\tau}} = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) \sim N(\mu_{\theta_{\boldsymbol{\tau}}}, \Sigma_{\theta_{\boldsymbol{\tau}}})$ then $\theta_{\boldsymbol{\tau}}|\mathbf{Y}_{\mathbf{u}}, \mathbf{Y}^{\perp}_{\mathbf{u}}, \mathbf{X}, W$ is normally distributed. Thus the $m+1$th MCMC draw is given by the following algorithm

1. Draw $W_i^{(m+1)} \sim W|\mathbf{Y}_{\mathbf{u}i}, \mathbf{Y}^{\perp}_{\mathbf{u}i}, \mathbf{X}_i, \theta_{\boldsymbol{\tau}}^{(m)} \sim GIG(\frac{1}{2}, \hat{\delta}_i, \hat{\phi})$ for $i \in \{1, ..., n\}$

2. Draw $\theta_{\boldsymbol{\tau}}^{(m+1)} \sim \theta_{\boldsymbol{\tau}}|\vec{\mathbf{Y}}_{\mathbf{u}}, \vec{\mathbf{Y}}^{\perp}_{\mathbf{u}}, \vec{\mathbf{X}}, \vec{W}^{(m+1)} \sim N(\hat{\theta}_{\boldsymbol{\tau}}, \hat{B}_{\boldsymbol{\tau}})$.

Where

$$\begin{aligned}
\hat{\delta}_i^2 &= \frac{1}{\gamma^2}(\mathbf{Y}_{\mathbf{u}i} - \beta'^{(m)}_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y}^{\perp}_{\mathbf{u}i} - \beta'^{(m)}_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}_i - \alpha_{\boldsymbol{\tau}}^{(m)})^2 \\
\hat{\phi}^2 &= 2 + \frac{\eta^2}{\gamma^2} \\
\hat{B}_{\boldsymbol{\tau}}^{-1} &= \Sigma_{\theta_{\boldsymbol{\tau}}}^{-1} + \sum_{i=1}^{n} \frac{[\mathbf{Y}^{\perp\prime}_{\mathbf{u}i}, \mathbf{X}'_i][\mathbf{Y}^{\perp\prime}_{\mathbf{u}i}, \mathbf{X}'_i]'}{\gamma^2 W_i^{(m+1)}} \\
\hat{\theta}_{\boldsymbol{\tau}} &= \hat{B}_{\boldsymbol{\tau}}\left(\Sigma_{\theta_{\boldsymbol{\tau}}}^{-1}\mu_{\theta_{\boldsymbol{\tau}}} + \sum_{i=1}^{n} \frac{[\mathbf{Y}^{\perp\prime}_{\mathbf{u}i}, \mathbf{X}'_i]'(\mathbf{Y}_{\mathbf{u}i} - \eta W_i^{(m+1)})}{\gamma^2 W_i^{(m+1)}}\right)
\end{aligned}$$

and $GIG(\nu, a, b)$ is the Generalized Inverse Gamma distribution whose density is

$$f(x|\nu, a, b) = \frac{(b/a)^\nu}{2K_\nu(ab)} x^{\nu-1} exp(-\frac{1}{2}(a^2x^{-1} + b^2x)), x > 0, -\infty < \nu < \infty, a, b \geq 0$$

and $K_\nu(\cdot)$ is the modified Bessel function of the third kind. An efficient sampler of the Generalized Inverse Gamma distribution was developed in Dagpunar (1989). Convergence speed can be improved by initializing the MCMC sequence at the frequentist estimate. The R package quantreg can provide such estimates (Koenker, 2018).

The Gibbs sampler is geometrically ergodic and thus the MCMC standard error is finite and the MCMC central limit theorem applies (Khare and Hobert, 2012). This guarantees that draws from this sampler are equivalent to random draws from the posterior after a long enough burn-in.

Numerous other algorithms can be used if the prior is not normally distributed. However, the researcher should take caution when using alternative MCMC algorithms because they could exhibit poor performance. Kozumi and Kobayashi (2011) provides a Gibbs sampler for when the prior is double exponential. Li et al. (2010) and Alhamzawi et al. (2012) provide algorithms for when regularization is desired. General purpose sampling schemes can also be used with arbitrary priors, such as the Metropolis-Hastings, slice sampling, or other algorithms (Hastings, 1970; Neal, 2003; Liu, 2008).

The Metropolis-Hastings algorithm can be implemented as follows. Define the likelihood to be $L_{\boldsymbol{\tau}}(\theta_{\boldsymbol{\tau}}) = \prod_{i=1}^n f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i, \alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}, 1)$. Let the prior for $\theta_{\boldsymbol{\tau}}$ have the density $\pi_{\boldsymbol{\tau}}(\theta_{\boldsymbol{\tau}})$. Define $g(\theta^\dagger|\theta)$ to be a proposal density. The $m+1$th MCMC draw is given by the following algorithm

1. Draw $\theta^\dagger_{\boldsymbol{\tau}}$ from $g(\theta^\dagger_{\boldsymbol{\tau}}|\theta^{(m)}_{\boldsymbol{\tau}})$

2. Compute $A(\theta^\dagger_{\boldsymbol{\tau}}, \theta^{(m)}_{\boldsymbol{\tau}}) = min\left(1, \frac{L(\theta^\dagger_{\boldsymbol{\tau}})\pi_{\boldsymbol{\tau}}(\theta^\dagger_{\boldsymbol{\tau}})g(\theta^{(m)}_{\boldsymbol{\tau}}|\theta^\dagger_{\boldsymbol{\tau}})}{L(\theta^{(m)}_{\boldsymbol{\tau}})\pi_{\boldsymbol{\tau}}(\theta^{(m)}_{\boldsymbol{\tau}})g(\theta^\dagger_{\boldsymbol{\tau}}|\theta^{(m)}_{\boldsymbol{\tau}})}\right)$

3. Draw u from $Uniform(0,1)$

4. If $u \leq A(\theta_{\boldsymbol{\tau}}^{\dagger}, \theta_{\boldsymbol{\tau}}^{(m)})$ set $\theta_{\boldsymbol{\tau}}^{(m+1)} = \theta_{\boldsymbol{\tau}}^{\dagger}$, else set $\theta_{\boldsymbol{\tau}}^{(m+1)} = \theta_{\boldsymbol{\tau}}^{(m)}$

Estimation of τ-quantile contours (see Appendix A) requires the simultaneous estimation of several different $\lambda_{\boldsymbol{\tau}}$. Simultaneous estimation of multiple $\lambda_{\boldsymbol{\tau}_m}$ ($m \in \{1, 2, ..., M\}$) can be performed by creating an aggregate likelihood. The aggregate likelihood is the product of the likelihoods for each m, $L_{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, ..., \boldsymbol{\tau}_M}(\alpha_{\boldsymbol{\tau}_1}, \beta_{\boldsymbol{\tau}_1}, \alpha_{\boldsymbol{\tau}_2}, \beta_{\boldsymbol{\tau}_2}, ..., \alpha_{\boldsymbol{\tau}_M}, \beta_{\boldsymbol{\tau}_M}) = \prod_{m=1}^{M} L_{\boldsymbol{\tau}_m}(\alpha_{\boldsymbol{\tau}_m}, \beta_{\boldsymbol{\tau}_m})$. The prior is then defined for the vector $(\alpha_{\boldsymbol{\tau}_1}, \beta_{\boldsymbol{\tau}_1}, \alpha_{\boldsymbol{\tau}_2}, \beta_{\boldsymbol{\tau}_2}, ..., \alpha_{\boldsymbol{\tau}_M}, \beta_{\boldsymbol{\tau}_M})$. The Gibbs algorithm can easily be modified for fixed τ to accommodate simultaneous estimation. To estimate the parameters from various τ, the values of η and γ need to be adjusted appropriately.

4 Simulation

This section verifies pointwise consistency of the parametric model by checking for estimator convergence to the population parameters. Asymptotic coverage probability using the results from Section 2.2 is also verified. Four DGPs are considered.

1. $\mathbf{Y} \sim Uniform\ Square$

2. $\mathbf{Y} \sim Uniform\ Triangle$

3. $\mathbf{Y} \sim N(\mu, \Sigma)$, where $\mu = \mathbf{0}_2$ and $\Sigma = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 9 \end{bmatrix}$

4. $\mathbf{Y} = \mathbf{Z} + \begin{bmatrix} 0 \\ X \end{bmatrix}$ where $\begin{bmatrix} X \\ \mathbf{Z} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_{\mathbf{Z}} \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{X\mathbf{Z}} \\ \Sigma_{X\mathbf{Z}}' & \Sigma_{\mathbf{ZZ}} \end{bmatrix} \right)$,

$$\Sigma_{XX} = 4,\ \Sigma_{X\mathbf{Z}} = \begin{bmatrix} 0 \\ 2 \end{bmatrix},\ \Sigma_{\mathbf{ZZ}} = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 9 \end{bmatrix},\ \mu_X = 0 \text{ and } \mu_{\mathbf{Z}} = \mathbf{0}_2$$

The first DGP has corners at $(-\frac{1}{2}, -\frac{1}{2}), (-\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, -\frac{1}{2}), (\frac{1}{2}, \frac{1}{2})$. The second DGP has corners at $(-\frac{1}{2}, -\frac{1}{2\sqrt{3}}), (\frac{1}{2}, -\frac{1}{2\sqrt{3}}), (0, \frac{1}{\sqrt{3}})$. DGPs 1,2, and 3 are location models. DGP 4 is a regression model. DGPs 1 and 2 satisfy Assumptions 1-6. DGPs 3 and 4 are cases when Assumption 4 is violated. In DGP 4, the unconditional distribution of $\mathbf{Y}$ is $\mathbf{Y} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1.5 \\ 1.5 & 17 \end{bmatrix}\right)$. Note the population parameters of the parametric model for DGP 4 might not correspond to τ-Tukey depth contours (Hallin et al., 2015; Koenker et al., 2018).

Two directions are considered, $\mathbf{u} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\mathbf{u} = (0, 1)$. The orthogonal directions are $\mathbf{\Gamma_u} = (1, 0)$ and $\mathbf{\Gamma_u} = (1/\sqrt{2}, -1/\sqrt{2})$. The first vector is a 45^o line between Y_2 and Y_1 in the positive quadrant and the second vector points vertically in the Y_2 direction. The depth is $\tau = 0.2$. The sample sample sizes are $n \in \{10^2, 10^3, 10^4\}$. The prior is $\theta_{\boldsymbol{\tau}} \sim N(\mu_{\theta_{\boldsymbol{\tau}}}, \Sigma_{\theta_{\boldsymbol{\tau}}})$ where $\mu_{\theta_{\boldsymbol{\tau}}} = \mathbf{0}_{k+p-1}$ and $\Sigma_{\theta_{\boldsymbol{\tau}}} = 1000\mathbf{I}_{k+p-1}$. The number of Monte Carlo simulations is 100 and for each Monte Carlo simulation 1,000 MCMC draws are used. The initial values are set to the frequentist estimate.

Checking for consistency or coverage probability requires knowledge of the population parameter. The population parameters are found by numerically minimizing the population objective function. The expectation in the objective function is calculated with a Monte Carlo simulation sample of 10^6.

4.1 Pointwise consistency

Consistency for the parametric model is verified by checking convergence of the Bayesian estimator to the population parameters, presented in Table 1.

		Data Generating Process			
$\mathbf{u}$	θ	1	2	3	4
$(1/\sqrt{2}, 1/\sqrt{2})$	$\alpha_{\boldsymbol{\tau}}$	-0.26	-0.20	-1.17	-1.16
	$\beta_{\boldsymbol{\tau}\mathbf{y}}$	0.00	0.44	-1.14	-1.17
	$\beta_{\boldsymbol{\tau}\mathbf{x}}$				-0.18
$(0, 1)$	$\alpha_{\boldsymbol{\tau}}$	-0.30	-0.20	-2.19	-2.02
	$\beta_{\boldsymbol{\tau}\mathbf{y}}$	0.00	0.00	1.50	1.50
	$\beta_{\boldsymbol{\tau}\mathbf{x}}$				1.50

Table 1: Parametric model population parameters

The Root Mean Square Error (RMSE) of the parameter estimates are presented in Tables 2, 3, and 4. The results show the Bayesian estimators are converging to the population parameters. Frequentist bias was also investigated and the bias showed convergence towards zero as sample size increased (no table presented).

		Data Generating Process			
θ	n	1	2	3	4
$\alpha_{\boldsymbol{\tau}}$	10^2	5.70e-02	4.41e-02	2.20e-01	1.83e-01
	10^3	1.49e-02	1.19e-02	6.80e-02	5.39e-02
	10^4	4.30e-03	3.66e-03	1.97e-02	1.85e-02
$\beta_{\boldsymbol{\tau}\mathbf{y}}$	10^2	9.63e-02	2.79e-01	9.61e-02	1.08e-01
	10^3	3.63e-02	6.58e-02	3.15e-02	3.15e-02
	10^4	1.19e-02	1.78e-02	1.07e-02	1.06e-02

Table 2: RMSE of parameter estimates ($\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$)

		Data Generating Process			
θ	n	1	2	3	4
$\alpha_{\boldsymbol{\tau}}$	10^2	3.57e-02	2.23e-02	3.47e-01	2.94e-01
	10^3	1.25e-02	5.59e-03	1.15e-01	1.13e-01
	10^4	4.23e-03	2.10e-03	3.27e-02	3.36e-02
$\beta_{\boldsymbol{\tau}\mathbf{y}}$	10^2	1.16e-01	7.03e-02	3.94e-01	2.78e-01
	10^3	3.96e-02	1.61e-02	1.18e-01	1.17e-01
	10^4	1.37e-02	6.73e-03	4.20e-02	3.13e-02

Table 3: RMSE of parameter estimates ($\mathbf{u} = (0, 1)$)

4.2 Coverage probability

Coverage probabilities for the parametric location model using the procedure in Section 2.2 are presented in Table 5. A correct coverage probability is 0.95. The number of Monte Carlo simulations is 300. The results show that the coverage probability tends to improve with sample size but has a slight undercoverage with sample size of 10^5. A naive interval constructed from the 0.025 and 0.975 quantiles of the MCMC draws produces coverage probabilities ranging from 0.980 to 1.000, with a majority at 1 (no table presented). This is clearly a strong overcoverage and thus the proposed procedure is preferred.

5 Application

The models are applied to educational data collected from the Project STAR public access database. Project STAR was an experiment conducted on 11,600 students in 300 classrooms from 1985-1989 to determine if reduced classroom size improved academic per-

		Data Generating Process
$\mathbf{u}$	n	4
	10^2	1.58e-01
$(1/\sqrt{2}, 1/\sqrt{2})$	10^3	4.86e-02
	10^4	1.48e-02
	10^2	1.49e-01
$(0, 1)$	10^3	5.82e-02
	10^4	1.83e-02

Table 4: RMSE of $\beta_{\boldsymbol{\tau}\mathbf{x}}$ estimates

formance.[4] Students and teachers were randomly selected in kindergarten to be in small (13-17 students) or large (22-26 students) classrooms. The students then stayed in their assigned classroom size through the fourth grade. The outcome measures were Stanford Achievement Test (SAT) scores for mathematics and reading tests observed each year.

This dataset has been analyzed many times before (see Finn and Achilles (1990); Folger and Breda (1989); Krueger (1999); Mosteller (1995); Word et al. (1990)). The previous analyses investigated either single-output test score measures or a single-output function (e.g., average) of mathematics and reading scores. Single-output analysis ignores important information about the relationship the mathematics and reading test scores might have with each other. Analysis on the average of scores better accommodates joint effects but obscures the source of effected subpopulations. Multiple-output quantiles provide information on the joint relationship between scores for the entire multivariate distribution (or several specified quantile subpopulations).

The treatment effect of classroom size is determined by inspecting the location τ-

[4] The data is publicly available at http://fmwww.bc.edu/ec-p/data/stockwatson.

		Data Generating Process					
		1	2	3	1	2	3
θ	n	$\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$			$\mathbf{u} = (0, 1)$		
$\alpha_{\boldsymbol{\tau}}$	10^2	.960	.950	.967	1.00	1.00	.937
	10^3	.963	.950	.940	.960	.963	.953
	10^4	.910	.947	.967	.930	.963	.957
$\beta_{\boldsymbol{\tau}\mathbf{y}}$	10^2	.810	1.00	.953	1.00	.987	.937
	10^3	.907	.967	.960	.978	.970	.933
	10^4	.933	.953	.937	.927	.960	.950

Table 5: Location model coverage probabilities

quantile contour estimates from the parametric model for small and large classrooms. The impact of teacher experience on child outcomes is also investigated by regressing test scores on teacher experience. The data is subset to first grade students (sample size of $n = 4,247$, after removal of missing data). The results for other grades were similar. Appendix F presents a fixed-$\mathbf{u}$ analysis and a sensitivity analysis.

Define the vector $\mathbf{u} = (u_1, u_2)$, where u_1 is the mathematics score dimension and u_2 is the reading score dimension. The $\mathbf{u}$ directions have an interpretation of relating how much relative importance the researcher wants to give to mathematics or reading. Define $\mathbf{u}^\perp = (u_1^\perp, u_2^\perp)$, where $\mathbf{u}^\perp$ is orthogonal to $\mathbf{u}$. The components $(u_1^\perp, u_2^\perp)$ have no meaningful interpretation. Define $mathematics_i$ to be the mathematics score of student i and $reading_i$ to be the reading score of student i.

5.1 Classroom size results

The (location) parametric model is

$$\begin{aligned}
\mathbf{Y}_{\mathbf{u}i} &= mathematics_i u_1 + reading_i u_2 \\
\mathbf{Y}^{\perp}_{\mathbf{u}i} &= mathematics_i u^{\perp}_1 + reading_i u^{\perp}_2 \\
\mathbf{Y}_{\mathbf{u}i} &= \alpha_{\boldsymbol{\tau}} + \beta_{\boldsymbol{\tau}} \mathbf{Y}^{\perp}_{\mathbf{u}i} + \epsilon_i \qquad (6) \\
\epsilon_i &\overset{iid}{\sim} ALD(0, 1, \tau) \\
\theta_{\boldsymbol{\tau}} = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) &\sim N(\mu_{\theta_{\boldsymbol{\tau}}}, \Sigma_{\theta_{\boldsymbol{\tau}}}).
\end{aligned}$$

Unless otherwise noted, $\mu_{\theta_{\boldsymbol{\tau}}} = \mathbf{0}_2$ and $\Sigma_{\theta_{\boldsymbol{\tau}}} = 1000\mathbf{I}_2$. This is interpreted as a weak ex-ante belief that the joint distribution of mathematics and reading has spherical τ-Tukey depth contours. The number of MCMC draws is 3,000 with a burn in of 1,000. The Gibbs algorithm is initialized at the frequentist estimate.

Figure 1 shows the τ-quantile contours for $\tau = 0.05$, 0.20 and 0.40. The data is stratified into smaller classrooms (blue) and larger classrooms (black) and separate models are estimated for each. The innermost contour is the $\tau = 0.40$ region, the middle contour is the $\tau = 0.20$ region, and the outermost contour is the $\tau = 0.05$ region. The τ-quantile contours for larger τ will always be contained in regions of smaller τ (if no numerical error and priors are not contradictory). All the points that lie on the contour have an estimated Tukey depth of τ. The contours for smaller τ capture the effects for the more outlying students (e.g., students who perform exceptionally well on mathematics or reading). The contours for larger τ capture the effects for the more central student (e.g., students who do not stand out from their peers).

The results show that all the τ-quantile contours shift up and to the right for the smaller classroom compared to the larger classroom. Thus a smaller classroom results in better test scores than a larger classroom for all inspected quantile subpopulations. Further, since the effect is present for a wide range of τ, it suggests that the entire distribution of scores might perform better in a smaller classroom.

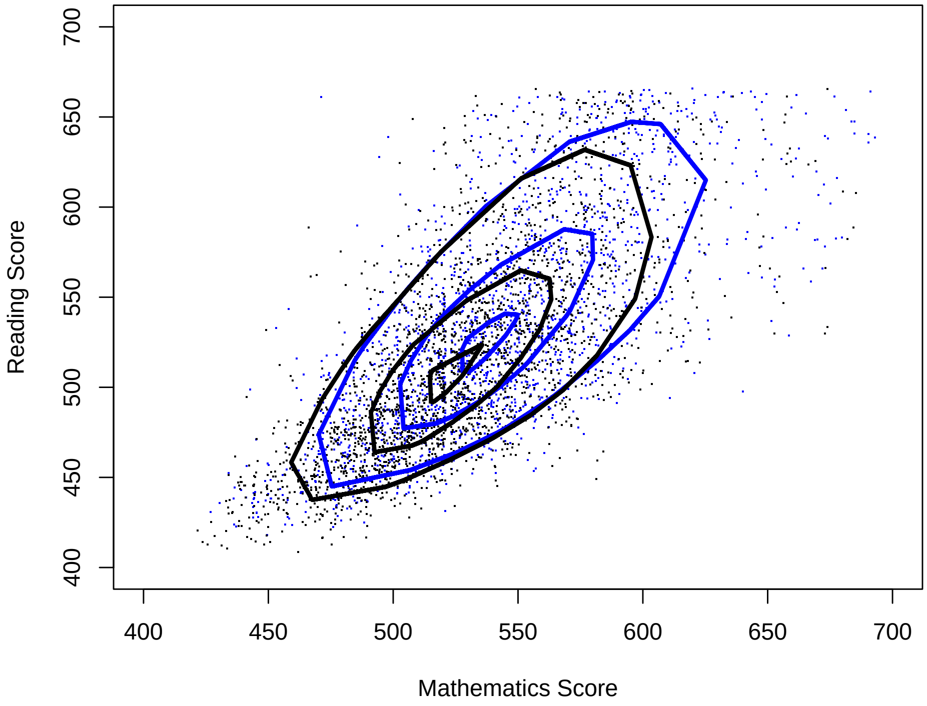

Figure 1: τ-quantile contours. Blue represents small and black represents large classrooms.

5.2 Teacher experience results

Teacher experience can be treated as exogenous due to random assignment of teachers which allows the impact of teacher experience on student outcomes to be estimated. Regression with continuous covariates, such as teacher experience, causes the τ-quantile contours to become tubes that travel through the covariate space. See Appendix F for a detailed specification of the models in this section.

Teacher experience can be modeled either parametrically or nonparametrically. As previously mentioned simply including experience as a linear term in the regression matrix ($\mathbf{X}$) of the parametric model will likely not lead to valid τ-quantile contours of the conditional distribution. However, including polynomial terms can improve the parametric model's approximation of the τ-quantile contours for the conditional distribution. Since both the data set and parameter space sizes are modest and only three slices of the regression tube are being inspected, the nonparametric model can suitably be used as well due to computational feasibility.

The Bayesian parametric linear and quadratic approaches are compared with the Bayesian nonparametric local constant and bilinear approaches in Figure 2 for $\tau = 0.05$ and Figure 3 for $\tau = 0.20$. The parametric linear model fails to uncover a strong impact or any non-linearities. However, the parametric quadratic and the nonparametric local constant models find that the marginal effect of teacher experience tends to be much larger for teachers that are at the beginning of their career than mid-career or late-career teachers, aligning with the results from other research (Rice, 2010). The nonparametric bilinear model is not the appropriate model for inspecting and comparing slices of the regression tube, but is still included for comparison.

Inspecting the τ-quantile regression contours more closely shows a heterogeneous treatment effect with respect to teacher experience. The reading and mathematics scores jointly increase with teacher experience for most student $\boldsymbol{\tau}$-quantile subpopulations except for the outlying students that perform poorly in both mathematics and reading (bottom left corner of $\tau = 0.05$). Those students showed no change with increasing teacher experience.

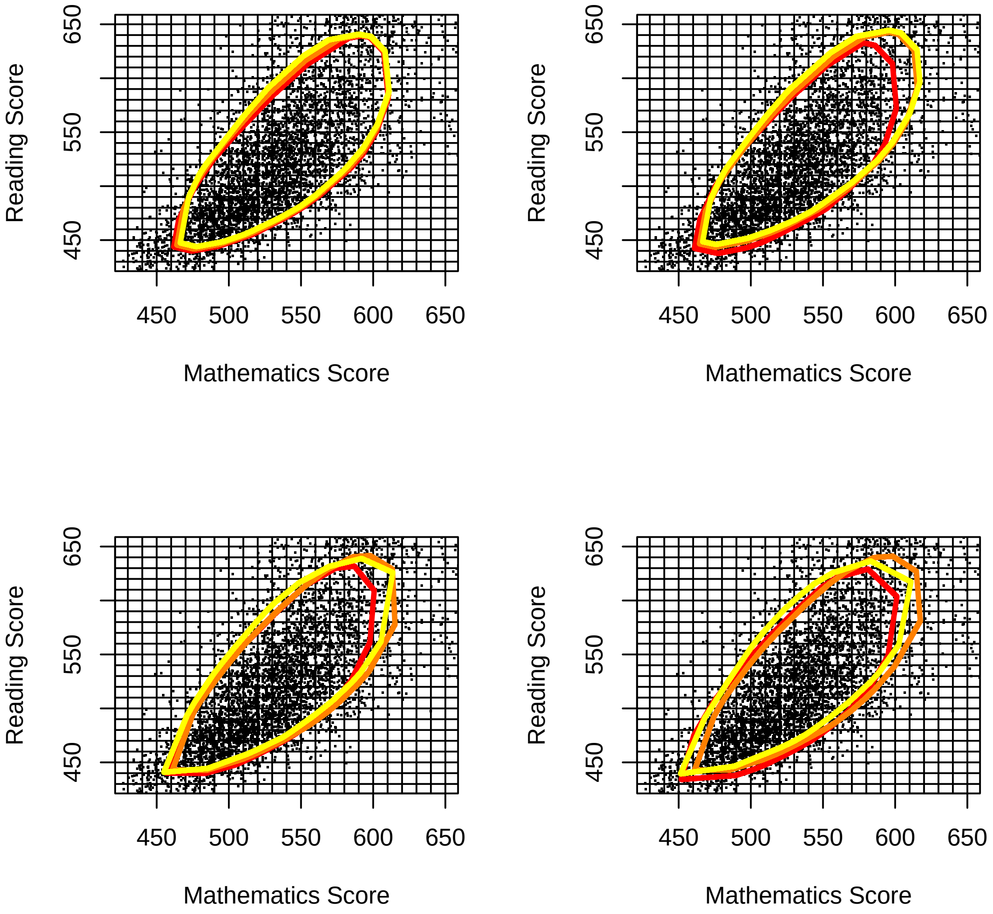

Figure 2: Regression tube slices for $\tau = 0.05$. Top left is parametric with linear experience. Top right is parametric with quadratic polynomial experience. Bottom left is nonparametric local constant. Bottom right is nonparametric local bilinear. Red, orange and yellow lines are one, ten, and twenty years of teacher experience respectively.

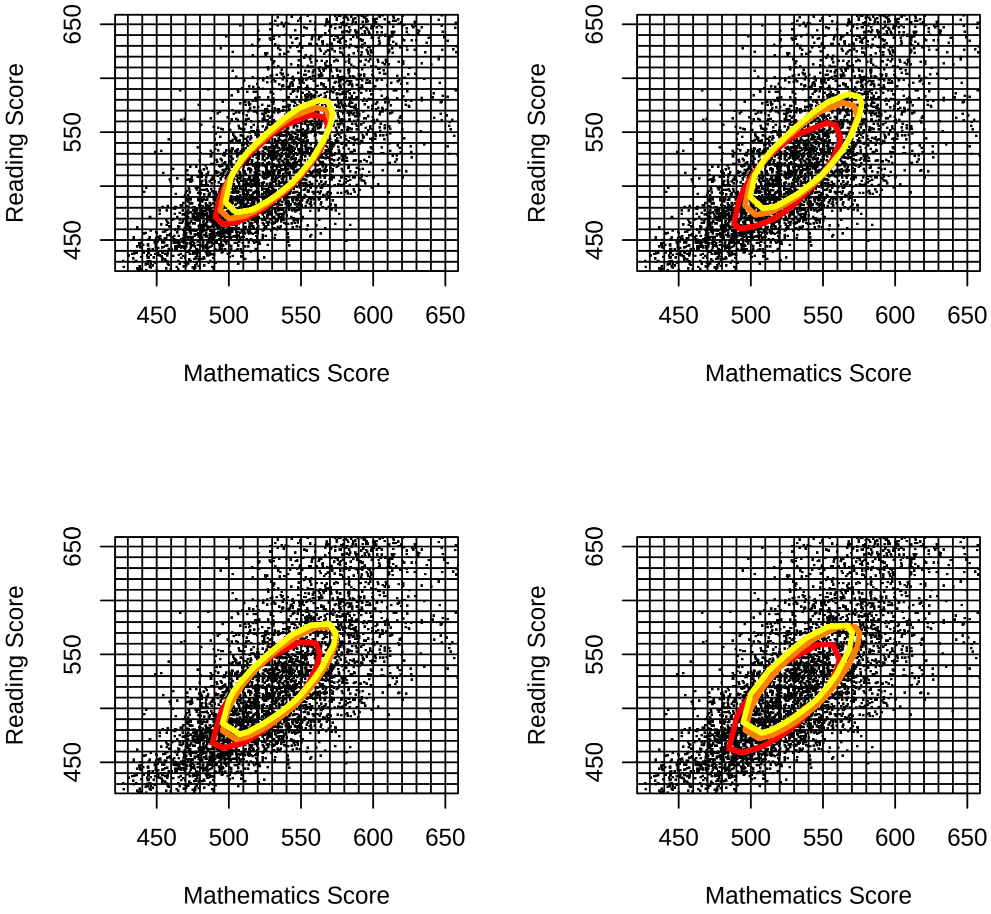

Figure 3: Regression tube slices for $\tau = 0.20$. Top left is parametric with linear experience. Top right is parametric with quadratic polynomial experience. Bottom left is nonparametric local constant. Bottom right is nonparametric local bilinear. Red, orange and yellow lines are one, ten, and twenty years of teacher experience respectively.

The Bayesian parametric quadratic polynomial and nonparametric local constant models are compared to their frequentist counterparts in Figure 4. Overall the Bayesian and frequentist estimates are similar (Hallin et al., 2010, 2015).

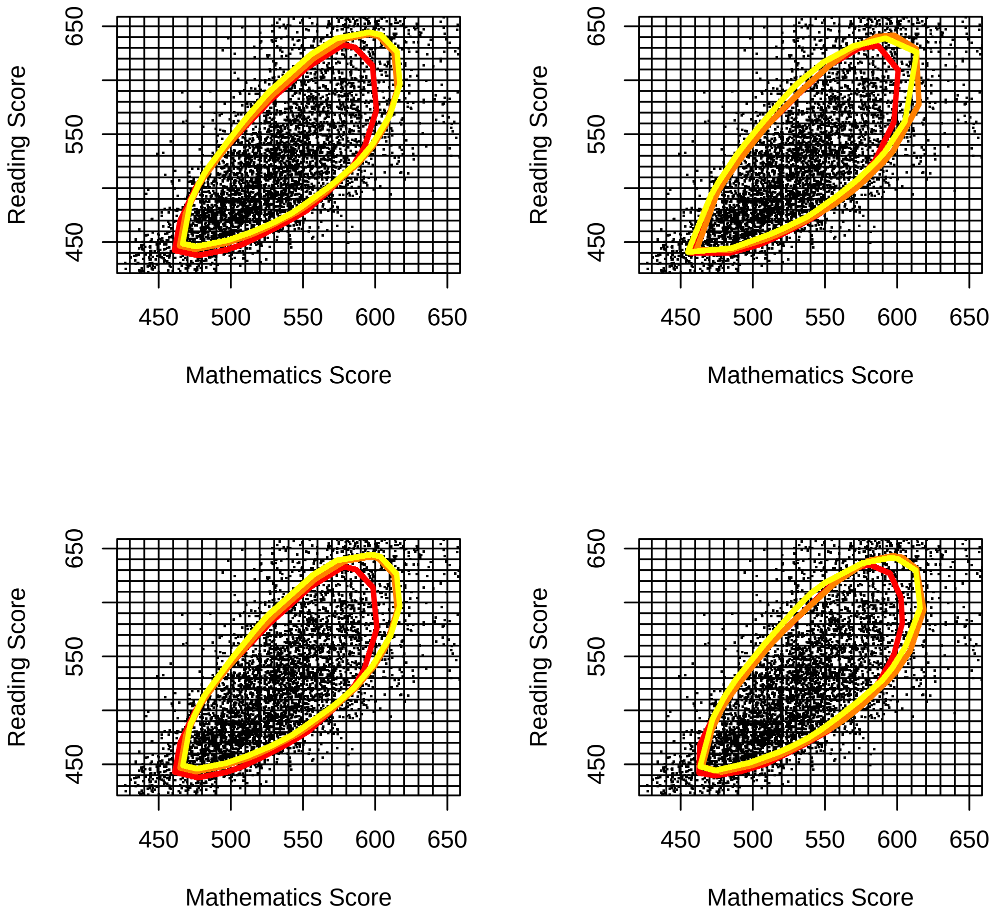

Figure 4: Regression tube slices for $\tau = 0.05$. Top plots are Bayesian and bottom plots are frequentist. The left plots are parametric with quadratic polynomial and right plots are nonparametric local constant. Red, orange and yellow lines are one, ten, and twenty years of teacher experience respectively.

6 Conclusion

A Bayesian framework for estimation of multiple-output quantiles was presented. Despite having a misspecified likelihood, the resulting posterior of the parametric model is consistent for the parameters of interest. The population parameters and prior are closely related to the τ-Tukey contours, the first prior of its kind. By performing inferences as a Bayesian, one inherits many of the strengths of a Bayesian approach. The models are applied to the Tennessee Project STAR experiment and it concludes that students in a smaller classroom perform better for every inspected quantile subpopulation than students in a larger classroom.

References

Alhamzawi, R., K. Yu, and D. F. Benoit (2012). Bayesian adaptive Lasso quantile regression. *Statistical Modelling 12*(3), 279–297.

Benoit, D. F. and D. Van den Poel (2012). Binary quantile regression: a Bayesian approach based on the asymmetric Laplace distribution. *Journal of Applied Econometrics 27*(7), 1174–1188.

Benoit, D. F. and D. Van den Poel (2017). bayesQR: A Bayesian approach to quantile regression. *Journal of Statistical Software 76*(7), 1–32.

Bhattacharya, I. and S. Ghosal (2020). Bayesian multivariate quantile regression using dependent Dirichlet process prior.

Carlier, G., V. Chernozhukov, and A. Galichon (2016). Vector quantile regression: An optimal transport approach. *Ann. Statist. 44*(3), 1165–1192.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association 91*(434), 862–872.

Chernozhukov, V. and H. Hong (2003). An MCMC approach to classical estimation. *Journal of Econometrics 115*(2), 293–346.

Dagpunar, J. (1989). An easily implemented generalised inverse Gaussian generator. *Communications in Statistics - Simulation and Computation 18*(2), 703–710.

Drovandi, C. C. and A. N. Pettitt (2011). Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis 55*(9), 2541–2556.

Dutta, S., A. K. Ghosh, P. Chaudhuri, et al. (2011). Some intriguing properties of Tukeys half-space depth. *Bernoulli 17*(4), 1420–1434.

Feng, Y., Y. Chen, and X. He (2015). Bayesian quantile regression with approximate likelihood. *Bernoulli 21*(2), 832–850.

Finn, J. D. and C. M. Achilles (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal 27*(3), 557–577.

Folger, J. and C. Breda (1989). Evidence from project star about class size and student achievement. *Peabody Journal of Education 67*(1), 17–33.

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

Hallin, M., D. Paindaveine, and M. Šiman (2010). Multivariate quantiles and multiple-output regression quantiles: from L1 optimization to halfspace depth. *The Annals of Statistics 38*(2), 635–703.

Hastings, W. K. (1970, 04). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Khare, K. and J. P. Hobert (2012). Geometric ergodicity of the Gibbs sampler for Bayesian quantile regression. *Journal of Multivariate Analysis 112*, 108 – 116.

Kleijn, B. and A. van der Vaart (2012). The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Statist. 6*, 354–381.

Koenker, R. (2018). *quantreg: Quantile Regression*. Comprehensive R Archive Network. R package version 5.38.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society 38*(1), 33–50.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2018). *Handbook of Quantile Regression*. Chapman & Hall/CRC. CRC Press, Taylor & Francis Group.

Kong, L. and I. Mizera (2012). Quantile tomography: using quantiles with multivariate data. *Statistica Sinica 22*(4), 1589–1610.

Kottas, A. and M. Krnjajić (2009). Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics 36*(2), 297–319.

Kotz, S., T. Kozubowski, and K. Podgorski (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Progress in Mathematics Series. Birkhäuser Boston.

Kozumi, H. and G. Kobayashi (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation 81*(11), 1565–1578.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics 114*(2), 497–532.

Laine, B. (2001). Depth contours as multivariate quantiles: A directional approach. Master's thesis, Univ. Libre de Bruxelles, Brussels.

Lancaster, T. and S. Jae Jun (2010). Bayesian quantile regression methods. *Journal of Applied Econometrics 25*(2), 287–307.

Li, Q., R. Xi, N. Lin, et al. (2010). Bayesian regularized quantile regression. *Bayesian Analysis 5*(3), 533–556.

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing.* Springer Science & Business Media.

McKeague, I. W., S. Lŏpez-Pintado, M. Hallin, and M. Šiman (2011). Analyzing growth trajectories. *Journal of Developmental Origins of Health and Disease 2*(6), 322329.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The future of children 5*(2), 113–127.

Müller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica 81*(5), 1805–1849.

Neal, R. M. (2003, 06). Slice sampling. *Ann. Statist. 31*(3), 705–767.

Rahman, M. A. (2016). Bayesian quantile regression for ordinal models. *Bayesian Analysis 11*(1), 1–24.

Rice, J. K. (2010). The impact of teacher experience: Examining the evidence and policy implications. brief no. 11. Technical report, National Center for Analysis of Longitudinal Data in Education Research.

Rousseeuw, P. J. and I. Ruts (1999). The depth function of a population distribution. *Metrika 49*(3), 213–244.

Santos, B. and T. Kneib (2020). Noncrossing structured additive multiple-output Bayesian quantile regression models. *Statistics and Computing 30*(4), 855–869.

Serfling, R. (2002). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica 56*(2), 214–232.

Serfling, R. and Y. Zuo (2010, 04). Discussion. *Ann. Statist. 38*(2), 676–684.

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique 58*(3), 263–277.

Sriram, K. (2015). A sandwich likelihood correction for Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Statistics & Probability Letters 107*, 18 – 26.

Sriram, K., R. Ramamoorthi, P. Ghosh, et al. (2013). Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Analysis 8*(2), 479–504.

Sriram, K., R. V. Ramamoorthi, and P. Ghosh (2016). On Bayesian quantile regression using a pseudo-joint asymmetric Laplace likelihood. *Sankhya A 78*(1), 87–104.

Taddy, M. A. and A. Kottas (2010). A Bayesian nonparametric approach to inference for quantile regression. *Journal of Business & Economic Statistics 28*(3), 357–369.

Tanner, M. and W. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*(398), 528–540.

Thompson, P., Y. Cai, R. Moyeed, D. Reeve, and J. Stander (2010). Bayesian nonparametric quantile regression using splines. *Computational Statistics & Data Analysis 54*(4), 1138–1150.

Tukey, J. W. (1975). Mathematics and the picturing of data.

Waldmann, E. and T. Kneib (2014). Bayesian bivariate quantile regression. *Statistical Modelling 15*(4), 326–344.

Wei, Y. (2008). An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts. *Journal of the American Statistical Association 103*(481), 397–409.

Word, E., J. Johnston, H. P. Bain, B. D. Fulton, J. B. Zaharias, C. M. Achilles, M. N. Lintz, J. Folger, and C. Breda (1990). The state of Tennessee's student/teacher achievement ratio (star) project: Technical report 1985 – 1990. Technical report, Tennessee State Department of Education.

Wu, T. and N. N. Narisetty (2021). Bayesian multiple quantile regression for linear models using a score likelihood. *Bayesian Analysis 1*(1), 1–29.

Yang, Y., H. J. Wang, and X. He (2015). Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *International Statistical Review 84*(3), 327–344. 10.1111/insr.12114.

Yu, K., Z. Lu, and J. Stander (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician) 52*(3), 331–350.

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