Environ Ecol Stat (2013) 20:519–531
DOI 10.1007/s10651-012-0232-1

Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology

Vito M. R. Muggeo · Mariangela Sciandra · Agostino Tomasello · Sebastiano Calvo

Received: 19 January 2012 / Revised: 29 October 2012 / Published online: 1 January 2013

Abstract We discuss a practical and effective framework to estimate reference growth charts via regression quantiles. Inequality constraints are used to ensure both monotonicity and non-crossing of the estimated quantile curves and penalized splines are employed to model the nonlinear growth patterns with respect to age. A companion R package is presented and relevant code discussed to favour spreading and application of the proposed methods.

Keywords Growth charts · Nonparametric regression quantiles · Penalized splines · P. oceanica modelling · R software

1 Introduction

Reference growth charts represent an essential mean to analyse and to monitor the well-being of the biological populations, communities or ecosystems more generally. For instance, in public health the growth charts are a well-known tool to assess the development of some target population, typically the children, and to establish whether a generic individual at a given age lies within the 'normal' range. Examples of typical growth charts include the analysis of rough measures, such as high or weight, or anthropometric indices, such as the body mass index where the corresponding percentiles are used to study the overweight and obesityfor a specific age (Hedley et al.

Handling Editor: Ashis SenGupta.

V. M. R. Muggeo (✉) · M. Sciandra
Dipartimento Scienze Statistiche e Matematiche 'Vianelli', Università di Palermo, Palermo, Italy
e-mail: vito.muggeo@unipa.it

A. Tomasello · S. Calvo
Dipartimento Scienze della Terra e del Mare, Università di Palermo, Palermo, Italy

2004). In Ecology, application and use of reference growth charts is far from being commonplace. Growth of biological systems is usually modelled by means of some known regression equation, for instance the Von Bertalanffy growth curve, relating to age the mean rather than the quantiles. From a statistical perspective, growth charts are curves relating the quantiles of the distributions of the response growth variable to age; for this reason these are more precisely referred as age-dependent growth charts (e.g. Wei et al. 2006). Estimation of age-dependent growth charts can be obtained via two major alternative methods. The LMS method (Cole and Green 1992) is based on Box Cox-type transformations of the observed data to obtain the desired 'target' distribution, such as the Normal, t or power exponential (Rigby and Stasinopoulos 2004, 2005). Different sub-regression models for the mean, variance and other additional parameters for the transformed responses are fitted by maximizing the penalized log likelihood from the target distribution, and estimates of the quantiles for the original variable are obtained accordingly. The LMS is very popular: it has been successfully used by the World Health Organization (Borghi et al. 2006) and it is currently implemented in several statistical packages such as R, SAS or Stata.

An alternative and probably more direct approach to obtain growth charts is via quantile regression, hereafter QR (Koenker 2005). QR models directly the observed data for the selected percentiles, it does not assume any distribution for the response, and it is particularly valuable in presence of heteroscedasticity that is typical of growth data. QR involves minimization of objective functions based on absolute values and thus, unlike the LMS method, it is more resistant to outliers and to possibly influential observations. Wei et al. (2006) discuss advantages of QR over the LMS by comparing reference growth charts obtained by both methods when additional confounding variables are included in the model as linear terms.

Regardless of the approach, the major concern in obtaining reliable estimates of the quantiles involves the following issues: i) flexibility, i.e. each fitted quantile curve should be flexible enough to account for the nonlinear growth pattern over age; ii) monotonicity, i.e. each fitted quantile curve should be non-decreasing over age; iii) non-crossing, i.e. the fitted quantile curves should not intersect each other. Crossing quantile curves is an annoying problem that could be possibly enhanced with nonlinear relationships, moderate samples and sparse/scarce data. The LMS method implicitly leads to non crossing curves via a scaling function, while QR needs some extra work to return non-crossing quantile curves. He (1997) discusses restricted QR based on location scale model with a proper scaling function preventing crossing; however the model, albeit sufficiently flexible, cannot capture adequately deviations from the assumed restricted model. Dette and Volgushev (2008) use an indirect approach by first estimating the conditional cumulative distribution function via local weighting, and then by inverting it to obtain the quantile curve. Their approach, however, does not allow to quantify the effect of covariates and therefore it can be of limited use in practice. Recently Bondell et al. (2010) propose an elegant approach based on simultaneous estimation of the conditional quantiles under inequality constraints. After reparameterization, the task is reduced to a sparse linear programming problem. Unfortunately the authors do not discuss monotonicity restrictions and therefore growth charts cannot be obtained easily; moreover they use simple regression B-splines to account for nonlinearity without discussing about number and location of knots. It is well known

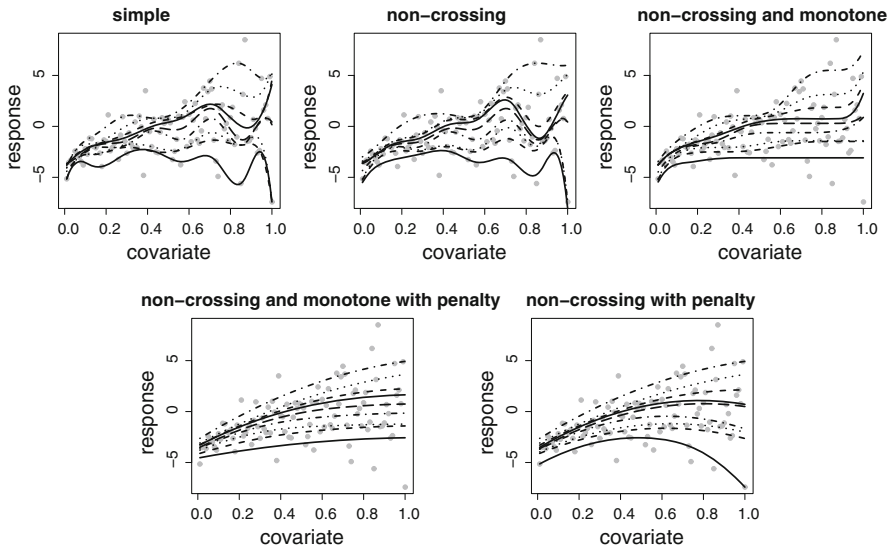

Fig. 1 Some examples of estimated quantile curves for $n = 100$ simulated data with/without monotonicity restrictions, with/without non-crossing constraints and with/without penalty term

that number and locations of knots is the key-point in spline modelling and typically two strategies are undertaken to bypass this important issue: knot removing (e.g. He and Shi 1998) or penalizing the spline coefficients (e.g. Koenker et al. 1994; Bosch et al. 1995; Bollaerts et al. 2006).

In this paper we propose an alternative and practical approach to estimate non-crossing conditional quantiles while ensuring flexibility and monotonicity over age. The goal is to provide a relatively simple algorithm that can be easily implemented into statistical softwares, such as R; at this aim we also discuss software availability and illustrate code of our new R package quantregGrowth to yield growth charts at specified percentiles. Figure 1 portrays some examples of quantile curves fitted on simulated data via the quantregGrowth package illustrated in Sect. 5.

2 Methods: estimation of noncrossing and monotone quantile curves

For the response random variable Y and percentile $\tau_k \in (0, 1)$ we consider the following quantile regression model

$$Q_{\tau_k}(Y|z_i, x_i) = x_i^T \beta_{\tau_k} + s_{\tau_k}(z_i) \tag{1}$$

where β_{τ_k} quantifies the linear effect of p covariates and $s_{\tau_k}(z_i)$ is a smooth but unspecified function relating the quantiles to the age variable z. $s_{\tau_k}(\cdot)$ describing the age-specific growth charts in corresponding of selected percentiles τ_k is expressed via B-splines, $s_{\tau_k}(\cdot) = \sum_j^J b_{jk} B_j(\cdot)$, namely a linear combination of the J basis

functions and the corresponding coefficients b_j to be estimated (He and Shi 1998; Bollaerts et al. 2006; Wei et al. 2006). Using B-splines allows to write model (1) in linear form $x_i^T \beta_{\tau_k} + B(z_i)^T b_k$ where the known basis functions can be considered as additional covariates entering the model in a linear way. Notice that the model intercept corresponding to the vector of ones is not requested as it is already accounted by the B-spline basis.

For the sake of simplicity, the jth basis function of the B-spline for the variable z will be indicated simply by B_j; moreover we will use Q_k, β_k and $s_k(\cdot)$ to mean $Q_{\tau_k}(Y|z_i, x_i)$, β_{τ_k}, and $s_{\tau_k}(\cdot)$ respectively. Also where necessary we will use $\theta_k = (\beta_k^T, b_k^T)^T$ to mean all the $(p + J)$ model parameters corresponding to the covariate vector $w_i = (x_i^T, B_i^T)^T$, thus the objective function to be minimized can be written as $\sum_i \rho_k(y_i - w_i^T \theta)$, where $\rho_k(u) = u(\tau_k - I(u < 0))$ is the so-called check function. In the next subsections we will describe how to modify this objective function to obtain non-crossing estimates of the Q_ks along with a flexible estimate of $s_k(\cdot)$ having monotonicity constraints.

2.1 Non-crossing curves

Let $Q_1, Q_2 \ldots, Q_{2K+1}$ be the regression quantiles corresponding to the ordered percentiles $\tau_1, \ldots, \tau_{2K+1}$ with generic regression equation $Q_k = w^T \theta_k$. Non-crossing means that for any couple of adjacent quantile curves Q_{k+1} and Q_k, it holds $Q_{k+1} > Q_k$ that implies $\theta_{k+1} > \theta_k$. Therefore given Q_k, the 'next' non-crossing Q_{k+1} can be estimated by imposing standard linear inequality constraints on relevant θ_{k+1} parameters. Linear inequality constraints are generically written in terms of a known matrix R and vector r, i.e. $R\theta_{k+1} \geq r$ (Koenker 2005, p. 214), thus to prevent crossing we set $R = I$ the identity matrix, and $r = \hat{\theta}_k$ the assumed known parameter vector of the 'reference' quantile curve Q_k. In the same way, given Q_k, the 'previous' non-crossing curve Q_{k-1} is obtained by considering the constraints $\theta_{k-1} < \hat{\theta}_k$, namely $R = -I$ and $r = -\hat{\theta}_k$. Therefore to prevent crossing across the fitted curves, an intuitive approach is to fit the different regression quantiles by imposing sequentially linear inequality constraints on the adjacent curve coefficients. For ease of exposition and without losing in generality we consider the 'centered' percentile indices $k = -K, -K+1, \ldots, -1, 0, 1, \ldots, K-1, K$, by assuming that the number of 'higher' and 'lower' quantiles is the same. Hence the algorithm can be schematized as follows

0. Initialize: Fit the model for Q_0, i.e. the QR for $k = 0$, and 'save' the relevant coefficients $\hat{\theta}_0$;
1. For $k = 1, \ldots, K$ ('higher' quantiles) fit the model $Q_k = w^T \theta_k$ with linear inequality constraints $\theta_k > \hat{\theta}_{k-1}$;
2. For $k = -1, \ldots, -K$ ('lower' quantiles) fit the model $Q_k = w^T \theta_k$ with linear inequality constraints $\theta_k < \hat{\theta}_{k+1}$.

The algorithm estimates a QR model for a starting percentile τ_0 and estimates the remaining quantile curves sequentially; lower ($\tau < \tau_0$) or higher ($\tau > \tau_0$) quantile curves are separately estimated starting from τ_0, and using at each step the parame-

ter estimates of the previous quantile curve as reference value for preserving non-crossing. While any percentile could be used as starting point, we suggest to use $\tau_0 = 0.5$ (or the percentile closest to 0.5) since the median has the lowest variance and therefore it represents the best estimate on which to 'anchor' the estimating procedure.

The aforementioned linear inequality constraints can be accounted straightforwardly when optimizing the L_1 objective function; for instance in R package quantreg by R. Koenker (Koenker 2011), these constraints are easily included in the function `rq()` by specifying `method=''fnc''` and supplying the matrix R and the vector r via the arguments `R` and `r`. Note, however, the linear inequality implemented in R are not strict, i.e. they are '$\geq$' rather than '$>$', leading to $Q_{k+1} \geq Q_k$ rather than $Q_{k+1} > Q_k$. Thus in order to guarantee strict inequality we simply add a small constant, namely we consider $\theta_k \geq (\hat{\theta}_{k-1} + .0001)$ for higher percentiles, and $\theta_k \leq (\hat{\theta}_{k+1} - .0001)$ for lower curves.

2.2 Monotone estimates

When B-splines are employed to express the nonlinear function $s_k(\cdot)$, monotonicity restrictions on the fitted curve may be accounted for straightforwardly. In fact the B-spline coefficients measure the amplitude of the bases, and therefore a sufficient condition to ensure monotonicity is that first-order differences of adjacent coefficients are non-negative, namely for the kth quantile curve $b_{k,j+1} - b_{k,j} \geq 0$ for $j = 1, \ldots, J-1$. These are again standard linear inequality constraints such as $Db_k \geq 0$ where D is the $(J-1) \times J$ first-order difference matrix

$$D = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ldots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & -1 & 1 \end{bmatrix}.$$

These linear inequality constraints can be included easily in the linear programming algorithms as discussed in the previous subsection; more specifically non-crossing quantile curves with monotonicity restrictions on the spline coefficients require the overall R matrix and vector r given by

$$R = \begin{bmatrix} I_{(p+J)\times(p+J)} \\ --- \\ 0_{(J-1)\times p} | D_{(J-1)\times J} \end{bmatrix} \quad r = \begin{bmatrix} \hat{\beta}_{p\times 1} \\ \hat{b}_{J\times 1} \\ 0_{(J-1)\times 1} \end{bmatrix},$$

where the subscripts have been added to clarify dimensions of vectors and matrices. As discussed in the previous subsection these constraints can be supplied via the arguments `R` and `r` in `rq()`.

2.3 Smoothing via P-splines

While nonparametric smoothing in QR can be performed using different techniques, we focus on flexible estimation based on the splines. The simplest approach is to use a B-spline basis with a fixed number of knots (Wei et al. 2006; Bondell et al. 2010); however this approach strongly depends on the number and position of knots and therefore it is not recommended in general, especially when there exist outlying regions with scarce and sparse data. Penalization is an effective remedy to deal with such 'unlucky' configuration of data and to bypass issues related to knot selection. Bosch et al. (1995) formulate the problem in terms of cubic smoothing splines with associated quadratic penalty $\int\{s''\}^2$ and make use of an interior point algorithm to solve the associated quadratic programming optimization problem. Koenker et al. (1994) also use smoothing splines but with a 'total variation penalty' on s', see also Ng (1996) for computational details and Koenker (2005, p. 222–248) for a wider discussion. Smoothing splines with a total variation penalization of the first derivative, implemented by the `qss()` function within `rqss()`, always produce a piecewise linear fit. This feature can be useful in some circumstances, for instance when interest lies in detecting breakpoints of the segmented relationships (Muggeo 2003), however when the underlying relationship is understood to be smooth, as in growth modelling, a piecewise fit is undesirable. Bollaerts et al. (2006) also employ a L_1 penalty on the first order differences of the B-spline coefficients, but while the final fit is smooth the resulting extended linear programming algorithm requires some additional programming effort and thus it is not favored in practice.

Similarly to Bollaerts et al. (2006) we also use a low-rank B-spline but with an L_2 penalty, namely a roughness measure given by the sum of d-order squared differences. The penalized objective function is

$$\sum_i \rho_k(y_i - w_i^T\theta_k) + \lambda \sum_{j=1}^{J-d} (\Delta^d b_k)_j^2, \tag{2}$$

possibly subject to the aforementioned inequality constraints to ensure monotonicity and non-crossing. $\Delta^d b_k$ means the d-order difference vector, for instance $\Delta^1 b_{kj} = (b_{k,j+1} - b_{kj})$, and λ is the tuning parameter regulating the amount of smoothing as in the usual mean regression. At $\lambda = 0$ a very wiggly curves is obtained, while as $\lambda \to \infty$ the fitted curve approaches to a polynomial of degree $d - 1$.

Notice we penalize the spline coefficients b_k, while possible inequality constraints refer to the whole parameter vector $\theta_k = (\beta_k^T, b_k^T)^T$ as illustrated previously.

Minimization of objective (2) looks awkward. However it turns out that it suffices to augment the design matrix and the response vector, namely

$$X^* = \begin{bmatrix} X_{n\times p} & B_{n\times J} \\ 0_{p\times p} & 0_{p\times J} \\ 0_{(J-d)\times p} & \sqrt{\lambda} D_{d,(J-d)\times J} \end{bmatrix} \qquad y^* = \begin{bmatrix} y_{n\times 1} \\ 0_{(J-d+p)\times 1} \end{bmatrix}, \tag{3}$$

where X, B, and D_d are respectively the design matrix of the p linear covariates, the rank-J B-spline basis, and the d-order difference matrix such that $b^T D_d^T D_d b = \sum_{j=1}^{J-d} (\Delta^d b_{kj})^2$ is the penalty term in (2). Therefore for given λ, estimation of QR model via the penalized objective function (2) is quite straightforward: it suffices to run a standard QR with response y^* and design matrix X^*. It is worth noting that a L_1 minimization of transformed variables results in a L_2 penalization. To clarify this point, we exploit the approximating framework described in Muggeo et al. (2012); here the absolute value function is replaced by a parametric smooth approximation depending on a constant c that leads to a L_2 objective function having an iterative weighted least square solution $\hat{\theta}(c) = (X^T W(c) X)^{-1} X^T z(c)$ where $W(c)$ and $z(c)$ are a proper weight matrix and working response vector both depending on c; as $c \to 0$ the approximate solution goes to the exact QR solution. The approximating L_2 objective function favours straightforward inclusion of a L_2 penalty with corresponding solution $\hat{\theta}_\lambda(c) = (X^T W(c) X + \lambda D_d^T D_d)^{-1} X^T z(c)$ equivalent to $(X^{*T} W^*(c) X^*)^{-1} X^{*T} z^*(c)$ where the stars indicate properly augmented versions. Thus a QR with augmented design matrix and response vector (3) is the limit of $\hat{\theta}_\lambda(c)$ as $c \to 0$ and it is actually a solution of the objective function (2).

When the model *degrees of freedom* (df) are available as a measure of model complexity, generalized cross validation, Akaike-type or Schwartz information criteria can be used (Koenker et al. 1994; Yuan 2006). However quantifying model dimension in penalized quantile regression is a difficult task when the penalty terms does not involve the 'natural' L_1 norm of Koenker et al. (1994). While one could exploit the work of Meyer and Woodroofe (2000) to compute $df = \sum_i \frac{\partial \hat{Q}_i}{\partial y_i}$, the computational burden could result important as numerical derivatives have to be used. A simpler yet consolidated approach is the G-fold cross validation that does not need the model df. The entire dataset is split into G subsets (folds) and each time $G-1$ subsets are used to fit the model and the remaining subset is employed to validate it via the cross validation statistic. For fixed λ the procedure is repeated G times, and the average 'error' across the G trials is computed. Running the process for a fine grid of candidate λ values allows to seek the best value minimizing the cross validation.

3 Application: growth curves for seagrass

For illustrative purposes we apply the proposed framework to estimate growth charts for the seagrass *Posidonia oceanica*. *P. oceanica* represents the key specie of the most important and productive ecosystem in subtidal habitats of the Mediterranean Sea. A peculiarity of *P. oceanica* is the presence of persistent morphological features characterizing its growth, that allows to obtain time determination, i.e. shoot age, associated with the growth history via back-dating techniques (Duarte et al. 1994). Being sensitive to changes in the environment, *P. oceanica* is considered a crucial indicator of the quality of coastal marine waters (Tomasello et al. 2007). It is therefore important to monitor the well-being of *P. oceanica* meadows and at this aim growth-charts represent a valid tool.

We analyse $n = 902$ plants sampled in 2000 in the central Mediterranean basin with non missing information. We consider two 'growth' variables, rhizome length

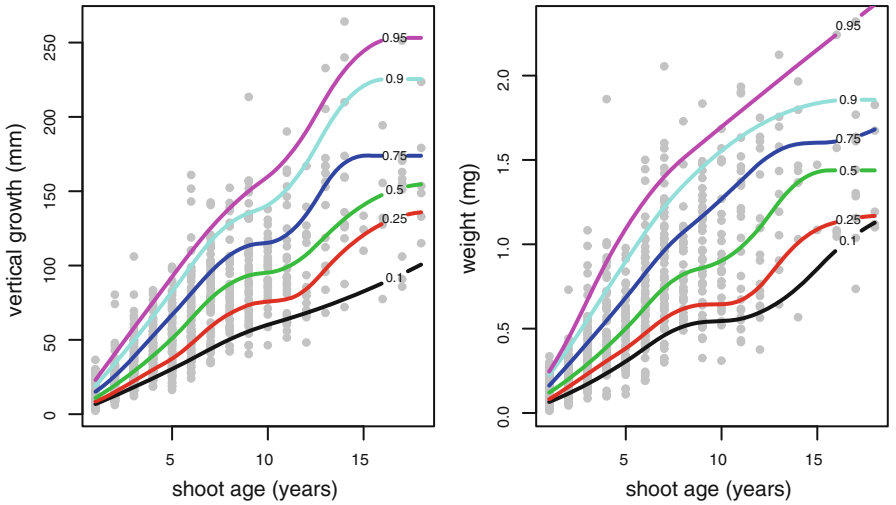

Fig. 2 Growth charts for height and weight of *P. oceanica* at specified percentiles displayed on the *right side* of each curve

(vertical growth) and total rhizome weight with respect to the shoot age. We apply our framework and perform the analyses using our R package quantregGrowth illustrated later. We employ a large B-spline basis ($J = 33$) with non-crossing and monotonicity constraints by imposing a penalty term on the spline coefficients to prevent under-smoothing. Moreover, since data are available from plants sampled at different depths, we also include in the predictor two linear variables with unpenalized coefficients representing a quadratic polynomial for the depth. The quadratic polynomial allows to assess possible better growth performance of the plants at middle depth. Therefore the regression equation for the response variable Y (growth or weight) and percentile τ_k is

$$Q_{\tau_k}(Y|\texttt{age}, \texttt{depth}) = \beta_{k,1}\texttt{depth} + \beta_{k,2}\texttt{depth}^2 + s_k(\texttt{age}) \tag{4}$$

where $s_{\tau_k}(\texttt{age}) = \sum_{j=1}^{33} b_{kj} B_{kj}(\texttt{age})$ and the spline coefficients b_{kj}s are penalized to obtain a smooth curve. Ten fold cross validation is used to select the best smoothing parameter separately for vertical growth and weight.

Figure 2 portrays the fitted quantile curves at percentiles $\tau_k = \{0.10, 0.25, 0.50, 0.70, 0.90, 0.95\}$ for plants at $\texttt{depth} = 8.9$ m.

4 Simulation experiments

The goal of simulation study is twofold. In the first experiment we focus on the non-crossing constraint, and in the second one we investigate on the penalty norm applied to the spline coefficients.

To assess how the proposed sequential approach performs with respect to the simultaneous approach of Bondell et al. (2010) to ensure non-crossing curves, we use two

different settings taken from Bondell et al. (2010), example 4 and example 5 of their paper. Both of them are special cases of the heteroskedastic error model, namely $y_i = f(x_i) + g(x_i)\epsilon_i$ with independent standard normal errors i.e. $\epsilon_i \sim \mathcal{N}(0, 1)$ and covariate $x_i \sim \mathcal{U}(0, 1)$. The first setting refers to a nonlinear mean function defined as $f(x_i) = 0.5 + 2x_i + \sin(2\pi x_i - 0.5)$ and variance function assumed to be $g(x) = 1$; this will result in parallel quantile curves $Q_\tau(x_i)$ shifting in the intercept. In the second setting $f(x_i) = 3x_i$ and the variance function defined as $g(x_i) = 0.5 + 2x_i + \sin(2\pi x_i - 0.5)$. The resulting quantile curves will be characterized by a linear median and nonlinear curves for the remaining percentiles with a different degree of smoothing. For each scenario and method, we use a B-spline basis with 6 degrees of freedom without penalization and monotonicity constraints. In fact the aim of the simulation studies is to assess how the proposed sequential approach performs with respect to the simultaneous one, and therefore we use the same B-spline basis without discussing the choice of the dimension, degree of the spline and number and position of knots. For each simulation scenario quantile curves are estimated at percentiles $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$ for three different sample sizes, $n = 50, 100$, and 500. Table 1 shows the results in terms of MISE (mean integrated squared error) given by $n^{-1}\sum_i(\widehat{Q}_\tau(x_i) - Q_\tau(x_i))^2$ for each percentile τ. As a benchmark we also include results from fitting QR curves separately, although this approach does not ensure noncrossing and therefore is not appropriate for estimating reference growth charts. Standard quantile regression (QR) is estimated via the `rq()` function while the joint estimation of Bondell et al. (2010) (NCQR) is implemented via the R function available at http://www4.stat.ncsu.edu/~bondell/Software/NoCross/NoCrossQuant.R at time of writing. Finally the proposed sequential approach (NCQRseq) is performed by means of our package discussed in the next section.

As expected, the values of MISE for the median curves are exactly the same for both the traditional QR and the proposed sequential approach NCQRseq, since NCQRseq starts just from the median; the simultaneous approach NCQR performs the best for $\tau = 0.5$. However for the remaining quantile curves, differences between NCQR and NCQRseq get negligible especially at larger sample sizes, while the standard QR exhibits the worst performance because each curve is estimated independently from the others. Differences among the three approaches in terms of the mean integrated absolute error, $n^{-1}\sum_i|\hat{Q}_\tau(x_i) - Q_\tau(x_i)|$ yields very similar results. Also the same experiments using a B-spline with 4 and 8 degrees of freedom again leaves the findings unchanged.

In the second simulation settings we focus on the penalty term used to prevent undersmoothing. As emphasized in formula (2) we use a quadratic penalty $\sum_j(\Delta^d b_{kj})^2$, but a lasso-type $\sum_j|\Delta^d b_{kj}|$ would be possible as discussed in Bollaerts et al. (2006). L_1 penalty on the first differences is sometimes referred as 'fused lasso' and its implementation is practicable, for instance, using algorithms and software available at http://www.public.asu.edu/~jye02/Software/SLEP/ (Liu et al. 2009).

We compare L_1 and L_2 penalty in fitting median regression for data $y_i = 2 + x_i - 2(x_i - .4)_+^2 + .3x_i\epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 1)$, three sample sizes $n = \{50, 100, 500\}$, and two different distributions for the covariate, $x_i \sim \mathcal{U}(0, 1)$ and $x_i \sim \mathcal{B}eta(1, 3.5)$; in

Table 1 Comparing different approaches to estimate multiple regression quantiles: QR, classical regression quantile; NCQR, non-crossing regression quantile by joint estimation; NCQRseq, non-crossing regression quantile by the proposed sequential approach

		τ				
n	Method	0.1	0.3	0.5	0.7	0.9
		Example 4				
50	QR	0.608	0.454	0.443	0.472	0.602
	NCQR	0.558	0.437	0.425	0.446	0.553
	NCQRseq	0.568	0.448	0.443	0.459	0.561
100	QR	0.444	0.334	0.316	0.336	0.447
	NCQR	0.400	0.313	0.297	0.316	0.405
	NCQRseq	0.416	0.328	0.316	0.335	0.428
500	QR	0.197	0.149	0.142	0.152	0.199
	NCQR	0.188	0.144	0.138	0.147	0.191
	NCQRseq	0.186	0.149	0.142	0.150	0.191
		Example 5				
50	QR	0.852	0.660	0.640	0.672	0.873
	NCQR	0.791	0.630	0.616	0.643	0.820
	NCQRseq	0.798	0.645	0.640	0.658	0.816
100	QR	0.640	0.480	0.458	0.485	0.621
	NCQR	0.598	0.462	0.440	0.468	0.589
	NCQRseq	0.590	0.467	0.458	0.475	0.585
500	QR	0.291	0.218	0.205	0.216	0.282
	NCQR	0.284	0.216	0.203	0.213	0.278
	NCQRseq	0.280	0.215	0.205	0.213	0.275

Entries in the Table are averages of root mean integrated squared errors over the simulated data sets (500 replicates)

the latter scenario the covariate distribution has a long right tail causing sparse data at large values of the x_is. To model the nonlinear relationship we use a rank-ten B-spline basis with penalty L_1 or L_2 on the first order differences of the spline coefficients. At each replicate the optimal value of the tuning parameter is selected by five fold robust cross validation (Yuan 2006).

Table 2 presents the values of the mean integrated squared errors based on 500 replicates.

Overall differences between the penalties are quite negligible, although the MISE values when a L_2 penalty is employed are slightly smaller; the same findings appear in terms of the mean integrated absolute error. Results only refer to the median, but we believe similar patters hold also when fitting other regression quantiles. We conclude that using quadratic penalty in fitting quantile regression does not worsen the performance of the estimators.

Table 2 Comparing L_1 and L_2 penalties on the first order differences on the spline coefficients in median regression with a B-spline basis

Averages of root mean integrated squared errors over the simulated data sets (500 replicates)

	Penalty	Covariate	
n	Norm	$\mathcal{U}(0, 1)$	$\mathcal{B}eta(1, 3.5)$
50	L_1	0.0619	0.0403
	L_2	0.0621	0.0349
100	L_1	0.0494	0.0309
	L_2	0.0490	0.0293
500	L_1	0.0272	0.0171
	L_2	0.0242	0.0167

5 Software considerations

Software availability is an essential requisite to spread statistical methods in non-statistical communities, such as ecologists or biologists. In fact the lack of user-oriented software makes less practical application of any methodology, even if useful and appealing. In this section we discuss the framework proposed in this paper within the R environment (R Development Core Team 2010) via our package quantregGrowth. The aim is to favor widespread use of quantile regression for applying and building growth charts. At time of writing there exist the package gamlss that allows fitting growth charts via the LMS method (Rigby and Stasinopoulos 2004), however it appears that there is no package that allows to estimate monotone and non-crossing regression quantiles.

The quantregGrowth package is based on the quantreg package by R. Koenker (Koenker 2011). The main function is `gcrq()` that stands for growth charts regression quantiles, and currently there are methods to display and to plot the results. To illustrate, the following code shows how to estimate growth charts for the growth variable 'y', age variable 'x' and conditional (linear) variable 'z'. These variables are stored in the dataframe `growthData` shipped with the package and, as usual, loaded in the current workspace via the function `data()`.

```
> library(quantregGrowth)
> data(growthData)
> taus<-c(.05,.1,.25,.5,.75,.9,.95)
> m<-gcrq(y~ps(x, monotone=TRUE, lambda=0)+z, tau=taus,
+   data=growthData)
```

The first argument of the `gcrq()` function is the usual formula with the response on the left side and covariates on the right side; here the user can optionally specify the conditional linear terms and the age variable via the function `ps()` that builds a B-spline basis to model its nonlinear relationship with the response. Several arguments can be set: `monotone` to use monotonicity constraints (default to `TRUE`), `lambda` to set the numeric value of the tuning parameter λ (default to zero for unpenalized B-splines), `pdiff` to choose the difference order d of the penalty term when $\lambda > 0$, and `ndx` and `deg` to choose the number of equally spaced intervals and the degree of

the piecewise spline basis. The percentile vector of interest is passed via the argument `tau`, and the optional (and usual) `data` argument may be used to specify the dataframe where the variables are stored.

Currently there are a summary and plot method for objects of class "gcrq". In particular `plot.gcrq()` plots the fitted curves on a new device, but it is also possible to add them to an existing plot by setting `add=TRUE`. By default all the fitted quantiles are drawn, but the argument `select.tau` can be used to display only some selected curves. Also, by setting `legend=TRUE` in `plot.gcrq()`, a legend is drawn on the right side reporting the percentiles of the fit; finally `y=TRUE` also portrays the data points when `gcrq()` has been called with the option `y=TRUE`. Therefore a possible call could be

```
> plot(m, legend=TRUE, y=TRUE, lwd=2)
```

and a plot similar to those reported in Fig. 2 is produced.

6 Conclusion

Quantile Regression is an useful exploratory and inferential tool for the regression analysis. Cade and Noon (2003) discuss quantile regression and its applications in Ecology by emphasizing related advantages over the most used mean regression. We have used the quantile regression framework to estimate growth charts. Proper linear inequality constraints are employed to ensure monotonicity and non-crossing constraints, while a low-rank B-spline basis with a L_2 penalty on its coefficients is set to guarantee flexible and smooth quantile curves; a single smoothing parameter is used for all the quantile curves, since as discussed in Bondell et al. (2010), percentile-specific smoothing parameters increase the computational burden without improving the performance of the estimators. Growth charts can be also build via the well known LMS method; however, unlike the LMS method, the proposed approach is robust to outliers because it is based on minimization of a L_1 objective function that is well-known to yield robust estimates, such as the median in the simplest case.

We have also presented our R package quantregGrowth with demonstrating example code; it is hoped that it can facilitate spread of the proposed approach among the ecologists and nonstatistical communities in general. In fact, widespread application of statistical methods crucially depends on availability of software, especially if it is public domain.

There are several sides that deserve further investigation, such as computation of standard errors and modelling with correlated data coming from longitudinal profiles. Flexible modelling also needs some study in depth, including automatic selection of the smoothing parameter, inclusion of additional penalty terms and/or a varying smoothing parameter to allow a different amount of smoothing over age. Finally concavity constraints could be added while preserving monotonicity and non-crossing. These extensions could enlarge applicability and generality of the proposed framework.

Acknowledgments The authors would like to thank the referee for his/her valuable comments.

References

Bollaerts K, Eilers PHC, Aerts M (2006) Quantile regression with monotonicity restrictions using P-splines and the L_1-norm. Stat Model 6:189–207

Bondell HD, Reich BJ, Wang H (2010) Non-crossing quantile regression curve estimation. Biometrika 97:825–838

Borghi E, de Onis M, Garza C, the WHO Multicentre Growth Reference Study Group (2006) Construction of the world health organization child growth standards: selection of methods for attained growth curves. Stat Med 25:247–265

Bosch RJ, Ye Y, Woodworth GG (1995) A convergent algorithm for quantile regression with smoothing splines. Comput Stat Data Anal 19:613–630

Cade BS, Noon BR (2003) A gentle introduction to quantile regression for ecologists. Front Ecol Environ 1:412–420

Cole TJ, Green P (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. Stat Med 11:1305–1319

Dette H, Volgushev S (2008) Non-crossing non-parametric estimates of quantile curves. J R Stat Soc B 70:609–627

Duarte CM, Marbà N, Agawin N, Cebrian J, Enriquez S, Fortes MD, Gallegos ME, Merino M, Olesen B, Sand-Jensen K, Uri J, Vermaat J (1994) Reconstruction of seagrass dynamics:age determinations and associated tools for the seagrass ecologist. Mar Ecol Prog Ser 107:195–209, http://www.int-res.com/articles/meps/107/m107p195.pdf

He X (1997) Quantile curves without crossing. Am Stat 51:186–192

He X, Shi P (1998) Monotone B-spline smoothing. J Am Stat Assoc 93:643–649

Hedley AA, Ogden CL, Johnson CL, Carroll MD, Curtin LR, Flegal KM (2004) Prevalence of overweight and obesity among US children, adolescents, and adults, 1999–2002. J Am Med Assoc 291:2847–2850

Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge

Koenker R (2011) quantreg: Quantile Regression. http://CRAN.R-project.org/package=quantreg, R package version 4.71

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

Liu J, Ji S, Ye J (2009) SLEP: sparse learning with efficient projections. Arizona State University, http://www.public.asu.edu/jye02/Software/SLEP

Meyer M, Woodroofe M (2000) On the degrees of freedom in shape-restricted regression. Ann Stat 28:1083–1104

Muggeo VMR (2003) Estimating regression models with unknown break-points. Stat Med 22:3055–3071

Muggeo VMR, Sciandra M, Augugliaro L (2012) Quantile regression via iterative least squares computations. J Stat Comput Simul 82:1557–1569

Ng PT (1996) An algorithm for quantile smoothing splines. Comput Stat Data Anal 22:99–118

R Development Core Team (2010) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org, ISBN 3-900051-07-0

Rigby RA, Stasinopoulos DM (2004) Smooth centile curves for skew and kurtotic data modelled using the box-cox power exponential distribution. Stat Med 23:3053–3076

Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape (with discussion). Appl Stat 54:507–554

Tomasello A, Calvo S, Maida GD, Lovison G, Pirrotta M, Sciandra M (2007) Shoot age as a confounding factor on detecting the effect of human-induced disturbance on posidonia oceanica growth performance. J Exp Mar Biol Ecol 343(2):166–175. doi:10.1016/j.jembe.2006.11.017, http://www.sciencedirect.com/science/article/pii/S0022098107000330

Wei Y, Pere A, Koenker R, He X (2006) Quantile regression methods for reference growth charts. Stat Med 25:1369–1382

Yuan M (2006) GACV for quantile smoothing splines. Comput Stat Data Anal 50:813–829