May, 2012

Smoothness Selection for Penalized Quantile Regression Splines

Philip T. Reiss
Lei Huang, *Johns Hopkins University*

The International Journal of Biostatistics

Manuscript 1381

Smoothness Selection for Penalized Quantile Regression Splines

Philip T. Reiss, *New York University and Nathan Kline Institute*
Lei Huang, *Johns Hopkins University*

Smoothness Selection for Penalized Quantile Regression Splines

Philip T. Reiss and Lei Huang

Abstract

Modern data-rich analyses may call for fitting a large number of nonparametric quantile regressions. For example, growth charts may be constructed for each of a collection of variables, to identify those for which individuals with a disorder tend to fall in the tails of their age-specific distribution; such variables might serve as developmental biomarkers. When such a large set of analyses are carried out by penalized spline smoothing, reliable automatic selection of the smoothing parameter is particularly important. We show that two popular methods for smoothness selection may tend to overfit when estimating extreme quantiles as a smooth function of a predictor such as age; and that improved results can be obtained by multifold cross-validation or by a novel likelihood approach. A simulation study, and an application to a functional magnetic resonance imaging data set, demonstrate the favorable performance of our methods.

KEYWORDS: asymmetric Laplace distribution, functional connectivity, generalized approximate cross-validation, growth chart, nonparametric quantile regression, smoothing parameter

Author Notes: The authors thank Eva Petkova for many valuable discussions; Adriana Di Martino, for her pivotal role in acquiring the fMRI data; Maarten Mennes, for assistance with preprocessing the data and visualizing the results; Xavier Castellanos, Mike Milham and Juan Zhou, for advice on applying our methodology to these data; and Roger Koenker, Doug Nychka and Giovanna Ranalli, for conversations and correspondence regarding some of their related work; and the Editor, Marten Wegkamp, and two reviewers, whose feedback led to significant improvements in the paper. The first author gratefully acknowledges the support of the National Science Foundation through grant DMS-0907017.

1 Introduction

Estimating percentiles of parameters such as height and weight as a smooth function of age is a well-studied problem, with roots in the investigations of Quetelet (1830). Statistical solutions to this problem underpin the standard growth charts that have been routinely used by pediatricians in the United States since the 1970s (Ogden et al., 2002). The data-rich character of today's biomedical research calls for a new, "high-throughput" variant of growth chart estimation: finding percentile curves for each of a large number of variables, with a view toward identifying some of these variables as potential markers of abnormal development.

This paper was motivated by functional magnetic resonance imaging (fMRI) experiments, in which the blood oxygen level dependent (BOLD) signal, an index of brain activity, is recorded at each of a dense grid of brain locations, known as voxels. Traditionally, subjects were scanned while attending to a series of stimuli; but a great deal of recent work has focused on resting-state fMRI (Biswal et al., 1995), in which individuals are scanned while at rest. A key objective of such studies is to understand functional connectivity, the temporal correlation between time courses of different brain regions (Friston, 1994). Some recent work has examined how functional connectivity develops with age (e.g., Fair et al., 2008), and has identified connections (pairs of regions of interest, or ROIs) for which abnormal developmental trajectories may be associated with psychiatric or neurological disorders (e.g., Church et al., 2009). In light of this work, our psychiatrist colleagues have expressed interest in functional connectivity growth charts that might be used to screen for risk of psychiatric disorders, in much the same way that pediatricians refer to growth charts to detect deviations from age-specific norms for height or weight. While such routine clinical applications are likely only a theoretical possibility for functional connectivity growth charts, these quantile curves may prove to be scientifically informative, as our application will illustrate.

A common approach to estimating percentile curves is the "LMS" method (Cole and Green, 1992), which fits a smoothly varying Box-Cox-transformed normal distribution to the data. But a number of authors, following a suggestion of Cox (1988), have opted instead for the quantile regression paradigm of Koenker and Bassett (1978), which is robust to departures from the LMS method's distributional assumptions, such as unimodality (Wei et al., 2006). In this paper we apply the nonparametric quantile regression framework, and more specifically a penalized spline approach, to functional connectivity in a set of connections between ROIs. We estimate quantile curves for the connections using a sample of normal individuals, and compare the results with data from individuals with attention deficit/hyperactivity disorder (ADHD). By identifying connections for which individuals with ADHD tend to have values in the tails of their age-specific conditional distribution, we may

gain insight into developmental anomalies associated with the disorder.

Our penalized spline method entails choosing a tuning parameter that controls the smoothness of the function estimate. In general the optimal degree of smoothness will depend on the quantile of interest. Because we are estimating growth charts for connectivity in a large number of connections, rather than for just a single variable as in conventional applications, it is critical to have a reliable automatic procedure to choose the optimal tuning parameter for each connection's quantile curve. This need is rendered even more acute by the relatively small size (128) of our normal sample. We have found, however, that in samples of this size, standard criteria for automatic smoothing parameter selection are unreliable for extreme quantiles, which are the quantiles of interest in our setting. Improved smoothness selection is therefore the major methodologic objective of this paper.

Section 2 outlines the penalized spline approach to nonparametric quantile regression. We review previous, prediction-error-based approaches to automatic smoothing parameter selection in Section 3, and introduce a new, likelihood-based approach in Section 4. Simulations in Section 5, and our analysis of the ADHD data in Section 6, point to advantages of the likelihood approach, and of multifold cross-validation, over more popular methods for smoothness selection. Section 7 offers concluding remarks.

2 Nonparametric quantile regression with penalized splines

In what follows we assume that we have a sample of n individuals with predictor values $x_1,\ldots,x_n$ (e.g., age), and responses $y_1,\ldots,y_n$ (e.g., functional connectivity for a particular pair of regions). Given a value $\tau \in [0,1]$, nonparametric quantile regression seeks to estimate the presumably smooth function $g(x)$ defined as the conditional $100\tau\%$ quantile of y given x. A popular general approach is to obtain an estimate $\hat{g} = \hat{g}_{\tau,\lambda}$ that minimizes

$$\sum_{i=1}^{n} \rho_\tau[y_i - g(x_i)] + \lambda J(g) \tag{1}$$

over an appropriate function space. Here ρ_τ is the "check function" of Koenker and Bassett (1978), given by

$$\rho_\tau(u) = \tau u^+ + (1-\tau)u^-$$

where $u^+ = \max(u,0)$ and $u^- = \max(-u,0)$, $J(g)$ is a roughness functional, and λ is a tuning parameter determining the extent to which roughness is penalized.

Koenker et al. (1994) take $J(g)$ to be a total variation penalty on g', for which linear programming can be used to find the minimizer of (1) over a particular function space that they define. Some subsequent work has retained the form (1) for the objective function but has differed from the approach of Koenker et al. (1994) in one or both of the following respects:

1. Some authors (e.g., Nychka et al., 1995, Bosch et al., 1995, Yuan, 2006) take $J(g) = \int [g''(x)]^2 dx$ (cf. Cox, 1983), a traditional roughness functional for ordinary nonparametric regression, which may enforce a more visually appealing form of "smoothness" than alternative functionals do.
2. A number of authors (e.g., Ng and Maechler, 2007, Pratesi et al., 2009) have taken the function space to be the span of a set of basis functions $b_1, \ldots, b_K$ such as B-splines, which combine favorable approximation-theoretic properties (De Boor, 2001) with computational efficiency. In other words, g is required to have the form $g(x) = b(x)^T \gamma$ for some $\gamma \in \mathscr{R}^K$, where $b(x) = [b_1(x), \ldots, b_K(x)]^T$.

In this paper we adopt both of these modifications of the quantile smoothing spline framework, and take the basis functions to be cubic B-splines. Thus our function estimate $\hat{g}_\lambda(x) = b(x)^T \hat{\gamma}$ (from here on we suppress the dependence on τ) will be found by solving the minimization problem

$$\begin{aligned} \hat{\gamma} &= \underset{\gamma \in \mathscr{R}^K}{\arg\min} \left[\sum_{i=1}^n \rho_\tau \{ y_i - b(x_i)^T \gamma \} + \lambda \int \{ b''(x)^T \gamma \}^2 dx \right] \\ &= \underset{\gamma \in \mathscr{R}^K}{\arg\min} \left[\sum_{i=1}^n \rho_\tau (y_i - b_i^T \gamma) + \lambda \gamma^T P \gamma \right], \end{aligned} \tag{2}$$

where $b_i = b(x_i)$ and $P = [\int b_i''(x) b_j''(x) dx]_{1 \le i,j \le K}$. For a given λ, the minimization can be performed by penalized iteratively reweighted least squares (PIRLS) (Nychka et al., 1995, Pratesi et al., 2009), yielding an estimate of the form

$$\hat{\gamma} = (B^T W B + \lambda P)^{-1} B^T W y, \tag{3}$$

where $B = (b_1 \ldots b_n)^T$, $y = (y_1, \ldots, y_n)^T$, and W is an $n \times n$ diagonal matrix whose diagonal elements are weights, described in Appendix A, which are iterated until convergence. As noted in the introduction, our main concern in this paper is optimal choice of λ.

We remark that an alternative to the minimizing (1) is the regression spline method (e.g., Wei and He, 2006, Wei et al., 2006), which omits the penalty and

restricts g to the span of a low-rank B-spline basis. This approach is more theoretically tractable than penalized splines, but is much more dependent on a suitable choice of the knots. In our application it is impractical to find a good choice of knots for each of the large number of connections for which growth charts are estimated.

3 Previous approaches to smoothing parameter selection

3.1 Schwarz information criterion

Koenker et al. (1994) propose to adapt the Schwarz (1978) information criterion (SIC) to the choice of λ in nonparametric quantile regression; that is, they choose λ in (1) to minimize

$$SIC(\lambda) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\{y_i - \hat{g}_\lambda(x_i)\}\right] + \frac{\log n}{2n}\mathrm{df}_\lambda, \tag{4}$$

where df_λ denotes the effective degrees of freedom of the fit. When (1) is minimized by PIRLS, it is conventional (e.g., Pratesi et al., 2009) to define the effective df as

$$\mathrm{df}_\lambda = \mathrm{tr}(H_\lambda), \tag{5}$$

where $H_\lambda = (h_{ij})_{1\le i,j\le n} = B(B^TWB + \lambda P)^{-1}B^TW$ is the "hat" matrix obtained at convergence such that

$$[\hat{g}_\lambda(x_1), \ldots, \hat{g}_\lambda(x_n)]^T = H_\lambda y$$

(see equation (3), and cf. Debruyne et al., 2008, Section 5.3). The df provide a useful index of the complexity of a fitted curve, with 2 df corresponding to a linear fit, and higher df implying bumpier fits. It can thus serve as a basis of comparison among methods, as in Figure 2 below. The application of SIC is based on an analogy with its use in mean regression, but to our knowledge has never been rigorously justified for nonparametric quantile regression.

3.2 Approximate versions of cross-validation

Cross-validation (CV) approaches to smoothing parameter selection for quantile spline smoothing are investigated by Yuan (2006). The starting point is to find a value λ that approximately minimizes the risk

$$\frac{1}{n}\sum_{i=1}^{n} E_z\rho_\tau[z_i - \hat{g}_\lambda(x_i)], \tag{6}$$

for a future sample $z_1, \ldots, z_n$ such that, for each i, the distribution of z_i conditional on x_i is the same as that of y_i. A very similar idea motivates the Akaike (1973) information criterion (AIC), but the loss function used there is the Kullback-Leibler information; thus Yuan (2006), following Wahba (1999), refers to (6) as the generalized comparative Kullback-Leibler distance (GCKL). Since the true distribution of the y_i's is unknown, (6) cannot be computed, so we instead minimize the leave-one-out CV criterion with check-function loss, i.e.

$$\frac{1}{n}\sum_{i=1}^{n}\rho_\tau[y_i - \hat{g}_\lambda^{[-i]}(x_i)], \tag{7}$$

where $\hat{g}_\lambda^{[-i]}$ is the function estimate based on all but the ith observation. This criterion is referred to in Nychka et al. (1995) as quantile CV and in Yuan (2006) as robust CV. There is a subtle difference here in that CV uses each left-out pair (x_i, y_i) as a proxy for entirely new data, whereas in the scenario underlying (6) we retain the original predictor data and generate a new set of responses. Nevertheless, as Yuan (2006) argues, (7) should be an approximately unbiased estimate of (6) in large samples (cf. Stone, 1977).

To avoid the computational expense of computing each leave-one-out function estimate $\hat{g}_\lambda^{[-i]}$, Nychka et al. (1995) propose the approximate cross-validation (ACV) criterion

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\rho_\tau[y_i - \hat{g}_\lambda(x_i)]}{1-h_{ii}}. \tag{8}$$

Appendix B explains why (8) is approximately equal to (7).

Yuan (2006) proposes to replace h_{ii} in (8) by its average value, yielding the generalized approximate cross-validation (GACV) criterion

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\rho_\tau[y_i - \hat{g}_\lambda(x_i)]}{1-\mathrm{tr}(H_\lambda)/n}. \tag{9}$$

He shows that this step, borrowed from the derivation of generalized cross-validation (GCV; Craven and Wahba, 1979), alleviates the failure of (8) to approximate the risk well when most of the h_{ii}'s are close to 0.

We have found, however, that GACV often severely overfits for extreme quantiles, i.e. τ near 0 or 1. Figure 1 illustrates the reason for this phenomenon, using a particular data set that we believe is representative of the general problem. The leverage h_{ii} for each of 150 observations is plotted against the corresponding summand $\rho_\tau[y_i - \hat{g}_\lambda(x_i)]/(1-h_{ii})$ in the ACV criterion (8), for quantiles $\tau = .01, .3$ and smoothing parameters $\lambda = .001, 1$. Observe that (a) the smaller λ results in more high-leverage ($h_{ii} \approx 1$) observations, and (b) for the more extreme τ, the summands

corresponding to high-leverage observations tend to be very large. Replacing each observation's leverage with the mean leverage downweights the high-leverage summands, so that when both (a) and (b) obtain, (9) will generally be smaller than (8). In other words, with extreme τ, GACV will be decreased relative to ACV for small λ, so GACV will tend to choose smaller λ than ACV does.

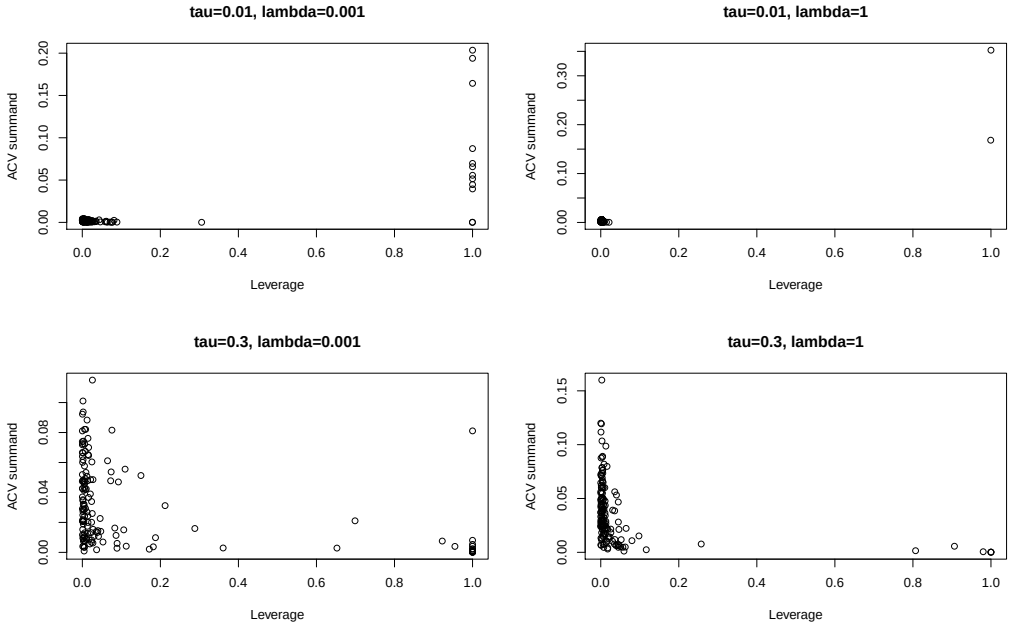

Figure 1: Illustration of why GACV favors a small smoothing parameter λ for extreme quantiles (see Section 3).

3.3 Multifold cross-validation

Multifold CV (Zhang, 1993) can greatly reduce the computational burden without appealing to the approximations motivating ACV and GACV. Here we divide the n observations into "validation sets" $V_1, \ldots, V_k$ of (approximately) equal size, and define the criterion

$$\frac{1}{n} \sum_{j=1}^{k} \sum_{i \in V_j} \rho_\tau [y_i - \hat{g}_\lambda^{[-V_j]}(x_i)], \tag{10}$$

where $\hat{g}_\lambda^{[-V_j]}$ is the function estimate based on the observations not belonging to V_j (the special case $k = n$ yields the leave-one-out CV criterion (7)). In general, small values of k produce downward-biased estimates of prediction error, whereas larger

values produce more variable results and impose a higher computational burden; $k = 5$ or 10 is often recommended as a compromise (Hastie et al., 2009).

While multifold CV is a widely used general technique, it seems not to be popular in the quantile smoothing context; but Section 5 below demonstrates that it can clearly outperform GACV and SIC, especially for extreme quantiles.

4 Likelihood-based smoothness selection

4.1 Background

In other roughness penalty smoothing contexts, likelihood-based smoothing parameter selection has been proposed as an alternative to prediction error-based approaches such as GCV or information criteria (Wahba, 1985, Ruppert et al., 2003, Reiss and Ogden, 2009, Wood, 2011). Given the difficulties we encountered with standard smoothness selection approaches for penalized quantile regression splines, we wondered whether a likelihood approach might work here. At first glance, this idea may seem unpromising, given that the check function is not generally thought of as arising from a likelihood, and accordingly (1) is not obviously related to a penalized log-likelihood. Some authors, however (e.g., Yu and Moyeed, 2001, Komunjer, 2005, Geraci and Bottai, 2007, Reich et al., 2010), have successfully approached certain quantile regression problems from a likelihood perspective, by making use of the asymmetric Laplace (AL) density

$$f(y;\mu,\theta,\tau) = \frac{\tau(1-\tau)}{\theta}\exp\left[-\rho_\tau\left(\frac{y-\mu}{\theta}\right)\right].$$

The connection between the check function and the asymmetric Laplace density enables us to formulate smoothing parameter selection in (2) as a mixed model problem, in which the coefficient vector γ arises from a multivariate normal distribution, while the distribution of the outcomes conditional on γ is asymmetric Laplace.

4.2 Formulating an "equivalent" mixed model

Ordinarily the null space of P, i.e., the space of coefficient vectors γ that are unpenalized, has dimension $d > 0$. Let Q_1, Q_2 be matrices of dimension $K \times d$ and $K \times (K-d)$, respectively, whose columns form orthonormal bases of this null space

and its orthogonal complement, respectively. Referring to (2),we can write

$$\begin{aligned} B\gamma &= B(Q_1\ Q_2)\begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix}\gamma \\ &= X\beta + Zu, \end{aligned} \tag{11}$$

where $B = (b_1 \ldots b_n)^T$ as above, $X = (x_1 \ldots x_n)^T = BQ_1$, $\beta = Q_1^T\gamma$, $Z = (z_1 \ldots z_n)^T = BQ_2$, and $u = Q_2^T\gamma$. (We use x_i here to echo standard mixed model notation; it should not be confused with the predictor values x_i.) It is then readily shown that

$$\gamma^T P\gamma = u^T Q_2^T P Q_2 u. \tag{12}$$

Consider the mixed model

$$\begin{aligned} y_i|u &\sim AL(x_i^T\beta + z_i^T u, \theta, \tau)\ (i = 1, \ldots, n); \\ u &\sim N\left[\mathbf{0}, (\theta/2\lambda)Q_2^T P^+ Q_2\right], \end{aligned} \tag{13}$$

where P^+ is a generalized inverse of P. The likelihood $L(\beta, \theta, \lambda)$ is then the integral with respect to u of the joint density

$$\begin{aligned} \prod_{i=1}^n f(y_i|u)f(u) &= \left[\frac{\tau(1-\tau)}{\theta}\right]^n \exp\left[-\sum_{i=1}^n \rho_\tau\left(\frac{y_i - x_i^T\beta - z_i^T u}{\theta}\right)\right] \times \\ &\quad \frac{\exp\left[-(\lambda/\theta)u^T Q_2^T P Q_2 u\right]}{(2\pi)^{(K-d)/2}|(\theta/2\lambda)Q_2^T P^+ Q_2|^{1/2}} \\ &= \left[\frac{\tau(1-\tau)}{\theta}\right]^n \frac{|(2\lambda/\theta)Q_2^T P Q_2|^{1/2}}{(2\pi)^{(K-d)/2}} \times \\ &\quad \exp\left[-\frac{1}{\theta}\left\{\sum_{i=1}^n \rho_\tau(y_i - b_i^T\gamma) + \lambda\gamma^T P\gamma\right\}\right], \end{aligned} \tag{14}$$

where the previous line used (11) and (12). The expression in curly brackets above is precisely the penalized sum of check-function loss criterion minimized in nonparametric quantile regression. As in Ruppert et al.'s (2003) presentation of likelihood-based smoothness selection, this correspondence motivates choosing λ by maximizing $L(\beta, \theta, \lambda)$.

4.3 Algorithm

Our algorithm for estimating λ by maximum likelihood, or more correctly maximum simulated likelihood, proceeds by "profiling out" (i.e., optimizing over) first

β and then θ, in an approximate sense. In the standard linear mixed model setting with y_i normal conditional on u, one can optimize with respect to β by noting that, for given variance parameter values, the likelihood is maximized by estimating β by generalized least squares. In our setting no such closed-form expression for $\hat{\beta}_{\theta,\lambda} = \arg\max_\beta L(\beta,\theta,\lambda)$ is available. However, it seems reasonable to assume that $\hat{\beta}_{\theta,\lambda}$ is well approximated by the (parametric) quantile regression estimate $\tilde{\beta}_\tau = \arg\min_\beta \sum_{i=1}^n \rho_\tau(y_i - x_i^T\beta)$. Thus, referring to joint density (14), we obtain the approximate profile likelihood

$$\tilde{L}_P(\theta,\lambda) = \int \left[\frac{\tau(1-\tau)}{\theta}\right]^n \exp\left[-\sum_{i=1}^n \rho_\tau\left(\frac{y_i - x_i^T\tilde{\beta}_\tau - z_i^T u}{\theta}\right)\right] \times \frac{\exp\left[-(\lambda/\theta)u^T Q_2^T P Q_2 u\right]}{(2\pi)^{(K-d)/2}|(\theta/2\lambda)Q_2^T P^+ Q_2|^{1/2}} du.$$

The above integral is intractable, but a Monte Carlo approximation can be obtained by sampling $u_1,\ldots,u_N$ from multivariate normal density (13) and calculating

$$\hat{\tilde{L}}_P(\theta,\lambda) = \left[\frac{\tau(1-\tau)}{\theta}\right]^n \frac{1}{N}\sum_{j=1}^N \exp\left[-\sum_{i=1}^n \rho_\tau\left(\frac{y_i - x_i^T\tilde{\beta}_\tau - z_i^T u_j}{\theta}\right)\right]. \qquad (15)$$

We cannot sample directly from distribution (13), since it depends on the parameters over which we wish to maximize. Instead, we base approximate maximum likelihood estimation of λ on the following algorithm, which samples from a distribution that does not depend on (θ,λ):

1. Obtain $\tilde{\beta}_\tau$ by parametric quantile regression with design matrix X. The R package `quantreg` (Koenker, 2011) can be used for this step.
2. For suitably large N, sample $u_1^*,\ldots,u_N^* \sim N(\mathbf{0}, Q_2^T P^+ Q_2/2)$.
3. For each candidate λ, let

$$m_\lambda(\theta) = \frac{1}{N\theta^n}\sum_{j=1}^N \exp\left[-\sum_{i=1}^n \rho_\tau\left(\frac{y_i - x_i^T\tilde{\beta}_\tau - \sqrt{\theta/\lambda}z_i^T u_j^*}{\theta}\right)\right]. \qquad (16)$$

   If we define $u_j = \sqrt{\theta/\lambda}u_j^*$ $(j=1,\ldots,N)$, then—ignoring the constant $[\tau(1-\tau)]^n$—$m_\lambda(\theta)$ equals $\hat{\tilde{L}}_P(\theta,\lambda)$, with the u_j's having the distribution (13) required for Monte Carlo approximation (15). We can thus apply a numerical optimization procedure to m_λ to find $\hat{\theta}_\lambda = \arg\max_\theta \hat{\tilde{L}}_P(\theta,\lambda)$.
4. Choose the candidate λ for which $\hat{\tilde{L}}_P(\hat{\theta}_\lambda,\lambda)$ is maximized.

In practice we use a modified version of step 3; see Appendix C.

5 Simulation study

Our simulation study consisted of 100 replications. In each, we sampled $x_1, \ldots, x_{200}$ independently from the $U(0,1)$ distribution, and then generated outcomes

$$y_i = f(x_i) + \varepsilon_i,\ i = 1, ..., 200,$$

where, as in Yuan (2006), we took $f(x) = \sin(2\pi x)$ and considered errors ε_i generated independently from the following five distributions:

1. the double exponential distribution, whose density function is $\frac{1}{2}\exp(-|\varepsilon|),\ \varepsilon \in (-\infty, \infty)$;
2. the standard normal distribution;
3. the t-distribution with 3 df;
4. the mixture $0.05N(0,25) + 0.95N(0,1)$; and
5. the so-called slash distribution $N(0,1)/U(0,1)$.

We then estimated quantile curves for $\tau = .01, .05, .2, .5$ by minimizing the penalized least squares (2), using 30 cubic B-spline functions with equally-spaced knots. (By symmetry, the results for each of these values of τ should be similar to what we would obtain for $1-\tau$.) We obtained the values of λ that minimized the GACV (9), SIC (4), and 5-fold CV (10) criteria, and that which maximized the approximate likelihood as in Section 4. (In a separate set of simulations [not shown], 10-fold CV performed virtually identically to 5-fold CV.) As in Li et al. (2007), performance of each criterion was evaluated by calculating

$$\text{prediction error} = \frac{1}{10000}\sum_{i=1}^{10000} \rho_\tau[y_i^* - \hat{g}_\lambda(x_i^*)], \tag{17}$$

where $(x_1^*, y_1^*), \ldots, (x_{10000}^*, y_{10000}^*)$ were independently generated from the same joint distribution as the (x_i, y_i)'s, and

$$\text{mean absolute deviation} = \frac{1}{200}\sum_{i=1}^{200} |g(x_i) - \hat{g}_\lambda(x_i)|, \tag{18}$$

where $g(x)$ and $\hat{g}_\lambda(x)$ are the true and estimated quantile functions, respectively.

The code for the simulations, written in R (R Development Core Team, 2010) and available from the authors, optimized each criterion over 30 equally spaced values of $\log\lambda$ from -32 to 0. Using a PC with an Intel Core 2 Duo 2.53 GHz processor with 3.45GB of RAM, optimizing GACV and SIC required less than

Table 1: Mean (SD), over 100 simulations, of 1000 times the prediction error (17), for the distributions listed on p. 10 (DE=double exponential).

	GACV	SIC	5-fold CV	Likelihood
$\tau = 0.01$				
DE	152 (28)	139 (30)	56 (8)	65 (13)
Normal	98 (20)	74 (26)	32 (3)	37 (7)
t_3	190 (29)	177 (30)	81 (9)	96 (17)
Mixture	488 (105)	362 (130)	147 (16)	170 (36)
Slash	3510 (1585)	3498 (1587)	2838 (354)	2972 (518)
$\tau = 0.05$				
DE	232 (23)	220 (26)	171 (8)	173 (8)
Normal	153 (16)	133 (22)	112 (6)	111 (4)
t_3	280 (27)	268 (32)	205 (8)	208 (8)
Mixture	768 (91)	627 (121)	540 (21)	545 (22)
Slash	4701 (3129)	4534 (2796)	3321 (137)	3304 (99)
$\tau = 0.2$				
DE	427 (28)	391 (9)	391 (14)	388 (8)
Normal	317 (21)	290 (6)	292 (8)	289 (5)
t_3	464 (37)	434 (8)	432 (9)	429 (7)
Mixture	1571 (140)	1432 (15)	1442 (21)	1439 (19)
Slash	3760 (67)	3745 (9)	3754 (35)	3746 (10)
$\tau = 0.5$				
DE	507 (13)	513 (13)	506 (7)	504 (4)
Normal	419 (22)	412 (10)	409 (8)	407 (6)
t_3	568 (21)	572 (13)	565 (8)	562 (5)
Mixture	2086 (104)	2043 (16)	2061 (43)	2048 (20)
Slash	3898 (7)	3898 (5)	3895 (14)	3897 (7)

5 seconds per simulation; 5-fold CV, about 14 seconds; and the likelihood method, about 2.5 minutes.

The results are given in Tables 1 and 2. Overall, the four methods perform similarly for $\tau = 0.5$. For $\tau = 0.2$, GACV performs less well than the other methods. For $\tau = 0.01, 0.05$, 5-fold CV and the likelihood method greatly outperform GACV and SIC, with a slight edge for 5-fold CV over likelihood in most cases.

Figure 2 offers further insight by presenting boxplots of the degrees of freedom (5) of the models fitted in the simulations for the double exponential distribution. Since we used basis dimension $K = 30$, the maximum df, implying no roughness penalization, is 30; df values anywhere near this value signal overfitting. Thus

Table 2: Mean (SD), over 100 simulations, of 100 times the estimation error given by the absolute mean deviation (18).

	GACV	SIC	5-fold CV	Likelihood
$\tau = 0.01$				
DE	196 (51)	189 (51)	109 (49)	120 (45)
Normal	92 (23)	79 (27)	45 (16)	49 (19)
t_3	241 (85)	235 (86)	170 (99)	195 (93)
Mixture	462 (108)	401 (125)	190 (99)	231 (96)
Slash	5156 (6538)	5145 (6541)	4912 (4956)	5954 (7049)
$\tau = 0.05$				
DE	108 (18)	102 (22)	56 (19)	58 (20)
Normal	59 (10)	46 (17)	31 (11)	30 (10)
t_3	135 (33)	127 (38)	61 (25)	71 (25)
Mixture	301 (48)	205 (102)	128 (60)	139 (57)
Slash	2921 (5362)	2589 (4513)	498 (404)	464 (341)
$\tau = 0.2$				
DE	53 (16)	33 (12)	32 (12)	30 (10)
Normal	37 (13)	21 (8)	22 (8)	20 (7)
t_3	48 (21)	34 (11)	30 (10)	27 (10)
Mixture	170 (76)	83 (36)	94 (36)	92 (36)
Slash	64 (44)	54 (15)	63 (29)	55 (15)
$\tau = 0.5$				
DE	17 (7)	22 (10)	17 (6)	15 (5)
Normal	22 (12)	20 (7)	18 (7)	16 (6)
t_3	23 (10)	28 (10)	22 (8)	20 (7)
Mixture	95 (54)	74 (30)	87 (37)	79 (31)
Slash	44 (8)	45 (7)	39 (14)	43 (9)

the boxplots for GACV and SIC suggest that these methods' poor performance for extreme quantiles—and even for $\tau = 0.2$, in the case of GACV—result from such overfitting. The likelihood method appears to be much more stable than the other three methods in terms of df, both within and among quantiles. Note that in our simulations, the true quantile curves are parallel to each other, so the fact that the four quantiles' df distributions are most alike for the likelihood method constitutes evidence of that method's efficacy.

These observations are illustrated in Figure 3, which shows fitted curves for the four quantiles, by each of the methods, for a typical replication. (This replication is "typical" in the sense that the df, averaged over the four quantiles and the four

methods, ranks 50th of 100.) The GACV and SIC fits for $\tau = 0.01, 0.05$, and to a less extent the GACV fit for $\tau = 0.2$, are implausibly bumpy. The marked difference in smoothness between the 5-fold CV fits for $\tau = 0.01$ and $\tau = 0.05$ mirror the contrasting df distributions for these two quantiles, as displayed in the lower left subfigure of Figure 2. Supplementary Appendix A, available at http://works.bepress.com/-phil_reiss/20/, provides analogues of Figures 2 and 3 for the other four distributions.

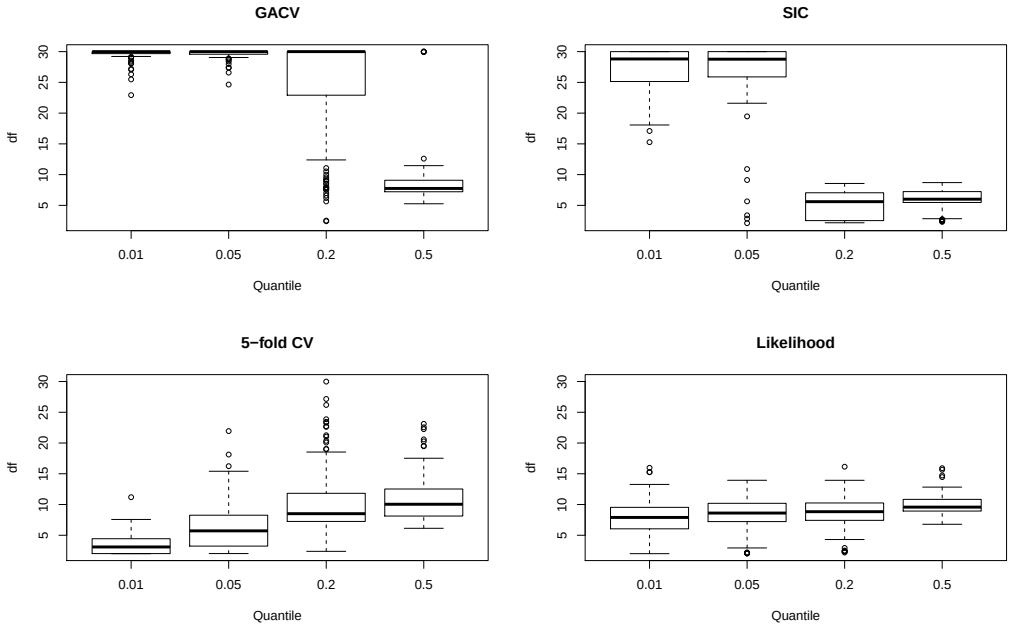

Figure 2: Degrees of freedom (5) in the 100 simulations for the double exponential distribution.

The above results assume a sample size of 200. Supplementary Appendix B, available at the above URL, reports the results of further simulations with larger sample sizes, 400 and 1000. As expected, the performance of GACV and SIC improves with larger samples, but 5-fold cross-validation and the likelihood method maintain a clear advantage, especially for $\tau = 0.01$ and $\tau = 0.05$.

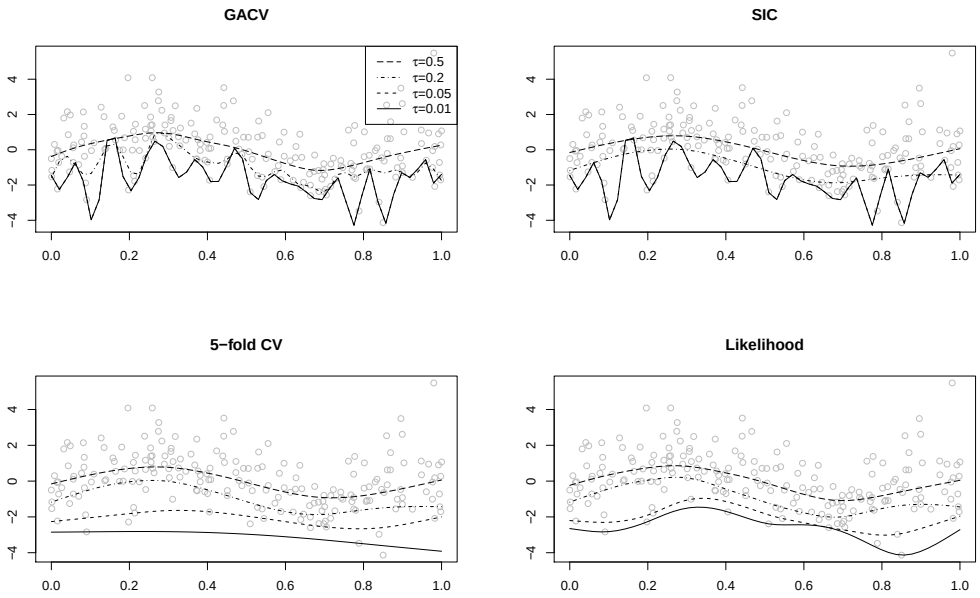

Figure 3: Four fitted quantile functions for each of the four methods, for a typical simulation with double exponential errors.

6 Application to the functional connectivity data

As discussed in the introduction, our motivating application was to investigate whether the distribution of functional connectivity conditional on age, for each of a large set of brain connections, differs between controls and individuals with ADHD. We explored this question using resting-state scans from 128 normal control participants, age 7–25, and 46 participants in the same age range with ADHD, all acquired at New York University. The 39 regions of interest we studied were identified by Dosenbach et al. (2007) as relevant to task control; as such it is reasonable to ask whether connections among these regions may tend to develop anomalously in ADHD. Dosenbach et al. (2007) found that the correlation matrix of resting state fMRI signals in these regions could be partitioned into a set of distinct networks. In particular, their analysis assigned 11 of the 39 ROIs to a "frontoparietal" network associated with active control, and 7 ROIs to a "cinguloopercular" network associated with stable maintenance; another 4 ROIs were found in the cerebellum, while the remaining ROIs were dispersed among five small clusters. For each scan we computed the mean BOLD time series for voxels in each of the ROIs, from which we obtained the 39×39 matrix of temporal correlations for each connection (pair

of ROIs). In what follows, "connectivity" refers to the Fisher (1921) z-transformed values of these correlations. Of the $\binom{39}{2} = 741$ connections for which connectivity growth charts could potentially be created, many show no significant change with age, and growth charts for such connections are unlikely to be useful for detecting abnormal development. We therefore prescreened for connections that change with age, as follows (cf. Church et al., 2009). For each of the 741 connections, we fit a penalized spline (mean) regression of connectivity on age using the R package `mgcv`, and tested the effect of age using the F test of Wood (2006); we retained the 100 connections found most significant by this test (approximate $p < .026$).

For $k = 1, \ldots, 100$, let $F_k(\cdot|x), G_k(\cdot|x)$ denote the cumulative distribution function of connectivity, conditional on age x, for the kth retained connection, in controls and ADHD individuals respectively. The global null hypothesis that we seek to test is

$$F_k(\cdot|x) = G_k(\cdot|x) \text{ for all ages } x, \text{ for } k = 1, \ldots, 100. \tag{19}$$

The maturational delay theory of ADHD (Rubia, 2007) suggests focusing on a particular type of departure from this null hypothesis: namely, that for some connections, many individuals with ADHD will have connectivities in either the left or the right tail of the control distribution for their age. We designed the following procedure to detect departures of this type. For $k = 1, \ldots, 100$,

1. apply nonparametric quantile regression to the 128 controls to construct 10th- and 90th-percentile growth charts, i.e., estimates of $F_k^{-1}(0.1|\cdot)$ and $F_k^{-1}(0.9|\cdot)$, for the kth connection;
2. determine l_k, the number of ADHD participants (out of 46) who fell below the estimated 10th percentile for their age, and u_k, the number above the 90th percentile;
3. calculate the test statistic

$$t_k = \max\{l_k, u_k\}. \tag{20}$$

Connections with $t_k \geq 12$ (see Appendix D for an explanation of how this threshold was chosen) were identified as those for which ADHD may be associated with abnormal development. The 10th and 90th percentiles were chosen heuristically. We could have chosen more extreme quantiles, and correspondingly a lower threshold than $t_k = 12$; but this would likely have reduced stability due to greater variability of the percentile curves.

Of the 100 connections tested, 8 were detected (i.e., met the $t_k \geq 12$ threshold) based on percentile curves with smoothing parameter selected by either likelihood or 5-fold CV (see the upper panels of Figure 4). In addition, 4 connections

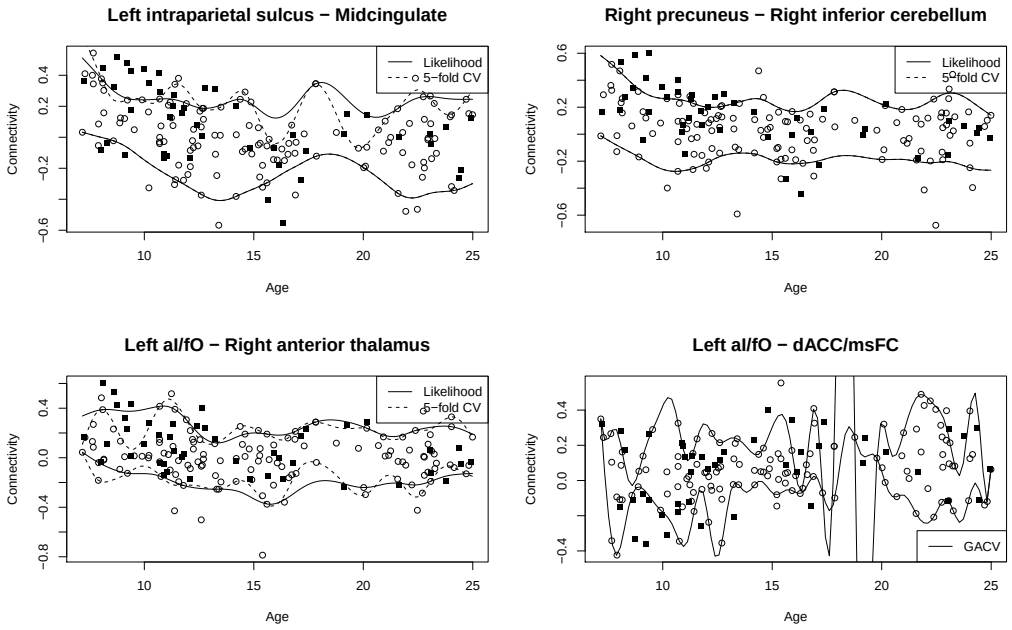

Figure 4: Estimated 10th- and 90th-percentile curves for the 128 controls, shown as light circles, with the 46 ADHD participants displayed as dark squares. The heading for each plot indicates the pair of ROIs whose connectivities are plotted (aI/fO = anterior insula/frontal operculum; dACC/msFC = dorsal anterior cingulate cortex/medial superior frontal cortex). Upper panels: two connections with $t_k \geq 12$, according to either likelihood or 5-fold CV; in both cases, many of the young ADHD participants have unusually high connectivity. Lower left: a connection identified by 5-fold CV, but not by the likelihood method, as having $t_k \geq 12$—evidently due to undersmoothing by 5-fold CV. Lower right: a connection spuriously identified by GACV (i.e., high t_k derived from the GACV-based curves), due to undersmoothed percentile curves.

were detected by 5-fold CV but not by the likelihood method, and the reverse was true for 1 connection. For each of these 5 connections, inspection of the fits revealed the reason for the disparate results: 5-fold CV chose very small λ for the 10th and/or the 90th percentile curve, leading to undersmoothing (see the lower left panel of Figure 4 for an example). In contrast to the likelihood and 5-fold CV methods, smoothness selection by GACV led to t_k values of very dubious utility, due to extreme undersmoothing; see the lower right panel of Figure 4 for an example. The 9 connections that were detected with likelihood-based smoothness selection are depicted in Figure 5. To some extent, this set of connections respects Dosen-

bach et al.'s (2007) assignment of the 39 ROIs to distinct networks. Eight of the 9 connections are within the frontoparietal network or between it and the cerebellum, or else within the cinguloopercular network or between it and the cerebellum. Further investigation and validation of these findings awaits future analyses with larger samples. See Fair et al. (2010) for a previous investigation of resting state network disparities between ADHD individuals and controls, using very different methodology and a different set of ROIs.

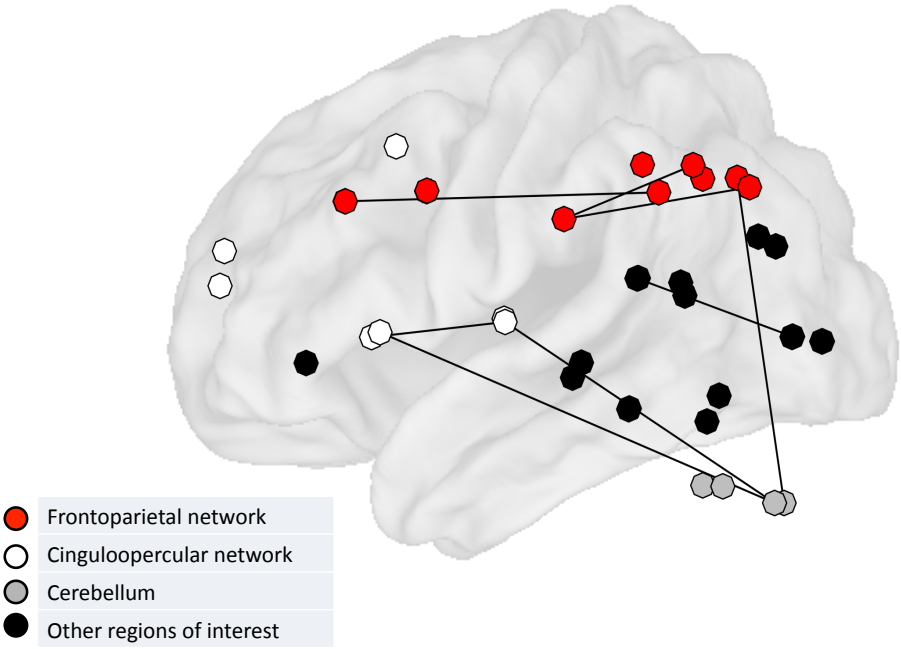

Figure 5: An approximate two-dimensional rendering of the 39 ROIs of Dosenbach et al. (2007). Line segments indicate connections for which our method, with likelihood-based smoothness selection, detected evidence of anomalous development in ADHD. (There are 9 such connections, but one is not shown because it involves two ROIs that are superimposed in this rendering.)

7 Discussion

We have adduced evidence, from simulations and a real-data example, that GACV and SIC, the two standard methods for automatic smoothness selection in nonparametric quantile regression, perform suboptimally for estimation of extreme quantiles. Consequently—in particular, for applications such as ours, which necessitate fitting quantile curves for a large number of response variables—we recommend either multifold CV or a likelihood approach. Each of these alternatives has its relative strengths. On the one hand, multifold CV is much faster and simpler than the likelihood method, and had slightly lower prediction and estimation error overall in the simulations (Tables 1 and 2). On the other hand, the distributions of df displayed in Figure 2 suggest that the likelihood method is the most stable, and this seems to be borne out by our analysis of the ADHD data. We recommend multifold CV for data analysts seeking a fast, straightforward, generally reliable approach to smoothness selection. We hope that future work will reduce the computational burden of the likelihood approach, and thereby make it more attractive as a practical option for routine use.

Li et al. (2007) propose a kernel approach to nonparametric quantile regression, which employs a quadratic penalty as in Nychka et al. (1995), Yuan (2006), and the present work, but solves the optimization by a new algorithm that finds the entire solution path (i.e., the solution for all smoothing parameter values) rather than by PIRLS. Their simulations, like ours, found that GACV and SIC (which are defined somewhat differently in their framework) performed better for the median than for extreme quantiles. It would be interesting to investigate whether our likelihood-based smoothness selection can be adapted to the Li et al. (2007) algorithm.

While a great deal of recent work (e.g., Krivobokova and Kauermann, 2007, Welham et al., 2007, Wood, 2011, and references therein) has advocated likelihood-based smoothness selection derived from mixed-model formulations of smoothing problems, to the best of our knowledge all such work has focused on linear or other exponential family models. The present work extends likelihood-based smoothing parameter selection into the novel domain of nonparametric quantile regression.

Our likelihood approach to smoothness selection can be viewed as empirical Bayes estimation of λ. A fully Bayesian method for nonparametric quantile regression might achieve similar results with greater computational efficiency than the algorithm of Section 4.3. After completing this paper we became aware of a non-spline-based approach of this type, due to Yue and Rue (2011), that uses integrated nested Laplace approximations with Gaussian Markov random field priors. Yue and Rue (2011) note that their method encounters some difficulties with extreme quantiles, on which the present paper has focused. Detailed comparisons of

empirical Bayes and fully Bayesian approaches await further work.

Appendix A Penalized iteratively reweighted least squares algorithm

We describe here a modified version of the algorithm of Nychka et al. (1995) to solve minimization problem (2) by PIRLS (see Wood, 2006). Given the kth-iteration estimate $\hat{\gamma}^{(k)}$, the updated estimate is

$$\hat{\gamma}^{(k+1)} = \arg\min_{\gamma} \left[\sum_{i=1}^{n} w_i^{(k)} (y_i - b_i^T \gamma)^2 + \lambda \gamma^T P \gamma \right],$$

with weights $w_1^{(k)}, \ldots, w_n^{(k)}$ chosen so that the estimating equation for this minimization, namely $\sum_{i=1}^{n} 2 w_i^{(k)} (y_i - b_i^T \gamma)(-b_i) + 2\lambda P\gamma = \mathbf{0}$, is approximately equivalent to the estimating equation for minimization (2). Supposing for the moment that all residuals are nonzero, the latter estimating equation is $\sum_{i=1}^{n} [\tau - I(y_i - b_i^T \gamma < 0)](-b_i) + 2\lambda P\gamma = \mathbf{0}$, and hence, for γ in the vicinity of $\hat{\gamma}^{(k)}$, the left sides of the last two equations can be approximately equated by setting

$$w_i^{(k)} = \frac{\tau - I\left[y_i - b_i^T \hat{\gamma}^{(k)} < 0\right]}{2\left[y_i - b_i^T \hat{\gamma}^{(k)}\right]} \tag{21}$$

for $i = 1, \ldots, n$.

In general, some estimated residuals may equal 0 (see Li et al., 2007), the only point at which ρ_τ is not differentiable. In Nychka et al. (1995), this problem is addressed by replacing ρ_τ with an approximating function that is differentiable everywhere. In our implementation, we set a large upper bound for the weights and truncate when the residual is very small and thus (21) is very large (Gentle, 2007, p. 233). Note that very small residuals imply negligible contributions to the sum in (2), and thus the effect of truncating the weights is to replace a negligible portion of that sum with an even smaller quantity.

The above heuristic argument suggests that the PIRLS iterates should converge to the spline coefficient vector γ solving (2). Nychka et al. (1995) suggest running the algorithm for each candidate λ in descending order, using the fit for each λ as the initial estimate for the next smaller λ.

Appendix B Derivation of ACV

By invoking two approximations, we arrive at a somewhat more streamlined variation on Yuan's (2006) derivation of (8) as approximately equal to (7).

The first approximation is based on the observation that

$$\rho_\tau[y_i - \hat{g}_\lambda^{[-i]}(x_i)] = \frac{y_i - \hat{g}_\lambda^{[-i]}(x_i)}{y_i - \hat{g}_\lambda(x_i)} \rho_\tau[y_i - \hat{g}_\lambda(x_i)],$$

provided that the leave-one-out residual $y_i - \hat{g}_\lambda^{[-i]}(x_i)$ and the full-data residual $y_i - \hat{g}_\lambda(x_i)$ are of the same sign. Since the signs should usually be the same for most i, and moreover the exceptions should tend to be i such that both residuals are relatively small, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} \rho_\tau[y_i - \hat{g}_\lambda^{[-i]}(x_i)] \approx \frac{1}{n}\sum_{i=1}^{n} \frac{y_i - \hat{g}_\lambda^{[-i]}(x_i)}{y_i - \hat{g}_\lambda(x_i)} \rho_\tau[y_i - \hat{g}_\lambda(x_i)]. \tag{22}$$

The second approximation is obtained by viewing $\hat{g}_\lambda(x_i)$ as a function of the responses, and considering two expressions for its partial derivative with respect to y_i. On the one hand, although $\partial \hat{g}_\lambda(x_i)/\partial y_i$ may not exist for all i, at least heuristically we can equate it with h_{ii} (as in Yuan, 2006). On the other hand, the leave-one-out lemma for quantile smoothing splines (Yuan, 2006, Lemma 3.1) says that, if the ith response y_i is replaced by $\hat{g}_\lambda^{[-i]}(x_i)$ and the model is refitted to this modified data set with smoothing parameter λ, then the new function estimate is precisely $\hat{g}_\lambda^{[-i]}$. Thus, we can approximate $\partial \hat{g}_\lambda(x_i)/\partial y_i$ by the slope of the secant from $(y_i, \hat{g}_\lambda(x_i))$ to $(\hat{g}_\lambda^{[-i]}(x_i), \hat{g}_\lambda^{[-i]}(x_i))$. Putting these together, we have

$$h_{ii} \approx \frac{\hat{g}_\lambda(x_i) - \hat{g}_\lambda^{[-i]}(x_i)}{y_i - \hat{g}_\lambda^{[-i]}(x_i)}.$$

Using this expression, (22) becomes

$$\frac{1}{n}\sum_{i=1}^{n} \rho_\tau[y_i - \hat{g}_\lambda^{[-i]}(x_i)] \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\rho_\tau[y_i - \hat{g}_\lambda(x_i)]}{1 - h_{ii}},$$

i.e., (7) is approximated by (8), as was to be shown.

Appendix C Maximizing with respect to θ

Step 3 of the algorithm given in Section 4.3 consists of defining a function $m_\lambda(\theta)$ and maximizing it with respect to θ. We have found that the procedure of Brent

(1973) implemented in the R function `optimize` works well for this purpose, provided the search interval $(\theta_{min}, \theta_{max}]$ is chosen carefully. We take $\theta_{min} = 0$ and choose θ_{max} as follows.

Let λ_0, θ_0 be arbitrary positive values. For any $\lambda > 0$ and any $\theta > m_{\lambda_0}(\theta_0)^{-1/n}$, we have

$$m_\lambda(\theta) \leq \theta^{-n} < m_{\lambda_0}(\theta_0), \tag{23}$$

where the first inequality follows from (16). Since our goal is to maximize the approximate likelihood over both λ and θ, (23) implies that when maximizing $m_\lambda(\theta)$ for each candidate λ, it suffices to consider θ between 0 and

$$m_{\lambda_0}(\theta_0)^{-1/n} = \theta_0 \left[\frac{1}{N} \sum_{j=1}^{N} \exp \left\{ -\frac{1}{\theta_0} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^T \tilde{\beta}_\tau - \sqrt{\theta_0/\lambda_0} z_i^T u_j^* \right) \right\} \right]^{-1/n}. \tag{24}$$

In particular, taking the limit of (24) as $\lambda_0 \to \infty$ and then minimizing with respect to θ_0 yields the value $m_\infty(\hat{\theta}_\infty)^{-1/n} = \frac{e}{n} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \tilde{\beta}_\tau)$; we take this as our initial θ_{max}. But since (23) and (24) are valid for any λ_0, θ_0, if at any point we find a (λ_0, θ_0) such that $m_{\lambda_0}(\theta_0) > \theta_{max}^{-n}$, we can restrict the search interval further by updating θ_{max} to the smaller value $m_{\lambda_0}(\theta_0)^{-1/n}$. This suggests a modified step 3 consisting of the following substeps:

(i) Set $\theta_{max} = \frac{e}{n} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \tilde{\beta}_\tau)$ and set λ to the largest candidate value.
(ii) Find $\tilde{\theta}_\lambda = \arg\max_{\theta \in (0, \theta_{max}]} m_\lambda(\theta)$.
(iii) Update θ_{max} to $\min\{\theta_{max}, m_\lambda(\tilde{\theta}_\lambda)^{-1/n}\}$.
(iv) If λ is the smallest of the candidate values, stop; otherwise set λ to the next largest candidate and return to substep (ii).

We would expect $\hat{\theta}_\lambda = \arg\max_{\theta > 0} m_\lambda(\theta)$ to tend to decrease as λ decreases. Consequently, considering the candidate λs in descending order should increase the probability that $\tilde{\theta}_\lambda$ found in substep (ii) is equal to $\hat{\theta}_\lambda$.

Appendix D Choice of threshold

We explain here why, in Section 6, connections for which $t_k \geq 12$ [see (20)] were taken to exhibit "significant" departure from the null hypothesis (19). Let $x_1, \ldots, x_{46}$ and be the ages of the 46 ADHD participants, and let $y_{k,1}, \ldots, y_{k,46}$ denote their connectivities for the kth of the 100 connections considered. Under the null hypothesis,

the values $z_{ki} \equiv F_k^{-1}(y_{ki}|x_i)$ are independent $U(0,1)$ variables. Ignoring error in estimating the quantile curves, we have

$$t_k \geq m \text{ if and only if either } z_{(m)} < 0.1 \text{ or } z_{(47-m)} > 0.9,$$

where $z_{(1)}, \ldots, z_{(46)}$ are the order statistics of the z_i's. Hence $Pr(t_k \geq m) = Pr\{u_{(m)} < 0.1 \text{ or } u_{(47-m)} > 0.9\}$ where $u_{(1)}, \ldots, u_{(46)}$ are the order statistics of 46 independent $U(0,1)$ variables. We estimated the latter probability for a range of values of m by simulation, and thereby obtained a table of p-values for t_k, from which the false discovery rate (FDR) can be estimated for the observed $t_1, \ldots, t_{100}$ by the step-up procedure of Benjamini and Hochberg (1995). The threshold $t_k = 12$ is the smallest value such that, with percentile curves estimated by either the likelihood method or 5-fold CV, we obtained FDR$< .05$. While our simulated p-values are asymptotically valid under assumptions guaranteeing the consistency of penalized quantile regression splines (Pratesi et al., 2009), we acknowledge that we have not studied their small-sample properties. Nevertheless, we consider our choice of threshold to be adequately justified given the exploratory nature of this analysis.

References

Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, Akademiai Kiado, 267–281.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B*, 57, 289–300.

Biswal, B., F. Z. Yetkin, V. M. Haughton, and J. S. Hyde (1995): "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI," *Magnetic Resonance in Medicine*, 34, 537–541.

Bosch, R. J., Y. Ye, and G. G. Woodworth (1995): "A convergent algorithm for quantile regression with smoothing splines," *Computational statistics & data analysis*, 19, 613–630.

Brent, R. P. (1973): *Algorithms for Minimization Without Derivatives*, Englewood Cliffs, NJ: Prentice-Hall.

Church, J. A., D. A. Fair, N. U. F. Dosenbach, A. L. Cohen, F. M. Miezin, S. E. Petersen, and B. L. Schlaggar (2009): "Control networks in paediatric Tourette syndrome show immature and anomalous patterns of functional connectivity," *Brain*, 132, 225.

Cole, T. J. and P. J. Green (1992): "Smoothing reference centile curves: the LMS method and penalized likelihood," *Statistics in Medicine*, 11, 1305–1319.

Cox, D. D. (1983): "Asymptotics for M-type smoothing splines," *The Annals of Statistics*, 11, 530–551.

Cox, D. R. (1988): "Discussion of T. J. Cole, 'Fitting smoothed centile curves to reference data'," *Journal of the Royal Statistical Society: Series A*, 151, 411.

Craven, P. and G. Wahba (1979): "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerische Mathematik*, 31, 317–403.

De Boor, C. (2001): *A Practical Guide to Splines*, New York: Springer-Verlag, revised edition.

Debruyne, M., M. Hubert, and J. A. K. Suykens (2008): "Model selection in kernel based regression using the influence function," *Journal of Machine Learning Research*, 9, 2377–2400.

Dosenbach, N. U. F., D. A. Fair, F. M. Miezin, A. L. Cohen, K. K. Wenger, R. A. T. Dosenbach, M. D. Fox, A. Z. Snyder, J. L. Vincent, M. E. Raichle, B. L. Schlaggar, and S. E. Petersen (2007): "Distinct brain networks for adaptive and stable task control in humans," *Proceedings of the National Academy of Sciences*, 104, 11073.

Fair, D. A., A. L. Cohen, N. U. F. Dosenbach, J. A. Church, F. M. Miezin, D. M. Barch, M. E. Raichle, S. E. Petersen, and B. L. Schlaggar (2008): "The maturing architecture of the brain's default network," *Proceedings of the National Academy of Sciences*, 105, 4028.

Fair, D. A., J. Posner, B. J. Nagel, D. Bathula, T. G. C. Dias, K. L. Mills, M. S. Blythe, A. Giwa, C. F. Schmitt, and J. T. Nigg (2010): "Atypical default network connectivity in youth with attention-deficit/hyperactivity disorder," *Biological Psychiatry*, 68, 1084–1091.

Fisher, R. A. (1921): "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron*, 1, 3–32.

Friston, K. J. (1994): "Functional and effective connectivity in neuroimaging: a synthesis," *Human Brain Mapping*, 2, 56–78.

Gentle, J. E. (2007): *Matrix Algebra: Theory, Computations, and Applications in Statistics*, New York: Springer.

Geraci, M. and M. Bottai (2007): "Quantile regression for longitudinal data using the asymmetric Laplace distribution," *Biostatistics*, 8, 140–154.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning*, New York: Springer-Verlag.

Koenker, R. (2011): *quantreg: Quantile Regression*, URL `http://CRAN.R-project.org/package=quantreg`, r package version 4.76.

Koenker, R. and G. Bassett (1978): "Regression quantiles," *Econometrica*, 46, 33–50.

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

Komunjer, I. (2005): "Quasi-maximum likelihood estimation for conditional quantiles," *Journal of Econometrics*, 128, 137–164.

Krivobokova, T. and G. Kauermann (2007): "A note on penalized spline smoothing with correlated errors," *Journal of the American Statistical Association*, 102, 1328–1337.

Li, Y., Y. Liu, and J. Zhu (2007): "Quantile regression in reproducing kernel Hilbert spaces," *Journal of the American Statistical Association*, 102, 255–268.

Ng, P. and M. Maechler (2007): "A fast and efficient implementation of qualitatively constrained quantile smoothing splines," *Statistical Modelling*, 7, 315.

Nychka, D., G. Gray, P. Haaland, D. Martin, and M. O'Connell (1995): "A nonparametric regression approach to syringe grading for quality improvement," *Journal of the American Statistical Association*, 90, 1171–1178.

Ogden, C. L., R. J. Kuczmarski, K. M. Flegal, Z. Mei, S. Guo, R. Wei, L. M. Grummer-Strawn, L. R. Curtin, A. F. Roche, and C. L. Johnson (2002): "Centers for Disease Control and Prevention 2000 growth charts for the United States: improvements to the 1977 National Center for Health Statistics version," *Pediatrics*, 109, 45–60.

Pratesi, M., M. G. Ranalli, and N. Salvati (2009): "Nonparametric M-quantile regression using penalised splines," *Journal of Nonparametric Statistics*, 21, 287–304.

Quetelet, A. (1830): "Sur la taille moyenne de l'homme dans les villes et dans les campagnes, et sur l'age où la croissance est complètement achevée," *Annales d'Hygiène Publique et le Médicine Légale*, 3, 24–26.

R Development Core Team (2010): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`, ISBN 3-900051-07-0.

Reich, B. J., H. D. Bondell, and H. J. Wang (2010): "Flexible Bayesian quantile regression for independent and clustered data," *Biostatistics*, 11, 337–352.

Reiss, P. T. and R. T. Ogden (2009): "Smoothing parameter selection for a class of semiparametric linear models," *Journal of the Royal Statistical Society: Series B*, 71, 505–523.

Rubia, K. (2007): "Neuro-anatomic evidence for the maturational delay hypothesis of ADHD," *Proceedings of the National Academy of Sciences*, 104, 19663–19664.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003): *Semiparametric Regression*, New York: Cambridge University Press.

Schwarz, G. (1978): "Estimating the dimension of a model," *The Annals of Statistics*, 461–464.

Stone, M. (1977): "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion," *Journal of the Royal Statistical Society: Series B*, 39, 44–47.

Wahba, G. (1985): "A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem," *The Annals of Statistics*, 1378–1402.

Wahba, G. (1999): "Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV," in *Advances in Kernel Methods*, Cambridge: MIT Press, 69–88.

Wei, Y. and X. He (2006): "Conditional growth charts," *The Annals of Statistics*, 34, 2069–2097.

Wei, Y., A. Pere, R. Koenker, and X. He (2006): "Quantile regression methods for reference growth charts," *Statistics in Medicine*, 25, 1369–1382.

Welham, S. J., B. R. Cullis, M. G. Kenward, and R. Thompson (2007): "A comparison of mixed model splines for curve fitting," *Australian & New Zealand Journal of Statistics*, 49, 1–23.

Wood, S. N. (2006): *Generalized Additive Models: An Introduction with R*, Boca Raton: Chapman & Hall/CRC.

Wood, S. N. (2011): "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," *Journal of the Royal Statistical Society: Series B*, 73, 3–36.

Yu, K. and R. A. Moyeed (2001): "Bayesian quantile regression," *Statistics & Probability Letters*, 54, 437–447.

Yuan, M. (2006): "GACV for quantile smoothing splines," *Computational Statistics & Data Analysis*, 50, 813–829.

Yue, Y. R. and H. Rue (2011): "Bayesian inference for additive mixed quantile regression models," *Computational Statistics & Data Analysis*, 55, 84–96.

Zhang, P. (1993): "Model selection via multifold cross validation," *The Annals of Statistics*, 21, 299–313.