Fractile Boosting: a novel approach to mode regression

Fabian Sobotka, Andreas Mayr, Thomas Kneib

1 Chair of Statistics, Georg August University Goettingen, GERMANY
2 IMBE, Friedrich Alexander University Erlangen-Nuremberg, GERMANY

E-mail for correspondence: fabian.sobotka@wiwi.uni-goettingen.de

Abstract: Mode regression would be very interesting in practice, if the estimation methods were easy to apply and could include semiparametric models. In the framework of boosting we are able to generalise quantile and expectile regression to fractiles, which minimise a loss function derived from error terms taken to flexible powers. This regression problem can also be modeled as a generalised model for location, scale and shape using an auxiliary likelihood. We combine both approaches to perform an approximation to and generalisation of mode regression, by inserting a low power to the losses and asymmetric weights. This approach is able to handle bimodal, highly skewed or truncated distributions and semiparametric models.

Keywords: mode regression; asymmetric loss regression; semiparametric models; componentwise functional gradient descent boosting; GAMLSS.

1 Introduction

Quantile and expectile regression (see Koenker, 2005, or Sobotka and Kneib, 2010, for example) are becoming increasingly popular. They provide easy ways to estimate more than the conditional location of a response variable. They also do so without many assumptions to the distribution of the aforementioned response. While expectiles allow for more flexible, semiparametric models, fast, easy and efficient estimation methods are nowadays available for both quantiles and expectiles. They aim on solving the following estimation problem

$$\hat{\beta}_\tau = \arg\min_{\beta} \sum_{i=1}^{n} w_i(\tau) |y_i - x_i'\beta|^k$$

with

$$w_i(\tau) = \begin{cases} \tau & \text{if } y_i > x_i'\beta \\ 1 - \tau & \text{if } y_i \leq x_i'\beta \end{cases}$$

for a response $y_i$, a covariate or design vector $x_i$, asymmetry $\tau \in (0,1)$ and power $k > 0$. The estimate is computed for a power of $k = 1$ using linear
programming techniques and provides regression quantiles, while \( k = 2 \) allows for a direct calculation of the estimate, resulting in expectiles. Setting \( \tau = 0.5 \) reduces the problem to classical median and mean regression, respectively.

In theory, a power of \( k = 0 \) would result in mode regression and the inclusion of weights \( w_i(\tau) \) in its generalisation. However, the practical execution is futile for finite samples with metric responses. Instead, mode regression was introduced as a form of kernel regression (Lee, 1989) and generalised to additive and nonlinear models without a penalty term by Kemp and Santos Silva (2010), but only for unimodal and to some extent only for symmetric distributions. In addition, semiparametric models with smooth spatial or random effects cannot be included here.

In order to offer a solution to this scenario, we propose the introduction of \( k \)-fractiles as solutions of the minimisation problem (1). Hence, quantiles and expectiles are also possible fractiles. However, for \( k \downarrow 0 \), fractiles can approximate and generalise mode regression while incorporating semiparametric models and smoothing. Due to the distribution-free formulation of the estimate, skewed and bimodal distributions for the response are also allowed.

Since a direct search of the minimum (1) might in general be a hard task, we offer two estimation algorithms based on boosting, a flexible framework for regression models and also provide some first examples.

2 Boosting

Component-wise gradient boosting as introduced by Bühlmann and Hothorn (2007) is a very flexible framework that allows for penalised semiparametric modelling. Smooth, spatial and random effects, e.g., are included by simple (least squares-like) base-learning procedures \( f_1(\cdot), \ldots, f_p(\cdot) \) independent from the applied loss function. It divides the minimisation problem in a fixed, large number of small steps where only the model part with the steepest gradient is included. Hence, automatic variable selection takes place during the estimation. The following estimation algorithm can be executed with the R-package \texttt{mboost} (Hothorn et al., 2013) for a pre-fixed number of iterations \( m_{\text{stop}} \).

1. **Initialization:** \( m = 0 \). Initialize the additive predictor \( \eta_i^{[m]} = 0 \) for \( i = 1, \ldots, n \). Specify a set of base-learners \( f_1(\cdot), \ldots, f_p(\cdot) \), one for each covariate.

2. **Negative gradient:** \( m = m + 1 \). Compute the negative gradient vector \( u^{[m]} \):

\[
    u^{[m]}_i = \begin{cases} 
        k \cdot \tau \cdot |y_i - \eta_i^{[m-1]}(k-1)| & (y_i - \eta_i^{[m-1]}(k-1)) \geq 0 \\
        k \cdot (1 - \tau) \cdot |y_i - \eta_i^{[m-1]}(k-1)| & (y_i - \eta_i^{[m-1]}(k-1)) < 0. \notag
    \end{cases}
\]
3. **Component-wise estimation:** Use the base-learners to fit the negative gradient vector $u^{[m]}$ to every possible covariate $x_1, \ldots, x_p$ separately

$$u^{[m]} = \text{base-leaner} \rightarrow \hat{f}_j(x_j) \quad \text{for} \quad j = 1, \ldots, p$$

4. **Update one component:** Select the component $j^*$ that best fits the negative gradient vector and update the additive predictor:

$$\hat{\eta}^{[m]} = \hat{\eta}^{[m-1]} + sl \cdot f_{j^*}(x_j),$$

where $sl$ is a small step-length ($0 < sl \ll 1$). Therefore, only the best-performing base-learner (and hence the best-performing covariate) contributes to the update.

5. **Iteration:** Iterate steps 2 to 4 until $m = m_{\text{stop}}$.

As penalised least squares regression is highly dependent on the optimal choice of the smoothing parameters, in this algorithm the optimal stopping iteration is also computed via cross-validation. This allows for the regulation of smoothness in our estimates.

The only requirement is the existence of a gradient, which is available for almost all $k > 0$. So, boosting theoretically allows for the fit of semiparametric models to fractiles, and especially also for $k \ll 1$.

3 **GAMLSS**

An alternative approach can be constructed by the combination of generalised additive models for location scale and shape (Rigby and Stasinopoulos, 2005) with boosting by Mayr et al. (2012). Instead of solving the minimisation problem (1) we choose a skew exponential power distribution

$$f_{\mu, \sigma, w, k}(y) = \frac{c}{\sigma} \left\{ I(y < \mu) \exp \left[ -\frac{1}{2} \left| \frac{y - \mu}{\sigma} \right|^k \right] + I(y \geq \mu) \exp \left[ -\frac{1}{2} \left| \frac{1}{w} \left( \frac{y - \mu}{\sigma} \right)^k \right| \right] \right\}$$

with $c = \frac{wk}{(1+w^2)^{1/2} \Gamma(1/k)}$ and $w = \left( \frac{1-\tau}{\tau} \right)^{1/k}$ as auxiliary likelihood. In the R-package `gamboostLSS` (Hofner et al., 2011) the boosting algorithm from the previous section is adapted to maximise such likelihoods for multiple parameters. While we fix $w$ and $k$ in order to rewrite our loss function (1), the algorithm iterates between a fit for $\mu$ and $\sigma$. Hence, we get a good fit of the distribution to our data and especially an estimate for the location $\mu$. By this reformulation we also end up with a maximisation problem instead of a minimisation problem which appears to be more stable for very small powers $k$ and extreme asymmetries $\tau \to 1$, $\tau \to 0$. 

4 Examples

In order to test our procedures we generate simple, bivariate data with \( n = 500 \) observations of the form \( y = f(x) + \varepsilon \) with \( f(x) = x^2 \) or \( f(x) = \exp(-x^2) \) and \( \varepsilon \sim \exp(0.5) \) or \( \varepsilon \sim 0.55N(0, 0.6^2) + 0.45N(3, 0.6^2) \). In these cases the conditional mode would be either at the lowest ends of the response’s values or in the lower half of the data with an additional, weaker mode in the top half. Both scenarios could not be estimated sensibly by kernel mode regression as the properties of the estimate are only known for symmetric and unbounded kernels.

However, as Figure 1 shows, boosting is possible down to \( k = 0.15 \), which seems to be a good approximation of the mode while still using a lot of information from the data. In comparison to the mean, median and (for fun) the 50\% 4-fractile we can also see that the mode approximation is a much better measure of location for the data at hand. Especially with bimodal errors (on the right) the mean represents the data rather poorly.

The addition of asymmetric weights as shown in Figure 2 also seems to work as it helps to uncover the second mode that is present in the data. While the first experiments were successful, we also uncovered an instability of the procedure for very small powers and extreme asymmetries. The combination might likely lead to artefacts in the estimate. The possibilities as well as the limits of fractiles are yet to be figured out. We also aim at a simulation study to assess the qualities of fractiles as replacement of mode regression and a comparison with GAMLSS by Rigby and Stasinopoulos (2005).
FIGURE 2. Example analysis of generated data with bimodal normal errors for fractiles with $k = 0.35$ and asymmetries between 0.01 and 0.99

References


