Regularized Bayesian Quantile Regression

Salaheddine El Adlouni[a,*], Garba Salaou[a], André St-Hilaire[b]

[a] Mathematics and Statistics departement, Universit de Moncton,
18 avenue Antonine-Maillet Moncton (NB) E1A 3E9 Canada ;
[b]Centre Eau Terre Environnement, INRS-ETE, 490 rue de la Couronne
Québec (Québec) G1K 9A Canada

Abstract A number of non-stationary models have been developed to estimate extreme events as function of covariates. Quantile Regression (QR) model is a statistical approach intended to estimate, and conduct inference about the conditional quantile functions. In this article, We focus on the simultaneous variable selection and parameter estimation through penalized quantile regression. We conducted a comparison of regularised Quantile Regression model with B-Splines in Bayesian framework. Regularisation is based on penalty and aims to favour parsimonious model, especially in the case of large dimension space. The prior distributions related to the penalties are detailed. Five penalties (Lasso, Ridge, SCAD0, SCAD1 and SCAD2) are considered with their equivalent expressions in Bayesian framework. The regularized quantile estimates are then compared to the maximum likelihood estimates with respect to the sample size. A Markov Chain Monte Carlo (MCMC) algorithms are developed for each hierarchical model to simulate the conditional posterior distribution of the quantiles. Results indicate that the SCAD0 and Lasso have the best performance for quantile estimation according to Relative Mean Biais (RMB) and the Relative Mean- Error (RME) criteria, especially in the case of heavy distributed errors. A case study of the annual maximum precipitation at Charlo, Eastern Canada, with the Pacific North Atlantic climate index as covariate is presented.

Keywords : asymmetric Laplace distribution ; Bayesian inference ; B-splines ; Lasso ; quantile regression ; Ridge ; SCAD.

[*] Corresponding author. Email : salah-eddine.el.adlouni@umoncton.ca

1 Introduction

Least absolute deviation regression (*LADR*) allows to estimate the conditional median function. Compared to ordinary least square (*OLS*), the *LADR* has the advantage to be more robust to outliers. Koenker and Basset (1978) generalized the idea of *LARD* and introduced quantile regression (*QR*) to estimate the conditional quantile function given the covariates. *QR* has attracted researcher and has undergone several development and been applied to many areas (Huang, 2015).

For given probability $p \in]0, 1[$, the quantile regression function is defined as : $\hat{Q}_p = g_{\hat{\beta}}(x_i)$ with

$$\hat{\beta} = \sum_{i=1}^{n} \rho_p(y_i - g_\beta(x_i)); \quad \beta \in R^d \tag{1}$$

where $x_i \in R^d$, $i = 1, ..., n$ the vector of covariates and $g_\beta : R^d \to R$ is the parametric function and ρ_p is the loss function.

$$\rho_p(t) = t(p - 1_{(t<0)}) = \begin{cases} tp, & \text{if} \quad t \geq 0 \\ t(p-1), & \text{if} \quad t < 0 \end{cases} \tag{2}$$

with $p \in]0, 1[$, and asymmetric weight on positive and negative residuals. Since its first introduction, *QR* has attracted researcher and has been applied in many area such as finance to estimate the value at risk (Engel and Manganelli, 2004) ; ecology in the presence of complex interaction between covariates (Cade and Noon, 2003) ; economics (Koenker and Hallock 2001) among others. However the use of the *QR* models in somme areas, such as hydro-meteorological studies, remains very limited.

In classical *QR* model, the parameter are estimated by solving the optimization problem 1. Indeed, Koenker and Bassett (1978) show that minimizing the loss function for given sample, leads to the p^{th} regression quantile. they converted problem 1 to linear program and give a detailed procedure to solve it. Even in the absence of any model error, Koenker and Bassett (1978) show the normality of estimators $\hat{\beta}$ and give its asymptotic variance.

Yu an Moyeed (2001) suggested the use of the asymmetric Laplace distribution (*ALD*) as residual distribution and show the equivalence between the quantile regression optimisation problem and the maximum likelihood estimates of the ALD distribution. This representation of the QR estimation problem offers several advantages that are also associated with the Generalized Linear Model. Indeed, $y_i = x_i^T\beta + \epsilon_i$, where $(\epsilon_i)_{i=1,...,n}$ are Independent and identically distributed (iid) with standard Asymmetric Laplace Distribution (ALD). This model can then be developed in a fully Bayesian framework.

When estimating the parameters of the QR model without any constraint, the estimates often have low bias but large variance. As in the case of Ordinary Least Squares (OLS) approach, Shrinkage, or setting to zero some coefficients, can improve the prediction accuracy. This is equivalent to the selection of a subset of factors with the strongest effects.

Tibshirani (1996) proposed ”Least Absolute Shrinkage and Selection Operator” ($Lasso$) to have a continuous process for parameter shrinkage and stable prediction. Indeed, the alternative based on variable selection may lead to an unstable model because of dichotomic decision to retain or remove a factor from the model. Very different models could be selected due to small change in the data.

Shrinkage allows to stabilize the solution especially when the dimension of β is large. Most popular shrinkage approaches add a regularization penalty to the objective function. The principal regularization techniques uses different penalties, for example the $Lasso$ estimator (Tibshirani, 1996) uses the L_1 norm based penalty $||\beta||_1 = \sum_{k=1}^{d} |\beta_k|$. The Ridge uses the L_2 norm based penalty $||\beta||_2 = \sum_{k=1}^{d} \beta_k^2$ (Friedman and Hastie, 2000). The elastic net, uses a mixture of the L_1 Lasso and L_2 (Ridge) penalties.

For a given penalty $P_l(\beta)$, the quantile regression estimation corresponds to the resolution of optimization problem :

$$\hat{\beta} = argmin_{\beta \in R^d} \sum_{i=1}^{n} \rho_p(y_i - g_\beta(x_i)) + \lambda P_l(\beta) \tag{3}$$

Where λ is a tuning parameter.

2 Bayesian Quantile Regression

2.1 The Asymmetric Laplace Distribution

As mentioned in the introduction, Yu and Moyeed (2001) suggest the use of the Asymmetric Laplace Distribution (*ALD*) where the shape parameter corresponds to the probability p of the quantile to be estimated. The probability density function pdf of the $ALD(\mu, \sigma, p)$ distribution is given by :

$$f(\mathbf{y}) = p(1-p)\exp(-\frac{1}{\sigma}\rho_p(y-\mu)) \tag{4}$$

For a given sample $(y_i)_{i=1,\ldots,n}$ and $(x_i^{'})_{i=1,\ldots,n}$ where $y_i \in R^d$ is a realization of the vector of covariates, and if $(y_i|g_\beta(x_i))_{i=1,\ldots,n}$ are *iid*, with $ALD(g(x_i), \sigma, p)$. Then the likelihood is given by :

$$l_n(y) = p^n(1-p)^n \exp\left\{-\frac{1}{\sigma}\sum_{i=1}^{n}\rho_p\left(y_i - g_\beta(x_i)\right)\right\} \tag{5}$$

Note that when $y_i|g(x_i) \sim ALD(g(x_i), \tau, p)$ then the $p - quantile$ is : $Q_p = g(x_i)$.

The log-likelihood is given by :

$$\log(l_n(y)) = \log p^n(1-p)^n - \frac{1}{\sigma}\sum_{i=1}^{n}\rho_p\left(y_i - g_\beta(x_i)\right)$$

This maximisation is equivalent to the minimisation of $\sum_{i=1}^{n}\rho_p\left(y_i - g_\beta(x_i)\right)$. This equivalence allows to make the inference for the *QR* problem through statistical properties of the *ALD* distribution. This pseudo-likelihood is considered by Yu and Moyeed (2001) to link the *QR* in Bayesian framework to the frequentist approach. They used an improper prior $\pi(\beta) \propto 1$ on the regression parameter β, and showed that the posterior distribution is proper. The posterior mode, which corresponds to the Bayesian estimator under absolute loss function, is the same as frequentist solutions.

2.1.1 ALD and regularization

In Bayesian framework, the regularization techniques can be introduced through the parameter priors. For example, in the case of Lasso penalty (Equation 3), Tibshirani (1996) suggests that

Lasso estimate can be interpreted as posterior mode, when the regression parameter have *iid* Laplace priors (i.e double-exponential (*DE*)).

Park and Casella (2008) consider a fully Bayesian analysis using a conditional Laplace prior. For the ridge-penalty, the corresponding prior is a scaled student distribution. Fahrmeir and al. (2010) presented a review on the regularization penalties and related Bayesian priors. Other examples of frequent regularization penalties see (Zou, 2006 ; Kneib et al., 2009 ; Bondell and Reich, 2008). Hanwen and Chen (2015) propose a Bayesian framework to combine weighted composite quantile and Lasso regularization together to perform estimation and variable selection simultaneously (Alhamzawi, 2015).

2.1.2 Non-linear dependence

We consider hereafter that the regression curve is modelled with a B-spline function of given degree l and m knots at the locations γ_k, $k = 1, ..., m$. Then the quantile regression model for probability $p \in]0, 1[$ can be represented by :

$$Y_i = g_\beta(x_i) + \epsilon_i, \quad i = 1, ..., n \tag{6}$$

Where$(\epsilon_i)_{i=1,...,n}$ are the *iid* draws from $ALD(0, \sigma, p)$.

Under a B-spline representation of the fitting curve of the *QR* model, 6 can be written as linear model :

$$g_\beta(x) = \beta_0 + \sum_{j=1}^{m+l} \beta_j B_{j,l}(x), \tag{7}$$

where,

m the numbers of knots, l the degree, β_j is the control points and in a regression setting will be the coefficients of the regression model, B is the spline basis function with :

– for $j = 0, m - 2$

$$B_{j,0} = \begin{cases} 1 & \text{si } x_j \leq x \leq x_{j+1} \\ 0 & \text{else} \end{cases} \tag{8}$$

– for $j = 0, m - l - 2$

$$B_{j,l}(x) = \frac{x - x_j}{x_{j+d} - x_j} B_{j,l-1}(x) + \frac{x_{j+l+1} - x}{x_{j+l+1} - x_{j+1}} B_{j+1,l-1}(x) \tag{9}$$

We denote : $B_j(x; q)$ the value at point x of j^{th} B-spline with the knots are equidistant and the dimension of the parameters' space is $d = m + l + 1$.

In the following sections, we present the penalties, for a regularized inference, with their corresponding prior distribution for Bayesian implementation.

2.2 Lasso penalty

2.2.1 Lasso prior

The Lasso is commonly used as regularization penalty and can be represented, in a Bayesian framework, as a double exponential (DE) prior distribution (Park and Casella, 2008) :

$$\pi(\beta) = \prod_{j=1}^{d} \frac{\lambda}{2} exp[-\lambda|\beta_j|] \tag{10}$$

where λ is regularization parameter.

The DE distribution prior can also be presented as two-level hierarchical model with $(\beta_j|\tau)_{j=1,\dots,d}$ are independent normal distribution and $\tau^2|\lambda$ has an exponential distribution with pdf :

$$\pi(\tau^2|\lambda) = \frac{\lambda^2}{2} exp[-\lambda^2\tau^2/2] \tag{11}$$

The (τ) is assumed to be the same for all the B-spline parameters. The marginal posterior distributions will be implicitly dependent.

2.2.2 Bayesian hierarchical model for Lasso

The hierarchical model corresponding to the quantile regression model with Lasso penalty is as follows :

Model

$$y \sim ALD(g_\beta(x), \sigma, p), \quad \text{for} \quad p \in]0, 1[,$$

Penalty

$$p_l(\beta) = \sum_{j=1}^{d} | \beta_j | \quad (L_1 - norm)$$

Prior distribution

$$\begin{aligned} \beta_j \mid \tau \sim N(0, \tau^2), \quad j = 1, ..., d \\ \tau^2 \mid \lambda \sim Exp\{\frac{\lambda^2}{2}\} \\ \lambda^2 \sim Gam(a, b) \end{aligned}$$

Posterior distribution

$$\pi(\beta, \tau, \lambda \mid \mathbf{y}) \propto \prod_{i=1}^{n} f_{ALD}(y_i; g_\beta(x), \sigma, p) \prod_{j=1}^{p} \pi(\sigma)\pi(\beta_j \mid \tau^2)\pi(\tau^2 \mid \lambda)\pi(\lambda^2).$$

The mixture of the first two priors define the Double-exponential distribution as prior of the parameters $(\beta_j)_{j=1,\dots,d}$ and prior distribution of the parameter λ is deduced from the conjugate formulation of an Exponential distribution.

The hyper-parameters a and b should be chosen adequately by model selection criteria. Generally, in the absence of any additional prior distribution, we attribute to the parameters a and b values that leads to a large variance of the parameter λ in order to cover the entire parameter space. The selection of these hyper-parameters, will be illustrated in the simulation study and the case study with meteorological data.

2.2.3 MCMC Algorithm for the Lasso prior

For given initial values of $\lambda(1)$, $\tau(1, :)$ and fixed values of the hyper-parameters a and b and the u^{th} iteration values, then the $(u + 1)^{th}$ iteration is given by :

Algorithm 1 Lasso

```
Step 1 : initial values
$\lambda(1)$;    $\tau(1)$ ;
for $k = 2 : N$ do
Step 2 : Proposal distributions
Generate $\lambda_0$ ; from $\mathcal{N}(\lambda(k-1), \sigma_\lambda^2)$
    for $j = 1 : d$ do
        Generate $\tau_0$ from $Exp(\lambda^2/2)$ ;
        Generate $\beta_0$ from $\mathcal{N}(0, {\tau_0}^2)$ ;
    end for
Step 3 : Calculate the Hastings ratio
    $$r(\beta(k-1), \beta_0) = \frac{\pi(\beta_0 \mid y, X, \sigma, p, \tau_0, \lambda_0)}{\pi(\beta(k-1) \mid y, X, \sigma, p, \tau(k-1), \lambda(k-1))}$$
    where
    $$\pi(\beta \mid y, X, \sigma, p, \tau, \lambda) = f_{ALD}(y \mid g_\beta(x), \sigma, p)\pi(\beta \mid \tau)\pi(\tau^2 \mid \lambda^2/2)\pi(\lambda^2; a, b)$$
Accept the proposed move $\beta_0$ with probability
    $$\alpha(\beta(k-1), \beta_0) = min(1, r(\beta(k-1), \beta_0))$$
or remain in the actual state $\beta(k-1)$ with probability $(1 - \alpha(\beta(k-1), \beta_0))$.
end for
With $N$ the length of the Markov Chain and $d$ the dimension of the parameter space
```

2.3 Ridge penalty

2.3.1 Ridge prior

As stated previously, the ridge regression the penalty correspond to the $l_2 - norm$.

$$P_l(\beta) = \sum_{j=1}^{d} \beta_j^2$$

The Bayesian formulation is then given by the posterior distribution :

$$l(\beta|data) \propto \log(p^n)(1-p)^n) - \sum_{i=1} n\rho_p(y_i - g_\beta(x_i)) - \lambda \sum_{j=1}^{d} \beta_j^2$$

Hence the last term is equivalent to a normal prior. Usually, the smoothing parameter λ is given by $\lambda = \frac{1}{\tau^2}$ (Fahmeir et al. 2010)). Criteria such as cross validation could be considered for smoothing selection.

For computational and analytical purposes, the ridge penalty prior can be present as a two-level hierarchical model (Griffin and Brown, 2005). The first level, assumes that the coefficients β_j follows independent normal distributions with mean zero and unknown variances τ^2. (suppose without loss of generality $\beta_0 = 0$)

$$\beta|\tau \sim \prod_{j=1}^{d} N(\beta_j|0, \tau^2) \tag{12}$$

At the second level, the variances τ^2 are assumed to follow the Inverse Gamma distribution $\mathcal{IG}(\nu, S^2)$ then :

$$\pi(\tau \mid \nu, S^2) \propto (\tau^2)^{-\nu/2+1} \exp\{-\nu S^2/\tau^2\} \tag{13}$$

Note that the marginal distribution of $(\beta_j)_{j=1,\ldots,d}$ will be given by a student t-distribution.

2.3.2 Bayesian hierarchical model for Ridge

Model

$$y \sim ALD(g_\beta(x), \sigma, p), \quad \text{for} \quad p \in]0, 1[,$$

Penalty

$$p_\lambda(\beta) = \lambda \sum_{j=1}^{d} \beta_j^2 \quad \text{with} \quad L_2 - norm$$

Prior distributions

$$\beta_j \mid \tau \sim \prod_{j=1}^{d} N(\beta_j \mid 0, \tau^2),$$

$$\tau^2 \sim \mathcal{IG}(\nu/2, S^2).$$

Posterior distribution

$$\pi(\beta, \tau \mid \mathbf{y}, \nu, S^2) \propto \prod_{i=1}^{n} f_{ALD}(y_i \mid g_\beta(x), \sigma, p) \prod_{j=1}^{d} \pi(\beta_j \mid \tau^2) \pi(\tau^2 \mid \nu, S^2)$$

The MCMC algorithm is equivalent to that given for the Lasso. It will be based on the conditional distribution of the posterior distribution of β given data and hyper-parameters.

Algorithm 2 Ridge

```
  Step 1 : Initial values

  As in the Lasso algorithm
  for $k = 2 : N$ do
  Step 2 : Proposal distributions
  Generate $\lambda_0$ ; from $\mathcal{N}(\lambda(k-1), \sigma_\lambda^2)$
      Generate $\tau_0$ from $IGam(\lambda^2/2)$ ;
      Generate $\beta_0$ from $\mathcal{N}(0, {\tau_0}^2 I_d)$ ;

      where $I_d$ is the identity matrix of dimension $d$.
  end for

  Step 3 : Calculate the Hastings ratio

  As in the Lasso algorithm
```

2.4 SCAD penalty

2.4.1 SCAD with fixed parameter

The Smoothimg Clipped Absolute Deviation (SCAD) penalty is defined by :

$$P_\lambda(0) = 0$$

$$\textit{and}$$

$$P'_\lambda(|\beta_j|) = \lambda I(|\beta_j| \leq \lambda) + \frac{a\lambda - |\beta_j|}{a-1} I(\lambda \leq |\beta_j| \leq \lambda a)$$

where a can be chosen using cross validation or generalize cross validation. Fan and Li (2001) show, through simulation study, that $a = 3.7$ is optimal.

Then the SCAD penalty is given by :

$$P_\lambda(|\beta_j|) = \begin{cases} -\lambda\,|\beta_j| & \text{if } 0 \leq |\beta_j| < \lambda \\ \frac{(a^2-1)\lambda^2 - (|\beta_j| - a\lambda)^2}{2(a-1)} & \text{if } \lambda \leq |\beta_j| \leq a\lambda \\ \frac{1}{2}(a+1)^2\lambda^2 & \text{if } |\beta_j| > a\lambda \end{cases}$$

According to Craven and Wahba (1978), a and λ can be estimated by cross validation criterion. Fan and Li (2001) proposed a prior distribution for the vector of the parameters β (of $\mathbf{R}^d$) conditional to a and λ. They assumed that $\beta \sim \mathcal{N}(0, (a\lambda)^2 I_d)$ where I_d is the identity matrix with dimension d and $\lambda = \sqrt{2log(d)}$ and $a = 3.7$ when $d < 100$.

Algorithm 3 SCAD0

```
Step 1 : Initial values

As in the Lasso algorithm
Step 2 : Proposal distributions
for given values of $\lambda = \sqrt{2log(d)}$ and $a = 3.7$
for $k = 2 : N$ do
    Generate $\beta_0$ from $\mathcal{N}(0, (a\lambda)^2 I_d)$ ;
Step 3 : Calculate the Hastings ratio

As in the Lasso algorithm
end for
```

2.4.2 SCAD with linear approximation

In this section we propose a Bayesian SCAD penalty approach based on the linearized SCAD version (Zou H. and Li R. 2008). Using Taylor expansion of $P_\lambda(|\beta_j|)$ we obtain the relationship in the neighborhood of an initial solution $\beta^{(0)}$ as follows :

$$P_\lambda(|\beta|) \approx P_\lambda(|\beta_j^{(0)}|) + P'_\lambda(|\beta_j^{(0)}|)(|\beta_j| - |\beta_j^{(0)}|),$$

for the given initial value $\beta^{(0)} \in \mathrm{R}^d$.

Then the QR problem with this local linear approximation of the SCAD penalty becomes :

$$\beta_j \min \rho_p(y_i - g_\beta(x_i)) + \sum_{j=1}^{d} P'_\lambda(|\beta_j^{(0)}|)\,|\beta_j|$$

This problem with linearized SCAD penalty is equivalent to the adaptive Lasso penalty (Alhamzawi et al. 2012). The Bayesian formulation is equivalent to Normal priors of the parameters β_j $j = 1, ..., d$ with means zeros and variances $\tau_j^2 = \frac{1}{2P'_\lambda(|\beta_j^{(0)}|)}$ for a given initial solution $\beta^{(0)}$. The Bayesian SCAD with linear approximation can be summarized by the following formulation :

Prior distributions

$$\beta_j \mid a, \lambda \sim \mathcal{N}(0, \frac{1}{2P'_\lambda(\mid \beta_j^{(0)} \mid)}),$$

$$\lambda = \sqrt{2log(d)} \quad and \quad a = 3.7.$$

Posterior distribution

$$\pi(\beta, a, \lambda \mid \mathbf{y}) \propto \prod_{i=1}^{n} f_{ALD}(y_i \mid g_\beta(x), \sigma, p) \prod_{j=1}^{d} \pi(\beta_j \mid a, \lambda)$$

The MCMC algorithm is equivalent to that given for the Lasso. It will be based on the conditional distribution of the posterior distribution of β given data, the initial solution of the vector $\beta^{(0)}$ and the hyper-parameters.

Algorithm 4 SCAD1

```
Step 1 : Initial values

for given initial solution $\beta^{(0)}$,
for $k = 2 : N$ do
Step 2 : Proposal distributions
    Generate $\beta_0$ from $\mathcal{N}(\beta(k-1), {\sigma_\beta}^2 I_d)$ ;

    where $I_d$ is the identity matrix of dimension $d$.
end for

Step 3 : Calculate the Hastings ratio

As in the Lasso algorithm
```

2.4.3 SCAD with quadratic approximation

In this section we present a Bayesian approach based on the local quadratic approximation of the SCAD penalty (Fan J., Li R. 2001) for the quantile regression model. For given initial value $\beta^{(0)} \in \mathrm{R}^d$ the local quadratic approximation is given by :

$$P_\lambda(|\,\beta\,|) \approx \sum_{j=1}^{n} P_\lambda(|\,\beta_j^{(0)}\,|) + \frac{P'_\lambda(|\,\beta_j^{(0)}\,|)}{2\,|\,\beta_j^{(0)}\,|}(\beta_j^2 - \beta_j^{(0)2}),$$

Then the QR problem related to local quadratic approximation of the SCAD penalty becomes :

$$\hat{\beta} = argmin \sum_{j=1}^{n} \rho_p(y_i - g_\beta(x_i)) + \sum_{j=1}^{d} \frac{P'_\lambda(|\,\beta_j^{(0)}\,|)}{2\,|\,\beta_j^{(0)}\,|}\beta_j^2$$

The Bayesian formulation is equivalent to Normal priors of the parameters $\beta_j \;\; j = 1, ..., d$ with means zeros and variances $\tau_j^2 = \frac{|\beta_j^{(0)}|)}{P'_\lambda(|\beta_j^{(0)}|)}$ for a given initial solution $\beta^{(0)}$. The Bayesian SCAD with quadratic approximation can be summarized by the following formulation :

Prior distributions

$$\beta_j \mid a, \lambda \sim \mathcal{N}(0, \frac{|\,\beta_j^{(0)}\,|)}{P'_\lambda(|\,\beta_j^{(0)}\,|)}),$$

$$\lambda = \sqrt{2log(d)} \quad and \quad a = 3.7.$$

Posterior distribution

$$\pi(\beta, a, \lambda \mid \mathbf{y}) \propto \prod_{i=1}^{n} f_{ALD}(y_i \mid g_\beta(x), \sigma, p) \prod_{j=1}^{d} \pi(\beta_j \mid a, \lambda)$$

The MCMC algorithm is equivalent to that given for the Lasso. It will be based on the conditional distribution of the posterior distribution of β given data, the initial solution of the vector $\beta^{(0)}$ and the hyper-parameters. All approaches are implemented in Matlab environment.

Algorithm 5 SCAD2

```
The same as SCAD1 algorithm. The unique modification concerns the variance of the prior
distribution of the parameters $\beta$.
```

3 Simulation studies

In this section, we perform a Monte Carlo simulations to investigate the performance of the parameter estimators of the quantile regression model by the regularisation penalties. All regularized estimation methods are compared to the maximum likelihood estimates in order to assess the estimator's behaviours for small and moderate sample sizes.

The first model (Pratesi, Ranalli and Salvati (2009)) is considered to generate the theoretical underlying relationship between the covariate X and the response variable $Y = g_\beta(X) + \epsilon$ and $\epsilon \sim GEV(\mu = 0, \sigma = 100, \xi = 0.2)$ with :

Cycle : $g_\beta(x) = 10. * sin(6X)$ and $X \sim Unif[0, 1]$

For all the studied penalties, let $\hat{\beta}_L$, $\hat{\beta}_R$, $\hat{\beta}_{S0}$, $\hat{\beta}_{S1}$, $\hat{\beta}_{S2}$ and $\hat{\beta}_M$ denote the parameter estimators corresponding to Lasso, Ridge, SCAD0, SCAD1 and SCAD2, respectively.

For a simulated example we consider a large number of Knots ($m = 20$) to assess the sparsity of the penalties. For a given probability p , the Relative Absolute Bias (RAB) and Relative Mean Square error (RMSE) are computed for performances comparison. We considered samples of size $n = 20 : 20 : 200$ and simulate $R = 1000$ replicates from M.

For given penalty P_λ, the estimate $(\hat{Q}_{P_\lambda})$ is then carried out through the MCMC algorithm. Table 1 summarizes the simulation results for the model M and with $p = 0.9$. Results show that the bias and the mean error decrease with sample size for all proposed approaches.

Results show that the SCAD0 penalty leads to the smaller RAB and RMSE especially for small sample sizes. Note that the SCAD0, as proposed by Fan and Li (2001), assumes a normal distribution as prior for the vector of the parameters β (of $\mathbf{R}^d$ and $d = 24$) with fixed values of $\lambda = \sqrt{2log(d)}$ and $a = 3.7$. The other penalties improve clearly the maximum likelihood (ML) estimation for small sample sizes.

We note also that for the Lasso and Ridge penalties, the RAB remains nearly the same for all n. The RMSE for the penalties SCAD1 et SCAD2 are similar for almost all sample sizes. All penalties perform better than the ML especially for very small sample sizes.

Figure 1 illustrates the quantile curves for the simulated model M and $p = 0.9$. We considered the same conditions with $m = 20$ knots and degree $l = 3$. The conditional curves illustrate clearly the superiority of the SCAD0 penalty to reproduce the same shape as the theoretical curve computed from the simulated model M.

In order to assess the estimator properties of the proposed approaches, with respect to the tail behaviour of the errors' distribution, three other scenarios have been added to the simulation study. The first scenario is generated with quadratic dependence and normally distributed errors. This model represents case with light tail, model $M1$. The two other models have the same dependence structure as $M1$ for the location parameter and the Generalized Extreme Value (GEV) distribution for the errors. Two tail behaviors of the GEV have been considered : (a) moderate tail with the shape parameter $\kappa = 0.1$ (model $M2$) and heavy tail with shape parameter $\kappa = 0.3$ ($M3$). A summary of the models for the tail behaviour study :

- Model M1 : $Y \sim N(a + bX^2, \sigma)$ with $X \sim N(0, 3)$; $a = 10$; $b = 1.2$;$\sigma = 1$.
- Model M2 : $Y \sim GEV(a + bX^2, \sigma, \kappa)$ with $X \sim N(0, 3)$; $a = 10$; $b = 1.2$;$\sigma = 1$ and $\kappa = 0.1$.
- Model M3 : $Y \sim GEV(a + bX^2, \sigma, \kappa)$ with $X \sim N(0, 3)$; $a = 10$; $b = 1.2$;$\sigma = 1$ and $\kappa = 0.3$.

For given penalty, two sample sizes ($n = 30andn = 100$) and two probabilities $p = 0.9$ and $p = 0.99$ the Relative Mean Bias (RMB) and the Relative Mean Error (RME) over all covariate values and based on $N_s = 1000$ simulated datasets.The RMB and RME are computed with respect to the theoretical values Q_{th}.

Table 2 summarizes the simulation results for these three models $M1$, $M2$ and $M3$ for $p = 0.9$ and $p = 0.99$ and two ample sizes $n = 30$ and $n = 100$.

Results show that the bias and the mean error decrease with sample size for all proposed approaches. However, in the case of high probability of non-exceedance, which corresponds to extreme quantiles, the RMB is negative. High values of RMB are obtained especially for heavy tailed distributions.

Lasso and SCAD0 lead to the best performances in terms of RMB and RME especially in the case of extremes ($p = 0.99$) and heavy tailed distribution ($M3, \kappa = 0.3$). The reduction of the bias is very important for these two methods.

Figure 2 illustrates results of this comparison for the three models for both probabilities of non-exceedance ($p = 0.9$ and 0.99). It shows the superiority of the Lasso and SCAD0 penalties especially in the case of model $M3$ and $p = 0.99$.

4 Case study : Annual maximum precipitation

In this section the regularized quantile regression with SCAD0 penalty is considered to estimate the effect of climate index on the variability of extreme precipitation. The main purpose of modeling extreme rainfall is to determine the frequency of certain exceptional values and deduce their return periods.

The concept of return period is very important for risk assessment in civil engineering. This is, on average, the time between two events of the same intensity. For example, to estimate the time separating two events with the same intensity Q_0, we consider the random variable Q with cumulative distribution F_Q then ; the probability to not exceed Q_0 is :

$$p_0 = F_Q(Q_0) = P(Q \leq Q_0)$$

Thus :

$$P(Q > Q_0) = 1 - p_0$$

This is equivalent to :

$$\mathcal{B}_0 = \begin{cases} 0 & \text{si} \quad Q \geq Q_0 \\ t(p-1) & \text{si} \quad sinon \end{cases}$$

where $\mathcal{B}_0$ parameter follows a Bernoulli distribution with parameter $P(Q > Q_0)$.

Let n_0 be a variable that corresponds to the number of time steps (year) separating two events such that "$Q > Q_0$".

N_0 follows a Geometric distribution with parameter $P(Q > Q_0) = 1 - p_0$. Then the expected number of years to observe, for the first time, the event $Q > Q_0$ is :

$$E[N_0] = \frac{1}{P(Q > Q_0)}$$

Thus the return period $T_0 = E[N_0] = \frac{1}{1-p_0}$, i.e. $p_0 = 1 - \frac{1}{T_0}$.

The return period T_0 is the average duration between two successive events $Q > Q_0$. In other words, the probability that the threshold Q_0 is exceeded on average once every T_0 years.

4.1 Precipitation at Charlo (NB)

Located in the northeast of the province of New Brunswick (Canada). The meteorological station Charlo is considered to illustrate the regularized quantile regression approach based on the SCAD0 penalty. This region of Canada is affected by several climatic phenomena. Disturbances in the Arctic north, those in the south-east Atlantic and the West from the great lakes. Such studies

aim to explain the inter-annual variability of extreme precipitation (Thiombiano et al. 2015).
For the Charlo station, when compared to other climate indices, the Pacific North American (PNA) leads to the largest value of the Spearman correlation coefficient. Figure 3 displays the time series of the annual maximum precipitation at Charlo and the Max(PNA) index. For details on this correlation study see Thiombiano et al. (2015).

Regularized quantile regression is performed using the SCAD0 penalty in order to estimate extreme events for fixed probability of non-exceedence. The results are shown through the figures below.

As mentioned in the methodology, the main advantage of the Bayesian framework is to include all uncertainty related to data and the parameters' estimation in the inference process through the posterior distribution. The empirical posterior distributions of the quantiles are deduced from that of the parameters provided by the MCMC algorithms. Figure 4 represents the outputs of the MCMC algorithm, for $N = 30.000$, corresponding to SCAD0 penalty of the shape, the scale and the four first parameters of the B-spline function with degree $l = 3$ and $m = 6$ knots. Thus the dimension of the estimated parameters' space is $d = 10$. All the generated Markov chains converge to their stationary distribution after some iterations (Figure 4). For more complex behaviour convergence assessment test should be considered. For more details on the convergence assessment of the MCMC chains see for example El Adlouni et al. (2006).

Figure 5 presents the the conditional curve of the precipitation event of return period $T_0 = 10$ years as function of the max(PNA) index at the Charlo station. The central curve constitutes the Maximum a Posteriori (MAP) estimates of the quantiles which correspond to the mode of the empirical conditional distributions. The credibility intervals corresponds to the 95% bounds, the 2.5% and 97.5% of the a posterior distribution of the quantile.

Unlike the frequency approach based on asymptotic normality, the Bayesian framework, allows to describe the all distribution of the quantiles. Thus the credibility intervals illustrates the properties of the quantile distribution such as, the skewness and the uncertainty for some particular value of the covariate (the PNA climate index). Indeed, these characteristics may depend on the covariate values and then on the dependence structure. Conditional quantile curve indicates a significative increasing of the precipitation events, with return period $T_0 = 10 years$, for the values of the max(PNA) that exceed $max(PNA) = 2$ (Figure 5).

5 Conclusion

Quantile regression allows estimating conditional distributions of extreme events as function of explanatory variables. The addition of covariates always improves the estimates in terms of bias of extreme events. However, the size of the parameter space increases estimators' variance. It is therefore important to have effective tools for choice of parameters of the most significant decline. Indeed, when the parameter space is higher dimension, the solution of the optimization problem for the parameters estimation, is not unique even in the case of negligible effects of certain components of the model. One solution may assign high values preserving a low bias value. However, the estimators' variance can be very high in the case of over-parameterizations. A solution to this problem is the integration of a penalty in the parameter estimation process by additional constraints to limit the parameter space.

The objective of this work was to compare the effect of the penalty constraints on the quantile estimators. Five penalties were considered in this work : Lasso, Ridge, SCAD0, SCAD1 and SCAD2 and were implemented with Bayesian approach. The introduction of the penalties is equi-

valent to the insertion of an additional term in the objective function to be maximised. In Bayesian framework, this term corresponds to the logarithm of the prior distribution of the parameters. The posterior distributions are then deduced for all the penalties with their hierarchical Bayesian models. MCMC Gibbs algorithms are developed to simulate the empirical posterior distributions and deduce the conditional predictive distribution of the quantiles. A simulated study shows that the SCAD0 penalty leads to very small bias and standard error even for small sample sizes. The Quantile Regression model with SCAD0 penalty has been considered for a case study to estimate the quantile of annual maximum precipitation at Charlo station in New-Brunswick, Canada. Results illustrates the use of the regularized quantile regression to estimate the conditional distribution of the extreme event corresponding to a given return period. Results are presented for a return period $T = 10$ years, i.e. for probability of non-exceedance $p = 0.9$. One of the advantages of the Bayesian framework is the possibility to estimate all of the quantile distribution of the quantile conditional to the covariates and then to estimate the uncertainty related to data and the parameters' estimation in the inference process. The empirical posterior distributions of the quantiles are deduced from that of the parameters provided by the MCMC algorithms and allow to estimate the credibility intervals.
Simulation study show that Lasso and SCAD0 penalties lead to the best performances in terms of RMB and RME especially in the case of extremes ($p = 0.99$) and heavy tailed distribution (GEV distribution with $\kappa = 0.3$). The reduction of the bias is very important for these two methods.

Note that in the Quantile Regression model, the inference is done for only one given probability. When, more than one quantile should be estimated, the inference is conducted separately. Thus the order condition may be violated. This situation, known as crossing problem, could be solved by simultaneous estimation of the conditional curve. An ongoing study is devoted to this problem for regularly varying tail behaviour.

Acknowledgements

Authors would like to thank IDRC (project FACE) and NSERC for their financial support. We would like to think Dr. Rahim Alhamzawi for the comments on an earlier version.

Références

[1] Alhamzawi R., Yu K., and Dries F. Benoit (2012). Bayesian adaptive Lasso quantile regression. Statistical Modelling, 12(3), 279-297.

[2] Alhamzawi R. (2015). Model selection in quantile regression models. Journal of Applied Statistics, 42(2), 445-458.

[3] Bondell HD and Reich B.J. (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. Biometrics, 64 :115123.

[4] Cade, B.S. and Noon B.R. (2003). A gentle introduction to quantile regression for ecologists. Front Ecol Environ ; 1(8), 412420.

[5] El Adlouni S., Favre A.C., and Bobée, B. (2006). Comparison of methodologies to assess the convergence of Markov Chain Monte Carlo methods. Computational Statistics and Data Analysis 50(10), 2685-2701.

[6] Engle R. and Manganelli S. (2004). Caviar : Conditional autoregressive value at risk by regression quantiles. Journal of Business and Economic Statistics 22, 367-381.

[7] Fahrmeir L., Kneib T. and Konrath S. (2010) Bayesian regularization in structured additive regression : A unifying perspective on shrinkage, smoothing and predictor selection. Statistics and Computing, 20, 203219.

[8] Fan J. and Li R. (2001). Variable Selection via Non concave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association, Vol. 96, No. 456, Theory and Methods.

[9] Friendman J., Hastie T. and Tibshirani R. (2000). Additive logistic regression : A statistical view of boosting. The annals of Statistics, 28(2), 337-407. With discussion.

[10] Griffin, J.E. and Brown, P.J. (2006). Alternative prior distributions for variable selection with very many more variables than observations. Technical report. University of Warwick, Coventry, UK.

[11] Hanwen H. and Chen Z., (2015). Bayesian composite quantile regression. Journal of Statistical Computation and Simulation Vol 85, Issue 18.

[12] Huang Y. (2015). Quantile regression-based Bayesian semi-parametric mixed-effects models for longitudinal data with non-normal, missing and mismeasured covariate. Journal of Statistical Computation and Simulation, http ://dx.doi.org/10.1080/00949655.2015.1057732.

[13] Kneib T., Hothorn T. and Tutz G.(2009). Variable Selection and Model Choice in Geoadditive Regression Models. In : Biometrics 65.2, 626634. DOI : 10.1111/j.1541-0420.2008.01112.x.

[14] Koenker R., Bassett G. (1978). Regression Quantiles. Journal of the Econometric Society, 33-50.

[15] Koenker R. and Hallock K.F. (2001) Quantile Regression. Journal of Economic Perspectives, 15, 143-156.

[16] Pratesi M., G. Ranalli and Salvati N., (2009). Nonparametric M-quantile regression using penalized splines, Journal of Nonparametric Statistics, 21 :3, 287-304, DOI : 10.1080/10485250802638290.

[17] Park T., Casella G. (2008). The Bayesian Lasso. Journal of the American Statistical Association June 2008, 103, No. 482, Theory and Methods.

[18] Thiombiano A. N., St-Hilaire A., El Adlouni S., Ouarda T. B.M.J and El-Jabi N. (2015). Nonstationary frequency analysis of extreme precipitation intensity in southeastern Canada using a peaks-over-threshold approach. Submitted.

[19] Tibshirani R. (1996). Regression Shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58, 267-288.

[20] Yu K. and Moyeed R.A. (2001), Bayesian quantile regression. Statistics & Probability Letters 54, 437447.

[21] Zou H. (2006). The adaptive Lasso and its oracle properties. Journal of the American Statistical Association, 101, 1418-1429.

[22] Zou H. and Li R.(2008). One-step sparse estimate in non concave penalized likelihood models. The Annals of Statistics 36(4), 1509-1533

n	Lasso	Ridge	SCAD0	SCAD1	SCAD2	ML
20	36 (37)	45 (50)	22 (32)	31 (46)	33 (53)	74 (93)
40	32 (29)	38 (32)	12 (21)	25 (28)	27 (31)	54 (72)
60	28 (26)	29 (29)	8 (18)	22 (25)	25 (23)	33 (43)
80	26 (19)	24 (27)	3 (10)	19 (18)	24 (15)	25 (31)
100	19 (18)	20 (26)	1 (7)	18 (16)	19 (9)	22 (22)
150	15 (15)	19 (19)	-2 (5)	12 (15)	17 (9)	21 (22)
200	14 (13)	19 (16)	-1 (4)	9 (15)	16 (9)	18 (22)

Table 1 – RAB (RMSE) of the quantile estimators with the studied penalties.

	$M1$ (p = 90%)		$M1$ (p = 99%)	
Approach	n = 30	n = 100	n = 30	n = 100
ML	19 (26)	9 (22)	-14 (20)	2 (13)
Lasso	5 (10)	3 (6)	-13 (18)	-3 (10)
Ridge	19 (26)	9 (22)	-14 (19)	2 (13)
SCAD0	9 (26)	5 (16)	-7 (13)	4 (8)
SCAD1	16 (31)	11 (23)	-10 (18)	8 (18)
SCAD2	20 (29)	11 (25)	-12 (19)	5 (16)
	$M2$ (p = 90%)		$M2$ (p = 99%)	
Approach	n = 30	n = 100	n = 30	n = 100
ML	9 (28)	3 (18)	-17 (26)	-16 (32)
Lasso	8 (25)	4 (13)	-5 (12)	-7 (10)
Ridge	9 (29)	3 (18)	-17 (27)	-16 (32)
SCAD0	5 (27)	3 (18)	-8 (15)	-3 (13)
SCAD1	10 (31)	4 (21)	-13 (24)	-11 (28)
SCAD2	9 (30)	2 (22)	-15 (24)	-10 (37)
	$M3$ (p = 90%)		$M3$ (p = 99%)	
Approach	n = 30	n = 100	n = 30	n = 100
ML	31 (48)	10 (22)	-39 (45)	-28 (34)
Lasso	17 (23)	12 (10)	-18 (30)	-7 (21)
Ridge	23 (39)	10 (22)	-39 (40)	-27 (34)
SCAD0	21 (28)	9 (12)	-12 (20)	-9 (15)
SCAD1	32 (50)	11 (16)	-40 (31)	-25 (23)
SCAD2	32 (49)	10 (15)	-36 (33)	-27 (25)

Table 2 – RAB (RMSE) of the quantile estimators for tail behaviour effects.

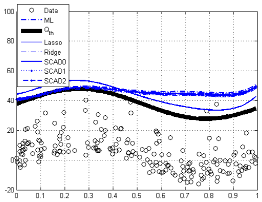

Figure 1 – Illustration of the quantile estimates for the ML, Lasso, Ridge, SCAD0, SCAD1 and SCAD2 penalties for the model M with probability $p = 0.9$

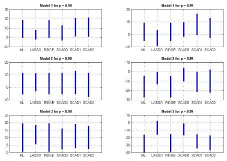

FIGURE 2 – RMB and RME for the estimated quantiles corresponding to models $M1$, $M2$ and $M3$.

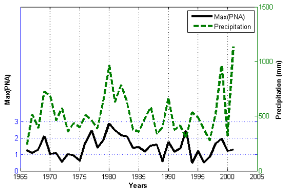

FIGURE 3 – Annual maximum precipitations at Charlo-NB and Max(PNA) time series.

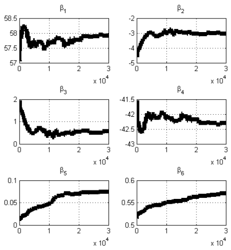

FIGURE 4 – MCMC output of the algorithm for estimating model parameters of the QR model with SCAD0 penalty : Annual maximum precipitations at Charlo-NB.

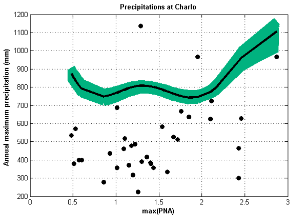

FIGURE 5 – Conditional curve of the precipitation event of return period $T_0 = 10 years$ as function of the max(PNA) index (Charlo-NB).