INSTITUTE FOR DEFENSE ANALYSES

A Bayesian Approach to Multiple-Output Quantile Regression

Michael Guggisberg

August 2019
Approved for public release;
distribution is unlimited.
IDA Paper NS P-10832
Log: H 19-000443

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication
The work was conducted by the Institute for Defense Analyses (IDA) under CRP C6523.

For More Information:
Dr. Julie Pechacek, Project Leader
jpechace@ida.org, 703-578-2858

ADM John C. Harvey, Jr., USN (Ret), Director, SFRD
jharvey@ida.org, 703-575-4530

INSTITUTE FOR DEFENSE ANALYSES

IDA Paper NS P-10832

A Bayesian Approach to Multiple-Output Quantile Regression

Michael Guggisberg

This page is intentionally blank.

A Bayesian Approach to Multiple-Output Quantile Regression

Michael Guggisberg*
Institute for Defense Analyses

August 13, 2019

Abstract

This paper presents a Bayesian approach to multiple-output quantile regression. The unconditional model is proven to be consistent and asymptotically correct frequentist confidence intervals can be obtained. The prior for the unconditional model can be elicited as the ex-ante knowledge of the distance of the τ-Tukey depth contour to the Tukey median, the first prior of its kind. A proposal for conditional regression is also presented. The model is applied to the Tennessee Project Steps to Achieving Resilience (STAR) experiment and it finds a joint increase in *τ-quantile subpopulations* for mathematics and reading scores given a decrease in the number of students per teacher. This result is consistent with, and much stronger than, the result one would find with multiple-output linear regression. Multiple-output linear regression finds the *average* mathematics and reading scores increase given a decrease in the number of students per teacher. However, there could still be subpopulations where the score declines. The multiple-output quantile regression approach confirms there are no quantile subpopulations (of the inspected subpopulations) where the score declines. This is truly a statement of 'no child left behind' opposed to 'no average child left behind.'

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, and Brian Bucks 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 quantile regression into a Bayesian framework. This advance opened the doors for Bayesian inference and generated a series of applied and methodological research.[1]

Multiple-output (i.e., multivariate) medians have been developing slowly since the early 1900s (Small, 1990). A multiple-output quantile can be defined in many different ways and there has been little consensus on which is the most appropriate (Serfling, 2002). The literature for Bayesian multiple-output quantiles is sparse; only two papers exist and neither use a commonly accepted definition for a multiple-output quantile (Drovandi and Pettitt, 2011; Waldmann and Kneib, 2014).[2]

This paper presents a Bayesian framework for multiple-output quantiles defined in Hallin et al. (2010). Their 'directional' quantiles have theoretic and computational properties not enjoyed by many other definitions. See McKeague et al. (2011) and Zscheischler (2014) for frequentist applications of multiple-output quantiles. The population parameters of these quantiles are unconditional on covariates since the population object function averages over the covariate space. The resulting $\lambda_{\boldsymbol{\tau}}$ hyperplanes are still a function of covariates (if included). This paper also presents a Bayesian framework for conditional multiple-output quantiles defined in Hallin et al. (2015). In the conditional model, both

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

[2]Drovandi and Pettitt (2011) uses a copula approach and Waldmann and Kneib (2014) uses a multiple-output asymmetric Laplace likelihood approach.

the population objective function and the $\lambda_{\boldsymbol{\tau}}$ hyperplanes are functions of covariates. These approaches use an idea similar to Chernozhukov and Hong (2003), which uses a likelihood that is not necessarily representative of the Data Generating Process (DGP). However, the resulting posterior converges almost surely to the true value.[3] 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 (i.e. online learning), 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.

The prior is a required component in Bayesian analysis where the researcher elicits their pre-analysis beliefs for the population parameters. The unconditional model prior is closely related to the Tukey depth of a distribution (Tukey, 1975). Tukey depth is a notion of multiple-output centrality of a data point. This is the first Bayesian prior for Tukey depth. The prior can be elicited as the Euclidean distance of the Tukey median from a (spherical) τ-Tukey depth contour. Once a prior is chosen, estimates can be computed using MCMC draws from the posterior. If the researcher is willing to accept prior joint normality of the model parameters, then a Gibbs MCMC sampler can be used. Gibbs samplers have many computational advantages over other MCMC algorithms, such as easy implementation,

[3]This is proven for the unconditional model and checked via simulation for the conditional model. Posterior convergence means that, as sample size increases, the probability mass for the posterior is concentrated in smaller neighborhoods around the true value, eventually converging to a point mass at the true value.

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 small 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.[4] The effect of classroom size on test scores is shown, comparing τ-quantile contours for mathematics and reading test scores for first grade students in small and large classrooms. The model finds that τ-quantile subpopulations of mathematics and reading scores improve for both central and extreme 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.[5] The multiple-output quantile regression approach confirms there are no quantile subpopulations where the score declines (of the inspected subpopulations). This is truly a statement of 'no child left behind', as opposed to 'no average child left behind.'

[4]Students were randomly selected to be in a small or large classroom for four years in their early elementary education. Every year the students were given standardized mathematics and reading tests.

[5]A plausible narrative is a poor performing student in a larger classroom might have more free time due to the teacher being busy with preparing, organizing, and grading. During this free time, the student might read more than they would have in a small classroom and might perform better on the reading test than they would have otherwise.

2 Bayesian multiple-output quantile regression

This section presents the unconditional and conditional Bayesian approaches to quantile regression. Notation common to both approaches is first presented, followed by the unconditional model and a theorem of consistency for the Bayesian estimator (section 2.1). Then a method to construct asymptotic confidence intervals is shown (section 2.2). The prior for the unconditional model is then discussed (section 2.3). Last, a proposal for conditional regression is presented (section 2.4). Expectations and probabilities in sections 2.1, 2.2, and 2.4 are conditional on parameters. Expectations in section 2.3 are with respect to prior parameters. Appendix A reviews frequentist single and multiple-output quantiles and Bayesian single-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 $\mathbf{Y_u} = \mathbf{u}'\mathbf{Y}$ and $\mathbf{Y_u^\perp} = \boldsymbol{\Gamma}_\mathbf{u}'\mathbf{Y}$. 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 Unconditional Quantiles

Define $\Psi^u(a, \mathbf{b}) = E[\rho_\tau(\mathbf{Y_u} - \mathbf{b_y}'\mathbf{Y_u^\perp} - \mathbf{b_x}'\mathbf{X} - a)]$ to be the objective function of interest. The $\boldsymbol{\tau}$th unconditional 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

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

The definition of the location case is embedded in definition (1), where $\mathbf{b}_{\mathbf{x}}$ and $\mathbf{X}$ are of null dimension. 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 unique 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.[6]

The population parameters satisfy two subgradient conditions

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

and

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

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

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

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

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.

[6]However, 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.

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. Appendix A provides more background on multiple-output quantiles defined in Hallin et al. (2010).

The Bayesian approach assumes

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

whose density is

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

The nuisance scale parameter, $\sigma_{\boldsymbol{\tau}}$, is fixed at 1.[7] 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 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* $i \in \{1, 2, ..., n, ...\}$.

[7] 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 will not attempt to prove consistency with a free nuisance scale parameter. Future research could follow the outline proposed in the single-output model and extend it to the multiple-output model (Sriram et al., 2013).

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.[8]

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* $i \in \{1, 2, ..., n, ...\}$.

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). Assumption 3 is satisfied using the joint normal prior suggested in section 2.3.

The next assumption bounds the covariates and response variables.

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* $\sup_{i,j} |[\boldsymbol{\Gamma}_{\mathbf{u}}]_{i,j}| < c_\Gamma$.

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. However, like $\mathbf{X}$, any

[8] This assumption can be weakened (Serfling and Zuo, 2010).

given dataset will satisfy this restriction. The assumption on $\mathbf{\Gamma_u}$ is innocuous; since $\mathbf{\Gamma_u}$ is chosen by the researcher, it is easy to choose such that all components are finite.

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 $B_n = \Pi_{\boldsymbol{\tau}}(U_n^c|(\mathbf{Y}_1,\mathbf{X}_1),...,(\mathbf{Y}_n,\mathbf{X}_n))$. To show convergence of B_n to

$B = 0$ almost surely, it is sufficient to show $\lim_{n\to\infty} \sum_{i=1}^{n} E[|B_n - B|^d] < \infty$ for some $d > 0$, using the Markov inequality and Borel-Cantelli lemma. 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 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 $\delta > 0$. Then the expected value of the posterior is less than $e^{-n\delta}$, which is summable.

2.2 Confidence Intervals

Asymptotic confidence intervals for the unconditional location case can be obtained using Theorem 4 from Chernozhukov and Hong (2003) and asymptotic results from Hallin et al. (2010).[9] 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

[9] A rigorous treatment would require verification of the assumptions of Theorem 4 from Chernozhukov and Hong (2003). Yang et al. (2015) and Sriram (2015) provide asymptotic standard errors for the single-output model.

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 $V_{\boldsymbol{\tau}}^{mcmc}$ is 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.

2.3 Choice of prior

A new model is estimated for each unique $\boldsymbol{\tau}$ and thus a prior is needed for each one. It might seem like there is an overwhelming amount of ex-ante elicitation required if one wants to estimate many models. For example, to estimate τ-quantile (regression) contours (see Appendix A).[10] However, simplifications can be made to make elicitation easier.

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 $\beta_{\boldsymbol{\tau}}$ be the parameters of the $\lambda_{\boldsymbol{\tau}}$ hyperplane for $\mathbf{Z}$. If the prior is centered over $H_0 : \alpha_{\boldsymbol{\tau}} = \alpha_\tau,\ \beta_{\boldsymbol{\tau}\mathbf{z}} = \mathbf{0}_{k-1}$ and $\beta_{\boldsymbol{\tau}\mathbf{x}} = \beta_{\tau\mathbf{x}}$ for all $\boldsymbol{\tau}$ (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 $\mathbf{Y}$ has spherical Tukey contours.[11] Under the belief H_0, $|\alpha_{\boldsymbol{\tau}} + \beta_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}|$

[10] Section 3.1 discusses how to estimate many models simultaneously.

[11] The null hypothesis $H_0 : \alpha_{\boldsymbol{\tau}} = \alpha_\tau,\ \beta_{\boldsymbol{\tau}\mathbf{z}} = \mathbf{0}_{k-1}$ and $\beta_{\boldsymbol{\tau}\mathbf{x}} = \beta_{\tau\mathbf{x}}$ for all $\boldsymbol{\tau}$ is a sufficient condition for spherical Tukey depth contours. It may or may not be necessary.

A sufficient condition for a density to have spherical Tukey depth contours is for the PDF to have spherical density contours and that the PDF, with a multivariate argument $\mathbf{Y}$, can be written as a monotonically decreasing function of $\mathbf{Y}'\mathbf{Y}$ (Dutta et al., 2011). This condition is satisfied for the location family for the

is the Euclidean distance of the τ-Tukey depth contour from the Tukey median. Since the contours are spherical, the distance is the same for all $\mathbf{u}$. This result is obtained using Theorem 2 (presented below) and the fact that the boundary of the intersection of upper quantile halfspaces corresponds to τ-Tukey depth contours (see equation (23) and the following text in Appendix A). The proof for Theorem 2 is presented in Appendix C. A notable corollary is 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 $|\alpha_\tau|$. Note if $\mathbf{X}$ has null dimension, $p = 2$, and $\mathbf{Z}$ has a zero vector Tukey median then for any $\mathbf{u} \in \mathcal{S}^{k-1}$ the population $\alpha_{\boldsymbol{\tau}0}$ is negative for $\tau < 0.5$ and the population $\alpha_{\boldsymbol{\tau}0}$ is positive for $\tau > 0.5$.

A prior for $(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})$ centered over H_0 expresses the researcher's confidence in the hypothesis of spherical Tukey depth contours. A large prior variance allows for large departures from H_0. If $\mathbf{X}$ is of null dimension, then the prior variance of α_τ represents the uncertainty of the distance of the τ-Tukey depth contour from the Tukey median. Further, if the parameter space for $\alpha_{\boldsymbol{\tau}}$ is restricted to $\alpha_{\boldsymbol{\tau}} = \alpha_\tau$ for fixed τ, then the prior variance of α_τ represents the uncertainty of the distance of the spherical τ-Tukey depth contour from the Tukey median.

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 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*

standard multivariate Normal, T and Cauchy. The distance of the Tukey median from the τ-Tukey depth contour for the multivariate standard normal is $\Phi^{-1}(1-\tau)$. Another distribution with spherical Tukey contours is the uniform hyperball. The distance of the Tukey median from the τ-Tukey depth contour for the uniform hyperball is the value r such that $arcsin(r) + r\sqrt{1-r^2} = \pi(0.5-\tau)$. This function is invertible for $r \in (0,1)$ and $\tau \in (0,.5)$ and can be computed using numerical approximations (Rousseeuw and Ruts, 1999).

$\tilde{\boldsymbol{\tau}} = \tau\tilde{\mathbf{u}}$ *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.*

Arbitrary priors not centered over 0 require a more detailed discussion. Consider the two-dimensional case ($k = 2$). There are two ways to think of appropriate priors for $(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})$. The first approach is a direct approach thinking of $(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})$ as the intercept and slope of $\mathbf{Y_u}$ against $\mathbf{Y_u^\perp}$ and $\mathbf{X}$.[12] The second approach is thinking of the implied prior of $\phi_{\boldsymbol{\tau}} = \phi_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})$ as the intercept and slope of Y_2 against Y_1 and $\mathbf{X}$. The second approach is presented in Appendix D.

In the direct approach, the parameters relate directly to the subgradient conditions (2) and (3) and their effect in $\mathbf{Y}$ space. A δ unit increase in $\alpha_{\boldsymbol{\tau}}$ results in a parallel shift in the hyperplane $\lambda_{\boldsymbol{\tau}}$ by $\frac{\delta}{u_2 - \beta_{\boldsymbol{\tau}\mathbf{y}} u_2^\perp}$ units. A δ unit increase in $\beta_{\boldsymbol{\tau}\mathbf{x}l}$ results in a parallel shift in the hyperplane $\lambda_{\boldsymbol{\tau}}$ by $\frac{\delta \mathbf{X}_l}{u_2 - \beta_{\boldsymbol{\tau}\mathbf{y}} u_2^\perp}$ units. When $\beta_{\boldsymbol{\tau}} = \mathbf{0}_{2+p-1}$ $\lambda_{\boldsymbol{\tau}}$ is orthogonal to $\mathbf{u}$ (and thus $\lambda_{\boldsymbol{\tau}}$ is parallel to $\Gamma_{\mathbf{u}}$). As $\beta_{\boldsymbol{\tau}\mathbf{y}}$ increases or decreases monotonically such that $|\beta_{\boldsymbol{\tau}\mathbf{y}}| \to \infty$, $\lambda_{\boldsymbol{\tau}}$ converges to $\mathbf{u}$ monotonically.[13] A δ unit increase in $\beta_{\boldsymbol{\tau}\mathbf{y}}$ tilts the $\lambda_{\boldsymbol{\tau}}$ hyperplane.[14] The direction of the tilt is determined by the vectors $\mathbf{u}$ and $\Gamma_{\mathbf{u}}$ and the sign of δ. The vectors $\mathbf{u}$ and $\Gamma_{\mathbf{u}}$ always form a 90° and 270° angle. For positive δ, the hyperplane travels monotonically through the triangle formed by $\mathbf{u}$ and $\Gamma_{\mathbf{u}}$. For negative δ, the hyperplane travels monotonically in the opposite direction.

[12]The value of $\mathbf{Y_u}$ is the scalar projection of $\mathbf{Y}$ in direction $\mathbf{u}$ and $\mathbf{Y_u^\perp}$ is the scalar projection of $\mathbf{Y}$ in the direction of the other (orthogonal) basis vectors.

[13]Monotonic meaning either the outer or inner angular distance between $\lambda_{\boldsymbol{\tau}}$ and $\mathbf{u}$ is always decreasing for strictly increasing or decreasing $\beta_{\boldsymbol{\tau}\mathbf{y}}$.

[14]Define *slope*(δ) to be the slope of the hyperplane when β is increased by δ. The slope of the new hyperplane is $slope(\delta) = (u_2 - (\beta + \delta)u_2^\perp)^{-1}(\delta u_1^\perp + (u_2 - \beta u_2^\perp)slope(0)$

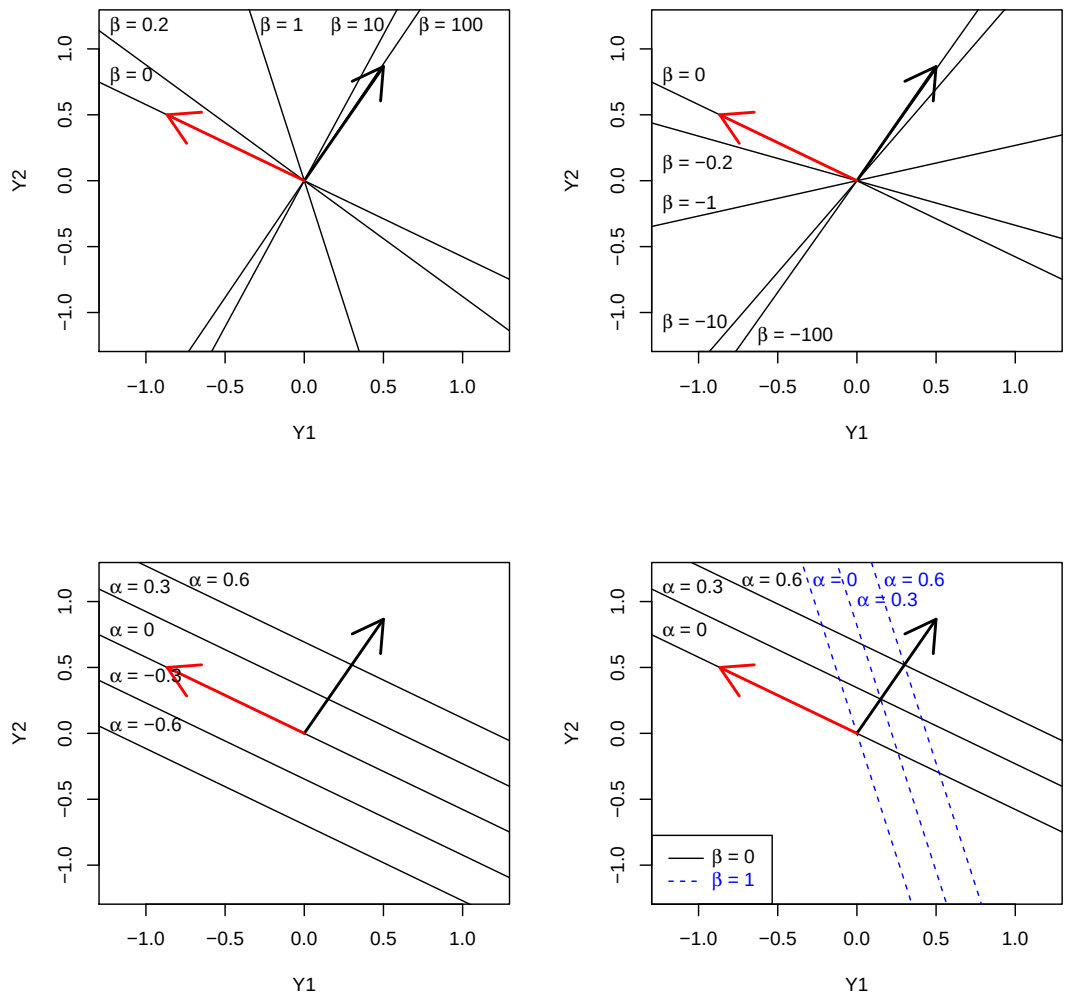

Figure 1: Implied $\lambda_{\boldsymbol{\tau}}$ from various hyperparameters ($\boldsymbol{\tau}$ subscript omitted). Top left, positive increasing β. Top right, negative decreasing β. Bottom left, different αs. Bottom right, different αs and βs.

Figure 1 shows prior $\lambda_{\boldsymbol{\tau}}$ implied from the center of the prior with various hyperparam-

eters. For all four plots $k = 2$, the directional vector is $\mathbf{u} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ (black arrow) and $\Gamma_{\mathbf{u}} = (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ (red arrow). The top left plot shows $\lambda_{\boldsymbol{\tau}}$ for $\beta_{\boldsymbol{\tau}}$ increasing from 0 to 100 for fixed $\alpha_{\boldsymbol{\tau}} = 0$. At $\beta_{\boldsymbol{\tau}} = 0$ the hyperplane is perpendicular to $\mathbf{u}$, as $\beta_{\boldsymbol{\tau}}$ increases $\lambda_{\boldsymbol{\tau}}$ travels counterclockwise until it becomes parallel to $\mathbf{u}$. The top right plot shows the $\lambda_{\boldsymbol{\tau}}$ for $\beta_{\boldsymbol{\tau}}$ decreasing from 0 to -100 for fixed $\alpha_{\boldsymbol{\tau}} = 0$. At $\beta_{\boldsymbol{\tau}} = 0$ $\lambda_{\boldsymbol{\tau}}$ is perpendicular to $\mathbf{u}$, as $\beta_{\boldsymbol{\tau}}$ decreases $\lambda_{\boldsymbol{\tau}}$ travels clockwise until it becomes parallel to $\mathbf{u}$. The bottom left plot shows $\lambda_{\boldsymbol{\tau}}$ with $\alpha_{\boldsymbol{\tau}}$ ranging from -0.6 to 0.6. If the Tukey median is the point $(0, 0)$, then $|\alpha_{\boldsymbol{\tau}}|$ is the distance of the intersection of $\mathbf{u}$ and $\lambda_{\boldsymbol{\tau}}$ from the Tukey median.[15] For positive $\alpha_{\boldsymbol{\tau}}$, $\lambda_{\boldsymbol{\tau}}$ is moving in the direction $\mathbf{u}$ and for negative $\alpha_{\boldsymbol{\tau}}$, $\lambda_{\boldsymbol{\tau}}$ is moving in the direction $-\mathbf{u}$. The bottom right plot shows $\lambda_{\boldsymbol{\tau}}$ for various $\alpha_{\boldsymbol{\tau}}$ and $\beta_{\boldsymbol{\tau}}$. The solid black $\lambda_{\boldsymbol{\tau}}$ are for $\beta_{\boldsymbol{\tau}} = 0$ and the dashed blue $\lambda_{\boldsymbol{\tau}}$ are for $\beta_{\boldsymbol{\tau}} = 1$ and $\alpha_{\boldsymbol{\tau}}$ takes on values 0, 0.3 and 0.6 for both values of $\beta_{\boldsymbol{\tau}}$. This plot confirms changes in $\alpha_{\boldsymbol{\tau}}$ result in parallel shifts of $\lambda_{\boldsymbol{\tau}}$ while $\beta_{\boldsymbol{\tau}}$ tilts $\lambda_{\boldsymbol{\tau}}$.

If one is willing to accept joint normality of $(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})$ then a Gibbs sampler can be used. The sampler is presented in Section 3.1. 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 analysis.

2.4 Conditional Quantiles

Quantile regression defined so far is unconditional on covariates. Thus, the quantiles are averaged over the covariate space. A conditional quantile provides local quantile estimates conditional on covariates. A Bayesian multiple-output conditional quantile can be defined

[15] The Tukey median does not exist in these plots since there is no data. If there was data, the point where $\mathbf{u}$ and $\Gamma_{\mathbf{u}}$ intersect would be the Tukey median.

from Hallin et al. (2015).[16] A possible Bayesian approach is outlined but no proof of consistency is provided. Consistency is checked via simulation in Section 4. The $\lambda_{\boldsymbol{\tau}}$ hyperplanes are separately estimated for each conditioning value, thus the approach can be computationally expensive. Define $\mathcal{M}_{\mathbf{u}} = \{(a, \mathbf{d}) : a \in \Re,\ \mathbf{d} \in \Re^k \text{ subject to } \mathbf{d}'\mathbf{u} = 1\}$ the population parameters are

$$(\alpha_{\boldsymbol{\tau};\mathbf{x}_0}, \delta_{\boldsymbol{\tau};\mathbf{x}_0}) = \underset{(a,\mathbf{d})\in\mathcal{M}_{\mathbf{u}}}{argmin}\ E[\rho_\tau(\mathbf{d}'\mathbf{Y} - a)|\mathbf{X} = \mathbf{x}_0].$$

If $\mathbf{d} = \mathbf{u} - \mathbf{b}\Gamma'_{\mathbf{u}}$ the population objective function can be rewritten as $\Psi^c(a, \mathbf{b}) = E[\rho_\tau(\mathbf{Y}_{\mathbf{u}} - \mathbf{b}'\mathbf{Y}^{\perp}_{\mathbf{u}} - a)|\mathbf{X} = \mathbf{x}_0]$. The population parameters are

$$(\alpha_{\boldsymbol{\tau};\mathbf{x}_0}, \beta_{\boldsymbol{\tau};\mathbf{x}_0}) = \underset{(a,\mathbf{b})\in\Re^k}{argmin}\ \Psi^c(a, \mathbf{b}). \tag{5}$$

The subgradient conditions are

$$\left.\frac{\partial\Psi^c(a,\mathbf{b})}{\partial a}\right|_{\alpha_{\boldsymbol{\tau};\mathbf{x}_0},\beta_{\boldsymbol{\tau};\mathbf{x}_0}} = Pr(\mathbf{Y}_{\mathbf{u}} - \beta'_{\boldsymbol{\tau};\mathbf{x}_0}\mathbf{Y}^{\perp}_{\mathbf{u}} - \alpha_{\boldsymbol{\tau};\mathbf{x}_0} \leq 0|\mathbf{X} = \mathbf{x}_0) - \tau = 0 \tag{6}$$

and

$$\left.\frac{\partial\Psi^c(a,\mathbf{b})}{\partial \mathbf{b}}\right|_{\alpha_{\boldsymbol{\tau};\mathbf{x}_0},\beta_{\boldsymbol{\tau};\mathbf{x}_0}} = E[\mathbf{Y}^{\perp}_{\mathbf{u}}1_{(\mathbf{Y}_{\mathbf{u}}-\beta'_{\boldsymbol{\tau};\mathbf{x}_0}\mathbf{Y}^{\perp}_{\mathbf{u}}-\alpha_{\boldsymbol{\tau};\mathbf{x}_0}\leq 0)}|\mathbf{X} = \mathbf{x}_0] - \tau E[\mathbf{Y}^{\perp}_{\mathbf{u}}|\mathbf{X} = \mathbf{x}_0] = \mathbf{0}_{k-1}. \tag{7}$$

Assuming the distribution of $\mathbf{X}$ is continuous, then the conditioning set has probability 0. Hallin et al. (2015) creates an empirical (frequentist) estimator using weights, providing larger weight to observations near $\mathbf{x}_0$. The estimator is

[16] This section omits theoretical discussion of multiple-output conditional quantiles. See Hallin et al. (2015) for a rigorous exploration of the properties (including contours) for multiple-output conditional quantiles.

$$\hat{\theta}_{\boldsymbol{\tau};\mathbf{x}_0} = \underset{\mathbf{b}}{argmin} \sum_{i=1}^{n} K_h(\mathbf{X}_i - \mathbf{x}_0)\rho_\tau(\mathbf{Y}_{\mathbf{u}i} - b\mathcal{X}^r_{\mathbf{u}i}) \text{ for } r = c, l. \tag{8}$$

The function K_h is a kernel function whose corresponding distribution has zero first moment and positive definite second moment (e.g., uniform, Epanechnikov or Gaussian). The parameter h determines bandwidth.[17] If $r = c$ then $\mathcal{X}^c_{\mathbf{u}i} = [1, \mathbf{Y}^{\perp\prime}_{\mathbf{u}i}]'$ and the estimator is called a local constant estimator. If $r = l$ then $\mathcal{X}^l_{\mathbf{u}i} = [1, \mathbf{Y}^{\perp\prime}_{\mathbf{u}i}]' \otimes [1, (\mathbf{X}_i - \mathbf{x}_0)']'$ and the estimator is called a local bilinear estimator. The space that $\mathbf{b}$ is minimized over is the real numbers of dimension equal to the length of $\mathcal{X}^r_{\mathbf{u}i}$. For either value of r, the minimization can be expressed as maximization of an asymmetric Laplace likelihood with a known (heteroskedastic) scale parameter.

The Bayesian approach assumes

$$\mathbf{Y}_\mathbf{u}|\mathcal{X}^r, \theta_{\boldsymbol{\tau};\mathbf{x}_0} \sim ALD(\theta'_{\boldsymbol{\tau}}\mathcal{X}^r, K_h(\mathbf{X} - \mathbf{x}_0)^{-1}, \tau)$$

whose density is

$$\begin{aligned} f_{\boldsymbol{\tau}}(\mathbf{Y}|\mathcal{X}^r, \theta_{\boldsymbol{\tau};\mathbf{x}_0}, K_h(\mathbf{X} - \mathbf{x}_0)^{-1}) &= \tau(1-\tau)K_h(\mathbf{X} - \mathbf{x}_0)exp(-K_h(\mathbf{X} - \mathbf{x}_0)\rho_\tau(\mathbf{Y} - \theta'_{\boldsymbol{\tau};\mathbf{x}_0}\mathcal{X}^r)) \\ &\propto exp(-K_h(\mathbf{X} - \mathbf{x}_0)\rho_\tau(\mathbf{Y} - \theta'_{\boldsymbol{\tau};\mathbf{x}_0}\mathcal{X}^r)) \end{aligned}$$

If the researcher assumes the prior distribution for $\theta_{\boldsymbol{\tau}}$ is normal, then the parameters can be estimated with a Gibbs sampler, which is presented in Section 3.2.

3 MCMC simulation

In this section, a Gibbs sampler to obtain draws from the posterior distribution is presented for unconditional regression quantiles (Section 3.1) and conditional regression quantiles

[17] To guarantee consistency of the frequentist estimator h must satisfy $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh_n^{p-1} = \infty$. Hallin et al. (2015) provides guidance for choosing h.

(Section 3.2).

3.1 Unconditional Quantiles

Assuming joint normality of the prior distribution for the parameters estimation can be performed using draws from the posterior distribution obtained from a Gibbs sampler 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 and an exponential, $\epsilon_i = \eta W_i + \gamma\sqrt{W_i}U_i$ 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 $\mathbf{Y}_{\mathbf{u}i}|\mathbf{Y}^{\perp}_{\mathbf{u}i}, \mathbf{X}_i, W_i, \beta_{\boldsymbol{\tau}}, \alpha_{\boldsymbol{\tau}}$ 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 $W_i^{(m+1)} \sim W|\mathbf{Y}_{\mathbf{u}i}, \mathbf{Y}^{\perp}_{\mathbf{u}i}, \mathbf{X}_i, \theta^{(m)}_{\boldsymbol{\tau}} \sim GIG(\frac{1}{2}, \hat{\delta}_i, \hat{\phi}_i)$ for $i \in \{1, ..., n\}$
2. Draw $\theta^{(m+1)}_{\boldsymbol{\tau}} \sim \theta_{\boldsymbol{\tau}}|\vec{\mathbf{Y}}_{\mathbf{u}}, \vec{\mathbf{Y}}^{\perp}_{\mathbf{u}}, \vec{\mathbf{X}}, \vec{W}^{(m+1)} \sim N(\hat{\theta}_{\boldsymbol{\tau}}, \hat{B}_{\boldsymbol{\tau}})$.

where

$$\begin{aligned}
\hat{\delta}_i &= \frac{1}{\gamma^2}(\mathbf{Y}_{\mathbf{u}i} - \beta'^{(m)}_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y}^{\perp}_{\mathbf{u}i} - \beta'^{(m)}_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}_i - \alpha^{(m)}_{\boldsymbol{\tau}})^2 \\
\hat{\phi}_i &= 2 + \frac{\eta^2}{\gamma^2} \\
\hat{B}^{-1}_{\boldsymbol{\tau}} &= B^{-1}_{\boldsymbol{\tau}0} + \sum_{i=1}^{n} \frac{[\mathbf{Y}^{\perp\prime}_{\mathbf{u}i}, \mathbf{X}'_i][\mathbf{Y}^{\perp\prime}_{\mathbf{u}i}, \mathbf{X}'_i]'}{\gamma^2 W_i^{(m+1)}} \\
\hat{\beta}_{\boldsymbol{\tau}} &= \hat{B}_{\boldsymbol{\tau}}\left(B^{-1}_{\boldsymbol{\tau}0}\beta_{\boldsymbol{\tau}0} + \sum_{i=1}^{n} \frac{[\mathbf{Y}^{\perp\prime}_{\mathbf{u}i}, \mathbf{X}'_i]'(\mathbf{Y}_{\mathbf{u}i} - \eta W_i^{(m+1)})}{\gamma^2 W_i^{(m+1)}}\right)
\end{aligned}$$

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

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

and $K_{\nu}(\cdot)$ is the modified Bessel function of the third kind.[18] To speed convergence, the MCMC sequence can be initialized with the frequentist estimate.[19] 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 after a long enough burn-in, draws from this sampler are equivalent to random draws from the posterior.

Numerous other algorithms can be used if the prior is non-normal. 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, 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 $Uniform(0, 1)$

[18] An efficient sampler of the Generalized Inverse Gamma distribution was developed in Dagpunar (1989). Implementations of the Gibbs sampler with a free σ parameter for R are provided in the package 'bayesQR' and 'AdjBQR' (Benoit and Van den Poel, 2017; Wang and Yang, 2016). However, the results presented in this paper use a fixed $\sigma = 1$ parameter.

[19] The R package 'quantreg' can provide such estimates (Koenker, 2018).

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

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.

3.2 Conditional Quantiles

Sampling from the conditional quantile posterior is similar to that of unconditional quantiles except the likelihood is heteroskedastic with known heteroskedasticity. The approach assumes $\mathbf{Y}_{\mathbf{u}i} = \theta'_{\boldsymbol{\tau}} \mathcal{X}^r_i + K_h(\mathcal{X}^r_i - \mathbf{x}_0)^{-1} \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} ALD(0, 1)$. The random component, $K_h(\mathcal{X}^r_i - \mathbf{x}_0)^{-1} \epsilon_i$, can be written as a mixture of a normal and an exponential, $K_h(\mathcal{X}^r_i - \mathbf{x}_0)^{-1} \epsilon_i = \eta V_i + \gamma \sqrt{K_h(\mathcal{X}^r_i - \mathbf{x}_0)^{-1} V_i} U_i$ where $V_i = K_h(\mathcal{X}^r_i - \mathbf{x}_0)^{-1} W_i$. If the prior is $\theta_{\boldsymbol{\tau}} = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) \sim N(\mu_{\theta_{\boldsymbol{\tau}}}, \Sigma_{\theta_{\boldsymbol{\tau}}})$ then a Gibbs sampler can be used. The $m + 1$th MCMC draw is given by the following algorithm

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

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

where

$$\hat{\delta}_i = \frac{K_h(\mathcal{X}_i^r - \mathbf{x}_0)}{\gamma^2}(\mathbf{Y}_{\mathbf{u}i} - \theta'^{(m)}_{\boldsymbol{\tau}}\mathcal{X}_i^r)^2$$

$$\hat{\phi}_i = 2K_h(\mathcal{X}_i^r - \mathbf{x}_0) + \frac{\eta^2 K_h(\mathcal{X}_i^r - \mathbf{x}_0)}{\gamma^2}$$

$$\hat{B}_{\boldsymbol{\tau}}^{-1} = B_{\boldsymbol{\tau}0}^{-1} + \sum_{i=1}^{n} \frac{K_h(\mathcal{X}_i^r - \mathbf{x}_0)\mathcal{X}_i^r\mathcal{X}_i^{r\prime}}{\gamma^2 W_i^{(m+1)}}$$

$$\hat{\beta}_{\boldsymbol{\tau}} = \hat{B}_{\boldsymbol{\tau}}\left(B_{\boldsymbol{\tau}0}^{-1}\beta_{\boldsymbol{\tau}0} + \sum_{i=1}^{n} \frac{K_h(\mathcal{X}_i^r - \mathbf{x}_0)\mathcal{X}_i^r(\mathbf{Y}_{\mathbf{u}i} - \eta W_i^{(m+1)})}{\gamma^2 W_i^{(m+1)}}\right).$$

The MCMC sequence can be initialized with the frequentist estimate.[20] A Metropolis-Hastings algorithm similar to the unconditional model can be used, where $L(\theta_{\boldsymbol{\tau}}) = \prod_{i=1}^{n} f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathcal{X}_i^r, \theta_{\boldsymbol{\tau}}, K_h(\mathcal{X}_i^r - \mathbf{x}_0)^{-1})$. Simultaneous estimation of many $\lambda_{\boldsymbol{\tau}_m}$ is similar to the unconditional model.

4 Simulation

This section verifies pointwise consistency of the unconditional and conditional models. Asymptotic coverage probability of the unconditional location model using the results from Section 2.2 is also verified. Pointwise consistency is verified by checking convergence to solutions of the subgradient conditions (population parameters). 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 $\Sigma = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 9 \end{bmatrix}$

[20] The R package 'quantreg' can provide such estimates using the weights option.

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

$\Sigma_{XX} = 4$, $\Sigma_{X\mathbf{Z}} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$, $\Sigma_{\mathbf{ZZ}} = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 9 \end{bmatrix}$, $\mu_X = 0$ 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 and 4 is a regression model. DGPs 1 and 2 conform to all the assumptions on the data generating process. DGPs 3 and 4 are cases when Assumption 4 is violated. In DGP 4, the unconditional distribution of $\mathbf{Y}$ is $\mathbf{Y} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1.5 \\ 1.5 & 17 \end{bmatrix} \right)$.

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.

4.1 Unconditional model pointwise consistency

Consistency for the unconditional model is verified by checking convergence to the solutions of the subgradient conditions (population parameters). Convergence of subgradient conditions (2) and (3) is verified in Appendix E.

The population parameters for the four DGPs are presented in Table 1.[21] The Root

[21] The population parameters are found by numerically minimizing the objective function. The expecta-

Mean Square Error (RMSE) of the parameter estimates are presented in Tables 2, 3 and 4. The results show the Bayesian estimator is converging to the population parameters.[22]

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

Table 1: Unconditional model population parameters

4.2 Unconditional location model coverage probability

Coverage probabilities for the unconditional 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 under-coverage 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 over-coverage and thus the proposed procedure is preferred.

tion in the objective function is calculated with a Monte Carlo simulation sample of 10^6.

[22]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
	10^2	5.70e-02	4.41e-02	2.20e-01	1.83e-01
$\alpha_{\boldsymbol{\tau}}$	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
	10^2	9.63e-02	2.79e-01	9.61e-02	1.08e-01
$\beta_{\boldsymbol{\tau}\mathbf{y}}$	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: Unconditional model RMSE of parameter estimates ($\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$)

4.3 Conditional model pointwise consistency

Convergence of the local constant conditional model is verified by checking convergence of the Bayesian estimator to the parameters minimizing the population objective function (5). The local constant specification is presented because that is the specification used in the application. The conditional distribution of DGP 4 is $Y|X = x_0 \sim N\left(\begin{bmatrix} 0 \\ x_0/2 \end{bmatrix}, \begin{bmatrix} 1 & 1.5 \\ 1.5 & 8 \end{bmatrix}\right)$. Thus, the population objective function can be calculated using Monte Carlo integration or with quadrature methods. The population parameters with $x_0 = 1$ are $(\alpha_{\boldsymbol{\tau};1}, \beta_{\boldsymbol{\tau};1}) = (-1.23, 1.167)$ for $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ and $(\alpha_{\boldsymbol{\tau};1}, \beta_{\boldsymbol{\tau};1}) = (-1.53, 1.49)$ for $\mathbf{u} = (0, 1)$, which are found by numerically minimizing the Monte Carlo estimated population objective function.[23] The weight function is set to $K_h(\mathbf{X}_i - \mathbf{x}_0) = \frac{1}{\sqrt{2\pi h^2}} exp\left(-\frac{1}{2h^2}(\mathbf{X}_i - \mathbf{x}_0)^2\right)$ where $h = \sqrt{9\hat{\sigma}^2_{\mathbf{X}} n^{-1/5}}$.

Table 6 shows the RMSE of the Bayesian estimator for the conditional local constant

[23]The relative error difference between the Monte Carlo and quadrature methods was at most $5 \cdot 10^{-3}$. The Monte Carlo approach used a simulation sample of 10^6.

		Data Generating Process			
θ	n	1	2	3	4
	10^2	3.57e-02	2.23e-02	3.47e-01	2.94e-01
$\alpha_{\boldsymbol{\tau}}$	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
	10^2	1.16e-01	7.03e-02	3.94e-01	2.78e-01
$\beta_{\boldsymbol{\tau}\mathbf{y}}$	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: Unconditional model RMSE of parameter estimates ($\mathbf{u} = (0, 1)$)

model. The estimator appears to be converging to the population parameter.

5 Application

The unconditional and conditional 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 performance.[24] Students and teachers were randomly selected in kindergarten to be in small (13-17 students) or large (22-26 students) classrooms.[25] The students then stayed in their assigned classroom size through the fourth grade. The outcome of the treatment was scores of mathematics and reading tests given each year. This dataset has been analyzed many times before (see Finn and Achilles (1990); Folger and Breda (1989);

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

[25]This analysis omits large classrooms that had a teaching assistant.

		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: Unconditional model RMSE of $\beta_{\boldsymbol{\tau}\mathbf{x}}$ estimates

Krueger (1999); Mosteller (1995); Word et al. (1990)).[26] The studies performed analyses on either single-output test score measures or on a functional 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 quantile regression provides information on the joint relationship between scores for the entire multivariate distribution (or several specified quantile subpopulations).

A student's outcome was measured using a standardized test called the Stanford Achievement Test (SAT) for mathematics and reading.[27] Section 5.1 inspects the τ-quantile con-

[26]Folger and Breda (1989) and Finn and Achilles (1990) were the first two published studies. Word et al. (1990) was the official report from the Tennessee State Department of Education. Mosteller (1995) provided a review of the study and Krueger (1999) performed a rigorous econometric analysis focusing on validity.

[27]The test scores have a finite discrete support. Computationally, this does not effect the Bayesian estimates; however; it prevents asymptotically unique estimators. Thus, each score is perturbed with a uniform(0,1) random variable.

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

Table 5: Unconditional location model coverage probabilities

tours on the subset first grade students (sample size of $n = 4,247$, after removal of missing data). The results for other grades were similar.[28] The treatment effect of classroom size is determined by inspecting the location τ-quantile contours of the unconditional model for small and large classrooms. The treatment effect for teacher experience is determined by inspecting the τ-quantile contours from the conditional model (conditional on teacher experience) pooling small and large classrooms. Section 5.2 shows a sensitivity analysis of the unconditional model by inspecting the posterior τ-quantile contours with different prior specifications. Appendix F presents fixed-$\mathbf{u}$ analysis and an additional 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

[28]This application explains the concepts of Bayesian multiple-output quantile regression and does not provide rigorous causal econometric inferences. In the latter case, a thorough discussion of missing data would be necessary. For the same reason, first grade scores were chosen. The first grade subset was best suited for pedagogy.

		u	
θ	n	$(1/\sqrt{2}, 1/\sqrt{2})$	$(0, 1)$
	10^2	2.90e-01	6.78e-01
$\alpha_{\boldsymbol{\tau},1}$	10^3	7.10e-02	3.04e-01
	10^4	2.86e-02	1.33e-01
	10^2	8.79e-02	3.22e-01
$\beta_{\boldsymbol{\tau},1}$	10^3	3.35e-02	1.29e-01
	10^4	1.53e-02	5.04e-02

Table 6: Local constant conditional model RMSE of parameter estimates

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 τ-quantile (regression) contours

The unconditional model is

$$\begin{aligned} \mathbf{Y}_{\mathbf{u}i} &= mathematics_i u_1 + reading_i u_2 \\ \mathbf{Y}_{\mathbf{u}i}^{\perp} &= mathematics_i u_1^{\perp} + reading_i u_2^{\perp} \\ \mathbf{Y}_{\mathbf{u}i} &= \alpha_{\boldsymbol{\tau}} + \beta_{\boldsymbol{\tau}} \mathbf{Y}_{\mathbf{u}i}^{\perp} + \epsilon_i \\ \epsilon_i &\overset{iid}{\sim} ALD(0, 1, \tau) \\ \theta_{\boldsymbol{\tau}} = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) &\sim N(\mu_{\theta_{\boldsymbol{\tau}}}, \Sigma_{\theta_{\boldsymbol{\tau}}}). \end{aligned} \tag{9}$$

Unless otherwise noted, $\mu_{\theta_{\boldsymbol{\tau}}} = \mathbf{0}_2$ and $\Sigma_{\theta_{\boldsymbol{\tau}}} = 1000\mathbf{I}_2$, meaning ex-ante knowledge is a weak belief that the joint distribution of mathematics and reading has spherical 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 2 shows the τ-quantile contours for $\tau = 0.05$, 0.20 and 0.40. A τ-quantile contour is defined as the boundary of 23 in Appendix A. The data is stratified into smaller classrooms (blue) and larger classrooms (black) and separate models are estimated for each. The unconditional regression model was used but the effective results are conditional, since separate models are estimated by classroom size. 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. Contour regions 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 Bayesian Tukey depth of τ. The contours for larger τ capture the effects for the more extreme students (e.g., students who perform exceptionally well on mathematics and reading or exceptionally poorly on mathematics but well on reading). The contours for smaller τ capture the effects for the more central or more 'median' student (e.g., students who do not stand out from their peers). All the contours shift up and to the right for the smaller classroom. This shows the centrality of mathematics and reading scores improves for smaller classrooms compared to larger classrooms. This also shows all inspected quantile subpopulations of scores improve for students in smaller classrooms.

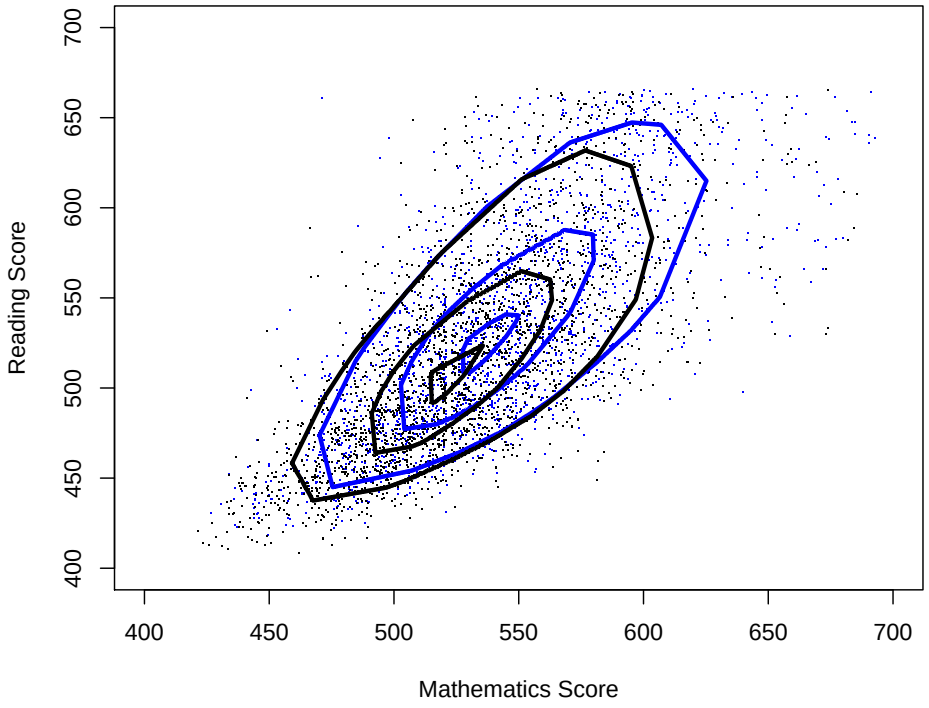

Figure 2: τ-quantile contours. Blue represents small and black represents large classrooms.

Up to this point, only quantile location models conditional on binary classroom size have been estimated.[29] When including continuous covariates, the τ-quantile regions become 'tubes' that travel through the covariate space. Due to random assignment of teachers, teacher experience can be treated as exogenous and the impact of experience on student outcomes can be estimated. Treating teacher experience as continuous, the appropriate model to use is the conditional regression model (8). The local constant specification is

[29]The unconditional model from (1) is used but is called conditional because separate models were estimated on small and large classrooms.

preferred if the researcher wishes to inspect slices of the regression tube. The local bilinear specification is preferred if the researcher wishes to connect the slices of the regression tube to create the regression tube. This analysis only looks at slices of the tube, thus the local constant specification is used.

The conditional model is

$$
\begin{aligned}
\mathbf{Y}_{\mathbf{u}i} &= mathematics_i u_1 + reading_i u_2 \\
\mathbf{Y}^{\perp}_{\mathbf{u}i} &= mathematics_i u^{\perp}_1 + reading_i u^{\perp}_2 \\
\mathbf{X}_i &= years\, of\, teacher\, experience_i \\
\mathcal{X}^{l}_{\mathbf{u}i} &= [1, \mathbf{X}_i - \mathbf{x}_0, \mathbf{Y}^{\perp}_{\mathbf{u}i}, (\mathbf{X}_i - \mathbf{x}_0)\mathbf{Y}^{\perp}_{\mathbf{u}i}]' \\
\hat{\sigma}^2_{\mathbf{X}} &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})^2 \\
h &= \sqrt{9\hat{\sigma}^2_{\mathbf{X}} n^{-1/5}} \\
K_h(\mathbf{X}_i - \mathbf{x}_0) &= \frac{1}{\sqrt{2\pi h^2}} exp\left(-\frac{1}{2h^2}(\mathbf{X}_i - \mathbf{x}_0)^2\right) \\
\mathbf{Y}_{\mathbf{u}i} &= \theta_{\boldsymbol{\tau};\mathbf{x}_0}\mathcal{X}_{\mathbf{u}i} + \epsilon_i \\
\epsilon_i &\overset{iid}{\sim} ALD(0, K_h(\mathbf{X}_i - \mathbf{x}_0)^{-1}, \tau) \\
\theta_{\boldsymbol{\tau};\mathbf{x}_0} &\sim N(\mu_{\theta_{\boldsymbol{\tau};\mathbf{x}_0}}, \Sigma_{\theta_{\boldsymbol{\tau};\mathbf{x}_0}}).
\end{aligned}
\tag{10}
$$

The prior hyperparameters are $\mu_{\theta_{\boldsymbol{\tau};\mathbf{x}_0}} = \mathbf{0}_4$ and $\Sigma_{\theta_{\boldsymbol{\tau};\mathbf{x}_0}} = 1000\mathbf{I}_4$. Small and large classrooms are pooled together. Figure 3 shows the τ-quantile regression regions with a covariate for experience. The values τ takes on are 0.20 (left plot) and 0.05 (right plot). The tubes are sliced at $\mathbf{x}_0 \in \{1, 10, 20\}$ years of teaching experience. The left plot shows reading scores increase with teacher experience for the more 'central' students but there does not seem to be a change in mathematics scores. The right plot shows a similar story for most of the 'extreme' students.

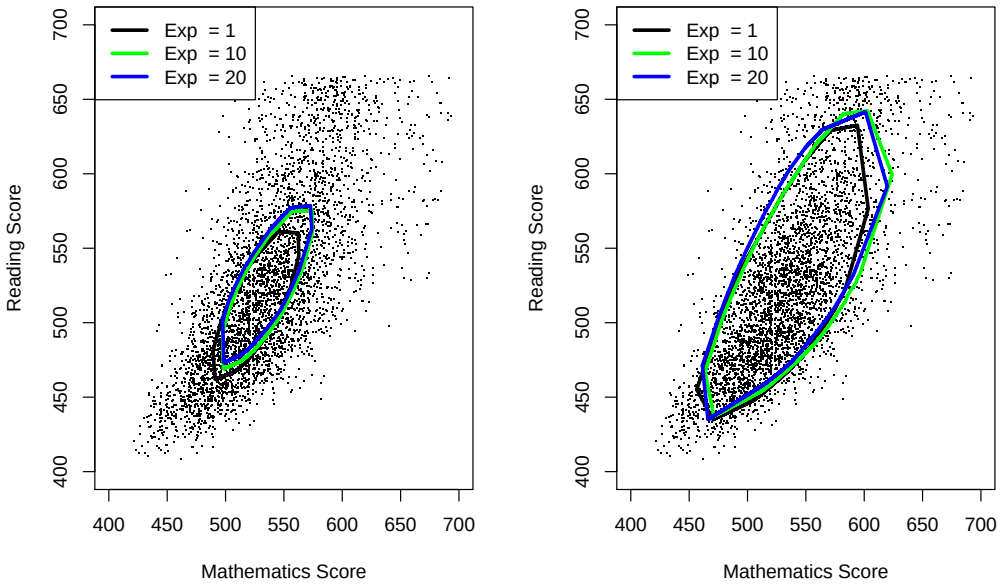

Figure 3: Regression tube slices. Left, $\tau = 0.2$ quantile regression tube. Right, $\tau = 0.05$ quantile regression tube.

A non-linear effect of experience is observed. It is clear there is a larger marginal impact on student outcomes going from 1 to 10 years of experience than from 10 to 20 years of experience. This marginal effect is more pronounced for the more central students ($\tau = 0.2$). Previous research has shown strong evidence that the effect of teacher experience on student achievement is non-linear. Specifically, the marginal effect of experience tends to be much larger for teachers that are at the beginning of their career than mid-career or late-career teachers (Rice, 2010). The more outlying students ($\tau = 0.05$) have a heterogeneous treatment effect with respect to experience. The best performing students in mathematics and reading show increased performance when experience increases from 1 to 10 years but

little change after that. All other outlying students are largely unaffected by experience.

5.2 Sensitivity analysis

Figure 4 shows posterior sensitivity of expected τ-quantile contours for weak and strong priors of spherical Tukey depth contours for $\tau \in \{0.05, 0.20, 0.40\}$.[30] The posterior from the weak prior is represented by the dashed red line and the posterior from the strong prior is represented by the dotted blue line. The posteriors are compared against the frequentist estimate represented by a solid black contour. The weak priors have covariance $\Sigma_{\theta_{\boldsymbol{\tau}}} = diag(1000, 1000)$ for all $\boldsymbol{\tau} \in \mathcal{B}^k$. For all plots, the posterior expected τ-contour with a weak prior is indistinguishable from the frequentist τ-contour. Appendix F presents a sensitivity analysis for a single $\lambda_{\boldsymbol{\tau}}$.

The top-left plot shows expected posterior τ-contours from a prior mean $\mu_{\theta_{\boldsymbol{\tau}}} = \mathbf{0}_2$ for all $\boldsymbol{\tau} \in \mathcal{B}^k$. The strong prior has covariance $\Sigma_{\theta_{\boldsymbol{\tau}}} = diag(1, 1000)$ for all $\boldsymbol{\tau} \in \mathcal{B}^k$. The strong prior represents a strong a priori belief that all τ-contours are near the Tukey median. The prior influence is strongest ex-post for $\tau = 0.05$. This is because the distance between the τ-contour and the Tukey median increases with decreasing τ. The strong prior in the top-right plot has covariance $\Sigma_{\theta_{\boldsymbol{\tau}}} = diag(1000, 0.0001)$ for all $\boldsymbol{\tau} \in \mathcal{B}^k$; all else is the same as the priors from the top-left plot. The top-right plot shows that a strong prior information on $\beta_{\boldsymbol{\tau}\mathbf{y}}$ provides little information ex-post in this setup.

The bottom-left plot shows expected posterior τ-contours from prior means $\mu_{\theta_{0.05\mathbf{u}}} = (-65, 0)$, $\mu_{\theta_{0.20\mathbf{u}}} = (-33, 0)$, and $\mu_{\theta_{0.40\mathbf{u}}} = (-10, 0)$ for all $\mathbf{u} \in \mathcal{S}^{k-1}$. The strong prior has covariance $\Sigma_{\theta_{\boldsymbol{\tau}}} = diag(1, 0.0001)$. This is a strong a priori belief that the τ-contours are spherical and the distance between the τ-contour and the Tukey median decreases with increasing τ. This plot shows the reduction of ellipticity of the posterior τ-contours. The

[30] To better show the effect of the prior, the dataset is reduced to the small classroom subset.

effect is strongest for $\tau = 0.05$.

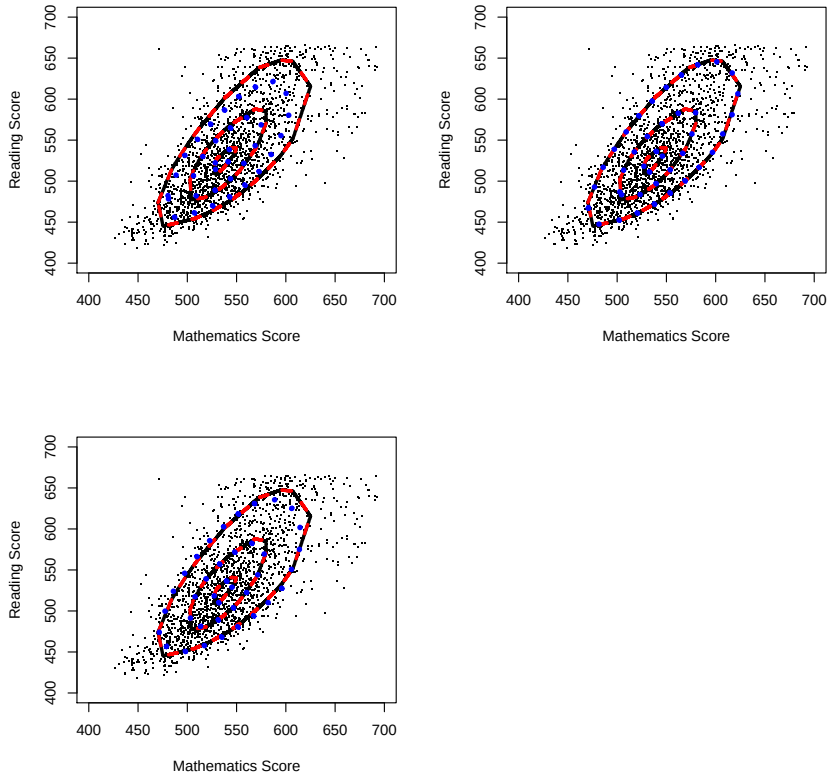

Figure 4: Tukey depth contour prior influence ex-post

6 Conclusion

A Bayesian framework for estimation of multiple-output directional quantiles was presented. The resulting posterior is consistent for the parameters of interest, despite having a misspecified likelihood. By performing inferences as a Bayesian, one inherits many of the

strengths of a Bayesian approach. The model is 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.

A possible avenue for future work is to find a structural economic model whose parameters relate directly to the subgradient conditions. This would give a contextual economic interpretation of the subgradient conditions. Another possibility would be developing a formalized hypothesis test for the distribution comparison presented in Figure 2. This would be a test for the ranking of multivariate distributions based off the directional quantile.

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.

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. iman (2015, 08). 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.

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

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

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.

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.

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.

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

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.

Wang, H. J. and Y. Yang (2016). *AdjBQR: Adjusted Bayesian Quantile Regression Inference*. R package version 1.0.

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.

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.

Zscheischler, J. (2014). *A global analysis of extreme events and consequences for the terrestrial carbon cycle*. Ph. D. thesis, ETH Zurich.

Appendix: A Bayesian Approach to Multiple-Output Quantile Regression

Michael Guggisberg*
Institute for Defense Analyses

August 13, 2019

*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, and Brian Bucks for their helpful comments.*

Appendix

A Review of single-output quantiles, Bayesian quantiles and multiple-output quantiles

A.1 Quantiles and quantile regression

Quantiles sort and rank observations to describe how extreme an observation is. In one dimension, for $\tau \in (0,1)$, the τth quantile is the observation that splits the data into two bins: a left bin that contains $\tau \cdot 100\%$ of the total observations that are smaller and a right bin that contains the rest of the $(1-\tau) \cdot 100\%$ total observations that are larger. The entire family of $\tau \in (0,1)$ quantiles allows one to uniquely define the distribution of interest. Let $Y \in \Re$ be a univariate random variable with Cumulative Density Function (CDF) $F_Y(y) = Pr(Y \leq y)$ then the τth population single-output quantile is defined as

$$Q_Y(\tau) = \inf\{y \in \Re : \tau \leq F_Y(y)\}. \tag{15}$$

If Y is a continuous random variable then the CDF is invertible and the quantile is $Q_Y(\tau) = F_Y^{-1}(\tau)$. Whether or not Y is continuous, $Q_Y(\tau)$ can be defined as the generalized inverse of $F_Y(y)$ (i.e., $F_Y(Q_Y(\tau)) = \tau$).[1] The definition of sample quantile is the same as (15) with $F_Y(y)$ replaced with its empirical counterpart $\hat{F}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} 1_{(y_i \leq y)}$ where $1_{(A)}$ is an indicator function for event A being true.

Fox and Rubin (1964) showed quantiles can be computed via an optimization based approach. Define the check function to be

$$\rho_\tau(x) = x(\tau - 1_{(x<0)}). \tag{16}$$

[1] There are several ways to define the generalized inverse of a CDF (Embrechts and Hofert, 2013; Feng et al., 2012).

The τth population quantile of $Y \in \Re$ is equivalent to $Q_Y(\tau) = \underset{a}{argmin}\, E[\rho_\tau(Y-a)]$. Note, this definition requires $E[Y]$ and $E[Y1_{(Y-a<0)}]$ to be finite. If the moments of Y are not finite, an alternative but equivalent definition can be used instead (Paindaveine and Šiman, 2011). The corresponding sample quantile estimator is

$$\hat{\alpha}_\tau = \underset{a}{argmin} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - a). \tag{17}$$

The commonly accepted definition of single-output linear conditional quantile regression (generally known as 'quantile regression') was originally proposed by Koenker and Bassett (1978). The τth conditional population quantile regression function is

$$Q_{Y|X}(\tau) = \inf\{y \in \Re : \tau \leq F_{Y|X}(y)\} = X'\beta_\tau \tag{18}$$

which can be equivalently defined as $Q_{Y|X}(\tau) = \underset{b}{argmin}\, E[\rho_\tau(Y - X'b)|X]$ (provided the moments $E[Y|X]$ and $E[Y1_{(Y-X'b<0)}|X]$ are finite). The parameter β_τ is estimated by solving

$$\hat{\beta}_\tau = \underset{b}{argmin}\, \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'b). \tag{19}$$

This optimization problem can be written as a linear programming problem and solutions can be found using the simplex or interior point algorithms.

There are two common motivations for quantile regression. First is quantile regression estimates and predictions can be robust to outliers and violations of model assumptions.[2] Second, quantiles can be of greater scientific interest than means or conditional means (as one would find in linear regression).[3] These two motivations extend to multiple-output

[2]For example, the median of a distribution can be consistently estimated whether or not the distribution has a finite first moment.

[3]For example, if one were interested in the effect of police expenditure on crime, one would expect there to be larger effect on high crime areas (large τ) and little to no effect on low crime areas (small τ).

quantile regression. See Koenker (2005) for a survey of the field of single-output quantile regression.

Several approaches to generalizing quantiles from a single-output to a multiple-output random variable have been proposed (Small, 1990; Serfling, 2002). Generalization is difficult because the inverse of the multiple-output CDF is a one-to-many mapping, hence a definition based off inverse CDFs can lead to difficulties. See Serfling and Zuo (2010) for a discussion of desirable criteria one might expect a multiple-output quantiles to have and Serfling (2002) for a survey of extending quantiles to the multiple-output case. Small (1990) surveys the special case of a median.

The proposed method uses a definition of multiple-output quantiles using 'directional quantiles' introduced by Laine (2001) and rigorously developed by Hallin et al. (2010). A directional quantile of $\mathbf{Y} \in \Re^k$ is a function of two objects: a direction vector $\mathbf{u}$ (a point on the surface of k dimension unit hypersphere) and a depth $\tau \in (0, 1)$. A directional quantile is then uniquely defined by $\boldsymbol{\tau} = \mathbf{u}\tau$. The $\boldsymbol{\tau}$ directional quantile hyperplane is denoted $\lambda_{\boldsymbol{\tau}}$ which is a hyperplane through $\Re^k$. The hyperplane $\lambda_{\boldsymbol{\tau}}$ generates two quantile regions: a lower region of all points below $\lambda_{\boldsymbol{\tau}}$ and an upper region of all points above $\lambda_{\boldsymbol{\tau}}$. The lower region contains $\tau \cdot 100\%$ of observations and the upper region contains the remaining $(1 - \tau) \cdot 100\%$. Additionally, the vector connecting the probability mass centers of the two regions is parallel to $\mathbf{u}$. Thus, $\mathbf{u}$ orients the regression and can be thought of as a vertical axis.

A.2 Bayesian single-output conditional quantile regression

Bayesian methods require a likelihood and hence an observational distributional assumption. Yet quantile regression avoids making strong distributional assumptions (a seeming contradiction). Yu and Moyeed (2001) introduced a Bayesian approach by using a possibly

misspecified Asymmetric Laplace Distribution (ALD) likelihood.[4] The Probability Density Function (PDF) of the $ALD(\mu, \sigma, \tau)$ is

$$f_\tau(y|\mu, \sigma) = \frac{\tau(1-\tau)}{\sigma} exp(-\frac{1}{\sigma}\rho_\tau(y-\mu)). \tag{20}$$

The Bayesian assumes $Y|X \sim ALD(X'\beta_\tau, \sigma_\tau, \tau)$, selects a prior, and performs estimation using standard procedures. The nuisance scale parameter can be fixed (typically at $\sigma_\tau = 1$) or freely estimated.[5] Sriram et al. (2013) showed posterior consistency for this model, meaning that as sample size increases, the probability mass of the posterior concentrates around the values of β_τ that satisfy (18). The result holds whether σ_τ is fixed at 1 or freely estimated. Yang et al. (2015) and Sriram (2015) provide a procedure for constructing confidence intervals with correct frequentist coverage probability. If one is willing to accept prior joint normality of β_τ, then a Gibbs sampler can be used to obtain random draws from the posterior (Kozumi and Kobayashi, 2011). Alhamzawi et al. (2012) proposed using an adaptive Lasso sampler to provide regularization. Nonparametric Bayesian approaches have been proposed by Kottas and Krnjajić (2009) and Taddy and Kottas (2010).

A.3 Unconditional multiple-output quantile regression

Any given $\lambda_{\boldsymbol{\tau}}$ quantile hyperplane separates $\mathbf{Y}$ into two halfspaces: an open lower quantile halfspace,

$$H^-_{\boldsymbol{\tau}} = H^-_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) = \{y \in \Re^k : \mathbf{u}'\mathbf{y} < \beta'_{\boldsymbol{\tau}\mathbf{y}}\boldsymbol{\Gamma}'_{\mathbf{u}}\mathbf{y} + \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X} + \alpha_{\boldsymbol{\tau}}\}, \tag{21}$$

and a closed upper quantile halfspace,

$$H^+_{\boldsymbol{\tau}} = H^+_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) = \{y \in \Re^k : \mathbf{u}'\mathbf{y} \geq \beta'_{\boldsymbol{\tau}\mathbf{y}}\boldsymbol{\Gamma}'_{\mathbf{u}}\mathbf{y} + \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X} + \alpha_{\boldsymbol{\tau}}\}. \tag{22}$$

[4]Note, the ALD maximum likelihood estimator is equal to the estimator from (19).

[5]Rahman (2016) and Rahman and Karnawat (2019) are two examples where the scale parameter is used in an ordinal model.

Under certain conditions, the distribution $\mathbf{Y}$ can be fully characterized by a family of hyperplanes $\Lambda = \{\lambda_{\boldsymbol{\tau}} : \boldsymbol{\tau} = \tau\mathbf{u} \in \mathcal{B}^k\}$ (Kong and Mizera, 2012, Theorem 5).[6] There are two subfamilies of hyperplanes: a fixed-$\mathbf{u}$ subfamily, $\Lambda_{\mathbf{u}} = \{\lambda_{\boldsymbol{\tau}} : \boldsymbol{\tau} = \tau\mathbf{u}, \tau \in (0,1)\}$, and a fixed-$\tau$ subfamily, $\Lambda_\tau = \{\lambda_{\boldsymbol{\tau}} : \boldsymbol{\tau} = \tau\mathbf{u}, \mathbf{u} \in \mathcal{S}^{k-1}\}$. The τ subfamily is called a τ quantile regression region (if no $\mathbf{X}$ is included then it is called a τ quantile region). The τ-quantile (regression) region is defined as

$$R(\tau) = \bigcap_{\mathbf{u}\in\mathcal{S}^{k-1}} \cap\{H^+_{\boldsymbol{\tau}}\}, \tag{23}$$

where $\cap\{H^+_{\boldsymbol{\tau}}\}$ is the intersection over $H^+_{\boldsymbol{\tau}}$ if (1) is not unique.

The boundary of $R(\tau)$ is called the τ-quantile (regression) contour. The boundary has a strong connection to Tukey (i.e., halfspace) depth contours. Tukey depth is a multivariate notion of centrality for some point $\mathbf{y} \in \Re^k$. Consider the set of all hyperplanes in $\Re^k$ that pass through $\mathbf{y}$. The Tukey depth of $\mathbf{y}$ is the minimum of the percentage of observations separated by each hyperplane passing through $\mathbf{y}$. Hallin et al. (2010) showed the τ quantile region is equivalent to the Tukey depth region.[7] This provides a numerically efficient approach to find Tukey depth contours.

If $\mathbf{Y}$ (and $\mathbf{X}$ for the regression case) is absolutely continuous with respect to Lebesgue measure, has connected support, and has finite first moments then $(\alpha_\tau, \beta_\tau)$ and λ_τ are unique (Paindaveine and Šiman, 2011).[8] Under this assumption, the subgradient conditions required for consistency are well defined. It follows that $\Psi(a, \mathbf{b})$ is continuously

[6]The conditions required are the directional quantile envelopes of the probability distribution of $\mathbf{Y}$ with contiguous support have smooth boundaries for every $\tau \in (0, 0.5)$

[7]Mathematically, the Tukey (or halfspace) depth of $\mathbf{y}$ with respect to probability distribution P is defined as $HD(\mathbf{y}, P) = \inf\{P[H] : H \text{ is a closed halfspace containing } \mathbf{y}\}$. Then the Tukey halfspace depth region is defined as $D(\tau) = \{\mathbf{y} \in \Re^k : HD(\mathbf{y}, P) \geq \tau\}$. Hallin et al. (2010) show $R(\tau) = D(\tau)$ for all $\tau \in [0, 1)$.

[8]This is Assumption 2, stated formally in Section 2.1.

differentiable with respect to a and $\mathbf{b}$ and convex. The population parameters $(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})$ are defined as the parameters that satisfy two subgradient conditions:

$$\left.\frac{\partial \Psi(a, \mathbf{b})}{\partial a}\right|_{\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}} = Pr(\mathbf{Y_u} - \beta'_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y}^{\perp}_{\mathbf{u}} - \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X} - \alpha_{\boldsymbol{\tau}} \leq 0) - \tau = 0 \tag{24}$$

and

$$\left.\frac{\partial \Psi(a, \mathbf{b})}{\partial \mathbf{b}}\right|_{\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}} = E[[\mathbf{Y}^{\perp\prime}_{\mathbf{u}}, \mathbf{X}']' 1_{(\mathbf{Y_u} - \beta'_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y}^{\perp}_{\mathbf{u}} - \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X} - \alpha_{\boldsymbol{\tau}} \leq)}] - \tau E[[\mathbf{Y}^{\perp\prime}_{\mathbf{u}}, \mathbf{X}']'] = \mathbf{0}_{k+p-1}. \tag{25}$$

The first condition can be written as $Pr(\mathbf{Y} \in H^-_{\boldsymbol{\tau}}) = \tau$, which retains the idea of a quantile partitioning the support into two sets, one with probability τ and one with probability $(1 - \tau)$. The second condition is equivalent to

$$\tau = \frac{E[\mathbf{Y}^{\perp}_{\mathbf{u}i} 1_{(\mathbf{Y} \in H^-_{\boldsymbol{\tau}})}]}{E[\mathbf{Y}^{\perp}_{\mathbf{u}i}]} \text{ for all } i \in \{1, ..., k-1\}$$
$$\tau = \frac{E[\mathbf{X}_i 1_{(\mathbf{Y} \in H^-_{\boldsymbol{\tau}})}]}{E[\mathbf{X}_i]} \text{ for all } i \in \{1, ..., p\}$$

Note that using the law of total expectations $E[\mathbf{Y}^{\perp}_{\mathbf{u}i} 1_{(\mathbf{Y} \in H^-_{\boldsymbol{\tau}})}] = E[\mathbf{Y}^{\perp}_{\mathbf{u}i} | \mathbf{Y} \in H^-_{\boldsymbol{\tau}}] Pr(\mathbf{Y} \in H^-_{\boldsymbol{\tau}}) + 0 Pr(\mathbf{Y} \notin H^-_{\boldsymbol{\tau}}) = E[\mathbf{Y}^{\perp}_{\mathbf{u}i} | \mathbf{Y} \in H^-_{\boldsymbol{\tau}}]\tau$. Then the second condition can be rewritten as

$$E[\mathbf{Y}^{\perp}_{\mathbf{u}i} | \mathbf{Y} \in H^-_{\boldsymbol{\tau}}] = E[\mathbf{Y}^{\perp}_{\mathbf{u}i}] \text{ for all } i \in \{1, ..., k-1\}$$
$$E[\mathbf{X}_i | \mathbf{Y} \in H^-_{\boldsymbol{\tau}}] = E[\mathbf{X}_i] \text{ for all } i \in \{1, ..., p\}.$$

Thus, the probability mass center in the lower halfspace for the orthogonal response is equal to the probability mass center in the entire orthogonal response space. Likewise for the covariates, the probability mass center in the lower halfspace is equal to the probability mass center in the entire covariate space.

The first $k-1$ dimensions of the second subgradient conditions can also be rewritten as $\Gamma'_{\mathbf{u}} E[\mathbf{Y} | \mathbf{Y} \in H^-_{\boldsymbol{\tau}}] = \Gamma'_{\mathbf{u}} E[\mathbf{Y}]$ or equivalently $\Gamma'_{\mathbf{u}}(E[\mathbf{Y} | \mathbf{Y} \in H^-_{\boldsymbol{\tau}}] - E[\mathbf{Y}]) = \mathbf{0}_{k-1}$, which is

satisfied if $E[\mathbf{Y}|\mathbf{Y} \in H_{\boldsymbol{\tau}}^-] = E[\mathbf{Y}]$. This sufficient condition interpretation states that the probability mass center of the response in the lower halfspace is equal to the probability mass center of the response in the entire space. However, this interpretation cannot be guaranteed.

Further note, $E[[\mathbf{Y}_{\mathbf{u}}^{\perp\prime}, \mathbf{X}']'] = E[[\mathbf{Y}_{\mathbf{u}}^{\perp\prime}, \mathbf{X}']' 1_{(\mathbf{Y} \in H_{\boldsymbol{\tau}}^+)}] + E[[\mathbf{Y}_{\mathbf{u}}^{\perp\prime}, \mathbf{X}']' 1_{(\mathbf{Y} \in H_{\boldsymbol{\tau}}^-)}]$. Then the second condition can be written as

$$diag(\boldsymbol{\Gamma}_{\mathbf{u}}', \mathbf{I}_p) \left[\frac{1}{1-\tau} E[[\mathbf{Y}', \mathbf{X}']' 1_{(\mathbf{Y} \in H_{\boldsymbol{\tau}}^+)}] - \frac{1}{\tau} E[[\mathbf{Y}', \mathbf{X}']' 1_{(\mathbf{Y} \in H_{\boldsymbol{\tau}}^-)}] \right] = \mathbf{0}_{k+p-1}.$$

The first $k-1$ components,

$$\boldsymbol{\Gamma}_{\mathbf{u}}' \left[\frac{1}{1-\tau} E[\mathbf{Y} 1_{(\mathbf{Y} \in H_{\boldsymbol{\tau}}^+)}] - \frac{1}{\tau} E[\mathbf{Y} 1_{(\mathbf{Y} \in H_{\boldsymbol{\tau}}^-)}] \right] = \mathbf{0}_{k-1},$$

show $\frac{1}{1-\tau} E[\mathbf{Y} 1_{(\mathbf{Y} \in H_{\boldsymbol{\tau}}^+)}] - \frac{1}{\tau} E[\mathbf{Y} 1_{(\mathbf{Y} \in H_{\boldsymbol{\tau}}^-)}]$ is orthogonal to $\boldsymbol{\Gamma}_{\mathbf{u}}'$ and thus, is parallel to $\mathbf{u}$. It follows the difference of the weighted probability mass centers of the two spaces ($H_{\boldsymbol{\tau}}^-$ and $H_{\boldsymbol{\tau}}^+$) is parallel to $\mathbf{u}$.[9]

Subgradient conditions 2 and 3 can be visualized in Figure 1. The data, $\mathbf{Y}$, are simulated with $1,000$ independent draws from the uniform unit square centered on $(0, 0)$. The directional vector is $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$, represented by the orange $45°$ degree arrow. The depth is $\tau = 0.2$. The hyperplane $\lambda_{\boldsymbol{\tau}}$ is the red dotted line. The lower quantile region, $H_{\boldsymbol{\tau}}^-$, includes the red dots lying below $\lambda_{\boldsymbol{\tau}}$. The upper quantile region, $H_{\boldsymbol{\tau}}^+$, includes the black dots lying above $\lambda_{\boldsymbol{\tau}}$. The probability mass centers of the lower and upper quantile regions are the solid blue dots in their respective regions. The first subgradient condition states that 20% of all points are red. The second subgradient condition states that the line joining the two probability mass centers is parallel to $\mathbf{u}$.

Figure 2 shows an example of a τ-quantile region (left) and fixed-$\mathbf{u}$ (right) halfspaces. The left plot shows fixed-τ-quantile upper halfspace intersections of 32 equally spaced

[9] Hallin et al. (2010) provides an additional interpretation in terms of Lagrange multipliers.

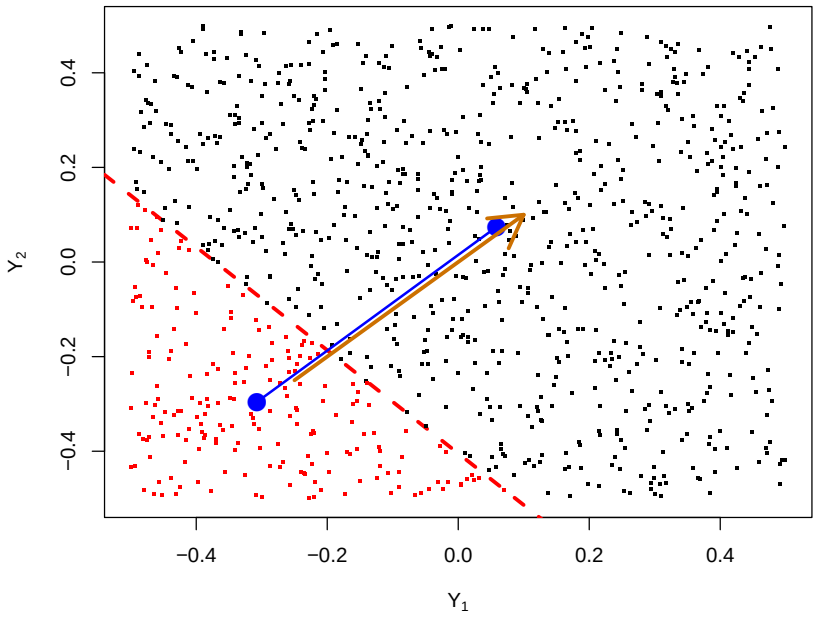

Figure 1: Lower quantile halfspace for $u = (1/\sqrt{2}, 1/\sqrt{2})$ and $\tau = 0.2$

directions on the unit circle for $\tau = 0.2$. Any points on the boundary have Tukey depth 0.2. All points within the shaded blue region have a Tukey depth greater than or equal to 0.2 and all points outside the shaded blue region have Tukey depth less than 0.2.

The right plot of Figure 2 plot shows 13 quantile hyperplanes $\lambda_{\boldsymbol{\tau}}$ for a fixed $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ with various τ (provided in the legend). The orange arrow shows the direction vector $\mathbf{u}$. The hyperplanes split the square such that $\tau \cdot 100\%$ of all points lie below the hyperplanes. The weighted probably mass centers (not shown) are parallel to $\mathbf{u}$. Note the hyperplanes do not need to be orthogonal to $\mathbf{u}$.

Figure 3 shows an example of an unconditional τ-quantile regression tube through a

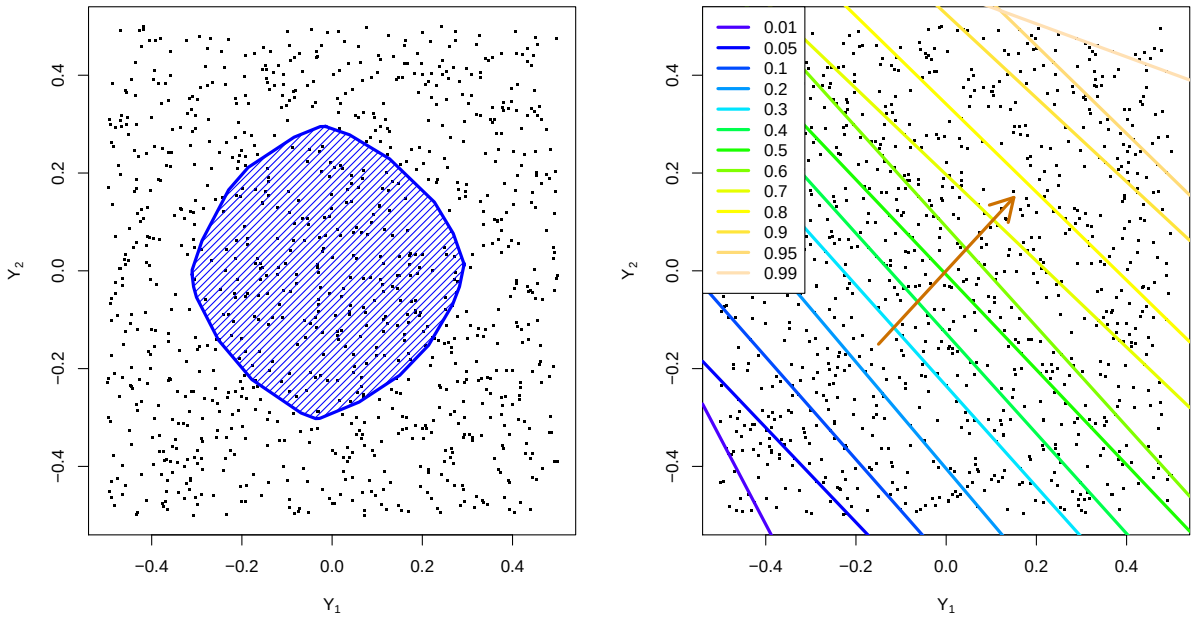

Figure 2: Example of a τ-quantile region and fixed-$\mathbf{u}$ halfspaces. Left, fixed $\tau = 0.2$ quantile region. Right, fixed $u = (1/\sqrt{2}, 1/\sqrt{2})$ quantile halfspaces.

random uniform regular square pyramid. The left plot is a three-dimension scatter plot of the uniform regular square pyramid.[10] The right plot shows the τ-quantile regression tube of Y_1 and Y_2 regressed on Y_3 with cross-section cuts at $Y_3 \in \{0, 0.15, 0.3\}$. As Y_3 increases, the tube travels from the base to the tip of the pyramid and the regression tube pinches.

As in the single-output conditional regression model, the regression tubes are susceptible to quantile crossing. Meaning if one were to trace out the entire regression tube along Y_3

[10] A uniform regular square pyramid is a regular right pyramid with a square base where, for a fixed ϵ, every ϵ-ball contained within the pyramid has the same probability mass. The measure is normalized to one.

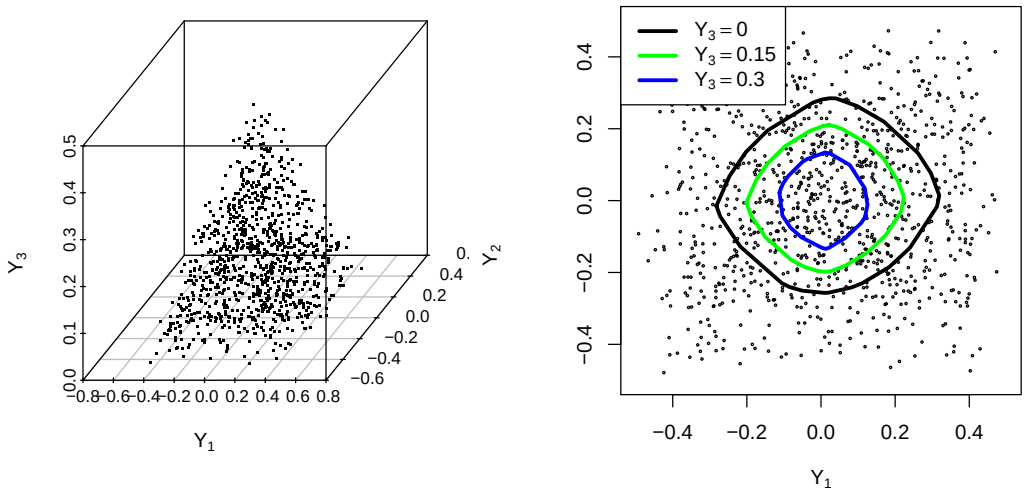

Figure 3: Example of an unconditional τ-quantile regression tube through a uniform regular square pyramid. Left, draws from random uniform regular square pyramid. Right, three slices of an unconditional $\tau = 0.2$ quantile regression tube.

for a given τ and $\tau^\dagger > \tau$, the regression tube for $\tau^\dagger$ might not be contained in the one for τ for all Y_3.

B Proof of Theorem 1

In this section, consistency of the posterior for the population parameters is proven. This proof is for the location case. The regression case should come with easy modification by concatenating $\beta_{\boldsymbol{\tau y}}$ with $\beta_{\boldsymbol{\tau x}}$ and $\mathbf{Y}^{\perp}_{\mathbf{u}i}$ with $\mathbf{X}_i$. Since $\mathbf{Y}$ and $\mathbf{X}$ rely on the same sets of assumptions, and expectations and probabilities are taken over $\mathbf{Y}$ and $\mathbf{X}$, there should not be any issue with these results generalizing to the regression case. For ease of readability

$\boldsymbol{\tau}$ is omitted from $\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}$ and $\Pi_{\boldsymbol{\tau}}$.

Define the population parameters (α_0, β_0) to be the parameters that satisfy (2) and (3). Note that the posterior can be written equivalently as

$$\Pi(U|(\mathbf{Y}_1, \mathbf{X}_1), (\mathbf{Y}_2, \mathbf{X}_2), ..., (\mathbf{Y}_n, \mathbf{X}_n)) = \frac{\int_U \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i, \alpha, \beta, \sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i, \alpha_0, \beta_0, \sigma_0)} d\Pi(\alpha, \beta)}{\int_\Theta \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i \alpha, \beta, \sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i, \alpha_0, \beta_0, \sigma_0)} d\Pi(\alpha, \beta)} \tag{26}$$

Writing the posterior in this form is for mathematical convenience. Define the posterior numerator to be

$$I_n(U) = \int_U \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha, \beta, \sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0, \beta_0, \sigma)} d\Pi(\alpha, \beta). \tag{27}$$

The posterior denominator is then $I_n(\Theta)$. The next lemma (presented without proof) provides several inequalities that are useful later.

Lemma 1. *Let* $b_i = (\alpha - \alpha_0) + (\beta - \beta_0)'\mathbf{Y}_{\mathbf{u}i}^{\perp}$, $W_i = (\mathbf{u}' - \beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i - \alpha_0$, $W_i^+ = max(W_i, 0)$ *and* $W_i^- = min(-W_i, 0)$. *Then a)* $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right) =$

$$\frac{1}{\sigma}\begin{cases} -b_i(1-\tau) & \text{if } (\mathbf{u}' - \beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i - \alpha \leq 0 \text{ and } (\mathbf{u}' - \beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i - \alpha_0 \leq 0 \\ -((\mathbf{u}' - \beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}{}')\mathbf{Y}_i - \alpha_0) + b_i\tau & \text{if } (\mathbf{u}' - \beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i - \alpha > 0 \text{ and } (\mathbf{u}' - \beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i - \alpha_0 \leq 0 \\ (\mathbf{u}' - \beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i - \alpha + b_i\tau & \text{if } (\mathbf{u}' - \beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i - \alpha \leq 0 \text{ and } (\mathbf{u}' - \beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i - \alpha_0 > 0 \\ b_i\tau & \text{if } (\mathbf{u}' - \beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i - \alpha > 0 \text{ and } (\mathbf{u}' - \beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i - \alpha_0 > 0 \end{cases}$$

b) $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right) \leq \frac{1}{\sigma}|b_i| \leq |\alpha - \alpha_0| + |(\beta - \beta_0)'||\boldsymbol{\Gamma}_{\mathbf{u}}'||\mathbf{Y}_i|$

c) $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right) \leq \frac{1}{\sigma}|(\mathbf{u}' - \beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i - \alpha_0| \leq \frac{1}{\sigma}(|(\mathbf{u}' - \beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')||\mathbf{Y}_i| + |\alpha_0|)$

d) $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right) = \frac{1}{\sigma}\begin{cases} -b_i(1-\tau) + \min(W_i^+, b_i) & \text{if } b_i > 0 \\ b_i\tau + \min(W_i^-, -b_i) & \text{if } b_i \leq 0 \end{cases}$

e) $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right) \geq -\frac{1}{\sigma}|b_i| \geq -|\alpha - \alpha_0| - |(\beta - \beta_0)'||\boldsymbol{\Gamma}_{\mathbf{u}}'||\mathbf{Y}_i|$

The next lemma provides more useful inequalities.

Lemma 2. *The following inequalities hold:*

a) $E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] \leq 0$

b) $\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] = E\left[-(W_i-b_i)1_{(b_i<W_i<0)}\right] + E\left[(W_i-b_i)1_{(0<W_i<b_i)}\right]$

c) $\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] \leq E\left[-(W_i-b_i)\right]Pr(b_i < W_i < 0) + E\left[(W_i-b_i)\right]Pr(0 < W_i < b_i)$

d) $\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] \leq -E\left[-\frac{b_i}{2}1_{(b_i<0)}\right]Pr(\frac{b_i}{2} < W_i < 0) - E\left[\frac{b_i}{2}1_{(0<b_i)}\right]Pr(0 < W_i < \frac{b_i}{2})$

e) if Assumption 4 holds then $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[|W_i|] < \infty$.

Proof. Note that $E[b_i] = (\alpha-\alpha_0)+(\beta-\beta_0)'E[\mathbf{Y}_{\mathbf{u}i}^{\perp}] = (\alpha-\alpha_0)+\frac{1}{\tau}(\beta-\beta_0)'E[\mathbf{Y}_{\mathbf{u}i}^{\perp}1_{((\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0\leq 0)}]$ from subgradient condition (12). Define A_i to be the event $(\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0 \leq 0$ and A_i^c it's complement. Define B_i to be the event $(\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha \leq 0$ and B_i^c it's complement.

$$
\begin{aligned}
&\sigma\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\\
&= b_i\tau - b_i1_{(A_i,B_i)} - ((\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0)1_{(A_i,B_i^c)} + ((\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i^c,B_i)}\\
&= b_i\tau - b_i1_{(A_i)} + (b_i - ((\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0))1_{(A_i,B_i^c)} + ((\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i^c,B_i)}\\
&= b_i\tau - b_i1_{(A_i)} - ((\mathbf{u}'_i-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i,B_i^c)} + ((\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i^c,B_i)}
\end{aligned}
$$

Since $E[(\alpha-\alpha_0)1_{(A_i)}] = \tau(\alpha-\alpha_0)$ then $E[b_i\tau - b_i1_{(A_i)}] = 0$. Then

$$\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] = E[-((\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i,B_i^c)}]+E[(\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i^c,B_i)}]$$

The constraint in the first term and second terms imply $-((\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha) < 0$ and $(\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha \leq 0$ over their respective support regions. It follows

$$\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] = E\left[-(W_i-b_i)1_{(b_i<W_i<0)}\right] + E\left[(W_i-b_i)1_{(0<W_i<b_i)}\right].$$

Note that $(W_i - b_i)1_{(0<W_i<b_i)} \leq (W_i - b_i)1_{(0<W_i<\frac{b_i}{2})} < -\frac{b_i}{2}1_{(0<W_i<\frac{b_i}{2})}$. Likewise, $-(W_i - b_i)1_{(b_i<W_i<0)} < \frac{b_i}{2}1_{(\frac{b_i}{2}<W_i<0)}$. Thus,

$$\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] \leq E\left[\frac{b_i}{2}1_{(\frac{b_i}{2}<W_i<0)}\right] + E\left[-\frac{b_i}{2}1_{(\frac{b_i}{2}>W_i>0)}\right].$$

Hölders inequality with $p = 1$ and $q = \infty$ implies $\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] \leq -E\left[-\frac{b_i}{2}1_{(b_i<0)}\right] Pr(\frac{b_i}{2} < W_i < 0) - E\left[\frac{b_i}{2}1_{(0<b_i)}\right] Pr(0 < W_i < \frac{b_i}{2})$. □

The next proposition shows that the KL minimizer is the parameter vector that satisfies the subgradient conditions.

Proposition 1. *Suppose Assumptions 2 and 5 hold. Then*

$$\inf_{(\alpha,\beta)\in\Theta} E\left[\log\left(\frac{p_0(\mathbf{Y}_i)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}\right)\right] \geq E\left[\log\left(\frac{p_0(\mathbf{Y}_i)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right]$$

with equality if $(\alpha,\beta) = (\alpha_0,\beta_0)$ *where* (α_0,β_0) *are defined in (11) and (12).*

Proof. This follows from the previous lemma and the fact that

$$E\left[\log\left(\frac{p_0(\mathbf{Y}_i)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}\right)\right] = E\left[\log\left(\frac{p_0(\mathbf{Y}_i)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right] + E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}\right)\right]$$

□

The next lemma creates an upper bound to approximate $E[I_n(B)^d]$.

Lemma 3. *Suppose Assumptions 3a or 3b hold and 4 holds. Let* $B \subset \Theta \subset \Re^k$. *For* $\delta > 0$ *and* $d \in (0,1)$, *let* $\{A_j : 1 \leq j \leq J(\delta)\}$ *be hypercubes of volume* $\left(\frac{\delta\frac{1}{k}}{1+c_\Gamma c_y}\right)^k$ *required to cover* B. *Then for* $(\alpha^{(j)},\beta^{(j)}) \in A_j$, *the following inequality holds*

$$E\left[\left(\int_B \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta)\right)^d\right] \leq \sum_{j=1}^{J(\delta)}\left[E\left[\left(\prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_j,\beta_j,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)^d\right] e^{nd\delta}\Pi(A_j)^d\right]$$

Proof. For all $(\alpha, \beta) \in A_j$, $|\alpha - \alpha^{(j)}| \leq \frac{\delta^{\frac{1}{k}}}{1+c_\Gamma c_y}$ and $|\beta - \beta^{(j)}| \leq \frac{\delta^{\frac{1}{k}}}{1+c_\Gamma c_y}\mathbf{1}_{k-1}$ compenentwise. Then $|\alpha - \alpha^{(j)}| + |\beta - \beta^{(j)}|'\mathbf{1}_{k-1}c_\Gamma c_y \leq \delta$. Using lemma 1b

$$\begin{aligned}
\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)}\right) &\leq |\alpha - \alpha^{(j)}| + |\beta - \beta^{(j)}|'|\Gamma_u'||\mathbf{Y}_i| \\
&\leq |\alpha - \alpha^{(j)}| + |\beta - \beta^{(j)}|'\mathbf{1}_{k-1}c_\Gamma c_y \\
&\leq \frac{\delta}{1 + c_\Gamma c_y} \\
&< \delta
\end{aligned}$$

Then $\int_{A_j} \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta) =$

$$\begin{aligned}
&\prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} \int_{A_j} \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)} d\Pi(\alpha,\beta) \\
&\leq \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} e^{n\delta}\Pi(A_j)
\end{aligned}$$

Then $E\left[\left(\int_B \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta)\right)^d\right] \leq$

$$\begin{aligned}
&E\left[\left(\sum_{j=1}^{J(\delta)} \left(\prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta)\right) e^{n\delta}\Pi(A_j)\right)^d\right] \\
&\leq \sum_{j=1}^{J(\delta)} E\left[\left(\prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta)\right)^d e^{nd\delta}(\Pi(A_j))^d\right].
\end{aligned}$$

The last inequality holds because $(\sum_i x_i)^d \leq \sum_i x_i^d$ for $d \in (0,1)$ and $x_i > 0$. □

Let $U_n^c \subset \Theta$ such that $(\alpha_0, \beta_0) \notin U_n^c$. The next lemma creates an upper bound for the expected value of the likelihood within U_n^c. Break U_n^c into a sequence of halfspaces,

$\{V_{ln}\}_{l=1}^{L(k)}$, such that $\bigcup_{l=1}^{L(k)} V_{ln} = U_n^c$, where

$$\begin{aligned} V_{1n} &= \{(\alpha,\beta) : \alpha - \alpha_0 \geq \Delta_n, \beta_1 - \beta_{01} \geq 0, ..., \beta_k - \beta_{0k} \geq 0\} \\ V_{2n} &= \{(\alpha,\beta) : \alpha - \alpha_0 \geq 0, \beta_1 - \beta_{01} \geq \Delta_n, ..., \beta_k - \beta_{0k} \geq 0\} \\ &\vdots \\ V_{L(k)n} &= \{(\alpha,\beta) : \alpha - \alpha_0 < 0, \beta_1 - \beta_{01} < 0, ..., \beta_k - \beta_{0k} \leq -\Delta_n\} \end{aligned}$$

for some $\Delta_n > 0$. This sequence makes explicit that the distance of at least one component of the vector (α, β) is larger than its corresponding component of (α_0, β_0) by at least $|\Delta_n|$. How the sequence is indexed exactly is not important. The rest of the proof will focus on V_{1n}, the arguments for the other sets are similar. Define $B_{in} = -E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right]$.[11]

Lemma 4. *Let $G \in \Theta$ be compact. Suppose Assumption 4 holds and $(\alpha,\beta) \in G \cap V_{1n}$. Then there exists a $d \in (0,1)$ such that*

$$E\left[\prod_{i=1}^{n}\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)^d\right] \leq e^{-d\sum_{i=1}^{n} B_{in}}$$

Proof. Define $h_d(\alpha,\beta) = \frac{1-E\left[\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)^d\right]}{d} - E\left[log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right]$. From the proof of Lemma 6.3 in Kleijn and van der Vaart (2006), $\lim_{d\to 0} h_d(\alpha,\beta) = 0$ and $h_d(\alpha,\beta)$ is a decreasing function of d for all (α,β). Note that $h_d(\alpha,\beta)$ is continuous in (α,β). Then by Dini's theorem $h_d(\alpha,\beta)$ converges to $h_d(0,\mathbf{0}_{k-1})$ uniformly in (α,β) as d converges to zero. Define $\delta = \inf_{(\alpha,\beta)\in G} \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)$ then there exists a d_0 such that $0 - h_{d_0}(\alpha,\beta) \leq \frac{\delta}{2}$. From

[11] I would like the thank Karthik Sriram for help with the proof of the next lemma.

lemma 2a $E\left[log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right] < 0$. Then

$$\begin{aligned} E\left[\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)^{d_0}\right] &\leq 1 + d_0 E\left[log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right] + d_0\frac{\delta}{2} \\ &\leq 1 + \frac{d_0}{2} E\left[log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right] \\ &\leq e^{\frac{d_0}{2}E\left[log\left(\frac{f_{\boldsymbol{\tau}}\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right]} \end{aligned}$$

The last inequality holds because $1 + t \leq e^t$ for any $t \in \Re$. □

The next lemma is used to show the numerator of the posterior, $I_n(U_n^c)$, converges to zero for sets U_n^c not containing (α_0, β_0).

Lemma 5. *Suppose Assumptions 3a, 4 and 6 hold. Then there exists a $u_j > 0$ such that for any compact $G_j \subset \Theta$,*

$$\int_{G_j^c \cap V_{jn}} e^{\sum_{i=1}^n \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta) \leq e^{-nu_j}$$

for sufficiently large n.

Proof. Let

$$C_0 = \frac{4 \lim_{n\to\infty} \frac{1}{m}\sum_{i=1}^m E[|W_i|]}{(1-\tau)c_p},$$

$\epsilon = \min(\epsilon_Z)$ and $A = kB\epsilon = 2C_0$, where c_p and ϵ_z are from Assumption 6. This limit exists by Lemma 2e. Define

$$G_1 = \{(\alpha,\beta) : (\alpha - \alpha_0, \beta_1 - \beta_{01}, ..., \beta_k - \beta_{0k}) \in [0,A] \times [0,B] \times ... \times [0,B]\}.$$

If $(\alpha,\beta) \in G_1^c \cap W_1$ then $(\alpha - \alpha_0) > A$ or $(\beta - \beta_0)_j > B$ for some j. If $\mathbf{Y}_{\mathbf{u}i}^{\perp} > \epsilon$ then

$b_i = (\alpha - \alpha_0) + (\beta - \beta_0)'\mathbf{Y}^{\perp}_{\mathbf{u}i} > 2C_0$. Split the likelihood ratio as

$$\sum_{i=1}^{n} \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) =$$
$$\sum_{i=1}^{n} \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) 1_{(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Zj}, \forall j)} + \sum_{i=1}^{n} \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) (1 - 1_{(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Zj}, \forall j)}).$$

Since $\min(W_i^+, b_i) \leq W_i^+ \leq |W_i|$ and using lemma 1 d,

$$\begin{aligned}\sum_{i=1}^{n} log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) 1(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Zj}, \forall j) &= \sum_{i=1}^{n}(-b_i(1-\tau) + min(W_i^+, b_i))1_{(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Zj}, \forall j)} \\ &\leq \sum_{i=1}^{n}(-2C_0(1-\tau) + |W_i|)1_{(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Zj}, \forall j)}.\end{aligned}$$

From lemma 1b and for large enough n then

$$\sum_{i=1}^{n} log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) 1_{(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Zj}, \forall j)} \leq \sum_{i=1}^{n} |W_i|(1 - 1_{(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Zj}, \forall j)}).$$

Then for large enough n

$$\begin{aligned}\sum_{i=1}^{n} log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) &\leq -nC_0(1-\tau)Pr(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Zj}, \forall j) + 2n \lim_{n\to\infty} \frac{1}{m}\sum_{i=1}^{m} E[|W_i|] \\ &= -2n \lim_{n\to\infty} \frac{1}{m}\sum_{i=1}^{m} E[|W_i|] \\ &= -\frac{1}{2}nC_0(1-\tau)Pr(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Zj}, \forall j)\end{aligned}$$

Thus the result holds when $u_i = \frac{1}{2}C_0(1-\tau)Pr(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Zj}, \forall j)$. □

The next lemma shows the marginal likelihood, $I_n(\Theta)$, goes to infinity at the same rate as the numerator in the previous lemma.

Lemma 6. *Suppose Assumptions 3a and 4 hold; then*

$$\int_{\Theta} e^{\sum_{i=1}^{n} \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta) \geq e^{-n\epsilon}.$$

Proof. From Lemma 1e $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) \geq -|b_i| \geq -|\alpha-\alpha_0| - |\beta-\beta_0|'|\Gamma_u||\mathbf{Y}_i|$. Define

$$D_\epsilon = \left\{(\alpha,\beta) : |\alpha-\alpha_0| < \frac{\frac{1}{k}\epsilon}{1+c_\Gamma c_y}, |\beta-\beta_0| < \frac{\frac{1}{k}\epsilon}{1+c_\Gamma c_y}\mathbf{1}_{k-1} \text{ componentwise}\right\}.$$

Then for $(\alpha,\beta) \in V_\epsilon$

$$\begin{aligned}
\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) &\geq -|\alpha-\alpha_0| - |\beta-\beta_0|'|\Gamma_u||\mathbf{Y}_i| \\
&\geq -|\alpha-\alpha_0| - |\beta-\beta_0|'\mathbf{1}_{k-1}c_\Gamma c_y \\
&\geq -\frac{\epsilon}{1+c_\Gamma c_y} \\
&> -\epsilon
\end{aligned}$$

Then $\sum_{i=1}^n \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) \geq -n\epsilon$. If $\Pi(\cdot)$ is proper, then $\Pi(D_\epsilon) \leq 1$.

□

The previous two lemmas imply the posterior is converging to zero in a restricted parameter space.

Lemma 7. *Suppose Assumptions 4 and 6 hold. Then for each $l \in \{1,2,...,L(k)\}$, there exists a compact G_l, such that*

$$\lim_{n\to\infty} \Pi(V_{ln} \cap G_l^c | \mathbf{Y}_1,...,\mathbf{Y}_n) = 0.$$

Proof. Let ϵ from Lemma 6 equal $\frac{u_i}{4}$ from Lemma 5. Then

$$\begin{aligned}
\int_\Theta e^{\sum_{i=1}^n log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta) &\geq \int_{D_\epsilon} e^{\sum_{i=1}^n log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta) \\
&\geq e^{-n\epsilon} d\Pi(D_\epsilon)
\end{aligned}$$

Then $\lim_{n\to\infty} \int_\Theta e^{\sum_{i=1}^n log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta) e^{nu_j/2} = \infty$ and
$\lim_{n\to\infty} \int_{V_{jn}\cap G_j^c} e^{\sum_{i=1}^n log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta) e^{nu_j/2} = 0.$ □

The next proposition bounds the expected value of the numerator, $E[I_n(V_{1n} \cap G)^d]$, and the denominator, $I_n(\Theta)$, of the posterior. Define $B_{in} = -E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right]$.

Lemma 8. *Suppose Assumptions 3a and 4 hold. Define* $D_{\delta_n} = \left\{(\alpha,\beta) : |\alpha - \alpha_0| < \frac{\frac{1}{k}\delta_n}{1+c_\Gamma c_y}, |\beta - \beta_0| < \frac{\frac{1}{k}\delta_n}{1+c_\Gamma c_y}\mathbf{1}_{k-1} \text{ componentwise}\right\}$. *Then for* $(\alpha,\beta) \in D_{\delta_n}$

1. *There exists a* $\delta_n \in (0,1)$ *and fixed* $R > 0$ *such that*

$$E\left[\left(\int_{V_{1n}\cap G}\prod_{i=1}^{n}\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}d\Pi(\alpha,\beta)\right)^d\right] \leq e^{d\sum_{i=1}^n B_{in}}e^{nd\delta_n}R^2/\delta_n^2$$

2.

$$\int_\Theta \prod_{i=1}^{n}\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}d\Pi(\alpha,\beta) \geq e^{-n\delta_n}\Pi(D_{\delta_n})$$

Proof. From Lemma 3 and 4 $E\left[\left(\int_{W_{1n}\cap G}\prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}d\Pi(\alpha,\beta)\right)^d\right]$

$$\begin{aligned}
&\leq \sum_{j=1}^{J(\delta_n)}\left[E\left[\left(\prod_{i=1}^{n}\frac{f_{u,\tau}(\mathbf{Y}_i|\alpha_j,\beta_j,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)^d\right]e^{nd\delta_n}\Pi(A_j)^d\right] \\
&\leq \sum_{j=1}^{J(\delta_n)}\left[e^{-d\sum_{i=1}^n B_{in}}e^{nd\delta_n}\Pi(A_j)^d\right] \\
&\leq e^{-d\sum_{i=1}^n B_{in}}e^{nd\delta_n}J(\delta_n)
\end{aligned}$$

Since G is compact, R can be chosen large enough so that $J(\delta_n) \leq R^2/\delta_n^2$. Line 2. is from Lemma 7. □

The proof of Theorem 1 is below.[12]

[12]I would like to thank Karthik Sriram for help with the proof improper prior case.

Proof. Suppose Π is proper. Lemma 5 shows we can focus on the case $W_{1n} \cap G$. Set $\Delta_n = \Delta$ and $\delta_n = \delta$. Then from Lemma 8, there exists a $d \in (0,1)$ such that for sufficiently large n

$$\begin{aligned} E\left[(\Pi(V_{1n} \cap G | \mathbf{Y}_1, ..., \mathbf{Y}_n))^d\right] &\leq \frac{R^2}{\delta^2 (\Pi(V_\delta))^d} e^{-d\sum_{i=1}^n B_{in}} e^{2nd\delta} \\ &\leq \frac{R^2}{\delta^2 (\Pi(V_\delta))^d} e^{-\frac{1}{2}dn \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^m B_{im}} e^{2nd\delta} \end{aligned}$$

Chose $\delta = \frac{1}{8}\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^m B_{im}$ and note that $C' = \frac{R^2}{\delta^2(\Pi(V_\delta))^d}$ is a fixed constant. Then $E\left[(\Pi(V_{1n} \cap G | \mathbf{Y}_1, ..., \mathbf{Y}_n))^d\right] \leq C' e^{-nd\delta/4}$. Since $\lim_{n\to\infty}\sum_{n=1}^\infty C' e^{-nd\delta/4} < \infty$ then the Markov inequality and Borel Cantelli imply posterior consistency a.s.

Now suppose the prior is improper but admits a proper posterior. Consider the posterior from the first observation $\Pi(\cdot|\mathbf{Y}_1)$. Under Assumption 3b, $\Pi(\cdot|\mathbf{Y}_1)$ is proper. Assumption 5 ensures that $f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0, \beta_0, 1)$ dominates p_0. Thus, the formal posterior exists on a set of $\mathbf{P}$ measure 1. Further, $\Pi(U|\mathbf{Y}_1) > 0$ for some open U containing (α_0, β_0). Thus, $\Pi(\cdot|\mathbf{Y}_1)$ can be used as a proper prior on the likelihood containing $\mathbf{Y}_2, ..., \mathbf{Y}_n$, which produces a posterior equivalent to the original $\Pi(\cdot|\mathbf{Y}_1, ..., \mathbf{Y}_n)$ and thus the same argument above using a proper prior can be applied to the posterior $\Pi(\cdot|\mathbf{Y}_2, ..., \mathbf{Y}_n)$ using $\Pi(\cdot|\mathbf{Y}_1)$ as a proper prior. □

C Proof of Theorem 2

Let $\mathbf{Z}'\mathbf{Z} = r_\tau^2$ represent the spherical τ-Tukey depth contour T_τ and $\mathbf{u}'\mathbf{Z} = d_\tau$ represent the $\lambda_{\boldsymbol{\tau}}$ hyperplane where $d_\tau = \alpha_\tau + \beta_{\tau\mathbf{x}}\mathbf{X}$.

1) Let $\hat{\mathbf{Z}}$ represent the point of tangency between T_τ and $\lambda_{\boldsymbol{\tau}}$. Then the normal vector to $\lambda_{\boldsymbol{\tau}}$ is $\mathbf{u}$. Then there exists a c such that $\hat{\mathbf{Z}} = c\mathbf{u}$. Let $c = \frac{-r}{\sqrt{\mathbf{u}'\mathbf{u}}}$, then $\hat{\mathbf{Z}}'\mathbf{u} = -r_\tau = d_\tau$. Thus $\sqrt{r_\tau^2} = |\alpha_\tau + \beta_{\tau\mathbf{x}}\mathbf{X}|$.

2) Let $\tilde{\mathbf{Z}}$ represent a point on T_τ and $\tilde{\mathbf{u}} = \frac{\tilde{\mathbf{Z}}}{\sqrt{\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}}}$. Then $\tilde{\mathbf{u}}'\tilde{\mathbf{u}} = 1$ implying $\tilde{\mathbf{u}} \in \mathcal{S}^{k-1}$. Note the normal of $\lambda_{\tilde{\tau}}$ is $\tilde{\mathbf{u}}$, which is a scalar multiple of $\tilde{\mathbf{Z}}$. Thus, there exists a $\mathbf{u}$ such that $\lambda_{\boldsymbol{\tau}}$ is tangent to T_τ at every point on T_τ.

3) Let $\mathbf{u} \in \mathcal{S}^{k-1}$ then the normal of $\lambda_{\boldsymbol{\tau}}$ is $\mathbf{u}$. Let $\mathbf{Z} = d_\tau \mathbf{u}$, which is normal to $\lambda_{\boldsymbol{\tau}}$. Further $\mathbf{Z}'\mathbf{Z} = d_\tau^2 \mathbf{u}'\mathbf{u} = d_\tau^2$ is a point on T_τ. Thus, there is a point on $\lambda_{\boldsymbol{\tau}}$ that is tangent to T_τ for every $\mathbf{u} \in \mathcal{S}^{k-1}$.

D Non-zero centered prior: second approach

The second approach is to investigate the implicit prior in the untransformed response space of Y_2 against Y_1, $\mathbf{X}$ and an intercept. Denote $\mathbf{\Gamma_u} = [u_1^\perp, u_2^\perp]'$. Note that $\mathbf{Y_{u}}_i = \beta_{\boldsymbol{\tau}\mathbf{y}} \mathbf{Y}_{\mathbf{u}i}^\perp + \beta'_{\boldsymbol{\tau}\mathbf{x}} \mathbf{X}_i + \alpha_{\boldsymbol{\tau}}$ can be rewritten as

$$\begin{aligned} Y_{2i} &= \frac{1}{u_2 - \beta_{\boldsymbol{\tau}\mathbf{y}} u_2^\perp} \left((\beta_{\boldsymbol{\tau}\mathbf{y}} u_1^\perp - u_1) Y_{1i} + \beta'_{\boldsymbol{\tau}\mathbf{x}} \mathbf{X}_i + \alpha_{\boldsymbol{\tau}} \right) \\ &= \phi_{\boldsymbol{\tau}y} Y_{1i} + \phi'_{\boldsymbol{\tau}\mathbf{x}} \mathbf{X}_i + \phi_{\boldsymbol{\tau}1} \end{aligned}$$

The interpretation of $\phi_{\boldsymbol{\tau}}$ is fairly straight forward since the equation is in slope-intercept form. It can be verified that $\phi_{\boldsymbol{\tau}y} = \phi_{\boldsymbol{\tau}y}(\beta_{\boldsymbol{\tau}\mathbf{y}}) = \frac{\beta_{\boldsymbol{\tau}\mathbf{y}} u_1^\perp - u_1}{u_2 - \beta_{\boldsymbol{\tau}\mathbf{y}} u_2^\perp} = \frac{1}{u_1(u_2^\perp \beta_{\boldsymbol{\tau}\mathbf{y}} - u_2)} + \frac{u_2}{u_1}$ for $\beta_{\boldsymbol{\tau}y} \neq \frac{u_2}{u_2^\perp}$ and $u_1 \neq 0$. Suppose prior $\theta_{\boldsymbol{\tau}} = [\beta_{\boldsymbol{\tau}\mathbf{y}}, \beta'_{\boldsymbol{\tau}\mathbf{x}}, \alpha_{\boldsymbol{\tau}}]' \sim F_{\theta_{\boldsymbol{\tau}}}(\theta_{\boldsymbol{\tau}})$ with support $\Theta_{\boldsymbol{\tau}}$. If $F_{\beta_{\mathbf{y}}}$ is a continuous distribution, the density of $\phi_{\boldsymbol{\tau}}$ is

$$f_{\phi_{\boldsymbol{\tau}y}} = f_{\beta_{\boldsymbol{\tau}\mathbf{y}}}(\phi_{\boldsymbol{\tau}y}^{-1}(\beta_{\boldsymbol{\tau}\mathbf{y}})) \left| \frac{d}{d\beta_{\boldsymbol{\tau}\mathbf{y}}} \phi_{\boldsymbol{\tau}y}^{-1}(\beta_{\boldsymbol{\tau}\mathbf{y}}) \right| = f_{\beta_{\boldsymbol{\tau}\mathbf{y}}} \left(\frac{1}{u_2^\perp} \left(\frac{1}{u_1 \phi_{\boldsymbol{\tau}y} - u_2} + u_2 \right) \right) \left| \frac{u_1}{u_2^\perp (u_1 \phi_{\boldsymbol{\tau}y} - u_2)^2} \right|$$

with support not containing $\left\{ -\frac{u_1^\perp}{u_2^\perp} \right\}$, for $u_2^\perp \neq 0$.

If $\beta_{\boldsymbol{\tau}\mathbf{y}} \sim N(\underline{\mu}_{\boldsymbol{\tau}y}, \underline{\sigma}^2_{\boldsymbol{\tau}y})$, then the density of $\phi_{\boldsymbol{\tau}y}$ is a shifted reciprocal Gaussian with density

$$f_{\phi_{\boldsymbol{\tau}y}}(\phi | \underline{a}, \underline{b}^2) = \frac{1}{\sqrt{2\pi \underline{b}_{\boldsymbol{\tau}}^2 (\phi - u_2/u_2^\perp)^2}} exp \left(-\frac{1}{2\underline{b}_\tau^2} \left(\frac{1}{\phi - u_2/u_2^\perp} - \underline{a} \right)^2 \right).$$

The parameters are $\underline{a} = \underline{\mu}_{\boldsymbol{\tau}} u_1 u_2^{\perp} - u_1 u_2$ and $\underline{b} = u_1 u_2^{\perp} \underline{\sigma}_{\boldsymbol{\tau}}$. The moments of $\phi_{\boldsymbol{\tau} y}$ do not exist (Robert, 1991). The density is bimodal with modes at

$$m_1 = \frac{-\underline{a} + \sqrt{\underline{a}^2 + 8\underline{b}^2}}{4\underline{b}^2} + \frac{u_2}{u_2^{\perp}} \text{ and } m_2 = \frac{-\underline{a} - \sqrt{\underline{a}^2 + 8\underline{b}^2}}{4\underline{b}^2} + \frac{u_2}{u_2^{\perp}}.$$

Elicitation can be tricky, since moments do not exist. However, elicitation can rely on the modes and their relative heights

$$\frac{f_{\phi_{\boldsymbol{\tau} y}}(m_1|\underline{a}, \underline{b}^2)}{f_{\phi_{\boldsymbol{\tau} y}}(m_2|\underline{a}, \underline{b}^2)} = \frac{\underline{a}^2 + \underline{a}\sqrt{\underline{a}^2 + 8\underline{b}^2} + 4\underline{b}^2}{\underline{a}^2 - \underline{a}\sqrt{\underline{a}^2 + 8\underline{b}^2} + 4\underline{b}^2} exp\left(\frac{\underline{a}\sqrt{\underline{a}^2 + 8\underline{b}^2}}{\underline{b}^4}\right)$$

Plots of the reciprocal Gaussian are shown in Figure 4. The left plot presents densities of the reciprocal Gaussian for several hyper-parameter values. The right plot shows contours of the log relative heights of the modes over the set $(\underline{a}, \underline{b}^2) \in [-5, 5] \times [10, 100]$.

The distribution of $\phi_{\boldsymbol{\tau}\mathbf{x}}$ and $\phi_{\boldsymbol{\tau}1}$ are ratio normals. The implied prior on $\phi_{\boldsymbol{\tau}1}$ is discussed. The distribution of $\phi_{\boldsymbol{\tau}\mathbf{x}}$ will follow by analogy. The implied intercept $\phi_{\boldsymbol{\tau}1} = \frac{\alpha_{\boldsymbol{\tau}}}{u_2 - \beta_{\boldsymbol{\tau}\mathbf{y}} u_2^{\perp}}$ is a ratio of normals distribution. The ratio of normals distributions can always be expressed as a location scale shift of $R = \frac{Z_1 + a}{Z_2 + b}$ where $Z_i \overset{iid}{\sim} N(0, 1)$ for $i \in \{1, 2\}$. That is, there exist constants c and d such that $\phi_{\boldsymbol{\tau}1} = cR + d$ (Hinkley, 1969, 1970; Marsaglia, 1965, 2006).[13] The density of $\phi_{\boldsymbol{\tau}1}$ is

$$f_{\phi_{\boldsymbol{\tau}1}}(\phi|\underline{a}, \underline{b}) = \frac{e^{-\frac{1}{2}(\underline{a}^2 + \underline{b}^2)}}{\pi(1 + \phi^2)} \left[1 + ce^{\frac{1}{2}c^2} \int_0^c e^{-\frac{1}{2}t^2} dt\right], \text{ where } c = \frac{\underline{b} + \underline{a}\phi}{\sqrt{1 + \phi^2}}.$$

Note, when $\underline{a} = \underline{b} = 0$, then the distribution reduces to the standard Cauchy distribution. The ratio of normals distribution, like the reciprocal Gaussian distribution, has no moments

[13] Proof: let $W_i \sim N(\theta_i, \sigma_i^2)$ for $i \in \{1, 2\}$ with $corr(W_1, W_2) = \rho$. Then $\frac{W_1}{W_2} = \frac{\sigma_1}{\sigma_2}\sqrt{1 - \rho^2}\left(\frac{\frac{\theta_1}{\sigma_1} + Z_1}{\frac{\theta_2}{\sigma_2} + Z_2} + \frac{\rho}{\sqrt{1 - \rho^2}}\right)$ where $Z_i \sim N(0, 1)$ for $i \in \{1, 2\}$ with $corr(Z_1, Z_2) = 0$. Thus $a = \frac{\theta_1}{\sigma_1}$, $b = \frac{\theta_2}{\sigma_2}$, $c = \frac{\sigma_1}{\sigma_2}\sqrt{1 - \rho^2}$ and $d = c\frac{\rho}{\sqrt{1 - \rho^2}}$ where $\theta_1 = \underline{a}_{\boldsymbol{\tau}1}$, $\theta_2 = u_2 - \underline{a}_{\boldsymbol{\tau} y} u_2^{\perp}$, $\sigma_1 = \underline{b}_{\boldsymbol{\tau}1}$ and $\sigma_2 = \underline{b}_{\boldsymbol{\tau} y} u_2^{\perp}$.

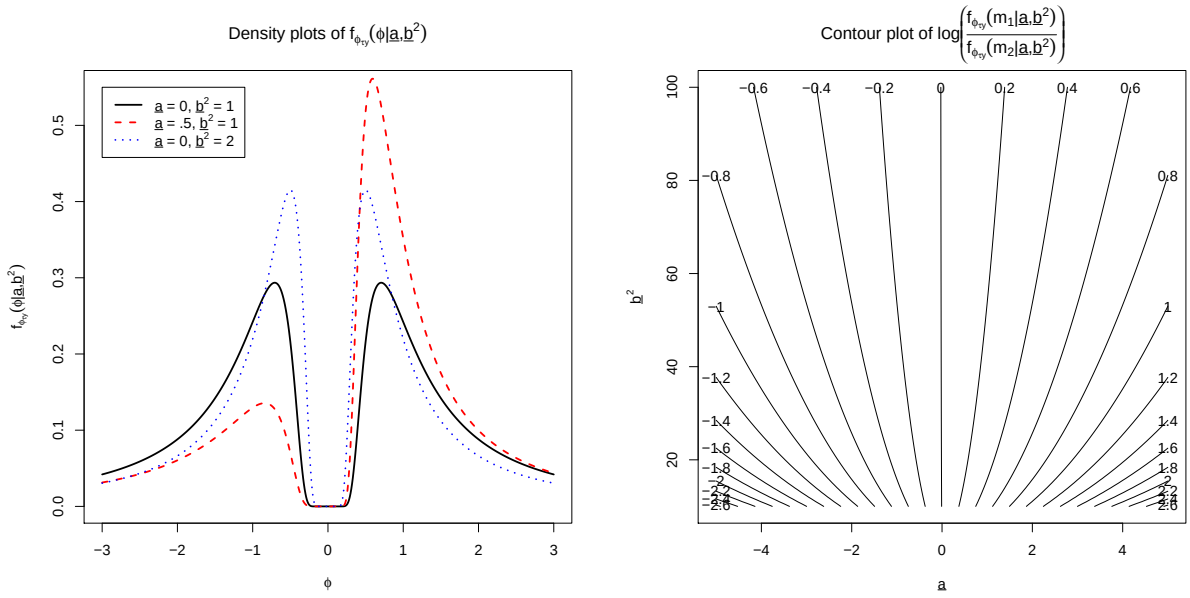

Figure 4: (left) Density of $f_{\phi_{\boldsymbol{\tau} y}}(\phi|\underline{a},\underline{b}^2)$ for hyper parameters $\underline{a}=0$, $\underline{b}^2=1$ (solid black), $\underline{a}=0.5$, $\underline{b}^2=1$ (dash red), $\underline{a}=0$, $\underline{b}^2=2$ (dotted blue). (right) A contour plot showing the log relative heights of the modes at m_1 over m_2 over the set $(\underline{a},\underline{b}^2)\in[-5,5]\times[10,100]$.

and can be bimodal. Focusing on the positive quadrant of $(\underline{a}, \underline{b})$, if $\underline{a} \leq 1$ and $\underline{b} \geq 0$, then ratio of normals distribution is unimodal. If $\underline{a} \gtrsim 2.256058904$, then the ratio of normals distribution is bimodal. There is a curve that separates the unimodal and bimodal regions.[14] Figure 5 shows three plots for the density of the ratio of normals distribution and the bottom right plot shows the regions where the density is unimodal and bimodal. The unimodal region is to the left of the presented curve and the bimodal region is to the right of the presented curve. If the ratio of normals distribution is bimodal, one mode will be to the left of $-\underline{b}/\underline{a}$ and the other to the right of $-\underline{b}/\underline{a}$. The left mode tends to be much lower than the right mode for positive $(\underline{a}, \underline{b})$. Unlike the reciprocal Gaussian, closed form solutions for the modes do not exist. The distribution is approximately elliptical with central tendency $\mu = \frac{\underline{a}}{1.01\underline{b}-0.2713}$ and squared dispersion $\sigma^2 = \frac{\underline{a}^2+1}{\underline{b}^2+0.108\underline{b}-3.795} - \mu^2$ when $\underline{a} < 2.256$ and $4 < \underline{b}$ (Marsaglia, 2006).

E Simulation

E.1 Convergence of subgradient conditions

This section verified convergence of subgradient conditions (2) and (3). For DGPs 1-4, $E[\mathbf{Y}_{\mathbf{u}}^{\perp}] = \mathbf{0}_2$ and for DGP 4, $E[\mathbf{X}] = 0$. Define $\hat{H}_{\boldsymbol{\tau}}^{-}$ to be the empirical lower halfspace where the parameters in (21) are replaced with their Bayesian estimates. Convergence of the first subgradient condition (2) requires

$$\frac{1}{n}\sum_{i=1}^{n} 1_{(\mathbf{Y}_i \in \hat{H}_{\boldsymbol{\tau}}^{-})} \to \tau. \tag{28}$$

[14] The curve is approximately $\underline{b} = \frac{18.621-63.411\underline{a}^2-54.668\underline{a}^3+17.716\underline{a}^4-2.2986\underline{a}^5}{2.256058904-\underline{a}}$ for $\underline{a} \leq 2.256....$

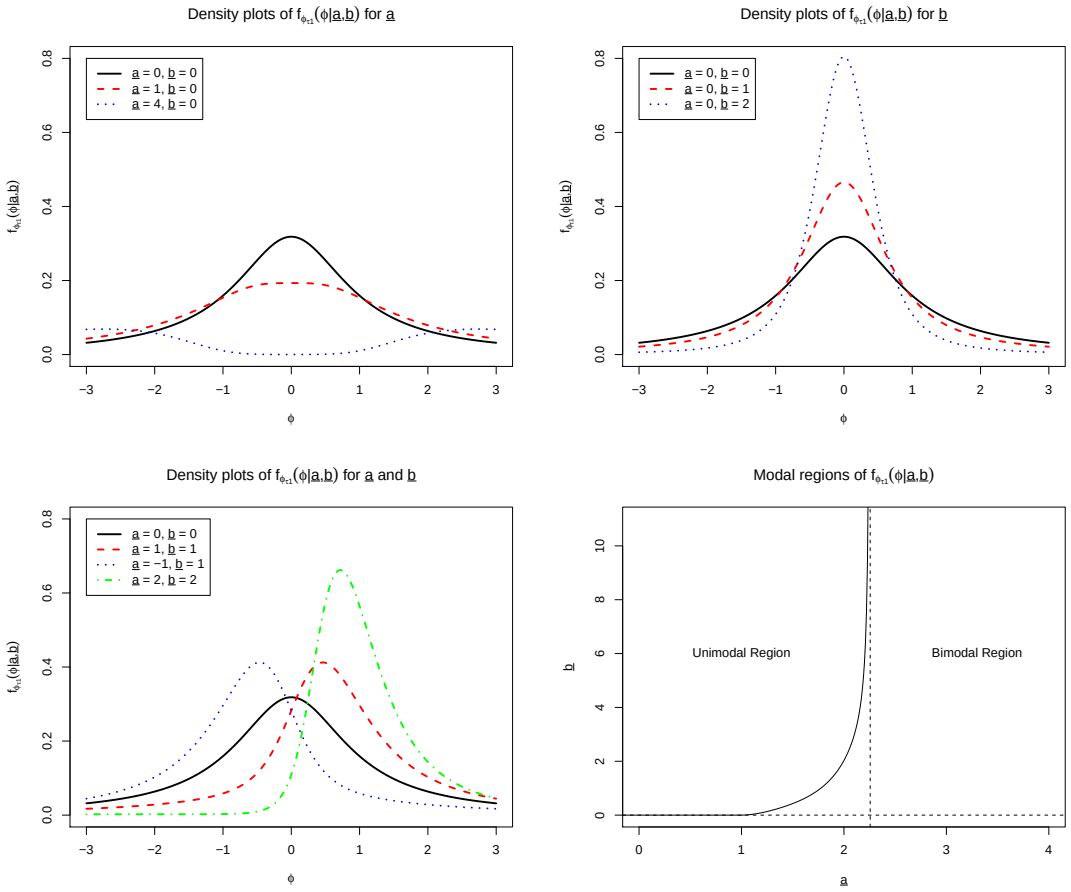

Figure 5: The top two plots and the bottom left plot show the density of the ratio normal distribution with parameters $(\underline{a}, \underline{b})$. The top left plot shows the density for different values of $\underline{a}$ with $\underline{b}$ fixed at zero. The parameters $(\underline{a}, \underline{b}) = (1, 0)$ and $(4, 0)$ result in the same density as $(\underline{a}, \underline{b}) = (-1, 0)$ and $(-4, 0)$. The top right plot shows the density for different values of $\underline{b}$ with $\underline{a}$ fixed at zero. The parameters $(\underline{a}, \underline{b}) = (0, 1)$ and $(0, 2)$ result in the same density as $(\underline{a}, \underline{b}) = (0, -1)$ and $(0, -2)$. The bottom left plot shows the density for different values of $\underline{a}$ and $\underline{b}$. The parameters $(\underline{a}, \underline{b}) = (1, 1)$, $(-1, 1)$and $(2, 2)$ result in the same density as $(\underline{a}, \underline{b}) = (-1, -1)$, $(1, -1)$and $(-2, -2)$. The bottom right graph shows the regions of the positive quadrant of the parameter space where the density is either bimodal or unimodal.

Computation of $1_{(\mathbf{Y}_i \in \hat{H}^-_{\boldsymbol{\tau}})}$ is simple. Convergence of the second subgradient condition (3) requires

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}^{\perp}_{\mathbf{u}i} 1_{(\mathbf{Y}_i \in \hat{H}^-_{\boldsymbol{\tau}})} \to \tau E[\mathbf{Y}^{\perp}_{\mathbf{u}}] \tag{29}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i 1_{(\mathbf{Y}_i \in \hat{H}^-_{\boldsymbol{\tau}})} \to \tau E[\mathbf{X}]. \tag{30}$$

Similar to the first subgradient condition, computation of $\mathbf{Y}^{\perp}_{\mathbf{u}i} 1_{(\mathbf{Y}_i \in \hat{H}^-_{\boldsymbol{\tau}})}$ and $\mathbf{X}_i 1_{(\mathbf{Y}_i \in \hat{H}^-_{\boldsymbol{\tau}})}$ is simple.

Tables 1, 2, and 3 show the results from the simulation. Tables 1 and 2 show the Root Mean Square Error (RMSE) of (28) and (29). Table 1 is using directional vector $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ and Table 2 is using directional vector $\mathbf{u} = (0, 1)$. For Tables 1 and 2, the first three rows show the RMSE for the first subgradient condition (28). The last three rows show the RMSE for the second subgradient condition (29). The second column, n, is the sample size. The next five columns are the DGPs previously described. It is clear that as sample size increases the RMSEs are decreasing, showing the convergence of the subgradient conditions.

Table 3 shows RMSE of (30) for the covariate subgradient condition of DGP 4. The three rows show sample size and the two columns show direction. It is clear that as sample size increases the RMSEs are decreasing, showing convergence of the subgradient conditions.

		Data Generating Process			
	n	1	2	3	4
Sub Grad 1	10^2	4.47e-02	2.91e-02	1.52e-02	1.75e-02
	10^3	5.44e-03	4.59e-03	2.48e-03	2.60e-03
	10^4	9.29e-04	8.66e-04	5.42e-04	5.12e-04
Sub Grad 2	10^2	6.34e-03	1.43e-02	4.34e-02	7.06e-02
	10^3	2.01e-03	3.29e-03	1.32e-02	2.05e-02
	10^4	5.82e-04	8.00e-04	3.59e-03	4.91e-03

Table 1: Unconditional model RMSE of subgradient conditions for $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$

F Application

F.1 Fixed-u hyperplanes

Figure 6 shows fixed-$\mathbf{u}$ $\lambda_{\boldsymbol{\tau}}$ hyperplanes for various τ along a fixed $\mathbf{u}$ direction with model (9). The values of τ are $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$. Two directions are presented: $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ (left) and $\mathbf{u} = (1, 0)$ (right). The direction vectors are represented by the orange arrows passing through the Tukey median (red dot). The hyperplanes to the far left of either graph are for $\tau = 0.01$. The hyperplanes along the direction of the arrow are for larger values of τ, ending with $\tau = 0.99$ hyperplanes on the far right.

		Data Generating Process			
	n	1	2	3	4
Sub Grad 1	10^2	2.02e-02	1.89e-02	1.16e-02	1.36e-02
	10^3	3.38e-03	3.61e-03	1.96e-03	1.98e-03
	10^4	7.71e-04	9.32e-04	3.87e-04	4.68e-04
Sub Grad 2	10^2	9.74e-03	1.35e-02	2.59e-02	2.29e-02
	10^3	2.08e-03	3.24e-03	7.11e-03	6.51e-03
	10^4	6.15e-04	9.89e-04	2.01e-03	1.83e-03

Table 2: Unconditional model RMSE of subgradient conditions for $\mathbf{u} = (0, 1)$

	Direction $\mathbf{u}$	
n	$(1/\sqrt{2}, 1/\sqrt{2})$	$(0, 1)$
10^2	5.17e-02	5.17e-02
10^3	1.41e-02	1.41e-02
10^4	3.90e-03	3.90e-03

Table 3: Unconditional model RMSE of covariate subgradient condition for DGP 4

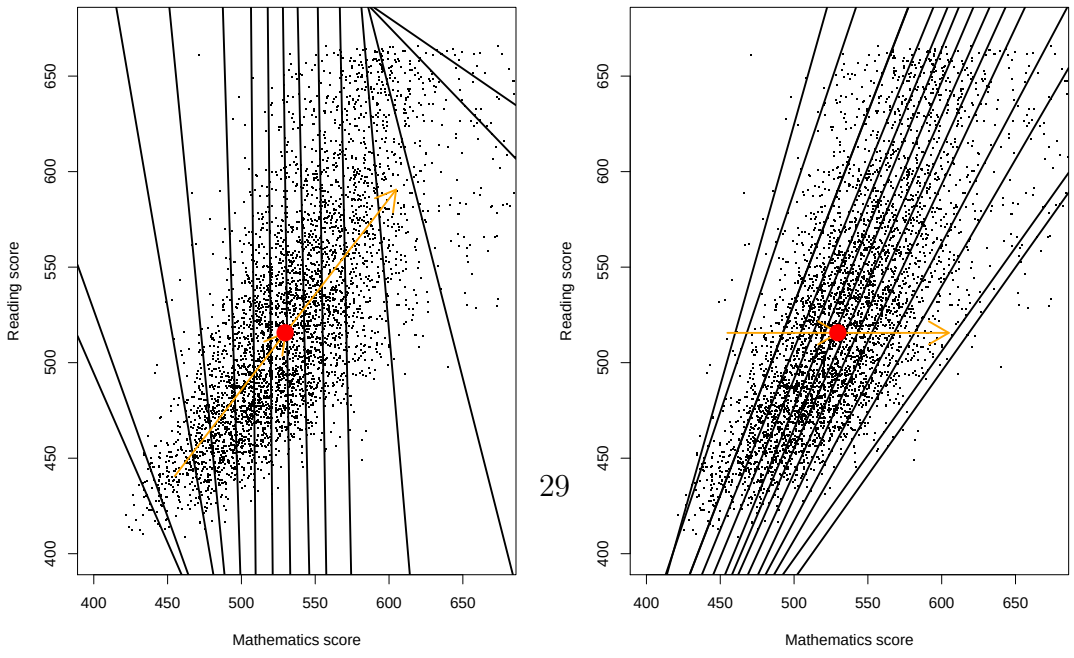

Figure 6: Left, fixed $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ hyperplanes. Right, fixed $\mathbf{u} = (1, 0)$ hyperplanes.

The left plot shows the hyperplanes initially tilt counter-clockwise for $\tau = 0.01$, tilt nearly vertical for $\tau = 0.5$, and then begin tilting counter-clockwise again for $\tau = 0.99$. The hyperplanes in the right plot are all almost parallel tilting slightly clockwise. To understand why this is happening, imagine traveling along the $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ vector through the Tukey median. Data can be thought of as a viscous liquid that the hyperplane must travel through. When the hyperplane hits a dense region of data, that part of the hyperplane is slowed down as it attempts to travel through it, resulting in the hyperplane tilting towards the region with less dense data. Since the density of the data changes as one travels through the $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ direction, the hyperplanes are tilting. However, the density of the data in the $\mathbf{u} = (1, 0)$ direction does not change much, so the tilt of the hyperplanes does not change.

F.2 Sensitivity analysis

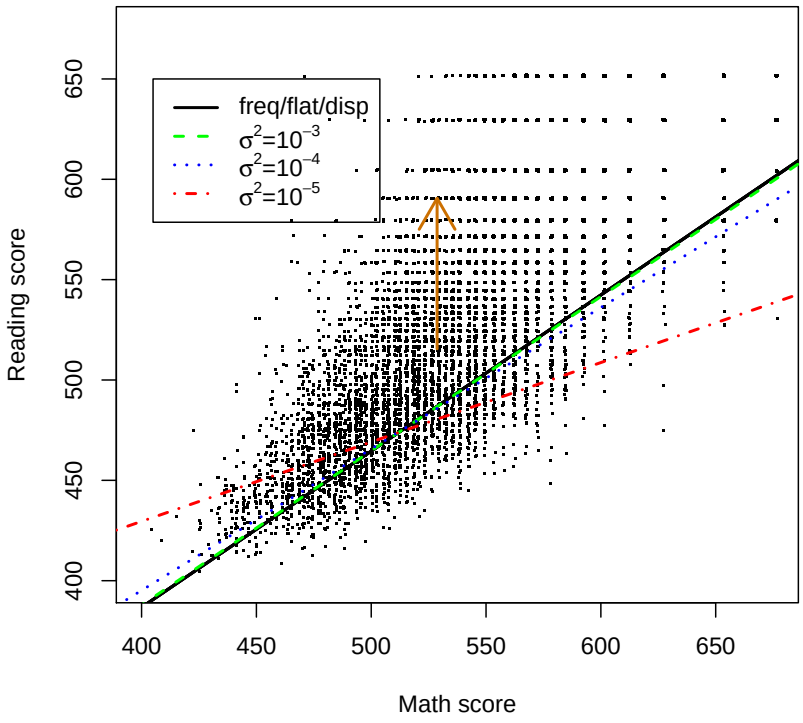

Figure 7: Prior influence ex-post

Figure 7 shows posterior sensitivity to different prior specifications of the location model with directional vector $\mathbf{u} = (0, 1)$ pointing 90° in the reading direction. The posteriors are compared against the frequentist estimate (solid black line). The first specification is the (improper) flat prior (i.e., Lebesgue measure) is represented by the solid black line and cannot be visually differentiated from the frequentist estimate. The rest of the specifications are proper priors with common mean, $\mu_{\theta_\tau} = \mathbf{0}_2$. The dispersed prior has covariance

$\Sigma_{\theta_{\boldsymbol{\tau}}} = 1000\mathbf{I}_2$ and is represented by the solid black line and cannot be visually differentiated from the frequentist estimate or the estimate from the flat prior. The next three priors have covariance matrices $\Sigma_{\theta_{\boldsymbol{\tau}}} = diag(1000, \sigma^2)$ with $\sigma^2 = 10^{-3}$ (dashed green), $\sigma^2 = 10^{-4}$ (dotted blue) and $\sigma^2 = 10^{-5}$ (dash dotted red). As the prior becomes more informative, $\beta_{\boldsymbol{\tau}}$ converges to zero with resulting model $\hat{reading}_i = \alpha_{\boldsymbol{\tau}}$.

References

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

Embrechts, P. and M. Hofert (2013). A note on generalized inverses. *Mathematical Methods of Operations Research 77*(3), 423–432.

Feng, C., H. Wang, X. M. Tu, and J. Kowalski (2012). A note on generalized inverses of distribution function and quantile transformation. *Applied Mathematics 3*(12A), 2098–2100.

Fox, M. and H. Rubin (1964, September). Admissibility of quantile estimates of a single location parameter. *Ann. Math. Statist. 35*(3), 1019–1030.

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.

Hinkley, D. V. (1969). On the ratio of two correlated normal random variables. *Biometrika 56*(3), 635–639.

Hinkley, D. V. (1970). Correction: 'on the ratio of two correlated normal random variables'. *Biometrika 57*(3), 683.

Kleijn, B. J. and A. W. van der Vaart (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics 38*(2), 837–877.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

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

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.

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

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

Marsaglia, G. (1965). Ratios of normal variables and ratios of sums of uniform variables. *Journal of the American Statistical Association 60*(309), 193–204.

Marsaglia, G. (2006, May). Ratios of normal variables. *Journal of Statistical Software 16*, 1–10.

Paindaveine, D. and M. Šiman (2011). On directional multiple-output quantile regression. *Journal of Multivariate Analysis 102*(2), 193 – 212.

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

Rahman, M. A. and S. Karnawat (2019, 09). Flexible Bayesian quantile regression in ordinal models. *Advances in Econometrics 40B*, 211–251.

Robert, C. (1991). Generalized inverse normal distributions. *Statistics & Probability Letters 11*(1), 37–41.

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.

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.

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