Electronic Journal of Statistics
Vol. 11 (2017) 3008–3038
ISSN: 1935-7524
DOI: 10.1214/17-EJS1307

Model selection in semiparametric expectile regression

Elmar Spiegel

University of Goettingen
Goettingen, Germany
e-mail: espiege@uni-goettingen.de

Fabian Sobotka

Carl von Ossietzky University
Oldenburg, Germany
e-mail: fabian.sobotka@uni-oldenburg.de

and

Thomas Kneib

University of Goettingen
Goettingen, Germany
e-mail: tkneib@uni-goettingen.de

Abstract: Ordinary least squares regression focuses on the expected response and strongly depends on the assumption of normally distributed errors for inferences. An approach to overcome these restrictions is expectile regression, where no distributional assumption is made but rather the whole distribution of the response is described in terms of covariates. This is similar to quantile regression, but expectiles provide a convenient generalization of the arithmetic mean while quantiles are a generalization of the median. To analyze more complex data structures where purely linear predictors are no longer sufficient, semiparametric regression methods have been introduced for both ordinary least squares and expectile regression. However, with increasing complexity of the data and the regression structure, the selection of the true covariates and their effects becomes even more important than in standard regression models. Therefore we introduce several approaches depending on selection criteria and shrinkage methods to perform model selection in semiparametric expectile regression. Moreover, we propose a joint approach for model selection based on several asymmetries simultaneously to deal with the special feature that expectile regression estimates the complete distribution of the response. Furthermore, to distinguish between linear and smooth predictors, we split nonlinear effects into the purely linear trend and the deviation from this trend. All selection methods are compared with the benchmark of functional gradient descent boosting in a simulation study and applied to determine the relevant covariates when studying childhood malnutrition in Peru.

Keywords and phrases: Expectiles, semiparametric regression, model selection, least asymmetrically weighted squares, boosting, non-negative garrote.

Received April 2016.

1. Introduction

Expectiles as introduced by Newey and Powell (1987) can be introduced in two ways, either as the generalization of the ordinary mean or as an alternative to quantiles. For the ordinary mean, the aim is to find the value that minimizes the average squared distance between the observed data points and the value such that the mean μ represents the center of gravity. In the following, we assume observations y_i, $i = 1, \ldots, n$, drawn from a random variable Y with finite expectation μ and finite variance. Then the ordinary mean of the data can be estimated via

$$\hat{\mu} = \underset{m}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - m)^2.$$

On the other hand, an α-quantile q_α is determined as the value where at least $\alpha \cdot 100\%$ of the data are located below and at least $(1-\alpha) \cdot 100\%$ of the data are located above. This value can be estimated as

$$\hat{q}_\alpha = \underset{q}{\operatorname{argmin}} \sum_{i=1}^{n} w_\alpha(y_i)|y_i - q|,$$

with weights $w_\alpha(y)$ depending on the data y and the asymmetry α via

$$w_\alpha(y) = \begin{cases} \alpha & \text{if } y \geq q \\ 1-\alpha & \text{if } y < q \end{cases}.$$

Expectiles are then introduced as a mixture of the mean and quantiles. On the one hand, they are determined by an asymmetrically weighted deviations criterion, where the L_1-norm of quantiles is replaced by the L_2-norm. On the other hand, they represent a weighted mean with weights depending on the asymmetry τ and the observed values of the data. Consequently, an estimate for the τ-expectile e_τ is determined via

$$\hat{e}_\tau = \underset{e}{\operatorname{argmin}} \sum_{i=1}^{n} w_\tau(y_i)(y_i - e)^2. \tag{1}$$

Since the expectile e_τ is the root of the first derivative of the loss function (1), the following characteristic equations can be derived:

$$\sum_{i \in I_1} (1-\tau)|y_i - e_\tau| = \sum_{i \in I_2} \tau |y_i - e_\tau| \tag{2}$$

$$\tau = \frac{\sum\limits_{i \in I_1} |y_i - e_\tau|}{\sum\limits_{i \in I_1 \cup I_2} |y_i - e_\tau|} \tag{3}$$

with sets of indices $I_1 = \{i | y_i < e_\tau\}$ and $I_2 = \{i | y_i \geq e_\tau\}$. As a consequence, expectiles e_τ represent the weighted center of gravity (Yao and Tong, 1996).

Furthermore, Equation (3) indicates that the fraction between the distances below the expectile and the total sum of distances is given by the corresponding asymmetry parameter τ. This statement replaces the corresponding statement on the number of data points for quantiles.

Both, expectiles and quantiles can be used to describe the complete distribution of a random variable when using a dense set of asymmetries / quantile levels τ. Even if they both differ in their definitions and therefore their properties, a bijective transformation between quantiles and expectiles exists (Yao and Tong, 1996; Schulze-Waltrup et al., 2015). Given the distribution function F of a continuous random variable Y, the bijective function $h: (0,1) \mapsto (0,1)$ converting the α-quantile q_α to the $h(\alpha)$-expectile $e_{h(\alpha)}$ is given by

$$h(\alpha) = \frac{-\alpha q_\alpha + \int_{-\infty}^{q_\alpha} y \mathrm{dF}(y)}{-\mu + 2\int_{-\infty}^{q_\alpha} y \mathrm{dF}(y) + (1-2\alpha)q_\alpha}.$$

This function can easily be calculated for standard types of distributions with finite variance, see Schulze-Waltrup et al. (2015) for further details and examples.

Similar as quantiles have been subjected to a regression problem (Koenker and Bassett, 1978), expectiles can also be used to estimate regression models (Newey and Powell, 1987), where the model specification is then given by

$$y_i = \eta_{i,\tau} + \varepsilon_{i,\tau}, \qquad i = 1, \ldots, n$$

with y_i being a continuous response and $\eta_{i,\tau}$ specifying the corresponding regression predictor. For the error term, no specific type of distribution is assumed but rather we assume that the error terms $\varepsilon_{\tau,i}$ are independent, have finite (but potentially different) variances and

$$0 = \underset{e}{\operatorname{argmin}}\, \mathbb{E}(w_\tau(\varepsilon_{i,\tau})(\varepsilon_{i,\tau} - e)^2)$$

holds, i.e. the τ-expectiles of the error terms are all zero. The resulting regression problem is typically solved using an iterative scheme called *least asymmetrically weighted squares* (*LAWS*), where estimation of the regression coefficients and the weights is done iteratively (compare Sobotka and Kneib, 2012, for further details). The main advantages of expectile regression are that no specific assumptions have to be made concerning properties of the error distribution such as homoscedasticity and that by estimating many expectiles the whole distribution of the response can be analyzed. In addition, expectile regression can easily be extended beyond the purely linear model specification when using semiparametric predictors (compare Section 2).

While the specification of regression models based on quantiles or expectiles has the clear advantage of reducing the required assumptions, it also makes inference and model choice questions more challenging since the models are no longer based on a likelihood. As a consequence, both likelihood-based inference (e.g. likelihood ratio tests or Akaike's information criterion, AIC) are no

longer immediately available. While either resampling approaches such as the bootstrap or asymptotic normality results can be used to conduct inferences in expectile regression models (see for example Sobotka et al., 2013), model choice questions have never been investigated in detail. This is even more relevant when considering semiparametric regression models where we do not only have to decide which covariates should be included but also whether effects should be purely linear or nonlinear. Furthermore, for expectile (and quantile) regression models, a common question is whether covariates should be selected for the different expectile levels separately or simultaneously for the complete response distribution.

We will mostly focus on two avenues for approaching model selection: (1) procedures based on information criteria such as stepwise selection based on AIC-type criteria that are very popular in ordinary least squares regression (Burnham and Anderson, 2002) and (2) shrinkage approaches such as the least absolute selection and shrinkage operator (LASSO) (Tibshirani, 1996) that augment a complexity penalty to the fit criterion. Since the L_1 penalty induced by the LASSO does not fit well with the L_2 geometry of expectiles, we will consider the non-negative garrote (Breiman, 1995) as an alternative.

An alternative to AIC-type criteria is provided by proper scoring rules derived from information-theoretic considerations Gneiting and Raftery (2007). Proper scoring rules can be applied for several regression types, including quantile regression and expectile regression (Gneiting, 2011). For quantile regression, further approaches on model selection have been introduced, starting with a goodness of fit criterion suggested in Koenker and Machado (1999). Alternatively, several authors including Li and Zhu (2008), Zou and Yuan (2008a), Zou and Yuan (2008b), Wu and Liu (2009), Koenker (2011) and Jiang, Bondell and Wang (2014), introduced LASSO-type or SCAD-type penalties to select variables in quantile regression (for a detailed review see Wu and Ma, 2015). All of them, except Koenker (2011), restrict their approaches to linear predictors while Koenker (2011) penalizes the variation of nonlinear effects but does not include the possibility to de/select them. Moreover, most of the former papers select the variables separately for each quantile level while Zou and Yuan (2008b) and Jiang, Bondell and Wang (2014) introduce methods to select the covariates jointly for a given set of asymmetries (albeit with the restriction that coefficients should vary smoothly over the quantile levels).

Among others (see Gijbels, Verhasselt and Vrinssen, 2015, for a review of recent approaches) Huang, Horowitz and Wei (2010), Greven and Kneib (2010), Marra and Wood (2011) and Chouldechova and Hastie (2015) introduce approaches to perform model selection in regression specifications with semiparametric predictors but a pre-specified type of response distribution. Greven and Kneib (2010) suggest criterion-based selection, while the others use shrinkage approaches.

Further research has also been done to implement semiparametric predictors in quantile regression (see Koenker, Ng and Portnoy, 1994, He and Ng, 1999, Doksum and Koo, 2000). In this context some ideas for deciding whether nonlinear effects are indeed necessary have been introduced. Most of them deal with

varying coefficient models and try to decide if the varying part is required (see for example Wang, Zhu and Zhou, 2009, Tang et al., 2012, Noh et al., 2012, Tang, Wang and Zhu, 2013). In order to select the models, they make use of asymptotic distributional results or shrinkage methods. Moreover, they select the covariates for each quantile level separately. Some approaches to overcome this assumption in varying coefficient models are introduced in Kai, Li and Zou (2011) and Guo, Yang and Lv (2015). Here the models of several quantile levels are selected jointly via a penalization approach. In addition, Guo et al. (2013), Lin et al. (2013) and Lv, Yang and Guo (2015) suggest methods to use truly additive models and apply model selection on the linear and the nonlinear predictors via penalization approaches.

A flexible alternative to criteria- or penalization-based approaches is functional gradient descent boosting (Bühlmann and Hothorn, 2007), where coefficient estimation and model selection is done in one estimation run. Boosting also has the advantage that selection of nonlinear and linear effects is readily available and can be combined with a variety of model specifications. Boosting has been introduced to semiparametric quantile regression by Fenske, Kneib and Hothorn (2011).

All told, we are not just adopting the above procedures to expectiles. Instead we combine several approaches, which have been used only separately in previous papers and add new ideas. First, we select variables based on AIC-type criteria and shrinkage methods. Thereby we introduce methods to select models for the whole distribution by combining multiple asymmetries. Whether or not performing model selection for the complete distribution or for single asymmetries very much depends on the specific context of the application. If we focus on specific parts of the distribution such as the tail, it will in general be preferable to select only asymmetries associated with this part of the distribution and to perform separate model selection. If, on the other hand, the complete distribution of the response should be studied, we expect benefits from performing simultaneous selection. In particular, this will have the advantage to provide a consistent model selection for the complete distribution, allows us to borrow strength from neighboring expectiles and facilitates interpretation of the results.

Second, to overcome the assumption, that a covariate influences the response linearly the specification of semiparametric regression is advantageous. Therefore all our suggested methods can also select nonlinear predictors. However, model selection with semiparametric predictors is more complex. In this paper, we make use of the fact, that P-splines can be separated in a linear trend and the wiggly/nonlinear deviation of this trend (Fahrmeir, Kneib and Lang, 2004) and select them separately. This decomposition is orthogonal in the parameters such that selection between the linear and the nonlinear part is not deterred by concurvity of the linear and the nonlinear effect.

The remainder of the article is organized as follows: In Section 2, we briefly review some basic concepts of semiparametric regression. Section 3 discusses model selection methods that allow to separate between linear and nonlinear predictor structures. Furthermore, the expectile-specific model selection approaches are introduced. The following Section 4 contains a simulation study on the empirical

performance of the different approaches while Section 5 presents the application of our methods on the estimation of determinants for chronic undernutrition of children in Peru. The concluding remarks are summarized in Section 6.

2. Semiparametric regression models

In the remainder of this paper, we will consider semiparametric regression specifications where the response variable y_i is related to a combination of linear effects $x_{il}\beta_l$ as well as smooth, nonlinear effects $f_s(x_{is})$ of continuous covariates such that

$$y_i = \beta_0 + \ldots + x_{il}\beta_l + \ldots + f_s(x_{is}) + \ldots + \varepsilon_i.$$

More generically, bivariate splines for interaction surfaces, Markov random fields or Kriging terms for spatial effects and a variety of other effect types can be included in a similar way (see Sobotka and Kneib, 2012). In the following we will focus on the special case of univariate nonlinear effects for the sake of illustration.

To further simplify the presentation, assume a model with only one covariate with nonlinear effect such that we obtain the specification

$$y_i = f(x_i) + \varepsilon_i$$

with smooth function f. This function is then approximated by a weighted sum of B-spline basis functions $B_r^{(d)}$ of a fixed degree d such that

$$f(x) = \sum_{r=1}^{R} \gamma_r B_r^{(d)}(x).$$

where γ_r are the corresponding coefficients and R is the dimensionality of the basis, i.e. the number of basis elements. If the matrix $\boldsymbol{Z}$ contains the values of the basis functions evaluated at the observed covariate values and $\boldsymbol{\gamma}$ is the vector of coefficients, the model can then be rewritten in matrix-vector notation as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{4}$$

To reduce the dependency on the dimension and the position of the basis functions, Eilers and Marx (1996) introduced P-splines where a large number of basis functions is combined with a difference penalty on the basis coefficients such that the estimated coefficients $\boldsymbol{\gamma}$ do not vary too much between neighboring basis functions. This leads to the penalized least squares criterion

$$(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}$$

where $\boldsymbol{K}$ is a penalty matrix representing squared differences of a predetermined order p and $\lambda \geq 0$ denotes the smoothing parameter that governs the trade-off

between smoothness of the function estimate ($\lambda \to \infty$) and fit to the data ($\lambda \to 0$).

To optimize λ, a first option is to use selection criteria like the AIC, or generalized cross-validation. Alternatively, the *Schall* algorithm (Schall, 1991) can be applied since P-splines can be cast in a mixed model framework (see for example Fahrmeir, Kneib and Lang, 2004). The Schall algorithm as well as optimization of selection criteria can also be adapted to the combination of P-splines with expectile regression, see Sobotka and Kneib (2012) and Schnabel and Eilers (2009).

When including nonlinear effects in a regression model, a typical model choice question relates to the decision whether a covariate should be included linearly or flexibly i.e. based on a P-spline. To facilitate this decision, we decompose the coefficients of the P-spline in its penalized and unpenalized part (following Lin and Zhang, 1999; Currie and Durban, 2002; Fahrmeir, Kneib and Lang, 2004)

$$\boldsymbol{\gamma} = \boldsymbol{V}\boldsymbol{\gamma}^{unp} + \boldsymbol{W}\boldsymbol{\gamma}^{pen}.$$

with matrices $\boldsymbol{V}$ and $\boldsymbol{W}$ spanning the null space and the orthogonal deviation from the null space of the penalty matrix $\boldsymbol{K}$. Plugging this reparameterization into Equation (4) yields

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Z}(\boldsymbol{V}\boldsymbol{\gamma}^{unp} + \boldsymbol{W}\boldsymbol{\gamma}^{pen}) + \boldsymbol{\varepsilon} \\ &= \tilde{\boldsymbol{V}}\boldsymbol{\gamma}^{unp} + \tilde{\boldsymbol{W}}\boldsymbol{\gamma}^{pen} + \boldsymbol{\varepsilon} \end{aligned} \tag{5}$$

With this representation, the first part contains the unpenalized coefficients which represent a linear function for second order differences ($p = 2$) and more generally a polynomial of order $p - 1$. The second part contains the penalized coefficients, which can be interpreted as the smooth deviation from the linear effect. Both parts can then be treated as separate model components in the model selection task (see Section 3.1 for details). Note that the orthogonal decomposition provided by the matrices $\boldsymbol{V}$ and $\boldsymbol{W}$ improves the model choice performance since concurvity between the linear effect and the nonlinear deviation is avoided by construction.

3. Selection methods

3.1. Model selection for P-splines

For continuous covariates, we consider four selection approaches:

- Linear vs. no effect: The covariate is either included as a linear function or not included at all. In this case, no nonlinear effects can be chosen.
- Nonlinear vs. no effect: The covariate is included as a P-spline (without decomposing it into penalized and unpenalized part) or not included at all. Note that the smoothing parameter still allows to reduce the effect to a polynomial of degree $p-1$ (and therefore a linear effect for the standard case of $p = 2$) if $\lambda \to \infty$.

- Nonlinear vs. linear vs. no effect: In this case, all three possibilities are compared, i.e. the covariate is included as a P-spline (without decomposing it into penalized and unpenalized part), as a linear effect or it is dropped from the model. We will consider this possibility in combination with best subset and stepwise forward methods (see Section 3) and only allow for one of the competing alternatives, i.e. once the covariate is included as either linear or nonlinear effect, the other inclusion variant is not possible any more. Accordingly we call this method *restricted*.
- Decomposition into linear and nonlinear effect: Here the decomposition of the P-spline in a linear part and the deviation from the linear effect is utilized as described in Equation (5). Hence the covariate can be included as a linear effect, the deviation from the linear function, or the combination of both effects which then yields the original P-spline again. Based on the orthogonality of the effects this corresponds to an independent treatment of penalized and unpenalized part such that both are included or excluded separately from the model. This method is called *complete* in the following and can be used in combination with all methods defined in the following including the non-negative garrote.

Other smooth function types such as Markov random fields can also be selected with the approaches defined in Section 3.2.1 to Section 3.3.3. However no decomposition is used for them such that the model choice is simplified considerably.

Besides the selection of semiparametric predictors, the main novelty of model selection for expectile regression is that selection can be done for each asymmetry separately or jointly for the whole distribution. In the next sections we introduce model selection methods for each asymmetry separately and discuss joint selection afterwards.

3.2. Selection methods for a single asymmetry

3.2.1. Best subset and stepwise selection

Classical model selection methods like stepwise or best subset selection rely on the definition of an appropriate selection criterion. We therefore introduce adaptations of selection criteria well known from mean regression.

The first criterion is obtained by defining a cross-validated measure corresponding to the *least asymmetrically weighted squares* (*LAWS*) criterion

$$\frac{1}{n}\sum_{i=1}^{n} w_\tau(y_i)(y_i - \hat{y}_{i,\tau})^2 \tag{6}$$

where y_i is the observed response and $\hat{y}_{i,\tau}$ is the predicted expectile for this observation. For the cross-validated LAWS criterion, we determine expectile regression estimates on a subset of the data and evaluate its predictive ability based on the predictive LAWS criterion. In the following, this predictive index

is called *mean weighted squared error* (*MWSE*, in generalization of the standard mean squared error, MSE).

As an alternative, we consider a generalization of Akaike's information criterion (AIC) (Akaike, 1974) defined as

$$AIC = n \ \log\left(\frac{1}{n}\sum_{i=1}^{n} w_\tau(y_i)(y_i - \hat{y}_{i,\tau})^2\right) + 2df$$

where the weighted sum of squared errors replaces the negative likelihood while the degrees of freedom $df = \text{trace}(\boldsymbol{H})$ are determined based on the hat matrix

$$\boldsymbol{H} = (\boldsymbol{W}_\tau)^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}_\tau\boldsymbol{X} + \boldsymbol{K})^{-1}\boldsymbol{X}'(\boldsymbol{W}_\tau)^{1/2},$$

where $\boldsymbol{X}$ is the complete design matrix of all effects, $\boldsymbol{K}$ is the (block diagonal) penalty matrix consisting of all individual penalty matrices and their associated smoothing parameters, and $\boldsymbol{W}_\tau$ is the diagonal matrix of the weights per observation for the current asymmetry τ. Similarly, an expectile version of the Bayesian information criterion *BIC* (Schwarz et al., 1978) can be defined by adjusting the weight on the degrees of freedom. Note that in expectile regression we did not make an explicit assumption on the distribution of the error term and therefore neither AIC nor BIC are strictly justified based on standard arguments as in Burnham and Anderson (2002) but they are rather defined as ad hoc analoga to these criteria.

3.2.2. Non-negative garrote

An alternative to best subset or stepwise selection procedures are shrinkage methods that avoid the necessity to estimate multiple models for a comparison. They rather combine estimation and model selection in one step. The most prominent example is the *LASSO* by Tibshirani (1996) which adds an L_1 penalty term to the fit criterion. Due to the specific geometry of the L_1 penalty of the LASSO, covariates are excluded from the model if their influence on the response is too small.

Since combining the LASSO with penalized predictors is not directly possible due to the ties in the coefficients, several attempts to build shrinkage methods for semiparametric predictors have been developed (see Marra and Wood, 2011, for an overview). In this article, we focus on the *non-negative garrote* introduced by Breiman (1995) that uses a two-step approach for shrinking. In a first step, the saturated model with all effects included is estimated. In a second step, the prediction accuracy is optimized by multiplying every estimated predictor component with an extra weight δ and checking which predicted value has a relevant influence on the response. We introduce the non-negative garrote for semiparametric expectile regression following the notation of Marra and Wood (2011). To get the special case of ordinary least squares, one has to set $\tau = 0.5$, such that τ can be ignored in the formulae.

In the following, we consider the non-negative garrote for models with semiparametric predictors $\eta_\tau = \sum_k f_{k,\tau}(x_k)$ where, for the sake of simplicity, also linear effects are represented as arbitrary functions $f_{k,\tau}(x_k) = x_k\beta_{k,\tau}$. In principle the aim of non-negative garrote is to optimize the coefficients not only due to the model fit, but also by the prediction accuracy. Therefore the MWSE is transformed to the following optimization criterion

$$\sum_{i=1}^{n} w_\tau(y_i)\left(y_i - \beta_{0,\tau} - \sum_{k=1}^{K}\hat{f}_{k,\tau}\delta_{k,\tau}\right)^2, \tag{7}$$

where $\delta_{k,\tau} \geq 0$ is an extra weight to optimize the predictive model fit. These weights $\delta_{k,\tau}$ are limited in size by the tuning parameter $\xi_\tau = \sum_{k=1}^{K}\delta_{k,\tau}$. To explain the procedure in more detail, we first show how to estimate the weights $\boldsymbol{\delta}_\tau = (\delta_{1,\tau},\ldots,\delta_{K,\tau})^T$ for a given tuning parameter ξ_τ. Next, we explain how to estimate the optimal tuning parameter ξ_τ. Finally we provide an algorithm to combine these steps and estimate the final weights $\boldsymbol{\delta}_\tau$ for a specific asymmetry τ.

For the estimation of $\boldsymbol{\delta}_\tau$, we assume a specific tuning parameter ξ_τ and estimate the saturated expectile regression with all effects included in the first step. Hence, the predictors $\hat{f}_{k,\tau}$ and the weights $w_\tau(y_i)$ are then available from this full model. Since we only consider cases with $K < n$, no specific form of the estimates $\hat{f}_{k,\tau}$ needs to be assumed. As the intercept is not subject to selection, the response $\boldsymbol{y} = (y_1,\ldots,y_n)^T$ is transformed to $\breve{\boldsymbol{y}} = \boldsymbol{y} - \beta_{0,\tau}$ for simplification. With this, Equation (7) can be rewritten in matrix notation as

$$\begin{aligned}\hat{\boldsymbol{\delta}}_\tau &= \underset{\boldsymbol{\delta}_\tau}{\operatorname{argmin}}\ \sqrt{\boldsymbol{W}}(\breve{\boldsymbol{y}} - \hat{\boldsymbol{F}}\boldsymbol{\delta}_\tau)^2 \\ &= \underset{\boldsymbol{\delta}_\tau}{\operatorname{argmin}}\ \breve{\boldsymbol{y}}^T\boldsymbol{W}\breve{\boldsymbol{y}} - 2\breve{\boldsymbol{y}}^T\boldsymbol{W}\hat{\boldsymbol{F}}\boldsymbol{\delta}_\tau + \boldsymbol{\delta}_\tau^T\hat{\boldsymbol{F}}^T\boldsymbol{W}\hat{\boldsymbol{F}}\boldsymbol{\delta}_\tau \end{aligned} \tag{8}$$

where $\hat{\boldsymbol{F}} = \left[\hat{f}_{1,\tau};\ \ldots\ ;\hat{f}_{K,\tau}\right]$ is the matrix of predicted values and $\boldsymbol{W} = \operatorname{diag}(w_\tau(y_i))$ is the diagonal matrix of expectile weights. Moreover it is possible to apply the methods for solving quadratic programming problems on Equation (8). This yields an estimate for $\hat{\boldsymbol{\delta}}_{\boldsymbol{\tau}}$ in the second step and the optimized model comprises the effects

$$\hat{f}_{k,\tau}^{new} = \hat{f}_{k,\tau}\cdot\hat{\delta}_{k,\tau}.$$

If the saturated model is indeed the true model, then $\xi_\tau = K$ and $\delta_{k,\tau} = 1$ for all $k = 1,\ldots,K$, thus the non-negative garrote changes nothing. However, if there is a covariate x_k, which is irrelevant for the prediction, then the prediction will be better if its weight $\delta_{k,\tau}$ is close to or even equal to zero such that the effect of x_k is excluded from the model. Furthermore, the weights $\delta_{k,\tau}$ can be larger than 1. This is, however unlikely, as the regression also minimizes the divergence between the observed values and the predictions. Moreover the value of ξ_τ is essential for the estimation, but it has to be specified in advance. In order to find the model with the best prediction accuracy, the tuning parameter ξ_τ is

determined via cross-validation out of a grid Ξ of possible tuning parameters ξ_τ. This leads to the following algorithm:

1. Build the grid Ξ of possible tuning parameters ξ_τ.
2. Split the data into G pairs of training and validation data sets.
3. Iterate over all pairs $g \in G$ and
   a. Estimate the full expectile regression model for the training data set to obtain $\hat{f}_{k,\tau}$ and $w_\tau(y_i)$.
   b. Iterate over all $\xi_\tau \in \Xi$ and
      i. Estimate the parameter $\hat{\boldsymbol{\delta}}_\tau$ based on the training data set to obtain the effect estimates $\hat{f}_{k,\tau}$.
      ii. Compute the updated coefficients $(\hat{f}_{1,\tau}^{new}, \ldots, \hat{f}_{K,\tau}^{new}) = (\hat{f}_{1,\tau}\hat{\delta}_{1,\tau}, \ldots, \hat{f}_{K,\tau}\hat{\delta}_{K,\tau})$.
      iii. Use these new coefficients to predict the expectiles for the corresponding validation data set.
      iv. Estimate the *MWSE* for this validation data set g and tuning parameter ξ_τ to obtain $MWSE_{g,\xi_\tau}$.
4. Build the cross-validation score for each ξ_τ separately $(score(\xi_\tau) = \frac{1}{G}\sum_{g=1}^{G} MWSE_{g,\xi_\tau})$.
5. Find the minimal *score* and use the corresponding ξ_τ as the optimal one (ξ_τ^{opt}).

With the algorithm described above, we can now define the complete algorithm for using the non-negative garrote:

I. Find the optimal tuning parameter ξ_τ^{opt} based on cross-validation.
II. Use this tuning parameter ξ_τ^{opt} to estimate the final weights $\hat{\boldsymbol{\delta}}_\tau$ of the complete data set.
III. Compute the new coefficients $\hat{f}_{k,\tau}^{new} = \hat{f}_{k,\tau}\hat{\delta}_{k,\tau}$ and treat the result as the final model.

We prefer using the non-negative garrote compared to the LASSO due to several reasons. First, it can easily be adopted for semiparametric models (see Marra and Wood, 2011). Second, the non-negative garrote does not limit the value of the coefficients, as LASSO does, but it rather measures the influence of the predicted values. This is appropriate for expectile regression, as we allow the coefficients to vary freely in between the different asymmetries τ. Last but not least, this method can be expanded to a joint estimation of the weights for all asymmetry parameters simultaneously (see Section 3.3.3).

3.3. Selection methods for the complete distribution

The methods considered in the previous sections select the optimal model for each asymmetry parameter τ separately. This is advantageous in order to analyze which covariate has an influence on specific parts of the distribution of

the response. Furthermore, the resulting models incorporate only relevant information for this specific asymmetry parameter. However, it might be easier to compare the effects of the covariates between the asymmetries if the same covariates are included for all asymmetries. Moreover, the probability of crossing expectiles is reduced with this joint approach. Our simulation studies show that with an approach for joint selection a superfluous covariate is excluded more often than otherwise. Still, a covariate with a specific influence on one tail can build up enough leverage to remain in a joint model.

3.3.1. Mean AIC

The first method we propose to select the optimal model for all asymmetry parameters jointly is the *mean AIC* or *area under the criteria curve*. This method is motivated by the fact that the whole distribution of the response can be estimated with expectile regression. Then the selection criterion for a given model, treated as a function of the asymmetry parameter τ, can be compared with the corresponding function of a competing model specification. To reduce the information to one single value (and therefore to facilitate the decision which of the models is "better"), the area under the criteria curve is determined and the model with the smaller area is preferred (for a negative orientation of the criterion as for example in case of the AIC). This principle is illustrated in Figure 1 where the model including the P-spline of x_2 performs better on the left side of the distribution while the simpler model would be preferred on the right side. When comparing the area under criterion curve, the more complex model is found to perform better. Of course the integrated criterion can easily be weighted such that specific parts of the distribution are more relevant than other when doing the comparison.

In practice, the whole distribution will be approximated by a grid of asymmetry parameters τ_j with $j = 1, \ldots, J$. For these parameters, the expectile regression will be estimated separately and the corresponding selection criterion $AIC(\tau_j)$ can be computed. Furthermore the *area under the criteria curve* is approximated with the simple Riemannian sum as

$$\text{area} = \sum_{j=1}^{J-1} AIC(\tau_j) \cdot (\tau_{j+1} - \tau_j).$$

If the asymmetry parameters τ_j are chosen on a homogeneous grid, the *area* criterion is proportional to the arithmetic mean of the individual contributions $AIC(\tau_j)$

$$\overline{AIC} = \frac{1}{J} \sum_{j=1}^{J} AIC(\tau_j).$$

Again, this index can also be replaced by a weighted mean with weights favoring outer or inner part of the distribution. Eventually this index can be used for stepwise model selection, for example, and the resulting model will contain the same covariates for all asymmetry parameters.

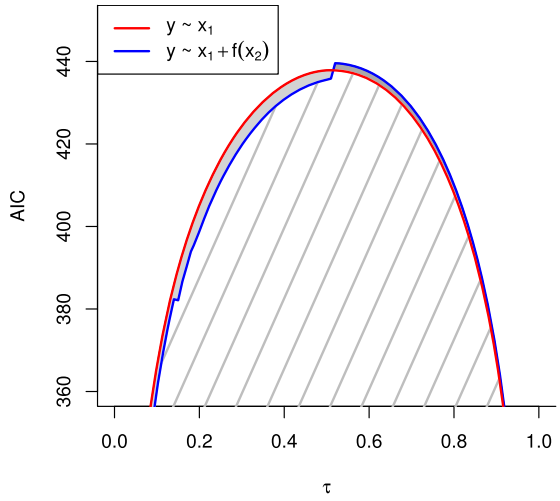


FIG 1. *Differences between models as area between the criteria curves. This graphic is based on the exponential simulation design of Section 4, but with more extreme parameters and fewer observations.*

3.3.2. Proper scoring rules

The purpose of model selection is often to identify models with good predictive performance. In the case of optimizing the model separately for each asymmetry, we therefore considered cross-validation via the MWSE criterion. For joint selection, we adopt the notion of scoring rules S (Gneiting and Raftery, 2007; Gneiting, 2011) that evaluate the fit between predictive distributions and actually observed realizations. More precisely, we assume that a forecast is probabilistic and we get a predictive distribution P as forecast for a given realization y. Then

$$S : (P, y) \mapsto s \in \mathbb{R}$$

is a scoring function and we write $S(P, Q)$ for the expected value of $S(P, \cdot)$ under Q. If Q is the best possible distributional forecast we say that S is a strictly proper scoring rule, if

$$S(P, Q) \geq S(Q, Q)$$
$$S(P, Q) = S(Q, Q) \Leftrightarrow P = Q.$$

For expectiles we assume to have asymmetries $\tau_1, \ldots, \tau_J \in (0, 1)$ with corresponding forecasts $e_{\tau_1}, \ldots, e_{\tau_J}$ for a given observation y and a probability measure P. The expected score is now given as

$$S(e_{\tau_1}, \ldots, e_{\tau_J}; P) = \int S(e_{\tau_1}, \ldots, e_{\tau_J}; y) dP(y) \tag{9}$$

similar to the quantile representation in Gneiting and Raftery (2007). Furthermore we know from Gneiting (2011) that a strictly proper scoring rule for expectiles is given by

$$S(e_\tau, y) = w_\tau(y)(\psi(y) - \psi(e_\tau) - \psi'(e_\tau)(y - e_\tau))$$

with ψ being a convex function and ψ' being its subgradient. The most prominent example is constructed with $\psi(x) = x^2$. Then the proper scoring rule for expectiles is similar to our asymmetrically weighted squared error:

$$S(e_\tau, y) = w_\tau(y)(y - e_\tau)^2.$$

To approximate (9), we utilize a set of asymmetry parameters to get the total distribution from the corresponding expectiles, i.e.

$$S(e_{\tau_1}, \ldots, e_{\tau_J}; y) = \sum_{j=1}^{J} w_{\tau_j}(y)(y - e_{\tau_j})^2 \tag{10}$$

similar as for quantiles (see Gneiting and Raftery, 2007, for further details on the quantile setting). To apply this on the model selection framework, we build the score for a complete data set as

$$score = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{n_g}\sum_{i=1}^{n_g}\sum_{j=1}^{J} w_{\tau_j}(y_i)(y_i - \hat{y}_{i,\tau_j})^2$$

where y_i is the true data point that materialized and $e_{i,\tau} = \hat{y}_{i,\tau}$ the predicted value for this point, G is the number of cross-validation folds and n_g is the size of the validation data sets in the individual folds. The cross-validated *score* can then again be used for stepwise model selection or for comparing a pre-selected set of models.

In applications, *scoring* can also be considered as a mixture of cross-validation and *mean AIC*, i.e. it is the prediction error integrated via asymmetries (as shown for quantiles in Gneiting and Ranjan, 2011). As a consequence, we cannot only look at the complete score but can also decompose it along the asymmetries similar as we did it for the AIC in Figure 1. Again, a weight vector emphasizing specific parts of the distribution can easily be included.

3.3.3. Non-negative garrote on a grid

To generalize the non-negative garrote from single asymmetries to the complete distribution, a grid of asymmetry parameters τ_j is used and expectile regressions for all τ_j are estimated, such that $\hat{f}_{k,\tau_j}$ and w_{τ_j} are given. With these estimates, we solve

$$(\hat{\delta}_1, \ldots, \hat{\delta}_K)^T = \underset{\delta_1,\ldots,\delta_K}{\operatorname{argmin}} \sum_{j=1}^{J}\sum_{i=1}^{n} w_{\tau_j}(y_i)\left(y_i - \hat{\beta}_{0,\tau_j} - (\hat{f}_{1,\tau_j}\delta_1 + \ldots + \hat{f}_{k,\tau_j}\delta_k)\right)^2 \tag{11}$$

such that the weights $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)^T$) are the same for all asymmetries while the same constraints as in Section 3.2.2 apply ($\delta_k \geq 0$ and $\sum_k \delta_k = \xi$). To solve this minimization problem with quadratic programming routines, it is necessary to rewrite (11) in matrix notation. With

$$\breve{y}_{i,\tau_j} = y_i - \hat{\beta}_{0,\tau_j} \qquad \breve{\boldsymbol{y}}_{\tau_j} = (\breve{y}_{1,\tau_j}, \ldots, \breve{y}_{n,\tau_j})^T$$
$$\boldsymbol{W}_{\tau_j} = \text{diag}(w_{\tau_j}(y_1), \ldots, w_{\tau_j}(y_n)) \qquad \hat{\boldsymbol{F}}_{\tau_j} = (\hat{f}_{1,\tau_j}, \ldots, \hat{f}_{K,\tau_j})$$

$$\underline{\breve{\boldsymbol{y}}} = \begin{pmatrix} \breve{\boldsymbol{y}}_{\tau_1} \\ \vdots \\ \breve{\boldsymbol{y}}_{\tau_J} \end{pmatrix}, \quad \underline{\boldsymbol{W}} = \begin{pmatrix} \boldsymbol{W}_{\tau_1} & & \\ & \ddots & \\ & & \boldsymbol{W}_{\tau_J} \end{pmatrix}, \quad \underline{\hat{\boldsymbol{F}}} = \begin{pmatrix} \hat{\boldsymbol{F}}_{\tau_1} \\ \vdots \\ \hat{\boldsymbol{F}}_{\tau_J} \end{pmatrix}$$

the values of the separate asymmetries are combined in a row-wise fashion such that $\underline{\breve{\boldsymbol{y}}}$ is an $((n \cdot J) \times 1)$ vector, $\underline{\hat{\boldsymbol{F}}}$ is an $((n \cdot J) \times K)$ matrix and $\underline{\boldsymbol{W}}$ is an $((n \cdot J) \times (n \cdot J))$ diagonal matrix. With these definitions Equation (11) can be transformed as

$$(\hat{\delta}_1, \ldots, \hat{\delta}_K)^T = \underset{\boldsymbol{\delta}}{\text{argmin}}(\underline{\breve{\boldsymbol{y}}}^T \underline{\boldsymbol{W}} \underline{\breve{\boldsymbol{y}}} - 2\underline{\breve{\boldsymbol{y}}}^T \underline{\boldsymbol{W}}\, \underline{\hat{\boldsymbol{F}}} \boldsymbol{\delta} + \boldsymbol{\delta}^T \underline{\hat{\boldsymbol{F}}}^T \underline{\boldsymbol{W}}\, \underline{\hat{\boldsymbol{F}}} \boldsymbol{\delta}) \tag{12}$$

and (12) can be solved with standard tools for solving quadratic problems. As in the separate case, the optimal ξ is computed via cross-validation.

3.4. Boosting

As a competitor for the approaches introduced in this paper, we consider functional gradient descent boosting (Bühlmann and Hothorn, 2007; Hofner et al., 2014) that provides a generic way of minimizing the empirical loss

$$f^* := \underset{f}{\text{argmin}} \sum_{i=1}^{n} \rho(y_i, f(\boldsymbol{x}_i))$$

with a pre-specified loss function ρ and a regression specification $f(\boldsymbol{x}_i)$. Boosting for each asymmetry parameter separately has been introduced to expectile regression by Sobotka and Kneib (2012) utilizing the weighted differences

$$\rho(y_i, f(\boldsymbol{x})) = w_\tau(y_i)(y_i - f(\boldsymbol{x}))^2$$

as loss function (for further details see Sobotka and Kneib, 2012). Similarly boosting was introduced to quantile regression by Fenske, Kneib and Hothorn (2011). They defined the loss function ρ as

$$\rho(y_i, f(\boldsymbol{x})) = w_\tau(y_i)|y_i - f(\boldsymbol{x})|$$

Boosting also allows to use semiparametric predictors and is therefore used as a benchmark in our simulation studies.

4. Simulation study

We conduct a simulation study that evaluates both the ability to identify relevant covariates and the ability to discriminate between linear and nonlinear effects for continuous covariates to determine the behavior of the different selection methods.

4.1. Design

For all scenarios considered in the following, we rely on the additive predictor

$$\eta = \beta_0 + f_{1,\tau}(x_1) + f_{2,\tau}(x_2) + f_{3,\tau}(x_3) + f_{4,\tau}(x_4)$$

where all covariates x_1, x_2, x_3, x_4 are randomly drawn from a uniform distribution $U(1,2)$ and x_4 is always a noise covariate, i.e. $f_{4,\tau}(x_4) \equiv 0$. We then designed three different settings (see also Figure 2):

1. *Parallel design*: The effects are equal for all asymmetry parameters τ (see Figure 2a):

$$\begin{gathered} y_i = \eta_i + \varepsilon_i, \quad \varepsilon_i \sim N(0,4) \\ \beta_0 = 2,\ f_1(x_1) = 6 \cdot x_1,\ f_2(x_2) = 2 \cdot x_2,\ f_3(x_3) = 0.75 \cdot x_3 \end{gathered}$$

2. *Linear design*: The effect size increases linearly with increasing asymmetry parameters τ, (see Figure 2b):

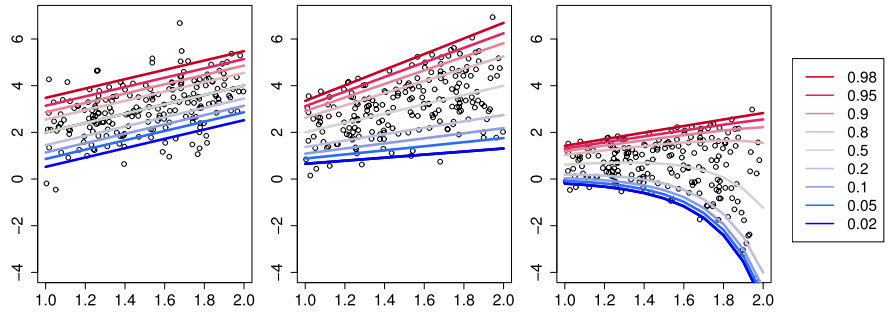
$$\begin{aligned} y_i &= \eta_i + \varepsilon_i, \quad \varepsilon_i = 0 \\ f_{1,\tau_i}(x_{i1}) &= x_{i1} \cdot \beta_{1,\tau_i} = x_{i1} \cdot 3 \cdot q_{\tilde{\tau}_i,2,0.5} \\ f_{2,\tau_i}(x_{i2}) &= x_{i2} \cdot \beta_{2,\tau_i} = x_{i2} \cdot 1 \cdot q_{\tilde{\tau}_i,2,0.5} \\ f_{3,\tau_i}(x_{i3}) &= x_{i3} \cdot \beta_{3,\tau_i} = x_{i3} \cdot 0.5 \cdot q_{\tilde{\tau}_i,1.5,0.5} \end{aligned}$$

   where $q_{\tilde{\tau},\text{mean},\text{sd}}$ is the $\tilde{\tau}$-quantile of $N(\text{mean}, \text{sd})$ and $\tilde{\tau}_i = h(\tau_i)$ where h is the bijective function converting the τ-quantile q_τ to the $h(\tau)$-expectile $e_{h(\tau)}$.
3. *Exponential design*: With increasing asymmetry parameters τ, the effect size increases. The effect is linear for x_1 and x_3, while it is exponential for x_2 (see Figure 2c):

$$\begin{aligned} y_i &= \eta_i + \varepsilon_i, \quad \varepsilon_i = 0 \\ f_{1,\tau_i}(x_{i1}) &= x_{i1} \cdot \beta_{1,\tau_i} = x_{i1} \cdot 3 \cdot q_{\tilde{\tau}_i,2,0.5} \\ f_{2,\tau_i}(x_{i2}) &= x_{i2} \cdot \beta_{3,\tau_i} + (\tilde{\tau}_i - 1) \cdot \frac{\exp\left((x_{i2})^2\right)}{10} \\ f_{3,\tau_i}(x_{i3}) &= x_{i3} \cdot \beta_{3,\tau_i} = x_{i3} \cdot 0.5 \cdot q_{\tilde{\tau}_i,1.5,0.5} \end{aligned}$$

Note that there is a fundamental difference in the way the data are generated in the parallel design as compared to the linear and the exponential design. For

the construction of data sets with a pre-specified structure for the expectiles, we rely on a generalization of importance sampling where the quantile function is replaced by the expectile function. More precisely, we draw quantile levels randomly from the uniform distribution, i.e. $\tau_i \sim U(0,1)$ for observation i. Then the quantile level is transformed to the expectile asymmetry using the transfer function h such that $\tilde{\tau}_i = h(\tau_i)$. The asymmetry is then plugged into the quantile function of the $N(\text{mean}, \text{sd})$ distribution, yielding $q_{\tilde{\tau},\text{mean},\text{sd}}$. The mean of the normal then controls the overall size of the effect, while the standard deviation controls the overall variation of the data. The main advantage of this complex design is that we get varying but predictable effects for each asymmetry parameter. With $\text{sd} = 0$ this approach reduces to the parallel design where an additional error term is still necessary.



(a) Parallel design (b) Linear design (c) Exponential design

FIG 2. *Simulation design for $f_{2,\tau}$ in all three data settings.*

All simulations are based on $n = 2000$ observations (results with $n = 500$ observations are available in the supplementary material at https://www.uni-goettingen.de/en//511092.html, but the basic conclusions did not change) and the asymmetry parameters $\{0.02, 0.05, 0.1, 0.2, 0.5, 0.8, 0.9, 0.95, 0.98\}$. For each scenario, we consider 100 replications. Cross-validation in all selection methods is based on 10 folds of equal size. For determining the area under the criterion curve and the average scores, we use a grid of asymmetry parameters consisting of 49 homogeneous points between 0 and 1. For the non-negative garrote, we used 49 points for the grid for ξ. Boosting is done with step size $\nu = 0.1$ and 10-fold in-bag cross-validation to find the optimal stopping iteration of the algorithm m_{stop} which has a maximum value of 4000. For the parallel design, we compare our methods with quantile boosting. Therefore we adjust the asymmetry parameters with the transformation function from expectiles to quantiles based on the $N(0,4)$ distribution. The resulting quantile levels for quantile boosting are $\{0.070, 0.127, 0.194, 0.291, 0.500, 0.709, 0.806, 0.873, 0.930\}$.

All covariates were included as cubic P-splines with 20 inner knots and a second order difference penalty. The smoothing parameter was estimated via the Schall algorithm (see Sobotka and Kneib (2012) and Schnabel and Eilers (2009) for further details). For a distinction between linear predictors and smooth func-

tions, we consider the different possibilities discussed in Section 3.1. Since the results for "restricted" and "complete" discrimination between linear and semiparametric predictors are similar, we will only discuss the "complete" separation in detail. The other results are presented in the supplementary material. The "complete" separation of the linear effect and the nonlinear deviation of the linear trend has the advantage that both can be selected independently by the algorithms. Thus it may appear that either both are included into the best model, or only one of them is included, or none of them is included into the best model. Therefore the sum of the selection frequencies of the linear and the nonlinear part of a covariate effect varies between 0 and 2.

All estimations and selection methods were implemented in R (R Core Team, 2017) using the R package `expectreg` (Sobotka et al., 2016). The applied version is available in the supplementary material.

4.2. Results

In the parallel design, the basic sensitivity of the selection methods is analyzed. In Figure 3, the frequencies of selection for the different coefficients are plotted depending on the asymmetry parameter.

The different characteristics of the selection approaches are discussed in the following, starting with the selection approaches for each asymmetry parameter separately.

- *Separate selection:*
  - *center vs. tails of the distribution:* Overall, AIC-based selection strongly depends on the current asymmetry parameter τ, while the non-negative garrote and boosting are less dependent on τ at least for the strong effects and CV is nearly independent of τ.
  - *informative vs. noise covariate:* The non-informative linear covariate x_4 is excluded for CV, non-negative garrote and boosting constantly in 80%-90% of the cases, while AIC excluded it in only 60%-70% of the replications. All approaches depend on the intensity of the informative covariates, i.e. the strong covariate x_1 is included almost always, while for the less influential covariate x_3 the general pattern of the dependence on the asymmetry parameter is most visible, even though the size of the parameter should be constant over all asymmetry parameters.
  - *linear vs. nonlinear covariate:* The data are simulated without any nonlinear part, so this should be excluded for all intensities of the covariates and all asymmetries. However the approaches behave differently concerning these nonlinear parts. While CV excludes the nonlinear part in most cases, boosting includes the nonlinear part relatively often for all asymmetry parameters. To the contrary, the non-negative garrote excludes the penalized part more likely in the center of the distribution than in the tails. This is similar for AIC

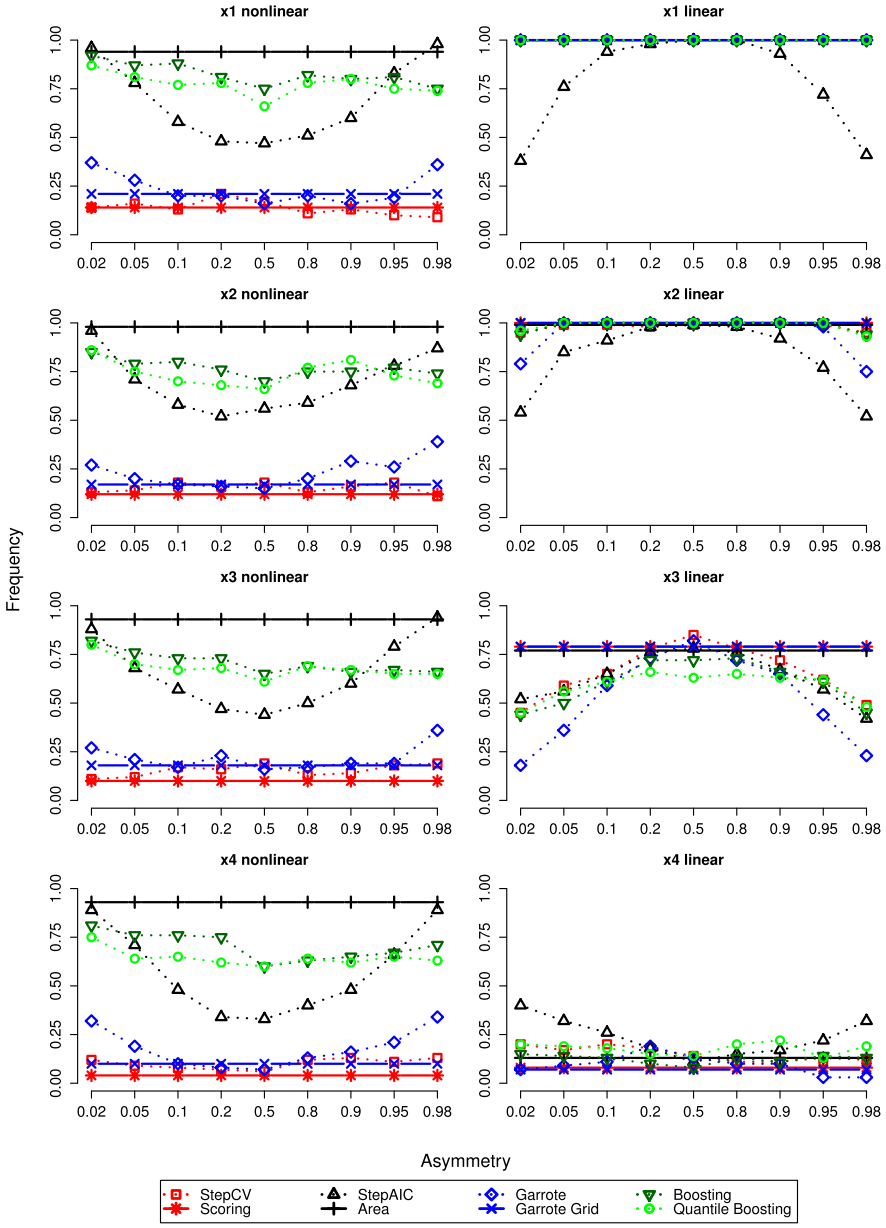


FIG 3. *Frequency of selected models for parallel design.*

selection, but there the frequency of excluding the noisy nonlinear part is rather low. So we can conclude that the known problem (see for example Greven and Kneib, 2010 and Saefken et al., 2014) of

nonlinear selection with AIC is even more problematic for expectiles beyond the mean.

- *Joint selection:* The joint approaches stabilize the selection results but the overall sensitivities of the separate selections methods stay the same. This means that joint CV, i.e. scoring includes few noise covariates and is most restrictive concerning nonlinear parts, while the area under the AIC curve includes the linear covariates on the level of scoring, it does nearly never exclude a nonlinear covariate. The non-negative garrote on the grid is also more stable, i.e. the exclusion rate of noisy linear covariates is also on the level of scoring, while the rate for the wiggly deviation is a little bit higher than for scoring, but much lower than for the AIC.
- *quantile boosting*: In this simulation study, the results of expectile boosting and quantile boosting with transformed asymmetry parameters are approximately the same.

The other design discussed here is the *exponential design*. Since the selection frequencies of the *linear design* behave similarly to the third design, their frequency plots can be found in the supplementary material (see Figure A.2). Other than in the *parallel design*, we cannot calculate the corresponding quantile levels for the heteroscedastic designs. Thus the results cannot be directly compared to quantile boosting.

As x_1 is a strong effect in the parallel case as well as in the one sided case, there are only minor changes in the behavior of the selection methods. The same is valid for the noise effect. Since the selection methods behave similarly as in the parallel setting, those frequencies are not shown in Figure 4, where the selection frequencies of the linear and nonlinear part of x_2 and x_3 are illustrated.

- *linear effect:* The strength of the small linear effect $x_3\beta_{3,\tau}$ is varying with the asymmetry parameter. Thus the selection frequency of the unpenalized part should increase with increasing asymmetry parameter (for the separate selection per asymmetry parameter). This behavior can indeed be detected in the plots. However, for the non-negative garrote, boosting and AIC, we observe a decrease in the selection frequencies for very large asymmetry parameters. This is caused by the general restriction of these methods in the tail of the distribution, similarly as detected in the parallel design.
- *nonlinear effects:* The nonlinear part of x_2 is also of major interest in this scenario. Using non-negative garrote and CV, the selection frequency for very small asymmetry parameters is not as high as expected. This is related to the fact that the selection frequency of the linear part is higher than expected. Thus the nonlinear trend is approximated with a linear function in 10–20% of the cases. With increasing asymmetry parameter, the selection frequency of the nonlinear part increases up to the 50% expectile. Beyond that, it is decreasing as expected for the non-negative garrote and CV. For AIC, the unpenalized part is selected into the best model more frequently on the upper part of the distribution, but this could

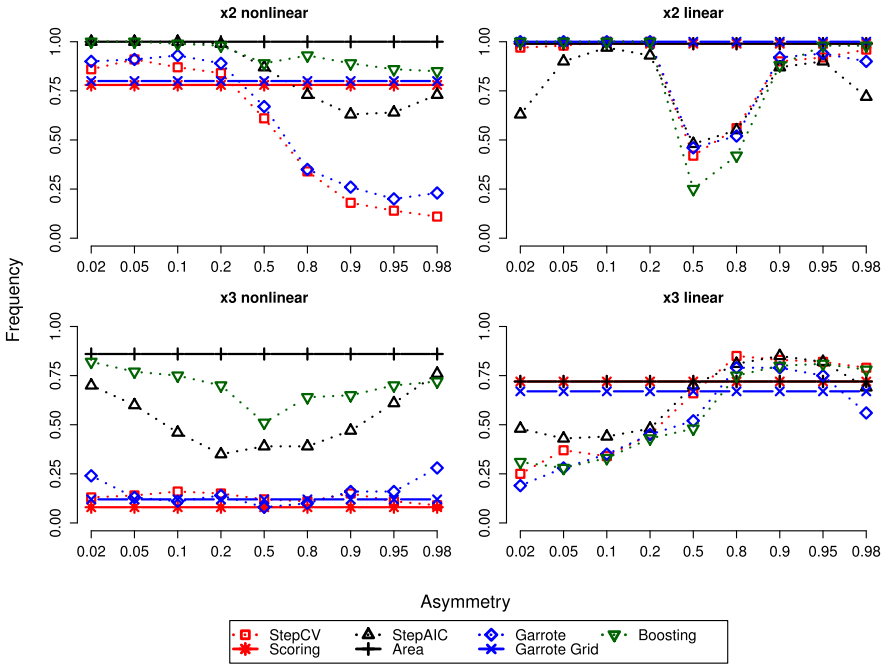


FIG 4. *Frequency of selected models for the exponential design.*

be expected from the parallel design. The unusual design of the selection frequency of the linear part of x_2 can be explained by the design of the data sets. For the upper part of the distribution, the linear part of x_2 and x_3 should be similar, which is confirmed by the simulations. For the bottom part, a higher selection frequency, as expected, is visible due to the usage of a linear trend in the estimation of the nonlinear part.

We conclude that for the selection of the linear part CV, boosting and non-negative garrote behave well, while for the nonlinear part only CV is approximately independent of the current asymmetry parameter. However the non-negative garrote is also reliable in the center of the distribution. The joint selection approach over all asymmetry parameters reduces linear parts of noise covariates for all selection types. Furthermore, it is advantageous for the selection of nonlinear parts with CV and non-negative garrote.

5. Application: Childhood malnutrition in Peru

5.1. Data structure

In the following, we apply the selection methods to a data set on childhood malnutrition in Peru. As an indicator for the nutritional status, we rely on the

stunting score

$$zscore_i = \frac{size_i - MedianSize_{(age_i, sex_i)}}{\sigma},$$

where the standardized height of a child is compared to the median size of children in a healthy comparison group given the child's age and gender and σ is a robust estimate for the standard deviation of the children's size in the comparison group. Stunting reflects chronic undernutrition which results in a lack of growth and a score of less than -2 indicates severe chronic undernutrition (WHO Expert Committee on Physical Status, 1995). As can be seen from Table 1, severe undernutrition occurs regularly but the mean nutritional status is above the cut-off value of -2.

TABLE 1
Distribution of zscores for children in Peru

Asymmetry	0.05	0.1	0.2	0.5	0.8	0.9	0.95
Expectile	-2.36	-2.06	-1.72	-1.15	-0.56	-0.23	0.08

When analyzing childhood malnutrition, it is particularly interesting to obtain covariates associated with the lower part of the stunting distribution since these are especially important determinants for chronic malnutrition. At the same time, discriminating between informative and uninformative covariates considerably assists in identifying sparse and interpretable models. Our analyzes rely on the *Demographic and Health Surveys* (*DHS*) data set of Peru from 2012 (Instituto Nacional de Estadistica e Informatica (INEI) Lima Peru and ICF International Calverton Maryland USA [Producer], 2012). Using only complete cases, we obtain a data set of $n = 8391$ children from the original 9620 observations.

From the original data base, we determined 21 potential covariates including (i) characteristics of the child such as age, gender, region of living (25 districts), duration of breastfeeding and birth order, (ii) characteristics of the mother such as age at birth, education, height and body mass index, (iii) household information such as the household size, the education of the partner, if there had been dead children in the family and (iv) variables related to the wealth of the family, i.e. several indicators for the presence/absence of specific assets.

Besides the estimation of effects beyond the mean, we would like to use flexible objects like P-splines or Gaussian Markov random fields to include nonlinear effects of continuous covariates and spatial effects. Those could also be estimated in quantile regression, but there it would be computationally demanding. Therefore we apply expectile regression where we can directly make use of the tools of standard least squares regression, thus we are computationally faster. For comparison, we also estimated a model based on quantile boosting but the results cannot be directly compared since the asymmetry levels of quantiles and expectiles are different. In the simulation study we knew the underlying distribution and could therefore correct the asymmetry parameters accordingly but this is not possible in the empirical example. Nevertheless, we applied quantile

boosting on the Peru data with asymmetry levels obtained when assuming Gaussian distributed data. The results are in general similar to the ones of expectile boosting and are therefore only included in the supplementary material.

As in the simulations, estimation is performed based on the LAWS criterion in combination with the Schall algorithm (see Sobotka and Kneib (2012) for further details). Spatial information was included as a Gauss Markov random field and the continuous covariates were included as cubic P-splines with 20 inner knots and a second order difference penalty. In the selection, the "complete" discrimination between linear trend and nonlinear deviation of this trend was used.

5.2. Selected models

We only discuss the results of stepwise forward 10-fold cross-validation (CV) and 10-fold scoring with a grid of 49 asymmetry parameters. The other results (stepwise forward selection with AIC, area under the AIC curve, non-negative garrote, non-negative garrote on the grid and boosting) are summarized as tables in the supplementary material.

Table 2 summarizes the selected covariates for CV and scoring. A black box indicates that the covariate is selected in the best model, while a blank signals that the covariate was not included. Obviously, several covariates are never or almost never included in the best model for the CV approach (e.g. bicycle, caesarian, electricity). These covariates are also not included in the best scoring model. Thus we can conclude that they do not have a relevant influence on the nutritional status. On the other hand, there are several covariates which are (almost) always included in the best CV model and also included in the scoring model (e.g. mother's education, refrigerator, television, region). Those covariates do have a relevant influence on the complete distribution of the nutritional status. Furthermore, there are several covariates that are selected in the scoring model, but only selected in some separate CV models (e.g. *sex, partner's education*). Those coefficients should then only be interpreted in the part of the distribution, where they are selected. Thus the sex of a child does only have an influence on the bottom part of the distribution, i.e. the undernourished children.

Additionally, there are covariates where no global trend is available, but a relevant deviation from the constant zero is detected, e.g. duration of breastfeeding. This means the influence of breastfeeding fluctuates around zero and has only a relevant size for the upper part of the distribution. Furthermore, the availability of a TV, a refrigerator, a motorcycle and a telephone are decisive indicators for the wealth of the family. The occurrence of a bicycle, a radio or electricity do not seem to explain the wealth of the family and thus the nutritional status of the child.

An advantage of interpreting the scoring approach as the area under the MWSE curve is that weights on more interesting parts of the distribution can be applied. In a second analysis, we applied a weight of 10 on all asymmetry parameters smaller than 0.11 to emphasize the importance of the lower part of the

TABLE 2
Selected covariates for stepwise forward selection with 10-fold cross-validation and scoring

Covariate	Type	0.02	0.05	0.1	0.2	0.5	0.8	0.9	0.95	0.98	grid
birth order		■	■	■	■	■	■	■	■	■	■
caesarian		■	■								
dead children		■									
household head											
household members			■	■	■	■	■	■	■	■	■
mother's education		■	■	■	■	■	■	■	■		■
partner's education						■	■	■	■	■	■
sex		■	■	■	■	■			■	■	■
bicycle											
electricity								■			
motorcycle			■	■	■	■	■	■	■	■	■
radio											
refrigerator		■	■	■	■	■	■	■	■	■	■
telephone			■	■	■	■	■	■	■	■	■
television		■	■	■	■	■	■	■	■	■	■
breastfeeding	linear										
breastfeeding	nonlinear					■	■	■	■	■	■
child's age	linear				■		■	■	■	■	■
child's age	nonlinear	■	■	■	■	■	■	■	■	■	■
mother's age	linear	■	■	■		■	■	■	■	■	■
mother's age	nonlinear		■	■	■	■			■	■	
mother's bmi	linear	■	■	■	■	■	■	■	■	■	■
mother's bmi	nonlinear										
mother's height	linear	■	■	■	■	■	■	■	■	■	■
mother's height	nonlinear	■	■								
region	GMRF	■	■	■	■	■	■	■	■	■	■

stunting distribution. The resulting scoring model does only differ marginally (see Table A.7 of the supplementary material). Here, the mother's age has a nonlinear effect in the best model. Due to the marginal differences, our unweighted optimal model also seems to be appropriate for the undernourished children.

An alternative to comparing the best scoring model with the separate CV model of all covariates would be to use the result of the scoring model and apply separate stepwise backward CV selections on the selected covariates. Then the selected covariates of the backward selection are treated as relevant for this part of the distribution. The result of this approach can be seen in Table A.8 in the supplementary material. However, the asymmetry parameters, where the covariates are excluded are very similar to the standard CV approach.

For a final comparison we also estimated bootstrap confidence intervals for the saturated model (Sobotka et al., 2013) and determined simultaneous confidence bands (SCB) following a method proposed in Krivobokova, Kneib and Claeskens (2010), which originally was designed to get simultaneous Bayesian credible intervals and is implemented in the R package `acid` (Sohn, 2016). Although the estimated confidence intervals provide more information about the variance of the estimated effects and are not designed for model selection (compare Burnham and Anderson, 2002), model selection was implemented by checking which confidence bands were completely covering the zero line. The resulting

"selection" table (Table A.6 in the supplementary material) is in general similar to the one of stepwise forward selection with 10-fold cross-validation (Table 2). However, minor differences appear and only the nonlinear part of the child's age is significant using the simultaneous confidence bands, while the other nonlinear functions are not significant.

5.3. *Effects*

In our analysis we first determine which covariates have a relevant influence on the response by model selection. Besides this abstract examination, the estimated effects also tell us the influence on the response. In Figure 5, 6 and 7 the estimated effects of the model optimized via scoring are plotted.

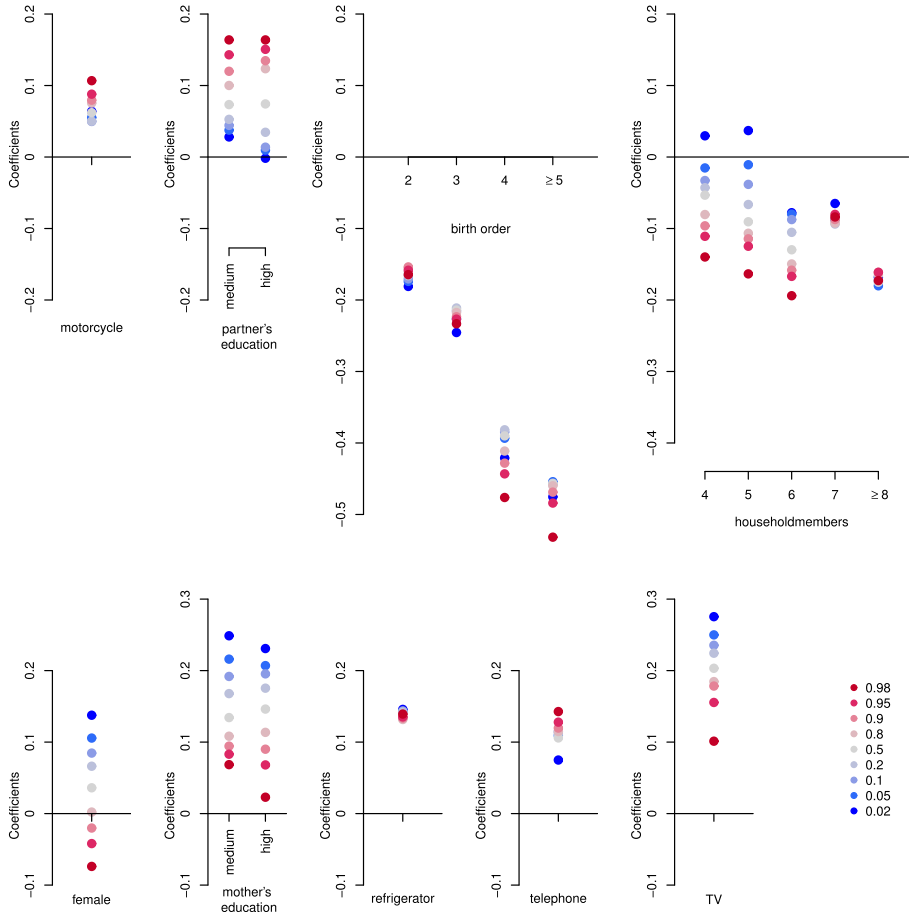


FIG 5. *Estimated coefficients of categorical covariates selected via scoring.*

The categorical covariate effects are visualized in Figure 5. Here we see that being a female is associated with an increase of the nutritional status for the bottom part of the distribution. Having a TV is also associated with an increase in the nutritional status more strongly on the bottom part of the distribution. On the other side, having a telephone has a larger influence on the upper part of the distribution. Moreover a higher eduction of the mother always increases the zscore, but is stronger for undernourished children. Aside from that, children in families with many members tend to have a worse zscore. This decreases even more if they are not the first child of this mother.

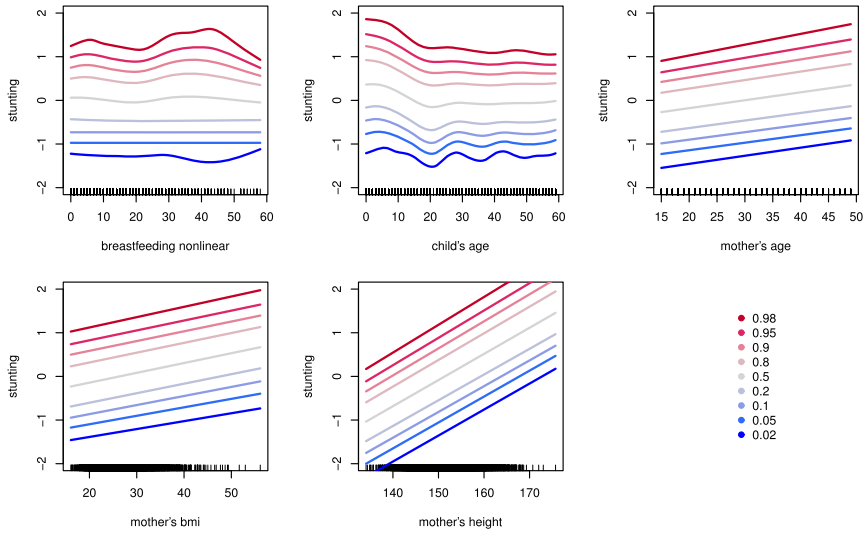


FIG 6. *Estimated coefficients of metric covariates selected via scoring.*

For the effects of continuous covariates shown in Figure 6, we find that the mother's age, bmi and height have a linear influence on the response. In addition, the estimated effects of these three covariates are almost parallel when comparing different asymmetry parameters such that their influence is approximately the same for all nutritional statuses. With increasing age, height and bmi, the zscore of the child will tend to be higher. Here the effect of bmi indicates the current nutritional status of the whole family, while the mother's height influence the zscore in two ways. First genetically, since large mothers will have large children and the zscore is a linear function of the child's height. Secondly, if the mother did not have enough food as a child she stayed smaller and often poor families stay poor, such that the child also does not get enough food. The child's age does have a strong influence on the nutritional status. Here we see a steep decrease between ten and 18 months. Afterwards the child's zscores are not affected by age anymore. This steep decrease is visible for all parts of the distribution, while it is considerably more expressed for the lower part of the stunting distribution.

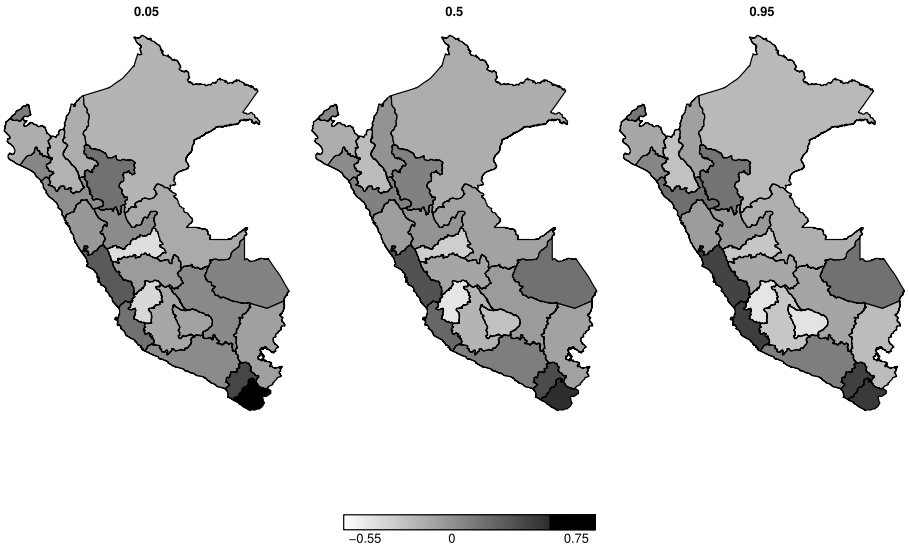


FIG 7. *Some estimated coefficients of regional information estimated via scoring.*

Finally, the regional effects of the 5%, 50% and 95% expectile are shown in Figure 7 (for the other asymmetry parameters see Figure A.25 in the supplementary material). For all asymmetry parameters, the regional distribution is nearly the same. Regions next to the Chilean border in the south and next to the capital Lima, on the middle of the Pacific coast line, are associated with a higher nutritional status. Furthermore, the zscore of children living in the Andes is generally found to be lower.

6. Conclusions

In this paper, we considered several approaches for model selection in semiparametric expectile regression differentiating between the separate selection for specific asymmetries and the selection of covariates for the complete response distribution. Moreover, we allowed for the separation between linear and nonlinear effects for continuous covariates. Generally we show that model selection strongly depends on the asymmetry parameter. Thus the joint model selection for the whole distribution may be advantageous if indeed the complete distribution of the response shall be analyzed. This approach also reduces the number of included noise covariates.

Overall, the different selection methods all have their advantages. Stepwise AIC selection is the most intuitive method and computationally moderately fast. It is reliable for the selection of linear predictors. However, it selects smooth predictors rather poorly. To the contrary, CV is restrictive in selecting linear and smooth predictors, but it is computationally the most demanding approach.

Here the non-negative garrote has its big advantage, because it is computationally much faster. The selection properties of the non-negative garrote are not as good as for CV, but it is a reliable alternative. Finally, boosting is a good way for selecting linear predictors, but it is not restrictive enough for smooth alternatives.

In summation, it is possible to decide if a spline is necessary by splitting the spline into its linear trend and the wiggly deviation of this trend. Those terms can then be selected separately with CV or the non-negative garrote. Again the joint selection approaches stabilizes the performance but for the area under the AIC curve it takes the upper bound, while for scoring and non-negative garrote on a grid the lower one.

Our example shows that our model selection approaches work for analyzes beyond the mean with a medium number of covariates.

Besides the approach based on the mixed model decomposition of P-splines, one could imagine further methods to check if the smooth part is necessary. So could a variable be transformed in a sequence of basis functions on which a L_1 penalty could be applied to select those which are relevant.

A sensible model selection technique is not only necessary to control the number of included covariates, but also overall model complexity. So far we can decide on how to include a continuous covariate, either restricted to a linear function or flexibly modeled by a P-spline basis. However, we find more and more spatio-temporal regression models. While an interaction between spatial / regional information and additional information for multiple points in time can offer a very flexible model, there is also a strong increase in model complexity. Especially if we aim to estimate more than the mean it will be very helpful to have a reliable measure to decide on additional model complexity.

Acknowledgements

We acknowledge financial support by the German Research Foundation (DFG), grant KN 922/4-2. Comments by an anonymous referee were very helpful in significantly improving the initial submission of this paper.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** 716–723. MR0423716

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384. MR1365720

Bühlmann, P. and Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science* **22** 477–505. MR2420454

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Verlag, New York. MR1919620

Chouldechova, A. and Hastie, T. (2015). Generalized Additive Model Selection. *arXiv preprint arXiv:1506.03850*.

Currie, I. and Durban, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modeling* **2** 333–349. MR1951589

Instituto Nacional de Estadistica e Informatica (INEI) Lima Peru and ICF International Calverton Maryland USA [Producer] (2012). Peru Demographic and Health Survey 2012 [Dataset]. `PEKR6IFL.SAV`. *ICF International [Distributor]*.

Doksum, K. and Koo, J.-Y. (2000). On spline estimators and prediction intervals in nonparametric regression. *Computational Statistics & Data Analysis* **35** 67–82. MR1815574

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11** 89–121. MR1435485

Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica* **14** 715–745. MR2087971

Fenske, N., Kneib, T. and Hothorn, T. (2011). Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression. *Journal of the American Statistical Association* **106** 494–510. MR2847965

Gijbels, I., Verhasselt, A. and Vrinssen, I. (2015). Variable selection using P-splines. *Wiley Interdisciplinary Reviews: Computational Statistics* **7** 1–20. MR3348718

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106** 746–762. MR2847988

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102** 359–378. MR2345548

Gneiting, T. and Ranjan, R. (2011). Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics* **29** 411–422. MR2848512

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **97** 773–789. MR2746151

Guo, C., Yang, H. and Lv, J. (2015). Robust variable selection in high-dimensional varying coefficient models based on weighted composite quantile regression. *Statistical Papers* 1–25. MR3304007

Guo, J., Tang, M., Tian, M. and Zhu, K. (2013). Variable selection in high-dimensional partially linear additive models for composite quantile regression. *Computational Statistics & Data Analysis* **65** 56–67. MR3064943

He, X. and Ng, P. (1999). Quantile splines with several covariates. *Journal of Statistical Planning and Inference* **75** 343–352. MR1678981

Hofner, B., Mayr, A., Robinzonov, N. and Schmid, M. (2014). Model-based boosting in R: a hands-on tutorial using the R package mboost. *Computational Statistics* **29** 3–35. MR3260108

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of statistics* **38** 2282. MR2676890

Jiang, L., Bondell, H. D. and Wang, H. J. (2014). Interquantile shrinkage and variable selection in quantile regression. *Computational Statistics & Data Analysis* **69** 208–219. MR3146889

Kai, B., Li, R. and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Annals of statistics* **39** 305. MR2797848

Koenker, R. (2011). Additive models for quantile regression: model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics* **25** 239–262. MR2832886

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society* **46** 33–50. MR0474644

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* **94** 1296–1310. MR1731491

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680. MR1326417

Krivobokova, T., Kneib, T. and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association* **105** 852–863. MR2724866

Li, Y. and Zhu, J. (2008). L1-Norm Quantile Regression. *Journal of Computational and Graphical Statistics* **17** 163–185. MR2424800

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61** 381–400. MR1680318

Lin, C.-Y., Bondell, H., Zhang, H. H. and Zou, H. (2013). Variable selection for non-parametric quantile regression via smoothing spline analysis of variance. *Stat* **2** 255–268.

Lv, J., Yang, H. and Guo, C. (2015). Smoothing combined generalized estimating equations in quantile partially linear additive models with longitudinal data. *Computational Statistics* 1–32. MR3528651

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis* **55** 2372–2387. MR2786996

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society* **55** 819–847. MR0906565

Noh, H., Chung, K., Van Keilegom, I. et al. (2012). Variable selection of varying coefficient models in quantile regression. *Electronic Journal of Statistics* **6** 1220–1238. MR2988445

WHO Expert Committee on Physical Status (1995). Physical status: The use and interpretation of anthropometry. *WHO technical report series* **854**.

Saefken, B., Kneib, T., van Waveren, C. and Greven, S. (2014). A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. *Electronic Journal of Statistics* **8** 201–225. MR3178544

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78** 719–727.

Schnabel, S. K. and Eilers, P. H. C. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis* **53** 4168–4177. MR2744314

Schulze-Waltrup, L., Sobotka, F., Kneib, T. and Kauermann, G. (2015). Expectile and Quantile Regression – David and Goliath? *Statistical Modelling* **15** 433–456. MR3403125

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics* **6** 461–464. MR0468014

Sobotka, F. and Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis* **56** 755–767. MR2888723

Sobotka, F., Kauermann, G., Schulze-Waltrup, L. and Kneib, T. (2013). On confidence intervals for semiparametric expectile regression. *Statistics and Computing* **23** 135–148. MR3016934

Sobotka, F., Schnabel, S., Schulze-Waltrup, L., Eilers, P., Kneib, T. and Kauermann, G. (2016). expectreg: Expectile and Quantile Regression R package version 0.50.

Sohn, A. (2016). acid: Analysing Conditional Income Distributions R package version 1.1.

Tang, Y., Wang, H. J. and Zhu, Z. (2013). Variable selection in quantile varying coefficient models with longitudinal data. *Computational Statistics & Data Analysis* **57** 435–449. MR2981100

Tang, Y., Wang, H. J., Zhu, Z. and Song, X. (2012). A unified variable selection approach for varying coefficient models. *Statistica Sinica* **22** 601–628. MR2954354

R Core Team (2017). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288. MR1379242

Wang, H. J., Zhu, Z. and Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *The Annals of Statistics* **37** 3841–3866. MR2572445

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica* **19** 801–817. MR2514189

Wu, C. and Ma, S. (2015). A selective review of robust variable selection with applications in bioinformatics. *Briefings in Bioinformatics* **16** 873–883.

Yao, Q. and Tong, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics* **6** 273–292. MR1383055

Zou, H. and Yuan, M. (2008a). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics* **36** 1108–1126. MR2418651

Zou, H. and Yuan, M. (2008b). Regularized simultaneous model selection in multiple quantiles regression. *Computational Statistics & Data Analysis* **52** 5296–5304. MR2526595