Paper with corrections: "Posterior Consistency of Bayesian Quantile Regression Based on the Misspecified Asymmetric Laplace Density"

K. Sriram * and R.V. Ramamoorthi † and P. Ghosh ‡

(Modified July 24, 2017)

Some errors were found in the original published paper: *Karthik Sriram, R.V. Ramamoorthi, Pulak Ghosh (2013) "Posterior Consistency of Bayesian Quantile Regression Based on the Misspecified Asymmetric Laplace Density", Bayesian Analysis, Vol 8, Num 2, pg 479–504*

This version of the paper is created for reference and incorporates the corrections: (i) First error is in the proof of Lemma 4 where we had stated and used an inequality viz, $\forall \quad t < 1, \quad e^t < 1/(1-t)$. We note that this is not true for $t < 0$. We acknowledge and thank Michael Guggisberg, a PhD candidate at University of California, Irvine for bringing this to our attention. Lemma 4 is restated with correct proof. Arguments following Lemma 4 will continue to hold with only minor modifications. These are incorporated in this version for easy reference. (ii) We also found that proof of part (b) of Theorem 2 had an error. We highlight the error and provide a suitable modification of the statement for Theorem 2 part (b).

Note: Places with changes are highlighted in blue. See Pages 7, 15,16, 18, 19, 20, 21.

Abstract. We explore an asymptotic justification for the widely used and empirically verified approach of assuming an asymmetric Laplace distribution (ALD) for the response in Bayesian Quantile Regression. Based on empirical findings, Yu and Moyeed (2001) argued that the use of ALD is satisfactory even if it is not the true underlying distribution. We provide a justification to this claim by establishing posterior consistency and deriving the rate of convergence under the ALD misspecification. Related literature on misspecified models focuses mostly on i.i.d. models which in the regression context amounts to considering i.i.d. random covariates with i.i.d. errors. We study the behavior of the posterior for the misspecified ALD model with independent but non identically distributed response in the presence of non-random covariates. Exploiting the specific form of ALD helps us derive conditions that are more intuitive and easily seen to be satisfied by a wide range of potential true underlying probability distributions for the response. Through simulations, we demonstrate our result and also find that the robustness of the posterior that holds for ALD fails for a Gaussian formulation, thus providing further support for the use of ALD models in quantile regression.

*Indian Institute of Management Ahmedabad, India. karthiks@iimahd.ernet.in
†Michigan State University, East Lansing, Michigan, USA. ramamoor@stt.msu.edu
‡Indian Institute of Management Bangalore, India. pulak.ghosh@iimb.ernet.in

Keywords: Asymmetric Laplace density, Bayesian Quantile Regression, Misspecified models, Posterior consistency

1 Introduction

Quantile Regression is a way to model different quantiles of the dependent variable as a function of covariates (see Koenker and Bassett Jr 1978; Koenker 2005). Given the response variable Y_i and covariate vector $\mathbf{X}_i$ $(i = 1, 2, \ldots, n)$, this involves solving for $\boldsymbol{\beta}$ in the following problem.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}),$$

where $\rho_\tau(u) = u(\tau - I_{(u \leq 0)})$ with $I_{(\cdot)}$ being the indicator function and $0 < \tau < 1$. This can be formulated as a maximum likelihood estimation problem by assuming an asymmetric Laplace distribution (ALD) for the response, *i.e.* $Y_i \sim ALD(., \mu_i^\tau, \sigma, \tau)$, where

$$ALD(y; \mu^\tau, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{ -\frac{(y - \mu^\tau)}{\sigma}(\tau - I_{(y \leq \mu^\tau)}) \right\}, \qquad (1)$$
$$\text{for } -\infty < y < \infty.$$

The ALD is a generalization of the Laplace double-exponential distribution, which is obtained as a special case by taking $\tau = .5$. It is a skewed distribution for $\tau \neq 0.5$. The parameter μ^τ happens to be the τ^{th} quantile of the ALD. For some other properties of ALD, see Yu and Zhang (2005). Yu and Moyeed (2001) proposed the idea of Bayesian quantile regression by assuming ALD for the response. Based on empirical findings, they argued that the use of ALD is satisfactory even if it is not the true underlying distribution. Since then, this method has been used in many problems involving Bayesian quantile regression (e.g. Yu et al. 2005; Yue and Rue 2011). However, to our knowledge, no theoretical justification has been put forward in support of this empirical finding. In this paper, we bridge the gap. We look at the problem where the likelihood is specified to be ALD in the presence of covariates, while the true underlying distribution may be different. We focus on the case of non-random covariates and study posterior consistency of the parameters under misspecification. The arguments can be easily extended to the case of random covariates. While the empirical findings of Yu and Moyeed (2001) restrict to the case of location models with i.i.d. errors, we do not impose such a restriction and attempt to derive more general conditions on the true underlying distribution. In other words, we allow the distribution of the response to be independent but non-identically distributed (i.n.i.d.).

More formally, suppose for observations $i = 1, 2, \ldots, n$, Y_i is the univariate response and $\mathbf{X}_i$ is the vector of p-dimensional covariates, whose components are non-random. Let $\tau \in (0, 1)$ be fixed. The specified model for the response conditional on $\mathbf{X}_i$ is given by $Y_i \sim ALD(., \mu_i^\tau, \sigma, \tau)$ with $\mu_i^\tau = \alpha + \mathbf{X}_i^T \boldsymbol{\beta}$, where α is univariate and $\boldsymbol{\beta}$ is p-dimensional. We denote by $f_{(i,\alpha,\boldsymbol{\beta},\sigma)}(y_i)$, the density function of $ALD(., \alpha + \mathbf{X}_i^T \boldsymbol{\beta}, \sigma, \tau)$

at y_i. However, the true (but unknown) probability distribution of (Y_i given $\mathbf{X}_i$) is P_{0i} with the τ^{th} conditional quantile given by $Q_\tau(Y_i|\mathbf{X}_i) = \alpha_0 + \mathbf{X}_i^T\boldsymbol{\beta}_0$. This also means that $(\alpha_0, \boldsymbol{\beta}_0)$ are the true values for the parameters $(\alpha, \boldsymbol{\beta})$. We note that the other quantiles and hence the distributions P_{0i} need not have an identical form across i as illustrated in section 3.3. We fix the parameter σ to be constant and without loss of generality at 1. We later comment on the case when σ is also endowed with a prior. Let $\Pi(\cdot)$ be a prior on the parameters $(\alpha, \boldsymbol{\beta}) \in \Theta$ where $\Theta \subseteq \Re^{1+p}$ (i.e. the $(p+1)$ dimensional Euclidean space).

Typically in misspecified models the posterior distribution concentrates on a neighborhood of $f_{(i,\alpha^*,\boldsymbol{\beta}^*,1)}$ that has the minimum Kullback-Leibler (KL) divergence from the true density p_{0i}. It will be seen in proposition 1 that in the ALD case, the minimum Kullback-Leibler divergence is attained at $\alpha = \alpha_0$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, which in turn yields consistency for the parameters of interest, namely, $(\alpha, \boldsymbol{\beta})$. Suppose $U_n \subset \Theta,\ n \geq 1$ are open sets such that $(\alpha_0, \boldsymbol{\beta}_0) \in\ U_n$. Then the posterior probability of the set U_n^c (i.e. the complement of U_n) under the specified likelihood is given by,

$$\Pi\left(U_n^c|(Y_1,\mathbf{X}_1),(Y_2,\mathbf{X}_2),\ldots,(Y_n,\mathbf{X}_n)\right) = \frac{\int_{U_n^c}\prod_{i=1}^n f_{(i,\alpha,\boldsymbol{\beta},1)}(Y_i)d\Pi(\alpha,\boldsymbol{\beta})}{\int_\Theta\prod_{i=1}^n f_{(i,\alpha,\boldsymbol{\beta},1)}(Y_i)d\Pi(\alpha,\boldsymbol{\beta})}.$$

In this paper, we derive sufficient conditions under which

$$\Pi\left(U_n^c|(Y_1,\mathbf{X}_1),(Y_2,\mathbf{X}_2),\ldots,(Y_n,\mathbf{X}_n)\right) \rightarrow\ 0\ a.s.\ [P], \tag{2}$$

where P is the true product measure $(P_{01} \times P_{02} \times \cdots \times P_{0n} \times \ldots)$.

Taking $U_n = U, \forall\ n$, (i.e. a fixed neighborhood for all n) gives posterior consistency and choosing a suitable sequence of U_n shrinking to $(\alpha_0, \boldsymbol{\beta}_0)$ gives the rate of convergence. We establish the main results by writing

$$\Pi\left(U_n^c|(Y_1,\mathbf{X}_1),(Y_2,\mathbf{X}_2),\ldots,(Y_n,\mathbf{X}_n)\right) = \frac{\int_{U_n^c}\prod_{i=1}^n \frac{f_{(i,\alpha,\boldsymbol{\beta},1)}(Y_i)}{f_{(i,\alpha_0,\boldsymbol{\beta}_0,1)}(Y_i)}d\Pi(\alpha,\boldsymbol{\beta})}{\int_\Theta\prod_{i=1}^n \frac{f_{(i,\alpha,\boldsymbol{\beta},1)}(Y_i)}{f_{(i,\alpha_0,\boldsymbol{\beta}_0,1)}(Y_i)}d\Pi(\alpha,\boldsymbol{\beta})}. \tag{3}$$

The idea is to then show that under certain conditions, $\exists\ d > 0$ such that the following inequality holds.

$$\sum_{n=1}^\infty E\left[\left(\Pi\left(U_n^c|(Y_1,\mathbf{X}_1),(Y_2,\mathbf{X}_2),\ldots,(Y_n,\mathbf{X}_n)\right)\right)^d\right] < \infty.$$

Markov's inequality along with Borel-Cantelli lemma would then imply that

$$\Pi\left(U_n^c|(Y_1,\mathbf{X}_1),(Y_2,\mathbf{X}_2),\ldots,(Y_n,\mathbf{X}_n)\right) \rightarrow\ 0\ a.s.\ [P].$$

Most studies of posterior consistency of model parameters under misspecification have been in i.i.d. models, which in the regression context amounts to considering random covariates with i.i.d. errors. An early work on this topic is Berk (1966). An exhaustive study of misspecification is carried out by Bunke and Milhaud (1998) and Kleijn and van der Vaart (2006). Bunke and Milhaud (1998) study parametric models. Kleijn and van der Vaart (2006) study L_1 convergence of the posterior, again in the i.i.d. case. Shalizi (2009) considers general non i.i.d. case. Since his results are in a general context, his conditions for the special model considered in this note turn out to be stringent. Besides, his results do not hold for the case of improper priors, which arise in the ALD models naturally as non-informative priors. More recently, Kleijn and van der Vaart (2012) have studied the Bernstein-von-Mises theorem for misspecified models.

In this note our focus is on the ALD model. This model is widely used and empirical studies support consistency of the posterior, or the formal posterior in the case of improper priors, even when ALD is not the true model. We study the behavior of the posterior for the misspecified ALD model with i.n.i.d. response in the presence of non-random covariates. The specific form of the ALD likelihood allows for a more direct derivation leading to simpler, more intuitive conditions and easily extends to the case of improper priors. We thus provide justification for the use of ALD models in quantile estimation and also provide an explanation for the consistency phenomenon observed empirically. Our note extends the earlier results to the non i.i.d. case in the context of ALD models. Our choice of ALD models was dictated by their wide use in applications. Besides, its mathematical tractability provides simple conditions on the "true distributions".

Our methods do have points of contact with Kleijn and van der Vaart (2006) and Ghosal and van der Vaart (2007) but do not directly follow from them. The fixed design misspecified model introduces some complexities. Another issue of interest is when there is a prior on σ. At a technical level, the KL minimizer now depends on i and a suitable point of posterior concentration is not obvious. We believe that the result in this case suggests a possible point of consistency in misspecified models in the general non i.i.d. case.

In what follows, we will first present our key assumptions and the main results in section 2. In section 3, we discuss some applications and demonstrate our results through simulations in section 4. We provide the detailed proof of our results in section 5. We briefly discuss the case of the σ parameter in section 6 and then conclude in section 7.

2 Assumptions and Main Results

In this section, we present our assumptions and the main results. By way of notation, probabilities $P(\cdot)$ and expectations $E(\cdot)$ will always be with respect to the true underlying product measure. To keep the exposition simple, we will work with the case of a univariate non-random covariate. The result is easily extendable to the case of multiple covariates as we remark later. The density function of $ALD(., \alpha + \beta X_i, \sigma, \tau)$ at y_i will be denoted by $f_{(i,\alpha,\beta,\sigma)}(y_i)$ and X_i will be a univariate non-random covariate. Again, for clarity of exposition we work with $\sigma = 1$ and later discuss the case when a prior may be imposed on σ. $\Pi(\cdot)$ is a prior on the parameters (α, β) and the parameter space is denoted by Θ. Without loss of generality, we consider open neighborhoods for (α, β) of the form $U_n = \{(\alpha, \beta) : |\alpha - \alpha_0| < \Delta_n, |\beta - \beta_0| < \Delta_n\}$, where $\Delta_n > 0$. The dependence of the neighborhood on the data size n allows for the derivation of posterior convergence rates along with posterior consistency.

Our assumptions broadly fall into three categories: assumptions on the prior, assumptions on the covariates and assumptions on the true model P. The first assumption is on the prior. As to be expected, the assumption on the prior for obtaining the rate of convergence needs to be a bit stronger than that for just posterior consistency. For clarity, it further helps to separate out the case of posterior consistency under improper priors. Hence, we split the assumption into three parts to cover these cases.

Assumption (1a) (posterior consistency under proper prior): $\Pi(\cdot)$ is proper and every open neighborhood of (α_0, β_0) has positive Π measure.

Assumption (1b) (posterior consistency under improper prior): $\Pi(\cdot)$ is improper, but with a proper posterior and every open neighborhood of (α_0, β_0) has positive Π measure.

Assumption (1c) (posterior consistency rate): $\Pi(\cdot)$ is proper with a probability density function with respect to Lebesgue measure, that is continuous and positive in a neighborhood of (α_0, β_0).

The next assumption is on the covariates.

Assumption 2: $\exists\ M > 0$, such that $|X_i| \leq M\ \forall\ i \geq 1$.

Such an assumption is not unreasonable in most practical situations. For example, in a clinical trial, X_i may be capturing different levels of an administered drug.

The rest of the assumptions involve both the true distribution and the covariates. The next assumption essentially assumes that the quantile is unique. Since the objective of the model is to estimate the τ^{th} quantile, it is reasonable to make such an assumption. Otherwise, the model will not be estimable. A possible way to state uniqueness would be to say that, $\forall\ \Delta > 0,\ P\left(0 < Y_i - \alpha_0 - \beta_0 X_i < \Delta\right) \neq 0$. Similarly, if the X_i's are all constant, then again the model will not be estimable. Therefore, it is reasonable to require that $\{X_i,\ \text{for}\ i \geq 1\}$ take on at least two distinct values each infinitely many times. Without loss of generality (by adjusting the location of the X_i's) this would mean that $\exists\ \epsilon_0 > 0$ such that $\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I_{(X_i > \epsilon_0)}\ > 0$ and $\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I_{(X_i < -\epsilon_0)}\ > 0$. Such a condition is used by Amewou-Atisso et al. (2003). It so happens that we need a combination of the above two types of assumptions. These ideas are captured in assumption 3.

Assumption 3: The below conditions hold.

(i) $\exists\, \epsilon_0 > 0$ such that

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I_{(X_i>\epsilon_0)} \;> 0 \text{ and } \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I_{(X_i<-\epsilon_0)} \;> 0.$$

(ii) $\exists\, C > 0$ such that for all sufficiently small $\Delta > 0$,

$$P\left(0 < Y_i - \alpha_0 - \beta_0 X_i < \Delta\right) > C\Delta,\ \forall\ i$$

and

$$P\left(-\Delta < Y_i - \alpha_0 - \beta_0 X_i < 0\right) > C\Delta,\ \forall\ i.$$

If the random variable $(Y_i - \alpha_0 - \beta_0 X_i)$ has a density that is continuous and positive in a neighborhood of 0, then $\exists\, C_i > 0$ such that

$$P\left(0 < Y_i - \alpha_0 - \beta_0 X_i < \Delta\right) > C_i\Delta$$

for small enough Δ. The second condition in assumption 3 is a stronger requirement where such an inequality needs to hold uniformly across i. However, such a condition will be satisfied if the density of $(Y_i - \alpha_0 - \beta_0 X_i)$ turns out to be a nice function w.r.t. X_i. For example, it is satisfied if the density can be bounded below by a positive continuous function in X_i.

The next assumption is somewhat technical and enables the application of Kolmogorov's Strong Law of Large Numbers (SLLN) for independent random variables.

Assumption 4: For $Z_i = Y_i - \alpha_0 - \beta_0 X_i$,

$$(a) \qquad \limsup_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} E\left(|Z_i|\right) < \infty$$

$$(b) \qquad \sum_{i=1}^{\infty} \frac{E\left(|Z_i|^2\right)}{i^2} < \infty.$$

The last assumption is to ensure that the Kullback-Leibler divergence is well defined between the ALD family and the true probability distribution. Interestingly, this assumption mainly comes into play when we extend the result to the case of improper priors. The true conditional density of Y_i given X_i is denoted by p_{0i}.

Assumption 5: $E\left(\log \frac{p_{0i}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)(Y_i)}}\right) < \infty, \forall\ i.$

Now, we state the main theorems of our paper. For both the theorems, the set up is as follows. $\{Y_i,\ i = 1, 2, \ldots, n\}$ are independent observations of a univariate response and $\{X_i,\ i = 1, 2, \ldots, n\}$ are 1-dimensional non-random covariates. P_{0i} denotes the true (but unknown) probability distribution of Y_i, with the true τ^{th} conditional quantile given by $Q_\tau(Y_i|X_i) = \alpha_0 + \beta_0 X_i$. Suppose however that the specified model for Y_i *is* $ALD(., \mu_i^\tau, \sigma = 1, \tau)$, where $\mu_i^\tau = \alpha + \beta X_i$. $\Pi(\cdot)$ is a prior on (α, β).

Theorem 1. *Under the set up described above, let* $U = \{(\alpha, \beta) : |\alpha - \alpha_0| < \Delta, |\beta - \beta_0| < \Delta\}$. *Let assumptions* **2**, **3** *and* **4** *hold. Also, suppose either*

A. assumption **(1a)** *holds, or*

B. assumption **(1b)** *holds along with assumption 5.*

Then $\Pi(U^c/Y_1, Y_2, \ldots, Y_n) \to 0$ *a.s.* $[P]$.

Theorem 2. *Under the set up described above, let* $U_n = \{(\alpha, \beta) : |\alpha - \alpha_0| < \Delta_n, |\beta - \beta_0| < \Delta_n\}$. *The following hold.*

(a) Let $\Delta_n = Mn^{-\delta}$ *where* $0 < \delta < 1/2$. *Then under assumptions* **(1c)**, **2**, **3** *and* **4**, *we have*

$$\Pi(U_n^c/Y_1, Y_2, \ldots, Y_n) \to 0 \ a.s.\ [P].$$

(b) Let $\Delta_n = M_n/\sqrt{n}$. *Under assumptions* **(1c)**, **2**, **3** *and* **4**, *there exists a large enough constant* C'' *such that for* $M_n^2 > C'' \log(n)$, *we have*

$$\Pi(U_n^c/Y_1, Y_2, \ldots, Y_n) \to 0 \ \text{ in probability } \ [P].$$

Note: There was an error found in proof of part(b) of original paper, where the proof would not go through for any general $M_n \to \infty$. *Hence, part (b) is suitably modified.*

The proofs of the theorems and the accompanying lemmas are presented in detail in section 5.

Remark 1. *It is straight forward to generalize the theorems to accommodate multiple non-random covariates. The conclusions will hold with the same assumptions as in section 2 with some minor modifications. Say,* $\mathbf{X}_i = (X_{i1}, X_{i2})$. *In assumption* **2**, *we just need to bound each component of the covariate vector. Assumption* **3** *needs to be written as follows.*

(i) $\exists\ \epsilon_0 > 0$ *such that* $\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n I_{(S_i)} > 0$, *where* $I_{(\cdot)}$ *is the indicator function and* (S_i) *denotes any one of the conditions:* $(X_{i1} > \epsilon_0, X_{i2} > \epsilon_0)$ *or* $(X_{i1} > \epsilon_0, X_{i2} < -\epsilon_0)$ *or* $(X_{i1} < -\epsilon_0, X_{i2} > \epsilon_0)$ *or* $(X_{i1} < -\epsilon_0,\ X_{i2} < -\epsilon_0)$.

(ii) $\exists\ C > 0$ *such that for all sufficiently small* $\Delta > 0$,

$$P\left(0 < Y_i - \alpha_0 - \mathbf{X}_i^T \boldsymbol{\beta}_0 < \Delta\right) > C\Delta,\ \forall\ i$$

$$P\left(-\Delta < Y_i - \alpha_0 - \mathbf{X}_i^T \boldsymbol{\beta}_0 < 0\right) > C\Delta,\ \forall\ i.$$

3 Applications

In this section, we will demonstrate that the results of the previous section will work for a wide range of possiblities for the true underlying likelihood. Basically, we will investigate part (ii) of assumption **3** and assumption **4**. Assumptions **1**, **2** and part (i) of assumption **3** are either on the prior or the covariates, which we will assume to hold for the purpose of this discussion. It is worth noting that the required assumptions are typically satisfied if the probabilities and expectations involved turn out to be bounded smooth functions of the non-random covariates. Here we analyze two classes of models, namely location models and scale models. We also look at an example of a model that is both a location and scale model.

3.1 Location Models

Let $Y_i = \alpha_0 + \beta_0 X_i + e_i$, where the error terms $\{e_i,\ i = 1, 2, \ldots, n\}$ are i.i.d. from some true unknown distribution P_0 with density p_0 and its τ^{th} quantile at 0. Note that $Z_i = Y_i - \alpha_0 - \beta_0 X_i = e_i$ are i.i.d. Assumption **3** (ii) will be satisfied if the P_0 has a density that is continuous and positive in a neighborhood of 0. In particular, the normal distribution with location shifted so as to make the τ^{th} quantile zero or even mixtures of such distributions would satisfy this condition. Similarly, one can consider location shifted gamma, beta, etc. Assumption **4** is satisfied if the distribution P_0 has finite variance.

3.2 Scale Models

An important feature of our result is that it can cover cases beyond location models for the true underlying likelihood. To demonstrate this, let us consider the case where the density function of Y_i conditional on X_i is given by $p_0\left(\frac{y_i \mu_0}{l(X_i)}\right) \cdot \frac{\mu_0}{l(X_i)}$, where p_0 is a probability density function on $(0, \infty)$ with τ^{th} quantile=μ_0 and $l(X_i) = \alpha_0 + \beta_0 X_i$. We assume that $l(X_i) > 0$. Note that the τ^{th} quantile of Y_i given X_i is $l(X_i)$. A gamma density would be an example of such a model.

We will investigate assumption **3** (ii) by considering one of the sub conditions, since the other one would be similar. Since assumption **2** implies $l(X_i) \le |\alpha_0| + |\beta_0| M$, we have,

$$\begin{aligned} & P(0 < Y_i - l(X_i) < \Delta) \\ = \;& P_0\left(\mu_0 < U < \frac{\Delta \mu_0}{l(X_i)} + \mu_0\right) \\ \ge \;& P_0\left(\mu_0 < U < \frac{\Delta \mu_0}{|\alpha_0| + |\beta_0| M} + \mu_0\right) \end{aligned}$$

where $U \sim P_0$ whose density is p_0. Clearly, assumption **3** (ii) will be satisfied if p_0 is continuous and positive in a neighborhood of μ_0. For assumption **4**, we just note that $Z_i = (U - \mu_0)\, l(X_i)/\mu_0$ and hence $|Z_i| \le |U - \mu_0| \cdot (|\alpha_0| + |\beta_0| M)/\mu_0$. Hence, the condition is satisfied if U has a finite second moment.

3.3 A Normal Location Scale Model

Here we demonstrate that the true likelihood can be more complicated than a purely location or purely scale model. Let $Y_i \sim P_{0i} = N(l(X_i) - \rho_\tau \sigma_i, \sigma_i^2)$, where $l(X_i) = \alpha_0 + \beta_0 X_i$ and ρ_τ is the τ^{th} quantile of the standard normal distribution. We assume that the σ_i can in general vary across i but are bounded, i.e., $0 < \sigma_i < \sigma \; \forall \; i$. Then the τ^{th} quantile of Y_i is $l(X_i)$.

For assumption **3** (ii), we again argue with one of the sub conditions since the argument for the other is similar.

$$
\begin{aligned}
& P(0 < Y_i - l(X_i) < \Delta) \\
= \; & \Phi\left(\frac{\Delta}{\sigma_i} + \rho_\tau\right) - \Phi\left(\rho_\tau\right) \\
\geq \; & \Phi\left(\frac{\Delta}{\sigma} + \rho_\tau\right) - \Phi\left(\rho_\tau\right)
\end{aligned}
$$

where $\Phi(\cdot)$ is the standard normal distribution function. Since the standard normal density is continuous and positive in any neighborhood of ρ_τ, assumption **3** (ii) is satisfied. To check assumption **4**, note that $Z_i = (S - \rho_\tau)\,\sigma_i$, where S is the standard normal random variable. Since σ_i is bounded, assumption **4** would be satisfied.

4 Simulation

We empirically verify the results of this paper by simulating from four different models and checking whether ALD based quantile regression indeed leads to reasonable results. We include two covariates (X_1, X_2) with X_1 being continuous and X_2 being 0-1 valued. For each model, conditional on (X_1, X_2) the $\tau = 75^{th}$ quantile is given by $\alpha_0 + \beta_{01} X_1 + \beta_{02} X_2$ where $(\alpha_0, \beta_{01}, \beta_{02}) = (1, 2, 3)$. For the Bayesian estimation, a normal prior with mean =0 and variance=100 is used for each of the quantile regression coefficients. This kind of a weakly informative prior is commonly used in practice. The four models conditioned on X_1, X_2 can be described as follows:

1. **Location shifted normal** : $Y \sim N(\alpha_0 + \beta_{01} X_1 + \beta_{02} X_2 - \rho_\tau, 1)$ where $\rho_\tau = \rho_{.75}$ is the 75^{th} percentile of standard normal distribution.
2. **Location shifted gamma** : $Y = \alpha_0 + \beta_{01} X_1 + \beta_{02} X_2 - \rho_\tau + e$, where $e \sim Gamma(scale = 1, shape = 1)$ and ρ_τ is the τ^{th} quantile of e.
3. **Scaled gamma** : $Y \sim Gamma(scale = \frac{\rho_\tau}{\alpha_0 + \beta_{01} X_1 + \beta_{02} X_2}, shape = 2)$ where ρ_τ is the τ^{th} quantile of $Gamma(scale = 1,\ shape = 2)$.
4. **Location shifted and scaled normal** : $Y \sim N(\alpha_0 + \beta_{01} X_1 + \beta_{02} X_2 - \rho_\tau |\alpha_0 + \beta_{01} X_1 + \beta_{02} X_2|, |\alpha_0 + \beta_{01} X_1 + \beta_{02} X_2|^2)$.

Bayesian estimation of the ALD model with the above mentioned prior can be done by formulating a Markov Chain Monte Carlo (MCMC) scheme. To facilitate a simple formulation of the MCMC scheme, we use the representation of ALD as a scaled mixture of normals (see Kozumi and Kobayashi 2011). Table 1 shows the 2.5^{th} percentile, mean and the 97.5^{th} percentile of the posterior distribution of the intercept term, covariates X_1 and X_2. In order to get a feel for the convergence of the estimates to the true parameter value, the estimation is done for different data sizes starting from as small as 100 data points to 25000 data points. For each case, the estimation is based on 1000 MCMC simulations after the burn-in period.

Table 1: Bayesian estimation using ALD

	Intercept				X_1				X_2				N
Model	Actual	Q2.5	Mean	Q97.5	Actual	Q2.5	Mean	Q97.5	Actual	Q2.5	Mean	Q97.5	
	1.00	0.32	1.03	1.53	2.00	1.86	2.03	2.26	3.00	2.79	3.15	3.51	100
	1.00	0.42	0.70	1.01	2.00	1.99	2.09	2.18	3.00	2.67	2.83	2.98	500
Location	1.00	0.99	1.16	1.34	2.00	1.90	1.96	2.01	3.00	2.94	3.06	3.19	1,000
Shifted	1.00	0.82	0.91	0.98	2.00	2.02	2.04	2.07	3.00	2.88	2.94	3.00	5,000
Normal	1.00	1.06	1.12	1.19	2.00	1.95	1.97	1.99	3.00	2.95	2.99	3.04	10,000
	1.00	1.01	1.06	1.11	2.00	1.97	1.99	2.00	3.00	2.95	2.99	3.03	15,000
	1.00	0.96	1.01	1.05	2.00	1.99	2.00	2.02	3.00	2.96	2.98	3.01	25,000
	1.00	0.18	0.80	1.34	2.00	1.95	2.10	2.27	3.00	2.21	2.63	3.08	100
	1.00	0.98	1.30	1.62	2.00	1.78	1.88	1.97	3.00	2.79	3.02	3.26	500
Location	1.00	1.05	1.27	1.51	2.00	1.84	1.91	1.98	3.00	2.68	2.81	2.94	1,000
Shifted	1.00	0.91	1.02	1.12	2.00	1.97	2.00	2.03	3.00	2.90	2.97	3.04	5,000
Gamma	1.00	1.03	1.11	1.19	2.00	1.94	1.97	2.00	3.00	2.95	3.00	3.05	10,000
	1.00	0.91	0.98	1.04	2.00	1.99	2.01	2.03	3.00	2.94	2.98	3.02	15,000
	1.00	1.02	1.06	1.11	2.00	1.96	1.98	1.99	3.00	2.98	3.01	3.05	25,000
	1.00	-3.68	-0.10	3.68	2.00	1.48	2.64	3.82	3.00	-0.47	1.60	3.66	100
	1.00	-1.75	-0.51	0.87	2.00	2.17	2.61	3.04	3.00	1.52	2.24	2.94	500
Scaled	1.00	0.74	1.59	2.45	2.00	1.34	1.61	1.90	3.00	3.29	3.96	4.71	1,000
Gamma	1.00	0.60	0.98	1.38	2.00	1.87	1.99	2.13	3.00	3.10	3.42	3.73	5,000
	1.00	0.79	1.07	1.35	2.00	1.93	2.02	2.12	3.00	2.98	3.23	3.46	10,000
	1.00	0.70	0.95	1.20	2.00	1.93	2.01	2.10	3.00	2.98	3.17	3.35	15,000
	1.00	0.82	1.00	1.22	2.00	1.91	1.98	2.04	3.00	2.91	3.05	3.18	25,000
	1.00	-2.89	1.92	6.19	2.00	0.74	2.10	3.52	3.00	0.23	2.93	6.12	100
	1.00	-1.32	1.32	3.76	2.00	1.60	2.30	3.08	3.00	0.04	1.88	3.71	500
Location	1.00	0.00	1.30	2.63	2.00	1.37	1.81	2.25	3.00	1.65	2.76	3.95	1,000
and	1.00	0.25	0.89	1.50	2.00	1.87	2.07	2.29	3.00	2.52	2.97	3.47	5,000
Scale	1.00	0.86	1.35	1.81	2.00	1.77	1.93	2.09	3.00	2.35	2.72	3.07	10,000
Normal	1.00	0.72	1.10	1.48	2.00	1.87	1.99	2.12	3.00	2.62	2.93	3.23	15,000
	1.00	0.86	1.12	1.43	2.00	1.89	1.98	2.07	3.00	2.73	2.96	3.19	25,000

For smaller data sizes, as expected, we see that the distance between the extreme percentiles is larger. However, as the data size increases the distance between the extreme percentiles narrows down towards the true parameter value. The location shift model is the simplest form for the true likelihood. In these cases (models 1 and 2), the convergence happens much faster. This is the type of misspecification considered by Yu and Moyeed (2001) in their empirical analysis. However, our results go beyond location models. We can see that posterior estimates from Bayesian quantile regression based on ALD converge to the true values even in the case when the true underlying likelihood is a scale or location-scale model (models 3 and 4).

Finally, we demonstrate the importance of the property from proposition 1, that

Table 2: Bayesian estimation using Gaussian likelihood (instead of ALD)

	Intercept				X_1				X_2				N
Model	Actual	Q2.5	Mean	Q97.5	Actual	Q2.5	Mean	Q97.5	Actual	Q2.5	Mean	Q97.5	
	1.00	0.62	1.23	1.82	2.00	1.79	1.96	2.15	3.00	2.63	3.02	3.36	100
	1.00	0.57	0.85	1.13	2.00	1.97	2.06	2.15	3.00	2.55	2.75	2.91	500
Location	1.00	1.00	1.20	1.41	2.00	1.89	1.96	2.02	3.00	2.74	2.87	3.01	1,000
Shifted	1.00	0.84	0.93	1.03	2.00	1.99	2.02	2.05	3.00	2.96	3.02	3.08	5,000
Normal	1.00	0.93	1.00	1.06	2.00	1.98	2.00	2.02	3.00	2.98	3.03	3.07	10,000
	1.00	0.96	1.02	1.08	2.00	1.97	1.99	2.01	3.00	2.98	3.01	3.05	15,000
	1.00	0.97	1.01	1.06	2.00	1.98	2.00	2.01	3.00	2.98	3.01	3.03	25,000
	1.00	0.37	0.94	1.50	2.00	1.95	2.12	2.28	3.00	2.38	2.75	3.09	100
	1.00	1.21	1.53	1.81	2.00	1.84	1.93	2.03	3.00	2.84	3.03	3.22	500
Location	1.00	1.18	1.40	1.60	2.00	1.92	1.98	2.04	3.00	2.78	2.93	3.06	1,000
Shifted	1.00	1.20	1.29	1.38	2.00	1.96	1.99	2.02	3.00	2.92	2.98	3.04	5,000
Gamma	1.00	1.22	1.29	1.36	2.00	1.98	2.00	2.02	3.00	2.95	3.00	3.04	10,000
	1.00	1.28	1.33	1.39	2.00	1.97	1.99	2.01	3.00	2.92	2.95	2.99	15,000
	1.00	1.30	1.34	1.38	2.00	1.97	1.99	2.00	3.00	2.97	3.00	3.03	25,000
	1.00	-1.26	0.36	1.95	2.00	1.51	2.06	2.63	3.00	0.07	1.38	2.67	100
	1.00	-0.12	0.92	2.01	2.00	1.17	1.51	1.83	3.00	1.81	2.51	3.24	500
Scaled	1.00	0.46	1.24	2.00	2.00	1.28	1.52	1.76	3.00	1.53	2.12	2.71	1,000
Gamma	1.00	0.66	1.07	1.49	2.00	1.48	1.60	1.73	3.00	1.77	2.05	2.32	5,000
	1.00	0.98	1.29	1.61	2.00	1.44	1.54	1.63	3.00	2.19	2.38	2.57	10,000
	1.00	1.28	1.51	1.73	2.00	1.39	1.46	1.53	3.00	2.03	2.17	2.31	15,000
	1.00	1.13	1.32	1.51	2.00	1.46	1.52	1.58	3.00	2.09	2.21	2.32	25,000
	1.00	-1.31	0.51	2.25	2.00	0.51	1.15	1.77	3.00	-0.95	0.67	2.18	100
	1.00	-1.20	0.26	1.70	2.00	0.21	0.68	1.18	3.00	-1.70	-0.55	0.59	500
Location	1.00	-1.48	-0.13	1.20	2.00	0.61	1.04	1.49	3.00	0.20	1.20	2.15	1,000
and	1.00	-0.28	0.38	1.05	2.00	0.66	0.87	1.09	3.00	0.63	1.11	1.58	5,000
Scale	1.00	0.35	0.87	1.38	2.00	0.56	0.71	0.87	3.00	0.34	0.69	1.02	10,000
Normal	1.00	0.38	0.79	1.27	2.00	0.58	0.73	0.86	3.00	0.53	0.80	1.08	15,000
	1.00	0.64	1.01	1.38	2.00	0.53	0.65	0.76	3.00	0.73	0.96	1.18	25,000

the minimum Kullback-Leibler divergence of ALD from the true likelihood is achieved at the true parameter values $(\alpha_0, \boldsymbol{\beta}_0)$. In order to see this, we carried out Bayesian quantile regression using a Gaussian likelihood instead of ALD. The approach is to assume that the likelihood of Y_i is a standard normal density with location adjusted so that its τ^{th} quantile is $\alpha_0 + \mathbf{X}_i^T \boldsymbol{\beta}_0$. Bayesian estimation of models 1 to 4 was done with this new likelihood specification instead of ALD. Table 2 shows the estimate of intercept, covariates X_1 and X_2 under the Gaussian formulation. Clearly, for model 1 this formulation is indeed correct and in that case the parameter estimates do converge to the true value showing more or less similar performance as the ALD case. However, unlike ALD, for the misspecified normal likelihood in the case of models 2, 3 and 4, the Kullback-Leibler divergence is not minimized at the true parameter values. For model 2, it can be checked that the Kullback-Leibler divergence is minimized for the true values of the slope parameters but not the intercept. Correspondingly, we see that the parameter estimates for the intercept do not converge to the true value while those of the slope parameters do. While the estimates for X_1 and X_2 converge to the true parameter values in the case of models 1 and 2, they break down for the case of scale and location-scale models (3 and 4). Therefore, the Kullback-Leibler divergence minimizing property of ALD seems to play a crucial role.

5 Details of the Proof of the Main Results

Here, we present the proof of the theorems presented in section 2. To keep the exposition simple, we will work with the case of a univariate non-random covariate. In the discussion that follows, we will often work with the log-ratio of ALD likelihood. The first lemma gives some identities and inequalities involving this ratio that are used throughout the paper.

Lemma 1. *Let* $b_i = (\alpha - \alpha_0) + (\beta - \beta_0)X_i$, $Z_i = Y_i - \alpha_0 - \beta_0 X_i$, $Z_i^+ = \max(Z_i, 0)$ *and* $Z_i^- = \max(-Z_i, 0)$. *Then, the following identities and inequalities hold true.*

(a) $\log\left(\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}\right)$

$$= \frac{1}{\sigma}.\begin{cases} -b_i(1-\tau), & \text{if } Y_i \le \min(\alpha+\beta X_i, \alpha_0+\beta_0 X_i) \\ (Y_i - \alpha_0 - \beta_0 X_i) - b_i(1-\tau), & \text{if } \alpha_0+\beta_0 X_i < Y_i \le \alpha+\beta X_i \\ b_i\tau - (Y_i - \alpha_0 - \beta_0 X_i), & \text{if } \alpha+\beta X_i < Y_i \le \alpha_0+\beta_0 X_i \\ b_i\tau, & \text{if } Y_i \ge \max(\alpha+\beta X_i, \alpha_0+\beta_0 X_i) \end{cases}$$

(b) $\left|\log\left(\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}\right)\right| \le |b_i|/\sigma$

(c) $\log\left(\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}\right) \le |Z_i|/\sigma$

(d) If $|X_i| <= M$ *then* $\left|\log\left(\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}\right)\right| \le (|\alpha-\alpha_0| + |\beta-\beta_0|M)/\sigma$

(e) $\log\left(\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}\right) = \frac{1}{\sigma}\cdot\begin{cases} -b_i(1-\tau) + \min(Z_i^+, b_i), & \text{if } b_i > 0 \\ b_i\tau + \min(Z_i^-, -b_i), & \text{if } b_i \le 0 \end{cases}$

We skip the proof of lemma 1 as it easily follows after some algebra. Note that lemma 1 holds for any (α', β') in place of (α_0, β_0).

Lemma 2. *Let* $b_i = (\alpha - \alpha_0) + (\beta - \beta_0)X_i$ *and* $Z_i = Y_i - \alpha_0 - \beta_0 X_i$. *The following identities and inequalities hold true.*

$$(a) \qquad E\left\{\log\left(\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}\right)\right\}$$
$$= E\left(\frac{(Z_i - b_i)}{\sigma} I_{(0<Z_i<b_i)}\right) + E\left(\frac{(b_i - Z_i)}{\sigma} I_{(b_i<Z_i<0)}\right).$$

$$(b) \qquad E\left\{\log\left(\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}\right)\right\} \le 0.$$

Further, equality is achieved if $\alpha = \alpha_0$ *and* $\beta = \beta_0$.

$$(c) \qquad E\left\{\log\left(\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}\right)\right\}$$
$$\le -\frac{b_i}{2\sigma} P\left(0 < Z_i < \frac{b_i}{2}\right) I_{(b_i>0)} + \frac{b_i}{2\sigma} P\left(\frac{b_i}{2} < Z_i < 0\right) I_{(b_i<0)}.$$

Proof. It follows from lemma 1(a) that when $\alpha_0 + \beta_0 X_i < \alpha + \beta X_i$,

$$\log\left(\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}\right)$$
$$= -\frac{b_i(1-\tau)}{\sigma} I_{(Y_i \le \alpha_0+\beta_0 X_i)} + \frac{(Y_i - \alpha - \beta X_i)}{\sigma} I_{(\alpha_0+\beta_0 X_i < Y_i \le \alpha+\beta X_i)}$$
$$+ \frac{b_i \tau}{\sigma} I_{(Y_i > \alpha_0+\beta_0 X_i)}.$$

Taking the expectation given X_i on both sides of the above equation and noting that $\alpha_0 + \beta_0 X_i$ is the τ^{th} quantile of $P(Y_i|X_i)$ and a change of variable gives the result in (a). The argument is similar for the case $\alpha_0 + \beta_0 X_i > \alpha + \beta X_i$. (b) follows from (a) since the expressions inside the expectation in both terms are negative. For (c), note that, if $b_i > 0$,

$$(Z_i - b_i) I_{(0<Z_i<b_i)} < (Z_i - b_i) I_{(0<Z_i<\frac{b_i}{2})} < -\frac{b_i}{2} I_{(0<Z_i<\frac{b_i}{2})}.$$

The case $b_i < 0$ follows similarly. □

Lemma 2 leads to the interesting consequence that for the family of ALD densities $\{f_{(i,\alpha,\beta,1)},\ (\alpha,\beta) \in \Theta\}$, the Kullback-Leibler divergence w.r.t. the true likelihood p_{0i} is minimized for $\alpha = \alpha_0$ and $\beta = \beta_0$. This is recorded in the next proposition.

Proposition 1. *If assumption* **5** *holds then,*

$$\inf_{(\alpha,\beta)\in\Theta} E\left\{\log\left(\frac{p_{0i}(Y_i)}{f_{(i,\alpha,\beta,1)}(Y_i)}\right)\right\} \ge E\left\{\log\left(\frac{p_{0i}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}\right)\right\}.$$

Proof. The result is an immediate consequence of lemma 2 and the following identity

$$E\left\{\log\left(\frac{p_{0i}(Y_i)}{f_{(i,\alpha,\beta,1)}(Y_i)}\right)\right\}$$
$$= E\left\{\log\left(\frac{p_{0i}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}\right)\right\} + E\left\{\log\left(\frac{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}{f_{(i,\alpha,\beta,1)}(Y_i)}\right)\right\}.$$

□

Next is another easy lemma to approximate the numerator of equation (3) over compact sets.

Lemma 3. *Let E be a compact subset of $\Theta \subseteq \Re^2$ and assumption* **2** *hold. For a $\delta > 0$, $0 < d < 1$, let $\{A_j : 1 \le j \le J(\delta)\}$ be squares of area $(\frac{\delta}{1+M})^2$ required to cover E. Then* *for* $(\alpha_j, \beta_j) \in A_j$*, the following inequality holds.*

$$E\left\{\left(\int_E \prod_{i=1}^{n} \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)} d\Pi(\alpha,\beta)\right)^d\right\} \leq \sum_{j=1}^{J(\delta)} \left\{E\left(\prod_{i=1}^{n} \frac{f_{(i,\alpha_j,\beta_j,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}\right)^d \cdot e^{nd\delta} \cdot (\Pi(A_j))^d\right\}.$$

Proof. Let $(\alpha_j, \beta_j) \in A_j$. Then $\forall\ (\alpha, \beta) \in A_j$ using part (d) of lemma 1 (with (α_j, β_j) in place of (α_0, β_0)) we get

$$\log\left(\frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_j,\beta_j,1)}(Y_i)}\right) \leq |\alpha - \alpha_j| + |\beta - \beta_j| M < \delta.$$

Therefore, we have, for each j,

$$\int_{A_j} \prod_{i=1}^{n} \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)} d\Pi(\alpha,\beta) = \prod_{i=1}^{n} \frac{f_{(i,\alpha_j,\beta_j,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)} \cdot \int_{A_j} \prod_{i=1}^{n} \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_j,\beta_j,1)}(Y_i)} d\Pi(\alpha,\beta) \leq \prod_{i=1}^{n} \frac{f_{(i,\alpha_j,\beta_j,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)} \cdot e^{n\delta} \cdot \Pi(A_j).$$

We thus have,

$$E\left\{\int_E \prod_{i=1}^{n} \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)} d\Pi(\alpha,\beta)\right\}^d \leq E\left\{\sum_j \left(\prod_{i=1}^{n} \frac{f_{(i,\alpha_j,\beta_j,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}\right) \cdot e^{n\delta} \cdot \Pi(A_j)\right\}^d \leq \sum_{j=1}^{J(\delta)} \left\{E\left(\prod_{i=1}^{n} \frac{f_{(i,\alpha_j,\beta_j,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}\right)^d \cdot e^{nd\delta} \cdot (\Pi(A_j))^d\right\}.$$

The last inequality follows because $(\sum t_i)^d \leq \sum t_i^d$ for $0 < d < 1, t_i > 0$. □

To establish the desired convergence result for $\Pi(U_n^c|Y_1, \ldots, Y_n)$ we split the set U_n^c as $U_n^c = \bigcup_{j=1}^{8} W_{jn}$ (similar to Amewou-Atisso et al. 2003), where

$$\begin{aligned}
W_{1n} &= \{(\alpha,\beta) : \alpha - \alpha_0 \geq \Delta_n, \beta \geq \beta_0\} \\
W_{2n} &= \{(\alpha,\beta) : \alpha - \alpha_0 \geq \Delta_n, \beta < \beta_0\} \\
W_{3n} &= \{(\alpha,\beta) : \alpha - \alpha_0 < -\Delta_n, \beta \geq \beta_0\} \\
W_{4n} &= \{(\alpha,\beta) : \alpha - \alpha_0 < -\Delta_n, \beta < \beta_0\} \\
W_{5n} &= \{(\alpha,\beta) : \alpha \geq \alpha_0, \beta - \beta_0 \geq \Delta_n\} \\
W_{6n} &= \{(\alpha,\beta) : \alpha < \alpha_0, \beta - \beta_0 \geq \Delta_n\} \\
W_{7n} &= \{(\alpha,\beta) : \alpha \geq \alpha_0, \beta - \beta_0 < -\Delta_n\} \\
W_{8n} &= \{(\alpha,\beta) : \alpha < \alpha_0, \beta - \beta_0 < -\Delta_n\}.
\end{aligned}$$

Then, it is enough to show the result for each of $\Pi(W_{jn}|Y_1, \ldots, Y_n)$. Towards this, we split the parameter space of (α, β) into two parts as $G \cup G^c$ such that G is a compact set and $\Pi(G^c|Y_1,\ldots,Y_n) \to\ 0\ a.s.[P]$.

We will prove the result for W_{1n}. The argument is similar for other W_{jn}.

Let $B_{in}^{(1)} = -\Delta_n \cdot P\left(0 < Z_i < \frac{\Delta_n}{2}\right) \cdot I_{(X_i > \epsilon_0)}$, with ϵ_0 as in assumption 3.

Note: Lemma 4 and proof have been updated

Lemma 4. *Let $G \subseteq \Theta$ be compact and assumption **2** hold. Let $\epsilon_0 > 0$ be as in assumption 3(i) and $C > 0$ be as in assumption 3(ii). Then $\exists\ 0 < d < 1$ such that for $K = \frac{C\tau(1-\tau)}{2} > 0$ and $\forall\ (\alpha,\beta) \in G \cap W_{1n}$,*

$$E\left[e^{dT_i}\right] \le e^{-dK\Delta_n^2 I_{X_i > \epsilon_0}}.$$

Proof. We will assume $b_i \ge 0$ as the argument is similar when $b_i < 0$. We note by Lemma 1a that when $\alpha_0 + \beta_0 X_i < Y_i \le \alpha + \beta X_i$,

$$\begin{aligned} T_i &= (Y_i - \alpha_0 - \beta_0 X_i) - b_i(1-\tau) = Y_i - q_i \\ &\quad \text{where } q_i = (\alpha_0 + \beta_0 X_i)\tau + (\alpha + \beta X_i)(1-\tau). \end{aligned}$$

$$\text{So, } (Y_i - q_i) \le \begin{cases} 0, & if\ \ \alpha_0 + \beta_0 X_i < Y_i \le q_i \\ (\alpha + \beta X_i - q_i) = b_i\tau, & if\ \ q_i < Y_i < \alpha + \beta X_i \end{cases}.$$

This observation along with Lemma 1a, implies

$$T_i \ \le\ -b_i(1-\tau) \times I_{Y_i \le \alpha_0 + \beta_0 X_i} + 0 \times I_{\alpha_0 + \beta_0 X_i < Y_i \le q_i} + b_i\tau \times I_{Y_i > q_i}. \tag{4}$$

Denoting $\tau_i^* = P(Y_i \le q_i)$ and recalling that $\tau = P(Y_i \le \alpha_0 + \beta_0 X_i)$,

$$E\left[e^{dT_i}\right] \le \tau e^{-db_i(1-\tau)} + (\tau_i^* - \tau) + e^{db_i\tau}(1 - \tau_i^*). \tag{5}$$

Let $g_i(t) = e^{-tb_i(1-\tau)}\tau + (\tau_i^* - \tau) + e^{tb_i\tau}(1 - \tau_i^*)$. By Taylor's formula,

$$g_i(t) \ =\ 1 + g_i'(0)t + g_i''(\xi)t^2/2, \quad \text{for some } 0 < \xi < t. \tag{6}$$

In equation (6), we first note that $g_i'(0) = -b_i\tau(\tau_i^* - \tau)$. Suppose, C, Δ_0 be as in Assumption 3(ii), i.e. $P(0 < Y_i - \alpha_0 - \beta_0 X_i < \Delta) > C\Delta\ \ \forall\ \ \Delta \le \Delta_0$. Defining $b_i^* = \min(b_i, \Delta_0)$ and noting that $q_i - \alpha_0 - \beta_0 X_i = b_i(1-\tau)$, we have

$$\begin{aligned} \tau_i^* - \tau &= P(\alpha_0 + \beta_0 X_i < Y_i \le q_i) = P(0 < Z_i \le b_i(1-\tau)) \\ &\ge P(0 < Z_i \le b_i^*(1-\tau)) > C\ b_i^*\ (1-\tau). \end{aligned}$$

$$\text{Hence } g_i'(0) \ \le\ -C\tau(1-\tau){b_i^*}^2. \tag{7}$$

Further, we note $g_i''(t) = b_i^2 \times \left(\tau(1-\tau)^2 e^{-tb_i(1-\tau)} + \tau^2(1-\tau_i^*)e^{tb_i\tau}\right)$. Since G is compact and hence b_i is uniformly bounded, say $b_i \le M_1\ \ \forall i$, the term within the parenthesis in the above expression can be bounded by some constant $K_1 > 0$. Further, note by definition that $b_i^* = b_i I_{b_i \leq \Delta_0} + \Delta_0 I_{b_i > \Delta_0}$. Therefore, if we choose $K_2 > 1$, such that $K_2\Delta_0 > M_1$, then we would have $K_2 b_i^* = (K_2 b_i I_{b_i \leq \Delta_0} + K_2\Delta_0 I_{b_i > \Delta_0}) \geq b_i$. In other words, $\exists\ K_2$ such that $b_i \leq K_2 b_i^*$ or $b_i^2 \leq K_2^2 {b^*}_i^2$. Therefore, by taking $2K_3 = K_1 \cdot K_2^2$, we get

$$g_i''(t) \leq 2K_3 \cdot {b_i^*}^2, \quad \forall\ 0 \leq t \leq 1. \tag{8}$$

Equations (6), (7) and (8) together give

$$g_i(t) \quad \leq \quad 1 - {b_i^*}^2 \cdot t \cdot (C\tau(1-\tau) - K_3 t)\,.$$

Let $t_0 < \min\left(\frac{1}{2}, \frac{1}{2}\frac{C\tau(1-\tau)}{K_3}\right)$ and $K = \frac{C\tau(1-\tau)}{2}$ then $\forall\ \ t < t_0$ we have,

$$g_i(t) \leq 1 - tK{b_i^*}^2 \leq e^{-tK{b_i^*}^2}. \tag{9}$$

We have $b_i^* \geq 0,\ \forall\ i$. Further, when $X_i > \epsilon_0$ and $(\alpha, \beta) \in W_{1n}$, we have $b_i \geq \Delta_n$. If we assume without loss of generality that $\Delta_0 > \Delta_n$, then $b_i^* \geq \Delta_n I_{(X_i > \epsilon_0)},\ \forall\ i$. It follows therefore that for $(\alpha, \beta) \in W_{1n} \cap G$,

$$\forall\ \ d < t_0, \quad E\left[e^{dT_i}\right] \leq e^{-dK{b_i^*}^2} \leq e^{-dK\Delta_n^2 I_{(X_i > \epsilon_0)}}.$$

□

The next two lemmas help construct a specific compact subset of the parameter space outside of which the posterior probability goes to zero almost surely.

Lemma 5. *Let* $\Pi(\cdot)$ *be proper and assumptions* **2**, **3** *and* **4** *hold. Then for* $j = 1, 2, \ldots, 8,\ \exists$ *a compact set* $G_j \subset \Theta$, $u_j > 0$ *such that for sufficiently large* n,

$$\int_{G_j^c \cap W_{jn}} e^{\sum_{i=1}^n \log \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}} d\Pi(\alpha, \beta) \leq e^{-nu_j}.$$

Proof. We will prove the result for the set W_{1n}. The argument is similar for other sets W_{jn} for j=2,...,8. Recall that $W_{1n} = \{(\alpha, \beta) : \alpha - \alpha_0 \geq \Delta_n, \beta \geq \beta_0\}$. Let ϵ_0 be as in assumption **3** and $Z_i = Y_i - \alpha_0 - \beta_0 X_i$.

$$\text{Let } C_0 = \frac{4\ \limsup_{m\to\infty} \frac{1}{m}\sum_{i=1}^m E(|Z_i|)}{(1-\tau)\liminf_{m\to\infty} \frac{1}{m}\sum_{i=1}^m I_{(X_i > \epsilon_0)}}.$$

Note that assumption **3** in particular implies that the denominator is well defined and assumption **4** ensures that the numerator is well defined. Now let $A = B\epsilon_0 = 2C_0$ and define

$$G_1 \quad = \quad \{(\alpha, \beta) : (\alpha - \alpha_0, \beta - \beta_0) \in [0, A] \times [0, B]\}.$$

Clearly G_1 is compact. Now if $(\alpha,\beta) \in G_1^c \cap W_{1n}$ then either $(\alpha-\alpha_0) > A$ or $(\beta-\beta_0) > B$. Further, if $X_i > \epsilon_0$ then in the former case we have $b_i = (\alpha-\alpha_0) + (\beta-\beta_0)X_i > A$ and in the latter case we would have $b_i > B\epsilon_0$. So, in either case when $X_i > \epsilon_0$, we have $b_i > 2C_0$. We can write

$$\begin{aligned}&\sum_{i=1}^{n}\log\left(\frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}\right)\\ &=\sum_{i=1}^{n}\log\left(\frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}\right)I_{(X_i>\epsilon_0)}+\sum_{i=1}^{n}\log\left(\frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}\right)I_{(X_i\leq\epsilon_0)}.\end{aligned}$$

Now, applying part (e) of lemma 1 to the first term in the right hand side (R.H.S.) and part (c) to the second term (for $(\alpha,\beta) \in G_1^c \cap W_1$), for sufficiently large n we have,

$$\begin{aligned}&\sum_{i=1}^{n}\log\left(\frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}\right)\\ &\leq\left(-2C_0(1-\tau)\sum_{i=1}^{n}I_{(X_i>\epsilon_0)}+\sum_{i=1}^{n}Z_i^{+}I_{(X_i>\epsilon_0)}+\sum_{i=1}^{n}|Z_i|I_{(X_i\leq\epsilon_0)}\right)\\ &\leq -nC_0(1-\tau)\liminf_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}I_{(X_i>\epsilon_0)}+2n\limsup_{m\to\infty}\frac{1}{m}\sum_{i=1}^{n}E\left\{|Z_i|\right\}\\ &\leq -\frac{nC_0(1-\tau)}{2}\liminf_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}I_{(X_i>\epsilon_0)}.\end{aligned}$$

The last but one inequality follows by using assumption **4**, which allows the application of the SLLN on the sequence $\{|Z_n|\}$ and the last step follows due to our specific choice of C_0. Now, the result follows by using propriety of prior and taking

$$u_1=\frac{C_0(1-\tau)}{2}\liminf_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}I_{(X_i>\epsilon_0)}.$$

□

Lemma 6. *Let assumptions* **2**, **3** *and* **4** *hold. Also suppose either assumption* **1a** *or* **1c** *holds. Then for each* $j = 1, 2, \ldots, 8,\ \exists$ *a compact set* $G_j \subset \Theta$ *such that*

$$\Pi(W_{jn}\cap G_j^c|Y_1,\ldots,Y_n)\to\ 0\ a.s.\ [P].$$

Proof.

$$\Pi(W_{jn}\cap G_j^c|Y_1,\ldots,Y_n)=\frac{\int_{G_j^c\cap W_{jn}}e^{\sum_{i=1}^{n}\log\frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}}d\Pi(\alpha,\beta)}{\int_{\Theta}e^{\sum_{i=1}^{n}\log\frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}}d\Pi(\alpha,\beta)}=\frac{I'_{1n}}{I'_{2n}}.$$

Lemma 5 implies that $e^{nu_j/2}I'_{1n} \to 0$ $a.s.$ $[P]$. To prove the lemma, it is therefore enough to show that $e^{nu_j/2}I'_{2n} \to \infty$ $a.s.$ $[P]$.

Let $\epsilon = u_j/4$ and $M > 0$ be as in assumption **2**. Define

$$V_\epsilon = \{(\alpha, \beta) : |\alpha - \alpha_0| < \epsilon/(1+M), |\beta - \beta_0| < \epsilon/(1+M)\}.$$

Using part (d) of lemma 1, note that for $(\alpha, \beta) \in V_\epsilon$, we have

$$\sum_{i=1}^{n} \log \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)} > -n\epsilon.$$

If follows that $I'_{2n} > \int_{V_\epsilon} e^{\sum_{i=1}^{n} \log \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}} d\Pi(\alpha, \beta) > e^{-n\epsilon}\Pi(V_\epsilon)$. Note that $\Pi(V_\epsilon) > 0$ holds under either of the assumptions **1a** and **1c**. Hence,

$$e^{nu_j/2}I'_{2n} = e^{2n\epsilon}I'_{2n} > e^{n\epsilon}\Pi(V_\epsilon) \to \infty \ a.s. \ [P].$$

□

Note: Following the change to the statement of Lemma 4, Proposition 2 is updated.

We summarize the final results as a proposition.

Proposition 2. *Let assumption* **2** *hold and either assumption* **1a** *or* **1c** *hold. Let* $\delta_n > 0$ *and* $V_{\delta_n} = \{(\alpha, \beta) : |\alpha - \alpha_0| < \delta_n/(1+M), |\beta - \beta_0| < \delta_n/(1+M)\}$. *Suppose* $G \subseteq \Theta$ *is compact and* $W_{1n} = \{(\alpha, \beta) : \alpha - \alpha_0 \geq \Delta_n, \beta \geq \beta_0\}$. *Then, the following inequalities hold.*

1. $\exists\ 0 < d < 1$ *and some constant* $R > 0$ *such that*

$$E\left\{\left(\int_{W_{1n}\cap G} \prod_{i=1}^{n} \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)} d\Pi(\alpha, \beta)\right)^d\right\} \leq e^{-dK\Delta_n^2 \sum_{i=1}^{n} I_{X_i > \epsilon_0}} \cdot e^{nd\delta_n} \cdot R^2/\delta_n^2$$

2.

$$\int_\Theta e^{\sum_{i=1}^{n} \log \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}} d\Pi(\alpha, \beta) \geq e^{-n\delta_n} \cdot \Pi(V_{\delta_n}).$$

Proof. From lemma 3 and lemma 4, we have

$$E\left\{\left(\int_{W_{1n}\cap G} \prod_{i=1}^{n} \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)} d\Pi(\alpha, \beta)\right)^d\right\} \leq e^{-dK\Delta_n^2 \sum_{i=1}^{n} I_{X_i > \epsilon_0}} \cdot e^{nd\delta_n} \cdot J(\delta_n).$$

The last step uses the propriety of prior from assumptions **1a** or **1c**. Further, we can choose $R > 0$ to be large enough such that G is contained within a square of area $R^2/(1+M)^2$ with $J(\delta_n) < R^2/\delta_n^2$.

(2) follows a similar argument as in the proof of lemma 6. □

Now, we prove the main theorems in the paper.

Proof of theorem 1.

We first prove it when Π is a proper prior. Taking $\Delta_n = \Delta, \delta_n = \delta$ for all n, lemma 5 shows that we can restrict to the case $W \cap G$ where $W = W_{1n}$ and G is compact. From proposition 2, we have that $\exists\ 0 < d < 1$ such that, for sufficiently large n,

$$E\left\{(\Pi(W \cap G|Y_1, \ldots, Y_n))^d\right\} \leq \frac{R^2}{\delta^2(\Pi(V_\delta))^d} \cdot e^{-dK\Delta^2 \sum_{i=1}^{n} I_{X_i > \epsilon_0}} \cdot e^{2nd\delta}.$$

Note that $\Pi(V_\delta) > 0$ by assumption (**1a**).So, by setting $L = \frac{K}{2} \liminf_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} I_{(X_i > \epsilon_0)}$ and choosing $\delta = \frac{L\Delta^2}{4}$, we get (for sufficiently large n),

$$E\left\{(\Pi(W_{1n} \cap G_1|Y_1, \ldots, Y_n))^d\right\} \leq C' e^{-\frac{ndL\Delta^2}{2}}, \text{for some constant } C'.$$

The R.H.S. of the above inequality is summable. Hence Markov's inequality along with the Borel - Cantelli lemma gives posterior consistency .

If we further make assumption **5**, posterior consistency generalizes easily to the case when the prior Π is improper but has a formal posterior and $\Pi(U) > 0$ for all neighborhoods U of (α_0, β_0). Recall the formal posterior density using a single observation is given by

$$\frac{f_{(1,\alpha,\beta,1)}(Y_1) d\Pi(\alpha, \beta, 1)}{\int_\Theta f_{(1,\alpha,\beta,1)}(Y_1) d\Pi(\alpha, \beta, 1)}$$

and for any set U, the probability given by the formal posterior is

$$\Pi(U|Y_1) = \frac{\int_U f_{(1,\alpha,\beta,1)}(Y_1) d\Pi(\alpha, \beta, 1)}{\int_\Theta f_{(1,\alpha,\beta,1)}(Y_1) d\Pi(\alpha, \beta, 1)}.$$

We argue that with P measure 1, the posterior density $\Pi(\cdot|Y_1)$ exists and satisfies assumption (**1a**). The set E, where the formal posterior is undefined has measure 0 under the "marginal" distribution of Y_1 and hence has measure 0 under some $f_{(1,\alpha,\beta,1)}$. Since the ALD densities are positive, this set also has 0 measure under all $f_{(1,\alpha,\beta,1)}$ and in particular when $\alpha = \alpha_0, \beta = \beta_0$. Assumption **5** ensures that $f_{(1,\alpha_0,\beta_0,1)}$ dominates p. Thus on E^c, a set of P measure 1, the formal posterior given by the above expression exists. Next, if U is a neighborhood of (α_0, β_0) then since $\Pi(U) > 0$ and the posterior density is positive everywhere, $\Pi(U|Y_1) > 0$ whenever $Y_1 \in E^c$.

□

Remark 2. *The result for improper priors is particularly interesting in view of theorem 1 of Yu and Moyeed (2001) where it is shown that the posterior based on ALD is always* *well defined for a flat prior w.r.t. (α, β) (i.e. $\Pi(\alpha, \beta) \propto 1$). Therefore, this would imply in particular that posterior consistency in theorem 1 will hold when the prior w.r.t. (α, β) is flat.*

Proof of theorem 2.
Note that

$$\Pi(U_n^c|Y_1, \ldots, Y_n) \to 0 \iff \Pi(W_{jn}|Y_1, \ldots, Y_n) \to 0 \ \forall\ j = 1, 2, \ldots, 8.$$

We will work with the case of W_{1n}. The argument is similar for $j = 2, \ldots, 8$. Further by lemma 6, it is enough to work with $W_{1n} \cap G_1$ where G_1 is compact. Let,

$$\Pi(W_{1n} \cap G_1|Y_1, \ldots, Y_n) = \frac{\int_{W_{1n} \cap G_1} e^{\sum_{i=1}^n \log \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}} d\Pi(\alpha, \beta)}{\int_\Theta e^{\sum_{i=1}^n \log \frac{f_{(i,\alpha,\beta,1)}(Y_i)}{f_{(i,\alpha_0,\beta_0,1)}(Y_i)}} d\Pi(\alpha, \beta)}.$$

Let $\delta_n \downarrow 0$ be a sequence (to be chosen later). Define

$$V_{\delta_n} = \{(\alpha, \beta) : |\alpha - \alpha_0| < \delta_n/(1 + M), |\beta - \beta_0| < \delta_n/(1 + M)\}.$$

Under assumption (**1c**), $\Pi(\cdot)$ has a density function that is positive and continuous in a neighborhood of (α_0, β_0). We can conclude that the density of Π is bounded away from 0 in a small neighborhood of (α_0, β_0) and hence for some constant $K > 0$,

$$\Pi(V_{\delta_n}) > K\delta_n^2 \text{ for sufficiently large } n. \tag{10}$$

Now, proposition 2 implies that there exists $0 < d < 1$ such that, for sufficiently large n,

$$\begin{aligned} &E\left\{(\Pi(W_{1n} \cap G_1|Y_1, \ldots, Y_n))^d\right\} \\ &\leq \frac{R^2}{K^d \delta_n^{2+2d}} \cdot e^{-dK\Delta^2 \sum_{i=1}^n I_{X_i > \epsilon_0}} \cdot e^{2nd\delta_n}. \end{aligned}$$

Further, by assumption **3** (ii), we have that for sufficiently large n,

$$\begin{aligned} &E\left\{(\Pi(W_{1n} \cap G_1|Y_1, \ldots, Y_n))^d\right\} \\ &\leq \frac{R^2}{K^d \delta_n^{2+2d}} \cdot e^{-dK\Delta_n^2 \sum_{i=1}^n I_{X_i > \epsilon_0}} \cdot e^{2nd\delta_n} \\ &\leq \frac{R^2}{K^d \delta_n^{2+2d}} \cdot e^{-\frac{n\Delta_n^2 dL}{2}} \cdot e^{2nd\delta_n} \\ &\text{where } L = \frac{K}{2} \liminf_{m\to\infty} \frac{1}{m} \sum_{i=1}^m I_{(X_i > \epsilon_0)}. \end{aligned}$$

By choosing $\delta_n = \frac{L\Delta_n^2}{4}$ (for sufficiently large n and some constant C'), we get

$$E\left\{(\Pi(W_{1n} \cap G_1|Y_1, \ldots, Y_n))^d\right\} \leq \frac{C'}{(\Delta_n^2)^{2+2d}} e^{-dL \cdot n\Delta_n^2}.$$

When $\Delta_n = Mn^{-\delta}$ for $0 < \delta < 1/2$, the R.H.S. of the above inequality is summable. Hence by Markov's inequality and the Borel-Cantelli lemma, we can reach conclusion (a) of theorem 2, which is $\Pi(W_{1n} \cap G_1|Y_1, \ldots, Y_n) \to\ 0\ a.s.[P]$.

Note: In the original paper, there was an error where denominator of R.H.S was written as $(n\Delta_n^2)^{2+2d}$ instead of $(\Delta_n^2)^{2+2d}$ (i.e. the factor n should not have been there).

With the above correction to R.H.S, the argument for part (a) Theorem 2 (i.e. when $\Delta_n = Mn^{-\delta}$, $0 < \delta < 1/2$) continues to hold. For part (b) of Theorem 2 (i.e. when $\Delta_n = M_n n^{-1/2}$), when $M_n^2 > C'' \log(n)$ for a sufficiently large C'' (to be precise $C'' > \frac{8(1+d)}{dL}$), we observe that R.H.S. of the above inequality converges to zero. Hence by Markov's inequality, we get

$$\Pi(W_{1n} \cap G_1|Y_1, \ldots, Y_n) \to\ 0 \ \text{ in probability } [P].$$

Note: The original paper had wrongly concluded that R.H.S converges to zero for any $M_n \to \infty$. The statement and proof are modified assuming a suitable rate of divergence for M_n.

□

6 Posterior Consistency with a Prior on σ

In applications involving Bayesian quantile regression using ALD, it is not uncommon to endow the parameter σ with a prior (e.g. Yue and Rue 2011; Hu et al. 2012). Since our focus is the estimation of a particular quantile of the true underlying distribution, the specific choice of σ parameter within the misspecified ALD has no direct interpretation. The advantages or disadvantages of using priors for σ are not well documented in literature and may be worthy of further research. However, in this paper, we seek to answer the question as to whether endowing a prior on σ would still preserve posterior consistency for (α, β). At a technical level, this provides an interesting example for the study of posterior consistency in the i.n.i.d. set up where the Kullback-Leibler divergence minimizing density from the specified family varies with i. There are two natural approaches to this problem. One way would be to work with a marginal likelihood obtained by integrating out the ALD density with respect to the prior on σ. Another approach would be to work out the consistency property of the combined vector (α, β, σ). We prefer the latter approach as it would allow us to exploit the form of the ALD density.

Let $\Pi(\cdot)$ be a prior on the parameters $(\alpha, \beta, \sigma) \in \Re^2 \times (0, \infty)$ which we write as $\Pi(\alpha, \beta|\sigma)\Pi(\sigma)$. We denote the support of $\Pi(\sigma)$ by Θ_σ and write Θ for the parameter space of (α, β). In misspecified models we would expect that the posterior distribution concentrates on a neighborhood of $f_{(i,\alpha^*,\beta^*,\sigma^*)}$ that has the minimum Kullback -Leibler divergence from the true density p_{0i}. It has been seen in proposition 1 that in the ALD case with a fixed σ, the minimum Kullback-Leibler divergence is attained at $\alpha = \alpha_0$ and $\beta = \beta_0$. It can be easily checked that the minimum KL is attained at (α_0, β_0) even if the parameter space is expanded to include σ. However, the parameter σ itself poses a challenge as the Kullback-Leibler minimizing value would depend on i. It turns out that the appropriate choice for the point of consistency for σ is given by

$$\sigma_0 = \arg\max_{\sigma \in \Theta_\sigma} \left\{ \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} E\left(\log f_{i,\alpha_0,\beta_0,\sigma}(Y_i)\right) \right\}. \tag{11}$$

Note that

$$\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} E\left(\log f_{i,\alpha_0,\beta_0,\sigma}(Y_i)\right) = \log\left(\frac{\tau(1-\tau)}{\sigma}\right) - \frac{C^*}{\sigma} \tag{12}$$

$$\text{where } C^* = \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} E(Z_i(\tau - I_{(Z_i \le 0)})).$$

It is easy to check that C^* and hence σ_0 is well defined if part (a) of assumption **4** is strengthened as below.

Assumption 4′: For $Z_i = Y_i - \alpha_0 - \beta_0 X_i$,

$$(a) \qquad \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} E\left(|Z_i|\right) \text{ and } \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} E\left(Z_i\right) \text{ are finite}$$

$$(b) \qquad \sum_{i=1}^{\infty} \frac{E\left(|Z_i|^2\right)}{i^2} < \infty.$$

We consider neighborhoods of the form $U = U_1 \times U_2$, where $U_1 = \{(\alpha, \beta) \in \Theta : |\alpha - \alpha_0| < \Delta, |\beta - \beta_0| < \Delta\}$ and $U_2 = \{\sigma \in \Theta_\sigma : |\sigma - \sigma_0| < \epsilon(\Delta)\}$. Note that the neighborhood of σ is expressed in terms of a specific monotonic function $\epsilon(\cdot)$, such that $\epsilon(\delta) \downarrow 0$ as $\delta \downarrow 0$. This is of course without loss of generality and is done to simplify the arguments. The function $\epsilon(\cdot)$ is defined as follows. Let

$$\eta(\sigma) = \log\left(\frac{\sigma_0}{\sigma}\right) - C^*\left(\frac{1}{\sigma} - \frac{1}{\sigma_0}\right)$$

where C^* is as in equation (12). Note that $\eta(\sigma)$ takes the value 0 at $\sigma = \sigma_0$ and is decreasing on either side of σ_0. Hence we can define a monotonic function $\epsilon(\delta)$ such that

$$|\sigma - \sigma_0| > \epsilon(\delta) \iff \eta(\sigma) < -\delta. \tag{13}$$

Our interest is in establishing convergence of the posterior probability, which we write as follows.

$$\Pi\left((U_1 \times U_2)^c | Y_1, Y_2, \dots, Y_n\right) = \frac{\int_{(U_1\times U_2)^c} \prod_{i=1}^n \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)} d\Pi(\alpha,\beta,\sigma)}{\int_{\Theta\times\Theta_\sigma} \prod_{i=1}^n \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)} d\Pi(\alpha,\beta,\sigma)}. \tag{14}$$

The assumption on the prior needs to be modified to include the σ parameter as follows.

> **Assumption (1′a)**: $\Pi(\cdot)$ is a proper prior and for any σ' in the support of $\Pi(\sigma)$, every open neighborhood of (α_0, β_0) has positive $\Pi(\cdot,\cdot|\sigma')$ measure.

Recall that the key to establishing the main theorems was proposition 2. Hence, we just state and outline the proof of a proposition analogous to proposition 2. Then consistency for the case involving a prior on σ would follow. For simplicity, we restrict to a compact subset G of Θ and to Θ_σ of the form $[\sigma_1, \sigma_2]$ where $0 < \sigma_1 \le \sigma_2 < \infty$. As in lemma 6, one can construct a compact set G outside of which the posterior probability goes to zero. Similarly, for the case $\Theta_\sigma = (0, \infty)$, the idea would be to construct a compact interval outside of which the posterior probability goes to zero.

Before stating the proposition it helps to note the following simple facts. Firstly, we have the following inequality for $0 < d < 1$.

$$\begin{aligned}
& E\left\{\left(\int_{((U_1\times U_2)^c\cap(G\times\Theta_\sigma))} \prod_{i=1}^n \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)} d\Pi(\alpha,\beta)\right)^d\right\} \\
\le\ & E\left\{\left(\int_{(U_1^c\cap G)\times\Theta_\sigma} \prod_{i=1}^n \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)} d\Pi(\alpha,\beta)\right)^d\right\} \\
& + E\left\{\left(\int_{G\times(U_2^c\cap\Theta_\sigma)} \prod_{i=1}^n \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)} d\Pi(\alpha,\beta)\right)^d\right\}.
\end{aligned}$$

Therefore, to obtain a bound on the L.H.S., we can focus on each of the terms on the R.H.S. of the above inequality. Secondly, as before, we can write $U_1^c = \cup_{j=1}^8 W_j$. Hence, we can obtain a bound on the first term of the R.H.S. by looking at each of the sets W_j separately. Now, we state proposition 3, which is analogous to proposition 2. Again, we state it for the case of W_1. Arguments are similar for other W_j. The parameter κ mentioned in proposition 3 is arbitrary. Its role is to help in choosing an appropriate $\delta > 0$ as done in the proof of theorem 1.

Proposition 3. *Let $G \subseteq \Theta$ be compact, $W_1 = \{(\alpha,\beta) : \alpha - \alpha_0 \ge \Delta, \beta \ge \beta_0\}$, $\Theta_\sigma = [\sigma_1, \sigma_2]$, $0 < \sigma_1 \le \sigma_2 < \infty$. Let assumptions* **1′a**, **2** *and* **4′** *hold. Let U_1 and U_2 be as defined above. Let $\delta > 0$ and $V_\delta = \{(\alpha,\beta) : |\alpha - \alpha_0| < \delta/(1+M), |\beta - \beta_0| < \delta/(1+M), |\sigma - \sigma_0| < \epsilon(\delta)\}$, where the function $\epsilon(\cdot)$ is as defined in (13). Also, let $\kappa > 0$ be arbitrary. Then, the following inequalities hold.*

1. $\exists\ 0 < d < 1$ *and some constant* $R > 0$ *such that*

$$(a) \qquad E\left\{\left(\int_{(W_1\cap G)\times\Theta_\sigma}\prod_{i=1}^{n}\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}d\Pi(\alpha,\beta,\sigma)\right)^d\right\}$$
$$\le e^{\frac{d\sum_{i=1}^{n}B_i^{(1)}}{\sigma_2}}\cdot e^{\frac{nd\delta}{\sigma_1}}\cdot e^{nd\kappa\delta}\cdot R^2/\delta^2$$
$$\text{where } B_i^{(1)} = -\Delta\cdot P\left(0 < Z_i < \frac{\Delta}{2}\right)\cdot I_{(X_i>\epsilon_0)}.$$

$$(b) \qquad E\left\{\left(\int_{G\times(U_2^c\cap\Theta_\sigma)}\prod_{i=1}^{n}\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}d\Pi(\alpha,\beta)\right)^d\right\}$$
$$\le e^{-nd\Delta}\cdot e^{nd\kappa\delta}.$$

2.

$$\int_{\Theta\times\Theta_\sigma} e^{\sum_{i=1}^{n}\log\frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}}d\Pi(\alpha,\beta) \ge e^{-\frac{n\delta}{\sigma_1}}\cdot e^{-n\delta}\cdot e^{-n\kappa\delta}\cdot\Pi(V_\delta).$$

Outline of the Proof: Note that

$$\frac{1}{m}\sum_{i=1}^{m}\log\left(\frac{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}\right) = \log\frac{\sigma_0}{\sigma} - \left(\frac{1}{\sigma}-\frac{1}{\sigma_0}\right)\cdot\frac{1}{m}\sum_{i=1}^{m}\left(Z_i\cdot(\tau - I_{Z_i\le 0})\right).$$

Assumption **4′** along with Kolmogorov's SLLN for independent random variables would imply

$$\frac{1}{n}\sum_{i=1}^{n}\left(Z_i\cdot(\tau - I_{Z_i\le 0}) - E\left\{Z_i\cdot(\tau - I_{Z_i\le 0})\right\}\right) \to\ 0\ a.s.\ [P].$$

Since $\sigma\in[\sigma_1,\sigma_2]$, we can conclude that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(\frac{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}\right) - E\left\{\log\left(\frac{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}\right)\right\}\right)$$
$$\to\ 0\ a.s.\ [P]\ \text{(uniformly in } \sigma).$$

Therefore, for any given $\kappa > 0$, for sufficiently large n, we have

$$\left|\sum_{i=1}^{n}\log\left(\frac{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}\right) - n\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}E\left[\log\left(\frac{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}\right)\right]\right| \qquad (15)$$
$$< n\kappa\delta.$$

For parts 1(a) and 1(b) of the result, from equation (15) we note that

$$\begin{aligned}
&\prod_{i=1}^{n} \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)} \\
= \;& e^{\sum_{i=1}^{n} \log \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}} \cdot e^{\sum_{i=1}^{n} \log \frac{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}} \\
\leq \;& e^{\sum_{i=1}^{n} \log \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}} \cdot e^{n \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} E\left[\log\left(\frac{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}\right)\right]} \cdot e^{n\kappa\delta}.
\end{aligned}$$

Part 1(a) can then be derived by first noting from definition (11) that

$$\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} E\left[\log\left(\frac{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}\right)\right] \leq 0$$

and then using a similar approach that leads up to proposition 2 for a fixed σ, along with the fact that $\sigma_1 \leq \sigma \leq \sigma_2$.

For Part 1(b), we note from the definition of $\epsilon(\cdot)$ in (13) that for $\sigma \in U_2^c$,

$$\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} E\left[\log\left(\frac{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}\right)\right] \leq -\Delta.$$

Then the inequality follows by choosing an appropriate $d > 0$, using an approach similar to lemmas 3 and 4, so that

$$E\left\{\left(\int_{(W_1\cap G)} \prod_{i=1}^{n} \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)} d\Pi(\alpha,\beta|\sigma)\right)^{d}\right\} \leq 1 \quad \forall\; \sigma \in [\sigma_1, \sigma_2].$$

For part 2 note that for $(\alpha, \beta, \sigma) \in V_\delta$, for sufficiently large n,

$$\begin{aligned}
&\prod_{i=1}^{n} \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)} \\
= \;& e^{\sum_{i=1}^{n} \log \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}} \cdot e^{\sum_{i=1}^{n} \log \frac{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}} \\
\geq \;& e^{\sum_{i=1}^{n} \log \frac{f_{(i,\alpha,\beta,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}} \cdot e^{n \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} E\left[\log\left(\frac{f_{(i,\alpha_0,\beta_0,\sigma)}(Y_i)}{f_{(i,\alpha_0,\beta_0,\sigma_0)}(Y_i)}\right)\right]} \cdot e^{-n\kappa\delta} \\
\geq \;& e^{-\frac{n\delta}{\sigma_1}} \cdot e^{-n\delta} \cdot e^{-n\kappa\delta}.
\end{aligned}$$

The first term in the product of the last expression uses part (d) of lemma 1 and the fact that $\sigma \in [\sigma_1, \sigma_2]$. The second term follows by the definition of $\epsilon(\cdot)$ as given in (13) and the third term from equation (15). □

7 Conclusion

The main contribution of this paper has been to provide an asymptotic justification for assuming ALD for the response in Bayesian Quantile Regression, although it could be a misspecification. The method is justified under some reasonable conditions on the covariates, the prior and the underlying true distribution. This is significant given the fact that this approach has been used extensively since the work of Yu and Moyeed (2001), but to our knowledge has only been checked empirically. We find that the use of ALD works for a variety of possibilities for the true likelihood, including location models, scale models, location-scale models and in fact any case where the probabilities and expectations appearing in the assumptions **1 to 4** are nicely behaved. ALD has the nice property (proposition 1) that the Kullback-Leibler divergence is minimized at the true values of the regression parameters. This is not true in general for any distribution. For example, if instead of ALD, we use a normal density function whose location is adjusted to make the τ^{th} quantile $= \alpha_0 + \mathbf{X}^T \beta_0$, this property is sometimes violated depending on the true underlying distribution. In such cases, Bayesian quantile regression based on such a normal likelihood does not necessarily lead to correct results.

References

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (2003). "Posterior consistency for semi-parametric regression problems." *Bernoulli*, 9(2): 291–312.

Berk, R. H. (1966). "Limiting behavior of posterior distributions when the model is incorrect." *The Annals of Mathematical Statistics*, 37(1): 51–58.

Bunke, O. and Milhaud, X. (1998). "Asymptotic behavior of Bayes estimates under possibly incorrect models." *The Annals of Statistics*, 26(2): 617–644.

Ghosal, S. and van der Vaart, A. W. (2007). "Convergence rates of posterior distributions for noniid observations." *The Annals of Statistics*, 35(1): 192–223.

Hu, Y., Gramacy, R. B., and Lian, H. (2012). "Bayesian quantile regression for single-index models." *Statistics and Computing*, 1–18.

Kleijn, B. J. K. and van der Vaart, A. W. (2006). "Misspecification in infinite-dimensional Bayesian statistics." *The Annals of Statistics*, 34(2): 837–877.

Kleijn, B. J. K. and van der Vaart, A. W. (2012). "The Bernstein-Von-Mises theorem under misspecification." *Electronic Journal of Statistics*, 6: 354–381.

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

Koenker, R. and Bassett Jr, G. (1978). "Regression quantiles." *Econometrica*, 46(1): 33–50.

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

Shalizi, C. R. (2009). "Dynamics of Bayesian updating with dependent data and misspecified models." *Electronic Journal of Statistics*, 3: 1039–1074.

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

Yu, K., van Kerm, P., and Zhang, J. (2005). "Bayesian quantile regression: an application to the wage distribution in 1990s Britain." *Sankhyā*, 67(2): 359–377.

Yu, K. and Zhang, J. (2005). "A three-parameter asymmetric Laplace distribution and its extension." *Communications in Statistics-Theory and Methods*, 34: 1867–1879.

Yue, Y. R. and Rue, H. (2011). "Bayesian inference for additive mixed quantile regression models." *Computational Statistics and Data Analysis*, 55: 84–96.

Acknowledgments

The authors thank the referees and the Associate editor for their insightful comments and suggestions.

We thank Michael Guggisberg, a PhD candidate at University of California, Irvine for pointing out the error in the proof of Lemma 4 in the original paper.