UC Davis

UC Davis Previously Published Works

Title

Segmented Model Selection in Quantile Regression Using the Minimum Description Length Principle

Permalink

<https://escholarship.org/uc/item/53p322nb>

Journal

Journal of the American Statistical Association, 109(507)

ISSN

0162-1459

Authors

Aue, A
Cheung, RCY
Lee, TCM
et al.

Publication Date

2014-07-03

DOI

10.1080/01621459.2014.889022

Peer reviewed

Segmented Model Selection in Quantile Regression Using the Minimum Description Length Principle

Alexander Aue, Rex C. Y. Cheung, Thomas C. M. Lee & Ming Zhong

To cite this article: Alexander Aue, Rex C. Y. Cheung, Thomas C. M. Lee & Ming Zhong (2014) Segmented Model Selection in Quantile Regression Using the Minimum Description Length Principle, Journal of the American Statistical Association, 109:507, 1241-1256, DOI: [10.1080/01621459.2014.889022](https://doi.org/10.1080/01621459.2014.889022)

To link to this article: <http://dx.doi.org/10.1080/01621459.2014.889022>

Accepted author version posted online: 07 Mar 2014.

Submit your article to this journal [↗](#)

Article views: 342

View Crossmark data [↗](#)

Segmented Model Selection in Quantile Regression Using the Minimum Description Length Principle

Alexander AUE, Rex C. Y. CHEUNG, Thomas C. M. LEE, and Ming ZHONG

This article proposes new model-fitting techniques for quantiles of an observed data sequence, including methods for data segmentation and variable selection. The main contribution, however, is in providing a means to perform these two tasks simultaneously. This is achieved by matching the data with the best-fitting piecewise quantile regression model, where the fit is determined by a penalization derived from the minimum description length principle. The resulting optimization problem is solved with the use of genetic algorithms. The proposed, fully automatic procedures are, unlike traditional break point procedures, not based on repeated hypothesis tests, and do not require, unlike most variable selection procedures, the specification of a tuning parameter. Theoretical large-sample properties are derived. Empirical comparisons with existing break point and variable selection methods for quantiles indicate that the new procedures work well in practice.

KEY WORDS: Genetic algorithms; Large p , small n ; Minimum description length; Piecewise modeling.

1. INTRODUCTION

Quantile methods are useful whenever observed data are not compatible with the standard symmetry and normality assumptions for which estimation procedures in linear regression models work best. Over the past few decades, a large body of literature has been developed providing a broad spectrum of statistical methodology for data with asymmetries, local persistence, and changes in location, scale, and shape. The relevant lines of this research along with many further techniques are given in the informative monograph by Koenker (2005). Most popular among all methods have been quantile regressions (QRs) pioneered by Koenker and Bassett (1978). QR assumes that the conditional quantile function of a response variable is related linearly with its predictor variables. Targeting a collection of quantiles enables statisticians to extract more comprehensive relationships between response and predictors. Estimation in a QR model can be performed through minimization of a convex optimization problem involving the so-called check function, the large-sample properties of which are thoroughly established (see Bassett and Koenker 1986; Pollard 1991, among others). Applications of QR can be found in a number of fields such as econometrics (Koenker and Xiao 2002), microarray studies (Wang and He 2007, 2008), and survival analysis (Koenker and Geling 2001). All of these contributions assume that the underlying random process does not suffer from structural breaks and that all predictors included in the model are significant. This may not always be realistic.

Research related to uncovering structural breaks has received considerable attention for at least three decades. While most of the contributions to regression models have been focused on structural breaks in the conditional mean, Qu (2008) and Oka

and Qu (2011) pointed out convincingly that changes, for example, in economic policies may affect only certain quantiles while others may not be subjected to structural breaks. They proposed a procedure based on repeated hypothesis tests to detect and estimate an unknown number of structural breaks, thereby extending existing work (e.g., Bai 1998; Su and Xiao 2008) to the quantile setting. A comprehensive review of research on structural breaks can be found in Csörgö and Horváth (1997) and a summary of more recent developments in Aue and Horváth (2013).

Another relevant work is the pioneer article of Li et al. (2011), which is one of the first articles that studied segmented QR models. To be more specific, this article allows the slope of one of the covariates to change at an unknown break point. The authors developed parameter estimates, established their asymptotic validity, and provided procedures for constructing parameter confidence intervals. They have also discussed methods for testing the existence of a break point.

Variable selection is another principal issue in the model building process for multiple regressions, whose importance has been growing with the complexity and size of the available data. A statistician often faces the task to remove redundant variables from a potentially large group of predictors for which the traditional stepwise selection and deletion framework may be computationally prohibitive. Instead variable selection has been rephrased in a regularization framework, with the many methods developed for this purpose including the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996), smoothly clipped absolute deviation (SCAD; Fan and Li 2001), and adaptive LASSO (Zou 2006) procedures for linear regression. These methods were extended for examples by Li, Liu, and Zhu (2007), Li and Zhu (2008), Zou and Yuan (2008), and Wu and Liu (2009) to QR models.

The present article adds to this research by combining the tasks of data segmentation and variable selection into a new joint piecewise QR model selection method. One major novelty is the use of the minimum description length (MDL) principle (Rissanen 1989, 2007) to determine the model complexity. The

A. Aue (E-mail: aaue@ucdavis.edu) is Associate Professor, R. C. Y. Cheung (E-mail: rccheung@ucdavis.edu) is graduate student, T. C. M. Lee (E-mail: tcmllee@ucdavis.edu) is Professor, and M. Zhong (E-mail: mzhong@ucdavis.edu) is graduate student, Department of Statistics, University of California, Davis, One Shields Avenue, Davis, CA 95616. The authors are most grateful to the constructive comments from the reviewers and the associate editor, which have led to a much improved version of the article. This research was partially supported by NSF grants DMS 1007520, DMS 1209226, DMS 1209232, and DMS 1305858.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jasa.

underlying philosophy of MDL will be discussed in greater detail in Section 3.2, but briefly, MDL defines the best-fitting model as the one that achieves the best compression rate for the observed data. This is achieved by constructing a parameter-free code length function for which the best-fitting model can be obtained as its minimizer. The MDL principle has been applied with great success to many model selection problems, but to our best knowledge, it has never been applied in the context of QR before.

If only data segmentation is considered, the proposed method competes with the hypothesis testing approach of Qu (2008) and Oka and Qu (2011). For this case, consistency for the estimation of relative break point locations is established. If only variable selection is considered, the new method becomes a fully automatic competitor to LASSO, SCAD, and adaptive LASSO that does not require the specification of a tuning parameter. In this case, consistency, sparsity, and asymptotic normality (oracle properties) are established.

For the practical minimization of the code length function, genetic algorithms (GAs, e.g., Davis 1991) are employed to solve the resulting optimization problem. Empirical results include simulation experiments, which show that the proposed method compares favorably with the respective competitors, and applications to hotel accommodation costs, Boston housing data, and U.S. interest rates.

The remainder of the article is organized as follow. In Section 2, QR models and their piecewise specifications are introduced. Section 3 presents the methodology for jointly and separately conducting data segmentation and variable selection. Section 4 develops large-sample properties. Section 5 elaborates on the implementation details of GAs for the proposed procedures. The results of a comprehensive simulation study are tabulated in Section 6. Section 7 presents real-data applications and Section 8 concludes. All proofs are relegated to the Appendix.

2. PIECEWISE QUANTILE REGRESSION

To circumvent some of the limitations of linear models discussed in the introduction, QR models have been introduced by Koenker and Bassett (1978). Their formal definition is as follows.

Definition 1. Given $((X'_t, y_t) : t \in \mathbb{Z})$, with $\mathbb{Z}$ being the set of integers, y_t a real-valued response variable, and $X_t = (1, x_{t,1}, \dots, x_{t,p})'$ a vector of dimension $p + 1$, QR formulates the τ th conditional quantile of y_t , $Q_{y_t}(\tau|X_t) = F_{y_t}^{-1}(\tau|X_t) = \inf\{s : F_{y_t}(s|X_t) \geq \tau\}$, as a linear function of parameters $\theta = (\theta_0, \theta_1, \dots, \theta_p)'$, that is,

$$Q_{y_t}(\tau|X_t) = X'_t\theta(\tau), \quad t \in \mathbb{Z}, \quad (1)$$

where $\tau \in (0, 1)$.

While a single QR model may serve as a useful tool for approximating local persistence and potentially more complicated global dynamics, a set of piecewise QR models should be more capable of modeling datasets with structural breaks. Some models, including location-shift models and more complicated location-scale-shift models, have been proposed to tackle such problems (see Koenker 2005). Moreover, Qu (2008) and Oka and Qu (2011) developed a rigorous testing and estimation

framework for multiple structural breaks in QR models. None of these works, however, allows the simultaneous selection of variables. In this section, the adoption of piecewise QR models that allow for the possibility of both estimating structural breaks and selecting variables is proposed. With the existence of break points, dividing a given dataset into smaller segments enables the fitting of separate QR models for which linearity could be more closely achieved. On the QR segments a more tailored approach to variable selection appears feasible.

These piecewise QR models are defined as follows. Assume that the observed data $(X'_t, y_t), t = 1, \dots, n$, can be decomposed into $m + 1$ segments, and that, for $\ell = 1, \dots, m + 1$, the ℓ th segment can be modeled by a $QR(s_\ell)$ process. Here s_ℓ denotes the number of predictors selected in the model, and each two adjacent QR models have different parameters. Note that the $m + 1$ QR models do not necessarily include the same predictor variables, and the details for conducting variable selection will be discussed later. For $\ell = 1, \dots, m$, the ℓ th break point (the time index at which the transition from the ℓ th to the $(\ell + 1)$ th segment occurs) is denoted by k_ℓ . Letting $k_0 = 0$ and $k_{m+1} = n$, the conditional τ th quantile function on the ℓ th segment is

$$Q_{y_t}(\tau|X_{\ell,t}) = X'_{\ell,t}\theta_\ell(\tau), \quad t = k_{\ell-1} + 1, \dots, k_\ell, \quad (2)$$

where $X_{\ell,t} = (1, x_{\ell,t,1}, \dots, x_{\ell,t,s_\ell})'$ and $\theta_\ell(\tau) = (\theta_{\ell,0}(\tau), \theta_{\ell,1}(\tau), \dots, \theta_{\ell,s_\ell}(\tau))'$. For the τ th quantile, model (2) is thus specified by the parameters $m, \mathcal{K} = (k_1, \dots, k_m)'$, and $\theta(\tau) = (\theta_1(\tau)', \dots, \theta_{m+1}(\tau)')'$. For the case $m = 0$, (2) reduces to (1) if all predictors are included in the model.

3. BREAK POINT DETECTION AND VARIABLE SELECTION USING MDL

3.1 Estimation of $\theta(\tau)$

Following Koenker and Bassett (1978), an estimate of $\theta(\tau)$ in QR models can be obtained by solving the convex optimization problem

$$\min_{\theta(\tau) \in \mathbb{R}^{p+1}} \sum_{t=1}^n \rho_\tau(y_t - X'_t\theta(\tau)), \quad (3)$$

where $\rho_\tau(\cdot)$ is the check function defined as $\rho_\tau(u) = u(\tau - I(u < 0))$. Extension to piecewise QR models can be achieved as follows. As stated in Section 2, at a particular quantile τ , a piecewise QR model is determined by the number of breaks m and their locations $\mathcal{K} = (k_1, \dots, k_m)'$, along with the parameters for each segment $\theta(\tau) = (\theta_1(\tau)', \dots, \theta_{m+1}(\tau)')'$. If both m and $\mathcal{K}$ are known, $\theta(\tau)$ can be estimated by solving

$$\min_{\theta(\tau)} \sum_{\ell=1}^{m+1} \sum_{t=k_{\ell-1}+1}^{k_\ell} \rho_\tau(y_t - X'_{\ell,t}\theta_\ell(\tau)). \quad (4)$$

The estimation of m and $\mathcal{K}$ will be discussed in Section 3.2.

Solving (3) can be related to maximum likelihood estimation for asymmetric Laplace (AL) distributions, see Geraci and Bottai (2007) and Yu, Lu, and Stander (2003). Assuming that at the τ th quantile the innovations in model (1) follow an $AL(\tau)$ distribution with density $f(z) = \tau(1 - \tau)\exp\{-\rho_\tau(z)\}$, then maximizing the likelihood function $L(\theta(\tau)) \propto \exp\{-\sum_{t=1}^n \rho_\tau(y_t - X'_t\theta(\tau))\}$ is equivalent to solving the problem in (3). The equivalent to (4) could be stated

in a similar fashion. The use of the AL likelihood enables the formulation of an MDL criterion for estimating m and $\mathcal{K}$, as to be discussed next.

3.2 MDL Criterion for Break Point Detection

The estimation of m and $\mathcal{K}$ is less trivial, as different values of m would lead to a different number of parameters in $\mathcal{K}$. In other words, this is a model selection problem and (4) cannot be applied for this task. This section applies the MDL principle to derive an estimation criterion for m and $\mathcal{K}$. For an introduction to MDL, see, for example, Rissanen (1989, 2007), Hansen and Yu (2001), and Lee (2001).

When applying the MDL principle, the “best” model is defined as the one that allows the greatest compression of the data $Y = (y_1, \dots, y_n)$. That is, the “best” model enables us to store Y in a computer with the shortest code length. There are different versions of MDL, and the so-called “two-part” version, for which the code length of Y is calculated as the sum of two parts, shall be used here. The first part represents the fitted values of the piecewise QR model under consideration, denoted by $\hat{\mathcal{F}}$. The second part represents the leftover information that is not explained by the first part, the residuals in other words, denoted by $\hat{\mathcal{E}} = Y - \hat{\mathcal{F}}$, where $\hat{Y}$ is the fitted value for Y . The notation $\text{CL}_{\mathcal{F}}(Y|\tau)$ is used to represent the code length of Y under the model $\mathcal{F}$ at a given τ th quantile. This leads to the decomposition

$$\text{CL}_{\mathcal{F}}(Y|\tau) = \text{CL}_{\mathcal{F}}(\hat{\mathcal{F}}|\tau) + \text{CL}_{\mathcal{F}}(\hat{\mathcal{E}}|\hat{\mathcal{F}}, \tau). \quad (5)$$

The MDL principle then defines the best piecewise model $\hat{\mathcal{F}}$ as the one that minimizes $\text{CL}_{\mathcal{F}}(Y|\tau)$.

To use (5), calculation of the two terms in the right-hand side is needed. Starting with $\text{CL}_{\mathcal{F}}(\hat{\mathcal{F}}|\tau)$, let $n_{\ell} = k_{\ell} - k_{\ell-1}$ be the number of observations in the ℓ th segment of $\hat{\mathcal{F}}$. Since $\hat{\mathcal{F}}$ is completely specified by m , k_{ℓ} 's, and $\theta_{\ell}(\tau)$'s, it follows that

$$\begin{aligned} \text{CL}_{\mathcal{F}}(\hat{\mathcal{F}}|\tau) &= \text{CL}_{\mathcal{F}}(m|\tau) + \text{CL}_{\mathcal{F}}(n_1, \dots, n_{m+1}|\tau) \\ &\quad + \text{CL}_{\mathcal{F}}(\theta_1(\tau), \dots, \theta_{m+1}(\tau)) \\ &= \text{CL}_{\mathcal{F}}(m|\tau) + \text{CL}_{\mathcal{F}}(n_1|\tau) + \dots + \text{CL}_{\mathcal{F}}(n_{m+1}|\tau) \\ &\quad + \text{CL}_{\mathcal{F}}(\theta_1(\tau)) + \dots + \text{CL}_{\mathcal{F}}(\theta_{m+1}(\tau)). \end{aligned} \quad (6)$$

According to Rissanen (1989), it requires approximately $\log_2 I$ bits to encode an integer I if its upper bound is unknown, leading to $\text{CL}_{\mathcal{F}}(m|\tau) = \log_2 m$. On the other hand, it requires approximately $\log_2 I_U$ bits if the upper bound I_U of I is known, which leads to $\text{CL}_{\mathcal{F}}(n_{\ell}|\tau) = \log_2 n$ for all ℓ , as $n_{\ell} \leq n$.

The code length calculation for $\theta_{\ell}(\tau)$ requires additional steps. Let $p+1$ be the total number of predictor variables in the observed data, and $s_{\ell} \leq p+1$ be the number of significant predictor variables in the ℓ th segment. To encode $\theta_{\ell}(\tau)$, one first needs to identify which of the s_{ℓ} predictors are involved. This takes $s_{\ell} \log(p+1)$ bits, as each of the s_{ℓ} predictors can be identified by an index upper bounded by $p+1$. Next, the corresponding parameter estimates need to be encoded, and using the result by Rissanen (1989), which states that effectively $\frac{1}{2} \log_2 N$ bits are required to encode a maximum likelihood estimate of a real parameter computed from N observations, one has $\text{CL}_{\mathcal{F}}(\theta_{\ell}(\tau)) = s_{\ell} \log(p+1) + \frac{s_{\ell}}{2} \log_2 n_{\ell}$.

For the moment assuming that $s_{\ell} = p+1$ for all ℓ , giving $\text{CL}_{\mathcal{F}}(\theta_{\ell}(\tau)) = (p+1) \log(p+1) + \frac{p+1}{2} \log_2 n_{\ell}$. That is, all

predictor variables will be used and no variable selection will be conducted. The cases for different s_{ℓ} values will be discussed later. Then, with this and the above results, (6) becomes

$$\begin{aligned} \text{CL}_{\mathcal{F}}(\hat{\mathcal{F}}|\tau) &= \log_2 m + (m+1) \log_2 n + (m+1)(p+1) \\ &\quad \times \log_2(p+1) + \sum_{\ell=1}^{m+1} \frac{p+1}{2} \log_2 n_{\ell}. \end{aligned} \quad (7)$$

Now it remains to calculate the last term of (5). Proved by Rissanen (1989), it is known that the code length of the residuals $\hat{\mathcal{E}}$ is equal to the negative of the log-likelihood of the fitted model $\hat{\mathcal{F}}$. With the application of the AL likelihood, this leads to

$$\begin{aligned} \text{CL}_{\mathcal{F}}(\hat{\mathcal{E}}|\hat{\mathcal{F}}, \tau) &= \sum_{\ell=1}^{m+1} \sum_{t=k_{\ell-1}+1}^{k_{\ell}} \rho_{\tau}(y_t - X'_t \theta(\tau)) - n \log(\tau(1-\tau)). \end{aligned} \quad (8)$$

Summing up the results obtained in Equations (7) and (8) and dropping the constant term $n \log(\tau(1-\tau))$, the best piecewise QR model at a single quantile $\tau \in (0, 1)$ is defined as the one that minimizes the MDL criterion

$$\begin{aligned} \text{MDL}(m, \mathcal{K}, \theta|\tau) &= \log_2 m + (m+1) \log_2 n + (m+1)(p+1) \log_2(p+1) \\ &\quad + \sum_{\ell=1}^{m+1} \frac{p+1}{2} \log_2 n_{\ell} + \sum_{\ell=1}^{m+1} \sum_{t=k_{\ell-1}+1}^{k_{\ell}} \rho_{\tau}(y_t - X'_{\ell,t} \theta_{\ell}(\tau)). \end{aligned} \quad (9)$$

Practical minimization of this criterion will be discussed in Section 5. Notice that this MDL criterion covers the case when $m = 0$. That is, it does not assume the existence of any break point.

3.3 MDL Criterion for Variable Selection

This section considers a different model selection problem for QR: it is assumed that there is no break point (i.e., $m = 0$) but the significant predictor variables are to be selected from the $p+1$ available predictor variables. Let the number of significant predictor variables be s . The above derivation of the MDL criterion (9) can be straightforwardly modified for this problem, setting first $m = 0$ and then $s_1 = s$ and $n_1 = n$. This gives the code length at a particular quantile τ as

$$\begin{aligned} \text{MDL}(s, \theta(\tau)|\tau) &= \log_2 n + s \log_2(p+1) + \frac{s}{2} \log_2 n \\ &\quad + \sum_{t=1}^n \rho_{\tau}(y_t - X'_t \theta(\tau)). \end{aligned} \quad (10)$$

Note that in the above the intercept term is included in the predictor list. This does not follow the common practice in the variable selection literature, but it should not cause any issue as for (10), as the intercept could enter the model if it is statistically significant.

3.4 MDL Criterion for Simultaneous Break Point Detection and Variable Selection

An MDL criterion can also be derived for simultaneously conducting break point detection and variable selection. In fact, the derivation is similar to (9), except $\text{CL}_{\mathcal{F}}(\theta_{\ell}(\tau)) =$

$s_\ell \log(p+1) + \frac{s_\ell+2}{2} \log_2 n_\ell$ for all ℓ is used instead of $\text{CL}_\mathcal{F}(\theta_\ell(\tau)) = (p+1) \log(p+1) + \frac{p+1}{2} \log_2 n_\ell$. With $\mathcal{S} = (s_1, \dots, s_{m+1})'$, this yields the following criterion at any given quantile τ :

$$\begin{aligned} \text{MDL}(m, \mathcal{K}, \mathcal{S}, \theta(\tau)|\tau) \\ = \log_2 m + (m+1) \log_2 n + \sum_{\ell=1}^{m+1} s_\ell \log_2(p+1) \\ + \sum_{\ell=1}^{m+1} \frac{s_\ell}{2} \log_2 n_\ell + \sum_{\ell=1}^{m+1} \sum_{t=k_{\ell-1}+1}^{k_\ell} \rho_\tau(y_t - X'_{\ell,t} \theta_\ell(\tau)). \end{aligned} \quad (11)$$

It is clear that (11) collapses to (9) and (10) if $s_\ell = p$ and $m = 0$, respectively.

4. LARGE-SAMPLE RESULTS

4.1 Consistency of Break Point Detection

To verify large-sample results for the MDL-based model selection procedure, a number of assumptions are needed on the underlying random process. To make asymptotic statements a true model is needed, so it is assumed that observations $(X'_1, y_1), \dots, (X'_n, y_n)$ have been taken and that they follow the piecewise QR model (2). True model parameters will be denoted with a superscript “0,” so that the true number of break points is written as m^0 and their locations as $k_\ell^0 = \lfloor \lambda_\ell^0 n \rfloor$, where $0 < \lambda_1^0 < \lambda_2^0 < \dots < \lambda_{m^0}^0 < 1$. It is worthwhile mentioning that for an application of the MDL-based procedures, the assumption of a true model is not necessary. To sufficiently separate neighboring break points in the true segmentation, assume that there is an $\epsilon > 0$ such that $\epsilon \ll \min_{\ell=1, \dots, m^0+1} (\lambda_\ell^0 - \lambda_{\ell-1}^0)$. Candidate segmentations will be given in the form

$$\Lambda_m = \{(\lambda_1, \dots, \lambda_m) : 0 < \lambda_1 < \dots < \lambda_m < 1, \lambda_\ell - \lambda_{\ell-1} \geq \epsilon, \ell = 1, \dots, m+1\},$$

where $\lambda_0 = 0$ and $\lambda_{m+1} = 1$. Fixing $\tau \in (0, 1)$, setting $\lambda = (\lambda_1, \dots, \lambda_m)$, the parameters m^0 , $\lambda^0 = (\lambda_1^0, \dots, \lambda_{m^0}^0)$, and $\theta^0(\tau) = (\theta_1^0(\tau)', \dots, \theta_{m^0+1}^0(\tau)')$ are then in view of (9) estimated by minimizing the MDL criterion

$$(\hat{m}, \hat{\lambda}, \hat{\theta}(\tau)) = \arg \min_{(m, \lambda, \theta(\tau)) \in \mathcal{M}} \frac{1}{n} \text{MDL}(m, \lambda, \theta(\tau)|\tau),$$

where the minimum is taken in the set $\mathcal{M} = \{(m, \lambda, \theta(\tau)) : m \leq M, \lambda \in \Lambda_m\}$ with M denoting an upper bound for m .

The next set of assumptions concerns the structure of the predictors and innovations in the piecewise QR model. For $\ell = 1, \dots, m^0 + 1$ and $t = k_{\ell-1}^0 + 1, \dots, k_\ell^0$ denote by $\varepsilon_{\ell,t} = y_t - X'_{\ell,t} \theta_\ell^0(\tau)$ the independent and identically distributed innovations of the ℓ th segment.

Assumption 1.

(a) There exists a positive definite matrix Σ such that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^{\lfloor n\kappa \rfloor} X_t X'_t = \kappa \Sigma \quad \text{for all } \kappa \in [0, 1].$$

(b) The innovations $\varepsilon_{\ell,t}$ have a continuous, positive density f_ℓ in a neighborhood of zero.

Part (a) of Assumption 1 is needed for identifiability, ensuring that the predictors $X_1, \dots, X_n$ are not concentrated on a linear subspace, and also used to restrict the rate of growth of the predictors. It extends standard conditions in the QR literature (see Pollard 1991) and makes them applicable to the piecewise QR model under consideration here. Part (b) ensures that the τ th quantile of y_t is unique. In the Appendix, the following theorem is proved.

Theorem 1. Assume that Assumption 1 is satisfied, and suppose the number of break points m^0 is known. Then estimating the piecewise QR model at any single quantile $\tau \in (0, 1)$ leads to

$$\hat{\lambda}_\ell \rightarrow \lambda_\ell^0 \quad \text{with probability one} \quad (n \rightarrow \infty)$$

for all $\ell = 1, 2, \dots, m^0$, where $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_{m^0})$ is the minimizer of the criterion function (9). Consequently, $\hat{\theta}(\tau) \rightarrow \theta^0(\tau)$ with probability one.

An inspection of the proof shows that as long as one can ensure that Lemma 1 holds, the results of Theorem 1 will be valid also under weaker assumptions permitting dependent regressors and time series settings. Lemma 1 holds therefore in particular if the regressors $(X_t : t \in \mathbb{Z})$ are stationary and ergodic. It would also hold under the conditions imposed in Oka and Qu (2011).

The assumption that m^0 is known does not seem essential. In fact, Bai (1998) developed a least absolute deviation (LAD) estimation procedure for the number of break points based on a heavier penalty of the form $g(m) = \sqrt{nm}$ for which he showed consistency under conditions even allowing m to grow to infinity with the sample size. The simulations in Bai (1998) indicate that $g(m)$ has a tendency to underestimate the true number of breaks, while the Bayesian information criterion (BIC)-based penalty has a tendency to overestimate. The MDL penalty is stronger than BIC and weaker than $g(m)$. Our simulations suggest that MDL correctly identifies m^0 in the finite sample scenarios that were tested, but whether this criterion leads to a consistent estimate of the number of breaks remains an open question. The case of unknown m^0 is generally much harder to work with and theoretical results have been restricted to special cases such as the normal and binomial distributions, see the discussion in Aue and Lee (2011), or impose strong smoothness conditions on the likelihood function, see the recent work of Davis and Yau (2013), that are not satisfied in the present framework.

4.2 Oracle Properties of Variable Selection

In this section, oracle properties of the MDL variable selection procedure in (10) are verified. Since there are no break points, the QR model formulation in (1) applies. The relevant true parameter vector at the τ th quantile is then denoted by $\theta^0(\tau) = (\theta_0^0(\tau), \theta_1^0(\tau), \dots, \theta_p^0(\tau))'$. It is without loss of generality assumed that the s^0 significant predictors correspond to the first s^0 components of the parameter vector, yielding the decomposition $\theta^0(\tau) = (\tilde{\theta}_1^0(\tau)', \tilde{\theta}_2^0(\tau)')$, where all s^0 entries of $\tilde{\theta}^0(\tau)$ are nonzero and all $p+1-s^0$ entries of $\tilde{\theta}^0(\tau)$ are zero. Performing the corresponding rearrangement in the predictors gives consequently

$$Q_{y_t}(\tau|X_t) = X'_t \theta^0(\tau) = X'_{t,1} \tilde{\theta}_1^0(\tau) + X'_{t,2} \tilde{\theta}_2^0(\tau), \quad t \in \mathbb{Z}. \quad (12)$$

Finally, denote the top-left $s \times s$ submatrix of the rearranged Σ specified in Assumption 1 by Σ_{11} and the right-bottom $(p+1-s) \times (p+1-s)$ submatrix of Σ by Σ_{22} . With these assumptions, $\sqrt{n}$ -consistency and oracle properties of the minimizers $\hat{\theta}$ and $\hat{\theta}(\tau)$ of the MDL criterion (10) are established in the next two theorems. Their proofs can be found in the Appendix.

Theorem 2 (Consistency). Consider a sample $(X'_1, y_1), \dots, (X'_n, y_n)$ of n observations from model (12). Under Assumption 1, there exists a local minimizer $\hat{\theta}(\tau)$ such that

$$\sqrt{n} \|\hat{\theta}(\tau) - \theta^0(\tau)\| = O_p(1) \quad (n \rightarrow \infty)$$

for any $\tau \in (0, 1)$, where $\|\cdot\|$ denotes Euclidean norm.

Theorem 3 (Oracle). Consider a sample $(X'_1, y_1), \dots, (X'_n, y_n)$ of n observations from model (12) and let $\hat{\theta}(\tau) = (\hat{\theta}_1(\tau)', \hat{\theta}_2(\tau)')'$ denote the $\sqrt{n}$ -consistent local minimizer of Theorem 2. Then, under Assumption 1,

- (a) $P(\hat{\theta}_2(\tau) = 0) \rightarrow 1$ (sparsity), and
- (b) $\sqrt{n}(\hat{\theta}_1(\tau) - \hat{\theta}_1^0(\tau)) \xrightarrow{D} N(0, \tau(1-\tau)\Sigma_{11}^{-1}/f(0)^2)$ (asymptotic normality),

for any $\tau \in (0, 1)$ as $n \rightarrow \infty$.

Theorems 2 and 3 are derived making heavy use of the convexity lemma stated in Pollard (1991). As is outlined there, the case of random and dependent regressors X_t can be incorporated (see his Theorem 2). In particular, X_t may collect the lags of the response variable y_t (see Example 2 of Pollard 1991) under appropriately modified assumptions and important time series are included in the framework presented here. We formulate this as a corollary.

Corollary 1. Let $(\mathcal{F}_t : t \in \mathbb{Z})$ be an increasing sequence of σ -fields such that X_t is $\mathcal{F}_{t-1}$ -measurable and $\varepsilon_{\ell,t}$ is independent of $\mathcal{F}_{t-1}$ for all $t \in \mathbb{Z}$, and $\max_{t \leq n} |X_t| = o_P(\sqrt{n})$. Then, under Assumption 1, the results of Theorems 2 and 3 are retained.

4.3 Properties of Simultaneous Break Point Detection and Variable Selection

Combining the results of the previous two subsections, the following asymptotic result can be established. It is proved in the Appendix.

Theorem 4. Assume that Assumption 1 is satisfied, and suppose the number of break points m^0 is known. Then estimating the piecewise QR model specified before at any single quantile $\tau \in (0, 1)$ leads to

- (a) $\hat{\lambda}_\ell \rightarrow \lambda_\ell^0$ with probability one as $n \rightarrow \infty$ for all $\ell = 1, 2, \dots, m^0$, where $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_{m^0})$ is the minimizer of the criterion function (11).
- (b) The variables selected in each segment satisfy the consistency and oracle properties of Theorems 2 and 3.
- (c) The results in (a) and (b) remain true if $(X_t : t \in \mathbb{Z})$ is stationary and ergodic and if the conditions of Corollary 1 are satisfied.

5. PRACTICAL MDL MINIMIZATION USING GENETIC ALGORITHMS

Minimizing any of the MDL criteria derived in Section 3 is a difficult task. This section proposes the use of *genetic algorithms* (GAs) to solve these optimization problems.

5.1 General Description

GAs are a class of stochastic optimization techniques that are founded on the idea of Darwin's theory of natural selection. A GA first generates, usually in a random fashion, a population of possible solutions to the optimization problem. These solutions are typically represented in vector form and are known as *chromosomes*. Such chromosomes are then combined to form the next generation of chromosomes in such a way that "good characteristics" (with respect to the optimization problem) will have a high chance of being preserved. Two methods for combining chromosomes will be discussed below: *crossover* and *mutation*. Once a sufficient number of second-generation chromosomes are formed, the same procedure for producing chromosomes will be applied to form the third generation. As this process continues, "good characteristics" will gradually dominate the population, and the evolution stops when a chromosome with desirable characteristics (i.e., a good optimizer) is obtained. Successful applications of GAs for solving various optimization problems can be found, for example, in Davis (1991).

5.2 Implementing Break Point Detection

This section provides implementation details of the GA that is specific for break point detection, that is, for minimizing (9).

Chromosome representation. For the problem of detecting break points within a linear QR model at a specific quantile τ , it is sufficient for a chromosome to carry information of the break points k_ℓ , as other parameters such as m will then be uniquely calculated. Following Davis, Lee, and Rodriguez-Yam (2006), a chromosome is expressed as a vector of n integers: a chromosome $\delta = (\delta_1, \dots, \delta_n)$ is of length n with gene values δ_t defined as

$$\delta_t = \begin{cases} -1, & \text{if no break point at time } t, \\ 1, & \text{if } t = k_{\ell-1} \text{ at quantile } \tau. \end{cases}$$

In practice, the following constraint on each δ is imposed: to have sufficiently many observations for parameter estimation, each piecewise QR process is required to have a minimum length of 10.

First population generation. The initialization of the algorithm requires the creation of a number of chromosomes to fill the first generation. The gene values of these chromosomes are randomly generated, as follows. Starting from $t = 1$, the first 10 genes δ_1 to δ_{10} are all set to -1 , as required by the minimum length constraint mentioned above. Then for the next gene in line (i.e., the 11th): with probability π_B it will be initialized as a break point, otherwise it will be assigned -1 with probability $1 - \pi_B$. Following Davis, Lee and Rodriguez-Yam (2006), $\pi_B = 10/n$ is used. If this gene is declared to be a break point, then the value 1 is assigned to it, and this implies that the next nine genes will have values -1 due to the minimum length constraint. On the other hand, if the 11th gene is initialized as a nonbreak point with value -1 , then the process will continue to

the next gene (i.e., the 12th gene) and decide whether this gene should be declared as a break point or not. This initialization process continues in the same way until a value is assigned to the last gene δ_n .

Crossover and mutation. Once a first generation of random chromosomes is generated, *crossover* and *mutation* operations are applied to generate offspring, which will form the second generation. The offspring tend to be better than their parents in the sense that they are better solutions to the optimization problem. In the present implementation, the probability for conducting a crossover operation is set as π_C and for conducting a mutation operation as $1 - \pi_C$, where $\pi_C = (n - 10)/n$.

In a crossover operation, two parent chromosomes are chosen from the current pool of chromosomes to produce one child chromosome. The parent chromosomes are chosen with probabilities inversely proportional to their ranks sorted by their values of the objective function, that is, their MDL values in the current problem. The goal of this operation is to allow the child chromosome to inherit good traits from its parents. A common “mating” strategy is that every child gene location has an equal probability of inheriting from its father gene or its mother gene. In the present problem, the gene values δ_i of the child chromosome will be selected as follows. Beginning with $t = 1$, δ_t will take the corresponding gene value either from its father chromosome or its mother chromosome equally likely. If its value is 1, then this location is declared to be a break point, and the next nine genes will be assigned -1 to satisfy the minimum length constraint. If this value is -1 , then the procedure will shift to the next gene in line, and continue until all child genes are allocated.

In a mutation operation, one parent chromosome is chosen to generate one child chromosome. The child is mostly identical to its parent, except that random changes are made to a small number of genes. This operation provides the child chromosome with additional freedom to explore the search space, and thus avoids the problem of overly fast convergence to a suboptimal solution. For each child gene, the mutation operation assigns one of the following three possible gene values: (i) with probability π_P it assigns the corresponding δ_i value from its parent, (ii) with probability π_N it assigns the value -1 , and (iii) with probability $1 - \pi_P - \pi_N$ it assigns the value 1. These assignments are subject to the minimum length constraint. In the present implementation, $\pi_P = \pi_N = 0.3$.

Elitist step and island model. To converge to an optimal solution at a faster rate, two additional steps, the *elitist* step and the *island model*, are performed. In the elitist step, the best chromosome of the current generation will replace the worst chromosome of the next generation. This conserves the best chromosome in each generation and guarantees the monotonicity in the search process. In conducting the island model, parallel implementations are applied and N_I (number of islands) GAs are simultaneously performed in N_I different subpopulations. After every J generations, the best H chromosomes from the $(j - 1)$ th island will replace the worst H chromosomes from the j th island, $j = 2, \dots, N_I$ (for $j = 1$ the worst H chromosomes are replaced by the best chromosomes from the N_I th island). The implementation here uses $N_I = 40$, $J = 5$, $H = 2$, and a sub-population size of 40.

Convergence. At the end of each generation the best chromosome that has the smallest MDL value is selected. If this best chromosome remains the same for 20 consecutive generations, then the GA stops and this best chromosome is taken as the final solution to the optimization problem.

5.3 Implementing Variable Selection

This subsection elaborates on the implementation of the GA tailored for variable selection, that is, for minimizing (10).

Chromosome representation. To choose significant variables from data $(X'_1, y_1), \dots, (X'_n, y_n)$, where $X'_i = (X_{i,0}, X_{i,1}, \dots, X_{i,p})'$, a chromosome is expressed as a vector of $p + 1$ integers: a chromosome $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_p)$ is of length $p + 1$ with gene values γ_j defined as

$$\gamma_j = \begin{cases} -1, & \text{if the } j\text{th variable is not chosen,} \\ 1, & \text{if the } j\text{th variable is chosen.} \end{cases}$$

In practice, the constraint is imposed that the number of variables selected is at least 1, and at most $\min\{p + 1, n\}$. It specifies that for the cases where the number of predictors $p + 1$ is greater than the number of observations n , at most n variables will be chosen.

First population generation. For the chromosome $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_p)$, γ_j will be assigned 1 with probability $\pi_B = 10/n$ and -1 with probability $1 - \pi_B$ for each j . If it turns out that all the γ_j 's are assigned -1 , then randomly choose a number j from the set $(1, \dots, p + 1)$ and set $\gamma_j = 1$. For the case where $p + 1 > n$, if it turns out that the number of generated 1's, say, κ , is greater than n in the generated chromosome $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_p)$, then randomly choose n 1's from the generated κ 1's, and let the rest $\kappa - n$ 1's to be -1 .

Crossover and mutation. The crossover and mutation step, including the corresponding probabilities π_C , π_P , and π_N , is the same as the one described for break point detection. It should be noticed that the constraint for the number of selected variables needs to be maintained. If after the crossover and mutation step, the new chromosome has all -1 's or has more than n 1's for the case where the dimension of predictors is greater than the sample size, then a modification process is applied as described in the previous paragraph.

If the number of predictors is small, the elitist step and the island model may not be required. However, they can be applied in a similar manner as before to help expedite searching for the optimal results for cases where there are a large number of predictors. Finally, the same convergence criterion in the previous section can be used.

5.4 Implementing Simultaneous Break Point Detection and Variable Selection

For simultaneous break point detection and variable selection (i.e., minimizing (11)), the above two GA implementations can be combined to form a “nested” GA and carry out the optimization.

As in Section 5.2, for a dataset of length n , the chromosome is represented as $\delta = (\delta_1, \dots, \delta_n)$ with gene value $\delta_t = -1$ if there is no break point at time t and $\delta_t = 1$ if there is a break at time t . The implementation details follow from Section 5.2, but within

each estimated segment in each generated chromosome, not all predictors are selected. Instead, the predictors are selected with the GA described in Section 5.3, and then the MDL value of the resulting chromosome is calculated by the combination of estimated breaks and selected predictors in each segment. In other words, a GA is first used to generate chromosomes with breaks, and then a second GA is applied to each segment of each chromosome to select predictors that minimize the MDL value conditional on the estimated break points. Such a procedure enables us to jointly find the combination of break points and the predictors that (approximately) minimizes (11).

A final remark: when using GAs to solve complicated optimization problems, such as those considered above, very often it is difficult to establish their convergence properties. However, for simpler problems, it has been shown that without using the elitist step GAs will *not* converge to the global optimum, while including this step will guarantee convergence to the global optimum if the algorithm is let run long enough, for example, see Rudolph (1994). Given our extensive numerical experience, we are confident that the above GAs (which use the elitist step) also achieve good convergence properties.

6. SIMULATION RESULTS

6.1 Break Point Detection

In this section, two sets of simulation experiments are conducted to evaluate the empirical performance of the proposed method for break point detection: that is, minimizing the MDL criterion (9) using the GA described in Section 5.2. The performance of the proposed method is compared with the one proposed by Oka and Qu (2011), which is a sequential testing procedure for locating break points. The results are based on $n = 400$ with 500 repetitions, and break points are estimated at the individual quantiles $\tau = 0.25, 0.50$, and 0.75 . For convenience the *relative* locations of break points $\hat{\lambda}_\ell = \hat{k}_\ell/n$, $\ell = 1, \dots, \hat{m}$ are reported.

Experiment 1: Location-shift model. The data are generated according to

$$y_t = \begin{cases} 1 + x_t + \varepsilon_t, & \text{if } 1 \leq t \leq n/2, \\ 1 + 3x_t + \varepsilon_t, & \text{if } n/2 < t \leq n, \end{cases} \quad (13)$$

where $x_t \sim \text{iid } \chi_3^2/3$ and $\varepsilon_t \sim \text{iid AL}(0.5)$. This data-generating process implies that the conditional dispersion of the model remains the same before and after the break point, while the break point occurs to all conditional quantiles at $t = n/2$. The scatterplot of a typical realization is shown in Figure 1, in which dots represent the observations before the break point and circles the observations after the break point.

The simulation results are summarized in Table 1 under the heading “independent x_t ’s.” It can be observed that the proposed method has a higher percentage of correctly identifying the number of break points at each quantile, and Oka and Qu’s method has a slight tendency of detecting an additional false break point.

The same experiment was repeated with correlated predictors. To be precise, the x_t ’s now follow an autoregressive process (AR) of order 1 with coefficient 0.3 and unit innovation variance. The results are summarized in Table 1 under the heading “correlated x_t ’s.” For this setting, Oka and Qu’s method also exhibits a slight tendency of overestimating the number of break points, while the proposed method’s performances deteriorate for $\tau = 0.25$ and 0.75 when comparing with the independent case.

Experiment 2: Location-scale-shift model. Next we consider the following data-generating process

$$y_t = \begin{cases} 1 + x_t + (1 + x_t)\varepsilon_t, & \text{if } 1 \leq t \leq n/2, \\ 1 + 5x_t + (1 + x_t)\varepsilon_t, & \text{if } n/2 < t \leq n, \end{cases} \quad (14)$$

where $\varepsilon_t \sim \text{iid } t(5)$. As in Experiment 1, both independent and correlated x_t ’s were considered: $x_t \sim \text{iid } \chi_3^2/3$ and x_t ’s follow an AR(1) process with coefficient 0.3 and unit innovation

Figure 1. The scatterplots for typical realizations of the data-generating processes (13) (left panel) and (14) (right panel). Dots (circles) represent observations before (after) the break point.

Table 1. Summary of the estimated break points by the proposed method and Oka and Qu’s procedure for model (13). Reported are the percentages of number of times that 0, 1, or 2 break points are detected; see columns under “%.” The means and standard errors (s.e.) of the relative break point locations are also reported

		Independent x_t ’s				Correlated x_t ’s				
		Number of detected break points				Number of detected break points				
		0		1	2	0		1	2	3
		%	%	mean (s.e.)	%	%	%	mean (s.e.)	%	%
Proposed	$\tau = 0.25$	0.6	99.4	0.5081 (0.0243)	0	9.4	90.6	0.5027 (0.0399)	0	0
	$\tau = 0.5$	0	100	0.5006 (0.0152)	0	0	100	0.5012 (0.0191)	0	0
	$\tau = 0.75$	0.6	99.2	0.4942 (0.0250)	0	13	87	0.5053 (0.0378)	0	0
	$\tau = 0.25$	0	92.8	0.5065 (0.0224)	7.2	0	92.8	0.4993 (0.0306)	6.6	0.6
Oka and Qu	$\tau = 0.5$	0	91.8	0.5003 (0.0136)	8.2	0	91.2	0.5000 (0.0229)	8.8	0
	$\tau = 0.75$	0	94.4	0.4945 (0.0219)	5.6	0	92.4	0.4995 (0.0054)	6.4	1.2

variance. Figure 1 shows the scatterplot of a typical realization for the independent case. From Table 2 in which the results are tabulated, it can be seen that the proposed method has a very high accuracy in detecting the number of break points at each quantile, for both independent and correlated cases.

6.2 Variable Selection

Next the practical performance of the proposed variable selection method is evaluated, which minimizes (10) with the GA detailed in Section 5.3. The proposed method is compared with the popular LASSO, SCAD, and adaptive LASSO procedures. For details of these other methods, see Wu and Liu (2009). In each set of experiments, the results are based on $n = 100$ observations with 500 repetitions at the quantiles $\tau = 0.25, 0.50$, and 0.75 . In what follows, the tuning parameters for LASSO, SCAD, and adaptive LASSO are chosen by BIC.

Experiment 3: Classical setup. The data in this example are generated from the linear model

$$y_t = X_t'\beta + \sigma \varepsilon_t,$$

(15)

where $X = (x_1, \dots, x_8)'$ and $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$, and the components of X_t and ε_t are standard normal. The correlation between x_i and x_j is set to $0.5^{|i-j|}$, and two values of σ are tested: 1 and 3. The model (15) has been considered by many authors such as Wu and Liu (2009).

Following Wu and Liu (2009), for each method we record the average number of insignificant predictors correctly removed from the model, and the average number of significant predictors wrongly removed from the model. These averaged numbers, together with their standard deviations, are summarized in Table 3. For $\sigma = 1$, it can be seen that the proposed method perfectly selects the correct number of predictors, while for $\sigma = 3$ the performance slightly deteriorates. However, the results seem to suggest that the proposed method is superior to the other methods compared.

To investigate if there is any advantage of using the proposed method over the other three methods when the error distribution is fat-tailed, the above experiment was repeated with two different error distributions: t_2 and $AL(0.5)$. The results are summarized in Table 4, which seems to suggest that the proposed method is also the preferred method for fat-tailed errors.

Experiment 4: “large p small n .” In this experiment the case is considered when the number of predictors is larger than the sample size. The data are also generated from model (15), but augmented with 112 additional insignificant predictors $x_9, x_{10}, \dots, x_{120}$. These insignificant predictors, as well as the error terms ε_t , follow the $N(0, 0.5^2)$ distribution and they are independent of each other. The results are tabulated in Table 5 in a similar manner as in Table 3. These results suggest that the proposed method outperforms all other variable selection methods.

Table 2. Similar to Table 1 but for model (14)

		Independent x_t ’s				Correlated x_t ’s				
		Number of detected break points				Number of detected break points				
		0		1	2	0		1	2	3
		%	%	mean (s.e.)	%	%	%	mean (s.e.)	%	%
Proposed	$\tau = 0.25$	0	100	0.5017 (0.0082)	0	0	100	0.5021 (0.0048)	0	0
	$\tau = 0.5$	0	100	0.4999 (0.0059)	0	0	100	0.5021 (0.0044)	0	0
	$\tau = 0.75$	0	99.2	0.4989 (0.0087)	0.8	0	100	0.5022 (0.0060)	0	0
	$\tau = 0.25$	0.8	91.0	0.5010 (0.0084)	8.2	0.8	82.6	0.4999 (0.0040)	14.4	3
Oka and Qu	$\tau = 0.5$	0.8	89.8	0.4994 (0.0082)	9.4	0.8	89.2	0.4997 (0.0034)	9	1.8
	$\tau = 0.75$	1.0	91.2	0.4991 (0.0086)	7.8	1	84.4	0.4995 (0.0054)	14.4	1.2

Table 3. Simulation results of variable selection for model (15). Columns with heading “correct”: average number of insignificant predictors correctly removed from the model. Columns with heading “wrong”: average number of significant predictors wrongly removed from the model. Numbers in parentheses are standard deviations

		$\sigma = 1$		$\sigma = 3$	
		Correct	Wrong	Correct	Wrong
$\tau = 0.25$	Proposed	5.00 (0.00)	0.00 (0.00)	4.954 (0.210)	0.384 (0.507)
	LASSO	3.934 (0.862)	0.000 (0.00)	3.304 (0.997)	0.014 (0.118)
	SCAD	3.990 (0.834)	0.000 (0.00)	3.310 (1.033)	0.014 (0.118)
	Adapt-LASSO	4.908 (0.390)	0.000 (0.00)	4.444 (0.827)	0.108 (0.317)
$\tau = 0.5$	Proposed	5.00 (0.00)	0.00 (0.00)	4.956 (0.254)	0.205 (0.445)
	LASSO	3.870 (0.866)	0.000 (0.00)	3.318 (1.039)	0.002 (0.045)
	SCAD	4.114 (0.885)	0.000 (0.00)	3.470 (1.088)	0.008 (0.089)
	Adapt-LASSO	4.936 (0.297)	0.000 (0.00)	4.436 (0.819)	0.086 (0.288)
$\tau = 0.75$	Proposed	5.00 (0.00)	0.00 (0.00)	4.952 (0.362)	0.232 (0.509)
	LASSO	3.884 (0.827)	0.000 (0.00)	3.258 (1.053)	0.016 (0.126)
	SCAD	4.010 (0.853)	0.000 (0.00)	3.270 (1.077)	0.012 (0.109)
	Adapt-LASSO	4.878 (0.451)	0.000 (0.00)	4.354 (0.875)	0.104 (0.318)

Table 4. Similar to Table 3 but for fat-tailed error distributions

		t_2		AL(0.5)	
		Correct	Wrong	Correct	Wrong
$\tau = 0.25$	Proposed	5.00 (0.00)	0.028 (0.165)	4.996 (0.063)	0.196 (0.397)
	LASSO	3.848 (0.859)	0.000 (0.00)	3.592 (0.959)	0.006 (0.077)
	SCAD	3.892 (0.882)	0.000 (0.00)	3.612 (1.020)	0.004 (0.063)
	Adapt-LASSO	4.838 (0.434)	0.002 (0.046)	4.606 (0.710)	0.032 (0.176)
$\tau = 0.5$	Proposed	5.00 (0.00)	0.00 (0.00)	4.99 (0.10)	0.044 (0.205)
	LASSO	3.776 (0.869)	0.000 (0.00)	3.560 (0.901)	0.000 (0.00)
	SCAD	4.064 (0.868)	0.000 (0.00)	3.756 (0.911)	0.002 (0.045)
	Adapt-LASSO	4.922 (0.323)	0.000 (0.00)	4.752 (0.543)	0.016 (0.126)
$\tau = 0.75$	Proposed	5.00 (0.00)	0.026 (0.159)	4.98 (0.140)	0.208 (0.411)
	LASSO	3.842 (0.869)	0.000 (0.00)	3.548 (0.960)	0.002 (0.045)
	SCAD	3.884 (0.892)	0.000 (0.00)	3.584 (0.996)	0.006 (0.996)
	Adapt-LASSO	4.852 (0.398)	0.000 (0.00)	4.612 (0.700)	0.050 (0.218)

6.3 Simultaneous Break Point Detection and Variable Selection

Finally, two sets of experiments are conducted to evaluate the empirical performance of the proposed method for simultaneous

break point detection and variable selection. That is, minimizing (11) using the GA in Section 5.4. As before, the number of repetitions is 500, and the proposed method is applied at the individual quantiles $\tau = 0.25, 0.50$, and 0.75 .

Experiment 5: One break point and two predictors. The data are generated from

$$y_t = \begin{cases} 2x_{t,1} + \varepsilon_t, & \text{if } 1 \leq t \leq n/2, \\ 3x_{t,2} + \varepsilon_t, & \text{if } n/2 < t \leq n, \end{cases} \quad (16)$$

where x_1 and x_2 are iid $\chi^2_3/3$, the error terms ε_t 's are iid standard normal, and $n = 200$. From Table 6, in which the results are reported, it can be seen that the proposed method performs very well in both break point detection and variable selection for this experiment.

To demonstrate the consistency and asymptotic normality of the parameter estimates, additional simulations were ran with $n = 400$ and 800 . For those cases where the models were correctly identified, the parameter estimates of the regression coefficients were recorded. Histograms of these estimates for the cases with $\tau = 0.5$ are given in Figure 2. Histograms for $\tau = 0.25$ and 0.75 are essentially the same and hence are omitted. These histograms provide strong evidence to support the consistency and asymptotic normality of the parameter estimates.

Table 5. Similar to Table 3 but for the “large p , small n ” experiment

		Number of insignificant predictors	
		Correct	Wrong
$\tau = 0.25$	Proposed	117.00 (0.00)	0.00 (0.00)
	LASSO	113.65 (4.30)	0.00 (0.00)
	SCAD	116.72 (1.09)	0.00 (0.00)
	Adapt-LASSO	115.47 (3.16)	0.00 (0.00)
$\tau = 0.5$	Proposed	117.00 (0.00)	0.00 (0.00)
	LASSO	112.37 (5.67)	0.00 (0.00)
	SCAD	116.38 (1.38)	0.00 (0.00)
	Adapt-LASSO	114.54 (4.58)	0.00 (0.00)
$\tau = 0.75$	Proposed	117.00 (0.00)	0.00 (0.00)
	LASSO	112.47 (6.50)	0.00 (0.00)
	SCAD	116.59 (1.18)	0.00 (0.00)
	Adapt-LASSO	115.02 (4.79)	0.00 (0.00)

Figure 2. Histograms of parameter estimates for model (16). Overlaid in each plot is the corresponding asymptotic density function given by Theorem 3. The true values for coefficients 1 and 2 are, respectively, 2 and 3.

Table 6. Simulation results for model (16). For all 500 repetitions the proposed method detected one break point. The mean and standard error (s.e.) of these estimated break point locations are reported. See the caption of Table 3 for the meanings of “correct” and “wrong”

	Estimated break point Mean (s.e.)	First piece		Second piece	
		Correct	Wrong	Correct	Wrong
$\tau = 0.25$	0.5073 (0.0242)	0.98 (0.14)	0.00 (0.00)	0.97 (0.17)	0.00 (0.00)
$\tau = 0.5$	0.4998 (0.0228)	0.99 (0.11)	0.00 (0.00)	0.98 (0.12)	0.00 (0.00)
$\tau = 0.75$	0.4963 (0.0219)	0.96 (0.18)	0.00 (0.00)	0.99 (0.11)	0.00 (0.00)

Table 7. Simulation results for model (17). For all 500 repetitions the proposed method detected two break points. The means and standard errors (s.e., in parentheses) of these estimated break point locations are reported

	First break point	Second break point
$\tau = 0.25$	0.33 (0.016)	0.66 (0.016)
$\tau = 0.5$	0.33 (0.014)	0.67 (0.018)
$\tau = 0.75$	0.33 (0.015)	0.67 (0.016)

Simulation 6: Two break points and three predictors. In this last experiment, the data are generated from

$$y_t = \begin{cases} 5x_{t,1} + \varepsilon_t, & \text{if } 1 \leq t \leq n/3, \\ x_{t,1} + 3x_{t,2} + \varepsilon_t, & \text{if } n/3 < t \leq 2n/3, \\ -2x_{t,2} + x_{t,3} + \varepsilon_t, & \text{if } 2n/3 < t \leq n, \end{cases} \quad (17)$$

where $x_{t,1}$ ’s are iid $\chi^2_3/3$, $x_{t,2}$ ’s and ε_t ’s are iid standard normal, and $n = 300$. The results are summarized in Tables 7 and 8, from which it can be seen that the proposed method also performs very well for this experiment with a more complicated model structure.

7. REAL DATA APPLICATIONS

7.1 Break Point Detection: Hotel Accommodation Cost

The Consumer Price Index (CPI) is an important index in economics. For example, changes in the CPI are known to affect other economic variables such as incomes, cost of living, and others. The dataset analyzed below is made publicly available by the Australian Bureau of Statistics. The response variable is the average monthly cost for one night’s accommodation at a Victoria hotel in Australia, from January 1980 to June 1995. The independent variable is the estimated 1 month percentage change of CPI. This dataset is plotted in the left panel of Figure 3. A simple linear regression model is fitted to the data, and some obvious patterns can be observed in the residual plot shown in Figure 3: if CPI is smaller than 65 or greater than 87, the residuals show a decreasing pattern, while for CPI between 65 and 87, the residuals are mostly negative. This suggests that it may not be appropriate to fit a global model to the whole dataset, and also indicates the existence of potential break points. When estimating at quantiles $\tau = 0.25, 0.50$, and 0.75 , the proposed method divides the series into three segments at each quantile, and the two estimated break points correspond to March 1984 (when $CPI = 65.1$) and March 1988 (when $CPI = 87$), respectively. The estimated coefficients of the fitted piecewise QR model are tabulated in Table 9. It can be seen

Figure 3. Hotel accommodation data. The left panel depicts the observations (circles) with fitted linear regression line (broken line), and the right panel depicts the residuals from the regression. The two vertical broken lines in the right panel indicate the locations of the detected break points.

Table 8. Simulation results for model (17). See the caption of Table 3 for the meanings of “correct” and “wrong”

	First piece		Second piece		Third piece	
	Correct	Wrong	Correct	Wrong	Correct	Wrong
$\tau = 0.25$	1.97 (0.16)	0.00 (0.00)	0.98 (0.14)	0.16 (0.29)	0.98 (0.14)	0.12 (0.22)
$\tau = 0.5$	1.97 (0.18)	0.00 (0.00)	0.95 (0.21)	0.04 (0.19)	0.97 (0.17)	0.03 (0.15)
$\tau = 0.75$	1.97 (0.18)	0.00 (0.00)	0.96 (0.20)	0.04 (0.20)	0.95 (0.22)	0.03 (0.15)

that the slopes are very similar in the first and the third segment, but that the slope of the second segment is much larger. The intercepts in the three pieces are very different.

The break points found by the proposed method can be explained by changes in the economic environment. Badly affected by the recession of 1982–1983, unemployment in Australia rose to almost 10%, to its highest level since the 1920s’ Great Depression. Following the election of a new government in 1983, the Australian dollar was floated, and its value was allowed to fluctuate, depending on the international money markets. Intended to diversify the national economy and make it more resilient, economic reforms were implemented after the mid-1980s decline, including a series of reductions in tariffs across most industry sectors introduced by the government in 1988.

7.2 Variable Selection: Boston Housing Data

To illustrate the usefulness of the proposed method for variable selection, the Boston housing data is considered. This dataset has been analyzed by many researchers, including Fan and Huang (2005), Wang and Xia (2009), Wu and Liu (2009), among others. It contains several variables for owner-occupied homes in 506 U.S. census tracts in the Boston area, back in 1970. Following Fan and Huang (2005), the variable MEDV (median value of owner-occupied homes in 1000 USD) is taken as the response, and the following seven variables are taken as explanatory variables: AGE (proportion of owner-occupied units built prior to 1940), CRIM (per capita crime rate by town), LSTAT (the percentage of lower status of the population), NOX (nitric oxides concentration), PTRATIO (pupil–teacher ratio by town), RM (average number of rooms per dwelling), and TAX (full-value property-tax rate per 10,000 USD).

The variables selected over different quantiles by the proposed method are listed in Table 10. The results indicate that at different quantiles, the median value of owner-occupied homes is associated with different predictors. For each quantile, RM, PTRATIO, and LSTAT are common predictors that are important for housing prices. This may be because size of the house, quality of the school districts, and income of the residents di-

rectly affect the housing prices, no matter at which price level. The variable TAX is selected only in the quantiles 0.25 and 0.50, arguably because the property tax rate is not an important factor for people who can afford to buy expensive homes. The variable CRIM is selected only at quantile 0.25. While it is generally considered to be a crucial factor for housing prices, its effect is here only significant for houses that are comparatively cheap. This could be explained by the possibility that expensive houses are mostly in parts of town where the crime rates are low, and hence the relationship between prices of expensive homes and high crime rates are not observable.

The variable selection methods LASSO and SCAD were also applied to this dataset. It turns out that LASSO selects all the variables except “NOX” at each tested quantile; SCAD selects all the variables except “NOX” at quantile 0.25, and selects all seven variables at quantiles 0.50 and 0.75. These results are in line with the observations from the simulation results in Section 6.2, which have shown that both LASSO and SCAD tend to keep more variables (sometimes insignificant ones) in the model. However, the variables selected by the proposed method seem to be more aligned with intuition.

7.3 Break Point Detection and Variable Selection: U.S. Interest Rates

In this last real data example, the data are quarterly interest rates in the United States from 1964 to 1991. The goal is to estimate break points and conduct variable selection concurrently. There are 110 observations in the dataset. As response variable (y) the interest rate in the United States is used, and there are two predictor variables: money supply (x_1) and federal cyclically adjusted budget (x_2). The proposed method was applied at quantiles $\tau = 0.25, 0.50$, and 0.75 , and one break point was detected for all quantiles at $t = 60$, which corresponds to the fourth quarter of 1978. From the correlation scatterplots shown in Figure 4, we see that for each segment, the correlations between the response variable and each predictor variable change dramatically. Before 1978, money supply and interest rate are positively correlated, and are negatively correlated thereafter.

Table 9. Estimated coefficients for the hotel accommodation data

	First piece		Second piece		Third piece	
	Intercept	Slope	Intercept	Slope	Intercept	Slope
$\tau = 0.25$	1.333	0.586	−20.271	0.935	13.008	0.628
$\tau = 0.5$	1.047	0.604	−20.559	0.954	9.818	0.674
$\tau = 0.75$	−0.061	0.635	−22.71	1.00	11.276	0.676

Table 10. Selected variables by the proposed method for the Boston housing data

Quantile	Selected variables
$\tau = 0.25$	RM, PTRATIO, LSTAT, TAX, CRIM
$\tau = 0.5$	RM, PTRATIO, LSTAT, TAX
$\tau = 0.75$	RM, PTRATIO, LSTAT

Figure 4. U.S. interest rate data. The left/right panel is the scatterplot matrix for cross-correlation among variables for the first/second segment.

Table 11. Selected variables for each segment in the interest rate data

	First piece		Second piece	
	x_1	x_2	x_1	x_2
$\tau = 0.25$	Yes	No	Yes	No
$\tau = 0.5$	Yes	No	Yes	No
$\tau = 0.75$	Yes	Yes	Yes	No

The variables selected are not the same at each tested quantile for each segment. The results are tabulated in Table 11. It can be seen that money supply x_1 is always important for the change of interest rate at each quantile, while federal cyclically adjusted budget x_2 is selected only at $\tau = 0.75$ in the QR model before the break point.

8. CONCLUSIONS

This article proposes a new methodology to simultaneously (or separately) detect break points, conduct variable selection, and estimate parameters in QR models. This is done through an application of the MDL principle and a GA. Several desirable large-sample properties have been established. Simulation results have demonstrated that the finite sample performance of the proposed procedure is quite good, and several data applications are also provided with satisfactory results.

The methodology can be straightforwardly modified to handle mean regression, say with Gaussian errors. The only change is to replace the likelihood term in the corresponding MDL criterion with a Gaussian likelihood. For example, the last term of (11) would be replaced by a standardized residual sum of squares term. For the mean regression case, theoretical results on the consistency of the break point estimator $\hat{m}$ may be derived more straightforwardly.

APPENDIX: PROOFS

A.1 Proofs of Section 4.1

Lemma 1. If $((X'_t, y_t) : t \in \mathbb{Z})$ follows the QR model (1) such that the conditions of Assumption 1 are satisfied, then with probability one and for all $\tau \in (0, 1)$ and $\kappa \in [0, 1]$,

$$\frac{1}{n} \sum_{t=1}^{\lfloor \kappa n \rfloor} \rho_\tau(\hat{\varepsilon}_t) \rightarrow \kappa E[\rho_\tau(\varepsilon_1)] \quad (n \rightarrow \infty),$$

where ρ_τ is the check function defined below (3).

Proof. The statement of the lemma is obviously correct if $\kappa = 0$. Fix $\tau \in (0, 1)$ and $\kappa \in (0, 1]$. Note that the QR equations imply that $\hat{\varepsilon}_t = \varepsilon_t + X'_t(\hat{\theta}(\tau) - \theta(\tau))$. It follows from Bassett and Koenker (1986) that $\hat{\theta}(\tau)$ is strongly consistent for $\theta(\tau)$. Therefore, with probability one,

$$\frac{1}{\lfloor \kappa n \rfloor} \left| \sum_{t=1}^{\lfloor \kappa n \rfloor} (\hat{\varepsilon}_t - \varepsilon_t) \right| = \frac{1}{\lfloor \kappa n \rfloor} \left| \sum_{t=1}^{\lfloor \kappa n \rfloor} X'_t(\hat{\theta}(\tau) - \theta(\tau)) \right| \rightarrow 0.$$

Consequently, $\frac{1}{\lfloor \kappa n \rfloor} \sum_{t=1}^{\lfloor \kappa n \rfloor} \hat{\varepsilon}_t \rightarrow E[\varepsilon_1]$ with probability one by the strong law of large numbers. Since $\rho_\tau(\cdot)$ is a continuous and measurable function, the latter limit relation is also true if ε_t is replaced by $\rho_\tau(\varepsilon_t)$. $\square$

Lemma 2. If $((X'_t, y_t) : t \in \mathbb{Z})$ follows a piecewise QR model such that Assumption 1 is satisfied for each of the segments and if $\lambda^0 = (\lambda_1^0, \dots, \lambda_{m^0}^0)$ denotes the true segmentation, then, with probability one and for all $\tau \in (0, 1)$,

$$\frac{1}{M - K} \sum_{t=K+1}^M \rho_\tau(\hat{\varepsilon}_t) \rightarrow L_\tau(\kappa, \mu) + B_\tau(\kappa, \mu),$$

where $K = \lfloor \kappa n \rfloor$, $M = \lfloor \mu n \rfloor$ with $0 \leq \kappa < \mu \leq 1$ and $L_\tau(\kappa, \mu)$ and $B_\tau(\kappa, \mu)$ being defined in the proof.

Proof. If, for $0 \leq \kappa < \mu \leq 1$, there are $1 \leq \ell < \ell' \leq m^0 + 1$ such that $\kappa \in [\lambda_{\ell-1}^0, \lambda_\ell^0)$ and $\mu \in (\lambda_{\ell'-1}^0, \lambda_{\ell'}^0]$, then $\rho_\tau(\hat{\varepsilon}_t) = [\rho_\tau(\hat{\varepsilon}_t) - \rho_\tau(\hat{\varepsilon}_{t,j})] + \rho_\tau(\hat{\varepsilon}_{t,j})$, $t = k_{j-1}^0 + 1, \dots, k_j^0$, and $j = 1, \dots, m^0 + 1$,

where $\hat{\varepsilon}_t$ is the residual from the fit-based observations $y_{K+1}, \dots, y_M$, $\hat{\varepsilon}_{t,j}$ the residual from a fit based on the j th true segment and $k_j^0 = \lfloor \lambda_j^0 n \rfloor$. An application of Lemma 1 yields that, with probability one,

$$\begin{aligned} & \frac{1}{M-K} \left[\sum_{t=K+1}^{k_\ell^0} \rho_\tau(\hat{\varepsilon}_{t,\ell}) + \sum_{j=\ell+1}^{\ell'-1} \sum_{t=k_{j-1}^0+1}^{k_j^0} \rho_\tau(\hat{\varepsilon}_{t,j}) + \sum_{t=k_{\ell'-1}^0+1}^M \rho_\tau(\hat{\varepsilon}_{t,\ell'}) \right] \\ & \rightarrow \frac{1}{\mu - \kappa} \left[(\lambda_\ell^0 - \kappa) \rho_{\tau,\ell} + \sum_{j=\ell+1}^{\ell'-1} (\lambda_j^0 - \lambda_{j-1}^0) \rho_{\tau,j} + (\mu - \lambda_{\ell'-1}^0) \rho_{\tau,\ell'} \right] \\ & = L_\tau(\kappa, \mu), \end{aligned}$$

where $\rho_{\tau,\ell} = E[\rho_\tau(\varepsilon_{t,\ell})]$. On the other hand, the average of the terms $\Delta \rho_{\tau,t,j} = \rho_\tau(\hat{\varepsilon}_t) - \rho_\tau(\hat{\varepsilon}_{t,j})$ converges to a nonnegative bias term on each of the segments, leading to

$$\begin{aligned} & \frac{1}{M-K} \left[\sum_{t=K+1}^{k_\ell^0} \Delta \rho_{\tau,t,\ell} + \sum_{j=\ell+1}^{\ell'-1} \sum_{t=k_{j-1}^0+1}^{k_j^0} \Delta \rho_{\tau,t,j} + \sum_{t=k_{\ell'-1}^0+1}^M \Delta \rho_{\tau,t,\ell'} \right] \\ & \rightarrow B_\tau(\kappa, \mu), \end{aligned}$$

where $B_\tau(\kappa, \mu)$ is a bias term whose exact form depends on the true segmentation and the distribution of the y_t on these segments. Note that $B_\tau(\kappa, \mu) > 0$ unless both κ and μ coincide with two adjacent true break locations and that $B_\tau(\kappa, \mu)$ is unique due to part (b) of Assumption 1. Similarly, if $\ell = \ell'$, it holds with probability one that

$$\frac{1}{M-K} \sum_{t=K+1}^M \rho_\tau(\hat{\varepsilon}_t) \rightarrow \rho_{\tau,\ell} = L_\tau(\kappa, \mu) + 0,$$

which completes the proof. $\square$

Proof of Theorem 1. Fix $\tau \in (0, 1)$. Denote the segmentation chosen by the MDL criterion (9) and the true segmentation by $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_{m^0})$ and $\lambda^0 = (\lambda_1^0, \dots, \lambda_{m^0}^0)$, respectively. Assume that $\hat{\lambda}$ does not converge with probability one to λ_0 . By boundedness, there is a subsequence along which $\hat{\lambda}$ converges with probability one to a limit, say, $\lambda^* = (\lambda_1^*, \dots, \lambda_{m^0}^*) \neq \lambda^0$. If $\lambda_{\ell'-1}^0 \leq \lambda_{\ell-1}^* < \lambda_\ell^* \leq \lambda_{\ell'}^0$, then Lemma 2 implies that, with probability one along the subsequence,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=k_{\ell-1}^0+1}^{k_\ell^0} \rho_\tau(\hat{\varepsilon}_t) = (\lambda_\ell^* - \lambda_{\ell-1}^*) \rho_{\tau,\ell'}, \quad \rho_{\tau,\ell'} = E[\rho_\tau(\varepsilon_{k_\ell^0})].$$

Similarly, if $\lambda_{\ell'-1}^0 \leq \lambda_{\ell-1}^* < \lambda_{\ell'}^0 < \dots < \lambda_{\ell'+L}^0 < \lambda_\ell^* \leq \lambda_{\ell'+L+1}^0$ for some positive integer L , then, with probability one,

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=k_{\ell-1}^0+1}^{k_\ell^0} \rho_\tau(\hat{\varepsilon}_t) \\ & > \frac{1}{\lambda_\ell^* - \lambda_{\ell-1}^*} \left[(\lambda_{\ell'}^0 - \lambda_{\ell-1}^*) \rho_{\tau,\ell'} + \sum_{j=\ell'+1}^{\ell'+L} (\lambda_j^0 - \lambda_{j-1}^0) \rho_{\tau,j} \right. \\ & \quad \left. + (\lambda_\ell^* - \lambda_{\ell'+L}^0) \rho_{\tau,\ell'+L+1} \right]. \end{aligned}$$

Combining the two preceding displays and using the fact that $\lambda_\ell^* - \lambda_{\ell-1}^* < 1$, it follows with probability one that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n} \text{MDL}(m^0, \hat{\lambda} | \tau) \\ & = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^{m^0+1} \sum_{\ell=k_{\ell-1}^0+1}^{k_\ell^0} \rho_\tau(\hat{\varepsilon}_t) \\ & > \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{\ell=1}^{m^0+1} \sum_{t=k_{\ell-1}^0+1}^{k_\ell^0} \rho_\tau(\hat{\varepsilon}_t) = \lim_{n \rightarrow \infty} \frac{1}{n} \text{MDL}(m^0, \lambda^0 | \tau), \end{aligned}$$

a contradiction. Hence $\hat{\lambda} \rightarrow \lambda^0$ with probability one and Theorem 1 is established. $\square$

A.2 Proofs of Section 4.2

The main difficulty in the proofs related to the variable selection are in the nondifferentiability of the check function $\rho_\tau(\cdot)$. To circumvent this issue, the convexity arguments put forward in Pollard (1991) are applied. As in Wu and Liu (2009), let $D_t = (1 - \tau)\{\varepsilon_t < 0\} - \tau\{\varepsilon_t \geq 0\}$. Since the τ th quantile of ε_t is zero, it follows that $E[D_t] = 0$. Let now further $W_n = \sum_{t=1}^n D_t X_t / \sqrt{n}$, and $W_{n,11} = \sum_{t=1}^n D_t X_{t,1} / \sqrt{n}$, then the multivariate central limit theorem (CLT) implies that $W_n \xrightarrow{D} N(0, \tau(1 - \tau)\Sigma)$ and $W_{n,11} \xrightarrow{D} N(0, \tau(1 - \tau)\Sigma_{11})$.

Proof of Theorem 2. It suffices to verify that for $\delta > 0$, there exists a constant C such that

$$P\left(\inf_{\|u\|=C} \text{MDL}\left(s, \theta^0(\tau) + \frac{u}{\sqrt{n}} \middle| \tau\right) > \text{MDL}(s^0, \theta^0(\tau) | \tau)\right) \geq 1 - \delta, \quad (\text{A.1})$$

which yields the existence of a local minimizer in the compact set $\{\theta^0(\tau) + u/\sqrt{n} : \|u\| \leq C\}$ and consequently implies the assertion of the theorem. Now, for $s \geq s^0$ and fixed u ,

$$\begin{aligned} & \text{MDL}\left(s, \theta^0(\tau) + \frac{u}{\sqrt{n}} \middle| \tau\right) - \text{MDL}(s^0, \theta^0(\tau) | \tau) \\ & \geq \sum_{t=1}^n \left[\rho_\tau\left(y_t - X_t' \left(\theta^0(\tau) + \frac{u}{\sqrt{n}}\right)\right) - \rho_\tau(y_t - X_t' \theta^0(\tau)) \right] \\ & = \frac{f(0)}{2} u' \left(\frac{1}{n} \sum_{t=1}^n X_t X_t' \right) u + W_n' u + o_p(1), \end{aligned}$$

where the last equation follows as in Lemma 3 of Wu and Liu (2009) and the $o_p(1)$ rate is pointwise in u . Note that part (a) of Assumptions 1 implies that, up to a $o(1)$ term, $Q_n(u) = 0.5 f(0) u' \left(\frac{1}{n} \sum_{t=1}^n X_t X_t' \right) u$ is the mean of $G_n(u) = \sum_{t=1}^n [\rho_\tau(y_t - X_t'(\theta^0(\tau) + u/\sqrt{n})) - \rho_\tau(y_t - X_t' \theta^0(\tau))]$. It follows with part (b) of Assumption 1 that $h_n(u) = G_n(u) - W_n' u$ converges in probability to $h(u) = 0.5 f(0) u' \Sigma u$. Noting that h_n is convex, an application of the convexity lemma (Pollard 1991) strengthens this statement to uniform convergence in probability on compact subsets of $\mathbb{R}^p$, that is, $\sup_{u \in K} |h_n(u) - h(u)| \rightarrow 0$ in probability for any compact $K \subset \mathbb{R}^p$. Since $E[W_n' u] = 0$ and $\text{var}(W_n' u) = \tau(1 - \tau) u' \left(\frac{1}{n} \sum_{t=1}^n X_t X_t' \right) u$, it follows that $\text{MDL}(s, \theta^0(\tau) + u/\sqrt{n} | \tau) - \text{MDL}(s^0, \theta^0(\tau) | \tau)$ is dominated with probability approaching one by the quadratic term $Q_n(u)$ if $C = \|u\|$ is sufficiently large. Part (b) of Assumption 1 yields then that (A.1) holds. $\square$

Lemma 3. Let $\tilde{\theta}(\tau) = (\tilde{\theta}_1(\tau)', \tilde{\theta}_2(\tau)')'$ be such that $\sqrt{n} \|\tilde{\theta}_1(\tau) - \tilde{\theta}_1^0(\tau)\| = O_p(1)$. If Assumption 1 is satisfied, then, with probability approaching one,

$$\text{MDL}(s^0, (\tilde{\theta}_1(\tau)', 0')' | \tau) = \min_{\sqrt{n} \|\tilde{\theta}_2(\tau)\| \leq C} \text{MDL}(s, \tilde{\theta}(\tau) | \tau)$$

for any $C > 0$.

Proof. Fix $\tau \in (0, 1)$. Let $G_n(u)$ and $Q_n(u)$ be defined as in the proof of Theorem 2 and set $u_1 = \sqrt{n}((\tilde{\theta}_1(\tau) - \tilde{\theta}_1^0(\tau))', 0')'$ and $u_2 =$

$\sqrt{n}((\tilde{\theta}_1(\tau) - \tilde{\theta}_1^0(\tau))', \tilde{\theta}_2(\tau))' | \tau)$. Then,

$$\begin{aligned} & \text{MDL}(s^0, (\tilde{\theta}_1(\tau)', 0')' | \tau) - \text{MDL}(s, \tilde{\theta}(\tau) | \tau) \\ &= [\text{MDL}(s^0, (\tilde{\theta}_1(\tau)', 0')' | \tau) - \text{MDL}(s^0, (\tilde{\theta}_1^0(\tau)', 0')' | \tau)] \\ &\quad - [\text{MDL}(s, \tilde{\theta}(\tau) | \tau) - \text{MDL}(s^0, (\tilde{\theta}_1^0(\tau)', 0')' | \tau)] \\ &= G_n(u_1) - G_n(u_2) - \left[\frac{s - s^0}{2} \log_2 n + (s - s^0) \log_2(p + 1) \right] \\ &= Q_n(u_1) - Q_n(u_2) + W_n'(u_1 - u_2) \\ &\quad - \left[\frac{s - s^0}{2} \log_2 n + (s - s^0) \log_2(p + 1) \right] + o_p(1). \end{aligned}$$

The $\sqrt{n}$ -consistency of $\tilde{\theta}_1(\tau)$ and the condition $0 < \sqrt{n} \|\tilde{\theta}_2(\tau)\| \leq C$ show that $G_n(u_1) = O_p(1)$, $G_n(u_2) = O_p(1)$, and $W_n'(u_1 - u_2) = -\sqrt{n} W_n'(0', \tilde{\theta}_2(\tau)') = (n\tau(1 - \tau)\Sigma_{22}\tilde{\theta}_2(\tau))^{1/2}(1 + o_p(1))$. Since $-\left[\frac{s - s^0}{2} \log_2 n + (s - s^0) \log_2(p + 1)\right] \rightarrow -\infty$ if $s > s^0$, the lemma is established. $\square$

Proof of Theorem 3.

- (a) This is a direct consequence of Lemma 3.
 (b) It can be seen from the proof of Theorem 2 that $\sqrt{n}(\tilde{\theta}_1(\tau) - \tilde{\theta}_1^0(\tau))$ minimizes the function

$$H_n(v) = G_n((v', 0')') + \log_2 n + \frac{s + 2}{2} \log_2 n + s \log_2(p + 1),$$

where G_n is as before and $v \in \mathbb{R}^{s^0}$. Setting $u = (v', 0')'$, the preceding proofs imply that

$$\begin{aligned} G_n(u) &= Q_n(u) + W_n'u + o_p(1) = Q_n(v) \\ &\quad + \frac{v'}{\sqrt{n}} \sum_{i=1}^n D_i X_{i,1} + o_p(1), \end{aligned}$$

where the convergence is uniform on compact subsets of $\mathbb{R}^{s^0}$. Consequently,

$$\begin{aligned} H_n(v) &= Q_n(v) + \frac{v'}{\sqrt{n}} \sum_{i=1}^n D_i X_{i,1} + \log_2 n + \frac{s + 2}{2} \log_2 n \\ &\quad + s \log_2(p + 1) + r_n(v) \\ &= Q_n(v - w_n) - Q_n(w_n) + \log_2 n + \frac{s + 2}{2} \log_2 n \\ &\quad + s \log_2(p + 1) + r_n(v), \end{aligned}$$

where $w_n = -n(f(0) \sum_{i=1}^n X_{i,1} X_{i,1}')^{-1} W_{n,11}$ and the residual $r_n(v) \rightarrow 0$ in probability uniformly on any compact subset of $\mathbb{R}^{s^0}$. Since the penalty term $\log_2 n + (s + 2) \log_2 n/2 + s \log_2(p + 1)$ does not depend on v , it follows that the local minimizer satisfies $\|v - w_n\| = o_p(1)$. This implies that $\sqrt{n}(\tilde{\theta}_1(\tau) - \tilde{\theta}_1^0(\tau)) = -n(f(0) \sum_{i=1}^n X_{i,1} X_{i,1}')^{-1} \sum_{i=1}^n D_i X_{i,1} / \sqrt{n} + o_p(1)$. After an application of Slutsky's theorem yields $\sqrt{n}f(0)\Sigma_{11}(\tilde{\theta}_1(\tau) - \tilde{\theta}_1^0(\tau)) \xrightarrow{D} N(0, \tau(1 - \tau)\Sigma_{11})$, the proof is complete. $\square$

Proof of Corollary 1. The arguments are similar to those employed in the proof of Theorem 2 in Pollard (1991). $\square$

A.3 Proofs of Section 4.3

Proof of Theorem 4. The assertion can be verified by combining the proofs of the previous two subsections. Details are omitted to conserve space. $\square$

REFERENCES

- Aue, A., and Horváth, L. (2013), "Structural Breaks in Time Series," *Journal of Time Series Analysis*, 34, 1–16. [1241]
- Aue, A., and Lee, T. C. M. (2011), "On Image Segmentation Using Information Theoretic Criteria," *The Annals of Statistics*, 39, 2912–2935. [1244]
- Bai, J. (1998), "Estimation of Multiple-Regime Regressions With Least Absolute Deviation," *Journal of Statistical Planning and Inference*, 74, 103–134. [1241, 1244]
- Bassett, G., and Koenker, R. (1986), "Strong Consistency of Regression Quantiles and Related Empirical Processes," *Econometric Theory*, 2, 191–201. [1241, 1253]
- Csörgő, M., and Horváth, L. (1997), *Limit Theorems in Change-Point Analysis*, New York: Wiley. [1241]
- Davis, L. (1991), *Handbook of Genetic Algorithms*, New York: Van Nostrand Reinhold. [1242, 1245]
- Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006), "Structural Break Estimation for Non-Stationary Time Series Models," *Journal of the American Statistical Association*, 101, 223–239. [1245]
- Davis, R. A., and Yau, Y. (2013), "Consistency of Minimum Description Length Model Selection for Piecewise Stationary Time Series Models," *Electronic Journal of Statistics*, 7, 381–411. [1244]
- Fan, J., and Huang, T. (2005), "Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models," *Bernoulli*, 11, 1031–1057. [1252]
- Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1241]
- Geraci, M., and Bottai, M. (2007), "Quantile Regression for Longitudinal Data Using the Asymmetric Laplace Distribution," *Biostatistics*, 8, 140–154. [1242]
- Hansen, M. H., and Yu, B. (2001), "Model Selection and the Principle of Minimum Description Length," *Journal of the American Statistical Association*, 96, 746–774. [1243]
- Koenker, R. (2005), *Quantile Regression*, New York: Cambridge University Press. [1241, 1242]
- Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50. [1241, 1242]
- Koenker, R., and Geling, R. (2001), "Reappraising Medfly Longevity: A Quantile Regression Survival Analysis," *Journal of the American Statistical Association*, 96, 458–468. [1241]
- Koenker, R., and Xiao, Z. (2002), "Inference on the Quantile Regression Process," *Econometrica*, 70, 1583–1612. [1241]
- Lee, T. C. M. (2001), "An Introduction To Coding Theory and the Two-Part Minimum Description Length Principle," *International Statistical Review*, 69, 169–183. [1243]
- Li, C., Wei, Y., Chappell, R., and He, X. (2011), "Bent Line Quantile Regression With Application to an Allometric Study of Land Mammals' Speed And Mass," *Biometrics*, 67, 242–249. [1241]
- Li, Y., Liu, Y., and Zhu, J. (2007), "Quantile Regression in Reproducing Kernel Hilbert Spaces," *Journal of the American Statistical Association*, 102, 255–268. [1241]
- Li, Y., and Zhu, J. (2008), " L_1 -Norm Quantile Regression," *Journal of Computational and Graphical Statistics*, 17, 163–185. [1241]
- Oka, T., and Qu, Z. (2011), "Estimating Structural Changes in Regression Quantiles," *Journal of Econometrics*, 162, 248–267. [1241, 1242, 1244, 1247]
- Pollard, D. (1991), "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory*, 7, 186–199. [1241, 1244, 1245, 1254, 1255]
- Qu, Z. (2008), "Testing for Structural Change in Regression Quantiles," *Journal of Econometrics*, 146, 170–184. [1241, 1242]
- Rissanen, J. (1989), *Stochastic Complexity in Statistical Inquiry*, Singapore: World Scientific. [1241, 1243]
- (2007), *Information and Complexity in Statistical Modeling*, New York: Springer. [1241, 1243]
- Rudolph, G. (1994), "Convergence Analysis of Canonical Genetic Algorithms," *IEEE Transactions on Neural Networks*, 5, 96–101. [1247]
- Su, L., and Xiao, Z. (2008), "Testing for Parameter Stability in Quantile Regression Models," *Statistics and Probability Letters*, 78, 2768–2775. [1241]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [1241]
- Wang, H., and He, X. (2007), "Detecting Differential Expressions in Genechip Microarray Studies: A Quantile Approach," *Journal of the American Statistical Association*, 102, 104–112. [1241]
- Wang, H., and Xia, Y. (2009), "Shrinkage Estimation of the Varying Coefficient Model," *Journal of the American Statistical Association*, 104, 747–757. [1241]

- Wang, H., and Xia, Y. (2009), "Shrinkage Estimation of the Varying Coefficient Model," *Journal of the American Statistical Association*, 104, 747–757. [1252]
- Wu, Y., and Liu, Y. (2009), "Variable Selection in Quantile Regression," *Statistica Sinica*, 19, 801–817. [1241, 1248, 1252, 1254]
- Yu, K., Lu, Z., and Stander, J. (2003), "Quantile Regression: Applications and Current Research Areas," *Journal of the Royal Statistical Society, Series D*, 52, 331–350. [1242]
- Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1241]
- Zou, H., and Yuan, M. (2008), "Regularized Simultaneous Model Selection in Multiple Quantiles Regression," *Computational Statistics and Data Analysis*, 52, 5296–5304. [1241]