qgam: Bayesian Nonparametric Quantile Regression Modeling in R

Matteo Fasiolo [iD]
University of Bristol

Simon N. Wood [iD]
University of Bristol

Margaux Zaffran [iD]
ENSTA Paris

Raphaël Nedellec [iD]
Électricité de France R&D

Yannig Goude [iD]
Électricité de France R&D

Abstract

Generalized additive models (GAMs) are flexible non-linear regression models, which can be fitted efficiently using the approximate Bayesian methods provided by the **mgcv** R package. While the GAM methods provided by **mgcv** are based on the assumption that the response distribution is modeled parametrically, here we discuss more flexible methods that do not entail any parametric assumption. In particular, this article introduces the **qgam** package, which is an extension of **mgcv** providing fast calibrated Bayesian methods for fitting quantile GAMs (QGAMs) in R. QGAMs are based on a smooth version of the pinball loss of Koenker (2005), rather than on a likelihood function, hence jointly achieving satisfactory accuracy of the quantile point estimates and coverage of the corresponding credible intervals requires adopting the specialized Bayesian fitting framework of Fasiolo, Wood, Zaffran, Nedellec, and Goude (2021b). Here we detail how this framework is implemented in **qgam** and we provide examples illustrating how the package should be used in practice.

Keywords: Bayesian quantile regression, generalized additive models, regression splines, calibrated Bayes, fast Bayesian inference, R.

1. Introduction: Additive quantile modeling in R

Generalized additive models (GAMs, Hastie and Tibshirani 1986) are flexible regression models, where the relation between the response distribution and several covariates is modeled nonparametrically, typically via spline bases expansions. In standard GAMs only one parameter of the response distribution (typically the location) is modeled additively, but Rigby

and Stasinopoulos (2005) developed methods for handling more flexible generalized additive models for location, scale and shape (GAMLSS), where potentially all the parameters of the response distribution are allowed to depend on the covariates. The quantile GAMs (QGAMs) described in this work provide even more flexibility by modeling the quantiles of the conditional response distribution individually, thus avoiding any parametric assumption on the distribution of the response variable.

The purpose of this article is to discuss methods and software for QGAM modeling in R (R Core Team 2021). In particular, we focus on the **qgam** R package (Fasiolo, Wood, Zaffran, Goude, and Nedellec 2021a), which implements the fast calibrated Bayesian fitting methods proposed by Fasiolo *et al.* (2021b) and which is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=qgam`. The **qgam** package is an extension of the recommended **mgcv** package (Wood 2017, 2021), which provides tools for building, fitting and visualizing GAMs and GAMLSS models. A new package is required, because QGAMs can not be handled using the standard GAM methods implemented by **mgcv**. In particular, QGAMs are based on the pinball loss function (Koenker and Bassett 1978), rather than on a probabilistic model of the response distribution, and the absence of a likelihood function impedes direct application of the Bayes' rule when updating the prior distribution on the regression coefficients given the observed responses. While this problem can be overcome by adopting the coherent Bayesian belief updating framework of Bissiri, Holmes, and Walker (2016), which effectively leads to the application of Bayes' rule using a loss-based pseudo-likelihood, naïvely plugging such a pseudo-likelihood into standard Bayesian fitting methods can lead to inaccurate quantile estimates and inadequate coverage of the corresponding credible intervals as discussed, for instance, in Yang, Wang, and He (2016) and Sriram (2015). To avoid such issues, **qgam** implements the calibrated Bayesian methods of Fasiolo *et al.* (2021b), which explicitly aim at selecting the "learning rate" tuning parameter of the loss so as to achieve near-nominal frequentist coverage of the quantile credible intervals. Furthermore, **qgam** bases quantile regression on a smoothed version of the pinball loss, which enables the adoption of fast maximum a posteriori (MAP) and empirical Bayes methods to estimate the regression coefficients and select their prior variance hyper-parameters, respectively. The smoothness of the new loss is determined by minimizing the asymptotic mean squared error (MSE) of the estimated regression coefficients, approximated using a location-scale GAM.

To our best knowledge and at the time of writing, the QGAM fitting framework proposed by Fasiolo *et al.* (2021b) is the only one able to estimate all the regression coefficients and prior hyper-parameters using fast direct optimization methods and to provide, at no extra computational cost, credible intervals which achieve adequate coverage for tail quantiles. It should be clarified that, when we describe **qgam** as "fast", we mean that it is the fastest additive quantile regression method we know of, which has the properties just described. Hence, for example, the `rqss` function provided by the **quantreg** package (Koenker 2021) might be faster, if the smoothing parameters are known or if the model of interest contains at most one or two such parameters, which could be selected by minimizing a model selection criterion numerically, as discussed in Koenker (2011, Section 3). Another alternative to the calibrated Bayesian estimation methods provided by **qgam** is additive quantile regression via gradient boosting, available in R via the **mboost** package (Hothorn, Bühlmann, Kneib, Schmid, and Hofner 2021). Fasiolo *et al.* (2021b) show that boosting can provide accurate point estimates, but selecting the optimal number of boosting steps requires running a computationally intensive cross-validation routine. Furthermore, the uncertainty of the estimated

quantiles must be obtained by bootstrapping. Waldmann, Kneib, Yue, Lang, and Flexeder (2013) propose Bayesian methods for estimating QGAMs via the **BayesX** stand-alone software (Brezger, Kneib, and Lang 2005), with a specific quantile regression family being accessible from R using the **bamlss** package (Umlauf, Klein, Simon, and Zeileis 2021). However, their proposal is based on Markov chain Monte Carlo (MCMC) methods, which are slower than the direct optimization methods adopted here, and the resulting credible intervals struggle to achieve nominal frequentist coverage for tail quantiles, as detailed in Waldmann *et al.* (2013). Another alternative is the **VGAM** (Yee 2010) R package but, as for **quantreg**, the prior smoothing hyper-parameters have to be selected manually. Lin, Bondell, Zhang, and Zou (2013) focus on variable selection, not smoothing, hence the corresponding software can not be considered an alternative to **qgam**.

The rest of the paper is structured as follows. In Section 2 we first present the basic structure of QGAMs, then we show how such models can be fitted using the fast calibrated Bayesian methods of Fasiolo *et al.* (2021b) and we explain how the fitting framework is implemented in the **qgam** package. In Section 3 we illustrate how the package can be used for quantile GAM modeling. In particular, the first example introduces the basic features of the package, while the second example considers a more realistic application, focused on electricity demand forecasting using functional effects. In Section 4 we outline a promising direction for future work on additive quantile modeling with **qgam**.

2. Models, methods and software

2.1. General structure of additive quantile regression models

Let y be a continuous random variable with conditional distribution $p(y|\boldsymbol{x})$, where $\boldsymbol{x}$ is a p-dimensional vector of covariates. Define also the conditional quantiles of $p(y|\boldsymbol{x})$ by $\mu_\tau(\boldsymbol{x}) = F^{-1}(\tau|\boldsymbol{x})$, where $\tau \in (0,1)$ and F^{-1} is the inverse conditional cumulative distribution function (CDF) of y. In quantile regression (Koenker and Bassett 1978) the conditional quantiles are modeled individually, without specifying a model for $p(y|\boldsymbol{x})$. Direct quantile estimation is generally achieved by exploiting the following alternative definition of a quantile

$$\mu_\tau(\boldsymbol{x}) = \underset{\mu}{\operatorname{argmin}} \, \mathbb{E}\{\rho_\tau(y - \mu)|\boldsymbol{x}\}, \tag{1}$$

where

$$\rho_\tau(z) = (\tau - 1)\frac{z}{\sigma}\mathbb{1}(z < 0) + \tau\frac{z}{\sigma}\mathbb{1}(z \geq 0)$$

is the scaled version of the so-called "pinball" or "check" loss. $\sigma > 0$ is a scale parameter, which can potentially depend on the covariates $\boldsymbol{x}$. Its role will be clarified in Section 2.2. In **qgam** the pinball loss, which is piecewise linear and has discontinuous derivatives, is replaced by the extended log-f (ELF) loss (Fasiolo *et al.* 2021b), that is

$$\tilde{\rho}_\tau(z) = (\tau - 1)\frac{z}{\sigma} + \lambda\log(1 + e^{\frac{z}{\lambda\sigma}}),$$

where $\lambda > 0$, and the pinball loss is recovered as $\lambda \to 0^+$. The ELF loss is a smoothed version of the pinball loss which, as will be explained later, has the advantages of leading to

more accurate quantile estimates (if λ is tuned adequately) and of enabling the use of efficient computational methods for model fitting.

In this work we assume that the conditional quantiles have an additive structure, that is

$$\mu(\boldsymbol{x}) = \sum_{j=1}^{J} f_j(\boldsymbol{x}),$$

where the f_j's are parametric, random or smooth effects. The latter are built using spline bases expansions, so the jth effect can be written

$$f_j(\boldsymbol{x}) = \sum_{k=1}^{K_j} b_k^j(\boldsymbol{x})\beta_k^j,$$

where $b_1^j, \ldots, b_{K_j}^j$ are the spline basis functions used to built the jth effect and $\beta_1^j, \ldots, \beta_{K_j}^j$ are the corresponding regression coefficients. The basis functions are known and fixed, while the regression coefficients must be estimated. Note that the quantile μ depends both on $\boldsymbol{x}$ and $\boldsymbol{\beta}$, but in the following we will refer to it using $\mu(\boldsymbol{x})$ or $\mu(\boldsymbol{\beta})$, depending on the context.

In a Bayesian framework, the complexity of the smooth and random effects is controlled using a prior distribution on the regression coefficients, which we indicate with $p(\boldsymbol{\beta})$. In this work we assume that the prior is an improper multivariate Gaussian distribution, centered at $\boldsymbol{0}$ and with positive semi-definite precision matrix $\mathbf{S}^\gamma = \sum_{l=1}^m \gamma_l \mathbf{S}_l$. Here the $\mathbf{S}_l$'s are positive semi-definite matrices, scaled by the positive parameters $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_m\}$. We assume that the $\mathbf{S}_l$'s are given, while the vector $\boldsymbol{\gamma}$ needs to be selected. The $\mathbf{S}_l$'s related to random effects are often simple diagonal matrices, while the $\mathbf{S}_l$'s related to smooth effects have more complex structures, aimed at penalizing departures from (some definition of) smoothness. Hence, increasing $\boldsymbol{\gamma}$ leads to a prior where the random effects coefficients are more concentrated around zero and the smooth effects less wiggly. In the following we refer to $\boldsymbol{\gamma}$ as the vector of smoothing parameters, with the understanding that some of its elements could instead be controlling the prior precision of the random effects. In general, there needs not to be a one-to-one correspondence between the smoothing parameters and the effects because, for example, a smoothing parameter might be determining the prior precision of several effects or the complexity of a single effect might be controlled using multiple smoothing parameters. See Wood (2017) for a detailed account on smooth and random effects model structures, in an additive modeling context.

This section outlined the basic QGAM modeling framework, the next two will detail how such models can be estimated within a fast calibrated Bayesian framework.

2.2. Performing the Bayesian update under the ELF loss

Let us indicate with $p(y|\boldsymbol{\beta})$ the true conditional distribution of the response, where the dependency on $\boldsymbol{x}$ has been temporarily suppressed to simplify the notation, and indicate with $p(\boldsymbol{\beta})$ the prior distribution, which is implicitly a function of the smoothing parameters $\boldsymbol{\gamma}$. Assume for the moment that the latter have been fixed to some value, so that the prior is given. Recall that we are basing additive quantile regression on the ELF loss, rather than on a probabilistic model for $p(y|\boldsymbol{\beta})$. This is an impediment to performing Bayesian inference for QGAMs, as the lack of a likelihood function prevents us from applying Bayes' rule to

update $p(\boldsymbol{\beta})$ to the corresponding posterior, $p(\boldsymbol{\beta}|y)$. We address the issue by adopting the belief updating (BU) framework of Bissiri *et al.* (2016), which allows to perform the Bayesian update using a general loss function, rather than a likelihood. In particular, as detailed in Fasiolo *et al.* (2021b), applying the BU framework to the ELF loss leads to the following update formula

$$p(\boldsymbol{\beta}|y) \propto \tilde{p}_\tau\{y - \mu(\boldsymbol{\beta})\}p(\boldsymbol{\beta}), \tag{2}$$

where

$$\tilde{p}_\tau(y - \mu) = \frac{e^{-\tilde{\rho}_\tau\{y-\mu\}}}{\int e^{-\tilde{\rho}_\tau\{y-\mu\}}dy} = \frac{e^{(1-\tau)\frac{y-\mu}{\sigma}}(1 + e^{\frac{y-\mu}{\lambda\sigma}})^{-\lambda}}{\lambda\sigma\mathrm{Beta}\big[\lambda(1-\tau), \lambda\tau\big]}, \tag{3}$$

is the probability density function (PDF) of the ELF distribution, which is an extension of the log-f distribution of Jones (2008). Here λ determines the smoothness of the loss, while $1/\sigma$ plays the role of a "learning rate", determining the relative weights of the loss-based pseudo-likelihood and the prior. In fact, letting $\lambda \to 0^+$ leads to $p(\boldsymbol{\beta}|y) \propto \exp\{-\rho_\tau\{y - \mu\}/\sigma\}p(\boldsymbol{\beta})$ and, as $1/\sigma$ increases, the loss-based likelihood progressively dominates the prior, thus leading to faster learning and a higher risk of overfitting. Decreasing $1/\sigma$ has the opposite effect. In contexts where the variance of y varies strongly with $\boldsymbol{x}$, it can be advantageous to let σ depend on $\boldsymbol{x}$ via the decomposition $\sigma(\boldsymbol{x}) = \sigma_0\tilde{\sigma}(\boldsymbol{x})$, where σ_0 is the baseline or average learning rate and $\tilde{\sigma}(\boldsymbol{x})$ is its $\boldsymbol{x}$-dependent component.

Having defined (2) which, following Syring and Martin (2015), we refer to as the Gibbs posterior, the next section outlines how the fast calibrated Bayesian methods proposed by Fasiolo *et al.* (2021b), and implemented by **qgam**, can be used to fit ELF-based QGAMs.

2.3. Fast calibrated Bayesian model fitting methods

We fit quantile GAMs using three nested optimization routines, in particular:

1. in the outer iteration, the baseline learning rate $1/\sigma_0$ is selected by minimizing a calibration loss function numerically;

2. for fixed σ_0, the loss smoothness λ and the $\boldsymbol{x}$-dependent component of the learning-rate $\tilde{\sigma}(\boldsymbol{x})$ are determined using closed-form expressions, while the smoothing parameters $\boldsymbol{\gamma}$ are selected by numerically optimizing an intermediate criterion;

3. for fixed $\sigma(\boldsymbol{x}) = \sigma_0\tilde{\sigma}(\boldsymbol{x})$, λ and $\boldsymbol{\gamma}$, the regression coefficients $\boldsymbol{\beta}$ are estimated by numerically optimizing an inner criterion.

Of course parameter $\tau \in (0, 1)$, which indicates the quantile of interest, is kept fixed throughout. The fact that the three iterations are nested, not sequential, means that evaluating the outer objective function requires solving the intermediate optimization problem and, in turn, each step of the latter entails running the inner optimization to convergence. In the following we provide a methodological outline of the three iterations, going from the inner to the outer one, and in Section 2.4 we describe how the whole procedure is implemented in **qgam**.

Inner iteration: MAP estimation of the regression coefficients

The inner iteration estimates the regression coefficients using efficient MAP methods, for fixed $\boldsymbol{\gamma}$, $\sigma(\boldsymbol{x})$ and λ. In particular, let $\mathbf{y} = \{y_1, \ldots, y_n\}$ be the vector of observed responses and

note that the conditional quantile is modeled by $\mu(\boldsymbol{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, where $\mathbf{x}_i$ is the ith row of the $n \times d$ design matrix $\mathbf{X}$, containing the spline basis functions evaluated at $\boldsymbol{x}_i$. It is easy to show that maximizing the logarithm of the ELF-based Gibbs posterior (2) is equivalent to minimizing the criterion

$$\tilde{V}_D\{\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma(\boldsymbol{x}), \lambda\} = \sum_{i=1}^n \mathrm{Dev}_i\{\boldsymbol{\beta}, \sigma(\boldsymbol{x}_i), \lambda\} + \sum_{l=1}^m \gamma_l \boldsymbol{\beta}^\top \mathbf{S}_l \boldsymbol{\beta}, \tag{4}$$

where $\mathrm{Dev}_i\{\boldsymbol{\beta}, \sigma(\boldsymbol{x}_i), \lambda\}$ is the ith deviance component, based on the ELF density (3). Criterion (4) is smooth and can be minimized efficiently with respect to (w.r.t.) $\boldsymbol{\beta}$ using a penalized iteratively re-weighted least squares (PIRLS) algorithm, as detailed in Fasiolo *et al.* (2021b).

Intermediate iteration: Selecting the smoothing parameters and ELF loss smoothness

The intermediate iteration selects the smoothing parameters by maximizing a Laplace approximation to the ELF-based marginal likelihood. The latter is

$$\log p\{\mathbf{y}|\boldsymbol{\gamma}, \sigma(\boldsymbol{x}), \lambda\} = \int \tilde{p}_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{\gamma}) d\boldsymbol{\beta}, \tag{5}$$

where we have made explicit the dependence of the prior on the smoothing parameters. Applying a Laplace approximation to the intractable integral (5) leads to the following Laplace approximate marginal likelihood (LAML) criterion

$$\begin{aligned} G\{\boldsymbol{\gamma}, \sigma(\boldsymbol{x}), \lambda\} = & -\frac{1}{2}\tilde{V}_D\{\hat{\boldsymbol{\beta}}, \boldsymbol{\gamma}, \sigma(\boldsymbol{x}), \lambda\} + \tilde{\mathrm{ll}}\{\sigma(\boldsymbol{x}), \lambda\} \\ & -\frac{1}{2}\Big[\log|\mathbf{X}^\top \mathbf{W}\mathbf{X} + \mathbf{S}^\gamma| - \log|\mathbf{S}^\gamma|_+\Big] + \frac{M_p}{2}\log(2\pi), \end{aligned} \tag{6}$$

where $\hat{\boldsymbol{\beta}}$ is the minimizer of (4), estimated using the inner iteration, $\tilde{\mathrm{ll}}\{\sigma(\boldsymbol{x}), \lambda\}$ is the saturated log-likelihood corresponding to the ELF density (3), $\mathbf{W}$ is a diagonal matrix such that $W_{ii} = 1/2\partial^2 Dev_i/\partial\mu_i^2$ and 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. Fasiolo *et al.* (2021b) focus on a marginal loss criterion, which is equal to the negative of (5) up to a normalization constant, to stress that σ_0 should not be selected by jointly maximizing (6) w.r.t. $\boldsymbol{\gamma}$ and σ_0. In fact, σ_0 is not the scale parameter of a standard GAM, but a learning rate that should be selected using an outer calibration procedure. Here we prefer to adopt a likelihood-based terminology to emphasize that (6) can be maximized w.r.t. $\boldsymbol{\gamma}$ using the numerically stable methods of Wood, Pya, and Sëfken (2016), which are aimed at standard probabilistic GAMs.

Before describing the outer iteration, we need to specify how $\tilde{\sigma}(\boldsymbol{x})$ and λ, which are held constant throughout the intermediate and inner iterations, are determined for fixed σ_0. Recall that $\sigma(\boldsymbol{x}) = \sigma_0 \tilde{\sigma}(\boldsymbol{x})$ and define the scaled loss bandwidth $h(\boldsymbol{x}) = \sigma(\boldsymbol{x})\lambda$. Assume that the responses follow the location-scale model $y_i|\boldsymbol{x}_i \sim \alpha(\boldsymbol{x}_i) + \kappa(\boldsymbol{x}_i)z_i$, where the z_i's are i.i.d with $\mathbb{E}(z|\boldsymbol{x}) = 0$ and $\mathrm{var}(z|\boldsymbol{x}) = 1$. Fasiolo *et al.* (2021b) show that, under the above location-scale model and further assumptions specified therein, the asymptotic MSE of the regression coefficients is minimized by

$$\tilde{h}^*(\boldsymbol{x}) = \left[\frac{d}{n}\frac{9f_z\{F_z^{-1}(\tau)\}}{\pi^4 f_z'\{F_z^{-1}(\tau)\}^2}\right]^{\frac{1}{3}}\kappa(\boldsymbol{x}), \tag{7}$$

where f_z, f_z' and F_z are the PDF of z, its first derivative and its CDF, $F_z^{-1}(\tau)$ is the τth quantile of z and d the dimension of $\boldsymbol{\beta}$. Having determined $\tilde{h}^*(\boldsymbol{x})$, we impose $n^{-1} \sum_i \tilde{\sigma}(\boldsymbol{x}_i) = 1$ so that $\lambda = n^{-1} \sum_i \tilde{h}^*(\boldsymbol{x}_i)/\sigma_0$. The idea is to use $\tilde{\sigma}(\boldsymbol{x})$ to modulate the baseline scaled loss smoothness, $\lambda\sigma_0$, and learning rate, $1/\sigma_0$, to make them respectively directly and inversely proportional to the conditional variance of y. Note that (7) implies that the asymptotic MSE is not minimized by imposing $h(\boldsymbol{x}) = 0$ or, equivalently, $\lambda = 0$. That is, under the assumptions detailed in Fasiolo *et al.* (2021b), adopting the original pinball loss leads to asymptotically less accurate quantile estimates, relative to those based on the ELF loss with smoothness determined via (7).

For fixed σ_0, we obtain $\tilde{\sigma}(\boldsymbol{x})$ and λ using the formulas just mentioned. This, of course, requires estimates of $\alpha(\boldsymbol{x})$ and $\kappa(\boldsymbol{x})$ in the location-scale model, as well as of f_z and F_z. Such estimates need to be obtained only once, before initiating the nested QGAM fitting procedure described here, hence we will discuss the specific approach we follow in Section 2.4, where we also describe its software implementation.

Outer iteration: Selection of the learning rate by calibrated Bayes

Recall that $1/\sigma_0$ is the baseline learning rate, which determines the relative weight of the loss and the prior in the Gibbs posterior (2). Increasing $1/\sigma_0$ leads to faster learning, that is wigglier fitted quantile curves and narrower posterior credible intervals. Fasiolo *et al.* (2021b) selects σ_0 using a calibration procedure aimed at guaranteeing that the credible intervals for $\mu(\boldsymbol{x})$ approximately achieve the correct frequentist coverage. This is achieved by minimizing

$$\widehat{\text{IKL}}(\sigma_0) = n^{-1} \sum_{i=1}^{n} \left[\frac{\hat{v}_s(\boldsymbol{x}_i)}{v(\boldsymbol{x}_i)} + \log \frac{v(\boldsymbol{x}_i)}{\hat{v}_s(\boldsymbol{x}_i)} \right]^{\zeta}, \tag{8}$$

which is an estimate of the integrated Kullback-Leibler (IKL) divergence

$$\text{IKL}(\sigma_0) = \int \text{KL}\big[\text{N}\{\mu(\boldsymbol{x}), v_s(\boldsymbol{x})\}, \text{N}\{\mu(\boldsymbol{x}), v(\boldsymbol{x})\}\big]^{\zeta} p(\boldsymbol{x})d\boldsymbol{x} \propto \int \left\{ \frac{v_s(\boldsymbol{x})}{v(\boldsymbol{x})} + \log \frac{v(\boldsymbol{x})}{v_s(\boldsymbol{x})} \right\}^{\zeta} p(\boldsymbol{x})d\boldsymbol{x}.$$

Here $\text{KL}(\cdot, \cdot)$ and $\text{N}(\cdot, \cdot)$ indicate respectively the KL divergence and the univariate Gaussian distribution, ζ is a positive parameter which we fix to $1/2$, $v(\boldsymbol{x}) = \mathbf{x}^\top \mathbf{V}\mathbf{x}$ and $v_s(\boldsymbol{x}) = \mathbf{x}^\top \mathbf{V}_s\mathbf{x}$ are the posterior variance of $\mu(\boldsymbol{x})$ under two alternative posterior covariance matrices for $\boldsymbol{\beta}$. In particular $\mathbf{V} = (\mathcal{I} + \mathbf{S}^\gamma)^{-1}$ and $\mathbf{V}_s = (\mathcal{I}\boldsymbol{\Sigma}_\nabla^{-1}\mathcal{I} + \mathbf{S}^\gamma)^{-1}$, where $\mathcal{I}$ is the negative Hessian of the ELF-based log-likelihood and $\boldsymbol{\Sigma}_\nabla = \text{cov}[\nabla_{\boldsymbol{\beta}}\tilde{\rho}\{y - \mu(\boldsymbol{\beta})\}]|_{\boldsymbol{\beta}=\hat{\beta}}$ is the covariance matrix of the gradient of the ELF loss, under the data generating process. $\hat{v}_s(\boldsymbol{x}_i)$ is an estimate of $v_s(\boldsymbol{x}_i)$, based on the regularized estimator of $\mathbf{V}_s$ described in Fasiolo *et al.* (2021b), who use a different notation and indicate $v_s(\boldsymbol{x}_i)$ and $\mathbf{V}_s$ by $\tilde{v}(\boldsymbol{x}_i)$ and $\tilde{\mathbf{V}}$, respectively. The IKL objective function is generally smooth and convex, hence we minimize it efficiently using Brent's method (Brent 2013).

Fasiolo *et al.* (2021b) propose also an alternative calibration procedure, based on another version of the IKL loss where $\hat{v}_s(\boldsymbol{x})$ is substituted by the sample pointwise variance of the estimated quantile $\hat{\mu}(\boldsymbol{x})$, estimated by bootstrapping. Such a procedure requires re-fitting the model to several bootstrap samples, hence it is computationally more expensive than minimization of the IKL loss defined above, but it can lead to better coverage in small samples. We refer to Fasiolo *et al.* (2021b) for explanations regarding how minimizing either

Function name	Description
qgam	Fits a QGAM for a single quantile τ. It is analogous to the gam function in **mgcv**, hence any smooth effect type available in gam is also available in qgam.
mqgam	Fits the same QGAM to a vector of k quantiles $\tau_1, \dots, \tau_k$ more efficiently than by calling qgam repeatedly.
tuneLearnFast	Selects the learning rate by minimizing (8) using Brent's method.
tuneLearn	Evaluates (8) on a grid of values of σ_0.
check	Given a model fitted using qgam, produces some diagnostics plots.
qdo	A wrapper for using standard generics (e.g., summary) on the output of mqgam. Needed because mqgam does not output an object of class 'gam'.

Table 1: Main functions provided by **qgam**.

version of the IKL loss w.r.t. σ_0 leads to better frequentist coverage of the posterior credible intervals, relative to joint maximization of LAML (6) w.r.t. σ_0 and $\boldsymbol{\gamma}$. Note, however, that the methods of Fasiolo *et al.* (2021b) do not propagate forward the uncertainty in the estimation of $\sigma(\mathbf{x})$, $\tilde{h}^*(\boldsymbol{x})$ and $\boldsymbol{\gamma}$ and when computing $\mathbf{V}$ and $\mathbf{V}_s$. The effect of overlooking such additional sources of uncertainty on the coverage of the resulting posterior credible intervals has not been studied in the context of QGAMs and might be non-negligible when dealing with small samples.

This section outlined the Bayesian QGAM fitting framework of Fasiolo *et al.* (2021b) from a methodological point of view. The next section explains how the fitting framework just described is implemented in the **qgam** package.

2.4. Software implementation of the fitting framework

The **qgam** package is an extension of **mgcv** providing additional tools for handling QGAMs. Here we focus mainly on how the package implements the fitting framework of Fasiolo *et al.* (2021b), while in Section 3 we provide usage examples.

Table 1 lists the main functions provided by **qgam**. The most commonly used one is itself called qgam and has the following arguments

```
qgam(form, data, qu, lsig = NULL, err = NULL,
  multicore = !is.null(cluster), cluster = NULL, ncores = detectCores() - 1,
  paropts = list(), control = list(), argGam = NULL)
```

Arguments form and data have the same meaning as in mgcv::gam (henceforth just gam), qu is the quantile of interest τ, lsig is $\log \sigma_0$ and err is a positive parameter allowing to set the ELF loss smoothness manually. By default qgam determines the loss smoothness automatically, using the methods described in Section 2.3, hence most users will not need to use argument err. However, Appendix A provides some guidelines for users who decide to select the loss smoothness manually. In qgam version 1.3.4, arguments multicore, cluster, ncores and paropts are relevant only when the learning rate is calibrated by bootstrapping, and can be used to perform bootstrap re-fitting in parallel. control is a list of control parameters useful, for instance, to choose the type of calibration used and to suppress the text output printed by qgam, while argGam is a list of arguments to be passed to gam.

To describe how **qgam** implements the fitting framework of Fasiolo *et al.* (2021b), assume that we want to fit a QGAM for quantile $\tau = 0.5$, with model formula `myForm = list(y ~ s(x1) + s(x2), ~ s(x3))` and using data from a 'data.frame' called `myDat`. Here `y ~ s(x1) + s(x2)` is the model used for the median, that is $\mu_\tau(\boldsymbol{x}) = f_1(x_1) + f_2(x_2)$, where f_1 and f_2 are two smooth effects constructed, by default, using ten thin plate splines basis functions. The second element of the formula, `~ s(x3)`, is the model used for the variance $\kappa(\boldsymbol{x})$ in the preliminary location-scale GAM fit. The first thing to point out is that, if σ_0 and the bandwidth $h(\boldsymbol{x})$ were known, QGAMs could be fitted using `gam` directly. In fact, we could fit a median quantile model using

```
gam(myForm[[1]], data = myDat,
  family = elf(theta = 1, qu = 0.5, co = rep(1, nrow(myDat))))
```

where `qgam::elf` is a family providing the ELF log-likelihood function and its derivatives. Its arguments `theta` and `co` correspond respectively to $\log \sigma_0$ and $h(\boldsymbol{x})$, which have been arbitrarily fixed to 1 here. Hence, given σ_0 and $h(\boldsymbol{x})$, the intermediate and inner iterations for selecting $\boldsymbol{\gamma}$ and estimating $\boldsymbol{\beta}$ take place inside `gam`. But these parameters are generally unknown, so **qgam** determines σ_0 using the outer iteration and $h(\boldsymbol{x})$ via the preliminary location-scale GAM fit. In particular, upon calling

```
qgam(myForm, data = myDat, qu = 0.5)
```

the `qgam` function executes the following pseudo-code:

1. Estimate $\alpha(\boldsymbol{x})$ and $\kappa(\boldsymbol{x})$ by fitting a Gaussian location-scale GAM

   ```
   gam(myForm, data = myDat, family = gaulss)
   ```

   where `mgcv::gaulss` is a Gaussian location-scale family.

2. Numerically fit the sinh-arch density of Jones and Pewsey (2009) to the standardized Gaussian GAM residuals, to provide estimates of f_z, f'_z and $F_z^{-1}(\tau)$.

3. Estimate $\tilde{h}^*(\boldsymbol{x})$ by plugging the estimates obtained in the previous two steps into (7), where d is set to the number of effective degrees of freedom used to model $\alpha(\boldsymbol{x})$ in step 1.

4. Call the `tuneLearnFast` function, which minimizes the IKL loss w.r.t. $\log \sigma_0$ using an outer Brent algorithm. For each trial value of $\log \sigma_0$:

   (a) Call `gam` with arguments `formula = myForm[[1]]` and `family = elf`, the latter having parameter `theta` fixed to the current value of $\log \sigma_0$ and `co` fixed to $\tilde{h}^*(\boldsymbol{x})$. ELF parameters λ and $\sigma(\boldsymbol{x})$ are obtained within `elf` by $\lambda = n^{-1} \sum_i \tilde{h}^*(\boldsymbol{x}_i)/\sigma_0$ and then $\sigma(\boldsymbol{x}_i) = \tilde{h}^*(\boldsymbol{x}_i)/\lambda$. The `gam` call performs the intermediate iteration for selecting $\boldsymbol{\gamma}$ and the inner iteration for estimating $\boldsymbol{\beta}$.

   (b) Estimate the IKL loss using (8), where $v(\boldsymbol{x})$ is easily obtained from the `gam` output, while $v_s(\boldsymbol{x})$ is estimated using the regularized estimator of Fasiolo *et al.* (2021b).

   Upon convergence, return the value of $\log \sigma_0$ minimizing the IKL loss.

5. Call `gam` as in step 4(a), with `elf` parameter `theta` fixed to the value of $\log \sigma_0$ returned by `tuneLearnFast`, and `co` set to $\tilde{h}^*(\boldsymbol{x})$.

6. Return the output of the last `gam` call, with the output of `tuneLearnFast` call stored in its `$calibr` component.

Note that a variety of different density estimators could be used in step 2 above. We use the sinh-arch distribution because it provides a good balance between flexibility and stability. However, future versions of `qgam` might allow users to provide their own density estimator. Similar considerations hold for the Gaussian GAM used in step 1. The `tuneLearnFast` function of step 4 has exactly the same arguments as `qgam`, and it returns a list containing diagnostic information regarding the outer iteration and the value of σ_0 which minimizes the IKL loss. The output of `qgam` is an object of class 'qgam' which inherits from class 'gam' which in turn inherits from classes 'glm' and 'lm' and which can be manipulated using any of the S3 methods available for objects of class 'gam'. Version 1.3.4 of the **qgam** package does not provide any S4 method or class.

3. Examples

3.1. A basic example: The motorcycle data set

Here we consider the classic motorcycle accident data set of Silverman (1985), available in the **MASS** package (Venables and Ripley 2002; Ripley 2021). We start by loading the data and fitting a QGAM for quantile $\tau = 0.9$ as follows

```
R> data("mcycle", package = "MASS")
R> library("qgam")
R> fitCycle1 <- qgam(list(form = accel ~ s(times, k = 20, bs = "ad"),
+    ~ s(times)), data = mcycle, qu = 0.9)
```

We are using the model $\mu_\tau(\text{time}) = f(\text{time})$, where $f(\cdot)$ is an adaptive smooth effect built using 20 P-splines basis functions (Eilers and Marx 1996). We use an adaptive smooth because, as shown in Figure 1, the shape of the conditional quantile varies sharply between roughly 10 and 40 milliseconds, but it is otherwise quite flat. Recall that the second part of the formula, `~ s(times)`, is not used to model the quantile, but the variance $\kappa(\boldsymbol{x})$ of the response within the location-scale GAM.

Given that `fitCycle1` inherits from class 'gam', generic function calls such as

```
R> head(predict(fitCycle1), 5)
```

```
          1           2           3           4           5
0.460372275 0.421737850 0.271621363 0.146774904 0.005606131
```

dispatch to the relevant method (here the `predict` method for 'gam' objects). Similarly, the fitted effects in the quantile model can be plotted using
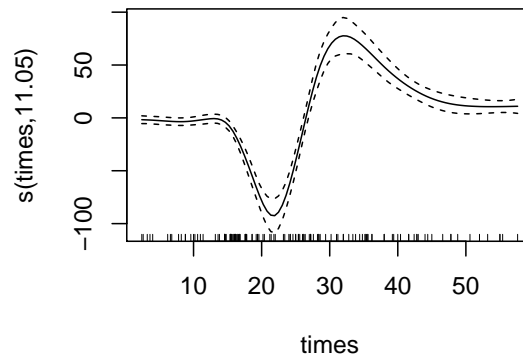
```
R> plot(fitCycle1)
```

Figure 1: Fitted smooth effect of time on quantile 0.9 of the `mcycle` data set, with 95% credible intervals obtained using a Gaussian approximation to the posterior.

which dispatches to the `plot` method for 'gam' objects, and produces the plot shown in Figure 1. To save memory, no component of the Gaussian location-scale fit is stored in the output of `qgam`, but the selected values of the parameters $\log \sigma_0$ and $h(\boldsymbol{x})$ can be extracted by

```
R> fitCycle1$family$getTheta()

      0.9
1.237221
```

```
R> head(fitCycle1$family$getCo(), 5)


[1] 0.2791352 0.2728345 0.2549963 0.2442300 0.2346063
```

Above we fitted a single quantile, but multiple quantiles can be estimated at once using the following function call

```
R> fitCycleM <- mqgam(form = list(accel ~ s(times, k = 20, bs = "ad"),
+     ~ s(times)), data = mcycle, qu = c(0.1, 0.25, 0.5, 0.75, 0.9))
```

where the `mqgam` function fits a QGAM to each of the five quantiles. The main advantage of using `mqgam`, rather than several calls to `qgam`, is that it saves memory by storing a single version of potentially large objects, such as the model matrix, for all the fitted QGAMs. The five fitted QGAMs can be found in the list `fitCycleM$fit`, but calling functions such as `plot(fitCycleM$fit[[1]])` would trigger an error because some object elements are missing. The output of `mqgam` should instead be manipulated using the `qdo` function, for example

```
R> qdo(fitCycleM, qu = 0.25, fun = summary)


Family: elf
Link function: identity


Formula:
```

```
accel ~ s(times, k = 20, bs = "ad")

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -42.739      2.016   -21.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
           edf Ref.df Chi.sq p-value
s(times) 10.55  11.83  837.3  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.735   Deviance explained = 89.2%
-REML = 570.51  Scale est. = 1          n = 133
```

shows the output of the `summary` method for 'gam' objects for the QGAM fitted to quantile $\tau = 0.25$. Similarly, the fitted effect for $\tau = 0.1$ can be plotted by `qdo(fitCycleM, qu = 0.1, fun = plot)`.

More convenient methods for plotting multiple QGAMs jointly are provided by the **mgcViz** package (Fasiolo and Nedellec 2020). In the following we will only use the **mgcViz** tools that are specific to quantile GAMs, and we refer to Fasiolo, Nedellec, Goude, and Wood (2020) and to the relevant package documentation for more details. To exploit the visualizations offered by **mgcViz**, it is necessary to transform the output of `mqgam` as follows

```
R> library("mgcViz")
R> fitCycleM <- getViz(fitCycleM)
```

which produces an object of class `c("mqgamViz", "mgamViz")`. This is simply a list of objects of class 'gamViz', which inherits from the 'qgam' class. Given that the elements of `fitCycleM` are full objects, without any missing component, they can be manipulated directly rather than via `qdo`. For example, we can use `plot(fitCycleM[[1]])` to plot the fitted effect for quantile $\tau = 0.1$. Of course, transforming `fitCycleM` using `getViz` has some memory cost, because `fitCycleM` now stores a copy of the model matrix for each quantile.

Having transformed `fitCycleM`, the effect of time on all quantiles can now be visualized jointly by

```
R> plot(fitCycleM)
```

which produces the plot shown in Figure 2, by dispatching to `plot` methods for 'mgamViz' objects in package **mgcViz**. The latter produces an object of class `c("plotGam", "gg")`, which is a wrapper around one or several objects of class 'ggplot', defined in the **ggplot2** package (Wickham 2016; Wickham *et al.* 2021). Note that Figure 2 shows the fitted effects for each τ which, in this case, are equal to the fitted quantiles up to some vertical shifts (the intercepts). The effects cross each other because they are all centered (see Wood 2017 for details on identifiability constraints in GAMs), hence this does not imply that the corresponding quantile
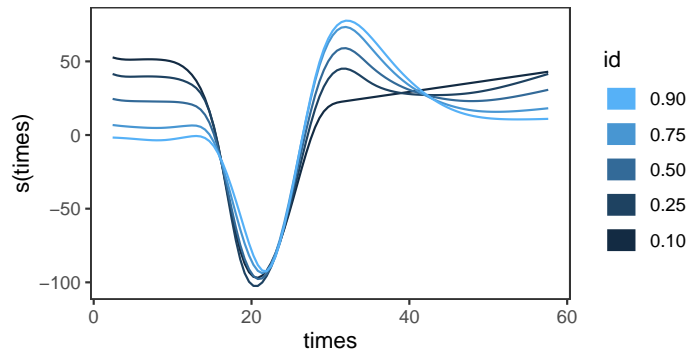
Figure 2: Plot showing the fitted additive effects of time on all five quantiles, obtained using the **mgcViz** package.
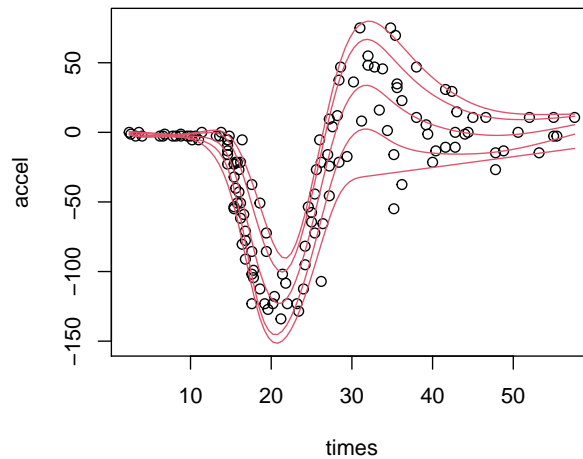

Figure 3: Plot showing the five conditional quantiles fitted using `qgam`.

estimates (which are not centered) cross within the range of the data, time $\in (0, 60)$. In fact, if we plot the estimated quantiles using

```
R> xseq <- with(mcycle, seq(min(times), max(times), length.out = 100))
R> preds <- sapply(fitCycleM, predict, newdata = data.frame(times = xseq))
R> plot(mcycle, ylim = range(preds))
R> for(ii in 1:5) lines(xseq, preds[, ii], col = 2)
```

we do not see any clear crossing in Figure 3. Indeed, the minimal distance between consecutive quantiles is

```
R> min(apply(preds, 1, diff))
```

```
[1] 0.1524413
```

hence there is no crossing within the range of the data. However, quantiles fitted with **qgam** often cross somewhere (outside the data range in this example), because they are fitted independently and without any non-crossing constraint. This issue is not specific to **qgam**,
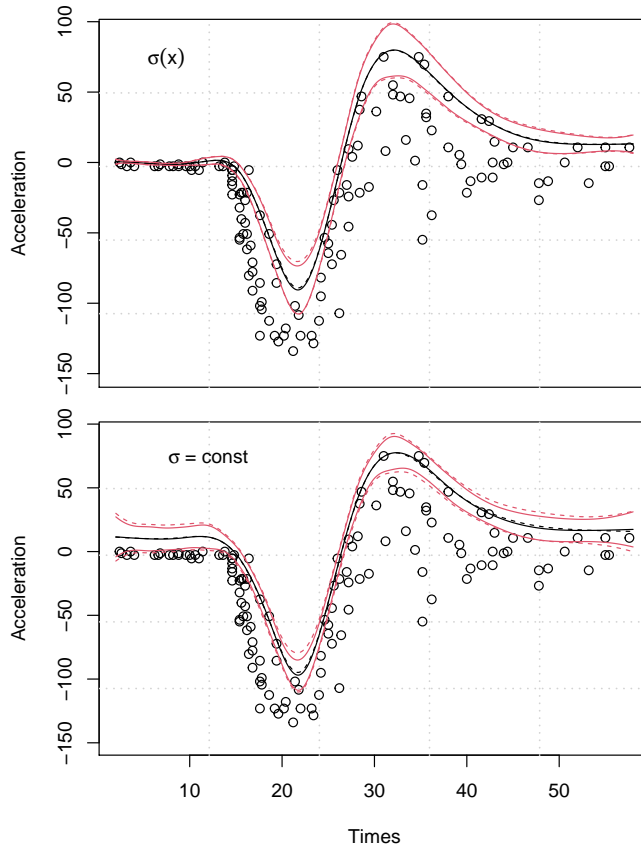
Figure 4: Fitted quantile $\tau = 0.9$ (solid black), with 95% credible intervals (solid red), when the learning rate and the loss smoothness are allowed to vary with the independent variable (top) or not (bottom). The dashed lines correspond to the fitted quantile and credible intervals obtained by using 100 bootstrap samples to estimate $v_s(\boldsymbol{x})$ during IKL-based calibration of σ_0 (see Section 2.3 for details). Bootstrap-based calibration is selected by setting the `control` argument of `qgam` to `list(loss = "cal", K = 100)`.

but is a well known drawback of distribution-free quantile regression (Koenker 2005). For solutions to the quantile crossing problem see, for example, Bondell, Reich, and Wang (2010) or, in a Bayesian framework, Rodrigues and Fan (2017).

In the `mcycle` data set, the variance of the response (`accel`) varies strongly with the independent variable (`times`), hence it is natural to let the learning rate and the ELF loss smoothness vary along the latter. This is achieved by providing a model formula which is a list with two elements, where the second, `~ s(times)`, is used by `qgam` to model the conditional variance $\kappa(\boldsymbol{x})$ in the location-scale GAM. However, it is possible to avoid modeling the variance and to assume that it is constant along the covariates. For example, in the following code

```
R> fitCycleConst <- qgam(accel ~ s(times, k = 20, bs = "ad"), data = mcycle,
+     qu = 0.9)
```

we are providing only the model for the quantile $\tau = 0.9$, rather than a list of two formulas, so `qgam` uses a constant learning rate $1/\sigma(\boldsymbol{x}) = 1/\sigma_0$ and loss bandwidth $\tilde{h}^*(\boldsymbol{x}) = \tilde{h}^*$. For the `mcycle` data set, ignoring the heteroscedasticity leads to the poor results shown in Figure 4.

One problem with the quantile fit shown on the right is that it lies far above all the responses, for time < 10ms. But we are fitting quantile 0.9, hence we should expect around 10% of the responses to lie above the fit. The issue is that the bias of the ELF loss used by `qgam` is inversely proportional to the conditional variance of the response (see Fasiolo *et al.* 2021b for details), hence not modeling the response variance leads to high bias for time < 10ms. Ignoring heteroscedasticity can also lead to credible intervals that have non-constant coverage. In fact, the width of the intervals should be proportional to the variance of the response, but on the right plot in Figure 4 they have nearly constant width, thus providing an incorrect representation of the uncertainty of the fit. Of course, in most data sets, heteroscedasticity is not as dramatic as in `mcycle` and using a location-only model often leads to satisfactory results. However, it is important to be aware of the issues highlighted by this example.

3.2. Functional QGAM modeling of electricity demand data

In this example we consider smart meter data from $n_c = 247$ anonymized residential customers from Sydney, Australia, covering the period between the 3rd of July 2010 and the 30th of June 2011. The data has been downloaded from https://www.ausgrid.com.au/, and it originally contained electricity demand from 300 customers, at 30min resolution. We discarded 53 customers because their demand was too irregular (e.g., they were absent from home for over 30 consecutive days), and we integrated the demand data with temperature data from the National Climatic Data Center (https://www.ncdc.noaa.gov/), covering the same period. The aim of this example is to illustrate how to use quantile GAMs in **qgam**, in the context of an electricity demand forecasting application. We start by loading the following data set:

```
R> data("AUDem", package = "qgam")
R> meanDem <- AUDem$meanDem
R> head(meanDem, 3)

  doy  tod       dem dow     temp                date     dem48
1 184 18.0 0.8248777 Sat 3.357407 2010-07-03 17:30:00 0.8978636
2 184 18.5 0.8686110 Sat 2.517073 2010-07-03 18:00:00 0.9417633
3 184 19.0 0.8519471 Sat 1.898399 2010-07-03 18:30:00 0.9148921
```

which contains the variables:

- `doy`: the day of the year, from 1 to 365;

- `tod`: the time of day, ranging from 18 to 22, where 18 indicates the period from 17:00 to 17:30, 18.5 the period from 17:30 to 18:00 and so on;

- `dem`: the demand (in kWh) during a 30min period, averaged over the n_c households;

- `dow`: factor variable indicating the day of the week;

- `temp`: the external temperature at Sydney airport, in degrees Celsius;

- `date`: local date and time;

- `dem48`: the lagged mean demand, that is the average demand (`dem`) during the same 30min period of the previous day.
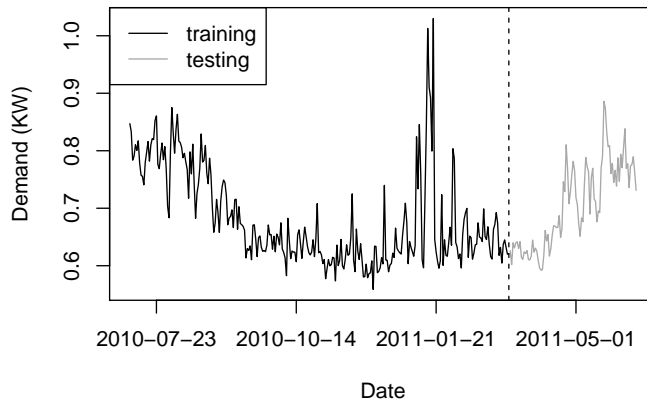
Figure 5:   Electricity demand between 19:30 and 20:00, averaged across the n_c customers.

Assume that we aim at producing a probabilistic forecast of the average demand one day ahead, using additive quantile regression. We start by dividing the data into a learning and a testing set:

```
R> cutDate <- as.Date("2011-04-01 00:00:00")
R> meanDemLearn <- subset(meanDem, as.Date(date) < cutDate)
R> meanDemTest <- subset(meanDem, as.Date(date) >= cutDate)
```

which leaves the last three months for testing. Figure 5 shows the average demand between 19:30 and 20:00, over the whole period. We focus only on the period between 17:30 and 21:30, because the demand dynamics change considerably before and after this time slot, hence modeling demand across the whole day would require a much more sophisticated and computationally expensive model than any of those considered below. We start by modeling the τth quantile of the average demand using the quantile model

$$\mu_\tau(\boldsymbol{x}_t) = \sum_{j=1}^{7} \beta_j \mathbb{I}(\mathrm{dow}_t = j) + \beta_8 \mathrm{dem}_{t-48} + f_1(\mathrm{tod}_t) + f_2(\mathrm{temp}_t) + f_3(\mathrm{doy}_t), \qquad (9)$$

where $\mathbb{I}(\mathrm{dow}_t = j) = 1$ when dow_t is the jth day of the week and zero otherwise, f_1 and f_2 are smooth effects built using thin plate splines, while f_3 is a cyclical effect constructed using cubic regression splines. We use a cyclic effect for dow to ensure that the seasonal effect has the same value on the 31st of December and on the 1st of January. Given that demand variability seems to spike during the austral summer (see Figure 5), we model the variance $\kappa(\boldsymbol{x})$ using a cyclic effect for doy in the location-scale Gaussian GAM.

We fit the model just described to quantiles $\tau = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ using the following code

```
R> library("qgam")
R> qusObj <- seq(0.1, 0.9, length.out = 5)
R> fitQ <- mqgam(list(dem ~ dow + dem48 +
+    s(tod, k = 6) + s(temp) + s(doy, bs = "cc"), ~ s(doy, bs = "cc")),
+    qu = qusObj, data = meanDemLearn)
```

We then plot the estimated smooth effects and the linear coefficients of lagged demand using
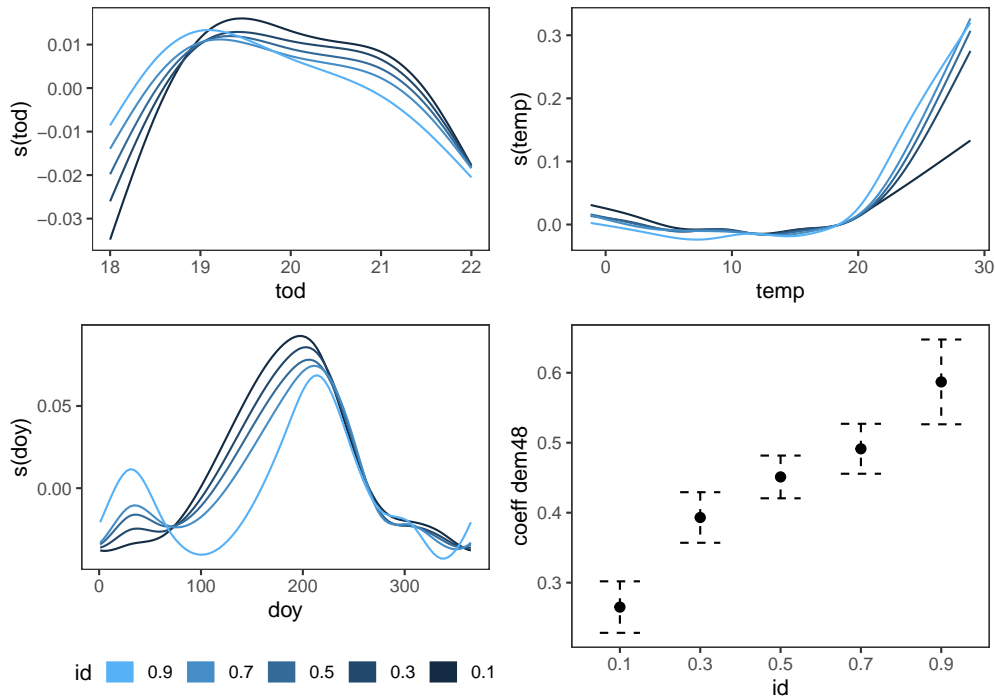
Figure 6: Estimated effects of time of day (`tod`), temperature (`temp`), day of year (`doy`) and lagged demand (`dem48`) from model (9), for the quantiles labelled using the `id` variable.

```
R> library("mgcViz")
R> fitQ <- getViz(fitQ)
R> print(plot(fitQ, allTerms = TRUE, select = c(1:3, 5)), pages = 1)
```

The output, up to some aesthetic manipulation, is shown in Figure 6. Note that we used the plotting tools provided by **mgcViz**, which required us to convert the output of `mqgam` using `getViz`. Then we used `plot` to plot all model terms (`allTerms = TRUE` is used to plot the parametric terms in addition to the smooth effects), we selected a subset of the effects using the `select` argument and we used `print` to arrange all the plots on a single page.

The shapes of the estimated effects seem quite reasonable: the `tod` effect increases as people get home and decreases later in the evening, the temperature plot shows a strong cooling effect and a negligible heating effect, and the effect of `doy` shows a main mode during the austral winter and a lower, narrower, mode during the summer. The discrepancy between the effects estimated on different quantiles can be interpreted in terms of the corresponding effect on the shape of the demand distribution. For example, at `tod` = 18 the effects f_1 are more spread out than at `tod` = 19 implying that, all else being equal, the variance of the demand is higher at `tod` = 18. Similarly, at `doy` ≈ 25 the effect f_3 is more positive for quantile $\tau = 0.9$ than for $\tau = 0.1$, while the positions are reversed at `doy` ≈ 120. Hence, the demand distribution is more skewed to the right in the first case. The linear coefficient of the lagged demand increases with τ; hence, all else being equal, the demand variance one day ahead increases with the lagged demand.

So far we considered only the mean demand over the portfolio of customers, however the original data set contains demand for each of the n_c households. Average demand is typically

much easier to forecast than individual demand, because the highly irregular consumption of individual customers is averaged out. However, the by-customer data should carry more information, hence it is interesting to verify whether it can be used to improve the quantile forecasts for the average demand. We capture part of the information contained in the by-customer demand data by considering a model which, for storage reasons, uses a functional summary of the full demand data, as explained in the following.

We consider the functional quantile model

$$\mu_\tau(\boldsymbol{x}_t) = \sum_{j=1}^{7} \beta_j \mathbb{I}(\text{dow}_t = j) + f_1(\text{tod}_t) + f_2(\text{temp}_t) + f_3(\text{doy}_t) + \int_0^1 f_4(p) G_{t-48}^{-1}(p) dp, \quad (10)$$

where $\mu_\tau(\boldsymbol{x}_t)$ is, as in (9), the τth quantile of the average demand, while we indicate with G_t the CDF of the individual demand across customers at time t and with $G_{t-48}^{-1}(p)$ the corresponding pth quantile, at the same instant of the previous day. Function f_4 is a smooth effect along $p \in [0, 1]$, constructed using a thin plate spline basis. Hence, the last term in (10) is a functional effect which takes as input the whole quantile function of the across-customers demand distribution at time $t - 48$, and integrates it over p using the weight function $f_4(p)$. The aim is to verify whether the shape of the distribution of the individual demand at time $t - 48$ can be used to improve the forecast of the average demand at time t. The regression coefficients of $f_4(p)$ are unknown, but they can be estimated using the methods presented in Section 2.3. In particular, we have that

$$\int_0^1 f_4(p) G_{t-48}^{-1}(p) dp = \int_0^1 \sum_{k=1}^{K_4} b_k^4(p) \beta_k^4 G_{t-48}^{-1}(p) dp = \sum_{k=1}^{K_4} \left\{ \int_0^1 b_k^4(p) G_{t-48}^{-1}(p) dp \right\} \beta_k^4 = \sum_{k=1}^{K_4} \tilde{b}_k^4 \beta_k^4,$$

where the transformed basis $\tilde{b}_1^4, \ldots, \tilde{b}_{K_4}^4$ implicitly depends on time t. Hence, once the transformed basis is obtained, the functional effect is linear in the regression coefficients and estimating them presents no extra difficulty, relative to standard smooth effects. The across-customers demand CDF G_{t-48} is unknown, but we have a sample of size n_c from it, consisting of the by-customer demand data. Hence, we can approximate the transformed basis using

$$\tilde{b}_k^4 = \int_0^1 b_k^4(p) G_{t-48}^{-1}(p) dp \approx \frac{1}{n_q} \sum_{l=1}^{n_q} b_k^4(p_l) \hat{G}_{t-48}^{-1}(p_l), \quad (11)$$

where $\hat{G}_{t-48}^{-1}(p_1), \ldots, \hat{G}_{t-48}^{-1}(p_{n_q})$ is a set of empirical quantiles of the n_c by-customer demand observations at time $t - 48$, corresponding to probability level $0 < p_1 < p_2 < \cdots < p_{n_q} < 1$.

To reduce the amount of data stored in the **qgam** package, we use $n_q = 20 < n_c$ quantiles equally spaced between $p_1 = 0.01$ and $p_{n_q} = 0.99$. The quantiles are stored in

```
R> qDem48 <- AUDem$qDem48
R> head(qDem48[, c(1, 2, 5, 15, 19, 20)], 3)

          1% 6.157895% 21.63158% 73.21053% 93.84211%       99%
[1,] 0.1995373 0.3128564 0.5125927  1.127776  1.655850 2.089360
[2,] 0.1595543 0.3368127 0.5569682  1.206294  1.587732 2.084527
[3,] 0.2091480 0.3571971 0.5652339  1.141480  1.632144 2.044097
```
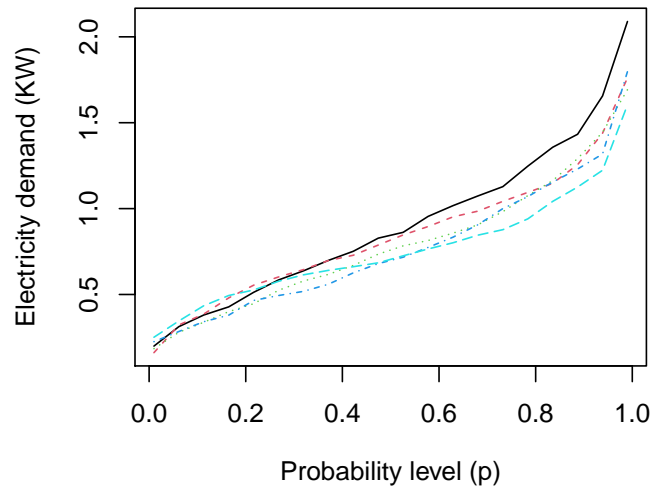
Figure 7: Estimates of the individual demand quantile function G_{t-48}^{-1}, for five values of t.

where each row contains $\hat{G}_{t-48}^{-1}(p_1), \ldots, \hat{G}_{t-48}^{-1}(p_{n_q})$, for some t. Figure 7 shows five selected rows of `qDem48`, each line being an estimate of the quantile function G_{t-48}^{-1}, for a different t. To fit model (10) with **qgam**, we need also a matrix where each row of `probLev` contains the probability levels $p_1, p_2, \ldots, p_{n_q}$. This is computed as follows

```
R> probLev <- matrix(seq(0.01, 0.99, length.out = 20), nrow = nrow(meanDem),
+    ncol = 20, byrow = TRUE)
R> head(probLev[, c(1, 2, 5, 15, 19, 20)], 3)

      [,1]       [,2]      [,3]      [,4]      [,5] [,6]
[1,] 0.01 0.06157895 0.2163158 0.7321053 0.9384211 0.99
[2,] 0.01 0.06157895 0.2163158 0.7321053 0.9384211 0.99
[3,] 0.01 0.06157895 0.2163158 0.7321053 0.9384211 0.99
```

We add the matrices just defined to the learning and testing sets by

```
R> ntrain <- nrow(meanDemLearn)
R> ntest <- nrow(meanDemTest)
R> meanDemLearn$qDem48 <- head(qDem48, ntrain)
R> meanDemTest$qDem48 <- tail(qDem48, ntest)
R> meanDemLearn$probLev <- head(probLev, ntrain)
R> meanDemTest$probLev <- tail(probLev, ntest)
```

We are now ready to fit model (10) using

```
R> fitFunQ <- mqgamV(list(dem ~ dow + s(temp) + s(doy, bs = "cc") +
+    s(tod, k = 6) + s(probLev, by = qDem48), ~ s(doy, bs = "cc")),
+    qu = qusObj, data = meanDemLearn)
```

where we are using the `mgcViz::mqgamV` function, which is simply a shortcut for fitting the model using `mqgam` and then transforming it using `getViz`. Here the functional effect is
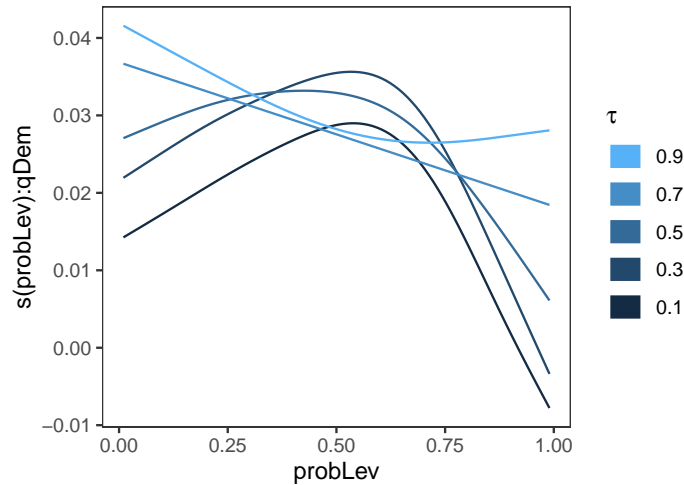
Figure 8: Estimated effect $f_4(p)$ from model (10), for each quantile.

constructed using `s(probLev, by = orDem48)`, which is the constructor typically used for by-variable or varying-coefficient effects in **mgcv**. The corresponding transformed spline basis is constructed by `gam`, which performs the summation in (11) using the rows of the `probLev` and `orDem48` matrices.

We can have a look at the fitted functional smooth effects $f_4(p)$ using

```
R> plot(fitFunQ, select = 4) + labs(color = expression(tau))
```

which produces the plot shown in Figure 8. It is interesting to note that the effects corresponding to low and high probabilities diverge as $p \to 0$ and $p \to 1$, suggesting that the extremes of the quantile function $G_{t-48}^{-1}(p)$ have a stronger linear effect on the high quantiles of the average demand (e.g., $\mu_\tau(\boldsymbol{x})$ with $\tau = 0.9$) than on the low ones. However, the `plot` method for 'mgamViz' objects does not include credible intervals when plotting the same effect for multiple quantiles, to avoid cluttering the plot. In Figure 9 we compare the effects for $\tau = 0.1$ and 0.9, and we include 95% credible intervals. The plot shows that the discrepancy between the effects is statistically stronger for $p \approx 1$, than for $p \approx 0$. Hence, low quantiles of the average demand are less sensitive than high quantiles to changes in the right tail of the distribution of the individual demand. This asymmetric behavior highlights the usefulness of modeling the effect of the changes in the individual demand distribution via functional effects, which take into account the entire shape of the distribution, rather than via a set of manually built summary statistics (e.g., mean and standard deviation) which will inevitably miss some of its features. To check whether model (10) leads to a better fit than model (9), we can compare the values of the Bayesian information criterion (BIC) corresponding to each quantile by

```
R> round(rbind(QGAM = sapply(fitQ, BIC), FuncQGAM = sapply(fitFunQ, BIC)))


            0.1   0.3   0.5   0.7   0.9
QGAM      -8255 -9513 -9700 -9137 -6651
FuncQGAM  -8458 -9617 -9698 -9152 -6680
```
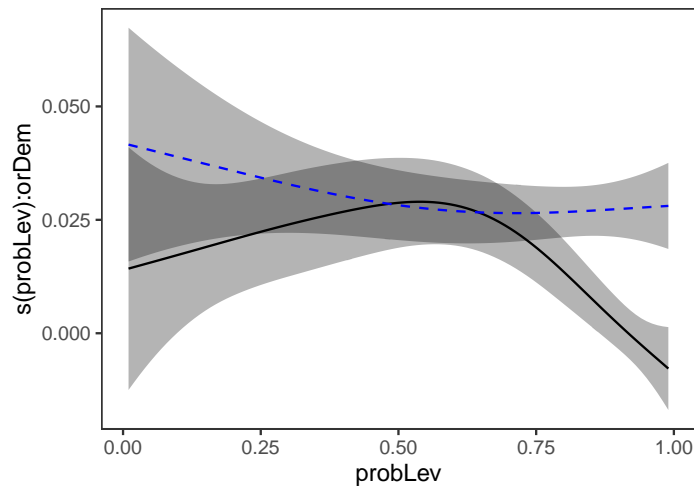
Figure 9: Estimated effect $f_4(p)$ from model (10) with 95% credible intervals, for quantile $\tau = 0.1$ (black solid) and 0.9 (blue dashed).

BIC favors slightly model (10), but the simpler model is preferred for one of the quantiles. Instead the Akaike information criterion (AIC), which penalizes model complexity less strongly than BIC, favors model (10) for all quantiles

```
R> round(rbind(QGAM = sapply(fitQ, AIC), FuncQGAM = sapply(fitFunQ, AIC)))
```

```
           0.1   0.3   0.5   0.7   0.9
QGAM     -8416 -9679 -9867 -9303 -6813
FuncQGAM -8637 -9800 -9881 -9325 -6858
```

We can also compare the predictive performance of the two models on the testing set, which comprises the last three months of the data set (1st of April to 30th of June, see Figure 5). For example, we can calculate the pinball loss for each model and quantile on such data as follows

```
R> predQ <- sapply(fitQ, predict, newdata = meanDemTest)
R> predFunQ <- sapply(fitFunQ, predict, newdata = meanDemTest)
R> round(rbind(QGAM = pinLoss(meanDemTest$dem, predQ, qusObj),
+     FuncQGAM = pinLoss(meanDemTest$dem, predFunQ, qusObj)), 1)
```

```
         0.1 0.3  0.5  0.7  0.9
QGAM     4.7 8.8 10.4 10.6 10.7
FuncQGAM 4.6 8.6 10.3 10.2 10.2
```

where `qgam::pinLoss` evaluates the pinball loss. The functional model does better on the testing set, but it is important to point out that the size of the testing set is rather limited, especially for the purpose of evaluating the performance of the extreme quantiles estimates.

Of course, model (10) could be improved further. For example, a call to

```
R> check(fitFunQ[[3]])

Theor. proportion of neg. resid.: 0.5    Actual proportion: 0.5155229
Integrated absolute bias |F(mu) - F(mu0)| = 0.02719397
Method: REML    Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-3.420505e-08,4.411812e-07]
(score -4857.326 & scale 1).
Hessian positive definite, eigenvalue range [0.5027064,3.678087].
Model rank =  39 / 39

Basis dimension (k) check: if edf is close to k' (maximum possible edf)
it might be worth increasing k.

                   k'  edf
s(temp)             9 8.21
s(doy)              8 7.65
s(tod)              5 4.69
s(probLev):qDem48  10 3.15
```

shows that, for $\tau = 0.5$, the effective degrees of freedom (`edf`) are quite close to their maximum possible value (`k'`) for the temperature effect. This means that the fit is using all the degrees of freedom available for that effect, hence we might want to consider increasing the number of basis functions by using `s(temp, k = 20)` in the model formula. However, one might be reluctant to increase the number of basis functions used to fit more extreme quantiles (e.g., $\tau = 0.9$), where the data is sparser. This highlights a disadvantage of using `mqgam`, rather than repeated calls to `qgam`, when fitting several QGAMs: the same model formula is used for all the quantiles. The model could also be improved by, for example, including the effect of lagged temperatures, often useful for capturing thermal inertia in buildings.

4. Conclusions

The **qgam** R package provides tools for building, fitting and checking quantile additive models. The main advantage of such models, relative to standard probabilistic GAMs, is that they do not make any parametric assumption on the distribution of the response variable. This is achieved by modeling and estimating the conditional quantiles directly, using the pinball loss. Model fitting is based on the fast calibrated Bayesian framework of Fasiolo *et al.* (2021b), which provides calibrated credible intervals, via IKL selection of the learning rate, and computational efficiency, by smoothing the pinball loss and by exploiting the marginal likelihood methods of Wood *et al.* (2016). Model building is handled by **mgcv**, hence the quantile GAMs built using **qgam** can include any of the effect types made available by **mgcv**, among others: cyclic, adaptive and multivariate tensor-product smooths, as well as functional and random effects.

At the time of writing, **qgam** version 1.3.4 can handle data sets and models of moderate size. In particular, the current fitting methods require explicit formation of the $n \times d$ model matrix **X**, which leads to high memory usage for large n. Wood, Goude, and Shaw (2015) proposed a

GAM fitting framework where explicit formation of $\mathbf{X}$ is avoided by adopting block-oriented iterations for smoothing parameters selection and regression coefficients estimation. The `bam` function in **mgcv** implements such methods, as well as the marginal discretization approach of Wood, Li, Shaddick, and Augustin (2017), which allows it to fit models with up to 10^4 coefficients and 10^8 data points. Future work will aim at improving the scalability of the Bayesian QGAM fitting framework currently implemented in **qgam**, by exploiting the block-oriented discretized fitting methods provided by **mgcv** for the inner and intermediate iterations, and developing a similarly scalable version of the IKL minimization routine currently used for learning rate selection.

Acknowledgments

References

Bissiri PG, Holmes CC, Walker SG (2016). "A General Framework for Updating Belief Distributions." *Journal of the Royal Statistical Society B*, **78**(5), 1103–1130. `doi:10.1111/rssb.12158`.

Bondell HD, Reich BJ, Wang H (2010). "Noncrossing Quantile Regression Curve Estimation." *Biometrika*, **97**(4), 825–838. `doi:10.1093/biomet/asq048`.

Brent RP (2013). *Algorithms for Minimization without Derivatives*. Courier Corporation.

Brezger A, Kneib T, Lang S (2005). "**BayesX**: Analyzing Bayesian Structured Additive Regression Models." *Journal of Statistical Software*, **14**(11), 1–22. `doi:10.18637/jss.v014.i11`.

Eilers PHC, Marx BD (1996). "Flexible Smoothing with *B*-Splines and Penalties." *Statistical Science*, **11**(2), 89–102. `doi:10.1214/ss/1038425655`.

Fasiolo M, Nedellec R (2020). *mgcViz: Visualisations for Generalized Additive Models*. R package version 0.1.6, URL `https://CRAN.R-project.org/package=mgcViz`.

Fasiolo M, Nedellec R, Goude Y, Wood SN (2020). "Scalable Visualization Methods for Modern Generalized Additive Models." *Journal of Computational and Graphical Statistics*, **29**(1), 78–86. `doi:10.1080/10618600.2019.1629942`.

Fasiolo M, Wood SN, Zaffran M, Goude Y, Nedellec R (2021a). *qgam: Smooth Additive Quantile Regression Models*. R package version 1.3.4, URL `https://CRAN.R-project.org/package=qgam`.

Fasiolo M, Wood SN, Zaffran M, Nedellec R, Goude Y (2021b). "Fast Calibrated Additive Quantile Regression." *Journal of the Americal Statistical Association*, **116**(535), 1402–1412. `doi:10.1080/01621459.2020.1725521`.

Hastie T, Tibshirani R (1986). "Generalized Additive Models." *Statistical Science*, **1**(3), 297–318. doi:10.1214/ss/1177013604.

Hothorn T, Bühlmann P, Kneib T, Schmid M, Hofner B (2021). **mboost**: *Model-Based Boosting*. R package version 2.9-5, URL https://CRAN.R-project.org/package=mboost.

Jones MC (2008). "On a Class of Distributions with Simple Exponential Tails." *Statistica Sinica*, **18**(3), 1101–1110.

Jones MC, Pewsey A (2009). "Sinh-Arcsinh Distributions." *Biometrika*, **96**(4), 761–780. doi:10.1093/biomet/asp053.

Koenker R (2005). *Quantile Regression*. Cambridge University Press, Cambridge.

Koenker R (2011). "Additive Models for Quantile Regression: Model Selection and Confidence Bandaids." *Brazilian Journal of Probability and Statistics*, **25**(3), 239–262. doi:10.1214/10-bjps131.

Koenker R (2021). **quantreg**: *Quantile Regression*. R package version 5.85, URL https://CRAN.R-project.org/package=quantreg.

Koenker R, Bassett G (1978). "Regression Quantiles." *Econometrica*, **46**(1), 33–50. doi:10.2307/1913643.

Lin CY, Bondell H, Zhang HH, Zou H (2013). "Variable Selection for Non-Parametric Quantile Regression via Smoothing Spline Analysis of Variance." *Stat*, **2**(1), 255–268. doi:10.1002/sta4.33.

R Core Team (2021). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rigby RA, Stasinopoulos DM (2005). "Generalized Additive Models for Location, Scale and Shape." *Journal of the Royal Statistical Society C*, **54**(3), 507–554. doi:10.1111/j.1467-9876.2005.00510.x.

Ripley B (2021). **MASS**: *Support Functions and Datasets for Venables and Ripley's MASS*. R package version 7.3-54, URL https://CRAN.R-project.org/package=MASS.

Rodrigues T, Fan Y (2017). "Regression Adjustment for Noncrossing Bayesian Quantile Regression." *Journal of Computational and Graphical Statistics*, **26**(2), 275–284. doi:10.1080/10618600.2016.1172016.

Silverman BW (1985). "Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting." *Journal of the Royal Statistical Society B*, **47**(1), 1–52. doi:10.1111/j.2517-6161.1985.tb01327.x.

Sriram K (2015). "A Sandwich Likelihood Correction for Bayesian Quantile Regression Based on the Misspecified Asymmetric Laplace Density." *Statistics & Probability Letters*, **107**, 18–26. doi:10.1016/j.spl.2015.07.035.

Syring N, Martin R (2015). "Scaling the Gibbs Posterior Credible Regions." arXiv:1509.00922 [stat.ME], URL https://arxiv.org/abs/1509.00922.

Umlauf N, Klein N, Simon T, Zeileis A (2021). "**bamlss**: A Lego Toolbox for Flexible Bayesian Regression (and Beyond)." *Journal of Statistical Software*, **100**(4). `doi:10.18637/jss.v100.i04`. 1–53.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York. `doi:10.1007/978-0-387-21706-2`. URL `https://www.stats.ox.ac.uk/pub/MASS4/`.

Waldmann E, Kneib T, Yue YR, Lang S, Flexeder C (2013). "Bayesian Semiparametric Additive Quantile Regression." *Statistical Modelling*, **13**(3), 223–252. `doi:10.1177/1471082x13480650`.

Wickham H (2016). **ggplot2**: *Elegant Graphics for Data Analysis*. Springer-Verlag, New York. `doi:10.1007/978-3-319-24277-4`. URL `https://ggplot2.tidyverse.org/`.

Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K, Yutani H, Dunnington D (2021). **ggplot2**: *Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.3.5, URL `https://CRAN.R-project.org/package=ggplot2`.

Wood SN (2017). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton.

Wood SN (2021). **mgcv**: *Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. R package version 1.8-36, URL `https://CRAN.R-project.org/package=mgcv`.

Wood SN, Goude Y, Shaw S (2015). "Generalized Additive Models for Large Data Sets." *Journal of the Royal Statistical Society C*, **64**(1), 139–155. `doi:10.1111/rssc.12068`.

Wood SN, Li Z, Shaddick G, Augustin NH (2017). "Generalized Additive Models for Gigadata: Modeling the UK Black Smoke Network Daily Data." *Journal of the American Statistical Association*, **112**(519), 1199–1210. `doi:10.1080/01621459.2016.1195744`.

Wood SN, Pya N, Säfken B (2016). "Smoothing Parameter and Model Selection for General Smooth Models." *Journal of the American Statistical Association*, **111**(516), 1548–1575. `doi:10.1080/01621459.2016.1180986`.

Yang Y, Wang HJ, He X (2016). "Posterior Inference in Bayesian Quantile Regression with Asymmetric Laplace Likelihood." *International Statistical Review*, **84**(3), 327–344. `doi:10.1111/insr.12114`.

Yee TW (2010). "The **VGAM** Package for Categorical Data Analysis." *Journal of Statistical Software*, **32**(10), 1–34. `doi:10.18637/jss.v032.i10`.

A. Choosing the loss bandwidth manually

The qgam function determines the ELF loss bandwidth $h(\boldsymbol{x})$ automatically, but this parameter can be chosen manually by specifying the err argument, which is set to NULL by default. Here we provide some guidelines for users who wish to choose this parameter manually. In particular, the examples in Section A.1 illustrate how parameter err affects the statistical and computational performance of the fitting methods provided by **qgam**.

A.1. Illustrating the effect of the loss bandwidth and some diagnostics

To clarify the interpretation of parameter err, let us consider a univariate context where we want to estimate the τth quantile of response y and there are no covariates. Define $\epsilon = |F(\mu^*) - F(\mu_0)|$, where F is the CDF of y, μ_0 is the true quantile at probability level τ and μ^* is the minimizer of $\mathbb{E}\{\tilde{\rho}_\tau(y - \mu)\}$. Given that μ_0 is the minimizer of the expected pinball loss (see (1)), ϵ is the asymptotic absolute bias induced by the use of the smooth ELF loss, with bandwidth $h = \lambda\sigma$. In Appendix A.2 we prove that

$$\epsilon = |F(\mu^*) - F(\mu_0)| \leq 2\log 2\, h \sup_y f(y) = \frac{2\log 2}{\sqrt{2\pi\kappa}}\, h,$$

where f is the PDF of y and the last equality holds if y is Gaussian with variance κ. Via argument err, the qgam function allows users to set the maximum tolerable bias ϵ, which is then used to determine the loss bandwidth by

$$h = \epsilon \frac{\sqrt{2\pi\kappa}}{2\log 2}. \tag{12}$$

Here κ is estimated via the location-scale GAM, and can vary with $\boldsymbol{x}$ as usual. Parameter $h = h(\boldsymbol{x})$, obtained via (12), is then used to determine λ and $\sigma(\boldsymbol{x})$ as in Section 2.3. Of course ϵ (err in the software) is an approximate bound, because the response is not normally distributed in practice.

In **qgam** we let users choose ϵ, rather than λ, because it does not depend on the scale of the response y. Indeed, considering the interpretation of ϵ, reasonable values of this parameter fall in $(0, 1)$ (but qgam checks only that err is positive). To illustrate how err influences the accuracy of the quantile estimates and the computational performance of qgam, we consider a simple data set simulated as follows

```
R> set.seed(5523)
R> x <- seq(-3, 3, length.out = 1e3)
R> X <- data.frame("x" = x, "y" = x + x^2 + rgamma(1e3, 4, 1))
```

Hence $y_i = x_i + x_i^2 + z_i$, where $z_i \sim \text{Gamma}(4, 1)$. Assuming that we want to estimate quantiles 0.05, 0.5 and 0.95, the following code estimates them using different values of err

```
R> fitGrid <- lapply(c(0.01, 0.05, 0.1, 0.3, 0.5),
+    function(.errVal) {
+    mqgam(y ~ s(x), data = X, qu = c(0.05, 0.5, 0.95), err = .errVal)
+ })
```
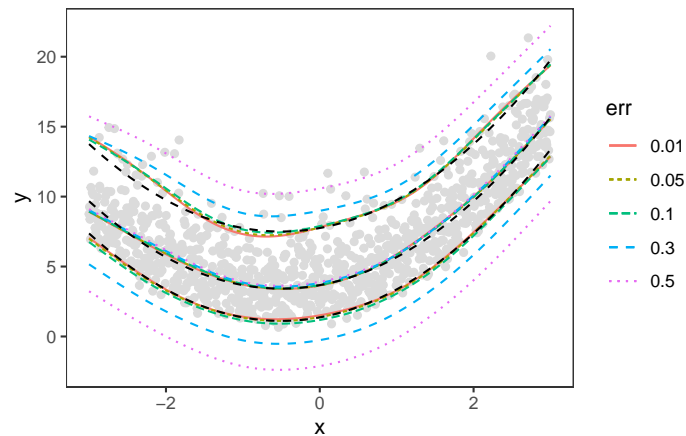
Figure 10: Quantile fits for $\tau = 0.05, 0.5$ and 0.95 for different values of `err`. The dashed lines are the true quantiles.

The resulting quantile fits are shown in Figure 10. It is clear that the bias is positive for $\tau = 0.95$, negative for $\tau = 0.05$ and that its absolute value increases with `err`, becoming clearly visible for `err` > 0.1. However, increasing `err` affects the median estimate only slightly. Heuristically, this is because the bias induced by the smoothed loss is greater when the response distribution is more asymmetric around the quantile of interest.

One might conclude that the loss smoothness should be kept as small as possible, but this leads to computational and numerical stability issues. In fact, decreasing `err` tends to slow down computation, and in this example using a very small tolerance increases the computing time dramatically.

```
R> system.time(fitBigErr <- qgam(y ~ s(x), data = X, qu = 0.95,
+    err = 0.05))[[3]]

Estimating learning rate. Each dot corresponds to a loss evaluation.
qu = 0.95.............done
[1] 0.48


R> system.time(fitSmallErr <- qgam(y ~ s(x), data = X, qu = 0.95,
+    err = 0.001))[[3]]

Estimating learning rate. Each dot corresponds to a loss evaluation.
qu = 0.95...................done
[1] 41.455
```

More importantly, using a very low value of `err` can lead to numerical problems. In fact, the following code

```
R> check(fitSmallErr$calibr, sel = 2)
```

produces the plot in Figure 11, which shows how the estimated IKL loss minimized by the outer iteration looks like. The dots indicate the points at which the loss has been evaluated
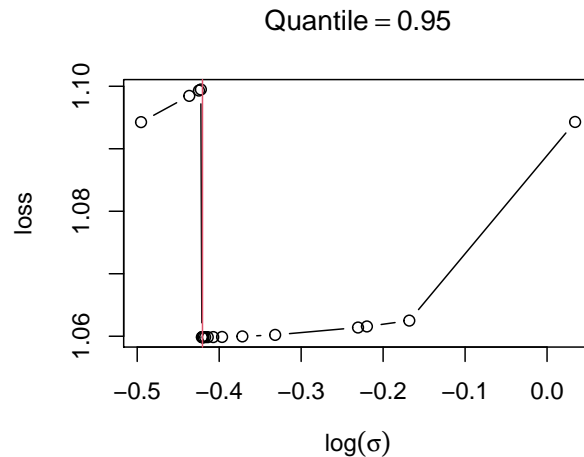
Quantile = 0.95

Figure 11: Plot produced by the `check` method for 'learnFast' objects, showing the estimated IKL loss at several values of $\log \sigma_0$, when `err = 0.001`. The vertical jump in the loss indicates numerical problems.

during the outer Brent optimization, and the vertical red line indicates the value of $\log \sigma_0$ which minimizes the loss. Here the loss seems to be discontinuous, which is due to numerical instabilities. In fact, the analogous plot for `fitBigErr` (not shown) looks smooth and convex. To produce the diagnostic plot in Figure 11 we used the generic `qgam::check` function, which dispatched to the `check` method for 'learnFast' objects. This is because the output of `qgam` contains the results of a call to `tuneLearnFast` in the `$calibr` component. The call to `check` generates two plots and here we are showing only the second one by choosing `sel = 2`.

We can use `check` to have also an estimate of the bias attributable to the smoothed loss and information regarding the convergence of the smoothing parameter estimation routine. This is achieved by

```
R> check(fitBigErr)
```

```
Theor. proportion of neg. resid.: 0.95   Actual proportion: 0.952
Integrated absolute bias |F(mu) - F(mu0)| = 0.004319735
Method: REML   Optimizer: outer newton
full convergence after 5 iterations.
Gradient range [2.628193e-05,2.628193e-05]
(score 3187.909 & scale 1).
Hessian positive definite, eigenvalue range [1.874667,1.874667].
Model rank =  10 / 10

Basis dimension (k) check: if edf is close to k' (maximum possible edf)
it might be worth increasing k.

     k'  edf
s(x)  9 5.27
```

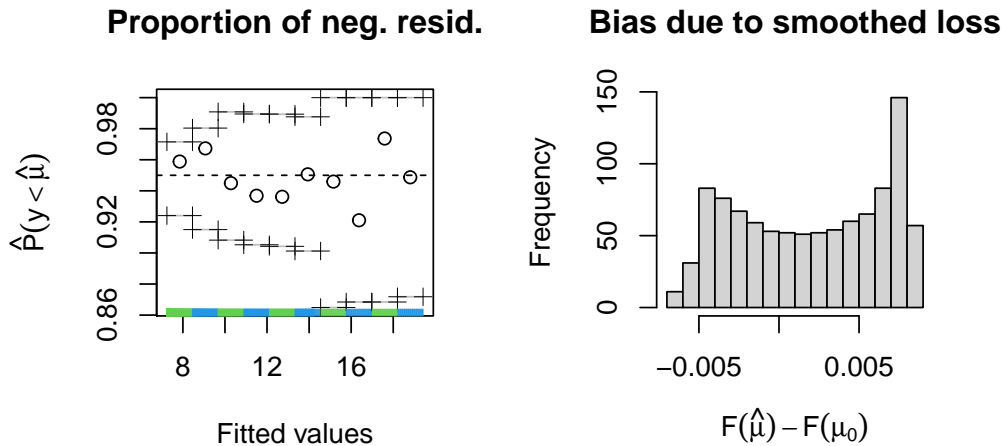which produces the preceding text output and the diagnostic plots shown in Figure 12. The

Figure 12: Diagnostic plots produced by the `check` method for 'qgam' objects. Left: observed proportion of negative residuals in ten bins, with 95% reference intervals. Right: histogram of estimated bias attributable to using a smooth loss.

text states that the intermediate optimization for selecting γ achieved full convergence in few iterations. It reports the range of the entries of the LAML (6) gradient w.r.t. γ at convergence, which we expect to be small, and confirms that the Hessian is positive definite, so that LAML is (at least locally) convex. The text also states that the model has full rank, hence all the regression coefficients β are identifiable. The first two lines of the printed text are concerned with the bias due to the use of the smooth ELF loss. The first line says that nearly 95% of the observations fall below the fitted quantile, which is the percentage we expect for $\tau = 0.95$. The second line reports an estimate of the absolute bias $|F\{\mu^*(\boldsymbol{x}_i)\} - F\{\mu_0(\boldsymbol{x}_i)\}|$, averaged over the observed $\boldsymbol{x}_i$'s. This is much lower than the bound `err = 0.05` used to fit this model. Figure 12 gives more details, with the left plot showing the proportion of negative residuals in a sequence of bins, ordered according to increasing value of $\mu_\tau(\boldsymbol{x})$. The crosses define 95% reference intervals for the proportions, obtained using binomial quantiles. The plot on the right is a histogram of all the estimated values of the bias $|F\{\mu^*(\boldsymbol{x}_i)\} - F(\boldsymbol{x}_i)|$, and it shows that the bias is low for all observations. Hence, the text and visual output of the `check` method for 'qgam' objects confirm that the LAML optimization achieved full convergence and that we should not be too concerned about the bias. The last few lines of the text output indicate that the effective degrees of freedom (`edf`) used to model the effect of `x` are well below the degrees of freedom available for this smooth effect (`k'`). This suggests that the number of basis functions used to model the effect of `x` was sufficiently large, while `edf` $\approx$ `k'` would indicate that the fit is using all the available degrees of freedom, and that we might want to increase the number of basis functions used.

The results presented in this section and our practical experience on loss smoothness selection and convergence checking with **qgam** can be summarized in the following set of suggestions:

- the automatic procedure for selecting the loss smoothness generally offers a good compromise between statistical bias, variance and numerical stability;

- the old default (used in versions of the **qgam** package lower than 1.3.0) was `err = 0.05`, which generally does not lead to unacceptably high levels of bias;

- if the calibration loss plotted by `check(fittedQGAM$learn)` is irregular, or the text printed by `check(fittedQGAM)` does not confirm that `full convergence` was achieved, try to increase `err`;

- if you have to increase `err` to 0.2 or higher to avoid convergence issues, then there might be something wrong with your model (for example, it is missing an important effect);

- the bias attributable to the adoption of a smoothed ELF loss can be estimated using `check(fittedQGAM)`;

- you might get messages saying that `outer Newton did not converge fully` during estimation. These are generated by `mgcv::gam` during LAML maximization, and should not be problematic as long as the calibration loss is smooth (which you can check using `check(fittedQGAM$calibr)`) and `check(fittedQGAM)` states that `full convergence` was achieved;

- in preliminary studies, that is when you are exploring different model structures and you are not yet interested in getting the most accurate estimates, do not decrease `err` too much as it considerably slows down computation;

- setting `err` too low is generally not a good idea. In fact, `err` is an approximate upper bound on the bias, the latter being generally much lower than this parameter suggests, and it is arguably better to have a small amount of bias than numerical problems. As stated above, the default loss smoothness selection procedure, used by default in `qgam` and `mqgam`, typically provides stable fits and a near optimal bias-variance tradeoff.

A.2. Derivation of the asymptotic bias

To simplify the notation, we indicate $\tilde{p}_\tau(y - \mu)$ with $\tilde{p}_\tau(y)$. We start from

$$
F(\mu^*) - F(\mu_0) = \int \mathbb{1}(y \leq \mu^*)f(y)dy - \tau = \int \left\{ \left.\frac{\partial \log \tilde{p}_\tau(y)}{\partial \mu}\right|_{\mu=\mu^*} - \left.\frac{\partial \rho_\tau(y)}{\partial \mu}\right|_{\mu=\mu^*} \right\} f(y)\, dy,
$$

where $f(y)$ is the PDF of y and we used the fact that $\int \partial \log \tilde{p}_\tau(y)/\partial \mu|_{\mu=\mu^*} f(y)dy = 0$, by definition of μ^*. We proceed to bound the right hand side from above. For any μ, simple manipulations lead to

$$
\int \left\{ \frac{\partial \log \tilde{p}_\tau(y)}{\partial \mu} - \frac{\partial \rho_\tau(y)}{\partial \mu} \right\} f(y)\, dy = \int \left\{ \Phi(y|\mu, \lambda\sigma) - \mathbb{1}(y > \mu) \right\} f(y)\, dy, \tag{13}
$$

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

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

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

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

Note that the right hand side of (13) makes it clear that, if $f(y)$ is symmetric around μ^*, then $|F(\mu^*) - F(\mu_0)| = 0$ and there is no bias.

Affiliation:

Matteo Fasiolo
School of Mathematics
University of Bristol
University Walk
BS8 1TW Bristol, United Kingdom
E-mail: matteo.fasiolo@gmail.com
URL: https://research-information.bristol.ac.uk/en/persons/matteo-fasiolo