Fast calibrated additive quantile regression

Matteo Fasiolo[1,†], Yannig Goude[2], Raphael Nedellec[2], and Simon N. Wood[1]

[1]School of Mathematics, University of Bristol, United Kingdom.
[2]EDF R&D, Clamart, France.
[†]Correspondence: matteo.fasiolo@bristol.ac.uk

June 22, 2016

Abstract

Quantile regression represents a flexible approach for modelling the impact of several covariates on the conditional distribution of the dependent variable, which does not require making any parametric assumption on is conditional density. However, fitting quantile regression models using the traditional pinball loss is computationally expensive, due to the non-differentiability of this function. In this work we reduce the computational burden by approximating the pinball loss with a differentiable function. This allows us to exploit the computationally efficient fitting methods described by Wood et al. (2016). Beside this, we show how the smoothing parameters can be selected in a robust fashion, and how reliable uncertainty estimated can be obtained, even for extreme quantiles. We demonstrate our approach in the context of probabilistic electricity load forecasting.

Keywords: Quantile Regression; Generalized Additive Models; Smoothing; Penalized regression splines; Calibrated Bayes; Gibbs Posterior.

[1]The methods described here are implemented by the *qgam* R package, which can be found at https://github.com/mfasiolo/qgam.

Contents

1 Introduction

In this work we develop methodology for fitting Bayesian quantile regression models, based on penalized splines. In particular, we exploit the general methods described by Wood et al. (2016), which allow us to fit quantile functions that are additive smooths of several covariates: this framework enables us to estimate the smoothing parameters of each smooth effect, using stable and efficient Marginal Likelihood methods. This is an advance relative to existing methods, because stable and computationally efficient methods implementing non-parametric additive quantile regression are otherwise lacking. For instance, the *quantreg* R package, which is based on the methods of Koenker (2013), only permits additive models that have at most two smooth terms, and it requires users to select the smoothing parameters manually. On the other hand, the gradient boosting quantile regression method implemented by the *mboost* R package (Hothorn et al., 2010) does not limit the number of smooth terms, but it requires users to manually choose the degrees of freedom used by each base model. In addition, *mboost* offers smooth effects that are at most bivariate and uses bootstrapping to estimate parameter uncertainty, while the approach proposed here allows for smooths of any dimension and quantifies uncertainty using analytic expressions.

In order to exploit the tools described by Wood et al. (2016), and implemented within the *mgcv* R package, we have to overcome several difficulties. Firstly, an appropriate probabilistic model for the observation process, $p(y|x)$, needs to be specified. In the Bayesian quantile regression literature (e.g. Yu and Moyeed, 2001), the Asymmetric Laplace (AL) density is often used for this purpose, but this density has the drawback of not being differentiable at its mode and of having no curvature on the log-scale. In related contexts, Yue and Rue (2011) and Oh et al. (2012) address this issue by proposing two different smooth approximations to the AL density. Here we propose to embed the AL density in a new family which, while being differentiable everywhere, can be used to approximate the AL density with arbitrary precision. The advantage of this approach is that it results in a normalized approximation to the AL density, which is critical for the purpose of model selection. Similarly to Oh et al. (2012), we provide results concerning the impact of the approximation on the estimated quantiles. Further, we exploit them to develop a scale invariant parametrization of the coefficients that controls the trade-off between accuracy and computational stability.

A second, unrelated, issue that needs to be addressed is the fundamental misspecification of the AL density. Indeed, this density rarely provides a plausible model for $p(y|x)$, but it is considered only because the resulting negative log-likelihood is proportional to the pinball loss (Yu and Moyeed, 2001). The presence of model misspecification can cause posterior credible intervals for model parameters to not achieve the nominal frequentist coverage, even asymptotically (Kleijn et al., 2012). Waldmann et al. (2013) point out that the misspecification is more severe for extreme quantiles, so that achieving adequate coverage is difficult in the tails of the response distribution.

To address these issues, we consider how Bayesian quantile regression fits into the general belief-updating framework of Bissiri et al. (2013). Looking at Bayesian quantile regression through this interpretation sheds light on why using a mis-specified density is appropriate. In addition, this general framework makes it clear that the AL scale parameter, σ, controls the complexity of the fit. Based on this interpretation, we propose a calibration-based approach for selecting σ, related to that of Syring and Martin (2015). This method allows us to avoid overly wiggly fits and to obtain appropriate credible interval coverage, even for extreme quantiles, while increasing the computation cost only mildly.

The rest of the paper is structured as follows. In Section 2 we briefly review additive quantile regression and how it can be set in a Bayesian framework using the AL density. Section 3 develops the proposed approximation to this density. In Section 4 we explain how additive quantile regression models can be fitted, using the methods proposed by Wood et al. (2016). In Section 5 we briefly describe the belief-updating framework of Bissiri et al. (2013), and give details of the calibration approach for selecting σ, which we test on simulated data. In Section 6 we demonstrate the performance of the proposed approach in the context of probabilistic electricity load forecasting.

2 Background

2.1 Quantile regression basics

Quantile regression aims at modelling the τ-th quantile (where $\tau \in (0,1)$) of a response, y, conditionally on a d-dimensional vector of covariates, $\mathbf{x}$. More precisely, is $F(y|\boldsymbol{x})$ is the conditional c.d.f. of y, then the τ-th conditional quantile is

$$F^{-1}(\tau|\boldsymbol{x}) = \inf\{y : F(y|\boldsymbol{x}) \geq \tau\}.$$

The τ-th conditional quantile can also be defined as the minimizer of the expected loss

$$L(\mu|\boldsymbol{x}) = E\big\{\ \rho_\tau(y-\mu)|\boldsymbol{x}\big\} = \int\ \rho_\tau(y-\mu)dF(y|\boldsymbol{x}), \tag{1}$$

w.r.t. $\mu = \mu(\mathbf{x})$, where

$$\rho_\tau(z) = (\tau - 1)z\mathbb{1}(z < 0) + \tau z\mathbb{1}(z \geq 0), \tag{2}$$

is the so-called pinball loss. Hence, given a sample of size n, one approximates $dF(y)$ with its empirical version, $dF_n(y)$, which leads to the quantile estimator

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau\{y_i - \mu(\mathbf{X}_{\mathbf{i},:})\},$$

where $\mathbf{X}$ is an $n \times d$ matrix of covariates, while $\mathbf{X}_{\mathbf{i},:}$ indicates its i-th row.

In this work we assume that $\mu(\mathbf{x})$ has an additive structure, such as

$$\mu(\mathbf{x}) = \sum_{j=1}^{d} f_j(x_j) + f_{12}(x_1, x_2) + f_{235}(x_2, x_3, x_5) + \cdots,$$

where the marginal, $f_j(x_j)$, and the joint, $f_{ij}(x_i, x_j)$, smooth functions are defined in terms of spline bases. For instance

$$f_j(\mathbf{x}) = \sum_{i=1}^{k} \beta_{ji} b_{ji}(x_j)$$

where $\boldsymbol{\beta}_{j,:}$ is a vector of unknown coefficients and $\boldsymbol{b}_{j,:}(x_j)$ is a vector of basis functions. Analogous expressions can be used to define the joint smooths. See Wood (2006) for an introduction to additive models.

2.2 Bayesian quantile regression

In order to set quantile regression in a Bayesian framework, it is necessary to specify an appropriate probabilistic model for the observation process. Yu and Moyeed (2001) assume that the observations follow an Asymmetric Laplace (AL) distribution, which has density

$$p_{AL}(y|\mu, \tau, \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{ -\rho_\tau\Big(\frac{y-\mu}{\sigma}\Big)\right\}, \tag{3}$$

where μ and σ are, respectively, a location and a scale parameter. It is simple to show that maximizing (3) wrt μ is equivalent to minimizing the $\rho_\tau[(y-\mu)/\sigma]$ wrt the same parameter. Hence, in a penalized spline regression context, we could (ideally) obtain a MAP estimate of μ by minimizing

$$-\sum_{i=1}^{n} \log p_{AL}\{y_i|\mu(\mathbf{X}_{i,:}), \tau, \sigma\} + \sum_{j=1}^{m} \gamma_j \boldsymbol{\beta}^T \mathbf{S}^j \boldsymbol{\beta},$$

where $\mathbf{S}^j$ are semi-definite matrices, used to penalize the wiggleness of the fit $\mu(\mathbf{x})$, while $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_m\}$ is a vector of positive smoothing parameters.

The problem with this approach is that p_{AL} is not differentiable at its mode, while $\log p_{AL}$ is piecewise linear. This means that the methods described in Wood et al. (2016) cannot be used to estimate $\mu(\boldsymbol{x})$, because they require the observation density to be differentiable and to have non-zero curvature. In Section 3 we address these issues by proposing a smooth extension of the AL density.

3 Extending the asymmetric Laplace density

We consider the family of densities with exponential tails described by Jones (2008)

$$p_G(y) = K_G^{-1}(\alpha, \beta) \exp\big\{\alpha y - (\alpha + \beta) G^{[2]}(y)\big\},$$

where $\alpha, \beta > 0$, $K_G(\alpha, \beta)$ is a normalizing constant,

$$G^{[2]}(y) = \int_{-\infty}^{y} \int_{-\infty}^{t} g(z) dz dt = \int_{-\infty}^{y} G(t) dt,$$

and $g(z)$ and $G(z)$ are respectively the p.d.f and c.d.f. of a (fictitious) random variable z.

Importantly, this family nests the AL distribution, which is recovered by choosing $g(z)$ to be the Dirac delta and by imposing $\alpha = 1 - \tau$, $\beta = \tau$, with $0 < \tau < 1$. Adding a location and a scale parameter is trivial. Obviously, we do not aim at recovering p_{AL} (which is non-differentiable) exactly, but we propose to approximate the Dirac delta using a smoother p.d.f.. We achieve this by choosing

$$G(y) = \frac{e^{\frac{z}{\lambda}}}{1 + e^{\frac{z}{\lambda}}},$$

which is the c.d.f. of a logistic random variable with location $\mu = 0$ and scale λ. Notice that, as $\lambda \to 0$, we have that $G(z) \to I(z > 0)$ which is the c.d.f. corresponding to a Dirac delta density. With this choice we have

$$G^{[2]}(y) = \int_{-\infty}^{y} \frac{e^{\frac{z}{\lambda}}}{1 + e^{\frac{z}{\lambda}}} dz = \lambda \log\big(1 + e^{\frac{y}{\lambda}}\big),$$

which leads to

$$p_L(y) = K_L^{-1}(\tau, \lambda) e^{(1-\tau)y} (1 + e^{\frac{y}{\lambda}})^{-\lambda}. \tag{4}$$

The normalizing constant is

$$K_L(\tau, \lambda) = \lambda \mathrm{Beta}\big[\lambda(1-\tau), \lambda\tau\big],$$

where $\mathrm{Beta}(\cdot, \cdot)$ is the beta function. The location-scale extension of (4) is simply

$$\tilde{p}_L(y|\mu, \tau, \sigma, \lambda) = \frac{1}{\sigma} p_L\Big(\frac{y-\mu}{\sigma}\Big) = \frac{e^{(1-\tau)\frac{y-\mu}{\sigma}} (1 + e^{\frac{y-\mu}{\lambda\sigma}})^{-\lambda}}{\lambda\sigma \mathrm{Beta}\big[\lambda(1-\tau), \lambda\tau\big]}, \tag{5}$$

We refer to (5) as the Extended Log-F (ELF) density, because imposing $\lambda = 1$ leads to the log-F density described by Jones (2008). Appendix A contains additional details regarding the new density. Most of these are necessary to fit semi-parametric additive models, using the methods described in Section 4.

Adopting the ELF density for quantile regression leads to a quantile estimator that asymptotically minimizes

$$\tilde{L}(\mu|\mathbf{x}) = E\big[-\log \tilde{p}_L(y - \mu)|\mathbf{x}\big], \tag{6}$$

The following theorem quantifies how minimizing (6), rather than (1), affects the resulting quantile estimates.

Theorem 3.1. *Let μ_0 be the minimizer of (1) and let μ^* be the minimizer of (6), for fixed τ. Indicate with $f(y|\mathbf{x})$ and $F(y|\mathbf{x})$ the conditional p.d.f. and c.d.f. of y. Then*

$$|F(\mu^*|\mathbf{x}) - F(\mu_0|\mathbf{x})| \leq 2 \log(2) \lambda\sigma \sup_y f(y|\mathbf{x}).$$

Proof. See Appendix C. □

The practical appeal of the above bound is that it can be used to select λ. In fact, assume that σ is fixed and that the user has defined a maximal approximation error $\epsilon \in (0, 1)$ for $|F(\mu^*|\mathbf{x}) - F(\mu_0|\mathbf{x})|$. Then, it is sufficient to approximate $\sup_y f(y|\mathbf{x})$ to determine λ. For instance, if a Gaussian model $y \sim \mathrm{N}\{\alpha(\mathbf{x}), \kappa^2\}$ is used, then $\sup_y f(y|\mathbf{x}) \approx 1/\sqrt{2\pi\kappa^2}$, which leads to

$$\lambda^* = \epsilon \frac{\sqrt{2\pi\kappa^2}}{2\log(2)\sigma}. \tag{7}$$

In an heteroscedastic setting, it might be desirable to let κ^2, and hence λ^*, depend on $\mathbf{x}$. This can achieved using by an additive model for both α and κ^2, as described in Wood et al. (2016).

4 Extended GAM fitting via PIRLS and LAML

Having defined a smooth extension of the AL density, we briefly describe how to perform semi-parametric additive quantile model fitting, using the extended GAM framework detailed in Wood et al. (2016). This class of models requires the log-likelihood to have the form

$$ll(\boldsymbol{y}) = \sum_{i=1}^{n} \log p(y_i|\mu_i, \boldsymbol{\theta}, \phi),$$

where $\boldsymbol{\theta}$ is a vector of model parameters, ϕ is a scale parameter and μ_i is, in many cases, $E(y_i)$. In our context, $\boldsymbol{\theta} = \{\sigma, \boldsymbol{\gamma}\}$, $\phi = 1$, while $\mu_i = \mathbf{X}_{i,:}\boldsymbol{\beta}$ is the quantile location, depending on regression coefficients $\boldsymbol{\beta}$. Also, here the likelihood is based on the ELF density, $\tilde{p}_L$, hence it implicitly depends on parameters τ and λ, which are considered fixed here.

Assuming for a moment that $\boldsymbol{\theta}$ and the smoothing parameters $\boldsymbol{\gamma}$ are fixed, the regression coefficients $\boldsymbol{\beta}$ of an extended GAM model are estimated by minimizing the following criterion

$$D(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} Dev_i(\boldsymbol{\beta}, \boldsymbol{\theta}) + \sum_{j=1}^{m} \gamma_j \boldsymbol{\beta}^T \mathbf{S}^j \boldsymbol{\beta}, \tag{8}$$

where $Dev_i(\boldsymbol{\beta}, \boldsymbol{\theta})$ is the deviance corresponding to y_i. $D(\boldsymbol{\beta}, \boldsymbol{\theta})$ can be minimized efficiently and stably using a Penalized Iteratively Re-weighted Least Square (PIRLS) algorithm. This consists of iteratively minimizing

$$\sum_{i=1}^{n} w_i (z_i - \mathbf{X}_{i,:}\boldsymbol{\beta})^2 + \sum_{j=1}^{m} \gamma_j \boldsymbol{\beta}^T \mathbf{S}^j \boldsymbol{\beta}, \tag{9}$$

where

$$z_i = \eta_i - \frac{1}{2w_i}\frac{\partial D_i}{\partial \eta_i}, \quad w_i = \frac{1}{2}\frac{d^2 D}{d\eta_i^2},$$

and, in our context, $\eta_i = \mathbf{X}_{i,:}\boldsymbol{\beta}$.

Ideally, the parameter vector $\boldsymbol{\theta}$ could be estimated by maximizing the marginal likelihoood

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\beta}|\boldsymbol{\theta})\, d\boldsymbol{\beta},$$

where $p(\boldsymbol{\beta}|\boldsymbol{\theta})$ is an improper multivariate normal distibution $\mathrm{N}(\mathbf{0}, \mathbf{S}^{\boldsymbol{\gamma}-})$, with $\mathbf{S}^{\boldsymbol{\gamma}-}$ being the Moore-Penrose pseudo-inverse of $\mathbf{S}^{\boldsymbol{\gamma}} = \sum_{j=1}^{m} \gamma_j \mathbf{S}^j$. However, given that the above

integral is generally intractable, Wood et al. (2016) maximize the Laplace Approximate Marginal Likelihood (LAML) instead

$$L(\boldsymbol{\theta}) = \frac{1}{2}D(\hat{\boldsymbol{\beta}}) - \tilde{ll}(\boldsymbol{\theta}) + \frac{1}{2}\Big[\log|\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \mathbf{S}^\gamma| - \log|\mathbf{S}^\gamma|_+\Big] + \frac{M_p}{2}\log(2\pi\phi), \tag{10}$$

where $\tilde{ll}(\boldsymbol{\theta})$ is the saturated log-likelihood, $\mathbf{W}$ is a diagonal matrix such that $\mathbf{W}_{ii} = w_i$, M_p is the dimension of the null space of $\mathbf{S}^\gamma$ and $|\mathbf{S}^\gamma|_+$ is the product of its non-zero eigenvalues.

Wood et al. (2016) maximize $L(\boldsymbol{\theta})$ using an outer Newton algoritihm, which requires mixed derivatives of the deviance up to order four in $\boldsymbol{\mu}$ and up to order two in $\boldsymbol{\theta}$. Obviously, these are model dependent, hence we provide these quantities for $\tilde{p}_L$ in Appendix A. In addition, given that the w_is can be very close to zero when fitting quantile regression models based on $\tilde{p}_L$, obtaining reliable and stable estimates required modifying the PIRLS iteration and the computation of LAML and its derivatives. This more stable implementation is described in Appendix B.

The procedure just described provides computationally stable and cheap estimates of $\boldsymbol{\theta} = \{\sigma, \boldsymbol{\gamma}\}$, as long as parameter λ in the log-F density is large enough to introduce sufficient curvature. However, we demonstrate empirically in Section 5.3, this approach often leads to unreliable uncertainty estimates and to wiggly fits, especially for extreme quantiles. Both issues originate from the fact that the AL density is used as an expedient to obtaining quantile estimates, but in most cases it is not an appropriate model for the observation density, $p(y|\mathbf{x})$. This misspecification often causes the resulting uncertainty estimates, such as posterior credible intervals $\boldsymbol{\beta}$, to be inadequate (Sriram, 2015).

Given that it is difficult to argue that the AL density provides a satisfactory description of the conditional distribution of y, it also hard to come up with an interpretation for scale parameter σ, at least from a probabilistic modelling perspective. But this parameter is quite important in determining the smoothness of the fit, as we clarify in Section 5.1. There, we describe a general framework through which it is possible to put the use of the AL density into context, and to provide a clear interpretation for the role of σ.

5 Dealing with model misspecification

5.1 The general belief-updating framework

Bissiri et al. (2013) propose a general framework for updating belief distributions. In particular, they show how a prior beliefs distribution can be updated to produce a posterior while using a loss function, rather than full likelihood function, to connect model parameters to the data. In particular, assume that we are interested in finding the vector of model parameters $\boldsymbol{\theta}$ minimizing

$$E\{L(\boldsymbol{\theta})\} = \int l(y, \boldsymbol{\theta}) f(y) dy, \tag{11}$$

where $l(\cdot, \cdot)$ is a general loss function and $f(y)$ is the p.d.f. of y. Suppose that we have a prior believe about $\boldsymbol{\theta}$, which is quantified by the prior density $p(\boldsymbol{\theta})$. Then Bissiri et al. (2013) argue that, given some data y, a coherent approach to updating $p(\boldsymbol{\theta})$ is represented by the posterior

$$p(\boldsymbol{\theta}|y) = \frac{e^{-l(y,\boldsymbol{\theta})}p(\boldsymbol{\theta})}{\int e^{-l(y,\boldsymbol{\theta})}p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}.$$

When multiple samples $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ are available this becomes

$$p(\boldsymbol{\theta}|y) = \frac{e^{-\sum_{i=1}^n l(y_i,\boldsymbol{\theta})}p(\boldsymbol{\theta})}{\int e^{-\sum_{i=1}^n l(y_i,\boldsymbol{\theta})}p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}, \tag{12}$$

where $-\sum_{i=1}^{n} l(y_i, \boldsymbol{\theta})$ is an estimate of (11). Following Syring and Martin (2015), we refer to (12) as the "Gibbs posterior".

Quantile regression, based on the AL density, fits squarely into this framework. In fact, the appropriate loss is the pinball loss, which is equivalent to the negative AL log-density, if σ is fixed to 1. In addition, Gibbs posteriors often include a learning rate or scaling factor $w > 0$, which determines the relative weight of the loss and of the prior. One way of creating a scaled Gibbs posterior is

$$p(\boldsymbol{\theta}|y) \propto we^{-w\sum_{i=1}^{n} l(y_i,\boldsymbol{\theta})}p(\boldsymbol{\theta}). \tag{13}$$

In our context $w = 1/\sigma$ and the pinball loss is approximated using the ELF density. Bissiri et al. (2013) discuss several approaches for selecting w, but in Section 5.2 we describe a calibration procedure which is able to provide adequate uncertainty and smoothing parameter estimates.

5.2 Selecting the learning rate

Syring and Martin (2015) propose selecting the learning rate w so that the resulting Gibbs posterior is well calibrated. In particular, let $C_\alpha(w, \boldsymbol{y})$ be the $100(1-\alpha)\%$ credible interval for $\boldsymbol{\theta}$, at level $\alpha \in (0, 1)$. Then, they propose to select w so that

$$\mathbb{P}\{\boldsymbol{\theta}^0 \in C_\alpha(w, \boldsymbol{y})\} \approx 1 - \alpha,$$

where $\mathbb{P}$ is the objective probability measure, based on the data-generating process, and $\boldsymbol{\theta}^0$ is the true parameter. We propose a similar approach but, rather than calibrating at a single level α, we consider all levels jointly. In particular, under a Gaussian approximation to the posterior of $\boldsymbol{\beta}$, we have that

$$\boldsymbol{\mu} \sim \mathrm{N}\{\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}_{\boldsymbol{\mu}}\},$$

where $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}}_{\boldsymbol{\mu}} = \mathbf{X}\hat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{X}^T$, with

$$\hat{\mathbf{V}}_{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{W}\mathbf{X}^T + \mathbf{S}^\gamma)^{-1}.$$

Hence, if $\boldsymbol{\mu}^0$ is the true quantile vector, the random variables $z_i = (\mu_i^0 - \hat{\mu}_i)(\hat{\mathbf{V}}_{\boldsymbol{\mu}})_{ii}^{-\frac{1}{2}}$, for $i = 1, \ldots, n$ should approximately follow a standard normal distribution, under a well calibrated posterior. Ideally, we could calibrate the posterior to the data-generating process by minimizing a criterion such as

$$A(\sigma) = n\int \{F(z) - \Phi(z)\}^2 v(z)\, d\Phi(z),$$

w.r.t. σ. Here $F(z)$ and $\Phi(z)$ are, respectively, the objective c.d.f. of z and a standard normal c.d.f., with the former being implicitly dependent on σ, while $v(z) > 0$ is a weighting function. We choose $v(z) = [\Phi(z)\{1 - \Phi(z)\}]^{-1}$, which results in $A(\sigma)$ being the distance used in the Anderson-Darling test for normality (Anderson and Darling, 1954). Notice that, if calibration at a single level α is the only object of interest, this can still be achieved by choosing $v(z) = \delta\{\Phi^{-1}(\alpha)\}$, where $\delta(\cdot)$ is the Dirac delta function.

Unfortunately $F(z)$ is unknown, but it can be substituted with its empirical version $F_n(z)$. Simulations from $F_n(z)$ can be obtained by boostrapping (that is, sampling with replacement) the observed data $\boldsymbol{y}$, fitting the model and then calculating the standardized deviations from the true quantile as above. Obviously $\boldsymbol{\mu}^0$ is unknown as well, but it can be substituted with $\hat{\boldsymbol{\mu}}^0$, which is the maximizer of the Gibbs posterior based on all the observed data, not on a bootstrap dataset.

Algorithm 1 Estimating $A(\sigma)$ for fixed σ

Let $\mathbf{X}^1, \ldots, \mathbf{X}^k$ and $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^k$ be k bootstrap datasets, obtained by resampling the n original observations. We assume that τ is fixed and that λ is a deterministic function of σ based, for instance, on (7). Then $A(\sigma)$ can be estimated as follows:

1: using the full dataset estimate the smoothing parameters, $\boldsymbol{\gamma}$, by maximizing LAML (10). Given $\hat{\boldsymbol{\gamma}}$, estimate $\hat{\boldsymbol{\beta}}$ by PIRLS (9), and obtain the reference conditional quantile estimate $\hat{\boldsymbol{\mu}}^0 = \mathbf{X}\hat{\boldsymbol{\beta}}$.

2: For $i = 1, \ldots, k$

1. Given $\hat{\boldsymbol{\gamma}}$, estimate $\boldsymbol{\beta}$ by PIRLS maximization of the penalized likelihood based on the bootstrapped dataset $\boldsymbol{y}^i$. Indicate the resulting estimate with $\hat{\boldsymbol{\beta}}^i$.

2. Predict the conditional quantile vector $\hat{\boldsymbol{\mu}}^i = \mathbf{X}\hat{\boldsymbol{\beta}}^i$, where $\mathbf{X}$ is the full (not boostrapped) design matrix.

3. Calculate the standardized deviations of the bootstrapped fit from the full data fit

$$z_{(i-1)n+j} = (\hat{\mu}_j^0 - \hat{\mu}_j^i)\{(\hat{\mathbf{V}}_{\boldsymbol{\mu}}^i)_{jj}\}^{-\frac{1}{2}}, \quad \text{for } j = 1, \ldots, n,$$

where $\hat{\mathbf{V}}_{\hat{\boldsymbol{\mu}}}^i$ is the posterior covariance matrix based on the i-th boostrapped sample. Notice that $\boldsymbol{z}$ is an nk-dimensional vector.

3: Calculate the Anderson-Darling statistic

$$A(\sigma)^2 = -nk - \sum_{l=1}^{nk} \frac{2l-1}{nk}[\log \Phi(z_l) + \log\{1 - \Phi(z_{nk+1-l})\}].$$

τ	0.01	0.05	0.5	0.95	0.99
CAL	28.7(0.7)	125.1(1.1)	670.4(3)	242.2(2.8)	77.4(3)
LAML	34(2.1)	127(1.6)	670.8(3.2)	256.6(3.5)	125.1(11)
BOOST	31.8(1.5)	126.7(1.8)	670.9(2.9)	243.2(3.4)	77.9(4)

Table 1: Mean(standard deviation) of the pinball loss for each quantile and method.

Given k bootstrap samples, Algorithm 1 details the steps needed to estimate $A(\sigma)$ for fixed σ. An important feature of Algorithm 1 is that we estimate the smoothing parameters only once, using the full dataset. This is critical, because re-estimating them using each boostrap sample would be very computationally expensive. Using Algorithm 1, minimizing $A(\sigma)$ w.r.t σ is quite simple. In fact, if the boostrapped samples are simulated only once, the objective is a deterministic function of σ, and it can be minimized using standard root-finding algorithms, such as bisection. In our experience, the objective is generally smooth and it has a unique minimum. Indeed, decreasing σ leads to more wiggly fits and hence to over-dispersion of z, relative to a standard normal. Increasing σ has the opposite effect.

5.3 An additive example

Before illustrating the proposed approach on a real example, we test it on simulated data. In particular, we consider the following additive model

$$y_i = x_i + x_i^2 - z_i + 2\sin(z_i) + 0.1v_i^3 + 3\cos(v_i) + e_i, \tag{14}$$

where $e_i \sim \text{Gamma}(3,1)$, $x_i \sim \text{Unif}(-4,4)$, $z_i \sim \text{Unif}(-8,8)$ and $v_i \sim \text{Unif}(-4,4)$. We aim at estimating the conditional quantile vectors corresponding to $\tau = 0.01, 0.05, 0.5, 0.95$ and 0.99. For this purpose, we fit an additive quantile regression model for each τ, using the ELF density and either LAML maximization or the calibration approach of Section 5.2 to select σ. We also consider quantile regression by gradient boosting, as implemented in the *mboost* R package (Hothorn et al., 2010).

The fitted model includes a smooth effect for each covariate, based on spline bases of rank 30. Beside selecting the rank of the bases, the boosting approach requires also selecting the degrees of freedom of each effect, which we set to 6. The number of boosting iterations was selected by minimizing the out-of-bag empirical risk, based on the pinball loss and on 100 boostrap datasets. To select σ using the calibration approach we used 100 boostrap datasets, and we minimized $A(\sigma)$, for each quantile, using Brent method (Brent, 2013). We selected λ using formula (7), with $\epsilon = 0.05$ and κ^2 set equal to the variance estimated using an initial Gaussian additive model fit. Fitting this model has a negligible impact on the computational cost, as it has to be done only once, before starting the calibration.

We firstly evaluate the accuracy of estimated quantile vectors using the pinball loss. To do this we simulated $n = 10^3$ data points from (14), and we fitted an additive model for each τ using each approach. We repeated the process 20 times, and Table 1 reports the average pinball loss and its standard deviation.

Quantile regression using the ELF density and calibration achieves the lowest loss for all quantiles. Also, the variabily of the loss is lower than with the remaining methods, suggesting that this method leads to smoother quantile estimates. LAML estimation of σ seems to perform very badly for the most extreme quantiles. We comment on this issue later. Gradient boosting does worse than calibrated quantile regression, but at least its relative performance seems to fairly constant with τ.

On an Intel 2.50GHz CPU, fitting an additive quantile regression model with the ELF density takes around 0.33s for $\tau = 0.5$ and 0.35s for $\tau = 0.01$, if σ is held fixed. Having

fixed the number of steps, gradient boosting takes around 1.2s and 17s to estimate the same quantiles. Gradient boosting is particularly slow for $\tau = 0.01$ because the empirical risk criterion used by *mboost* is minimized at a large number of steps ($\approx 3 \times 10^4$). The relative cost of calibrating σ and of selecting the number of boosting steps is roughly proportional to the timings given above.

Table 2 reports the empirical coverage, at 95%, 70% and 50% level, achieved by the credible intervals for μ, using calibration or LAML maximization to select σ. The coverage was calculated using 100 simulations from model (14). We do not check the coverage achieved by gradient boosting, because analytic formulas are, to our best knowledge, unavailable and confidence intervals must be obtained by boostrapping.

Notice that the coverage achieved using LAML for selecting σ is well below nominal levels for all quantiles. In particular, for $\tau = 0.99$, the credible intervals are so narrow that the coverage is negligible. Using calibration leads to empirical coverages matching nominal levels exaclty, for $\tau = 0.01$, 0.05 and 5. However, some under-coverage might be occurring for $\tau = 0.95$ and it certainly is for $\tau = 0.99$. Arguably, here we are dealing with a worst case scenario for quantile regression. Indeed, the observation density is highly skewed to the right but, when $\tau \approx 1$, the log-F density is extremely skewed to the left. Hence, the model is severely mispecified, which results in few observation located above the quantile having a strong influence on the fit. When σ is selected by LAML, these points are essentially interpolated, which results in severe overfitting. Calibrating leads to much higher values of σ, which in turn lead to higher estimates of $\boldsymbol{\gamma}$ in the nested LAML optimization. Indeed, when calibration is used, the average effective number of degrees of freedom is between 30 and 40 for $\tau = 0.01, 0.05$ and 0.5, but only 13 for $\tau = 0.99$.

τ	0.01	0.05	0.5	0.95	0.99
CAL95	0.950	0.947	0.946	0.927	0.774
LAML95	0.440	0.752	0.903	0.330	0.069
CAL75	0.754	0.748	0.746	0.708	0.525
LAML75	0.270	0.505	0.673	0.197	0.040
CAL50	0.504	0.501	0.497	0.468	0.326
LAML50	0.160	0.312	0.435	0.116	0.024

Table 2: Empirical coverage achieved by calibration and LAML maximization, for each τ and confidence level.

6 Application to probabilistic load forecasting

We apply the proposed methodology to the dataset considered in the probabilistic electricity load forecasting track of the Global Energy Competition 2014 (GEFCom2014). The dataset covers the period between January 1, 2005 and December 31, 2011 and it includes hourly load consumption and temperatures, the latter being measured at 25 weather stations. In this work we aim at predicting 20 quantiles, equally spaced between $\tau = 0.05$ and $\tau = 0.95$. Given that load consumption is strongly dependent on the time of the day, it is common practice (e.g. Gaillard et al. (2016)) to fit a different model for each hour. To limit the computational burden, here we predict load one week ahead only for the period between 11am and 12am, and we use the year 2005-09 for training, leaving the last two for testing.

Gaillard et al. (2016) describe a quantile regression methodology that ranked 1st on both the load and the price forecasting track of GEFCom2014. They proposed a two-step procedure, called quantGAM, which was partially motivated by the lack of reliable software for fitting quantile additive models. Very briefly, their method firstly fits a

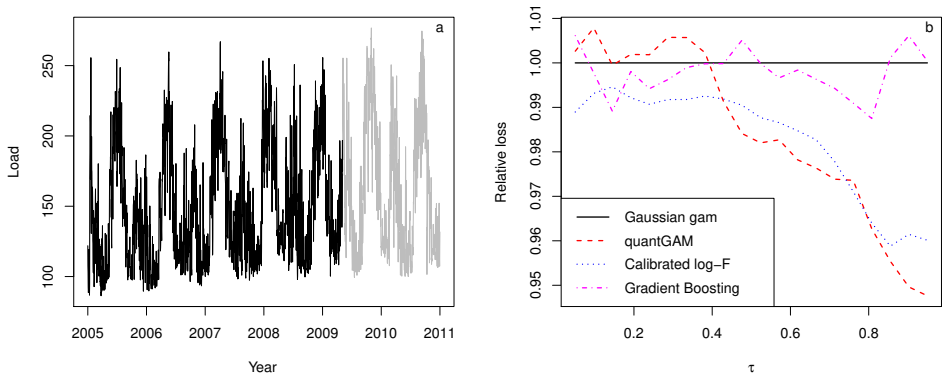

Figure 1: a: observed loads divided between training (black) and testing (grey) sets. b: relative pinball losses for each τ and method.

Gaussian additive model to model mean load and, optionally, a second one to model the variance of the residuals. Then, for each quantile, they fit a linear quantile regression model to the load, using the effects estimated by the Gaussian fits as covariates.

We compare their method to our proposal and to gradient boosting, using the set of covariates proposed by Gaillard et al. (2016): hourly temperatures, averaged over four weather stations (T_t); smoothed temperature (T_t^s), obtained by exponentially smoothing T_t; a cyclic variable indicating the position within the year (S_t); a factor variable indicating the day of the week (D_t); a sequential index representing time (t). See Gaillard et al. (2016) for details.

All variables, apart from D_t, are modelled using splines. For T_t, T_t^s and S_t we use bases of rank 30, while we limit the rank to 4 for the effect of t, which captures the long term trend. A periodic smooth is used for S_t. For gradient boosting we use 6 degrees of freedom for T_t and T_t^s, 15 for S_t and 4 for t. Notice that these parameters need to be tuned manually, which can be time consuming. To calibrate σ and to select the number of boosting steps we use 100 bootstrapped datasets. We use $\epsilon = 0.05$ in (7) to determine λ. Decreasing ϵ further lead to numerical instabilities for the most extreme quantiles. These problems are easily detected in the course of the calibration, because they result in discontinuities in $A(\sigma)$.

Figure 1b shows, for each τ, the pinball loss incurred by each method on the testing set, divided by the pinball loss of a Gaussian additive fit. It is satisfactory to notice that the calibrated quantile regression approach does better than a Gaussian fit for all quantile. The proposed approach does better than quantGAM for $\tau < 0.4$, but it incurs slightly larger losses for higher quantiles. Gradient boosting is marginally better than a Gaussian model, but it does not seem competitive with our approach or with quantGAM for $\tau > 0.4$. In terms of computing time, fitting an additive model for the median of the full dataset takes around 1.7s, if σ is held fixed to the value selected by the calibration. Gradient boosting takes around 50s to fit the same model, using the number of boosting steps that minimize the empirical risk.

7 Conclusion

The main contribution of this work has been to provide a stable and computationally efficient framework for fitting semi-parametric additive quantile models. We have addressed

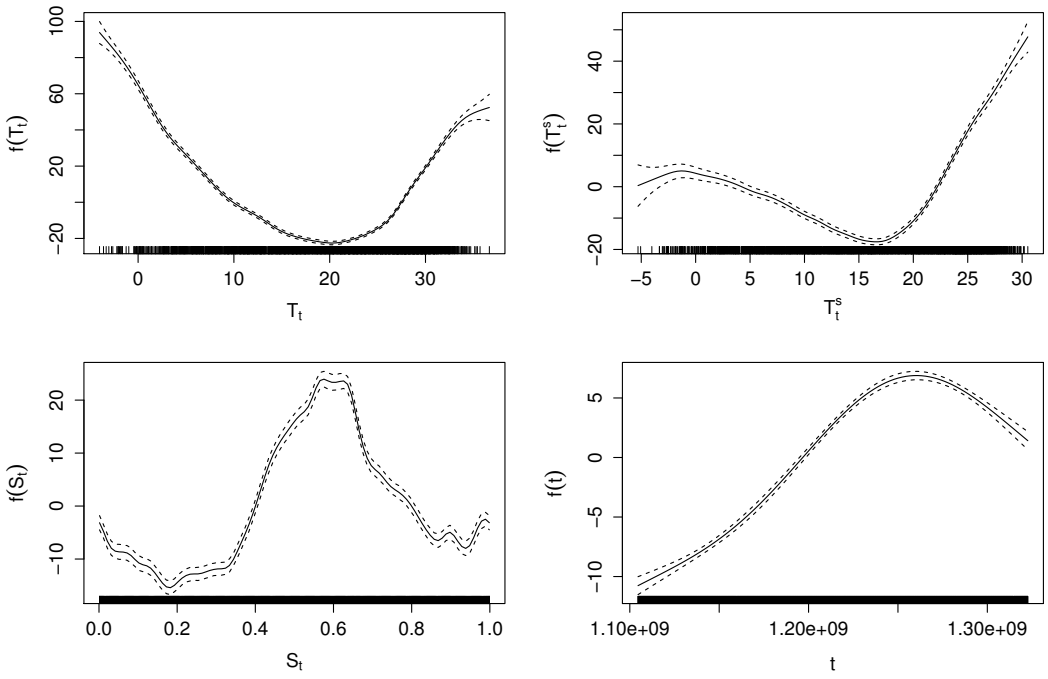

Figure 2: Smooth effects for the median, estimated using calibrated quantile regression with the full GEFCom2014 dataset.

the lack of curvature of the AL density by embedding it into a modified log-F density. While this requires choosing a tuning parameter, λ, we have illustrated how this can be expressed in terms of a more interpretable, and scale invariant, tolerance ϵ. Secondly, we have described a calibration approach for selecting the learning rate σ, which leads to smoother fits and more reliable credible intervals than LAML maximization. In the electricity load forecasting example we have demonstrated that practical utility of the proposed approach, whose performance is very close to that of the competition-winning method of Gaillard et al. (2016).

References

Anderson, T. W. and D. A. Darling (1954). A test of goodness of fit. *Journal of the American statistical association 49*(268), 765–769.

Bissiri, P., C. Holmes, and S. Walker (2013). A general framework for updating belief distributions. *arXiv preprint arXiv:1306.6430*.

Brent, R. P. (2013). *Algorithms for minimization without derivatives*. Courier Corporation.

Gaillard, P., Y. Goude, and R. Nedellec (2016). Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting*.

Hothorn, T., P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner (2010). Model-based boosting 2.0. *The Journal of Machine Learning Research 11*, 2109–2113.

Jones, M. (2008). On a class of distributions with simple exponential tails. *Statistica Sinica 18*(3), 1101–1110.

Kleijn, B., A. van der Vaart, et al. (2012). The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics 6*, 354–381.

Koenker, R. (2013). Quantreg: quantile regression. *R package version 5*.

Mächler, M. (2012). Accurately computing $\log(1-\exp(-|a|))$. *URL http://cran. r-project. org/web/packages/Rmpfr/vignettes/log1mexp-note. pdf*.

Oh, H.-S., T. C. Lee, and D. W. Nychka (2012). Fast nonparametric quantile regression with arbitrary smoothing methods. *Journal of Computational and Graphical Statistics*.

Sriram, K. (2015). A sandwich likelihood correction for bayesian quantile regression based on the misspecified asymmetric laplace density. *Statistics & Probability Letters 107*, 18–26.

Syring, N. and R. Martin (2015). Scaling the gibbs posterior credible regions. *arXiv preprint arXiv:1509.00922*.

Waldmann, E., T. Kneib, Y. R. Yue, S. Lang, and C. Flexeder (2013). Bayesian semiparametric additive quantile regression. *Statistical Modelling 13*(3), 223–252.

Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC press.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(1), 3–36.

Wood, S. N., N. Pya, and B. Safken (2016). On smooth modelling with regular likelihoods. *In press*.

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

Yue, Y. R. and H. Rue (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics & Data Analysis 55*(1), 84–96.

Appendices

A Details regarding the new density

A.1 Derivatives of the log-likelihood

The logarithm of the proposed density (5) is

$$lf(y) = \log \tilde{p}_L(y|\mu,\tau,\sigma) = (1-\tau)\frac{y-\mu}{\sigma} - \lambda \log\left[1+e^{\frac{y-\mu}{\lambda\sigma}}\right] - \log\left\{\lambda\sigma \mathrm{Beta}\big[\lambda(1-\tau),\lambda\tau\big]\right\},$$

When evaluating this numerically, it is important to approximate $\log(1+e^z)$ with $z+e^{-z}$ when $z=(y-\mu)/\lambda\sigma > 18$, as suggested by Mächler (2012). The gradient is

$$\frac{\partial lf(y)}{\partial \mu} = \frac{1}{\sigma}\left[\Phi(y|\mu,\lambda\sigma) - 1 + \tau\right], \quad \frac{\partial lf(y)}{\partial \sigma} = \frac{y-\mu}{\sigma^2}\left[\Phi(y|\mu,\lambda\sigma) - 1 + \tau\right] - \frac{1}{\sigma},$$

where $\Phi(y|\mu,\lambda\sigma)$ is the c.d.f. of a logistic density with location μ and scale $\lambda\sigma$.

The Hessian is

$$\frac{\partial^2 lf(y)}{\partial \mu^2} = -\frac{1}{\sigma}\phi(y|\mu,\lambda\sigma),$$

$$\frac{\partial^2 lf(y)}{\partial \sigma^2} = 2\frac{y-\mu}{\sigma^3}\left[1-\tau-\Phi(y|\mu,\lambda\sigma) - \frac{1}{2}(y-\mu)\phi(y|\mu,\lambda\sigma)\right] + \frac{1}{\sigma^2},$$

$$\frac{\partial^2 lf(y)}{\partial \mu \partial \sigma} = -\frac{1}{\sigma^2}\left[(y-\mu)\phi(y|\mu,\lambda\sigma) + \Phi(y|\mu,\lambda\sigma) - 1 + \tau\right],$$

where $\phi(y|\mu,\lambda\sigma)$ is the p.d.f of a logistic density with location μ and scale $\lambda\sigma$.

Now, define $z=(y-\mu)/(\lambda\sigma)$ so that

$$\Phi(y|\mu,\lambda\sigma) = \Phi(z|0,1) = \Phi(z) = \frac{1}{1+e^{-z}},$$

is the sigmoid function. Also, define $\Phi^{(k)}(y) = \partial\Phi^{(k)}(y)/\partial y^k$. Then, derivatives of higher order are

$$\frac{\partial^3 lf(y)}{\partial \mu^3} = \frac{\Phi^{(2)}(z)}{\lambda^2\sigma^3}, \qquad \frac{\partial^4 lf(y)}{\partial \mu^4} = -\frac{\Phi^{(3)}(z)}{\lambda^3\sigma^4},$$

$$\frac{\partial^3 lf(y)}{\partial \sigma^3} = -\frac{3}{\sigma}\frac{\partial^2 lf(y)}{\partial \sigma^2} + \frac{\lambda z^2}{\sigma^3}\left[3\Phi^{(1)}(z) + z\Phi^{(2)}(z) + \frac{1}{\lambda z^2}\right],$$

$$\frac{\partial^4 lf(y)}{\partial \sigma^4} = -\frac{4}{\sigma}\left[2\frac{\partial^3 lf(y)}{\partial \sigma^3} + \frac{3}{\sigma}\frac{\partial^2 lf(y)}{\partial \sigma^2}\right] - \frac{\lambda z^3}{\sigma^4}\left[4\Phi^{(2)}(z) + z\Phi^{(3)}(z) - \frac{2}{\lambda z^3}\right],$$

$$\frac{\partial^3 lf(y)}{\partial \mu^2 \partial \sigma} = \frac{1}{\lambda\sigma^3}\big[z\Phi^{(2)}(z) + 2\Phi^{(1)}(z)\big],$$

$$\frac{\partial^4 lf(y)}{\partial \mu^3 \partial \sigma} = -\frac{1}{\lambda^2\sigma^4}\big[z\Phi^{(3)}(z) + 3\Phi^{(2)}(z)\big],$$

$$\frac{\partial^3 lf(y)}{\partial \mu \partial \sigma^2} = \frac{1}{\sigma^3}\left\{2\big[\Phi(z) - 1 + \tau\big] + 4z\Phi^{(1)}(z) + z^2\Phi^{(2)}(z)\right\},$$

$$\frac{\partial^4 lf(y)}{\partial \mu \partial \sigma^3} = -\frac{3}{\sigma}\frac{\partial^3 lf(y)}{\partial \mu \partial \sigma^2} - \frac{z}{\sigma^4}\big[6\Phi^{(1)}(z) + 6z\Phi^{(2)}(z) + z^2\Phi^{(3)}(z)\big],$$

$$\frac{\partial^4 lf(y)}{\partial\mu^2\partial\sigma^2} = -\frac{1}{\lambda\sigma^4}\Big\{z^2\Phi^{(3)}(z) + 6z\Phi^{(2)}(z) + 6\Phi^{(1)}(z)\Big\},$$

where

$$\Phi^{(1)}(z) = \frac{\partial\Phi(z)}{\partial z} = \Phi(z)\big[1-\Phi(z)\big],$$
$$\Phi^{(2)}(z) = \Phi^{(1)}(z) - 2\Phi^{(1)}(z)\Phi(z),$$
$$\Phi^{(3)}(z) = \Phi^{(2)}(z) - 2\Phi^{(2)}(z)\Phi(z) - 2\Phi^{(1)}(z)^2.$$

A.2 Saturated log-likelihood, deviance and expected Fisher information

To find the saturated log-likelihood lf_s we need to maximize $\tilde{p}_L(y|\mu,\tau,\sigma)$ wrt μ. Hence we need to impose

$$\frac{\partial lf(y)}{\partial\mu} = \frac{1}{\sigma}\left[\frac{1}{1+e^{-\frac{y-\mu}{\lambda\sigma}}} - 1 + \tau\right] = 0,$$

which leads to

$$\hat{\mu} = \lambda\sigma\log\Big(\frac{\tau}{1-\tau}\Big) + y.$$

Hence, the saturated log-likelihood is

$$lf_s(y) = (1-\tau)\lambda\log\big(1-\tau\big) + \lambda\tau\log(\tau) - \log\Big\{\lambda\sigma\mathrm{Beta}\big[\lambda(1-\tau),\lambda\tau\big]\Big\},$$

and has derivatives

$$\frac{\partial lf_s(y)}{\partial\sigma} = -\frac{1}{\sigma}, \quad \frac{\partial^2 lf_s(y)}{\partial\sigma^2} = \frac{1}{\sigma^2}.$$

This implies that the deviance is

$$dev(y) = 2\big[lf_s(y) - lf(y)\big] = 2\Big\{(1-\tau)\lambda\log\big(1-\tau\big) + \lambda\tau\log(\tau) - (1-\tau)\frac{y-\mu}{\sigma} + \lambda\log\Big[1+e^{\frac{y-\mu}{\lambda\sigma}}\Big]\Big\}.$$

The entry of the expected Fisher information matrix corresponding to μ is

$$E\left[\frac{\partial^2 lf(y)}{\partial\mu^2}\right] = -\frac{1}{\lambda\sigma^2}\int_{-\infty}^{\infty}\frac{e^{-\frac{y-\mu}{\lambda\sigma}}}{\big(1+e^{-\frac{y-\mu}{\lambda\sigma}}\big)^2}\frac{e^{(1-\tau)\frac{x-\mu}{\sigma}}(1+e^{\frac{x-\mu}{\lambda\sigma}})^{-\lambda}}{\lambda\sigma\mathrm{Beta}\big[\lambda(1-\tau),\lambda\tau\big]}dy,$$

and some algebra leads to

$$E\left[\frac{\partial^2 lf(y)}{\partial\mu^2}\right] = -\frac{1}{\sigma^2}\frac{\tau(1-\tau)}{\lambda+1}.$$

A.3 Random variables generation and moments

Notice that, if a r.v. y has density $\tilde{p}_L(y|\mu,\sigma)$, then

$$z \sim B\big[\lambda(1-\tau),\lambda\tau\big],$$

where $B(\alpha,\beta)$ is a Beta distribution and $z = \Phi\{(y-\mu)/\sigma\lambda\}$. Hence, to simulate a random variables y from $\tilde{p}_L(y|\mu,\sigma)$ we could use

$$z \sim B\big[\lambda(1-\tau),\lambda\tau\big], \qquad y = \sigma\lambda\Phi^{-1}(z) + \mu = \sigma\lambda\log\left(\frac{z}{1-z}\right) + \mu.$$

Unfortunately, *rbeta*() in R produces many 1s when the shape parameters are fairly small, leading to the overflow of $z/(1-z)$. A better scheme is

$$u \sim \Gamma\{\lambda(1-\tau), 1\}, \quad v \sim \Gamma(\lambda\tau),$$

$$y = \sigma\lambda(\log u - \log v) + \mu.$$

The former parametrization is useful for deriving the first two moments of $y \sim \tilde{p}_L(y)$, in fact

$$E(y) = E\left[\sigma\lambda \log\left(\frac{z}{1-z}\right) + \mu\right] = \sigma\lambda \log E\left[\left(\frac{z}{1-z}\right)\right] + \mu = \sigma\lambda\big\{\Gamma[\lambda(1-\tau)] - \Gamma(\lambda\tau)\big\} + \mu,$$

where $\psi(x)$ is the digamma function

$$\psi(x) = \frac{dlog\Gamma(x)}{dx} = \frac{\Gamma'(x)}{\Gamma(x)}.$$

The variance is

$$var(x) = var\left[\sigma\lambda \log\left(\frac{z}{1-z}\right) + \mu\right] = \sigma^2\lambda^2\big\{\psi'[\lambda(1-\tau)] + \psi'(\lambda\tau)\big\},$$

where $\psi'(x) = d\psi(x)/dx$ is the trigamma function.

B Stabilizing parameter estimation using the ELF density

B.1 Dealing with zero weights in PIRLS

Quantile regression with the ELF density requires that we work with many weights that can be very close to zero, while the corresponding log-likelihood or deviance derivative is far from zero. This can lead to a situation in which the vector containing $w_i z_i$ is well scaled while the vector containing $\sqrt{|w_i|}z_i$ is very poorly scaled. This scaling problem can reverse the usual stability improvement of QR-based least squares estimation over direct normal equation solution.

Using the notation of Wood (2011) (Section 3), we have

$$\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{V}(\mathbf{I} - 2\mathbf{D}^2)\mathbf{V}^\mathsf{T}\mathbf{Q}_1^\mathsf{T}\sqrt{\bar{\mathbf{W}}}\bar{\mathbf{z}} = \mathbf{R}^{-1}\mathbf{f},$$

by definition of $\mathbf{f}$. Now we can test for stability of the computation to the scaling of $\sqrt{\bar{\mathbf{W}}}\bar{\mathbf{z}}$ by testing whether

$$\mathbf{R}\mathbf{Q}_1^\mathsf{T}\sqrt{\bar{\mathbf{W}}}\bar{\mathbf{z}} = \mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{z}$$

to sufficient accuracy. If it does not, then we recompute $\mathbf{f}$ using

$$\mathbf{f} = \mathbf{V}(\mathbf{I} - 2\mathbf{D}^2)\mathbf{V}^\mathsf{T}\mathbf{R}^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{z}.$$

Another possibility, that may be more convenient when using $\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{K}^\mathsf{T}\sqrt{\bar{\mathbf{W}}}\bar{\mathbf{z}}$, is to test whether $\mathbf{K}^\mathsf{T}\sqrt{\bar{\mathbf{W}}}\bar{\mathbf{z}} = \mathbf{P}^\mathsf{T}\mathbf{W}\mathbf{z}$ to sufficient accuracy, and to use $\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{P}^\mathsf{T}\mathbf{W}\mathbf{z}$ if not.

B.2 Dealing with zero weights in LAML

In this section we indicate the smoothing parameters with $\boldsymbol{\lambda}$, rather than with $\boldsymbol{\gamma}$, in order to be consistent with Wood (2011), and we define $\boldsymbol{\rho} = \log \boldsymbol{\lambda}$. Consider the derivatives of $\log|\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}}|$. In the notation of Wood (2011) (Section 3), we have $(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}})^{-1} = \mathbf{P}\mathbf{P}^\mathsf{T}$ and

$$\begin{aligned}\frac{\partial \log|\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}}|}{\partial \rho_k} &= \mathrm{tr}\left\{(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{X}^\mathsf{T}\frac{\partial \mathbf{W}}{\partial \rho_k}\mathbf{X}\right\} + \lambda_k \mathrm{tr}\left\{(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{S}_k\right\} \\ &= \mathrm{tr}\left\{\mathbf{P}^\mathsf{T}\mathbf{X}^\mathsf{T}\frac{\partial \mathbf{W}}{\partial \rho_k}\mathbf{X}\mathbf{P}\right\} + \lambda_k \mathrm{tr}\left\{\mathbf{P}^\mathsf{T}\mathbf{S}_k\mathbf{P}\right\}.\end{aligned}$$

Then

$$\begin{aligned}\frac{\partial^2 \log|\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}}|}{\partial \rho_k \partial \rho_j} &= \mathrm{tr}\left\{(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{X}^\mathsf{T}\frac{\partial^2 \mathbf{W}}{\partial \rho_k \partial \rho_k}\mathbf{X}\right\} + \delta_k^j \lambda_j \mathrm{tr}\left\{(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{S}_j\right\} \\ &- \mathrm{tr}\left\{(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}})^{-1}\left(\mathbf{X}^\mathsf{T}\frac{\partial \mathbf{W}}{\partial \rho_k}\mathbf{X} + \lambda_j \mathbf{S}_j\right)(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{X}^\mathsf{T}\frac{\partial \mathbf{W}}{\partial \rho_j}\mathbf{X}\right\} \\ &- \lambda_k \mathrm{tr}\left\{(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}})^{-1}\left(\mathbf{X}^\mathsf{T}\frac{\partial \mathbf{W}}{\partial \rho_k}\mathbf{X} + \lambda_j \mathbf{S}_j\right)(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{S}_k\right\}\end{aligned}$$

so that

$$\begin{aligned}\frac{\partial^2 \log|\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_{\boldsymbol{\lambda}}|}{\partial \rho_k \partial \rho_j} &= \mathrm{tr}\left\{\mathbf{P}^\mathsf{T}\mathbf{X}^\mathsf{T}\frac{\partial^2 \mathbf{W}}{\partial \rho_k \partial \rho_j}\mathbf{X}\mathbf{P}\right\} + \lambda_k \mathrm{tr}\left\{\mathbf{P}^\mathsf{T}\mathbf{S}_k\mathbf{P}\right\} \\ &- \mathrm{tr}\left\{\mathbf{P}^\mathsf{T}\mathbf{X}^\mathsf{T}\frac{\partial \mathbf{W}}{\partial \rho_j}\mathbf{X}\mathbf{P}\mathbf{P}^\mathsf{T}\mathbf{X}^\mathsf{T}\frac{\partial \mathbf{W}}{\partial \rho_k}\mathbf{X}\mathbf{P}\right\} \\ &- \lambda_j \mathrm{tr}\left\{\mathbf{P}^\mathsf{T}\mathbf{S}_j\mathbf{P}\mathbf{P}^\mathsf{T}\mathbf{X}^\mathsf{T}\frac{\partial \mathbf{W}}{\partial \rho_k}\mathbf{X}\mathbf{P}\right\} - \lambda_k \mathrm{tr}\left\{\mathbf{P}^\mathsf{T}\mathbf{X}^\mathsf{T}\frac{\partial \mathbf{W}}{\partial \rho_j}\mathbf{X}\mathbf{P}\mathbf{P}^\mathsf{T}\mathbf{S}_k\mathbf{P}\right\} \\ &- \lambda_j \lambda_k \mathrm{tr}(\mathbf{P}^\mathsf{T}\mathbf{S}_j\mathbf{P}\mathbf{P}^\mathsf{T}\mathbf{S}_k\mathbf{P}).\end{aligned}$$

Defining $\mathbf{K} = \mathbf{X}\mathbf{P}$, $T_j = \mathrm{diag}(\partial w_i/\partial \rho_j)$ and $T_{jk} = \mathrm{diag}(\partial^2 w_i/\partial \rho_j \partial \rho_k)$, then this last expression corresponds to the equivalent in Wood (2011) and can be computed in the same way. Notice also that $\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{K}^\mathsf{T}\mathbf{W}\mathbf{z}$.

The point of all this is that we avoid dividing by zero in the definition of T_j and T_{jk}.

C Proof of Theorem 3.1

The starting point is

$$F(\mu^*|\mathbf{x}) - F(\mu_0|\mathbf{x}) = \int \left\{ -\frac{\partial \log \tilde{p}_F(y)}{\partial \mu} - \frac{\partial \rho_\tau(y)}{\partial \mu}\right\} f(y|\mathbf{x})\, dy.$$

which is implied as part of the proof of proposition 1 in Oh et al. (2012). We proceed to bound for the r.h.s.. Simple manipulations lead to

$$\int \left\{ -\frac{\partial \log \tilde{p}_F(y)}{\partial \mu} - \frac{\partial \rho_\tau(y)}{\partial \mu}\right\} f(y|\mathbf{x})\, dy = \int \left\{ \mathbb{1}(y > \mu) - \Phi(y|\mu, \lambda\sigma)\right\} f(y|\mathbf{x})\, dy, \quad (15)$$

where $\mathbb{1}(\cdot)$ is an indicator function and $\Phi(y|\mu, \lambda\sigma)$ is the c.d.f. of a logistic random variable, with location μ and scale $\lambda\sigma$. Then we have

$$\begin{aligned}|F(\mu^*|\mathbf{x}) - F(\mu_0|\mathbf{x})| &\leq \int \left|\mathbb{1}(y > \mu) - \Phi(y|\mu, \lambda\sigma)\right| \sup_y f(y|\mathbf{x})\, dy \\ &= 2 \sup_y f(y|\mathbf{x}) \int_{-\infty}^{\mu} \Phi(y|\mu, \lambda\sigma)\, dy,\end{aligned}$$

where the second equality holds due to the symmetry of the integrand around μ. Using the substitution $z = (y - \mu)/\lambda\sigma$ leads to

$$\begin{aligned} |F(\mu^*|\mathbf{x}) - F(\mu_0|\mathbf{x})| &\leq 2\lambda\sigma \sup_y f(y|\mathbf{x}) \int_{-\infty}^{0} \frac{1}{1+e^{-z}}\, dz \\ &= 2\log(2)\lambda\sigma \sup_y f(y|\mathbf{x}). \end{aligned}$$

Finally, notice that the r.h.s. of (15) makes it clear that, if $f(y|\mathbf{x})$ is symmetric around μ, then $|F(\mu^*|\mathbf{x}) - F(\mu_0|\mathbf{x})| = 0$.