Horseshoe Prior Bayesian Quantile Regression

David Kohns†
Department of Economics, Heriot-Watt University
and
Tibor Szendrei‡
Department of Economics, Heriot-Watt University

June 16, 2020

Abstract

This paper extends the horseshoe prior of Carvalho et al. (2010) to the Bayesian quantile regression (HS-BQR) and provides a fast sampling algorithm that speeds up computation significantly in high dimensions. The performance of the HS-BQR is tested on large scale Monte Carlo simulations and an empirical application relevant to macroeconomics. The Monte Carlo design considers several sparsity structures (sparse, dense, block) and error structures (i.i.d. errors and heteroskedastic errors). A number of LASSO based estimators (frequentist and Bayesian) are pitted against the HS-BQR to better gauge the performance of the method on the different designs. The HS-BQR yields as just as good, or better performance than the other estimators considered when evaluated using coefficient bias and forecast error. We find that the HS-BQR is particularly potent in sparse designs and when estimating extreme quantiles. The simulations also highlight how the high dimensional quantile estimators fail to correctly identify the quantile function of the variables when both location and scale effects are present. In the empirical application, in which we evaluate forecast densities of US inflation, the HS-BQR provides well calibrated forecast densities whose individual quantiles, have the highest pseudo $R^2$, highlighting its potential for Value-at-Risk estimation.

Keywords: Global-Local Priors, Shrinkage, Machine Learning, Quantile Regression

JEL codes: C01, C11, C21, C53

*The authors would like to thank Arnab Bhattacharjee, Mark Schaffer, Atanas Christev, Aubrey Poon, Gary Koop and all participants of the Scottish Graduate Program in Economics conference in Crieff for their invaluable feedback.

†Email: dek1@hw.ac.uk
‡Email: ts136@hw.ac.uk
1 Introduction

Quantile regression has been an important tool in the econometricians’ toolkit when estimating heterogenous effects across the conditional response distribution since the seminal work of Koenker and Bassett (1978). In contrast to least squares methods, it estimates directly quantiles of the dependent variables’ conditional distribution which allows for richer inference than solely focusing on estimating the conditional mean. Since this approach does not put any symmetry restrictions on the conditional distribution, it has proven to be particularly useful in the macroeconomics and finance literature, where the quantile regression approach is used to compute probabilities of recessions (e.g. Adrian et al. (2019) and Giglio et al. (2016)), risk measures such as VaR (Engle and Manganelli, 2004; Chen et al., 2012; Taylor, 2019) which financial institutions are obliged to report in adherence with the Basel Committee on Banking Supervision (see Basel II), and more generally forecast density construction (e.g. Korobilis (2017) and Carriero et al. (2020)). The Bayesian approach to estimating the quantile regression has been used extensively for these purposes as it has been shown to provide more robust tail inferences compared to frequentist estimation (Chen et al., 2012).

Especially for modern forecast density applications in macroeconomics, it has become customary to use some form of dimension reduction, such as factor based (e.g. Stock and Watson (2002) or Bai and Ng (2008)) or sparsity based reduction (e.g. Tibshirani (1996)) as the dimensionality of the feature space renders the traditional least squares or quantile regression approach imprecise or infeasible. While there are well established frequentist sparse high-dimensional applications and theory for quantile regression (see Belloni and Chernozhukov (2011) and Wang (2017) for an overview), high-dimensional Bayesian quantile regression is less established. The Bayesian quantile regression approach, as popularised by Yu and Moyeed (2001), is based on the asymmetric Laplace likelihood (ALL), which has a special connection to the frequentist quantile regression solution, in that its maximum likelihood estimates are equivalent to the quantile regression with a check-loss function (Koenker, 2005). A hurdle in the Bayesian literature has been that the ALL based methods, result in improper posteriors with any but non-informative or exponential Laplace priors, where the latter results in the popular Bayesian Lasso quantile regression (Li et al., 2010; Alhamzawi and Yu, 2013; Alhamzawi et al., 2012; Chen et al., 2013). The broader Bayesian sparsity literature has shown, however, that global-local shrinkage priors

1Reason being that numerical integration through MCMC simulation often provides more robust inferences for extreme quantiles compared to numerical optimisation routines.
such as the horseshoe (Carvalho et al., 2010) and Dirichlet-Laplace prior (Bhattacharya et al., 2016) offer asymptotic as well as computational advantages over the former methods (Bhadra et al., 2019). These methods have not yet been considered for the Bayesian quantile regression. The aim of this paper is to bridge this gap and extend the global local prior to quantile regression.

This paper’s contribution is threefold. First, we derive the horseshoe prior of Carvalho et al. (2010) for the Bayesian quantile regression framework (BQR) of Yu and Moyeed (2001). Second, we develop an efficient posterior sampler for the regression coefficients based on data augmentation akin to Bhattacharya et al. (2016) which speeds up computation significantly for high dimensional quantile problems. Third, we show in large scale Monte Carlo studies that the proposed Horseshoe BQR (HS-BQR) provides more stable and at worst, similar performance compared to Bayesian lasso quantile regression methods as well as the widely popular L1-QR of Belloni and Chernozhukov (2011) in terms of coefficient bias and forecast accuracy. We find that, particularly, tails of the distributions are consistently better estimated by the HS-BQR.

For our empirical application, we construct forecast densities of quarterly US inflation. We show that the HS-BQR produces better calibrated forecast densities compared to frequentist and Bayesian alternatives. Importantly, this yields better calibrated risk measures as the $p^{th}$ fitted quantile can be interpreted as the $p^{th}$ VaR. The framework provided in this paper has the additional advantage that the derived algorithms can be directly applied to other global-local priors$^2$ that can be expressed as scale mixture of normals.

In what follows, we will first review the generic quantile regression framework. Then, we will present the main results in the Bayesian quantile as well as shrinkage literature which have motivated the form of the model. Following this, we will develop posteriors as well as the sampling algorithm. Lastly, we will provide evidence from Monte Carlo simulations and an empirical application of the favourable performance of the HS-BQR compared to alternative methods. We conclude with further generalisations of the algorithms provided and a discussion of our results.

$^2$For an overview of GL priors see Polson and Scott (2010).
2 Methodology

2.1 Quantile Regression

Quantile regression is a framework to extend econometric analysis beyond the conditional mean (Koenker and Bassett, 1978). The value added of quantile regression is that it does not exclusively focusing on the mean (as is the case with OLS) but allows for inference about the conditional distribution of the variable of interest directly, through modeling of its conditional quantiles.

Taking the linear model \( Y = X\beta + \epsilon \) as our starting point, the conditional quantile function of \( Y \) can be defined as

\[
Q_p(Y \mid X) = X\beta(p)
\]  

where \( p \in (0, 1) \), \( X \) is a \( T \times K \) matrix of covariates, \( \beta(p) \) is a \( K \times 1 \) vector of quantile specific regression coefficients, and \( \epsilon \) is \( T \times 1 \) vector of residuals which follow some unspecified distribution. While it is possible to estimate an infinite amount of quantiles, in practice one only estimates a finite number of quantiles which are of interest, as the data requirements of QR increase roughly linearly with the number of quantiles being estimated (Buchinsky, 1998; Davino et al., 2013).

The objective function of the quantile regression can be represented by the following minimisation problem:

\[
\min_{\beta} \sum_{t=1}^{n} \rho_p(y_t - x'_t \beta)
\]  

where \( \rho_p(.) \) is a loss function with the following form:

\[
\rho_p(y) = \begin{cases} 
  p - I(y < 0) & \text{if } y < 0 \\
  (1-p)I(y \leq 0) + pI(y > 0) & \text{if } y \geq 0 
\end{cases}
\]  

and \( I(.) \) is an indicator function taking on a value of 0 or 1 depending on whether the condition is satisfied. Equation (3) determines the weight each observation receives in the minimisation problem. It is often referred to as the check-loss function due to the weight profile it assigns depending on the quantile being estimated (Koenker, 2005). Note how \((y_t - x'_t \beta)\) is the residual of a regression model. The interpretation of the coefficients is thus similar to the classical
regression case: $\beta_j(p)$ is the rate of change of the $p^{th}$ quantile of the dependent variable’s distribution to a unit change in the $j^{th}$ regressor.

Unlike in classical regression analysis, QR does not make any parametric assumption about $\epsilon$ (Koenker, 2005) which allows for rich, non-symmetric inference about the conditional distribution of $Y$.

2.2 The Bayesian Quantile Regression

Three key findings enable extending the quantile regression to the horseshoe prior. (i) Yu and Moyeed (2001) have shown that maximising an asymmetric Laplace likelihood under mild conditions is equivalent to minimising the standard frequentist quantile loss function as in (3). (ii) Kozumi and Kobayashi (2011) showed that while conjugate Normal-Inverse-Gamma (N-IG) priors result in intractable posteriors with the asymmetric Laplace likelihood, independent N-IG priors paired with a latent data representation of the AL likelihood, result in tractable conditional posteriors which can be sampled from a straight forward Gibbs sampler. (iii) Khare and Hobert (2012) show that the Markov chain of this sampler is geometrically ergodic and also valid in $K >> T$ settings which gives theoretical justification to apply this sampler to high dimensional settings.

In particular, we assume the quantile regression model (1) and a fixed design $X$. As shown by Yu and Moyeed (2001), $\beta(p)$ can be obtained as the maximum likelihood estimator for $\beta$ under the fully parametric model $y_t = x_t' \beta + \epsilon$ where $\{\epsilon_t\}_{t=1}^T$ are assumed i.i.d. with common density given by

$$\{\epsilon\}_{t=1}^T \sim g(\epsilon; p) = \frac{p(1-p)}{\sigma} e^{(1-p)\epsilon/\sigma} I_{\mathbb{R}_-} (\epsilon) + e^{-p\epsilon/\sigma} I_{\mathbb{R}_+} (\epsilon)$$ (4)

where $\mathbb{R}_+ := (0, \infty)$ and $\mathbb{R}_- := (-\infty, 0]$. The errors follow an asymmetric Laplace density with the $p^{th}$ quantile equal to zero. Assuming the linear model as above with error density (4), the joint likelihood $f(Y|\beta, \sigma)$ becomes:

$$f(Y|\beta, \sigma) = (p^T)(1-p)^T \sigma^{-T} \prod_{t=1}^T \left[ e^{(1-p)(y_t - x_t' \beta)/\sigma} I_{\mathbb{R}_-} (y_t - x_t' \beta) + e^{-p(y_t - x_t' \beta)/\sigma} I_{\mathbb{R}_+} (y_t - x_t' \beta) \right]$$ (5)

It is apparent that using any non-trivial prior for $(\beta, \sigma)$, will result in an intractable posterior which will necessitate inefficient accept and reject sampling algorithms (Yu and Moyeed, 2001). However, Kozumi and Kobayashi (2011) showed using the mixture representation of the
asymmetric Laplace distribution provided by Kotz et al. (2012), that the likelihood in (5) can be obtained by formulating the error process as:

$$
\epsilon = \sigma \theta z + \sigma \tau \sqrt{z} u
$$

(6)

Where \( z \sim \exp(1) \), \( u \sim N(0, 1) \), and \( \theta = \frac{1-2p}{\rho(1-p)} \) and \( \tau^2 = \frac{2}{\rho(1-p)} \) are deterministic quantile specific parameters. Let \( \theta = \theta(p) \) and \( \tau^2 = \tau^2(p) \) be defined as above and let the tuples \( \{ (y_t, z_t) \}_{t=1}^T \) be independent random pairs. Now, to simplify the Gibbs sampler, we instead assume \( z_t \sim \exp(\sigma) \) such that given \( z_t \), \( y_t \) is in normally distributed as \( y_t|z_t \sim N(x_t' \beta + \theta z_t, z_t \sigma^2) \). The joint density of \( Y|Z \) is given by:

$$
\begin{align*}
 f(Y|\beta, \sigma, Z) \propto \left( \prod_{t=1}^T \frac{1}{\sqrt{z_t}} \right) \times \exp \left[ -\frac{1}{2} \sum_{t=1}^T \frac{(y_t - x_t' \beta - \theta z_t)^2}{\sigma \tau^2 z_t} \right]
\end{align*}
$$

(7)

2.3 Global-Local Priors

In order to identify the posterior of a large dimensional coefficient vector in small samples, informative priors are needed. Ideally, they are able to separate noise variables from signals such that the noise is shrunk towards zero and signals attain their unrestricted parameter values. In the frequentist setting, this is usually achieved through penalised likelihoods which force variables to threshold to 0 or not. In the Bayesian approach, it is important to note that the assumption about sparsity is different in that proper prior distributions have non-zero probability over sparse and non-sparse regions in the posterior (Batencourt, 2018). In order, therefore, to achieve the desired separation between shrunk and unshrunk variables, the amount of shrinkage on a 0-1 scale should approach a bi-modal distribution where most of the mass is on 0 and 1 respectively. The horseshoe prior of Carvalho et al. (2010) achieves such a shrinkage profile, while double-exponential based lasso priors do not.

To illustrate, we use the linear regression model as above, but for simplicity assume a Gaussian error term. As we are leveraging a conditionally normal likelihood for the Bayesian quantile regression, the following discussion also holds for the Bayesian quantile regression. Global-local priors as defined by Polson and Scott (2010) are a scale mixture of normals prior on the coefficient vector and otherwise unspecified priors for the variance parameters which we
write as:

\[ \sigma^2 \sim \pi(\sigma^2) d\sigma^2 \]
\[ \beta_j | \lambda_j^2, \nu^2, \sigma^2 \sim N(0, \lambda_j^2 \nu^2 \sigma^2), j \in (1, \cdots, K) \]
\[ \lambda_j^2 \sim \pi(\lambda_j^2) d\lambda_j^2, j \in (1, \cdots, K) \]
\[ \nu^2 \sim \pi(\nu^2) d\nu^2 \]

The idea of this family of shrinkage priors is that the global scale-prior \( \nu^2 \), controls the overall shrinkage applied to the regression, while the local scale \( \lambda_j^2 \) allows for the local possibility of regressors to escape shrinkage when they have large effects on the response. It can be shown that the conditional posterior is normal with the following mean and variance:

\[
p(\beta|\Lambda, \nu, \sigma^2, X) = N(\beta | \bar{\beta}, \Sigma),
\]
\[
\bar{\beta} = \nu^2 \Lambda(\nu^2 \Lambda + \sigma^2 (X'X)^{-1})^{-1} \hat{\beta},
\]
\[
\Sigma = (\nu^{-2} \Lambda^{-1} + \frac{1}{\sigma^2} X'X)^{-1}
\]

where \( \Lambda = diag(\lambda_1^2, \cdots, \lambda_K^2) \) and \( \hat{\beta} = (X'X)^{-1}X'\bar{y} \) is the maximum likelihood solution (given that the inverse exists). Assuming for simplicity that the covariates are uncorrelated with zero mean and variances \( Var(x_j) = s_j^2 \), then \( X'X \approx ndiag(s_1^2, \cdots, s_K^2) \) such that the mean of the posterior coefficient vector can be approximated as:

\[
\bar{\beta}_j = (1 - \kappa_j) \hat{\beta}_j
\]

and

\[
\kappa_j = \frac{1}{1 + n\sigma^{-2}\nu^2 s_j^2 \lambda_j^2}
\]

\( \kappa_j \) is called the shrinkage factor which is bounded to be between 0 and 1 depending on the values of the scales, \( \nu \) and \( \lambda \). This result is independent of the prior used for \( \nu \) and \( \lambda \). Their priors, however, determine the shape of the implicit prior on the shrinkage factor. The vocal point of this paper, the horse shoe prior of Carvalho et al. (2011), employs two half Cauchy distributions for \( \lambda \) and \( \tau \):

\[
\lambda_j^2 \sim C_+(0,1)
\]
\[\nu^2 \sim C_+(0,1)\]
The benefit of using two fat tailed distributions as the Cauchy for scale parameters is that each scale has most mass near 0, which entails shrinkage, however, has fat enough tails to allow certain features to escape the pull toward zero. It can be shown that these priors imply a Beta(0.5,0.5) distribution on the shrinkage factors \( \text{Carvalho et al.} \ 2010 \). When \( \tau \) and \( \lambda \) are strongly identified, this prior results in complete or no shrinkage for each coefficient in the limit, as can be visually confirmed from figure [1], left panel.

The same cannot be said for the lasso prior. This prior originated from the fact that the L1 norm penalised least squares objective function of \( \text{Tibshirani} \ 1996 \) is equivalent to the maximum a posteriori (MAP) value of a normal linear regression model with independent double-exponential priors on the regression coefficients and thus named the Bayesian lasso. Cast into the global local form of equation (8), the prior takes the following form\(^3\)

\[
\begin{align*}
\pi(\lambda_j^2) &= \frac{u^2}{2} e^{-u^2\lambda_j^2/2} \\
\pi(\sigma^2) &\propto \sigma^{-2}
\end{align*}
\]

Under this prior, the shrinkage coefficient can be shown to have the shape as in the right panel of figure [1]. The double exponential has the unfortunate trait that large signals can escape shrinkage, but noise variables are not shrunk aggressively enough. This will result in too little shrinkage in large dimensional problems with many noise variables.

2.4 Horseshoe Bayesian Quantile Regression

In order to generalise the horseshoe prior to the AL likelihood in (7), the HS prior needs to be formulated under the assumption of independence between the \( \beta \) and \( \sigma \) prior so that the posterior takes a conditionally normal form \( \text{Kozumi and Kobayashi} \ 2011 \). While prior independence between regression coefficients and error variance might seem a strong assumption, \( \text{Moran et al.} \ 2018 \) have shown that in high-dimensional settings, the independence assumption aids inference of the error variance. This is due to the fact that conjugate priors act mathematically as additional observations which artificially bias the error variances downwards when \( K >> T \)

\(^3\)Note here that that an auxiliary variable \( u \) is introduced which after integration yields the desired double exponential Laplace density on the coefficient vector (see \( \text{Park and Casella} \ 2008 \)).
Figure 1: Distribution of $\kappa_j$, the shrinkage coefficient implied by (a) the horseshoe prior and (b) the Lasso prior.

The general independent GL prior takes the following hierarchical form:

$$
\sigma^2 \sim \pi(\sigma^2) d\sigma^2
$$

$$
\beta_j | \lambda_j^2, \nu^2 \sim N(0, \lambda_j^2 \nu^2), j \in (1, \cdots, K)
$$

$$
\lambda_j^2 \sim \pi(\lambda_j^2) d\lambda_j^2, j \in (1, \cdots, K)
$$

$$
\nu^2 \sim \pi(\nu^2) d\nu^2
$$

(13)

As recommended by Gelman et al. (2006), we select a weakly informative inverse Gamma distribution as the prior for the error variance $\sigma^2$ and two independently distributed half Cauchy distributions on the positive support for the scale parameters of the $\beta$ prior:

$$
\sigma^2 \sim IG(a,b)
$$

(14)

$$
\lambda_j^2 \sim C_+(0, 1)
$$

(15)

$$
\nu^2 \sim C_+(0, 1)
$$

(16)

Under the above priors, the posteriors take the following form:

$$
\beta | \sigma, \lambda^2, \nu^2, \theta, \tau, X, Y, Z \sim N(\tilde{\beta}, \tilde{V}),
$$

where $\tilde{V} = (X'UX + V)^{-1}, U = diag(1/2^2\sigma)$ and $\tilde{V} = \nu^2 diag(\lambda_1^2, \cdots, \lambda_K^2)$. Using the usual Bayesian computations, $\tilde{\beta}$ is defined as $\tilde{\beta} = (X'UX + \tilde{V}^{-1})^{-1}(X'Uy + \tilde{V}^{-1}\beta)$ with $\beta = 0_K$.

The conditional posterior of the scale parameter is given by

$$
\sigma^2 | \beta, \lambda^2, \nu^2, \theta, \tau, X, Y, Z \sim IG(\bar{a}, \bar{b})
$$

(18)
and \( \bar{a} = a + \frac{3\tau}{2}, \bar{b} = b + \sum_{t=1}^{T} \frac{(y_t-x_t')^2}{2\tau + \sigma^2} + \sum_{t=1}^{T} z_t. \)

As the conditional posteriors for the scales \((\lambda, \nu)\) follow Cauchy distributions [Carvalho et al. 2010], and therefore no moments exist which would enable sampling, the literature has proposed Gibbs samplers which rely either on slice sampling [Polson et al. 2014] or mixture representations [Makalic and Schmidt 2015]. However, both rely on conjugate formulations. We use the fact that the posterior distribution of \( \lambda_j \) conditional on \( \nu \) remains independent of all other parameters by assumption, to formulate a block slice sampling algorithm for \( \lambda_j = (\lambda_1, \cdots, \lambda_K)' \) akin to Polson et al. (2014). Slice sampling generates pseudo-random numbers from any distribution function \( f(y) \) by sampling uniformly from horizontal slices through the PDF. Advantages of the algorithm include its simplicity, that it involves no rejections, and that it requires no external parameters to be set. Define \( \eta_j = 1/\lambda_j^2 \) and \( \mu_j = \beta_j/\nu \). The conditional posterior distribution of \( \eta_j \), given all other parameters is given by

\[
p(\eta_j|\nu, \sigma, \mu_j, X, Y, Z) \propto \exp\left\{ -\frac{\mu_j^2}{2} \eta_j \right\} \frac{1}{1 + \eta_j} \tag{19}
\]

Slice sampling can now be implemented to draw from (14):

1. Sample \((u_j|\eta_j)\) uniformly in the interval \((0, 1/(1 + \eta_j))\).

2. Sample \(\eta_j|\mu_j, u_j \sim Ex(2/\mu_j^2)\) from an exponential density truncated to have zero probability outside \((0, (1 - u_j)/u_j))\).

Taking the inverse square root of the sample of 2., one receives back the estimate for \( \lambda_j \). By replacing \( \eta = 1/\nu \) and \( \mu_j^2 \) by \( \sum_{j=1}^{K} (\beta_j/\lambda_j)^2/2 \), \( \nu \) can be sampled in a similar manner.

Finally, since the latent \( z_t \) are sampled independently, the conditional posterior follows the reciprocal of the inverse Gaussian:

\[
z_t|\beta, \sigma, \lambda^2, \nu^2, \theta, \tau, X, Y \sim I - G(\bar{c}_t, \bar{d}_t) \tag{20}
\]

Where I-G stands for the inverse Gaussian distribution with location and rate parameters respectively, \( \bar{c}_t = \frac{\sqrt{\theta^2 + 2\sigma^2}}{|y_t - x_t'\beta|} \) and \( \bar{d}_t = \frac{\theta^2 + 2\sigma^2}{\sigma\tau^2} \).
2.5 Gibbs Sampler

With these conditional posteriors at hand, we utilise a standard Gibbs sampler. The dynamics of the Markov chain \( \{(\beta_m, \sigma_m, \lambda^2_m, \nu^2_m, z_m)\}_{m=0}^{\infty} \) are implicitly defined through the following steps

1. Draw \( Z \sim \pi(.|\beta, \sigma, \lambda^2, \nu^2, \theta, \tau, X, Y) \) from \( I-G(\alpha_t, d_t) \) for all \( t \) and call the \( T \times 1 \) vector \( z_{n+1} \)
2. Draw \( \sigma_{n+1} \sim \pi(.|\beta, \lambda^2, \nu^2, \theta, \tau, X, Y, z_{n+1}) \) from \( IG(\alpha, \tilde{b}) \)
3. Draw \( \beta_{n+1} \sim \pi(.|\sigma_{n+1}, \lambda^2, \nu^2, \theta, \tau, X, Y, z_{n+1}) \) from \( N(\beta, \bar{V}) \)
4. Simulate \( \lambda^2_{n+1} \) and \( \nu^2_{n+1} \) through slice sampling as in (11)
5. Iterate (1-4) until convergence

A computational bottleneck is however present in very high dimensions in evaluating the \( K \times K \) dimensional inverse for the conditional posterior of \( \beta \). Cholesky decomposition based methods will generally be of order \( O(K^3) \). Taking into consideration that in quantile settings, one is usually interested in obtaining more than one expected quantile, this can result in prohibitively long computation times. We therefore provide a more efficient sampling algorithm for \( \beta \) which leverages data augmentation similar to the algorithm developed by Bhattacharya et al. (2016) and is shown to be of order \( O(T^2K) \) which is especially beneficial in high dimensional settings.

As derived above, using the scale mixture representation in (13), the conditional posterior of \( \beta \) given all other parameters can be written as:

\[
\beta|\sigma, \lambda, \nu, \tau, X, Y, Z \sim N(A^{-1}X'Uy, A^{-1}), \quad A = (X'UX + \Lambda^{-1}_r), \quad \Lambda^{-1}_r = \nu^2diag(\lambda^2_1, \cdots, \lambda^2_K)
\]

(21)

Suppose, we want to sample from \( N_K(\mu, \Sigma) \), where

\[
\Sigma = (\Phi'\Phi + D)^{-1}, \quad \mu = \Sigma \Phi' \alpha.
\]

(22)

Assume \( D \in \mathbb{R}^{K \times K} \) is a positive definitive matrix, \( \phi \in \mathbb{R}^{T \times K} \), and \( \alpha \in \mathbb{R}^{T \times 1} \). Then (21) is a special case of (22) when setting \( \Phi = \sqrt{U}X, \quad D = \Lambda_r \) and \( \alpha = \sqrt{U}y \). An exact algorithm to sample from (21) is thus given by:
Algorithm 1. Fast HS-BQR sampler

1. Sample independently \( u \sim N(0, D) \) and \( \delta \sim N(0, I_T) \)

2. Set \( \xi = \Phi u + \delta \)

3. Solve \((\Phi D\Phi' + I_T)w = (\alpha - \xi)\)

4. Set \( \theta = u + D\Phi'w \)

Proposition 1. Suppose \( \theta \) is obtained through algorithm 1. Then \( \theta \sim N(\mu, \Sigma) \).

Proof. Using the Sherman-Morrison-Woodbury identity, \( \mu = D\Phi' (\Phi D\Phi' + I_T)^{-1} \alpha \). Plugging in \(2.\) into \(3.\), we obtain \( \theta = u + D\Phi' (\Phi D\Phi' + I_T)^{-1} (\alpha - \xi) \). Since by definition \( \xi \sim N(0, \Phi D\Phi' + I_K) \), \( \theta \) follows a normal distribution with mean \( D\Phi' (\Phi D\Phi' + I_K) \alpha = \mu \). As \( \text{cov}(u, \xi) = D\Phi' \), it follows that \( \text{cov}(\theta) = D - D\Phi' (\Phi D\Phi' + I_K)^{-1} D \) which by the Sherman-Morrison-Woodbury identity is equal to \( \Sigma \). More details are provided in supplementary material. The provided algorithm is not specific to the horseshoe prior and follows through for any prior of the form in (13). The computational advantage provided in algorithm 1 compared to Cholesky based decompositions is that we can cheaply sample from \((u, \xi)'\) which via data augmentation yields samples from the desired distributions.

3 Simulation setup

Now we set out to compare exponential based shrinkage priors to the proposed HS-BQR in order to verify the theoretical advantages laid out above. We consider 3 variants of the original lasso prior which have been adapted to the Bayesian Quantile regression and the frequentist lasso quantile regression of Belloni and Chernozhukov (2011):

1. Bayesian Lasso QR (LBQR): The lasso prior is derived by noticing that the L1 norm penalised check loss function

\[
\min_{\beta} \sum_{t=1}^{T} \rho_p(y_t - x'_t\beta) + \lambda \sum_{j=1}^{K} |\beta_j| \tag{23}
\]

can be obtained as the MAP estimate of the AL likelihood with a Laplace prior on the regression coefficients, \( \pi(\beta|\sigma, \lambda) = (\sigma \lambda / 2)^p \exp\{ -\sigma \lambda \sum_{j=1}^{K} |\beta_j| \} \). The posterior takes the
following form:

$$\beta| y, X, \sigma, \lambda \propto \exp\left(-\sigma \sum_{t=1}^{T} \rho_{\theta}(y_t - x_t' \beta) - \sigma \lambda \sum_{j=1}^{K} |\beta_j|\right)$$  \tag{24}$$

To estimate estimate (24), we utilise the Gibbs sampler of Li et al. (2010) with their recommended hyperpriors. Due to the shrinkage coefficient profile discussed above, we expect the LBQR to do well in sparse designs with well identified signal and noise.

2. Bayesian Elastic Net QR (BQRENET): The elastic net estimator quantile regression differs from the lasso in that it adds a L2 norm of the regression coefficients to the minimisation problem. This is the ridge component which allows to shrink coefficients in a less aggressive way than the L1 norm. This makes it useful when dealing with correlated or dense designs. Assuming the elastic net estimator for the quantile regression, as

$$\min_{\beta} \sum_{t=1}^{T} \rho_p(y_t - x_t' \beta) + \lambda_1 \sum_{j=1}^{K} |\beta_j| + \lambda_2 \sum_{j=1}^{K} \beta_j^2$$  \tag{25}$$

the prior can, similarly to above, be derived with the AL likelihood and an exponential prior as, \(\pi(\beta_k| \lambda_1, \lambda_2, \sigma) \propto \frac{\sigma \lambda_j}{2} \exp\left(-\sigma \lambda_1 |\beta_j| - \sigma \lambda_2 \beta_j^2\right)\). The posterior is then:

$$\beta| y, X, \sigma, \lambda \propto \exp\left(-\sigma \sum_{t=1}^{T} \rho_p(y_t - x_t' \beta) - \sigma \lambda_1 \sum_{j=1}^{K} |\beta_j| - \sigma \lambda_1^2 \sum_{j=1}^{K} \beta_j^2\right)$$  \tag{26}$$

We use the same hyperpriors as recommended by Li et al. (2010).

3. Bayesian Adaptive Lasso QR (BALQR). The adaptive lasso as proposed by Alhamzawi et al. (2012) uses the same setup as the LBQR, however, allows for the shrinkage coefficient to vary with each covariate. The prior can then be formulated as follows:

$$\pi(\beta| \lambda_j) = (\sigma \lambda_j/2)^p \exp\left(-\sigma \lambda_j \sum_{j=1}^{K} |\beta_j|\right)$$. Since this estimator allows for coefficient specific shrinkage we expect it to outperform the LBQR.

4. Lasso QR (L1-QR). The seminal L1-QR of Belloni and Chernozhukov (2011) applies a L1 penalised term to the standard frequentist quantile regression with check-loss function as in (3). Quantile specific regression coefficients are obtained as:

$$\min_{\beta} \sum_{t=1}^{T} \rho_p(y_t - x_t' \beta(p)) + \lambda_p \sqrt{\frac{p(1-p)}{T}} \sum_{j=1}^{K} |\beta(p)|$$  \tag{27}$$
All simulation setups have more explanatory variables than observations. The aim of considering $K \gg T$ scenarios is to verify that the proposed method remains valid in high dimensional settings. In total 100 Monte Carlo datasets, each with 200 observations, are generated. The first 100 observations are used to retrieve the $\hat{\beta}(p)$ vector to calculate bias, while the last 100 observations are used to construct forecast errors.

We consider 12 designs in total which vary along two different dimensions: the degree of sparsity and the error generating process. We test the following sparsity profiles:

- Sparse with $\beta = (1, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0_{1 \times 2T})$,
- Dense with $\beta = (1, 0.85_{1 \times T})$,
- Block structure with $\beta = (1, 0.85_{1 \times T}, 0_{1 \times T}, 0.85_{1 \times T})$.

Consider a linear model as in (1). To retrieve the true quantile regression coefficients, $\beta(p)$, we make use of Koenker (2005)’s alternative representation of the quantile regression:

$$y_t = x'_t \beta + (x'_t \vartheta) u_t$$

where $u_t$ is assumed to be i.i.d. having some CDF, $F$. The dimensionality of $\vartheta$ is $K \times 1$ and determines which covariates have non constant quantile functions. This can be seen from the solution for $\beta(p)$ to equation (28):

$$\beta(p) = \beta + \vartheta F^{-1}(p)$$

Hence, the true $\beta$ profile of a quantile regression model has a random coefficient model interpretation, where the vector of coefficients can be decomposed into a fixed plus a random component. In particular, the random component depends on the inverse CDF of the error, $F^{-1}(p)$. One can therefore think of $\vartheta$ as determining which variable is correlated with the error, where by default the first entry, $\vartheta_0$, is set to 1. This entails that location effects will always be present.\footnote{While it is possible for $\vartheta$ to take on any value, for simplicity we assume that the elements of $\vartheta$ only to take on the values $\{0, 1\}$.}

From a frequentist’ perspective Equation (29) is our oracle estimator for $\beta$ for a given quantile $p$, which, given that the AL likelihood approximation in equation (7) holds, can be compared to the mean of the posterior of equation (17) (Kozumi and Kobayashi, 2011). With this in mind, it is trivial to calculate the true $\beta$’s for the error generating processes considered.
DGP | Error distributions | Quantile functions
---|---|---
$y_1 = X\beta + \epsilon$ | $\epsilon \sim N(0, 1)$ | $\beta_0(p) = \beta_0 + F^{-1}_{N(0,1)}(p)$
$y_2 = X\beta + \epsilon$ | $\epsilon \sim T(3)$ | $\beta_0(p) = \beta_0 + F^{-1}_{T(3)}(p)$
$y_3 = X\beta + (1 + X_2)\epsilon$ | $\epsilon \sim N(0, 1)$ | $\beta_0(p) = \beta_0 + F^{-1}_{N(0,1)}(p)$
 | $\epsilon_1 \sim N(0, 1)$ | $\beta_1(p) = \beta_1 + F^{-1}_{N(0,1)}(p)$
$y_4 = X\beta + \epsilon_1 + X_2\epsilon_2$ | $\epsilon_1 \sim N(0, 1)$ | $\beta_0(p) = \beta_0 + F^{-1}_{N(0,1)}(p)$
 | $\epsilon_2 \sim U(0, 2)$ | $\beta_1(p) = \beta_1 + F^{-1}_{U(0,2)}(p)$

Table 1: Summary of simulation setups

The second dimension along which the DGPs differ is in their error process. The proposed DGPs can be grouped into two broad cases: (1) i.i.d. errors ($y_1$ and $y_2$); and (2) heteroskedastic errors ($y_3$ and $y_4$). In $y_1$, we assume that the error distribution follows a standard normal distribution and in $y_2$, the error has student-t distributed errors with 3 degrees of freedom. For the other cases, we assume simple heteroskedasticity caused by correlation between the second covariate and $\epsilon$ which will influence $\beta_1$. In contrast to $y_3$, $y_4$ can be thought of as containing a mixture between a uniform and a standard normal error distribution. In all simulations, the design matrix is simulated using a multivariate normal distribution with mean 0 and a covariance matrix with its $(i,j)$th element defined as $0.5|i-j|$. 

Relating the assumed error processes to the random coefficient representation (29), it is clear that, under i.i.d. errors, only the constant has a non-constant quantile function caused by $F^{-1}$ (hereinafter called location shifters). Under the heteroskedastic designs, apart from the constant, $\beta_1$ will have a non-constant quantile function as well. Hence, $\beta_1$ in $y_3$ is determined by $F^{-1}_{N(0,1)}$ across p, and $\beta_1$ in $y_4$ follows $F^{-1}_{U(0,2)}$, i.e., increases linearly with p. The simulation designs (and the expected quantile functions) are summarised in table (1).

We evaluate the performance of the estimators in terms of bias in the coefficients and forecast error. Using the true quantile profile in $\beta(p)$ in (29), we calculate root mean coefficient bias (RMCB) and root mean squared forecast error (RMSFE) as:

1. Root Mean Coefficient Bias = $\sqrt{\frac{1}{iter}||\hat{\beta}(p) - \beta(p)||}$
2. Root Mean Squared Forecast Error = $\sqrt{\frac{1}{iter}||X\hat{\beta}(p) - X\beta(p)||}$

$iter$ refers to the number of Monte Carlo repetitions. For the Bayesian estimators we define $\hat{\beta}(p)$ as the mean of the posterior for the $p^{th}$ quantiles model. Note that the way the simulations
were designed, \textit{iter} is 100 for both RMSFE and RMCB, for all estimators.\footnote{The only estimator where there is a deviation from 100 is the RMCB and RMSFE of BALQR where the variance covariance matrix of the posterior coefficients was not invertible for some of the cases. This is indicative that the BALQR prior did not shrink enough.}

### 3.1 I.i.d. distributed random error simulation results

The different error distributions test whether the quantile is truly robust to changes in their specification as the literature suggests (see Koenker and Bassett (1982)). As described above, the conditional distributions are completely described with location shifters in the case of i.i.d. errors, where $y_1$'s $\beta_0$ has a profile of an inverse normal CDF shifted up by the respective constant coefficient, and $y_2$'s $\beta_0$ has a profile of an inverse student-t distribution with 3 degrees of freedom, shifted up by the respective constant coefficient. In both these cases $\vartheta$ is equal 1. The results for the bias for the three designs (sparse, dense, block) across a selection of quantiles are presented in table (2) and the results of the forecast performance are presented in table (3). The way the DGP’s were constructed, only a few variables have a non constant quantile curve. While the tables present a good overview of overall performance it does not provide information whether the estimator captures the correct quantile curves. To alleviate this, boxplots were created for each quantile for the variables with non constant quantile curves. The HS-BQR’s boxplots are presented in figure (2).\footnote{The same boxplots were constructed for the other estimators and are presented in an online appendix.}

Table (2) shows that the HS-BQR performs admirably compared to the considered estimators in all i.i.d designs regardless of what type of sparsity structure is considered. In particular, for the sparse case the HS-BQR provides the lowest coefficient bias for both $y_1$ and $y_2$ for all the quantiles. The forecast results from table (3) corroborate these findings with the HS-BQR providing the lowest root mean squared forecast errors among the estimators considered for the sparse design of $y_1$ and $y_2$.

The HS-BQR’s performance is competitive for the dense and block cases, but other estimators yield lower coefficient bias for these designs as can be seen in table (2). For the dense design the HS-BQR rivals the performance of the BQRENET and even provides lower coefficient bias for some of the quantiles for both $y_1$ and $y_2$. The same cannot be said for the block design where the BALQR yields superior performance, except for the most extreme quantiles considered for which the HS-BQR gives the lowest coefficient bias. The forecast error results of table (3) are in line with the bias results where the HS-BQR offers comparable performance compared to the
Figure 2: \( \beta_0 \) profiles for \( y_1 \) and \( y_2 \) across quantiles for the different sparsity settings

BQRENET for the dense designs, while the block design is dominated by the BALQR. This coheres with the theoretical properties of the priors. The ridge component in the BQRENET provides more stable inference for dense designs, while the BALQR benefits in block structures from adaptive shrinkage without having to identify a global shrinkage parameter. Note, how the coefficient bias and forecast errors are the lowest for the sparse cases for all the estimators considered. This result is expected, as the priors favour sparse posterior solutions.

Both the normally distributed \( y_1 \) and t-distributed \( y_2 \) showcase a situation where the extreme quantiles (0.1 and 0.9) have higher bias than the central quantile (0.5) for all the estimators considered. This is a common finding in quantile regressions which is on account of more extreme quantiles being ”data sparse” as a few observations get large weights. While it is expected that there is a U-shape in the coefficient bias as we move across the quantiles, the slope of this shape is not uniform across the estimators. In particular, it can be seen in table (2) that the HS-BQR’s bias does not increase as much as the other estimators.\(^7\) Similarly, extreme quantiles generally tend to have higher forecast errors for all estimators, but the HS-BQR’s extreme quantiles don’t suffer as much as it’s competition as shown in table (3). This property cannot be overstated, as quantile regression is often employed for extreme quantiles.

\(^7\)Apart from the LI-QR and HS-BQR in the block design, where the estimators have lower coefficient bias and forecast error for its extreme low quantiles than its central quantiles.
| p  | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|----|-----|-----|-----|-----|-----|
| Sparse |
| HS-BQR | 0.045 | 0.036 | 0.034 | 0.038 | 0.050 |
| L1-QR | 0.070 | 0.056 | 0.051 | 0.056 | 0.069 |
| LBQR | 0.051 | 0.044 | 0.050 | 0.074 | 0.146 |
| BQRnet | 0.046 | 0.042 | 0.053 | 0.080 | 0.113 |
| BALQR | 0.075 | 0.049 | 0.043 | 0.052 | 0.080 |
| Dense |
| HS-BQR | 0.711 | 0.710 | 0.709 | 0.716 | 0.722 |
| L1-QR | 0.897 | 0.890 | 0.889 | 0.891 | 0.910 |
| LBQR | 0.780 | 0.731 | 0.728 | 0.773 | 0.816 |
| BQRnet | 0.739 | 0.676 | 0.679 | 0.716 | 0.781 |
| BALQR | 1.271 | 1.233 | 1.250 | 1.246 | 1.265 |
| Block |
| HS-BQR | 0.747 | 0.752 | 0.754 | 0.760 | 0.766 |
| L1-QR | 0.860 | 0.879 | 0.880 | 0.879 | 0.858 |
| LBQR | 0.821 | 0.737 | 0.716 | 0.783 | 0.870 |
| BQRnet | 0.776 | 0.690 | 0.696 | 0.730 | 0.835 |
| BALQR | 0.682 | 0.660 | 0.671 | 0.670 | 0.687 |
### Table 3: Root mean squared forecast error

|          | $p$ | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Sparse   |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| HS-BQR   | 0.860 | 0.680 | 0.642 | 0.708 | 0.971 | 1.184 | 0.928 | 0.871 | 0.943 | 1.372 | 3.292 | 3.074 | 3.054 | 3.146 | 3.682 | 2.550 | 1.786 | 1.386 | 1.769 | 2.566 |
| L1-QR    | 1.408 | 1.120 | 1.011 | 1.109 | 1.380 | 1.915 | 1.605 | 1.563 | 1.617 | 1.940 | 5.938 | 6.565 | 7.591 | 7.961 | 7.539 | 3.701 | 4.174 | 4.521 | 4.939 | 5.222 |
| LBQR     | 1.145 | 0.996 | 1.104 | 1.542 | 2.950 | 1.616 | 1.214 | 1.416 | 1.904 | 3.441 | 234.886 | 140.067 | 374.485 | 181.890 | 124.801 | 233.683 | 138.670 | 373.608 | 185.163 | 125.013 |
| BQRENET  | 1.006 | 0.925 | 1.086 | 1.546 | 2.168 | 1.424 | 1.106 | 1.215 | 1.666 | 2.648 | 2.905 | 3.302 | 4.411 | 7.266 | 10.059 | 3.078 | 3.607 | 4.854 | 6.211 | 9.148 |
| BALQR    | 1.500 | 0.981 | 0.869 | 1.055 | 1.610 | 3.079 | 2.762 | 2.756 | 2.770 | 3.187 | 20.606 | 19.906 | 20.136 | 20.365 | 22.049 | 11.017 | 10.935 | 10.492 | 11.061 | 11.715 |
| Dense    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| HS-BQR   | 11.003 | 10.980 | 10.946 | 11.044 | 11.230 | 11.183 | 11.078 | 11.049 | 11.125 | 11.368 | 89.282 | 88.411 | 89.460 | 91.045 | 95.320 | 240.697 | 233.984 | 231.751 | 236.006 | 243.771 |
| L1-QR    | 14.294 | 13.236 | 13.004 | 13.110 | 14.361 | 14.303 | 13.303 | 13.140 | 13.212 | 14.480 | 126.723 | 98.655 | 94.843 | 109.917 | 130.619 | 127.593 | 101.577 | 97.671 | 102.995 | 130.733 |
| LBQR     | 10.983 | 10.278 | 10.333 | 10.790 | 11.067 | 10.944 | 10.392 | 10.218 | 10.626 | 11.625 | 70.888 | 72.796 | 70.737 | 72.803 | 74.851 | 71.326 | 71.024 | 72.722 | 70.954 | 71.660 |
| BQRENET  | 10.206 | 9.699 | 9.541 | 9.968 | 10.515 | 10.261 | 9.892 | 9.808 | 10.246 | 10.821 | 68.021 | 70.640 | 69.962 | 71.698 | 72.486 | 67.980 | 68.151 | 69.333 | 69.734 | 70.895 |
| BALQR    | 17.183 | 16.514 | 16.826 | 16.817 | 16.955 | 17.108 | 16.754 | 16.782 | 17.039 | 17.276 | 94.984 | 96.213 | 96.598 | 98.022 | 100.397 | 95.155 | 95.543 | 98.125 | 99.915 | 101.325 |
| Block    |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| HS-BQR   | 24.309 | 24.306 | 24.339 | 24.457 | 24.607 | 24.248 | 24.357 | 24.479 | 24.508 | 24.627 | 161.618 | 162.538 | 163.134 | 163.784 | 164.756 | 49.476 | 49.699 | 50.382 | 50.141 | 50.380 |
| L1-QR    | 26.625 | 26.695 | 26.743 | 26.748 | 26.783 | 26.643 | 26.639 | 26.681 | 26.737 | 26.832 | 398.370 | 397.449 | 380.420 | 382.532 | 401.150 | 401.824 | 349.831 | 349.718 | 348.566 | 349.646 |
| LBQR     | 25.202 | 23.456 | 23.160 | 24.373 | 26.090 | 25.294 | 23.511 | 22.860 | 23.877 | 26.381 | 438.545 | 349.283 | 348.442 | 349.462 | 351.545 | 347.024 | 349.831 | 349.718 | 348.566 | 349.646 |
| BQRENET  | 23.998 | 22.610 | 22.919 | 23.252 | 25.065 | 24.531 | 22.887 | 22.411 | 23.471 | 25.309 | 547.150 | 548.995 | 444.694 | 444.198 | 438.850 | 347.864 | 345.601 | 348.192 | 349.530 | 348.173 |
| BALQR    | 22.934 | 22.803 | 22.719 | 22.858 | 23.163 | 21.115 | 22.835 | 22.858 | 22.823 | 23.259 | 357.677 | 358.503 | 360.752 | 362.253 | 365.787 | 359.939 | 360.265 | 362.389 | 363.765 | 366.789 |
While tables (2) and (3) provide a good overview of the general performance of the HS-BQR, it does not give an indication of how well the estimator captures the true quantile function. To tackle this, figure (2) presents the boxplots of all the estimated quantiles for all designs, for \( y_1 \) and \( y_2 \). The figure underpins the findings of the tables: The HS-BQR captures the normal inverse CDF shape for \( y_1 \) and inverse t-distribution for \( y_2 \) for the sparse design. In the dense design the location shift effect is not as well captured as for the sparse case, but the HS-BQR is capable of identifying location shift’s for the more extreme quantiles. The figure also highlights how the HS-BQR suffers the most in block designs, where the boxplots highlight how for many monte carlo cases the estimated \( \beta_0 \) coefficient is an outlier, several magnitudes above the mean \( \beta_0 \). This finding underpins, that in designs with unmodeled block structures and, hence, badly identified global shrinkage, quantile effect might be shrunk away. Implementation of group-level shrinkage along with prior information about the sparsity pattern in the data might be able to alleviate this problem, which we leave for future research.

### 3.2 Heteroskedastic error simulation results

As correlation between a covariate and the error causes a quantile profile, \( y_3 \) and \( y_4 \) will have both location shifters and scale shifters. In particular, \( \beta_1 \) of \( y_3 \) follows an inverse Normal CDF shifted up by the respective constant, and \( \beta_1 \) of \( y_4 \) is a linear line from 1 to 3 across the quantiles. It is worth noting, that to make sure there is no kink in the data\(^8\) the variable multiplied by the error is restricted to take on positive values only. This was achieved by taking the exponential of that particular variable\(^9\).

As with the homoskedastic DGPs, we see that for all estimators, the error rate increases when moving away from the central quantiles and that coefficient bias as well as forecast accuracy worsens for dense and block designs compared to the sparse design. Further, the bias and forecast results in tables (2) and (3) show that the HS-BQR provides better or similar performance to the alternative estimators, where it consistently outperforms all Bayesian estimators for \( y_4 \) in sparse designs. Similar to the previous discussion, the HS-BQR stands out in that it provides consistently more stable inference of extreme quantiles independent of the sparsity structure.

---

\(^8\)When allowing for \( X_2 \) to take on negative values as well, the relationship between \( Y \) and \( X_2 \) will involve a kink at \( X_2 = 0 \) which causes non-linearity and therefore violates the assumption of linearity in parameters.

\(^9\)The exponential of the \( 7^{th} \) variable for the sparse and dense case and the \( T + 2^{nd} \) variable for the block case is also taken. This does not influence the results.
Figure 3: $\beta_0$ and $\beta_1$ profiles for $y_3$ across quantiles for the different sparsity settings

Figure 4: $\beta_0$ and $\beta_1$ profiles for $y_4$ across quantiles for the different sparsity settings
Confirming the homoskedastic simulation results, in dense designs, the BQRENET aided by the ridge component in the prior, provides lower coefficient bias and forecast error, than the HS-BQR, whereas in block DGPs, the BALQR marginally outperforms the HS-BQR for $y_3$. A different picture emerges for block DGPs of $y_4$. Here, the HS-BQR provides a gain in the precision of coefficient estimation between 2-20% and in excess of 100% in terms of forecast accuracy.

The boxplots in figure (4) also provide an explanation as to why the HS-BQR’s forecast performance is much better for the block case of $y_4$, which is that it captured the quantile function for $\beta_0$. The HS-BQR’s lackluster performance in forecasting the dense design of $y_3$ is revealed in figure (3): the HS-BQR failed to capture the quantile function as precisely it was able to do so for the same design of $y_4$. This highlights that to obtain good performance for these estimators, it is not sufficient to shrink the unimportant variables to 0, it is also imperative to identify the correct quantile function for the variables that are heteroskedastic.

Generally, as can be seen from (3) and (4), The HS-BQR seems to have difficulty capturing both location and scale effects present in the DGP. It often isolates only one of the effects, while shrinking the other to a constant quantile function. To check whether this is a general problem for Bayesian quantile regression, we estimated the HS-BQR on a DGP determined by only one explanatory variable (and a constant). As can be seen in figure (5), quantile profiles are well...
All the estimators have difficulty simultaneously identifying the true regressors and partialling out the location ($\beta_0(p)$) and scale ($\beta_1(p)$) effects in high dimensional setting. This difficulty of capturing the quantile profiles for both location and scale effects could stem from estimating the quantiles independently. Namely, there is no monotonicity imposed on the profile of $\beta_0(p)$. As such, further improvements in high dimensional quantile regression can be achieved, by finding ways where the methods don’t shrink one of the location of scale shifters to a constant quantile profile.

4 Inflation Density Forecasting

We now compare the above estimator’s ability to construct forecast densities of US quarterly inflation. Density forecasts are an important aspect of macroeconomics that are extensively used in institutions such as central banks which construct fan-chart estimates of key economic indicators and is getting increasing attention in academics as well (Adrian et al., 2019; Aastveit et al., 2018). Rather than simply forecasting the mean, density forecasts also provide information about tail risk. Quantile regression is particularly useful for characterising these tail risks, as it is capable of giving asymmetric forecast densities where upside and downside risk are estimated independently. Traditional conditional mean models, in contrast, assume a normal distribution around the forecasted values, which characterises the uncertainty in a symmetric way.

We break from the forecast density literature a bit, by not exclusively focusing on testing the whole density, but also evaluating specific quantiles’ performance. This is done to show the HS-BQR’s power to create Value-at-Risk estimates on top of creating intuitive density forecasts.

We use data provided by Korobilis (2017) which uses 16 macro variables as potential regressors. These include measures of real economic activity (e.g. unemployment, investment, housing starts), financial market information (e.g. default yield spreads), money supply variables (e.g. M1 money stock) and expectation indices (e.g. PMI) (see: Korobilis (2017) for more information on the data set) as well as two lags of the dependent variable. The dependent variable is US quarterly CPI inflation, seasonally adjusted, spanning the period from 1947Q1-2015Q3.

To obtain the forecasts, we use the general linear model:

$$ y_{t+h} = x'_{t+h} \beta + \epsilon_{t+h} $$ (30)

10The small design follows the same structure as outlined in table (I).
for \( t = 1, \cdots, T - h \), where \( h \) refers to the forecast horizon. For brevity, in this paper we only consider one-step-ahead forecasts such that \( h \) is set to 1. Using the quantile setup, forecasts from each quantile are denoted as \( y_{T+1|T}^p \). Note, that these one step ahead forecasts are equivalent to the one step ahead \( p^{th} \) Value-at-Risk. Forecasts are computed on a rolling basis, meaning that each \( T+1 \) forecast is constructed from a one time period expanding window of in-sample observations \( t = 1, \cdots, T \). The initial in-sample period uses the first 40 observations of the sample, which makes for 233 rolling forecast windows. Just like in Korobilis (2017), we estimate a grid of 19 equidistant quantiles to construct the predictive density \( p(\hat{y}_{T+1|T}) \).

Forecast densities are evaluated along Kolmogorov-Smirnov (KS) statistics based on Probability Integral Transform (PIT).\(^{11}\) The PIT is often used when evaluating density forecasts and hinges on the idea that any given forecast distribution \( p(y_{t+h} | y_t) \) can be converted to random variables having a standard uniform distribution. In particular, the PIT is the corresponding CDF of the density function evaluated at the actual observation of the out-of-sample periods, \( y_{t+h} \), and is constructed using the following formula:

\[
    g_{t+h} = \int_{-\infty}^{y_{t+h}} p(u | y_t) du = P(y_{t+h} | y_t) \quad (31)
\]

The estimated predictive density is consistent with the true density when the \( g_{t+h} \) are i.i.d. uniform. As such, the CDF should be a 45 degree line (Diebold et al., 1998). The additional benefit of this test is that the theoretical true PIT distribution is independent of the econometrician’s loss function. With this in mind, any test can be employed that compares the distance between the empirical and theoretical PIT distribution. We opt for using the KS test to compare the PITs of the estimators with the uniform CDF.

As mentioned above, this paper does not exclusively focus on the forecast densities but also tries to evaluate the specific quantile forecasts. The two most popular tests to verify the performance of a specific quantile are the Dynamic Quantile test of Engle and Manganelli (2004) and the VQR test of Gaglianone et al. (2011). These tests provide a principled way of testing the null hypothesis of the selected quantile being correct. However, they do not offer a comparative measure as to which method provides better fit for a specific quantile. For this reason this paper instead computes the pseudo \( R^2 \) for the quantiles for each estimator, following Koenker and Machado (1999). The pseudo \( R^2 \) of the following regression is obtained:

\(^{11}\)There are a plethora of tests to evaluate distributions based on QQ-plot of the PIT. The choice of the KS was based solely on its simplicity to compute and any other test would suffice for evaluation.
where $V_t(p)$ is the fitted value of of the estimator for the $p^{th}$ quantile. Running the regression in equation (32) for the $p^{th}$ quantile gives an intuitive test for the ability of the estimated fitted value to capture the dynamics we are interested in. In particular the pseudo $R^2$ shows how much information $V_t(p)$ adds to the regression compared to a Quantile regression with only a constant. Ideally the coefficient of $\beta_0$ should be 0 while the coefficient of $\beta_1$ should be 1.

In addition to the estimators considered in the Monte Carlo study, we also estimate the Bayesian quantile regression (BQR) without any shrinkage priors. This serves as a benchmark. The results of the KS test, and the pseudo $R^2$'s are presented in table (4), while the forecasted densities are shown in figures (6) and (7) where the shaded regions show the fill between the $5^{th}$ quantile and the $95^{th}$ quantile after sorting the quantiles as suggested by Chernozhukov et al. (2010).

Comparing the forecast distributions on a visual basis in figures 6 and 7, reveals certain properties of the estimators. First, the horsehoe prior increases calibration of the density forecasts compared to the plain BQR. This can be seen from figure 6 where the blue shaded area is often too narrow with the central tendency not capturing sharp movements unlike the HS-BQR. Second, L1QR and BALQR perform similarly to the HS-BQR, but offer wider uncertainty bands. The BALQR in particular consistently has larger 95th and 5th quantile estimates than the other estimators considered in figure 6 which from an economic standpoint are unreasonably large. This can be seen by the lowest band of the BALQR which consistently provides estimates that are below 0% inflation. The BALQR’s forecasts imply a consistent 5% risk of deflation for the past 50 years. Third, the LBQR and BQRENET provide the worst calibrated forecast densities which are too wide compared to the other estimators. As shown in figure (7), both estimators have considerable downward skew with lower quantiles estimating in excess of 20%-30% deflation at different parts of the sample. Upper quantiles for the BQRENET and LBQR after 2009 include the possibility of 30% inflation. These forecast densities seem far too wide to be informative.

The visual inspection however is not confirmed by the KS statistics in table (4): The lasso based estimators have a lower KS statistic than the HS-BQR. Since the KS-statistic has a null hypothesis that the empirical CDF is no different than a uniform distribution, the higher
Figure 6: One-step-ahead forecast distributions for the L1QR, BQR, BALQR and HS-BQR. Shaded areas correspond to plots of all 19 quantiles.

Figure 7: One-step-ahead forecast distributions for the BQR, BQRENET, and LBQR models. Shaded areas correspond to plots of all 19 quantiles.
the statistic is, the worse the estimator performs. Note, however that with the 19 quantiles used, the critical values for the KS statistic is 0.374 for the 1% significance level, 0.312 for the 5% significance level, and 0.280 for the 10% significance level. As such according to the KS statistics, none of the presented models have PIT’s that are significantly different from the uniform distribution. Hence, it cannot be concluded that any of the estimators density forecasts PIT are inaccurate.

The pseudo $R^2$ is able to remedy this conundrum by comparing the explanatory power of each individual quantile and corroborates the visual inspection. The HS-BQR yields the best performance, when looking at the pseudo $R^2$, with it beating the other estimators considered for all the quantiles considered. The L1-QR and BQR also provide good results with pseudo $R^2$’s just below that of the HS-BQR. The BALQR’s pseudo $R^2$ for the lower quantiles are worse than that of the HS-BQR, BQR, and L1-QR, but its higher quantiles perform just as well, which is is in line with figure [6]. Finally, the LBQR and BQRENET perform the worst, especially for lower quantiles. The pseudo $R^2$ results highlight the potential of HS-BQR for not just creating density forecasts, but also in providing Value-at-Risk estimates.

### 5 Conclusion

In this paper, we have extended the widely popular horseshoe prior of Carvalho et al. (2010) to the Bayesian quantile regression and provided a new algorithm to sample the shrinkage coefficients via slice sampling for the independent prior and a fast sampling algorithm that speeds up computation significantly in high dimensions.
In our simulations, we considered a variety of sparse, dense and block designs with different error distributions. We then tested a selection of quantile regression methods to see how the HS-BQR fares in comparison. The simulations revealed three points about the HS-BQR. First, the HS-BQR provides better or comparable performance in terms of both coefficient bias and forecast risk. For the sparse design it beat all other estimators, while in the dense and block designs the HS-BQR came close to the best estimator even beating it for some quantiles. Second, the HS-BQR was more stable when estimating the tails of the distribution, having the lowest bias and forecast error for the extreme quantiles (0.1 and 0.9). This shows how the shrinkage profile of the Horseshoe prior is more efficient than the lasso based priors. Finally, the HS-BQR has difficulties simultaneously identifying the correct location and scale effects in high-dimensional setting, but this is an issue for the other estimators as well. When the HS-BQR was applied to a comparable ‘small data’ design, it had no issues simultaneously identifying $\beta_0$’s and $\beta_1$’s quantile function.

In the empirical application, we tested the HS-BQR’s performance in creating density forecasts as well as Value-at-Risk estimations. The forecast density of the HS-BQR provides economically more intuitive forecast densities than the LBQR, BALQR, and BQRENENET, all of which provide forecast bands that are too wide. The KS statistic reveals that the PIT CDF of the HS-BQR is not significantly different from the expected uniform distribution. The pseudo $R^2$ revealed that the the HS-BQR’s fitted quantiles provide the best goodness of fit of all the estimators. The only other estimator that comes close to the HS-BQR is the L1-QR. This shows that the HS-BQR is an adequate method to give credible Value-at-Risk estimates.

The results show that the HS-BQR is a competitive estimator for which especially good behaviour can be expected in sparse designs with few observations. However, there are multiple fronts on which the proposed HS-BQR can be improved upon. For instance, the simulations highlighted that in dense and block designs, the HS prior tends to shrink the constant too aggressively. Hence, extensions which allow for differing shrinkage terms for subsets of the regressors might be able to alleviate this problem. The results also highlight that the estimators have difficulties simultaneously capturing location and scale shifts in high dimensions. Extensions to the HS-BQR should tackle this because to attain oracle properties in quantile regression it is not enough to shrink unimportant variables to 0, but also to identify the correct quantile functions.
References

Aastveit, K. A., J. Mitchell, F. Ravazzolo, and H. K. Van Dijk (2018). The evolution of forecast density combinations in economics. Technical report, Tinbergen Institute Discussion Paper.

Adrian, T., N. Boyarchenko, and D. Giannone (2019). Vulnerable growth. *American Economic Review* 109(4), 1263–89.

Alhamzawi, R. and K. Yu (2013). Conjugate priors and variable selection for bayesian quantile regression. *Computational Statistics & Data Analysis* 64, 209–219.

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

Bai, J. and S. Ng (2008). *Large dimensional factor analysis*. Now Publishers Inc.

Batencourt, M. (2018). Bayes sparse regression.

Belloni, A. and V. Chernozhukov (2011). l1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* 39(1), 82–130.

Bhadra, A., J. Datta, N. G. Polson, B. Willard, et al. (2019). Lasso meets horseshoe: A survey. *Statistical Science* 34(3), 405–427.

Bhattacharya, A., A. Chakraborty, and B. K. Mallick (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, asw042.

Buchinsky, M. (1998). Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources*, 88–126.

Carriero, A., T. E. Clark, and M. G. Marcellino (2020). Nowcasting tail risks to economic activity with many indicators.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika* 97(2), 465–480.

Chen, C. W., D. B. Dunson, C. Reed, and K. Yu (2013). Bayesian variable selection in quantile regression. *Statistics and its Interface* 6(2), 261–274.
Chen, C. W., R. Gerlach, B. B. Hwang, and M. McAleer (2012). Forecasting value-at-risk using nonlinear regression quantiles and the intra-day range. *International Journal of Forecasting* 28(3), 557–574.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica* 78(3), 1093–1125.

Davino, C., M. Furno, and D. Vistocco (2013). *Quantile regression: theory and applications*, Volume 988. John Wiley & Sons.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review* 39(4), 863–883.

Engle, R. F. and S. Manganelli (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* 22(4), 367–381.

Gaglianone, W. P., L. R. Lima, O. Linton, and D. R. Smith (2011). Evaluating value-at-risk models via quantile regression. *Journal of Business & Economic Statistics* 29(1), 150–160.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis* 1(3), 515–534.

Giglio, S., B. Kelly, and S. Pruitt (2016). Systemic risk and the macroeconomy: An empirical evaluation. *Journal of Financial Economics* 119(3), 457–471.

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. (2005). *Quantile regression*. New York: Cambridge University Press.

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

Koenker, R. and G. Bassett (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society*, 43–61.

Koenker, R. and J. A. Machado (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association* 94(448), 1296–1310.

Korobilis, D. (2017). Quantile regression forecasts of inflation under model uncertainty. *International Journal of Forecasting* 33(1), 11–20.
Kotz, S., T. Kozubowski, and K. Podgorski (2012). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media.

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

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

Makalic, E. and D. F. Schmidt (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters* 23(1), 179–182.

Moran, G. E., V. Ročková, E. I. George, et al. (2018). Variance prior forms for high-dimensional bayesian variable selection. *Bayesian Analysis*, 1091–1119.

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association* 103(482), 681–686.

Polson, N. G. and J. G. Scott (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics 9*, 501–538.

Polson, N. G., J. G. Scott, and J. Windle (2014). The bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(4), 713–733.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association* 97(460), 1167–1179.

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. *Journal of Business & Economic Statistics* 37(1), 121–133.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1), 267–288.

Wang, L. (2017). Nonconvex penalized quantile regression: A review of methods, theory and algorithms. In *Handbook of Quantile Regression*, pp. 273–292. Chapman and Hall/CRC.

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

We now give further details on the derivation of Algorithm 1. The goal of the algorithm is to circumvent having to compute large $K \times K$ matrices by redefining auxiliary variables which under certain linear combination result in draws of the desired distribution $N(\beta, V)$. As above, by the Sherman-Morrison-Woodbury theorem (see e.g. Hager 1989), $\Sigma$ and $\mu$ can be expanded as:

$$
\Sigma = (\Phi' \Phi + D^{-1})^{-1} = D - D\Phi'(\Phi D \Phi + I_T)^{-1}\Phi D \\
\mu = D\Phi'(\Phi D \Phi + I_T)\alpha
$$

This expansion per-se won’t help in sampling from $N(0, \Sigma)$. Letting $\xi$ and $u$ being defined as above, $\omega = (\xi', u')' \in \mathbb{R}^{T+K}$ follows a multivariate normal distribution centred on 0 with covariance

$$
\Omega = \begin{pmatrix} P & S \\ S' & R \end{pmatrix}
$$

where it is easily verified that $P = (\Phi D \Phi' + I_T)$, $R = D$ and $S$ can be derived as:

$$
\text{Cov}(\xi, u) = \text{Cov}(\sqrt{D}\epsilon, \Phi u) \\
= E(\sqrt{D}\epsilon u'X'\sqrt{U}) \\
= E(\sqrt{D}\epsilon\epsilon'X'\sqrt{U}) \\
= DX'\sqrt{U} \\
= D\Phi'
$$

$\epsilon$ is defined here following $N(0,1)$ distribution. Rewriting $\Omega$ into its LDU decomposition (see e.g. Hamilton, 1994) as:

$$
\begin{pmatrix} P & S \\ S' & R \end{pmatrix} = \begin{pmatrix} I_T & 0 \\ S'P^{-1} & I_K \end{pmatrix} \begin{pmatrix} P & 0 \\ 0 & R - S'P^{-1}S \end{pmatrix} \begin{pmatrix} I_T & P^{-1}S \\ 0 & I_K \end{pmatrix}
$$

(33)

Where the lower $K \times K$ block in $\Gamma$ is equal to $\Sigma$. To retrieve the lower part, we isolate $\Gamma$ which is easily obtained because $L$ is lower triangular and thus the inverse is readily available as:

$$
\begin{pmatrix} I_T & 0 \\ -S'P^{-1} & I_K \end{pmatrix}
$$
Since \( \omega \) has already been sampled from \( N(0, \Omega) \) in steps 2 and three of the algorithm, the transformation \( \omega_* = L^{-1} \) is distributed \( N(0, \Gamma) \). Collecting the lower block of \( \omega_* \) yields a sample from \( N(0, \Sigma) \).