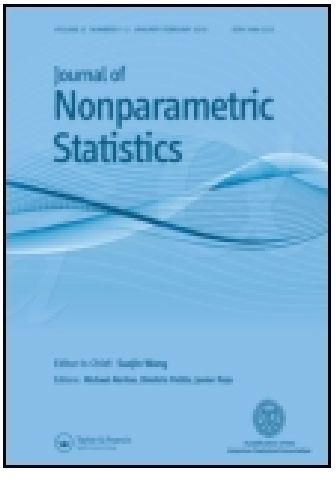
This article was downloaded by: [Portland State University]
On: 01 December 2014, At: 22:38
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gnst20

Simultaneous multiple non-crossing quantile regression estimation using kernel constraints

Yufeng Liu [a] & Yichao Wu [b]

[a] Department of Statistics and OR, Carolina Center for Genome Sciences , University of North Carolina , 354 Hanes Hall, CB 3260, Chapel Hill, NC, 27599, USA

[b] Department of Statistics , North Carolina State University , Raleigh, NC, 27695, USA

Published online: 12 Jan 2011.

To cite this article: Yufeng Liu & Yichao Wu (2011) Simultaneous multiple non-crossing quantile regression estimation using kernel constraints, Journal of Nonparametric Statistics, 23:2, 415-437, DOI: 10.1080/10485252.2010.537336

To link to this article: http://dx.doi.org/10.1080/10485252.2010.537336

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms &

Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Downloaded by [Portland State University] at 22:38 01 December 2014

Journal of Nonparametric Statistics
Vol. 23, No. 2, June 2011, 415–437

Taylor & Francis
Taylor & Francis Group

Simultaneous multiple non-crossing quantile regression estimation using kernel constraints

Yufeng Liu[a] and Yichao Wu[b]*

[a]*Department of Statistics and OR, Carolina Center for Genome Sciences, University of North Carolina, 354 Hanes Hall, CB 3260, Chapel Hill, NC 27599, USA;* [b]*Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA*

(*Received 23 January 2010; final version received 26 October 2010*)

Quantile regression (QR) is a very useful statistical tool for learning the relationship between the response variable and covariates. For many applications, one often needs to estimate multiple conditional quantile functions of the response variable given covariates. Although one can estimate multiple quantiles separately, it is of great interest to estimate them simultaneously. One advantage of simultaneous estimation is that multiple quantiles can share strength among them to gain better estimation accuracy than individually estimated quantile functions. Another important advantage of joint estimation is the feasibility of incorporating simultaneous non-crossing constraints of QR functions. In this paper, we propose a new kernel-based multiple QR estimation technique, namely simultaneous non-crossing quantile regression (SNQR). We use kernel representations for QR functions and apply constraints on the kernel coefficients to avoid crossing. Both unregularised and regularised SNQR techniques are considered. Asymptotic properties such as asymptotic normality of linear SNQR and oracle properties of the sparse linear SNQR are developed. Our numerical results demonstrate the competitive performance of our SNQR over the original individual QR estimation.

Keywords: asymptotic normality; kernel; multiple quantile regression; non-crossing; oracle property; regularisation; variable selection

Downloaded by [Portland State University] at 22:38 01 December 2014

1. Introduction

Regression is central to statistics. Different from ordinary least squares regression, quantile regression (QR) tries to estimate the conditional quantile function. It was originally introduced by Koenker and Bassett (1978) and has been extensively studied afterwards. It has been applied in many different areas. Interested readers are referred to Koenker (2005) for a comprehensive review on QR.

Many real applications ask for a complete understanding of the conditional distribution of the response given covariates. One approach is to estimate multiple conditional quantile functions. A naive method is to individually estimate different conditional quantile functions. This individual estimation method is simple and easy to carry out. Theoretically different conditional quantile

*Corresponding author. Email: wu@stat.ncsu.edu

ISSN 1048-5252 print/ISSN 1029-0311 online
© American Statistical Association and Taylor & Francis 2011
DOI: 10.1080/10485252.2010.537336
http://www.informaworld.com

functions should not cross each other according to the basic principle of conditional distribution functions. However, the naive individual estimation may lead to estimated conditional quantile functions that cross each other. Thus, it is desirable to jointly estimate multiple QR with non-crossing constraints embedded. Another important motivation of joint estimation is that multiple quantile functions may share strength among them (Zou and Yuan 2008a). As a result, simultaneous estimation may help to improve the estimation accuracy of an individual quantile function.

In the literature, there exist some techniques addressing the crossing issue of multiple quantile function estimation. He (1997) proposed the location-scale shift model to impose monotonicity across the quantile functions. However, as noted by Neocleousa and Portnoy (2007), even for linear regression quantiles, corresponding models can be much more general. Thus, a more general development of the estimation of non-crossing regression quantiles is needed. Shim, Hwang and Seok (2009) also considered the location-scale model and proposed to estimate both location and scale functions simultaneously by doubly penalised kernel machines to achieve non-crossing of quantiles. Takeuchi, Le, Sears and Smola (2006) proposed to impose non-crossing constraints on the data points. Although their approach can help to reduce the chance of crossing, their data constraints may not ensure non-crossing in the entire covariate space. Takeuchi and Furuhashi (2004) further extended the method of Takeuchi et al. (2006) by using the ϵ-insensitive check function in the support vector machine framework. Recently, Wu and Liu (2009) proposed a stepwise procedure to perform the estimation of multiple non-crossing QR functions. Despite improvement over individually estimated quantile functions, the stepwise procedure does not produce a simultaneous estimation. In a recent paper, Bondell, Reich and Wang (2010) proposed an method for non-crossing quantile regression curve estimation using spline-based constraints.

For nonparametric non-crossing quantile estimation, several people have proposed to first estimate the conditional cumulative distribution function via local weighting and then invert it to obtain the quantile curve. Yu and Jones (1998) suggested a double kernel smoothing method with a minor modification of this estimate in a second step, so that the corresponding quantile curves are monotone. Hall, Wolff and Yao (1999) proposed an adjusted Nadaraya–Watson estimate, which modifies the weights of the Nadaraya–Watson estimate such that the resulting estimate of the conditional distribution function is monotone. Chernozhukov, Fernandez-Val and Galichon (2009) proposed to estimate non-crossing quantile curves via a monotonic rearrangement of the original non-monotone function. They also studied the asymptotic behaviour of their bootstrap-type method. Dette and Volgushev (2008) proposed a similar approach to achieve non-crossing quantile curves via solving the problem of inversion and monotonisation on the initial estimates. Although these indirect approaches are effective in obtaining nonparametric quantile curves without crossing, it can be difficult to quantify the effect of the predictors. For instance, if variable selection is a desirable goal, a direct approach is needed to estimate multiple non-crossing quantiles.

In this paper, we propose a new method to perform simultaneous estimation of multiple non-crossing conditional quantile functions. We call the method simultaneous non-crossing quantile regression (SNQR). We employ simple constraints on the kernel coefficients which can guarantee the estimated conditional quantile functions never cross each other. This kernel formulation covers both linear and nonlinear models. Furthermore, we demonstrate that through sharing strength among different quantiles, SNQR can produce better estimation than individually estimated quantile functions. We have also developed asymptotic normality of linear SNQR and oracle properties of the sparse linear SNQR.

To illustrate the effect of quantile crossing and the benefit of joint estimation, we consider a simple illustrating one-dimensional toy example. Consider the underlying model $Y = X + \epsilon$, where $X \sim \text{Uniform}[-1, 1]$ and $\epsilon \sim N(0, 0.25)$ are independent of each other. Figure 1 displays the true and estimated quantile functions based on a simulated data set of size 40 using individual and joint estimations, respectively. The Gaussian kernel was used for the estimation. From the plots, we can clearly see that individual estimation has severe quantile crossing, while our SNQR

Downloaded by [Portland State University] at 22:38 01 December 2014

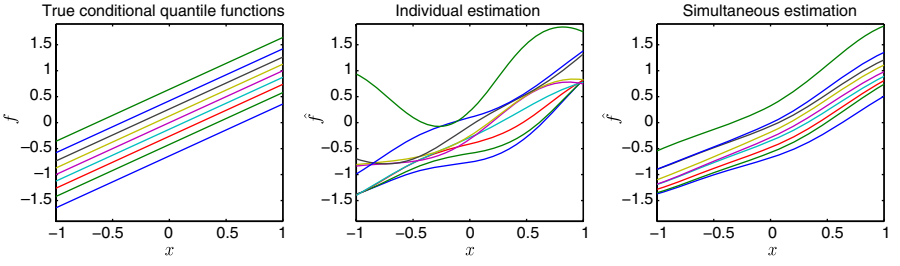


Figure 1. Illustration plot of quantile crossing of individual estimation and quantile non-crossing of the proposed SNQR estimation on the one-dimensional toy example. The left panel displays the true quantile functions for $\tau = 0.1, 0.2, \ldots, 0.9$. The middle and right panels display the estimation results of the original individual and proposed simultaneous estimation of the nine quantiles for one data realisation.

Downloaded by [Portland State University] at 22:38 01 December 2014

does not. More importantly, it appears that the individual estimation performs poorly for small or large τ values such as 0.1 and 0.9. In contrast, through the joint estimation, our proposed SNQR gives much improvement on the estimation of all quantile functions.

The rest of this article is organised as follows. In Section 2, we briefly review the standard QR and then introduce the proposed SNQR. In Section 3, we develop the asymptotic properties of a linear SNQR. We demonstrate the numerical performance of our proposed SNQR using simulations in Section 4 and the baseball data example in Section 5. Some final discussion is given in Section 6. Proofs of theoretical results are collected in the appendix.

2. Methodology

In this section, we first briefly review the standard QR and then introduce the proposed SNQR. In this paper, we use the kernel representation for quantile functions and embed non-crossing constraints on the kernel coefficients. Due to the use of kernel formulation, we first introduce the nonlinear version in Section 2.1, followed by the linear case in Section 2.2.

Suppose that we are given a sample $\{(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, n\}$ with covariates $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and the response $y_i \in \mathbb{R}$. The conditional τth quantile function $f_\tau(\boldsymbol{x})$ is defined such that

$$P(Y \leq f_\tau(\boldsymbol{x}) | \boldsymbol{X} = \boldsymbol{x}) = \tau \tag{1}$$

for $0 < \tau < 1$. By tilting the absolute loss function, Koenker and Bassett (1978) introduced the check function which is defined as $\rho_\tau(z) = z(\tau - I(z < 0))$ and illustrated in Figure 2. Here $I(\cdot)$ denotes the indicator function. Further they demonstrated in their seminal paper (Koenker and Bassett 1978) that the τth conditional quantile function can be estimated by solving

$$\min_{f_\tau \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - f_\tau(\boldsymbol{x}_i)). \tag{2}$$

Depending on how large the function space $\mathcal{F}$ is, a regularisation term may be necessary to avoid over-fitting and improve generalisation ability as considered in Koenker, Ng and Portnoy (1994) and Koenker (2004). Namely, we add a roughness penalty term $J(f_\tau)$ and solve

$$\min_{f_\tau \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - f_\tau(\boldsymbol{x}_i)) + \lambda J(f_\tau), \tag{3}$$

where λ is a tuning parameter balancing the data fitting measured by the check function and the complexity of $f_\tau(\cdot)$ measured by the roughness penalty $J(f_\tau)$. The kernel QR by Li, Liu and Zhu (2007) fits in the form of Equation (3).

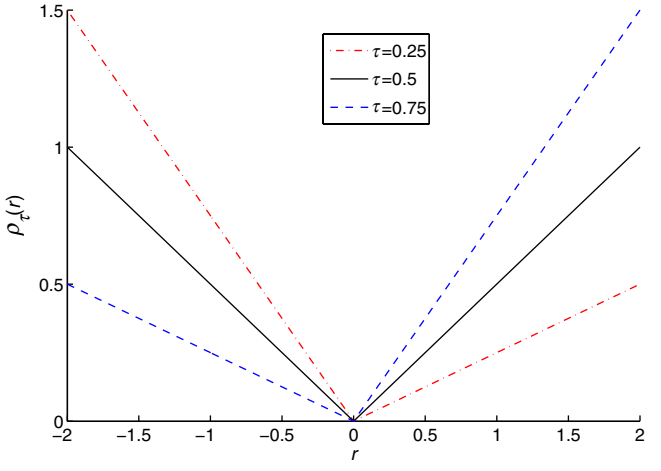


Figure 2. Plot of the check function for three different values of τ.

Although QR works well for estimating a quantile function for any particular τ, in certain scenarios, it is desirable to estimate multiple conditional quantile functions simultaneously. For example, one may be interested in estimating K quantile functions for $0 < \tau_1 < \tau_2 < \cdots < \tau_K < 1$. A naive way is to estimate $f_{\tau_k}(\cdot)$ individually by solving Equation (2) or (3) one at a time to get estimates $\hat{f}_{\tau_k}$, $k = 1, 2, \ldots, K$. Despite its simple implementation, there are some drawbacks with the naive approach. First of all, in theory, different quantiles should not cross each other. However, the naive estimates may suffer from quantile crossing for the finite sample case, especially when the sample size is small. Secondly, the naive estimation cannot share the strength of other quantile estimation due to the individual estimation scheme. Therefore, it is desirable to have a joint estimation technique which can ensure non-crossing of different quantiles and also improve the estimation accuracy of the quantile functions.

In this section, we propose a new general method that guarantees non-crossing of the estimated multiple quantile functions. Our method is based on the use of kernel representation of quantile functions. To introduce the proposed technique, we first discuss the nonparametric case using a Mercer kernel. Then we extend our method to the parametric linear case. For both cases, we assume that our input domain $\mathcal{X}$ is bounded. This bounded domain assumption is natural and necessary for our nonparametric technique. Even for the linear case, the bounded domain assumption is very reasonable due to the fact that two linear lines will eventually cross each other in $\mathbb{R}^d$ unless they are parallel.

2.1. *Nonlinear case*

For a Mercer kernel function $K(\cdot, \cdot)$, the representer theorem (Kimeldorf and Wahba 1971) allows us to represent the τth quantile function by $f_\tau(\boldsymbol{x}) = \sum_{i=1}^n w_{\tau,i} K(\boldsymbol{x}_i, \boldsymbol{x}) + b_\tau$. Our key observation is that for two quantile functions f_{τ_1} and f_{τ_2}, if the kernel function is non-negative with $K(\cdot, \cdot) \geq 0$, then we have $f_{\tau_1}(\boldsymbol{x}) \leq f_{\tau_2}(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$ if $w_{\tau_1,i} \leq w_{\tau_2,i}$; $i = 1, \ldots, n$ and $b_{\tau_1} \leq b_{\tau_2}$. One typical example of non-negative kernels is the Gaussian kernel with $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2/\sigma^2)$, where σ^2 is the kernel parameter. Using this observation, we can use simple constraints on the kernel coefficients to jointly estimate K kernel-based quantile functions without crossing.

Downloaded by [Portland State University] at 22:38 01 December 2014

Using the additional constraints, our SNQR technique estimates the QR coefficients by solving the following joint optimisation problem

$$\min \quad \sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}\left(y_i-\sum_{j=1}^{n}w_{\tau_k,j}K(\boldsymbol{x}_j,\boldsymbol{x}_i)-b_{\tau_k}\right)+\lambda\sum_{k=1}^{K}\boldsymbol{w}_{\tau_k}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{w}_{\tau_k}, \tag{4}$$

$$\text{subject to} \quad b_{\tau_k}\le b_{\tau_{k+1}} \quad \text{for } k=1,2,\ldots,K-1, \tag{5}$$

$$w_{\tau_k,i}\le w_{\tau_{k+1},i} \quad \text{for } i=1,2,\ldots,n;\ \ k=1,2,\ldots,K-1, \tag{6}$$

where $\boldsymbol{w}_{\tau_k}=(w_{\tau_k,1},w_{\tau_k,2},\ldots,w_{\tau_k,n})^{\mathrm{T}}$ and $\boldsymbol{K}$ is a matrix of size $n\times n$ with its (i,j) element being $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$. Here the regularisation term for the kth quantile function is $\boldsymbol{w}_{\tau_k}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{w}_{\tau_k}$ as a consequence of the representer theorem (Kimeldorf and Wahba 1971).

Note that the set of simple constraints (5) and (6) can guarantee the non-crossing of the estimated quantile functions as long as $K(\cdot,\cdot)\ge 0$. Here we want to note that the non-negativity assumption on the kernel $K(\cdot,\cdot)$ is not essential. According to Scholkopf and Smola (2002), $K(\cdot,\cdot)+C$ is a Mercer kernel as long as $K(\cdot,\cdot)$ is a Mercer kernel and $C\ge 0$. Thus, for any Mercer kernel $K(\cdot,\cdot)$, we define $K_+(\cdot,\cdot)=K(\cdot,\cdot)-K_{\mathcal{X}}$, where $K_{\mathcal{X}}=\min\{0,\inf_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{x}'\in\mathcal{X}}K(\boldsymbol{x},\boldsymbol{x}')\}$. Then the new kernel $K_+(\cdot,\cdot)$ satisfies the non-negativity assumption. Denote the solution to Equation (4) by $\hat{w}^+_{\tau_k,i}$ and $\hat{b}^+_{\tau_k}$ when the new kernel $K_+(\cdot,\cdot)$ is used. Our estimated conditional quantile functions are given by $\hat{f}_{\tau_k}(\boldsymbol{x})=\sum_{i=1}^{n}\hat{w}^+_{\tau_k,i}K_+(\boldsymbol{x}_i,\boldsymbol{x})+\hat{b}^+_{\tau_k}=\sum_{i=1}^{n}\hat{w}^+_{\tau_k,i}K(\boldsymbol{x}_i,\boldsymbol{x})-\left(\sum_{i=1}^{n}\hat{w}^+_{\tau_k,i}\right)K_{\mathcal{X}}+\hat{b}^+_{\tau_k}$ for $k=1,2,\ldots,K$. Note that our estimate $\hat{f}_{\tau_k}(\boldsymbol{x})$ can still be expressed in terms of the original kernel $K(\cdot,\cdot)$ that we begin with.

As a remark, we note that the objective function (4) aggregates the check function losses for different τ's and treats them equally. However, the value of $\sum_{i=1}^{n}\rho_{\tau_k}(y_i-\sum_{j=1}^{n}w_{\tau_k,j}K(\boldsymbol{x}_j,\boldsymbol{x}_i)-b_{\tau_k})$ may not be on the same scale for different τ's. Equal treatments of the loss function for different τ may be suboptimal. The following proposition gives the expected value of the check function when the error term is normally distributed.

PROPOSITION 1 *Let $\epsilon\sim N(0,1)$ and denote $\Phi^{-1}(\tau)$ as the τth quantile of ϵ, where $\Phi(\cdot)$ is the CDF of $N(0,1)$. Then, we have*

$$E[\rho_\tau(\epsilon-\Phi^{-1}(\tau))]=\phi(\Phi^{-1}(\tau)),$$

where $\phi(\cdot)$ is the density of $N(0,1)$.

Proposition 1 indicates that the expected value of the check function can vary greatly with different τ's. In the Gaussian case, the expected check function varies in the same way as the Gaussian density. In particular, the value for $\tau=0.5$ is the largest and it decreases as τ gets closer to 0 or 1. If we treat them equally, then those with τ around 0.5 will receive much larger emphasis than other quantiles. The quantiles with very small or large τ's tend to be ignored. To fix this problem, one can use weight adjustment for different quantiles. In particular, we can extend the objective function in Equation (4) to a weighted version as follows:

$$\sum_{k=1}^{K}W_k\sum_{i=1}^{n}\rho_{\tau_k}\left(y_i-\sum_{j=1}^{n}w_{\tau_k,j}K(\boldsymbol{x}_j,\boldsymbol{x}_i)-b_{\tau_k}\right)+\lambda\sum_{k=1}^{K}\boldsymbol{w}_{\tau_k}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{w}_{\tau_k}, \tag{7}$$

where W_k is the weight for the τ_kth quantile. In this paper, we consider two different weight vectors: equal weights and Gaussian-induced weights with $W_k=1/\phi(\Phi^{-1}(\tau))$. The Gaussian-induced weight can help to correct the scale difference of the check loss function for different quantiles

Downloaded by [Portland State University] at 22:38 01 December 2014

when the error is normal. Even when the error distribution is not normal, the Gaussian-induced weight provides a helpful adjustment for different quantile estimations. Furthermore, if some prior knowledge on the error distribution is available, then the corresponding weight can be adjusted accordingly.

2.2. *Linear case*

Different from nonlinear learning, we assume a parametric conditional quantile function $f_\tau = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau + \beta_{0\tau}$ in linear learning. However, the linear conditional quantile estimation can be achieved in the kernel representation framework using the linear kernel $L(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2$ and assuming $f_{\tau_k}(\boldsymbol{x}) = \sum_{i=1}^n w_{\tau_k,i} L(\boldsymbol{x}_i, \boldsymbol{x}) + b_{\tau_k}$. These two representations are equivalent in the sense that $\boldsymbol{\beta}_{\tau_k} = \sum_{i=1}^n w_{\tau_k,i}\boldsymbol{x}_i$ and $\beta_{0\tau_k} = b_{\tau_k}$.

Note that the linear kernel $L(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2$ does not satisfy the non-negativity assumption in general. As discussed above, we can define a new kernel $L_+(\cdot,\cdot) = L(\cdot,\cdot) - L_{\mathcal{X}}$, where $L_{\mathcal{X}} = \min\{0, \inf_{\boldsymbol{x}_1\in\mathcal{X},\boldsymbol{x}_2\in\mathcal{X}} L(\boldsymbol{x}_1, \boldsymbol{x}_2)\}$. Then $L_+(\cdot,\cdot)$ is a well-defined Mercer kernel and also satisfies the non-negativity assumption. With the new kernel $L_+(\cdot,\cdot)$, we can formulate our linear QR by defining $f_\tau(\boldsymbol{x}) = \sum_{i=1}^n w_{\tau,i} L_+(\boldsymbol{x}_i, \boldsymbol{x}) + b_\tau$ with slight abuse of notations. In this way, linear QR without crossing can be achieved by solving

$$\min \quad \sum_{k=1}^{K} W_k \sum_{i=1}^{n} \rho_{\tau_k}\left(y_i - \sum_{j=1}^{n} w_{\tau_k,j} L_+(\boldsymbol{x}_j, \boldsymbol{x}_i) - b_{\tau_k}\right), \tag{8}$$

$$\text{subject to} \quad b_{\tau_k} \le b_{\tau_{k+1}} \quad \text{for } k = 1, 2, \ldots, K-1, \tag{9}$$

$$w_{\tau_k,i} \le w_{\tau_{k+1},i} \quad \text{for } i = 1, 2, \ldots, n;\ \ k = 1, 2, \ldots, K-1. \tag{10}$$

In terms of the original linear kernel $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2$, the linear quantile function can be rewritten as $f_\tau(\boldsymbol{x}) = (\sum_{i=1}^n w_{\tau,i}\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{x} - (\sum_{i=1}^n w_{\tau,i}) L_{\mathcal{X}} + b_\tau$.

One interesting point is that our kernel representation of linear quantile functions is equivalent to the other parametric representation $f_\tau(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau + \beta_{0\tau}$ via the connection $\beta_\tau = (\sum_{i=1}^n w_{\tau,i}\boldsymbol{x}_i)$ and $\beta_{0\tau} = -(\sum_{i=1}^n w_{\tau,i}) L_{\mathcal{X}} + b_\tau$. This connection allows us to apply techniques for linear QR. For example, we can incorporate various penalties in linear QR that are capable of variable selection.

Another approach to estimate linear non-crossing quantile functions is to use the parametric representation $f_\tau(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau + \beta_{0\tau}$ directly. Suppose $\mathcal{X} = [0, \infty)^d$. Then linear non-crossing QR functions can be obtained by solving

$$\min \quad \sum_{k=1}^{K} W_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\tau_k} + \beta_{0\tau_k}), \tag{11}$$

$$\text{subject to} \quad \beta_{0\tau_k} \le \beta_{0\tau_{k+1}} \quad \text{for } k = 1, 2, \ldots, K-1, \tag{12}$$

$$\beta_{\tau_k,j} \le \beta_{\tau_{k+1},j} \quad \text{for } j = 1, 2, \ldots, d;\ \ k = 1, 2, \ldots, K-1. \tag{13}$$

The constraints here ensure quantile functions with larger τ's to be always above of those with smaller τ's to prevent crossing.

As a remark, we note that the kernel representation (8) requires $(K-1)(n+1)$ inequality constraints while the linear parametric representation (11) requires $(K-1)(d+1)$ inequality constraints. For low-dimensional problems with $d < n$, the formulation (11) can be easier to solve as it involves fewer constraints. In contrast, the formulation (8) is more preferable for high-dimensional low sample-size problems with $d > n$.

Downloaded by [Portland State University] at 22:38 01 December 2014

Downloaded by [Portland State University] at 22:38 01 December 2014

2.3. *Variable selection for linear quantile functions*

Variable selection plays an important role in the model-building process. In practice, it is very common to have a large number of candidate predictor variables available. These variables can be included in the initial stage of modelling for the consideration of removing potential modelling bias (Fan and Li 2001). However, it is undesirable to keep irrelevant predictors in the final model since this makes it difficult to interpret the resulting model and may decrease its predictive ability.

In the regularisation framework, many different types of penalties have been introduced to achieve variable selection. The L_1 penalty was used in the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996) for variable selection. Zou (2006) proposed the adaptive LASSO to improve the original LASSO. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) function and also studied its oracle properties in the penalised likelihood setting. For the QR, Koenker (2004) applied the LASSO penalty to the mixed-effect QR model for longitudinal data to encourage shrinkage in estimating the random effects. One important special case of QR with $\tau = 0.5$, the least absolute deviation regression, was studied by Wang, Li and Jiang (2007). Li and Zhu (2008) developed an algorithm to derive the entire solution path of linear L_1 QR. Wu and Liu (2008) studied both the adaptive L_1 and SCAD QR and developed the corresponding oracle properties. They also developed the difference convex algorithm (Liu, Shen and Doss 2005) for the SCAD penalised methods.

For variable selection in multiple quantile estimation, Zou and Yuan (2008b) proposed a hybrid of L_1 and L_∞ penalties to perform variable selection. The sup-norm is applied on the coefficients of the same variable for multiple quantile functions to encourage simultaneous sparsity. A similar sup-norm penalty was used by Zhang, Liu, Wu and Zhu (2008) for variable selection in multicategory support vector machines. To achieve simultaneous variable selection for multiple quantile functions, we also consider a sup-norm type of penalty to achieve simultaneous variable selection. One fundamental difference of our approach from the approach by Zou and Yuan (2008b) is that their approach cannot guarantee non-crossing of different quantile functions. Using the connection of $\boldsymbol{\beta}_\tau = \sum_{i=1}^n w_{\tau,i}\boldsymbol{x}_i$, we propose to solve the penalised version of Equations (8)–(10) as follows:

$$\min \quad \sum_{k=1}^{K} W_k \sum_{i=1}^{n} \rho_{\tau_k}\left(y_i - \sum_{j=1}^{n} w_{\tau_k,j} L_+(\boldsymbol{x}_j, \boldsymbol{x}_i) - b_{\tau_j}\right) + \sum_{j=1}^{d} p_\lambda\left(\max_{k=1}^{K}\left|\sum_{i=1}^{n} w_{\tau_k,i} x_{ij}\right|\right), \tag{14}$$

$$\text{subject to} \quad b_{\tau_k} \le b_{\tau_{k+1}} \quad \text{for } k = 1, 2, \ldots, K-1, \tag{15}$$

$$w_{\tau_k,i} \le w_{\tau_{k+1},i} \quad \text{for } i = 1, 2, \ldots, n; \;\; k = 1, 2, \ldots, K-1, \tag{16}$$

where $p_\lambda(\cdot)$ is a general penalty function with the regularisation parameter λ. Similar to the nonlinear case in Section 2.1, constraints (15) and (16) can guarantee that our estimated linear conditional quantile functions do not cross each other in the bounded input space $\mathcal{X}$.

Similar to the kernel version, we can also extend the parametric linear formulation in Equations (11)–(13) with penalties as follows:

$$\min \quad \sum_{k=1}^{K} W_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\tau_k} + \beta_{0\tau_k}) + \sum_{j=1}^{d} p_\lambda\left(\max_{k=1}^{K}|\beta_{\tau_k,j}|\right), \tag{17}$$

$$\text{subject to} \quad \beta_{0\tau_k} \le \beta_{0\tau_{k+1}} \quad \text{for } k = 1, 2, \ldots, K-1, \tag{18}$$

$$\beta_{\tau_k,j} \le \beta_{\tau_{k+1},j} \quad \text{for } j = 1, 2, \ldots, d; \;\; k = 1, 2, \ldots, K-1. \tag{19}$$

In this paper, we used the SCAD penalty (Fan and Li 2001); however, many other penalty functions can be adopted here. The SCAD penalty is mathematically defined in terms of its first-order derivative and is symmetric around the origin. For $\theta > 0$, its first-order derivative is given by

$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\}, \tag{20}$$

where $a > 2$ and $\lambda > 0$ are tuning parameters. Note that the SCAD penalty function is symmetric, non-convex on $[0, \infty)$ and singular at the origin.

2.4. *Computation*

Computation of the proposed SNQR can be done in a similar way as the original unregularised and regularised QR. For example, problems (4) and (8) can be implemented using quadratic programming and linear programming (LP), respectively. For problem (14), in order to handle the SCAD penalty, we make use of the local linear approximation algorithm proposed by Zou and Li (2008). In particular, at each step with the current solution $\tilde{w}_{\tau_k,i}$, we replace $p_\lambda(\max_{k=1}^K | \sum_{i=1}^n w_{\tau_k,i} x_{ij}|)$ by

$$p'_\lambda \left(\max_{k=1}^{K} \left| \sum_{i=1}^{n} \tilde{w}_{\tau_k,i} x_{ij} \right| \right) \left(\max_{k=1}^{K} \left| \sum_{i=1}^{n} w_{\tau_k,i} x_{ij} \right| \right). \tag{21}$$

To simplify Equation (21), we introduce a slack variable η_j to simplify the max function. In particular, we modify Equation (21) as

$$p'_\lambda \left(\max_{k=1}^{K} \left| \sum_{i=1}^{n} \tilde{w}_{\tau_k,i} x_{ij} \right| \right) \eta_j, \tag{22}$$

subject to

$$\eta_j \ge \sum_{i=1}^{n} w_{\tau_k,i} x_{ij}, \quad \eta_j \ge -\sum_{i=1}^{n} w_{\tau_k,i} x_{ij}; \quad \text{for } k = 1, \ldots, K.$$

Then using approximation (22), problem (14) can be solved using the iterative LP. Similarly, the parametric penalised version (17)–(19) can also be computed using the iterative LP.

3. Theoretical properties

In this section, we consider the theoretical properties of our non-crossing linear conditional quantile estimates presented above. To that end, we first consider the standard unpenalised and penalised linear QR without non-crossing constraints. Then we investigate the behaviour of the constraints as n grows to infinity to explore the theoretical properties of the new proposed technique.

3.1. *Asymptotic normality of unconstrained and constrained linear QR*

We first consider the unpenalised version by establishing asymptotic properties of the solution to Equation (8). Without the non-crossing constraints, it is equivalent to the naive individual estimate

Downloaded by [Portland State University] at 22:38 01 December 2014

by solving

$$\min_{\boldsymbol{\beta}_{\tau_k},\ \beta_{0\tau_k}} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \beta_{0\tau_k} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\tau_k}) \tag{23}$$

one at a time for each $k = 1, 2, \ldots, K$. Denote the optimal solution of Equation (23) by $\tilde{\beta}_{0\tau_k}$ and $\tilde{\boldsymbol{\beta}}_{\tau_k}$.

Define $\mathcal{C}_n$ to be an event that individually estimated conditional quantile functions, obtained by solving Equation (23) with a random sample of size n, cross each other, namely, there exist k and $\boldsymbol{x} \in \mathcal{X}$ such that $\tilde{\beta}_{0\tau_k} + \tilde{\boldsymbol{\beta}}_{\tau_k}^{\mathrm{T}}\boldsymbol{x} > \tilde{\beta}_{0\tau_{k+1}} + \tilde{\boldsymbol{\beta}}_{\tau_{k+1}}^{\mathrm{T}}\boldsymbol{x}$. We prove that $P(\mathcal{C}_n) \to 0$ as $n \to \infty$ by showing $P(\mathcal{C}_n)$ decays exponentially in n.

As in Koenker (2005, p. 120), we consider a general form of the linear quantile model. Let Y_1, $Y_2, \ldots$ be independent random variables with distribution functions F_1, $F_2, \ldots$ and suppose that the τth conditional quantile function is linear in the covariate vector $\boldsymbol{x}$ by assuming

$$Q_{Y_i}(\tau|\boldsymbol{x}) = \beta_0(\tau) + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}(\tau).$$

The conditional distribution functions of the Y_i's will be written as $P(Y_i < y|\boldsymbol{x}_i) = F_{Y_i}(y|\boldsymbol{x}_i) = F_i(y)$, and then

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = F_{Y_i}^{-1}(\tau|\boldsymbol{x}_i) \equiv \xi_i(\tau).$$

To proceed, we assume that the following two conditions are satisfied.

Condition A: The distribution functions $\{F_i\}$ are absolutely continuous, with continuous densities $f_i(\cdot)$ uniformly bounded away from 0 and ∞ at points $\{\xi_i(\tau_1), \xi_i(\tau_2), \ldots, \xi_i(\tau_K)\}$, $i = 1, 2, \ldots$.

Condition B: There exist positive-definite matrices Σ_0 and $\Sigma_1(\tau_k)$ for $k = 1, 2, \ldots, K$ such that

(1)

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^{\mathrm{T}} = \Sigma_0.$$

(2)

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} f_i(\xi_i(\tau_k)) \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^{\mathrm{T}} = \Sigma_1(\tau_k).$$

(3)

$$\max_{i=1,2,\ldots,n} \frac{\| \boldsymbol{x}_i \|}{\sqrt{n}} \to 0,$$

where $\tilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i^{\mathrm{T}})^{\mathrm{T}}$.

Recall that the naive individual estimate is denoted by $\tilde{\beta}_{0\tau_k}$ and $\tilde{\boldsymbol{\beta}}_{\tau_k}$. Denote the corresponding true parameters by $\beta_0(\tau_k)$ and $\boldsymbol{\beta}(\tau_k)$. Our non-crossing estimates are denoted by $\hat{\beta}_{0\tau_k}$ and $\hat{\boldsymbol{\beta}}_{\tau_k}$.

LEMMA 1 *Under conditions A and B, as $n \to \infty$, the naive individual estimates have the following asymptotic normality*

$$\sqrt{n}\left[\begin{pmatrix} \tilde{\beta}_{0\tau_k} \\ \tilde{\boldsymbol{\beta}}_{\tau_k} \end{pmatrix} - \begin{pmatrix} \beta_0(\tau_k) \\ \boldsymbol{\beta}(\tau_k) \end{pmatrix}\right] \longrightarrow N(\mathbf{0}, \tau_k(1-\tau_k)\Sigma_1(\tau_k)^{-1}\Sigma_0\Sigma_1(\tau_k)^{-1}). \tag{24}$$

Downloaded by [Portland State University] at 22:38 01 December 2014

PROPOSITION 2 *When the domain $\mathcal{X}$ is bounded, under the conditions of Lemma 1, there exists a constant $c > 0$ such that $P(\mathcal{C}_n) \leq e^{-nc}$ asymptotically.*

Proposition 2 shows that the quantile crossing phenomenon is only a finite sample behaviour. As the sample size n increases, the probability of quantile crossing decreases exponentially in n. Thus, we can expect that the non-crossing quantile technique shares the same asymptotic behaviour as the corresponding QR methods without non-crossing constraints if the constraints are necessary for non-crossing under certain cases.

As discussed earlier, constraints (9) and (10) or (12) and (13) are sufficient to ensure the non-crossing of the resulting estimated multiple quantile functions. The following proposition states the necessity of the constraints for non-crossing.

PROPOSITION 3 *Suppose $f_{\tau_k}(\boldsymbol{x}) = \beta_{0\tau_k} + \boldsymbol{\beta}_{\tau_k}^{\mathrm{T}}\boldsymbol{x}$ with $(\beta_{0\tau_k}, \boldsymbol{\beta}_{\tau_k})$; $k = 1, \ldots, K$, bounded and $\boldsymbol{x} \in \mathcal{X} = [0, \infty)^d$. Then* (i) *constraints* (12) *and* (13) *are necessary and sufficient to ensure the non-crossing of f_{τ_k} in $\mathcal{X}$*; (ii) *if $d > n$, constraints* (9) *and* (10) *are also necessary and sufficient to ensure the non-crossing of f_{τ_k}'s in $\mathcal{X}$.*

Our next theorem states the same asymptotic normality of the non-crossing estimators as the unconstrained estimators. Since the probability of the crossing event goes to 0 asymptotically as shown in Proposition 2, the non-crossing estimators asymptotically behave the same as the unconstrained estimators if the constraints are sufficient and necessary for non-crossing.

THEOREM 1 *Assume that the non-crossing constraints are necessary and sufficient. Under the conditions of Proposition 2, then with the probability tending to* 1, *the simultaneous non-crossing estimates obtained by solving Equation* (8) *have the asymptotic normality*

$$\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_{0\tau_k}\\ \hat{\boldsymbol{\beta}}_{\tau_k}\end{pmatrix} - \begin{pmatrix}\beta_{0\tau_k}\\ \boldsymbol{\beta}_{\tau_k}\end{pmatrix}\right] \longrightarrow N(\mathbf{0}, \tau_k(1-\tau_k)\Sigma_1(\tau_k)^{-1}\Sigma_0\Sigma_1(\tau_k)^{-1}). \tag{25}$$

3.2. *Oracle properties of sparse constrained linear SNQR*

In this section, we develop the oracle properties of our sparse penalised linear SNQR in the notion of Fan and Li (2001). With a non-concave penalty $p_\lambda(\cdot)$, similar to the development in Section 3.1, we first consider the version without non-crossing constraints by solving

$$\min \sum_{k=1}^{K} W_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\tau_k} - \beta_{0\tau_k}) + n\sum_{j=1}^{p} p_\lambda\left(\max_{k=1}^{K}|\beta_{j\tau_k}|\right). \tag{26}$$

Without loss of generality, in this section, we set $W_k = 1; k = 1, \ldots, K$. The results can be directly generalised to other weights W_k's. The corresponding optimiser is denoted by $\tilde{\boldsymbol{\beta}}_{\tau_k}^{S}$ and $\tilde{\beta}_{0\tau_k}^{S}$.

Recall that the true parameters are denoted by $\boldsymbol{\beta}(\tau_k) = (\beta_1(\tau_k), \beta_2(\tau_k), \ldots, \beta_p(\tau_k))^{\mathrm{T}}, \beta_0(\tau_k)$ for $k = 1, 2, \ldots, K$. Denote $u_{ik} = y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau_k) - \beta_0(\tau_k)$ for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, K$. The behaviour of $\sqrt{n}\{\tilde{\boldsymbol{\beta}}_{\tau_k}^{S} - \boldsymbol{\beta}(\tau_k)\}$ and $\sqrt{n}\{\tilde{\beta}_{0\tau_k}^{S} - \beta_0(\tau_k)\}$ follows from the consideration of the following objective function

$$Q(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_K, a_1, a_2, \ldots, a_K) = Z_{nk}(\boldsymbol{\alpha}_k, a_k) + n\sum_{j=1}^{p} p_\lambda\left(\max_{k=1}^{K}\left|\beta_j(\tau_k) + \frac{\alpha_{jk}}{\sqrt{n}}\right|\right), \tag{27}$$

where $\boldsymbol{\alpha}_k = (\alpha_{1k}, \alpha_{2k}, \ldots, \alpha_{pk})^{\mathrm{T}}$ and $Z_{nk}(\boldsymbol{\alpha}_k, a_k) = \sum_{i=1}^{n}[\rho_{\tau_k}(u_{ik} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\alpha}_k/\sqrt{n} - a_k/\sqrt{n}) - \rho_{\tau_k}(u_{ik})]$. This minimiser of Equation (27) is given by $\sqrt{n}(\tilde{\boldsymbol{\beta}}_{\tau_k} - \boldsymbol{\beta}(\tau_k))$ and $\sqrt{n}(\tilde{\beta}_{0\tau_k} - \beta_0(\tau_k))$.

Downloaded by [Portland State University] at 22:38 01 December 2014

Downloaded by [Portland State University] at 22:38 01 December 2014

By reordering if necessary, we assume without loss of generality that the first s predictors are important, which means that $\max_{1\le k\le K}|\beta_j(\tau_k)|>0$ for any $j\le s$. It also implies that $\beta_j(\tau_k)=0$ for $j>s$ and $k=1,2,\ldots,K$. Denote $\mathcal{A}=\{j:\max_{k=1,2,\ldots,K}|\beta_j(\tau_k)|\neq 0\}=\{1,2,\ldots,s\}$ to be the true active set. Note that the set $\mathcal{A}$ includes all variables that have at least one non-zero coefficient among all quantile functions considered. We do not require all quantile functions to have the same non-zero coefficients. Denote $c_n=\max_{j\in\mathcal{A}}p'_\lambda(\max_{k=1,2,\ldots,K}|\beta_j(\tau_k)|)$. Set $\mathcal{A}^c=\{s+1,s+2,\ldots,p\}$, $\mathcal{B}=\{1,2,\ldots,s+1\}$, and $\mathcal{B}^c=\{s+2,s+3,\ldots,p+1\}$. Use $\boldsymbol{\beta}_{\mathcal{A}k}$ to denote the subvector of $\boldsymbol{\beta}_k$ with indices in $\mathcal{A}$ and $\Sigma_{1,\mathcal{B},\mathcal{B}^c}$ to denote the submatrix of Σ_1 with a row index in $\mathcal{B}$ and a column index in $\mathcal{B}^c$.

LEMMA 2 *If $\lambda=\lambda_n\to 0$, $c_n=O(n^{-1/2})$ and $\max_{j\in\mathcal{A}}|p''_\lambda(\max_{k=1,2,\ldots,K}|\beta_j(\tau_k)|)|\to 0$, under conditions A and B, then there exits a local minimiser $\hat{\boldsymbol{\alpha}}_k$ and $\hat{a}_k$, $k=1,2,\ldots,K$, for Equation* (27) *such that $\|\hat{\boldsymbol{\alpha}}_k\|=O_p(n^{-1/2})$ and $\hat{a}_k=O_p(n^{-1/2})$.*

LEMMA 3 *If $\liminf_{\lambda\to 0+}\liminf_{\theta\to 0+}p'_\lambda(\theta)/\lambda>0$, under the conditions of Lemma 2, then with the probability tending to 1, for any a_k and $\boldsymbol{\alpha}_{\mathcal{A}k}$ satisfying $\sqrt{\sum_{k=1}^K(\|\boldsymbol{\alpha}_{\mathcal{A}k}\|^2+a_k^2)}=O_p(n^{-1/2})$ and for any constant $C>0$,*

$$Q\left(\begin{pmatrix}\boldsymbol{\alpha}_{\mathcal{A}1}\\ \mathbf{0}\end{pmatrix},\begin{pmatrix}\boldsymbol{\alpha}_{\mathcal{A}2}\\ \mathbf{0}\end{pmatrix},\ldots,\begin{pmatrix}\boldsymbol{\alpha}_{\mathcal{A}K}\\ \mathbf{0}\end{pmatrix},a_1,a_2,\ldots,a_K\right)$$
$$=\min_{\sqrt{\sum_{k=1}^K\|\boldsymbol{\alpha}_{\mathcal{A}^c k}\|}\le Cn^{-1/2}}Q\left(\begin{pmatrix}\boldsymbol{\alpha}_{\mathcal{A}1}\\ \boldsymbol{\alpha}_{\mathcal{A}^c 1}\end{pmatrix},\begin{pmatrix}\boldsymbol{\alpha}_{\mathcal{A}2}\\ \boldsymbol{\alpha}_{\mathcal{A}^c 2}\end{pmatrix},\ldots,\begin{pmatrix}\boldsymbol{\alpha}_{\mathcal{A}K}\\ \boldsymbol{\alpha}_{\mathcal{A}^c K}\end{pmatrix},a_1,a_2,\ldots,a_K\right).$$

THEOREM 2 *If $\lambda\to 0$ and $\sqrt{n}\lambda\to\infty$, under the conditions of Lemma 3, then with the probability tending to* 1, *the $\sqrt{n}$ consistent local minimiser $\tilde{\boldsymbol{\beta}}_{\tau_k}$ and $\tilde{\beta}_{0\tau_k}$, $k=1,2,\ldots,K$, of Lemma 2 satisfies that*

(1) *$\tilde{\beta}_{j\tau_k}=0$ for $j\notin\mathcal{A}$,*
(2) *the optimiser $\tilde{\beta}_{j\tau_k}$ for $j\in\mathcal{A}$ and $\tilde{\beta}_{0\tau_k}$ has the same asymptotic property of the minimiser of the following objective function*

$$\min\sum_{k=1}^K\sum_{i=1}^n\rho_{\tau_k}\left(y_i-\sum_{j\in\mathcal{A}}x_{ij}\beta_{j\tau_k}-\beta_{0\tau_k}\right)+\sum_{j\in\mathcal{A}}p_\lambda\left(\max_{k=1}^K|\beta_{j\tau_k}|\right). \tag{28}$$

As a remark, we note that the absolute value of the true parameters may have a tie for some $j\le s$, namely, $|\beta_j(\tau_k)|=|\beta_j(\tau_{k'})|$ for some $1\le k,k'\le K$ and $1\le j\le s$. Thus, it is not easy to derive the asymptotic properties of the minimiser of Equation (28) for a general non-concave penalty. When the SCAD penalty by Fan and Li (2001) is used, we have the following result as stated in Proposition 4.

PROPOSITION 4 *When the SCAD penalty is used, under the conditions of Theorem 2, as $n\to\infty$, with the probability tending to* 1, *we have that*

(1) *$\tilde{\beta}_{j\tau_k}=0$ for $j\notin\mathcal{A}$.*
(2) *the optimiser $\tilde{\beta}_{j\tau_k}$ for $j\in\mathcal{A}$ and $\tilde{\beta}_{0\tau_k}$ satisfies*

$$\sqrt{n}\left[\begin{pmatrix}\tilde{\beta}_{0\tau_k}\\ \tilde{\boldsymbol{\beta}}_{\mathcal{A}\tau_k}\end{pmatrix}-\begin{pmatrix}\beta_0(\tau_k)\\ \boldsymbol{\beta}_{\mathcal{A}}(\tau_k)\end{pmatrix}\right]\longrightarrow N(\mathbf{0},\tau_k(1-\tau_k)\Sigma_{1,\mathcal{B},\mathcal{B}}(\tau_k)^{-1}\Sigma_{0,\mathcal{B},\mathcal{B}}\Sigma_{1,\mathcal{B},\mathcal{B}}(\tau_k)^{-1}). \tag{29}$$

The following theorem states the same oracle property of the constrained sparse SNQR with the unconstrained sparse SNQR.

THEOREM 3 *Assume that the non-crossing constraints are necessary and sufficient for non-crossing. With the probability tending to* 1*, asymptotic results in Theorem 2 and Proposition 4 apply to the proposed non-concave penalised non-crossing quantile estimation* (14) *under the same conditions.*

As a remark, we note that model selection techniques that enjoy the oracle property may have unsatisfactory asymptotic behaviours in the 'uniform sense' with respect to the unknown parameter as one referee pointed out. The pointwise asymptotic distribution of the estimator may not be representative for the finite sample performance of the estimator (see, e.g. Leeb and Potscher 2008; Potscher and Leeb 2009; Potscher and Schneider 2010). We will not further explore this aspect on the proposed SNQR in this paper.

Downloaded by [Portland State University] at 22:38 01 December 2014

4. Simulations

In our simulated examples, our training sample size is denoted by n. An independent tuning set of size n and an independent test set of size $10n$ are generated in the same way to tune the regularisation parameter and calculate test errors, respectively. The tuning parameter λ is selected via a grid search by minimising $\sum_{k=1}^{K} W_k \sum_{i=1}^{n} \rho_{\tau_k}(\check{y}_i - \hat{f}_{\tau_k}(\check{\boldsymbol{x}}_i))$, where $(\check{\boldsymbol{x}}_i, \check{y}_i)$ denotes an pair of observations in our tuning set, $\hat{f}(\cdot)$ denotes an estimate of the conditional quantile function and W_k is the weight for τ_k. We evaluate the test error, $\mathrm{TE}(\hat{f}) = \sum_{k=1}^{K} W_k \sum_{i=1}^{10n} \rho_{\tau_k}(\tilde{y}_i - \hat{f}_{\tau_k}(\tilde{\boldsymbol{x}}_i))$, to compare the performance of our new method with competitive estimators, where $(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)$ denotes a pair of observations in our test set.

To examine the performance of the proposed SNQR, we compare it with the individual QR. For individual penalised QR as in Examples 4.1 and 4.2, we carry out two different tuning procedures. One is to separately tune λ for different QR functions. The other one is to jointly tune λ as in SNQR so that all different quantile terms use the same λ. Besides the unconstrained QR, we also compare SNQR with the QR with constraints on the training data only as suggested by Takeuchi et al. (2006). While comparing two different methods, we report the pairwise t-statistic between test errors over 100 repetitions for each example, namely $t_{\mathrm{M2,M1}} = \sqrt{100}\mathrm{mean}(\mathrm{TE}_i(\hat{f}_{\mathrm{M2}}) - \mathrm{TE}_i(\hat{f}_{\mathrm{M1}}), i = 1, 2, \ldots, 100)/\mathrm{std}(\mathrm{TE}_i(\hat{f}_{\mathrm{M2}}) - \mathrm{TE}_i(\hat{f}_{\mathrm{M1}}), i = 1, 2, \ldots, 100)$. For the nonlinear quantile estimation, we use the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2)$.

Example 1 (*Nonlinear example with i.i.d. noise*) In this example, our predictor is univariate and uniformly distributed over $[-1, 1]$, namely, $X \sim \mathrm{Uniform}[-1, 1]$. Conditional on X, $Y = 2\sin(\pi X) + 0.5\epsilon$, where $\epsilon \sim N(0, 1)$ denotes the independent noise. We set $\tau_k = 0.1k$ for $k = 1, 2, \ldots, 9$. We compare different estimators with the Gaussian kernel. The training sample size is set to be $n = 100$. Results over 100 repetitions are given in Table 1, which reports the pairwise t-test statistics for comparing test errors of different methods. The weight and error options indicate the types of weight used in model-building and calculation of the testing error, respectively. The results show that the proposed SNQR (M3) gives the best performance among all methods considered here. When we use the joint tuning procedure for λ, the simultaneous method with data point restriction (M2) works better than the individually estimated QR (M1). Interestingly, when we perform separate tuning for individual multiple quantile estimation (M1′), the results are better than the simultaneous method with data point restriction (M2). Furthermore, the types of weights and errors do not appear to have much influence on the methods in this example.

Table 1. Pairwise t-test for the test error of nonlinear i.i.d Example 1.

Weight	Error	M2 versus M1	M3 versus M2	M3 versus M1	M1 versus M1′	M2 versus M1′	M3 versus M1′
Uniform	Uniform	−2.9895	−11.3217	−10.4820	4.6798	3.1192	−9.7830
	Normal	−3.2488	−12.5514	−11.6472	6.1261	4.5797	−9.3860
Normal	Uniform	−2.8038	−12.3808	−11.2452	4.5684	3.4622	−10.4665
	Normal	−3.1294	−13.2998	−11.9478	5.7226	4.8629	−10.4114

Notes: M1, individual estimation with joint tuning; M1′, individual estimation with separate tuning; M2, simultaneous estimation with data point restriction; M3, SNQR.

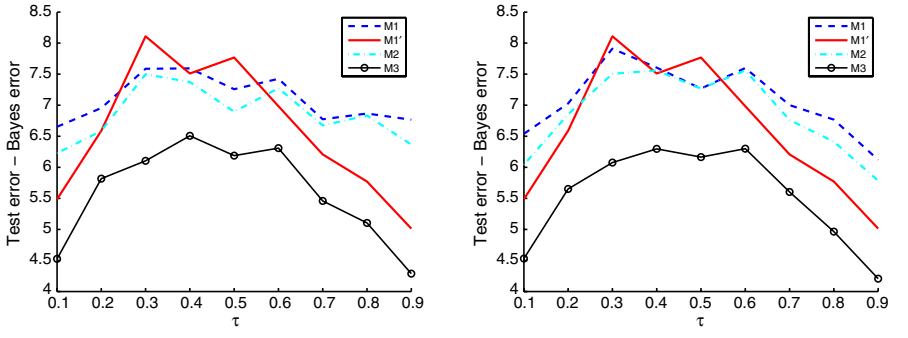


Figure 3. Plot of the average differences between the test errors and Bayes errors for Example 1. The left and right panels correspond to the uniform and normal weights respectively.

In Figure 3, we plot the average individual difference of the test error $\sum_{i=1}^{10n} \rho_{\tau_k}(\tilde{y}_i - \hat{f}_{\tau_k}(\tilde{\boldsymbol{x}}_i))$ and the Bayes error $\sum_{i=1}^{10n} \rho_{\tau_k}(\tilde{y}_i - f_{\tau_k}(\tilde{\boldsymbol{x}}_i))$ with respect to individual τ_k's for different methods, where $f_{\tau_k}(\cdot)$ denotes the true conditional quantile function. It clearly shows the improvement of our method.

Example 2 (*Nonlinear example with non-i.i.d. noise*) In this example, the predictor is the same as in the previous example, namely, $X \sim \text{Uniform}[-1, 1]$. Conditional on X, $Y = 2\sin(\pi X) + 0.5(1 + X^2)\epsilon$, where $\epsilon \sim N(0, 1)$ denotes the independent noise. The sample size is chosen to be $n = 100$. Results over 100 repetitions are reported in Table 2. The results are similar to that of Example 1, although the differences among methods are smaller in this example. In particular, the proposed SNQR (M3) works the best, followed by individual estimation with separate tuning (M1′), simultaneous estimation with data point restriction (M2) and individual estimation with joint tuning (M1).

Table 2. Pairwise t-test for the test error of Example 2.

Weight	Error	M2 versus M1	M3 versus M2	M3 versus M1	M1 versus M1′	M2 versus M1′	M3 versus M1′
Uniform	Uniform	−1.6692	−5.4152	−5.4005	2.7011	1.6137	−3.6387
	Normal	−1.4907	−5.0432	−4.9881	3.9332	2.9301	−2.4405
Normal	Uniform	−2.6016	−5.4650	−6.6603	2.8115	0.2473	−5.0223
	Normal	−3.0040	−4.6702	−6.0799	3.7212	0.8714	−3.8507

Notes: M1, individual estimation with joint tuning; M1′, individual estimation with separate tuning; M2, simultaneous estimation with data point restriction; M3, SNQR.

Downloaded by [Portland State University] at 22:38 01 December 2014

Example 3 (*Linear example with i.i.d. noise*) Data are generated from

$$Y = X_1 + X_2 + 0.5\epsilon,$$

with $X_1 \sim \text{Uniform}[0, 1]$, $X_2 \sim \text{Uniform}[0, 1]$, $\epsilon \sim N(0, 1)$ being independent of each other. We set $n = 100$, $d = 2$ and $\tau_k = k/10$ for $k = 1, 2, \ldots, 9$. For this example, we compare unpenalised QR methods, i.e. individual estimation (M1), simultaneous estimation with data point restriction (M2) and SNQR (M3). Results over 100 repetitions are reported in Table 3. The results show that SNQR (M3) works the best, then followed by the data restriction method (M2). The individual estimation (M1) gives the worst estimation accuracy.

Example 4 (*Linear example with non-i.i.d. noise*) Consider the following location-scale model

$$Y = 1 + X_1 + X_2 + \left(1 + \frac{1 + X_3}{2}\right)\epsilon,$$

where $X_j \sim \text{Uniform}[-1, 1]$, $j = 1, 2, 3$, and $\epsilon \sim N(0, 1)$ are independent of each other. Similar to Example 3, we compare three unpenalised QR methods: individual estimation (M1), simultaneous estimation with data point restriction (M2) and SNQR (M3). We set $n = 100$, $d = 3$ and $\tau_k = k/10$ for $k = 1, 2, \ldots, 9$. Results over 100 repetitions are reported in Table 4. The results once again demonstrate that SNQR (M3) works the best, followed by the data restriction method (M2) and then the individual estimation (M1).

Example 5 (*SCAD linear example with i.i.d. noise*) In this example, we simulate predictors $\boldsymbol{X} \sim N(\boldsymbol{0}, \Sigma)$ with $\Sigma = (\sigma_{ij})$, where $\sigma_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le p$. Data are generated from the model

$$Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon,$$

where $\epsilon \sim N(0, 1)$ is the independent error. Here we consider two settings, $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^{\mathrm{T}}$ as in Tibshirani (1996) and $\boldsymbol{\beta} = (1.5, 0.75, 0, 0, 1, 0, 0, 0)^{\mathrm{T}}$ which has a

Table 3. Pairwise t-test for test errors of linear i.i.d of Example 3.

Weight	Error	M2 versus M1	M3 versus M2	M3 versus M1
Uniform	Uniform	−4.3214	−6.6111	−7.2070
	Normal	−4.3348	−5.3270	−6.0527
Normal	Uniform	−3.9829	−7.4806	−8.0344
	Normal	−3.9335	−6.3636	−7.0218

Notes: M1, individual estimation; M2, simultaneous estimation with data point restriction; M3, SNQR.

Table 4. Pairwise t-test for test errors of linear non-i.i.d Example 4.

Weight	Error	M2 versus M1	M3 versus M2	M3 versus M1
Uniform	Uniform	−6.8340	−10.9368	−12.2511
	Normal	−6.6010	−10.4753	−11.8933
Normal	Uniform	−6.4172	−10.7988	−11.7562
	Normal	−6.2999	−10.7351	−11.9027

Notes: M1, individual estimation; M2, simultaneous estimation with data point restriction; M3, SNQR.

Downloaded by [Portland State University] at 22:38 01 December 2014

lower signal level. Among the eight covariates, three are important variables and the remaining five are noise variables. We use this example to examine the performance of sparse penalised QR.

For comparison, we consider five different methods: the individual QR estimation with joint tuning on λ (M1), the individual QR estimation with separate tuning on λ (M1′), the simultaneous SCAD-max QR estimation without non-crossing constraints (M1″), the simultaneous SCAD-max QR estimation with non-crossing restrictions on training data (M2) and the simultaneous SCAD-max SNQR (M3). Tables 5 and 6 report the pairwise t-statistics for the comparison of these five methods. For example, the first entry 1.0840 in Table 5 is the pairwise t-statistic $t_{\mathrm{M1}',\mathrm{M1}}$ which shows that M1 gives a smaller test error than that of M1′. Overall, we can conclude that the simultaneous SCAD-max SNQR (M3) works the best in terms of test errors. Between the uniform and normal weights, the results are similar although the improvement of SNQR over other methods appears to be larger when we use the normal weight than that of the uniform weight.

Similar to Example 1, in Figures 4 and 5, we plot the individual average differences of test errors and the Bayes errors with respect to individual τ_k for five different methods. Once again, the plot clearly demonstrates the competitiveness of the proposed SNQR for both settings of $\boldsymbol{\beta}$.

Table 5. Pairwise t-tests for test errors of Example 5 with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^{\mathrm{T}}$.

	Uniform error				Normal error			
Weight	M1	M1′	M1″	M2	M1	M1′	M1″	M2
Uniform								
M1′	1.0840				1.9523			
M1″	−3.6614	−4.6880			−4.0600	−4.9097		
M2	−6.9470	−7.8040	−5.6231		−7.7608	−8.2131	−5.7375	
M3	−12.1945	−13.1992	−11.9942	−9.5200	−12.9902	−13.9023	−12.7855	−10.3205
Normal								
M1′	0.4965				1.4542			
M1″	−3.3025	−3.8849			−3.3873	−4.4074		
M2	−6.5717	−6.6360	−4.4325		−7.4641	−7.3642	−4.7444	
M3	−12.5578	−13.5503	−13.0412	−11.0871	−13.7023	−14.6033	−14.5311	−11.9845

Notes: M1, individual estimation with joint tuning; M1′, individual estimation with separate tuning; M1″, simultaneous SCAD-max estimation without constraints; M2, simultaneous SCAD-max estimation with data point restriction; M3, simultaneous SCAD-max SNQR.

Table 6. Pairwise t-tests for test errors of Example 5 with $\boldsymbol{\beta} = (1.5, 0.75, 0, 0, 1, 0, 0, 0)^{\mathrm{T}}$.

	Uniform error				Normal error			
Weight	M1	M1′	M1″	M2	M1	M1′	M1″	M2
Uniform								
M1′	−0.8772				−1.3600			
M1″	−4.1814	−4.5246			−4.3181	−4.6945		
M2	−5.0955	−5.8740	−1.8677		−5.0897	−6.0284	−2.1718	
M3	−12.3679	−15.9101	−10.8048	−8.7489	−11.7193	−16.1446	−12.1681	−9.6643
Normal								
M1′	−0.4428				−0.4475			
M1″	−4.6602	−5.4521			−4.7435	−5.8945		
M2	−5.7025	−7.2761	−3.7704		−5.5784	−7.4678	−3.6670	
M3	−10.9239	−13.0084	−10.0445	−8.1136	−10.8430	−13.5719	−11.3901	−9.3674

Notes: M1, individual estimation with joint tuning; M1′, individual estimation with separate tuning; M1″, simultaneous SCAD-max estimation without constraints; M2, simultaneous SCAD-max estimation with data point restriction; M3, simultaneous SCAD-max SNQR.

Downloaded by [Portland State University] at 22:38 01 December 2014

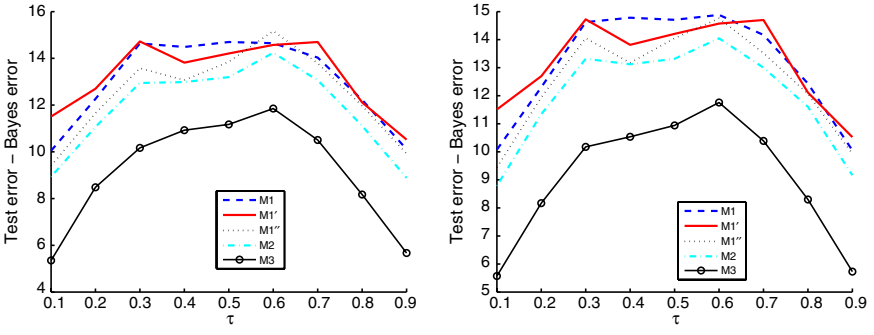


Figure 4. Plot of the average differences between the test errors and Bayes errors for Example 5 with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^{\mathrm{T}}$. The left and right panels correspond to the uniform and normal weights, respectively.

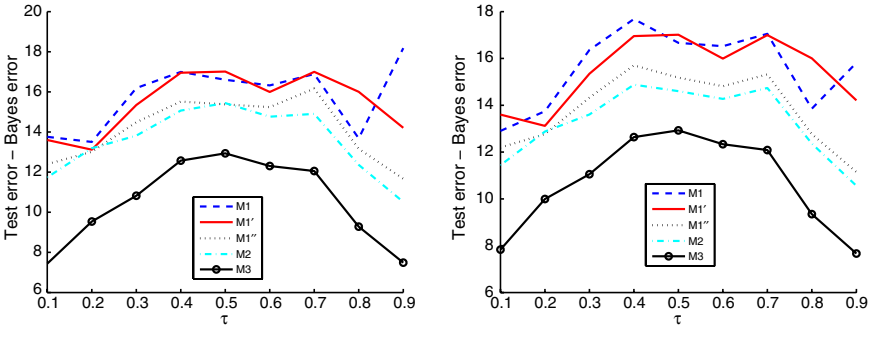


Figure 5. Plot of the average differences between the test errors and Bayes errors for Example 5 with $\boldsymbol{\beta} = (1.5, 0.75, 0, 0, 1, 0, 0, 0)^{\mathrm{T}}$. The left and right panels correspond to the uniform and normal weights, respectively (lower signal- to-noise ratio).

Table 7. Variable selection results of Example 5 with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^{\mathrm{T}}$.

Weight	M1	M1′	M1″	M2	M3
Uniform					
Average wrong 0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Average correct 0	3.41 (0.14)	1.00 (0.12)	4.49 (0.08)	4.40 (0.09)	3.91 (0.15)
Normal					
Average wrong 0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Average correct 0	3.45 (0.14)	1.00 (0.12)	4.42 (0.09)	4.31 (0.10)	3.69 (0.15)

Notes: M1, individual estimation with joint tuning; M1′, individual estimation with separate tuning; M1″, simultaneous SCAD-max estimation without constraints; M2, simultaneous SCAD-max estimation with data point restriction; M3, simultaneous SCAD-max SNQR.

Tables 7 and 8 show the results on variable selection of Example 5. We report the average correct and wrong zero coefficients across all quantiles. Since there are three important variables and five noise variables, the true model has five zero coefficients and three non-zero coefficients for each QR function. As expected, the performance for the weaker signal setting is worse than that of the stronger signal setting. For the individual estimation, joint tuning appears to work better than separate tuning in terms of variable selection. Interestingly, for simultaneous estimation methods, the method M1″ without non-crossing constraints works better in variable selection than the methods with constraints. Nevertheless, in view of the big advantage of SNQR in terms of test errors, SNQR is more preferable for multiple QR estimation.

Downloaded by [Portland State University] at 22:38 01 December 2014

Table 8. Variable selection results of Example 5 with $\boldsymbol{\beta} = (1.5, 0.75, 0, 0, 1, 0, 0, 0)^{\mathrm{T}}$.

Weight	M1	M1′	M1″	M2	M3
Uniform					
Average wrong 0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Average correct 0	2.76 (0.16)	0.64 (0.11)	4.39 (0.09)	4.27 (0.10)	3.35 (0.19)
Normal					
Average wrong 0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Average correct 0	2.84 (0.15)	0.64 (0.11)	4.26 (0.10)	4.16 (0.11)	3.54 (0.17)

Notes: M1, individual estimation with joint tuning; M1′, individual estimation with separate tuning; M1″, simultaneous SCAD-max estimation without constraints; M2, simultaneous SCAD-max estimation with data point restriction; M3, simultaneous SCAD-max SNQR.

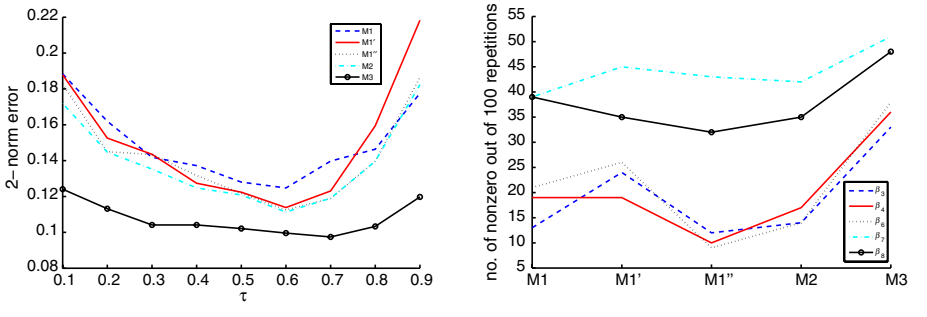


Figure 6. The left panel shows the average squared differences between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ for five different methods using normal weights. The right panel shows the corresponding number of non-zero estimates of β_3, β_4, β_6, β_7 and β_8 for the quantile function with $\tau = 0.4$ based on 100 replications.

One reviewer suggested another setting of the parameter vector $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0.1, 0.1)^{\mathrm{T}}$. In this case, the last two parameters are replaced by small non-zero values, $1/\sqrt{n} = 0.1$ with $n = 100$. A similar example was considered by Leeb and Potscher (2008). Since the last two parameters are close to 0, it can be more difficult to select them compared with other non-zero parameters. On the other hand, a model with only X_1, X_2, X_5 could be a reasonable model as well in terms of prediction and interpretability.

We examine the performance of five different methods, M1, M1′, M1″, M2 and M3, on this example with the new parameter setting. The results with normal weights are displayed in Figure 6. The left panel shows the average squared differences between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, $\sum_{j=0}^{8}(\hat{\beta}_j - \beta_j)^2$, based on 100 replications. Similar as before, our proposed SNQR works the best in terms of parameter estimation. The right panel shows the number of non-zero estimates of β_3, β_4, β_6, β_7, and β_8 for the quantile function with $\tau = 0.4$ among these 100 replications. Notice that all methods have higher percents of non-zero estimates for β_7, β_8 than those of β_3, β_4 and β_6. This is expected since β_7 and β_8 are non-zero while the other three are zero. Due to the small values of these two parameters, all methods estimate β_7 and β_8 as zero over 50% times. We do not plot the selection results for β_1, β_2 and β_5 since the corresponding estimates are non-zero for all replications. Overall, the performance of the proposed SNQR is very reasonable compared with other methods.

5. Real data

In this section, we apply our proposed SNQR to analyse the Annual Salary of Baseball Players Data provided by He, Ng and Portnoy (1998). This data set consists of $n = 263$ North American major

Downloaded by [Portland State University] at 22:38 01 December 2014

Downloaded by [Portland State University] at 22:38 01 December 2014

league baseball players for the 1986 season. Following He et al. (1998), we use the number of home runs in the latest year (performance measure) and the number of years played (seniority measure) as predictor variables. The response variable is the annual salary of each player (measured in thousands of dollars). We first standardise both predictor variables to have mean zero and variance one. We apply the nonlinear QR using the Gaussian kernel with data width parameter σ chosen to be the median pairwise Euclidean distance of the standardised predictor variables. Similar recommendation on data width parameter selection was previously provided by Brown et al. (2000). We use 10-fold cross-validation to select the regularisation parameter λ.

The conditional quantile function is estimated at $\tau = 0.1, 0.2, \ldots, 0.9$. In Figure 7, we plot the individually estimated median function and the Gaussian weighted SNQR estimated median function on the top left and right panels, respectively. To visualise quantile crossing, we plot the difference $\hat{f}_{0.8}(\boldsymbol{x}) - \hat{f}_{0.7}(\boldsymbol{x})$ on the bottom row. The one from the individual estimation is shown on the bottom left panel, and the one from SNQR is displayed on the bottom right panel. Several interesting remarks can be made from the plots. First of all, the conditional median plots suggest that players with large numbers of home runs and moderate numbers of years played have the highest median salaries. This matches our expectation since that group of players have

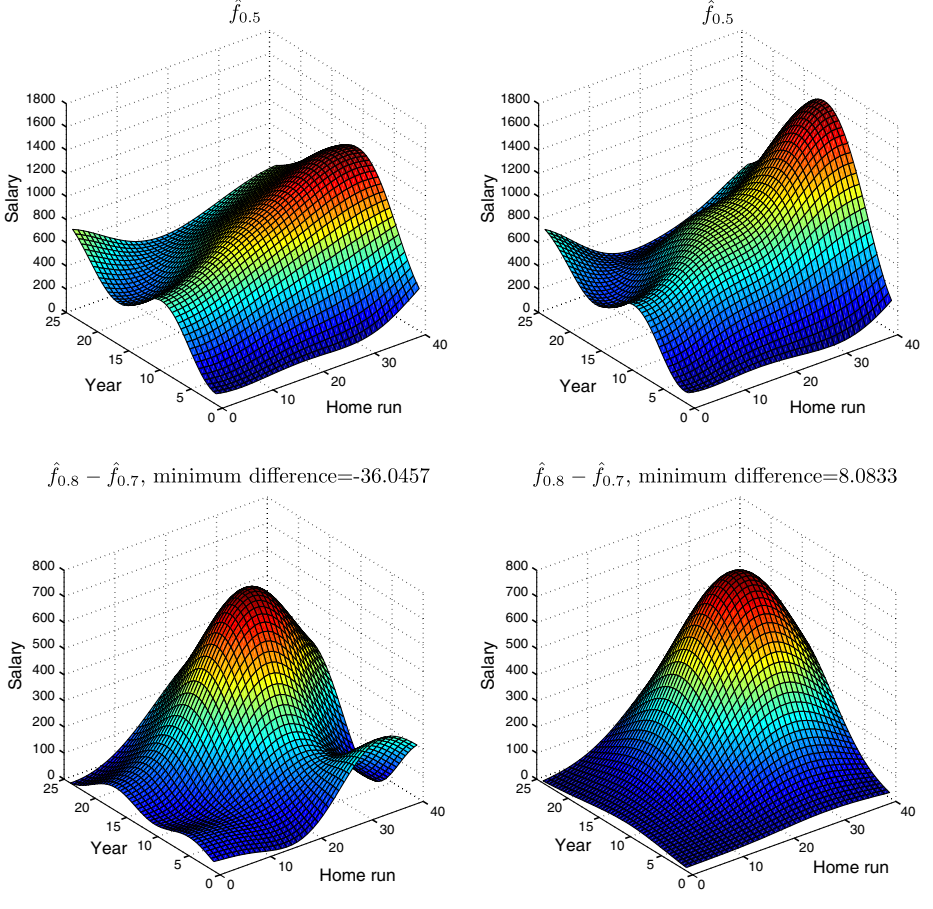


Figure 7. Plots for the Baseball data example. Top left panel: individually estimated median function; top right panel: SNQR estimated median function; bottom left panel: the difference between the individually estimated quantile functions of $\tau = 0.8$ and $\tau = 0.7$; bottom right panel: the difference between the SNQR estimated quantile functions of $\tau = 0.8$ and $\tau = 0.7$.

relatively better skills than other players and are possibly in the peak time of their Baseball career. Between the individually estimated median function and the SNQR median function, the shapes are quite similar although the SNQR median function appears to be slightly more peaked. As to quantile crossing, we can see from the bottom left panel of Figure 7 that the individually estimated 70% quantile function can be higher than that of the 80% quantile function. This undesirable phenomenon disappears when the SNQR is applied. Furthermore, due to the joint estimation, the difference curve of our SNQR is smoother than that from the individual estimation.

6. Discussion

In this paper, we study the problem of multiple conditional quantile function estimation. When individual optimisation is performed, the obtained quantile functions may cross each other and as a result violate the basic property of quantiles. A new method SNQR, which avoids quantile crossing via simple constraints, is proposed. We demonstrate that SNQR can not only help to obtain more interpretable quantile functions, it can also help to improve the estimation efficiency.

As in other regularisation problems, the choice of the regularisation parameter λ is very important for the performance of QR. It is common for one to select a finite set of representative values for λ and then use a separate validation data set or certain model selection criterion to select a value for λ. In this article, we have used separate validation sets for simulation and cross-validation for the real data analysis. As an alternative, one can use certain model selection criterion to choose λ. Two commonly used criteria are the Schwarz information criterion (Schwarz 1978; Koenker et al. 1994) and the generalised approximate cross-validation criterion (Yuan 1978). These criteria are well studied for unconstrained QR and require further developments for our constrained methods.

Our asymptotic study is restricted to the linear SNQR. It will be interesting to explore the asymptotic behaviour of the nonlinear SNQR as well. The existing asymptotic results (e.g. Yu and Jones 1998; Hall et al. 1999; Dette and Volgushev 2008; Chernozhukov et al. 2009) can shed some light here. Further investigation is needed.

Downloaded by [Portland State University] at 22:38 01 December 2014

Acknowledgements

Liu's research was supported in part by NSF grant DMS-0747575 and NIH grant 1R01CA149569-01. Wu's research was supported in part by NSF grant DMS-0905561 and NIH grant 1R01CA149569-01. The authors are indebted to the editor, the associate editor and two referees, whose helpful comments and suggestions led to a much improved presentation.

References

Bondell, H.D., Reich, B.J., and Wang, H. (2010), 'Non-Crossing Quantile Regression Curve Estimation', *Biometrika*, in press.

Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., and Haussler, D. (2000), 'Knowledge-Based Analysis of Microarray Gene Expression Data by using Support Vector Machines', *Proceedings of the National Academy of Science*, 97, 262–267.

Chernozhukov, V., Fernandez-Val, I., and Galichon, A. (2009), *Quantile and Probability Curves without Crossing*. arXiv:0704.3649.

Dette, H., and Volgushev, S. (2008), 'Non-crossing Non-parametric Estimates of Quantile Curves', *Journal of Royal Statistical Society*, Ser. B, 70, 609–627.

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.

Hall, P., Wolff, R.C.L., and Yao, Q. (1999), 'Methods for Estimating a Conditional Distribution Function', *Journal of American Statistical Association*, 94, 154–163.

He, X. (1997), 'Quantile Curves without Crossing', *American Statistician*, 51, 186–192.

He, X., Ng, P., and Portnoy, S. (1998), 'Bivariate Quantile Smoothing Splines', *Journal of the Royal Statistical Society*, Ser. B, 60, 537–550.

Kimeldorf, G., and Wahba, G. (1971), 'Some Results on Tchebycheffian Spline Functions', *Journal of Mathematical Analysis and Applications*, 33, 82–95.
Koenker, R. (2004), 'Quantile Regression for Longitudinal Data', *Journal of Multivariate Analysis*, 91, 74–89.
Koenker, R. (2005), *Quantile Regression* (*Econometric Society Monographs*), New York, NY: Cambridge University Press.
Koenker, R., and Bassett, G. (1978), 'Regression Quantiles', *Econometrica*, 46, 33–50.
Koenker, R., Ng, P., and Portnoy, S. (1994), 'Quantile Smoothing Splines', *Biometrika*, 81, 673–680.
Leeb, H., and Potscher, B.M. (2008), 'Sparse Estimators and the Oracle Property, or the Return of Hodge's Estimator', *Journal of Econometrics*, 142, 201–211.
Li, Y., and Zhu, J. (2008), 'L1-norm Quantile Regression', *Journal of Computational and Graphical Statistics*, 17, 163–185.
Li, Y., Liu, Y., and Zhu, J. (2007), 'Quantile Regression in Reproducing Kernel Hilbert Spaces', *Journal of the American Statistical Association*, 102, 255–268.
Liu, Y., Shen, X., and Doss, H. (2005), 'Multicategory ψ-learning and Support Vector Machine: Computational Tools', *Journal of Computational and Graphical Statistics*, 14, 219–236.
Neocleousa, T., and Portnoy, S. (2007), 'On Monotonicity of Regression Quantile Functions', *Statistics and Probability Letters*, 78, 1226–1229.
Potscher, B.M., and Leeb, H. (2009), 'On the Distribution of Penalized Maximum Likelihood Estimators: The Lasso, Scad, and Thresholding', *Journal of Multivariate Analysis*, 100, 2065–2082.
Potscher, B.M., and Schneider, U. (2010), 'Confidence Sets Based on Penalized Maximum Likelihood Estimators in Gaussian Regression', *Electronic Journal of Statistics*, 4, 334–360.
Scholkopf, B., and Smola, A. (2002), *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond*, Cambridge, MA: MIT Press.
Schwarz, G. (1978), 'Estimating the Dimension of a Model', *Annals of Statistics*, 6, 461–464.
Shim, J., Hwang, C., and Seok, K.H. (2009), 'Non-crossing Quantile Regression via Doubly Penalized Kernel Machine', *Computational Statistics*, 24, 83–94.
Takeuchi, I., and Furuhashi, T. (2004), 'Non-crossing Quantile Regressions by svm', in *Proceedings of the International Joint Conference on Neural Networks*, pp. 401–406.
Takeuchi, I., Le, Q.V., Sears, T.D., and Smola, A.J. (2006), 'Nonparametric Quantile Estimation', *Journal of Machine Learning Research*, 7, 1231–1264.
Tibshirani, R.J. (1996), 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.
Wang, H., Li, G., and Jiang, G. (2007), 'Robust Regression Shrinkage and Consistent Variable Selection Through the Lad-Lasso', *Journal of Business and Economic Statistics*, 25, 347–355.
Wu, Y., and Liu, Y. (2008), 'Variable Selection in Quantile Regression', *Statistica Sinica*, 19, 801–817.
Wu, Y., and Liu, Y. (2009), 'Stepwise Multiple Quantile Regression Estimation using Non-crossing Constraints', *Statistics and Its Interface*, 2, 299–310.
Yu, K., and Jones, M.C. (1998), 'Local Linear Quantile Regression', *Journal of American Statistical Association*, 93, 228–237.
Yuan, M. (1978), 'Gacv for Quantile Smoothing Splines', *Computational Statistics and Data Analysis*, 5, 813–829.
Zhang, H.H., Liu, Y., Wu, Y., and Zhu, J. (2008), 'Multicategory Sup-norm Support Vector Machines', *Electronic Journal of Statistics*, 2, 149–167.
Zou, H. (2006), 'The Adaptive Lasso and Its Oracle Properties', *Journal of the American Statistical Association*, 101, 1418–1429.
Zou, H., and Li, R. (2008), 'One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models (with Discussion)', *Annals of Statistics*, 36, 1509–1566.
Zou, H., and Yuan, M. (2008a), 'Composite Quantile Regression and the Oracle Model Selection Theory', *Annals of Statistics*, 36, 1108–1126.
Zou, H., and Yuan, M. (2008b), 'Regularized Simultaneous Model Selection in Multiple Quantiles Regression', *Computational Statistics and Data Analysis*, 52, 5296–5304.

Downloaded by [Portland State University] at 22:38 01 December 2014

Appendix

Proof of Proposition 1 The result can be shown directly using integration by parts. The details are not included here to save space. ■

Proof of Lemma 1 The result is straightforward by applying Theorem 4.1 of Koenker (2005) to each τ_k. ■

Proof of Proposition 2 In theory, it is guaranteed that

$$d_k \triangleq \inf_{\boldsymbol{x}\in\mathcal{X}} \{[\beta_0(\tau_{k+1}) - \beta_0(\tau_k)] + \boldsymbol{x}^{\mathrm{T}}[\boldsymbol{\beta}(\tau_{k+1}) - \boldsymbol{\beta}(\tau_k)]\} > 0$$

Downloaded by [Portland State University] at 22:38 01 December 2014

due to Condition A. Using the triangle inequality, we have

$$\begin{aligned}\inf_{x\in\mathcal{X}}\{\tilde{\beta}_{0\tau_{k+1}}-\tilde{\beta}_{0\tau_k}+\boldsymbol{x}^{\mathrm{T}}(\tilde{\boldsymbol{\beta}}_{\tau_{k+1}}-\tilde{\boldsymbol{\beta}}_{\tau_k})\} &\geq \inf_{x\in\mathcal{X}}\{\boldsymbol{x}^{\mathrm{T}}[\tilde{\boldsymbol{\beta}}_{\tau_{k+1}}-\boldsymbol{\beta}(\tau_{k+1})]\}+(\tilde{\beta}_{0\tau_{k+1}}-\beta_0(\tau_{k+1}))+(\beta_0(\tau_{k+1})-\beta_0(\tau_k))\\ &\quad+\inf_{x\in\mathcal{X}}\{\boldsymbol{x}^{\mathrm{T}}[\boldsymbol{\beta}(\tau_{k+1})-\boldsymbol{\beta}(\tau_k)]\}+(\beta_0(\tau_k)-\tilde{\beta}_{0\tau_k})+\inf_{x\in\mathcal{X}}\{\boldsymbol{x}^{\mathrm{T}}[\boldsymbol{\beta}(\tau_k)-\tilde{\boldsymbol{\beta}}_{\tau_k}]\}\\ &\geq -\sup_{x\in\mathcal{X}}|\boldsymbol{x}^{\mathrm{T}}[\tilde{\boldsymbol{\beta}}_{\tau_{k+1}}-\boldsymbol{\beta}(\tau_{k+1})]|-|\tilde{\beta}_{0\tau_{k+1}}-\beta_0(\tau_{k+1})|+(\beta_{0\tau_{k+1}}-\beta_{0\tau_k})\\ &\quad+\inf_{x\in\mathcal{X}}\{\boldsymbol{x}^{\mathrm{T}}[\boldsymbol{\beta}(\tau_{k+1})-\boldsymbol{\beta}(\tau_k)]\}-|\beta_0(\tau_k)-\tilde{\beta}_{0\tau_k}|-\sup_{x\in\mathcal{X}}|\boldsymbol{x}^{\mathrm{T}}[\boldsymbol{\beta}(\tau_k)-\tilde{\boldsymbol{\beta}}_{\tau_k}]|.\end{aligned}$$

Another application of the triangle inequality leads to

$$\sup_{x\in\mathcal{X}}|\boldsymbol{x}^{\mathrm{T}}[\tilde{\boldsymbol{\beta}}_{\tau_k}-\boldsymbol{\beta}(\tau_k)]|\leq\left(\sup_{x\in\mathcal{X}}\|\boldsymbol{x}\|\right)\|\tilde{\boldsymbol{\beta}}_{\tau_k}-\boldsymbol{\beta}(\tau_k)\|. \tag{A1}$$

Denote $M=\sup_{x\in\mathcal{X}}\|\boldsymbol{x}\|$. Consequently, we have

$$\begin{aligned}&P\left(\sup_{x\in\mathcal{X}}\{\tilde{\beta}_{0\tau_{k+1}}-\tilde{\beta}_{0\tau_k}+\boldsymbol{x}^{\mathrm{T}}(\tilde{\boldsymbol{\beta}}_{\tau_{k+1}}-\tilde{\boldsymbol{\beta}}_{\tau_k})\}<0\right)\\ &\leq P\left(\sup_{x\in\mathcal{X}}|\boldsymbol{x}^{\mathrm{T}}[\tilde{\boldsymbol{\beta}}_{\tau_{k+1}}-\boldsymbol{\beta}(\tau_{k+1})]|>\frac{d_k}{4}\right)+P\left(\sup_{x\in\mathcal{X}}|\boldsymbol{x}^{\mathrm{T}}[\tilde{\boldsymbol{\beta}}_{\tau_k}-\boldsymbol{\beta}(\tau_k)]|>\frac{d_k}{4}\right)\\ &\quad+P\left(|\beta_0(\tau_k)-\tilde{\beta}_{0\tau_k}|>\frac{d_k}{4}\right)+P\left(|\beta_0(\tau_{k+1})-\tilde{\beta}_{0\tau_{k+1}}|>\frac{d_k}{4}\right)\\ &\leq P\left(\|\tilde{\boldsymbol{\beta}}_{\tau_{k+1}}-\boldsymbol{\beta}(\tau_{k+1})\|>\frac{d_k}{2M}\right)+P\left(\|\tilde{\boldsymbol{\beta}}_{\tau_k}-\boldsymbol{\beta}(\tau_k)\|>\frac{d_k}{2M}\right)\\ &\quad+P\left(|\beta_0(\tau_k)-\tilde{\beta}_{0\tau_k}|>\frac{d_k}{4}\right)+P\left(|\beta_0(\tau_{k+1})-\tilde{\beta}_{0\tau_{k+1}}|>\frac{d_k}{4}\right).\end{aligned} \tag{A2}$$

Based on Lemma 1, the sum of probabilities in Equation (A2) decays exponentially. Thus, we have $P(\sup_{x\in\mathcal{X}}\{\tilde{\beta}_{0\tau_{k+1}}-\tilde{\beta}_{0\tau_k}+\boldsymbol{x}^{\mathrm{T}}(\tilde{\boldsymbol{\beta}}_{\tau_{k+1}}-\tilde{\boldsymbol{\beta}}_{\tau_k})\}<0)<e^{-na_k}$ asymptotically for some $a_k>0$. This completes the proof by noting that

$$P(\mathcal{C}_n)\leq\sum_{k=1}^{K}P\left(\sup_{x\in\mathcal{X}}\{\tilde{\beta}_{0\tau_{k+1}}-\tilde{\beta}_{0\tau_k}+\boldsymbol{x}^{\mathrm{T}}(\tilde{\boldsymbol{\beta}}_{\tau_{k+1}}-\tilde{\boldsymbol{\beta}}_{\tau_k})\}<0\right).$$

■

Proof of Proposition 3 The sufficiency of the constraints is straightforward. For necessity, we prove parts (i) and (ii) separately. For (i), the necessity of constraint (12) can be shown by letting $\boldsymbol{x}^*=(0,\ldots,0)$ for $f_{\tau_k}(\boldsymbol{x})$. For Equation (13), let $\boldsymbol{x}^*=(0,\ldots,0,M,0,\ldots,0)$, i.e. all elements are 0 except the jth element being $M>0$. Then $f_{\tau_k}(\boldsymbol{x}^*)=\beta_{0\tau_k}+\beta_{\tau_k,j}M$ and $f_{\tau_{k+1}}(\boldsymbol{x}^*)=\beta_{0\tau_{k+1}}+\beta_{\tau_{k+1},j}M$. Since $\beta_{0\tau_k}$ and $\beta_{0\tau_{k+1}}$ are bounded, the constraint $\beta_{\tau_k,j}\leq\beta_{\tau_{k+1},j}$ is necessary to ensure $f_{\tau_k}(\boldsymbol{x}^*)\leq f_{\tau_{k+1}}(\boldsymbol{x}^*)$ for arbitrarily large M. The conclusion in (i) then follows.

For (ii), $f_{\tau_k}(\boldsymbol{x})=\sum_{i=1}^{n}w_{\tau_k,i}\langle\boldsymbol{x}_i,\boldsymbol{x}\rangle+b_{\tau_k}$. Without loss of generality, assume that the design matrix is of rank n. Since $d>n$, there exists $\boldsymbol{x}^*\in\mathcal{X}$ such that $\boldsymbol{x}^*\perp\boldsymbol{x}_i$ for $\forall i\neq i'$ and $\langle\boldsymbol{x}^*,\boldsymbol{x}_{i'}\rangle=M$. Then $f_{\tau_k}(\boldsymbol{x}^*)\leq f_{\tau_{k+1}}(\boldsymbol{x}^*)$ implies that $w_{\tau_k,i'}M+b_{\tau_k}\leq w_{\tau_{k+1},i'}M+b_{\tau_{k+1}}$. When $M=0$, we have $b_{\tau_k}\leq b_{\tau_{k+1}}$. Moreover, we have $w_{\tau_k,i'}\leq w_{\tau_{k+1},i'}$ with M being arbitrarily large. Then part (ii) follows. ■

Proof of Theorem 1 The desired result is straightforward by combining Lemma 1 and Proposition 2. ■

Proof of Lemma 2 It is enough to show that for any $\delta>0$, there exists a large constant C such that

$$P\left(\inf_{\sum_{j=1}^{p}(\|\boldsymbol{\alpha}_j\|^2+a_j^2)=C}Q(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\ldots,\boldsymbol{\alpha}_K,a_1,a_2,\ldots,a_K)>Q(\mathbf{0},\mathbf{0},\ldots,\mathbf{0},0,0,\ldots,0)\right)>1-\delta.$$

This will imply that with probability at least $1-\delta$ there exists a local minimum inside the ball of $\boldsymbol{\beta}(\tau_k)+\boldsymbol{\alpha}_k/\sqrt{n}$, $\beta_0(\tau_k)+a_k/\sqrt{n}$, $k=1,2,\ldots,K$ with $\boldsymbol{\alpha}_k$ and a_k satisfying $\sum_{k=1}^{K}(\|\boldsymbol{\alpha}_k\|^2+a_k^2)=C$.

Downloaded by [Portland State University] at 22:38 01 December 2014

Note that

$$\begin{aligned}
&Q(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_K, a_1, a_2, \ldots, a_K) - Q(\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, 0, 0, \ldots, 0) \\
&= \sum_{k=1}^{K} Z_{nk}(\boldsymbol{\alpha}_k, a_k) + n \sum_{j=1}^{p} p_\lambda \left(\max_{k=1}^{K} \left| \beta_j(\tau_k) + \frac{\alpha_{jk}}{\sqrt{n}} \right| \right) - n \sum_{j=1}^{p} p_\lambda \left(\max_{k=1}^{K} |\beta_j(\tau_k)| \right) \\
&\geq \sum_{k=1}^{K} Z_{nk}(\boldsymbol{\alpha}_k, a_k) + n \sum_{j \in \mathcal{A}} \left[p_\lambda \left(\max_{k=1}^{K} \left| \beta_j(\tau_k) + \frac{\alpha_{jk}}{\sqrt{n}} \right| \right) - p_\lambda \left(\max_{k=1}^{K} |\beta_j(\tau_k)| \right) \right],
\end{aligned}$$

where the last inequality is due to the fact that $\max_{k=1}^{K} |\beta_j(\tau_k)| = 0$ for $j \notin \mathcal{A}$ and $p_\lambda(\cdot)$ is non-decreasing on $[0, \infty)$.

Note further that

$$\begin{aligned}
&n \sum_{j \in \mathcal{A}} \left[p_\lambda \left(\max_{k=1}^{K} \left| \beta_j(\tau_k) + \frac{\alpha_{jk}}{\sqrt{n}} \right| \right) - p_\lambda \left(\max_{k=1}^{K} |\beta_j(\tau_k)| \right) \right] \\
&\leq n \sum_{j \in \mathcal{A}} \left[p'_\lambda \left(\max_{k=1}^{K} |\beta_j(\tau_k)| \right) \sum_{k=1}^{K} \frac{|\alpha_{jk}|}{\sqrt{n}} \right] + n \frac{1 + o(1)}{2} \sum_{j \in \mathcal{A}} \left[p''_\lambda \left(\max_{k=1}^{K} |\beta_j(\tau_k)| \right) \sum_{k=1}^{K} \frac{\alpha_{jk}^2}{n} \right] \\
&\leq \sqrt{n} c_n \sum_{j \in \mathcal{A}} \sum_{k=1}^{K} |\alpha_{jk}| + \frac{1 + o(1)}{2} \left[\sum_{j \in \mathcal{A}} \sum_{k=1}^{K} \alpha_{jk}^2 \right] \max_{j \in \mathcal{A}} \left| p''_\lambda \left(\max_{k=1}^{K} |\beta_j(\tau_k)| \right) \right|.
\end{aligned}$$

According to Koenker (2005), we have

$$Z_{nk}(\boldsymbol{\alpha}_k, a_k) \longrightarrow -(a_k, \boldsymbol{\alpha}_k^{\mathrm{T}}) \mathbf{W}_k + \tfrac{1}{2}(a_k, \boldsymbol{\alpha}_k^{\mathrm{T}}) \Sigma_1(\tau_k)(a_k, \boldsymbol{\alpha}_k^{\mathrm{T}})^{\mathrm{T}} \text{ in distribution, as } n \to \infty, \tag{A3}$$

where $\mathbf{W}_k \sim N(\mathbf{0}, \tau_k(1 - \tau_k)\Sigma_0)$.

Recall that we assume that $c_n = O(n^{-1/2})$ and $\max_{j \in \mathcal{A}} |p''_\lambda(\max_{k=1}^{K} |\beta_j(\tau_k)|)| \to 0$. Thus asymptotically,

$$Z_{nk}(\boldsymbol{\alpha}_k, a_k) + n \sum_{j \in \mathcal{A}} \left[p_\lambda \left(\max_{k=1}^{K} \left| \beta_j(\tau_k) + \frac{\alpha_{jk}}{\sqrt{n}} \right| \right) - p_\lambda \left(\max_{k=1}^{K} |\beta_j(\tau_k)| \right) \right]$$

is dominated by the quadratic term $\frac{1}{2} \sum_{k=1}^{K} (a_k, \boldsymbol{\alpha}_k^{\mathrm{T}}) \Sigma_1(\tau_k)(a_k, \boldsymbol{\alpha}_k^{\mathrm{T}})^{\mathrm{T}}$ when C is large enough. This completes the proof. ∎

Proof of Lemma 3 Note that

$$\begin{aligned}
&Q\left(\begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}1} \\ \boldsymbol{\alpha}_{\mathcal{A}^c 1} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}2} \\ \boldsymbol{\alpha}_{\mathcal{A}^c 2} \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}K} \\ \boldsymbol{\alpha}_{\mathcal{A}^c K} \end{pmatrix}, a_1, a_2, \ldots, a_K \right) - Q\left(\begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}1} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}2} \\ \mathbf{0} \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}K} \\ \mathbf{0} \end{pmatrix}, a_1, a_2, \ldots, a_K \right) \\
&= \sum_{k=1}^{K} Z_{nk}\left(\begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}k} \\ \boldsymbol{\alpha}_{\mathcal{A}^c k} \end{pmatrix}, a_k \right) + n \sum_{j=s+1}^{p} p_\lambda \left(\max_{k=1}^{K} \left| \beta_j(\tau_k) + \frac{\alpha_{jk}}{\sqrt{n}} \right| \right) \\
&\quad - \sum_{k=1}^{K} Z_{nk}\left(\begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}k} \\ \mathbf{0} \end{pmatrix}, a_k \right) + n \sum_{j=s+1}^{p} p_\lambda \left(\max_{k=1}^{K} |\beta_j(\tau_k)| \right).
\end{aligned}$$

Note that

$$Z_{nk}\left(\begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}k} \\ \boldsymbol{\alpha}_{\mathcal{A}^c k} \end{pmatrix}, a_k \right) - Z_{nk}\left(\begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}k} \\ \mathbf{0} \end{pmatrix}, a_k \right) \longrightarrow -\boldsymbol{\alpha}_{\mathcal{A}^c k}^{\mathrm{T}} \mathbf{W}_{\mathcal{B}^c k} + (a_k, \boldsymbol{\alpha}_{\mathcal{A}k}^{\mathrm{T}}) \Sigma_{1, \mathcal{B}, \mathcal{B}^c} \boldsymbol{\alpha}_{\mathcal{A}^c k}.$$

Recall that $\beta_j(\tau_k) = 0$ for $j > s$. Thus

$$\begin{aligned}
&n \sum_{j=s+1}^{p} p_\lambda \left(\max_{k=1}^{K} \left| \beta_j(\tau_k) + \frac{\alpha_{jk}}{\sqrt{n}} \right| \right) - n \sum_{j=s+1}^{p} p_\lambda \left(\max_{k=1}^{K} |\beta_j(\tau_k)| \right) \\
&= n \sum_{j=s+1}^{p} p_\lambda \left(\max_{k=1}^{K} \left| \frac{\alpha_{jk}}{\sqrt{n}} \right| \right) \\
&\geq n \sum_{j=s+1}^{p} \lambda \left(\liminf_{\lambda \to 0} \liminf_{\theta \to 0+} \frac{p'_\lambda(\theta)}{\lambda} \right) \left(\max_{k=1}^{K} \left| \frac{\alpha_{jk}}{\sqrt{n}} \right| \right) (1 + o(1)) \\
&= (1 + o(1)) \sqrt{n} \lambda \sum_{j=s+1}^{p} \left(\max_{k=1}^{K} |\alpha_{jk}| \right) \left(\liminf_{\lambda \to 0} \liminf_{\theta \to 0+} \frac{p'_\lambda(\theta)}{\lambda} \right).
\end{aligned}$$

This completes the proof by noting that $\sqrt{n}\lambda \to \infty$ and as a result

$$n \sum_{j=s+1}^{p} p_\lambda \left(\max_{k=1}^{K} \left| \beta_j(\tau_k) + \frac{\alpha_{jk}}{\sqrt{n}} \right| \right) - n \sum_{j=s+1}^{p} p_\lambda \left(\max_{k=1}^{K} |\beta_j(\tau_k)| \right)$$

dominates

$$\sum_{k=1}^{K} \left(Z_{nk} \left(\begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}k} \\ \boldsymbol{\alpha}_{\mathcal{A}^c k} \end{pmatrix}, a_k \right) - Z_{nk} \left(\begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{A}k} \\ \mathbf{0} \end{pmatrix}, a_k \right) \right)$$

asymptotically as $n \to \infty$. ■

Proof of Theorem 2 This is straightforward due to Lemmas 2 and 3. ■

Proof of Proposition 4 Note for the SCAD penalty, $p_\lambda(\theta)$ is flat as long as $|\theta| > a\lambda$. Lemma 2 implies that $\tilde{\boldsymbol{\beta}}_{j\tau_k}$ is consistent. Thus we are solving Equation (28) in a neighbourhood of true $\beta_j(\tau_k)$ and, consequently, when n is large enough, $p_\lambda(\max_{k=1}^{K} |\beta_{j\tau_k})$ is flat by noting that $\lambda \to 0$ as $n \to \infty$. Then the asymptotical normality (29) is valid. ■

Proof of Theorem 3 This can be proved in the same way as how Theorem 1 is proved using Proposition 2. ■

Downloaded by [Portland State University] at 22:38 01 December 2014