Bayesian Iterative Adaptive Lasso Quantile Regression

Rahim Alhamzawi[1], Haithem Taha Mohammad Ali[2]

[1]Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Iraq
[2]College of Computers and Information Technology, Nawroz University

Abstract
Based on the Bayesian adaptive Lasso quantile regression (Alhamzawi et al., 2012), we propose the iterative adaptive Lasso quantile regression, which is an extension to the Expectation Conditional Maximization (ECM) algorithm (Sun et al., 2010). The proposed method is demonstrated via simulation studies and a real data set. Results indicate that the new algorithm performs quite good in comparison to the other existing methods.
Keywords*: ECM, Bayesian inference, Prior distribution, Posterior inference, Quantile regression.*

I. Introduction

Quantile regression (QRe) of Koenker and Bassett (1978) has appeared as a useful supplement to standard regression. It can be applied for inference about the relationship between quantiles of the outcome of interest and a set of predictors. This technique provides a more complete picture than standard regression due to a group of quantiles of the outcome distribution offers a more complete description of the outcome distribution than the center. Median regression is a special case of QRe. QRe puts very minimal assumptions on the error term and therefore is able to use in the case of non-normal error terms, which are popular in many real world applications. A comprehensive review of the QRe and its current applications is introduced by Al-Hamzawi (2013) and Koenker (2005).

Suppose that y_i denote the response variable, and suppose that x_i denote the corresponding predictor vector, for $i = 1, ..., n$. Then, the pth quantile regression (QRe) model ($Q_{yi}|x_i(p)$) is given by

$$Q_{y_{i|x_i}}(p) = x_i^{'}\beta, \quad i = 1, ..., n \quad and \ p \in (0,1). \tag{1}$$

According to Koenker and Bassett (1978), the unknown quantity $\beta = (\beta_1, ..., \beta_n)$ can be evaluated by minimizing

$$\sum_{i=1}^{n} \rho_p(y_i - Q_{y_{i|x_i}}(p)), \tag{2}$$

where $\rho_p(\varepsilon) = \varepsilon\{p - I(p \leq 0)\}$ is the Check Function (CF) of Koenker and Bassett (1978).
Equivalently, we may write the CF as

$$\rho_p(\varepsilon) = \frac{|\varepsilon| + (2p - 1)\varepsilon}{2} \tag{3}$$
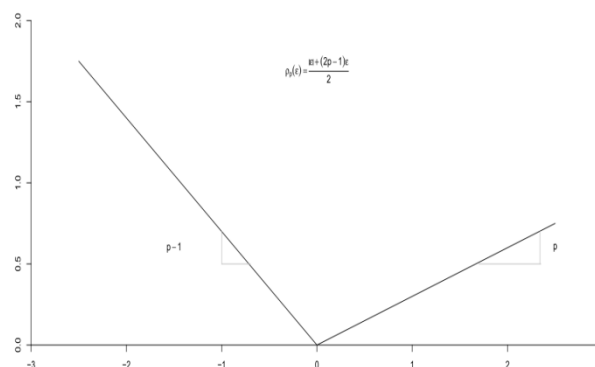

Figure 1: The panel shows the check function at p = 0.30.

Figure 1 shows the CF at *p = 0.30*. Since the CF is not differentiable at zero, a closed form solution is not available for the the unknown quantity *β* (Koenker and Bassett, 1978). However, the minimisation of the CF can be achieved via an algorithm proposed by Koenker and d'Orey (1987). From a computational perspective, many statistical softwares such as R, Matlab STATA and SAS can be used to estimate the QRe coefficients. The

CF (3) is closely related to the skewed Laplace Distribution *(LD); see Yu and Moyeed (2001). The density function of a skewed LD is given by*

$$f(\varepsilon) = \frac{p(1-p)}{\tau} \exp\{-\rho_p(\varepsilon)\}. \qquad (4)$$

According to Yu and Moyeed (2001), minimizing the CF in (3) is equvelent to maximizing the likelihood of the skewed LD in (4). This relationship between the CF and the skewed LD can be employed to rewrite the QRe technique in the usual likelihood framework. One of the attractive features of the skewed LD is that it can be reformulated as a location-scale mixture of normals (Kozumi and Kobayashi, 2011). This idea links the QRe model for response variable to the standard mean regression model. In addition, under this link, all the parameters of the model have desirable conjugacy features for building a simple MCMC algorithm for fitting the model to the data. This algorithm is implemented in an R package called bayesQR (Benoit et al., 2011). According to Kozumi and Kobayashi (2011), if $\varepsilon_i \sim N(\theta v_i, 2v_i)$, where $\theta = 1 - 2p$ and v_i is a mixing latent variable, then the skewed LD for ε_i arises when $v_i \sim Exp(p(1-p))$. This formulation appears in papers by Benoit et al. (2013), Waldmann et al. (2013) Hashem et al. (2015) and Nassiri and Loris (2014).

One important issue in constructing a multiple QRe model is the selection of active regressors in the regression. The prediction accuracy and model stability can be increased by selecting an active set of predictors. However, classical model selection approaches, such as AIC (Akaike, 1998) and BIC (Schwarz, 1978) are consuming alot of time and often instability. Recently, there has been promise research on regularization in linear regression. See for example, Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), and group-Lasso (Yuan and Lin, 2006).

With regard to QRe, Wang and Leng (2012) proposed the least absolute deviation method, Zou and Yuan (2008) proposed composite QRe, Li and Zhu (2008) presented the QRe with l_1 penalty and Wu and Liu (2009) developed QRe employing the SCAD. Recently, from a Bayesian point of view, Li et al. (2010) proposed Bayesian Lasso QRe and Alhamzawi et al. (2012) suggested the adaptive Lasso QRe. In this paper, based on the Bayesian adaptive Lasso QRe (Alhamzawi et al., 2012), I propose the iterative adaptive Lasso QRe, which is an extension to the Expectation Conditional Maximization algorithm (Sun et al., 2010). The proposed method is demonstrated via simulation studies and a real data set. Results indicate that the new algorithm performs quite good in comparison to the other existing methods.

II.   Bayesian Adaptive Lasso Qre

In this section, we briefly summarize the Bayesian adaptive Lasso QRe reported in Alhamzawi et al. (2012). The Lasso method of Li and Zhu (2008) is defined as

$$min \sum_{i=1}^{n} \rho_p(y_i - x_i'\beta) + \lambda ||\beta||_1, \quad \lambda \geq 0, \qquad (5)$$

where $\lambda ||\beta||_1$ is the l_1 penalty. From a Bayesian point of view, Alhamzawi et al. (2012) put a Laplace prior on β_j takes the form

$$p(\beta_j | \tau, \lambda_j) = \frac{1}{2\sqrt{\tau}\lambda_j} \exp\{-\frac{|\beta_j|}{\sqrt{\tau}\lambda_j}\}, \qquad (6)$$

which can be reformulated as

$$p(\beta_j | \tau, \lambda_j) = \int_0^\infty N(\beta_j; 0, s_j) Exp(s_j; 2\tau\lambda_j^2) ds_j. \qquad (7)$$

Then, the authors assign an inverse gamma prior on the regularization parameter λ_j^2 with shape parameter ς and scale parameter ι. Here, ς and ι are positive hyperparameters and $p(\iota, \varsigma) \propto \iota^{-1}$. The authors also assume that the prior of τ is gamma with shape parameter $a_{01} = 0.1$ and rate parameter $a_{02} = 0.1$.

To summarise, the Bayesian hierarchical model of Alhamzawi et al. (2012) is given by

$$p(y_i, v_i | \boldsymbol{\beta}, \tau) = N(y_i; \boldsymbol{x}_i'\boldsymbol{\beta} + \xi v_i, 2\tau v_i) Exp(v_i; p(1-p)), \qquad (8)$$

$$p(\beta_j, s_j | \tau, \lambda_j^2) = N(\beta_j; 0, s_j) Exp(s_j; 2\tau\lambda_j^2), \qquad (9)$$

$$p(\lambda_j^2 | \varsigma, \iota) = InvGamma(\lambda_j^2; \varsigma, \iota), \qquad (10)$$

$$p(\tau) = Gamma(\tau; a_{01}, a_{02}), \qquad (11)$$

$$p(\iota, \varsigma) \propto \iota^{-1}. \qquad (12)$$

III.   Bayesian Iterative Adaptive Lasso Qre

One shortcoming of the Bayesian regularization techniques is that no point mass at zero is assigned in the Bayesian regularization techniques, thus updating the insignificant coefficients from the conditional distribution are never set exactly to zero. Therefore, some ad hoc techniques should be used to find the active coefficients in the regression. To overcome this shortcoming in linear regression models, Sun et al. (2010) suggested the iterative adaptive Lasso. In this paper, we use the same hierarchical model of Alhamzawi et al. (2012) and propose the iterative adaptive Lasso QRe, which is an ECM algorithm. In this paper, we set $\tau = 1$ and $\iota = \varsigma = 0.1$, (i.e., we treat ι and ς as fixed values) . The iterative adaptive Lasso QRe is implemented as follows:

• Fix the quantile level so that the pth quantile is modelled.

• Generate initial values for $(\beta_1, \cdots, \beta_k), (s_1, \cdots, s_k), (\lambda_1^2, \cdots, \lambda_k^2)$ and $(v_1, \cdots, v_n)$.

• Sample β_j by using a conditional maximization step:

The full conditional distribution of each β_j for $j = 1, \ldots, k$, is $N(\beta_j, \tilde{\sigma}_j^2)$, where $\tilde{\sigma}_j^2 = (\sum_{i=1}^n x_{ij}^2/(2v_i) + s_j^{-1})^{-1}$, and $\beta_j = \tilde{\sigma}_j^2 \sum_{i=1}^n x_{ij}(y_i - \sum_{l \neq j} x_{il}\beta_l - \xi v_i)/(2v_i)$

1. $\hat{\beta}_j = 0$ if $-\tilde{\sigma}_j^2/\lambda_j^2 \leq \tilde{\beta}_j \leq \tilde{\sigma}_j^2/\lambda_j^2$

2. $\hat{\beta}_j = \tilde{\beta}_j - \tilde{\sigma}_j^2/\lambda_j^2$ if $\tilde{\beta}_j > \tilde{\sigma}_j^2/\lambda_j^2$

3. $\hat{\beta}_j = \tilde{\beta}_j + \tilde{\sigma}_j^2/\lambda_j^2$ if $\tilde{\beta}_j < \tilde{\sigma}_j^2/\lambda_j^2$

• Updating s_j^{-1}

The full conditional distribution of each s_j^{-1} is an inverse Gaussian $IG(\mu_j', \lambda_j')$ wher $\mu_j' = \sqrt{1/(\lambda_j^2\beta_j^2)}$ and $\lambda_j' = 1/\lambda_j^2$.

• Updating λ_j^2

The full conditional distribution of λ_j^2 is an inverse Gamma $InvGa(1 + \varsigma, \frac{s_j}{2} + \iota)$.

• Updating v_j^{-1}

The full conditional distribution of each v_j^{-1} is $IG(\mu_i', \lambda')$, where $\mu_i' = |y_i - x_i'\beta|^{-1}$ and $\lambda' = 1/2$.

IV.   Simulation Studies

In this section, we carry out Monte Carlo simulations to study the performance of the proposed method (BIAL) with comparison to the standard frequentist QRe using the rq function (RQ) and the Bayesian Lasso QRe method (BLQ). Methods are evaluated based on median of mean absolute deviations (MMAD), i.e. MMAD=median$((\sum_{i=1}^{200}(|x_i'\hat{\beta} - x_i'\beta|))/200)$, where the median is taken over the 100 simulations.

4.1 Simulation 1

In this simulation study, we consider a very sparse model with k = 8 covariates from samples of size n_t = {30, 60, 120} for the training set and 100 for the testing set. Predictors were generated independently from a multivariate normal distribution N(0, Σ), where the (i, j)th element of Σ is $0.5^{|i-j|}$. Data are simulated from the model

$$y_i = 3x_{1i} + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ with $\sigma = \{2, 4\}$. Thus the true regression coefficients, are $\beta =$ *(3, 0, 0, 0, 0, 0, 0, 0)*. We centered the response variable to have mean zero, while the covariates have been standardized. The proposed MCMC algorithm is run for 17000 iterations, after a burn-in period of 1000 iterations. The results are summarized in Table 1 which clearly suggest that the proposed method outperforms the other methods in terms of prediction accuracy.

4.2 Simulation 2

In this simulation, the setup is the same as the Simulation 1, except we set β =(0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85) so that the outcome of interest was simulated according to the model

$$y_i = x_i'\beta + \varepsilon_i,$$

The results are summarized in Table 2. From Table 2, the performance of our proposed method appears quite good compared to the other two methods. We can observe that the BIAL tends to produce lower MMAD and standard deviations than the other methods.

Table 1: MMADs for Simulation 1. In the parentheses are standard deviations of the MMADs obtained by 500 bootstrap resampling.

n_t	σ	BIAL	RQ	BLQ
30	2	0.6434 (0.2455)	1.2095 (0.3485)	0.7203 (0.2535)
60	2	0.4961 (0.0749)	0.7333 (0.1987)	0.5680 (0.1403)
120	2	0.3226 (0.1168)	0.4446 (0.1248)	0.3879 (0.1241)
30	4	1.5801 (0.4373)	2.0468 (0.5479)	2.2256 (0.6694)
60	4	0.7438 (0.2345)	1.5851 (0.3161)	1.0657 (0.2794)
120	4	0.6643 (0.2053)	0.9536 (0.3815)	0.7988 (0.2226)

Table 2: MMADs for Simulation 2. In the parentheses are standard deviations of the MMADs obtained by 500 bootstrap resampling.

n_t	σ	BIAL	RQ	BLQ
30	2	0.7907 (0.1381)	1.0317 (0.5289)	0.8879 (0.2016)
60	2	0.5801 (0.1705)	0.6418 (0.1933)	0.5673 (0.1720)
120	2	0.3671 (0.1041)	0.5177 (0.1118)	0.3854 (0.0956)
30	4	1.3701 (0.3825)	2.2797 (0.8209)	1.8353 (0.5443)
60	4	1.2887 (0.1938)	1.6256 (0.4394)	1.4400 (0.3482)
120	4	0.7413 (0.2100)	0.9430 (0.2480)	0.7939 (0.2212)

4**.3 Simulation 3**
In this simulation, the setup is the same as the Simulation 1, except we set β =(1, 1, 0, 0, 1, 0, 0, 0) so that the outcome of interest was simulated according to the model

$$y_i = x_i'\beta + \varepsilon_i,$$

The results are summarized in Table 3. Table 3 summarizes our experimental results for this example. It can be seen from the table that the proposed method outperform the other methods.

Table 3: MMADs for Simulation 3. In the parentheses are standard deviations of the MMADs obtained by 500 bootstrap resampling.

n_t	σ	BIAL	RQ	BLQ
30	2	0.6566 (0.3583)	0.9812 (0.5085)	0.8496 (0.4289)
60	2	0.5266 (0.1607)	0.7419 (0.1807)	0.5850 (0.1425)
120	2	0.3048 (0.0747)	0.5127 (0.0887)	0.3449 (0.0843)
30	4	1.2627 (0.5597)	3.0051 (0.9707)	2.2745 (0.4675)
60	4	0.9154 (0.3034)	1.4902 (0.4770)	1.2459 (0.4326)
120	4	0.6547 (0.1466)	1.0747 (0.1809)	0.8144 (0.2079)

V. Stack Loss Data

In this paper, the popular data of Brownlee and Brownlee (1965), previously analysed by Yu and Moyeed (2001) and Reed (2011), is considered to demonstrate the performance of the new method. This data from the operation of a plant for the oxidation of ammonia to nitric acid. The data consists of 3 predictors and one outcome of interest. For each of these variables there are 21 observations. The three predictors are air flow (x_1), temperature of the cooling water (x_2) and acid concentration (x_3). The third variable (x_3) has been shown as inactive variable in the regression by many authors. The outcome of interest is the percentage of ammonia lost and is called "stack loss". Owing to the presence of outliers on this dataset, this dataset is of specific interest to us.

Table 4: Estimates and 95% intervals for QRe coefficients of the stack loss data, when $p \in \{0.50, 0.75, 0.95\}$. The proposed method (BIAL) is compared with the frequentist QRe approach (RQ) and the Bayesian Lasso QRe approach (BLQ).

		RQ	BLQ	BIAL
p	Parameters	Mean (lower bd, upper bd)	Mean (95% CrI)	Mean (95% CrI)
0.50				
	Intercept	-39.690 (-41.620, -29.678)	-38.669 (-54.022, -23.512)	-39.661 (-41.198, -36.561)
	x_1	0.832 (0.523, 1.141)	0.836 (0.611, 1.064)	0.841 (0.742, 1.116)
	x_2	0.574 (0.322, 1.411)	0.726 (0.218, 1.360)	0.587 (0.571, 1.284)
	x_3	-0.061 (-0.213, -0.029)	-0.113 (-0.326, 0.082)	0.000 (-0.002, 0.001)
0.75				
	Intercept	-54.190 (-61.163, 8.484)	-48.504 (-69.183, -22.674)	-54.229 (-57.293, -33.674)
	x_1	0.871 (0.533, 1.206)	0.861 (0.576, 1.142)	0.870 (0.671, 0.993)
	x_2	0.983 (-0.538, 1.782)	1.031 (0.193, 1.834)	0.981 (0.793, 1.130)
	x_3	0.000 (-0.517, 0.053)	-0.064 (-0.383, 0.193)	0.000 (-0.045, 0.003)
0.95				
	Intercept	-58.462 (-72.733, 1.797e+308)	-41.856 (-89.834, 42.046)	-58.453 (-73.119, -20.431)
	x_1	0.525 (-1.797e+308, 1.797e+308)	0.755 (0.142, 1.414)	0.525 (0.322, 0.720)
	x_2	1.858 (-1.797e+308, 6.211)	1.486 (-0.168, 2.921)	1.832 (1.479, 2.180)
	x_3	0.107 (-1.797e+308, 0.454)	-0.131 (-1.097, 0.554)	0.000 (-0.073, 0.019)

The proposed MCMC algorithm is run for 17000 iterations, after a burn-in period of 1000 iterations. In Table 4, we compare three methods: the standard frequentist QRe using the rq function, the Bayesian Lasso QRe method, and our approach. The methods are evaluated based on 95% intervals for p ∈ {0.50, 0.75, 0.95}. It can be observed that our QRe coefficients estimates are close to the standard QRe estimated and our credible intervals are much narrower than the intervals given by RQ and BLQ. In addition, it can be seen from the table that the third variable (x_3) has been ranked to zero using our method, while the other two approaches have different performance. Thus, the performance of our method consists with the literature.

VI.   Summary

Using the skewed Laplace distribution, which offers a mechanism for Bayesian quantile regression, we propose a fully Bayesian method for variable selection and estimation in quantile regression. We designed a simple and efficient algorithm for inference and illustrated the performance of the proposed approach compared with other existing approaches. Results show that our approach performs well compare to the other approaches.

References

[1]   Akaike, H. (1998). "Information theory and an extension of the maximum likelihood principle." In Selected Papers of Hirotugu Akaike, 199–213. Springer.
[2]   Alhamzawi, R. (2014). Bayesian elastic net tobit quantile regression. Communications in Statistics-Simulation and Computation (just-accepted), 00-00.
[3]   Alhamzawi, R. (2015). Model selection in quantile regression models. Journal of Applied Statistics 42 (2), 445-458.
[4]   Alhamzawi, R. (2017). Brq: Bayesian Analysis of Quantile Regression Models. R package version 2.0.
[5]   Alhamzawi, R. and H. T. M. Ali (2017). Bayesian quantile regression for ordinal longitudinal data. Journal of Applied Statistics, 1-14.
[6]   Alhamzawi, R. and K. Yu (2013). Conjugate priors and variable selection for bayesian quantile regression. Computational Statistics & Data Analysis 64, 209-219.
[7]   Al-Hamzawi, R. J. T. (2013). "Prior elicitation and variable selection for Bayesian quantile regression." Ph.D. thesis, Brunel University, School of Information Systems, Computing and Mathematics.
[8]   Alhamzawi, R., Yu, K., and Benoit, D. F. (2012). "Bayesian adaptive Lasso quantile regression." Statistical Modelling, 12(3): 279–297.
[9]   Benoit, D., Al-Hamzawi, R., Yu, K., and Van den Poel, D. (2011). "bayesQR: Bayesian quantile regression."
[10]   Benoit, D. F., Alhamzawi, R., and Yu, K. (2013). "Bayesian lasso binary quantile regression." Computational Statistics, 28(6): 2861–2873.
[11]   Brownlee, K. A. and Brownlee, K. A. (1965). Statistical theory and methodology in science and engineering, volume 150. Wiley New York.
[12]   Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." Journal of the American statistical Association, 96(456): 1348–1360.
[13]   Hashem, H., Vinciotti, V., Alhamzawi, R., and Yu, K. (2015). "Quantile regression with group lasso for classification." Advances in Data Analysis and Classification, 1–16.
[14]   Koenker, R. (2005). Quantile regression. 38. Cambridge university press.
[15]   Koenker, R. and Bassett, G. (1978). "Regression quantiles." Econometrica: journal of the Econometric Society, 33–50.
[16]   Koenker, R. W. and d'Orey, V. (1987). "Algorithm AS 229: Computing regression quantiles." Journal of the Royal Statistical Society. Series C (Applied Statistics), 36(3): 383–393.
[17]   Kozumi, H. and Kobayashi, G. (2011). "Gibbs sampling methods for Bayesian quantile regression." Journal of statistical computation and simulation, 81(11): 1565–1578.
[18]   Li, Q., Xi, R., and Lin, N. (2010). "Bayesian regularized quantile regression." Bayesian Analysis, 5(3): 533–556.
[19]   Li, Y. and Zhu, J. (2008). "L1-norm quantile regression." Journal of Computational and Graphical Statistics.
[20]   Nassiri, V. and Loris, I. (2014). "An efficient algorithm for structured sparse quantile regression." Computational statistics, 29(5): 1321–1343.
[21]   Reed, C. (2011). "Bayesian parameter estimation and variable selection for quantile regression." Ph.D. thesis, Brunel University, School of Information Systems, Computing and Mathematics.
[22]   Schwarz, G. (1978). "Estimating the dimension of a model." The annals of statistics, 6(2): 461–464.
[23]   Sun, W., Ibrahim, J. G., and Zou, F. (2010). "Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression." Genetics, 185(1): 349–359.
[24]   Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological), 267–288.
[25]   Waldmann, E., Kneib, T., Yue, Y. R., Lang, S., and Flexeder, C. (2013). "Bayesian semiparametric additive quantile regression." Statistical Modelling, 13(3): 223–252.
[26]   Wang, H. and Leng, C. (2012). "Unified LASSO estimation by least squares approximation." Journal of the American Statistical Association.
[27]   Wu, Y. and Liu, Y. (2009). "Variable selection in quantile regression." Statistica Sinica, 801–817.
[28]   Yu, K. and Moyeed, R. A. (2001). "Bayesian quantile regression." Statistics & Probability Letters, 54(4): 437–447.
[29]   Yuan, M. and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables." Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1): 49–67.
[30]   Zou, H. and Yuan, M. (2008). "Composite quantile regression and the oracle model selection theory." The Annals of Statistics, 1108–1126.