Noname manuscript No.
(will be inserted by the editor)

Quantile Regression with Group Lasso for Classification

H. Hashem · V. Vinciotti · R. Alhamzawi · K. Yu

Received: date / Accepted: date

Abstract Applications of regression models for binary response are very common and models specific to these problems are widely used. Quantile regression for binary response data has recently attracted attention and regularized quantile regression methods have been proposed for high dimensional problems. When the predictors have a natural group structure, such as in the case of categorical predictors converted into dummy variables, then a group lasso penalty is used in regularized methods. In this paper, we present a Bayesian Gibbs sampling procedure to estimate the parameters of a binary quantile regression model under a group lasso penalty. Simulated and real data show a good performance of the proposed method in comparison to mean-based approaches and to quantile-based approaches which do not exploit the group structure of the predictors.

Keywords quantile regression · binary regression · regularized regression · Gibbs sampling.

Mathematics Subject Classification (2000) 62H12 · 62F15

H. Hashem
Department of Mathematics, Brunel University London, UK

V.Vinciotti
Department of Mathematics, Brunel University London, UK
Tel.: +44 (0) 1895267469
E-mail: veronica.vinciotti@brunel.ac.uk

R. Alhamzawi
College of Arts, University of Al-Qadisiyah, Iraq

K. Yu
Department of Mathematics, Brunel University London, UK

1 Introduction

Quantile regression was introduced by Koenker and Bassett (1978) and has since then become the method of choice for regression problems where the data do not satisfy the normal distributional assumptions underlying traditional methods or when the data are subject to some form of contamination. One line of research has extended the original quantile regression model to the case where the response is binary, as an alternative to traditional mean-based models, such as logistic and probit regression models. The methods were originally developed in the frequentist estimation setting by Manski (1975, 1985) and were subsequently extended also to the Bayesian counterpart (Yu and Moyeed 2001; Benoit and Poel 2012; Miguéis et al 2012) as a means to avoid large-sample based asymptotic results for inference and at the same time take regression parameter uncertainty into account.

In recent years, with the advent of highly dimensional and complex datasets in many application areas, regularized regression methods have attracted a lot of attention. These methods impose a penalty on the size of the parameters, thus making it possible to estimate regression coefficients in the presence of a large number of variables and a relatively small number of observations. In the popular lasso regression model, this penalty takes the form of an L_1 penalty, which has the advantage of providing simultaneous parameter estimation and variable selection (Tibshirani 1996). The original lasso method was extended in a number of directions, amongst which adaptive lasso (Zou 2006; Alhamzawi et al 2012) and Cox regularized regression (Tibshirani 1997). In some cases, the predictors have a natural group structure, such as in the case of a categorial variable being converted into dummy variables. In these cases, the selection of groups of variables is of interest, rather than of individual variables. In order to address this type of problems, Yuan and Lin (2006) developed the group lasso method and a number of authors have subsequently extended it and studied its theoretical properties (Bach 2008; Huang and Zhang 2010; Wei and Huang 2010; Lounici et al 2011; Sharma et al 2013; Simon et al 2013). Given the merits of the regularized methods just described, regularized methods for binary response variables have also been developed. In particular, Bae and Mallick (2004); Genkin et al (2007); Gramacy and Polson (2012) developed Bayesian logistic regression models under a lasso or ridge penalty, Meier et al (2008) developed the classical logistic regression model under a group lasso penalty, and Krishnapuram et al (2005) developed a sparse multinomial logistic regression model.

The references above refer to the estimation of mean-based regression models. A small line of research has explored the link between the robust quantile regression models and the regularized models for high-dimensional data. In particular, Li and Zhu (2008) have developed quantile regression models under a lasso penalty, the theoretical properties of which are derived in Belloni and Chernozhukov (2011). Li et al (2010) provide a Bayesian formulation of the same problem. Finally, Ji et al (2012) have developed a quantile regression model under an L_1 penalty and for a binary response. In this paper, we extend

the work of Ji et al (2012) on binary quantile regression models with the use of a group lasso penalty. Our model is derived in the framework of probit binary regression and offers an alternative to the mean-based logistic regression model with group lasso penalty (Meier et al 2008), when the response is binary, the predictors have a natural group structure and quantile estimation is of interest. In section 2 we describe the model, in section 3 we describe the estimation of the parameters in a Bayesian setting, in section 4 we discuss how the model is used for prediction, in sections 5 and 6 we compare the performance of the method with related mean-based and quantile-based regression approaches on simulated and real data. Finally, in section 7, we draw some conclusions.

2 Binary quantile group lasso

Similar to a probit regression model, binary quantile regression models can be viewed as linear quantile regression models with a latent continuous response variables, e.g. Ji et al (2012). In particular, let y be the binary response variable, taking values 0 and 1, let x be the vector of p predictors, β the vector of unknown regression coefficients and (x_i, y_i), $i = 1, \ldots, n$ a sample of n observations on X and Y. Given a quantile θ, $0 < \theta < 1$, we consider the model:

$$y_i^\star = x_i^T \beta_\theta + u_i, \quad i = 1, \ldots, n \text{ and } y_i = h(y_i^\star),$$

where u_i are the errors, satisfying $P(u_i \leq 0 | x_i) = \theta$, and h is a link function. For binary response data, the link function is given by $h(y^\star) = I(y^\star > 0)$, with I the indicator function. In real applications, y is the observed binary response and the interest is to predict y from knowledge of x. $y^\star$ is unobserved and used mainly for modelling purposes. Some examples of $y^\star$ include the actual birth weight of babies in a study where the aim is to investigate the factors behind the birth of premature babies, the credit risk of a customer in a study where the aim is to discriminate between good and bad customers (Kordas 2002) or the willingness to participate to work in a study where the factors behind the decision to work or not are investigated (Kordas 2006).

The attractive property of this latent model is that there is a correspondence between the quantiles of y and the quantiles of $y^\star$, which are directly modelled. In particular, using the equivariance properties of quantile functions (Kordas 2006), it holds that

$$Q_{y|x}(\theta) = Q_{h(y^\star|x)}(\theta) = h(Q_{y^\star|x}(\theta)),$$

with $Q_{y|x}(\theta)$ denoting the θ conditional quantile of y given x. From this, since $Q_{y^\star|x}(\theta) = x^T \beta_\theta$ under a linear quantile regression model, it follows that

$$Q_{y|x}(\theta) = h(x^T \beta_\theta) = I(x^T \beta_\theta > 0).$$

So the estimation of the parameters β_θ leads to the knowledge about the θ quantile of y. In the next section, we describe how to estimate β_θ under a group lasso penalty.

3 Bayesian parameter estimation

In a binary quantile regression model, the parameter β_θ is found by the following minimization problem (Manski 1985):

$$\min_{||\beta||=1} \sum_{i=1}^{n} \rho_\theta(y_i - h(x_i^T \beta)), \tag{1}$$

where ρ_θ is the check function defined by

$$\rho_\theta(t) = \begin{cases} \theta t & \text{if } t \geq 0, \\ -(1-\theta t) & \text{if } t \leq 0 \end{cases}$$

or equivalently $\rho_\theta(t) = \dfrac{|t| + (2\theta - 1)t}{2}$. The restriction on $||\beta|| = 1$ is motivated by the fact that the scale of the parameter is not identifiable, being $y^\star$ a latent variable. Yu and Moyeed (2001) have shown how minimizing (1) is equivalent to maximising the likelihood function, under the assumption that the error comes from an asymmetric Laplace distribution with density given by $f_\theta(u) = \theta(1-\theta)\exp(-\rho_\theta(u))$. That is, minimising (1) is equivalent to maximising the likelihood

$$f(y|x, \beta, \theta) = \theta^n (1-\theta)^n \exp\Big(-\sum_{i=1}^{n} \rho_\theta(y_i - h(x_i^T \beta))\Big). \tag{2}$$

This fact has created a straightforward working model for Bayesian inference quantile regression.

When the predictors have a natural group structure, the methodology above can be extended to the use of a group lasso penalty. In particular, suppose that the predictors are grouped into G groups and β_g is the vector of coefficients of the g^{th} group of predictors. We denote with x_{ig} the i^{th} observation of the predictors in group g. Let $\beta = (\beta_1^T, \ldots, \beta_G^T)^T$ and $x_i = (x_{i1}^T, \ldots, x_{iG}^T)^T$, $i = 1, \ldots, n$. Under a group lasso constraint, the minimization in (1) becomes

$$\min_{||\beta||=1} \sum_{i=1}^{n} \rho_\theta(y_i - h(x_i^T \beta)) + \lambda \sum_{g=1}^{G} ||\beta_g||_{H_g}, \tag{3}$$

where λ is a non-negative regularization parameter, controlling the sparsity of the solution, and $||\beta_g||_{H_g} = (\beta_g^T H_g \beta_g)^{1/2}$ with $H_g = d_g I_{d_g}$ and d_g the dimension of the vector β_g. The choice of d_g in H_g has been suggested by Yuan and Lin (2006) to ensure that the penalty term is of the order of the variables in the group. Under an appropriate choice of prior distribution, the minimization problem in (3) can be shown to be equivalent to a maximum a posteriori solution. In particular, a Laplace prior on β_g is chosen, that is

$$\pi(\beta_g|\lambda) = C_{d_g} \sqrt{\det(H_g)} \lambda^{d_g} \exp(-\lambda ||\beta_g||_{H_g}), \tag{4}$$

where $C_{d_g} = 2^{-(d_g+1)/2}(2\pi)^{-(d_g-1)/2}/\Gamma((d_g+1)/2)$ and Γ is the gamma function. Then, using the same asymmetric Laplace distribution for the residuals u, the minimization in (3) is equivalent to the maximum of the posterior distribution

$$f(\beta|y,x,\lambda,\theta) \propto \exp\Big(-\sum_{i=1}^{n}\rho_\theta(y_i - h(x_i^T\beta)) - \lambda\sum_{g=1}^{G}||\beta_g||_{H_g}\Big),$$

under the constraint that $||\beta|| = 1$.

3.1 Gibbs sampling procedure

We extend the Gibbs sampling procedure of Ji et al (2012) to the case of a group lasso penalty. As a first step we rewrite the prior of β_g using the equality (Andrews and Mallows 1974)

$$\frac{a}{2}\exp(-a|z|) = \int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\Big(\frac{-z^2}{2s}\Big)\frac{a^2}{2}\exp\Big(-\frac{a}{2}s\Big)ds,$$

which holds for any $a \geq 0$. In particular, we take $a = \lambda$ and $z = ||\beta_g||_{H_g} = (\beta_g^T H_g \beta_g)^{1/2}$. Then the prior in (4) can be rewritten as

$$\begin{aligned}\pi(\beta_g|\lambda) &= C_{d_g}\sqrt{\det(H_g)}\lambda^{d_g}\exp(-\lambda||\beta_g||_{H_g}) \qquad (5)\\ &= \frac{(\lambda^2/2)^{(d_g+1)/2}}{\Gamma((d_g+1)/2)}\int_0^\infty \frac{\exp\Big(-\frac{1}{2}\beta_g^T(s_g H_g^{-1})^{-1}\beta_g\Big)s_g^{\frac{d_g-1}{2}}\exp\Big(\frac{-\lambda^2}{2}s_g\Big)}{\sqrt{\det(2\pi s_g H_g^{-1})}}ds_g.\end{aligned}$$

As a second step, we use the fact that an asymmetric Laplace distributed random variable can be written as a mixture of a $N(0,1)$ distributed random variable and an exponentially distributed random variable with rate parameter $\theta(1-\theta)$ (Alhamzawi and Yu 2013; Kozumi and Kobayashi 2011; Lum and Gelfand 2012). This allows to rewrite the likelihood (2) as:

$$\begin{aligned}f(y|x,\beta,\theta) \propto \exp\Big(-\sum_{i=1}^{n}\rho_\theta(u_i)\Big) &= \exp\Big(-\sum_{i=1}^{n}\frac{|u_i| + (2\theta-1)u_i}{2}\Big)\\ &= \prod_{i=1}^{n}\int_0^\infty \frac{1}{\sqrt{4\pi v_i}}\exp\Big(-\frac{(u_i-\xi v_i)^2}{4v_i} - \zeta v_i\Big)dv_i\end{aligned}$$

with $u_i = y_i - h(x_i^T\beta)$, $\xi = (1-2\theta)$ and $\zeta = \theta(1-\theta)$.

From this, we can derive the following conditional distributions:

$$f(\beta|y,x,\lambda,\theta) \propto \prod_{i=1}^{n}\int_0^\infty \frac{1}{\sqrt{4\pi v_i}}\exp\Big(-\frac{(u_i-\xi v_i)^2}{4v_i}-\zeta v_i\Big)dv_i \prod_{g=1}^{G}\exp\Big(-\lambda||\beta_g||_{H_g}\Big)$$

$$f(\beta|y,x,v,\lambda,\theta) \propto \exp\Big(-\sum_{i=1}^{n}\frac{(u_i-\xi v_i)^2}{4v_i}-\sum_{g=1}^{G}\lambda||\beta_g||_{H_g}\Big)$$

$$f(\beta_g|y,x,v,\beta_{-g},\lambda,\theta) \propto \exp\Big(-\sum_{i=1}^{n}\frac{(u_i-\xi v_i)^2}{4v_i}-\lambda||\beta_g||_{H_g}\Big).$$

If we write $\tilde{y}_{ig} = y_i - h(\sum_{k=1,k\neq g}^{G} x_{ik}^T\beta_k) - \xi v_i$, then using (5), we can write the conditional distribution of β_g as

$$f(\beta_g|y,x,v,\beta_{-g},\lambda,\theta) \propto \exp\Big(-\sum_{i=1}^{n}\frac{(\tilde{y}_{ig}-h(x_{ig}^T\beta_g))^2}{4v_i}-\frac{1}{2}\beta_g^T(s_gH_g^{-1})^{-1}\beta_g\Big).$$

The derivations above lead to the following Gibbs sampling procedure for the quantile θ:

1. Sample $y^\star$ from a truncated Normal distribution:
$$y_i^\star|y_i,x_i,\beta,v_i \sim \begin{cases} N(x_i^T\beta+\xi v_i, 2v_i)I(y_i^\star>0) & \text{if } y_i=1,\\ N(x_i^T\beta+\xi v_i, 2v_i)I(y_i^\star<0) & \text{if } y_i=0.\end{cases}$$
2. Sample v_i^{-1}, given $y_i^\star$, x_i and β, from an inverse Gaussian distribution with mean and shape parameters given by, respectively,
$$\mu = \sqrt{\frac{1}{(y_i^\star - x_i^T\beta)^2}} \text{ and } \eta = \frac{1}{2}.$$
3. Sample s_g, given β_g and λ, from an inverse Gaussian distribution with mean and shape parameters given by, respectively,
$$\mu = \sqrt{\frac{\lambda^2}{\beta_g^T H_g \beta_g}} \text{ and } \eta = \lambda^2.$$
4. Sample β_g, given $y^\star, x, \beta_{-g}, s_g, v$, from a multivariate normal distribution with mean and covariance given by
$$\mu_g = \Sigma_g x_g V(y^\star - (1-2\theta)v - x_{-g}^T\beta_{-g}) \text{ and } \Sigma_g = (x_gVx_g^T + s_g^{-1}H_g)^{-1},$$
respectively, where $V = \text{diag}\Big(\frac{1}{2v_i}\Big)$, $i=1,\ldots,n$, and x_g is the $d_g \times n$ matrix of observations for group g.
5. Sample λ^2, given s_g, from a Gamma distribution with shape and rate parameters given by
$$\alpha = \frac{p+G}{2}+b_1 \text{ and } \beta = \sum_{g=1}^{G}\frac{s_g}{2}+b_2$$
respectively, with b_1 and b_2 two non-negative constants which we set equal to 0.1.

4 Class prediction

The estimation of the regression coefficients indicates the most influential variables for the prediction of the binary outcome y. As with any classification problem, the main interest is in the prediction of y from a new instance x for which the binary outcome, or class, is unknown. In this section, we describe how the method that we propose is used to this purpose. The classification of an instance x is based on the estimated probability $P(y = 1|x)$. For our model:

$$\begin{aligned} P(y = 1|x, \beta, \theta) &= P(y^\star > 0|x, \beta, \theta) = P(x^T\beta_\theta + u > 0|x, \beta, \theta) \\ &= P(u > -x^T\beta_\theta|x, \beta, \theta) = 1 - P(u < -x^T\beta_\theta|x, \beta, \theta) \\ &= 1 - \Phi_{ALD}(-x^T\beta_\theta|x, \beta, \theta), \end{aligned}$$

where Φ_{ALD} is the cdf of an asymmetric Laplace distribution. Using the estimated β_θ from the binary quantile regression model in the formula above, we get a natural estimate of the posterior probability of x belonging to class 1. Since $P(u < 0|x) = \theta$, it follows that (Kordas 2006):

$$x^T\beta_\theta \gtreqless 0 \;\Leftrightarrow\; P(y = 1|x, \beta, \theta) \gtreqless 1 - \theta.$$

So there is a direct link before the estimated β_θ and the probability that $P(y = 1|x) = 1 - \theta$. In general, we can expect the error to have a median around 0, which motivates the choice of $\theta = 0.5$.

In Kordas (2006), a second approach is also considered, where $P(y = 1|x, \beta)$ is computed as an average over different quantiles θ. In particular, it holds that

$$P(y = 1|x, \beta) = \int_0^1 I(x^T\beta_\theta > 0)d\theta.$$

This probability can be estimated using a grid of values $\theta_1, \ldots, \theta_m$ and then taking

$$P(y = 1|x, \beta) \approx \frac{1}{m}\sum_{k=1}^{m} P(y = 1|x, \hat{\beta}_{\theta_k}),$$

with $\hat{\beta}_{\theta_k}$ the estimate of β for quantile θ_k.

As a final step in predicting y, we set a threshold t and classify a new object x to class 1 if

$$P(y = 1|x) > t.$$

The threshold t is normally chosen according to the relative misclassification costs for class 0 and 1 and corresponds to the case $t = 0.5$ for equal misclassification costs (Hand and Vinciotti 2003).

5 Simulation study

In this section, we investigate the performance of our method with a simulation study. As typical for these applications, we simulate the data from

$$y_i^\star = x_i^T \beta_\theta + u_i, \quad i = 1, \ldots, n \text{ and } y_i = h(y_i^\star),$$

with β chosen to have a group structure and with different choices of the error distribution. Similar to Yu et al (2013), we consider the following distributions for the error:

- Normal: $N(0, 1)$
- t-distribution with 1 degree of freedom (Cauchy): t_1
- Laplace distribution with location 0 and scale 10
- Skewed (`skew`): $\frac{1}{5}\, N\Big(-\frac{22}{25}, 1\Big) + \frac{1}{5}\, N\Big(-\frac{49}{125}, \Big(\frac{3}{2}\Big)^2\Big) + \frac{3}{5} N\Big(\frac{29}{250}, \Big(\frac{5}{9}\Big)^2\Big)$
- Kurtotic (`kur`): $\frac{2}{3}\, N(0, 1) + \frac{1}{3}\, N\Big(0, \Big(\frac{1}{10}\Big)^2\Big)$
- Bimodal (`bim`): $\frac{1}{2}\, N\Big(-1, \Big(\frac{2}{3}\Big)^2\Big) + \frac{1}{2}\, N\Big(1, \Big(\frac{2}{3}\Big)^2\Big)$
- Bimodal, with separate modes (`bim-sep`): $\frac{1}{2} N\Big(-\frac{3}{2}, \Big(\frac{1}{2}\Big)^2\Big) + \frac{1}{2} N\Big(\frac{3}{2}, \Big(\frac{1}{2}\Big)^2\Big)$
- Skewed Bimodal (`skew-bim`): $\frac{3}{4}\, N\Big(-\frac{43}{100}, 1\Big) + \frac{1}{4}\, N\Big(\frac{107}{100}, \Big(\frac{1}{3}\Big)^2\Big)$
- Trimodal (`tri`): $\frac{9}{20}\, N\Big(-\frac{6}{5}, \Big(\frac{3}{5}\Big)^2\Big) + \frac{9}{20}\, N\Big(\frac{6}{5}, \Big(\frac{3}{5}\Big)^2\Big) + \frac{1}{10} N\Big(0, \Big(\frac{1}{4}\Big)^2\Big)$.

These distributions were chosen to have a median close to or equal to zero. Figure 1 shows a plot of the density functions for the different cases considered. For the simulation, we set the sample size to $n = 50$.

For the β vector, we consider the case of a large number of predictors, i.e. $p \gg n$. Similar to Li et al (2010), we create a group structure by simulating 10 groups, each consisting of 10 covariates. The 100 variables are assumed to follow a multivariate normal distribution $N(0, \Sigma)$, with Σ having a diagonal block structure. Each block corresponds to one group and is defined by the matrix $r^{|i-k|}$, $i = 1, \ldots, 10$, $k = 1, \ldots, 10$. For the correlation r, we experiment both with $r = 0.95$ (well-defined group structure) and $r = 0.5$. For the β values, we consider two cases:

1. The values for the first three groups are given by

   $$(0.5, 1, 1.5, 2, 2.5, 2, 2, 2, 2, 2), (2, 2, 1, 1, 1, 1, 3, 3, 3, 3), (1, 1, 1, 2, 2, 2, 3, 3, 3, 3)$$

   and they are set to zero for all other groups.
2. $\beta_j = 0.85$ for all j.

We compare our method, Bayesian binary quantile regression with group Lasso penalty (`BBQ.grplasso`), with a frequentist mean-based logistic regression model under a lasso penalty (R package `glmnet` (Friedman et al 2010)), a frequentist mean-based logistic regression model under a group lasso penalty

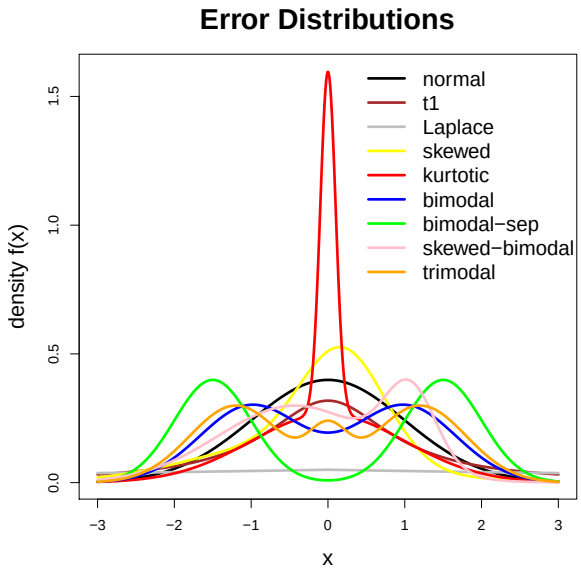

Fig. 1 Density functions of the errors considered in the simulation study.

(R package `grpreg` (Breheny and Huang 2014)) and a Bayesian binary quantile regression with a lasso penalty (R package `bayesQR` (Benoit and Poel 2012)). For `grpreg` and `glmnet`, the penalty parameter λ is selected using 5-fold cross-validation. For the Bayesian quantile methods and their Gibbs sampling procedures, we use 13000 iterations with the first 3000 iterations kept as burn-in. Furthermore, in the quantile methods, we use two methods to make class predictions, as described in Section 4: in the first case, we use the median ($\theta = 0.5$); in the second case we take an average of three quantiles, which we set as $\theta = 0.25, 0.5, 0.75$.

Tables 1 and 2 report the Area Under the Curve (AUC) values for the different methods and the different error distributions, with the AUC values averaged over 40 iterations and computed on a test set of the same size as the training set. In Table 1, we consider the first scenario for the β values and we set $r = 0.5$, whereas in Table 2 we consider the case of all βs equal to 0.85 and $r = 0.95$. Similar results were obtained in the other cases. No significant differences were found between the two approaches for prediction used for the Bayesian methods, namely that based on $\theta = 0.5$ and that based on the average of three quantiles. Tables 1 and 2 show how the `BBQ.grplasso` proposed in this paper significantly outperforms the other methods in all cases considered. Furthermore, the results show how `grpreg` is the worst performing method in all cases, surprisingly performing worse than `glmnet`, which is of a same nature but does not exploit the group structure of the predictors. The main competitor to `BBQ.grplasso` seems to be `bayesQR` which in fact differs with

Table 1 AUC values, averaged over 40 iterations (with standard deviations in brackets) for the case: n=50, p=100, r=0.5 and β values as in case (1). `BBQ.grplasso`: Bayesian binary quantile regression model proposed in this paper (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles); `grpreg`: frequentist mean-based logistic regression model with group lasso penalty, `glmnet`: frequentist mean-based logistic regression model under a group lasso penalty; `bayesQR`: Bayesian binary quantile regression with a lasso penalty (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles).

	`BBQ.grplasso` ($\theta = 0.5$)	`BBQ.grplasso` ($\theta = 0.25, 0.5, 0.75$)	`grpreg`	`glmnet`	`bayesQR` ($\theta = 0.5$)	`bayesQR` ($\theta = 0.25, 0.5, 0.75$)
N(0,1)	0.879	0.88	0.773	0.804	0.83	0.838
	(0.055)	(0.053)	(0.103)	(0.078)	(0.066)	(0.068)
t_1	0.838	0.838	0.725	0.765	0.787	0.797
	(0.06)	(0.06)	(0.116)	(0.116)	(0.066)	(0.068)
Laplace	0.766	0.764	0.64	0.718	0.722	0.727
	(0.089)	(0.089)	(0.125)	(0.113)	(0.087)	(0.089)
skew	0.885	0.886	0.792	0.811	0.832	0.837
	(0.043)	(0.043)	(0.087)	(0.086)	(0.049)	(0.051)
kur	0.89	0.888	0.775	0.816	0.843	0.846
	(0.049)	(0.05)	(0.112)	(0.082)	(0.058)	(0.052)
bim	0.898	0.898	0.782	0.789	0.853	0.857
	(0.05)	(0.05)	(0.098)	(0.117)	(0.066)	(0.065)
bim-sep	0.881	0.881	0.784	0.819	0.826	0.829
	(0.053)	(0.053)	(0.096)	(0.078)	(0.066)	(0.072)
skew-bim	0.885	0.886	0.797	0.814	0.838	0.848
	(0.054)	(0.055)	(0.099)	(0.09)	(0.067)	(0.066)
tri	0.879	0.879	0.774	0.822	0.82	0.829
	(0.053)	(0.054)	(0.117)	(0.074)	(0.064)	(0.061)

the proposed method only in the use of the lasso penalty in contrast to the group lasso penalty. The tables also show how the superiority of the Bayesian methods version the frequentist methods is particularly pronounced for the case of non-sparse coefficients (Table 2). This is well known in the literature, as Bayesian regularized methods do not return exact zeros for the parameter estimates.

Figure 2 confirms the results of the tables by showing the average ROC curve of the methods considered for two cases of error distributions. The figures show how the `BBQ.grplasso` outperforms the frequentist mean-based methods for all classification thresholds and has `bayesQR` as its main competitor.

6 Real application

In this section, we investigate the performance of the new method on five real applications:

– **Birth weight dataset:** This dataset is available in the R package `grpreg` and was used in Yuan and Lin (2006). The data record the birth weights of 189 babies, together with eight predictors. Among the predictors, two are continuous (mother's age and weight) and six are categorical (mother's race, smoking status during pregnancy, number of previous premature labours, history of hypertension, presence of uterine irritability, number

Table 2 AUC values, averaged over 40 iterations (with standard deviations in brackets) for the case: n=50, p=100, r=0.95 and β values as in case (2). `BBQ.grplasso`: Bayesian binary quantile regression model proposed in this paper (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles); `grpreg`: frequentist mean-based logistic regression model with group lasso penalty, `glmnet`: frequentist mean-based logistic regression model under a group lasso penalty; `bayesQR`: Bayesian binary quantile regression with a lasso penalty (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles).

	`BBQ.grplasso` ($\theta = 0.5$)	`BBQ.grplasso` ($\theta = 0.25, 0.5, 0.75$)	`grpreg`	`glmnet`	`bayesQR` ($\theta = 0.5$)	`bayesQR` ($\theta = 0.25, 0.5, 0.75$)
N(0,1)	0.962	0.962	0.606	0.895	0.928	0.943
	(0.026)	(0.027)	(0.104)	(0.046)	(0.042)	(0.046)
t_1	0.943	0.943	0.572	0.878	0.911	0.923
	(0.033)	(0.033)	(0.096)	(0.07)	(0.048)	(0.044)
Laplace	0.872	0.872	0.57	0.784	0.834	0.848
	(0.048)	(0.048)	(0.08)	(0.1)	(0.047)	(0.047)
skew	0.96	0.958	0.602	0.888	0.925	0.944
	(0.025)	(0.026)	(0.104)	(0.063)	(0.038)	(0.039)
kur	0.954	0.955	0.646	0.876	0.917	0.941
	(0.03)	(0.031)	(0.1)	(0.087)	(0.043)	(0.037)
bim	0.963	0.962	0.561	0.901	0.924	0.94
	(0.033)	(0.034)	(0.086)	(0.067)	(0.045)	(0.048)
bim-sep	0.967	0.966	0.609	0.899	0.935	0.948
	(0.029)	(0.028)	(0.119)	(0.068)	(0.036)	(0.041)
skew-bim	0.969	0.968	0.592	0.912	0.935	0.951
	(0.019)	(0.02)	(0.107)	(0.048)	(0.033)	(0.025)
tri	0.96	0.959	0.601	0.89	0.928	0.94
	(0.036)	(0.036)	(0.101)	(0.071)	(0.047)	(0.042)

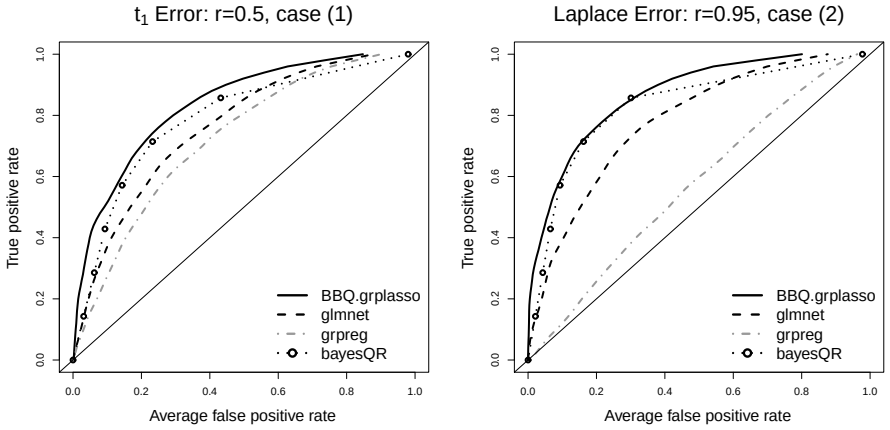

Fig. 2 Average ROC curves (over 40 iterations) of Bayesian binary quantile regression with group lasso (`BBQ.grplasso`, $\theta = 0.5$), compared with `grpreg`, `glmnet` and `bayesQR`, under a t_1 (left) and a Laplace (right) error distribution.

of physician visits). Through the use of orthogonal polynomials and dummmy variables, the data is converted into 17 predictors. The goal of this study is to identify the risk factors associated with giving birth to a low birth weight baby (defined as weighing less than 2500g).

- **Colon dataset:** This dataset is available in the R package `gglasso` and was used by Yang and Zou (2013). The data report the expression level of 20 genes from 62 colon tissue samples, of which 40 are cancerous and 22 normal. In Yang and Zou (2013), the 20 expression profiles are expanded using 5 basis B-splines, creating a dataset with 100 predictors and a group structure.
- **Labor force participation dataset:** This dataset is available in the R package `AER` and was used by Liu et al (2013). The data come from the panel study of income dynamics in 1976 and contain 753 observations on women's labour supply and 18 variables. The response variable, wife's participation in work, is a binary variable and the predictors have a group structure, as defined in (Liu et al 2013). The aim of this analysis is to assess if there is a relation between several social factors (wife's age, husband's wage, wife's father education etc.) and wife's participation in work.
- **Splice site detection dataset**: This dataset is available in the R package `grplasso` and is a random sample of the data used in (Meier et al 2008). It contains information on 200 true human donor splice sites and 200 false splice sites. For each site, the data report the last three bases of the exon and the third to sixth bases of the intron. Thus, the data contain 7 categorical predictors, with values A,C, G and T. These are converted into dummy variables, creating a natural group structure.
- **Cleveland heart dataset**: This dataset is available from the UCI machine learning repository. The data report information on 297 patients, 160 of whom have been diagnosed with heart disease and the remaining 137 have not been diagnosed with heart disease. The goal of the study is to predict heart disease from 13 predictors, related to patients' characteristics (age, sex,etc) and clinical information (blood pressure, cholesterol level, etc). Four of the predictors are categorical and have been converted into dummy variables, creating a group structure.

Table 3 reports the AUC values of 5-fold cross validation ROC curves, averaged over 30 iterations. As before, we compare the binary quantile regression method presented in this paper, `BBQ.grplasso`, with `grpreg`, `glmnet` and `BayesQR`. The results show how `BBQ.grplasso` is superior to `BayesQR` on all datasets, it outperforms the frequentist methods in the Birth and Colon datasets, but has comparable performances on the remaining datasets. Combined with the simulation study, this is probably a reflection of high levels of sparsity in the underlying model.

Table 3 AUC values, averaged over 30 iterations (with standard deviations in brackets) on the real data. `BBQ.grplasso`: Bayesian binary quantile regression model proposed in this paper (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles); `grpreg`: frequentist mean-based logistic regression model with group lasso penalty, `glmnet`: frequentist mean-based logistic regression model under a group lasso penalty; `bayesQR`: Bayesian binary quantile regression with a lasso penalty (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles).

Dataset	`BBQ.grplasso` ($\theta = 0.5$)	`BBQ.grplasso` ($\theta = 0.25, 0.5, 0.75$)	`grpreg`	`glmnet`	`bayesQR` ($\theta = 0.5$)	`bayesQR` ($\theta = 0.25, 0.5, 0.75$)
Birth	0.593	0.595	0.573	0.577	0.563	0.589
	(0.027)	(0.027)	(0.038)	(0.04)	(0.029)	(0.029)
Colon	0.662	0.662	0.649	0.631	0.626	0.638
	(0.06)	(0.059)	(0.072)	(0.07)	(0.064)	(0.059)
Labor	0.696	0.696	0.695	0.699	0.508	0.563
	(0.016)	(0.015)	(0.018)	(0.024)	(0.011)	(0.049)
Splice	0.697	0.696	0.695	0.693	0.669	0.687
	(0.017)	(0.019)	(0.021)	(0.019)	(0.019)	(0.022)
Heart	0.666	0.666	0.665	0.665	0.512	0.592
	(0.025)	(0.025)	(0.027)	(0.025)	(0.017)	(0.038)

7 Conclusion

In this paper, we present a novel method for binary regression problems where the predictors have a natural group structure, such as in the case of categorical variables. In contrast to existing methods for group-typed variables, we model the quantiles of the response variable, in order to account for possible departures from normality in the latent variable. In particular, we focus on class prediction and show how the probability of a new object x belonging to class 1, $p(1|x)$, is directly linked to the quantile of the latent variable, since $P(1|x) = P(y^\star > 0|x)$. This motivates the use of quantile-based regression for probit regression models.

We compare our method with a frequentist mean-based logistic regression model, under a lasso and a group lasso penalty, and with a Bayesian quantile-based regression model under a lasso penalty, on simulated and real data. The simulation shows a number of scenarios where the method outperforms the mean-based and quantile-based approaches. The R code of the method described in this paper is available from `http://people.brunel.ac.uk/~mastvvv/Software`. Future research will consider an extension of this method to include a variable selection prior, similarly to the method of Alhamzawi and Yu (2013) for Bayesian quantile regression.

References

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

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

Andrews DF, Mallows CL (1974) Scale mixtures of normal distributions. Journal of the Royal Statistical Society, Series B 36:99–102

Bach F (2008) Consistency of the group lasso and multiple kernel learning. Journal of Machine Learning Research 9:1179–1225
Bae K, Mallick B (2004) Gene selection using a two-level hierarchical bayesian model. Bioinformatics 20(18):3423–3430
Belloni A, Chernozhukov V (2011) Post l_1-penalized quantile regression in high-dimensional sparse models. Annals of Statistics 39:82–130
Benoit D, Poel D (2012) Binary quantile regression: a bayesian approach based on the asymmetric laplace density. Journal of Applied Econometrics 27(7):1174–1188
Breheny P, Huang J (2014) Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. To appear
Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 31(1):1–22
Genkin A, Lewis DD, Madigan D (2007) Large-scale bayesian logistic regression for text categorization. Technometrics 49(14):291–304
Gramacy R, Polson N (2012) Simulation-based regularized logistic regression. Bayesian Analysis 7(3):503–770
Hand D, Vinciotti V (2003) Local versus global models for classification problems: fitting models where it matters. The American Statistician 57(2):124–131
Huang J, Zhang T (2010) The benefit of group sparsity. The Annals of Statistics 38(4):1978–2004
Ji Y, Lin N, Zhang B (2012) Model selection in binary and tobit quantile regression using the gibbs sampler. Computational Statistics and Data Analysis 56(4):827 – 839
Koenker R, Bassett GW (1978) Regression quantiles. Econometrica 46:33–50
Kordas G (2002) Credit scoring using binary quantile regression. Statistics for Industry and Technology pp 125–137
Kordas G (2006) Smoothed binary regression quantiles. Journal of Applied Econometrics 21(3):387–407
Kozumi H, Kobayashi G (2011) Gibbs sampling methods for bayesian quantile regression. Journal of Statistical Computation and Simulation 81:1565–1578
Krishnapuram B, Carin L, Figueiredo MA, Hartemink AJ (2005) Sparse multinomial logistic regression: Fast algorithms and generalization bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence 27:957–968
Li Q, Xi R, Lin N (2010) Bayesian regularized quantile regression. Bayesian Analysis 5:1–24
Li Y, Zhu J (2008) L1-norm quantile regressions. Journal of Computational and Graphical Statistics 17:163–185
Liu X, Wang Z, Wu Y (2013) Group variable selection and estimation in the tobit censored response model. Computational Statistics and Data Analysis 60:80–89
Lounici K, Pontil M, Tsybakov A, van de Geer S (2011) Oracle inequalities and optimal inference under group sparsity. Annals of Statistics 39:2164–2204
Lum K, Gelfand A (2012) Spatial quantile multiple regression using the asymmetric laplace process. Bayesian Analysis 7(2):235–258
Manski C (1975) Maximum score estimation of the stochastic utility model of choice. Journal of Econometrics 3(3):205–228
Manski C (1985) Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. Journal of Econometrics 27(3):313–333
Meier L, van de Geer S, Bühlmann P (2008) The group lasso for logistic regression. Journal of the Royal Statistical Society, Serie B 70(1):53–71
Miguéis LV, Benoit DF, Van den Poel D (2012) Enhanced decision support in credit scoring using bayesian binary quantile regression
Sharma D, Bondell H, Zhang H (2013) Consistent group identification and variable selection in regression with correlated predictors. Journal of Computational and Graphical Statistics 22(2):319–340
Simon N, Friedman J, Hastie T, Tibshirani R (2013) A sparse-group lasso. Journal of Computational and Graphical Statistics 22(2):231–245
Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58:267–288
Tibshirani R (1997) The lasso method for variable selection in the cox model. Statistics in Medicine 16:385–395

Wei F, Huang J (2010) Consistent group selection in high-dimensional linear regression. Statistics in Medicine 16:1369–1384

Yang Y, Zou H (2013) A fast unified algorithm for solving group-lasso penalized learning problems

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

Yu K, Cathy C, Reed C, Dunson D (2013) Bayesian variable selection in quantile regression. Statistics and Its Interface 6:261–274

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68:49–67

Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101:1418–1429