Drift Removal for Time Series Data Using Quantile Trend Filtering

Halley Brantley* Joseph Guinness† and Eric C. Chi‡

Abstract

Abstract

Keywords: Key words

*Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203 (E-mail: hlbrantl@ncsu.edu)

†Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203 (E-mail: jsguinne@ncsu.edu)

‡Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203 (E-mail: eric_chi@ncsu.edu).

1 Introduction

1.1 Background

- Oh et al. (2011) unified framework for non-parametric quantile regression by approximating check loss function with quadratic loss near zero.
- Koenker et al. (1994) introduce quantile smoothing splines. We need to discuss how our approach is different.
- Kim et al. (2009) introduce the concept of ℓ_1-trend filtering.
- Tibshirani (2014) describes properties of trend filtering using quadratic loss. Shows that trend filtering estimates adapt to the local level of smoothness much better than smoothing splines, and exhibit a remarkable similarity to locally adaptive regression splines. Prove that (with the right choice of tuning parameter) the trend filtering estimate converges to the true underlying function at the minimax rate for functions whose kth derivative is of bounded variation.
- Ning et al. (2014) address problem of estimating a smooth baseline in noisy data with drift.
- Takeuchi et al. (2006) Nonparametric quantile regression using SVM with Gaussian RBF kernels and check (pinball) loss.
- Yuan (2006) Comparison of cross-validation methods for quantile smoothing splines.

We propose to use the trend filtering penalty with the check loss function to produce a non-parametric quantile regression estimate that can be computes using a linear time algorithm for removing trends in time series. The formulation was proposed by Kim et al. (2009) as a possible extension of ℓ_1-trend filtering but not studied. Moreover we extend the basic framework to model multiple quantiles and ensure non-crossing.

1.2 Application

Examples:

Figure 1: Raw Data - three collocated sensors

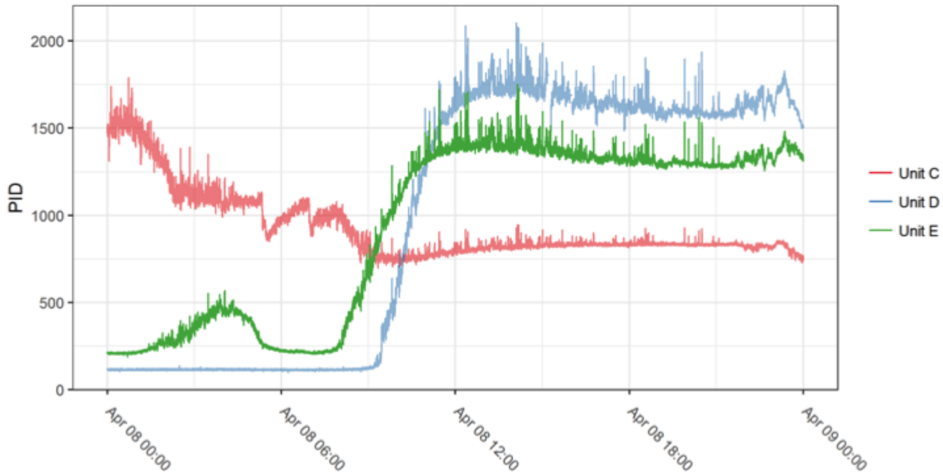
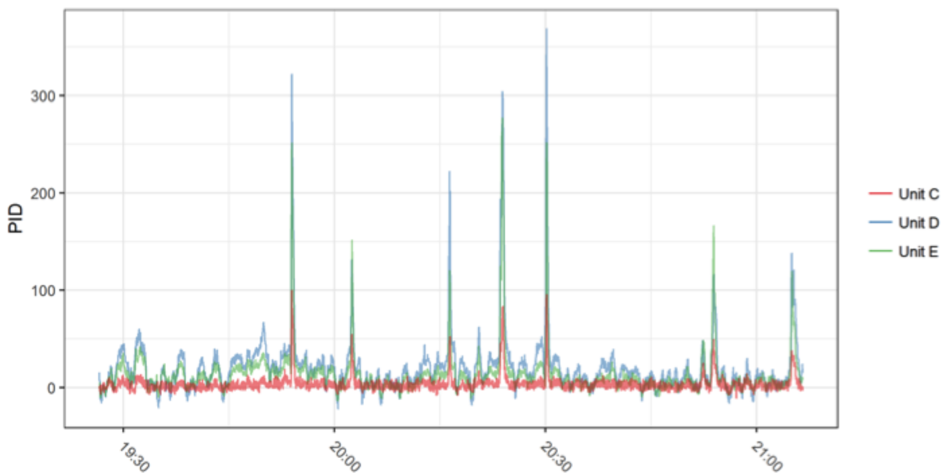


Figure 2: Detrended data - three collocated sensors



- Air quality (Apte et al., 2017)
- ECG (Sanyal et al., 2012)
- Chromatogram baseline estimation (Ning et al., 2014; Ilewicz et al., 2015)
- Galaxy spectrum baseline estimation (Ilewicz et al., 2015; Bacher et al., 2016)
- Identifying absorption dips from black body radiation

Need to decide between detrending versus detrending + denoising. The BEADS (Ning et al., 2014) method does both. We may wish to focus on detrending and then use wavelet

SURE denoising as a postprocessing step, i.e. do a two-stage procedure, both of which can be done in linear time. On paper this should be faster than the BEADS procedure. Or we may just want to stick with detrending. There's Matlab code for BEADS, and the BEADS paper also points to two other popular methods in chromotography.

Things to do:

- Convergence of the algorithm
- Convergence rate (He and Yuan, 2012, 2015; Davis, 2017)
- Timing experiments of LP versus Spingarn
- Make an R package - detrendr
- Compare on synthetic data quality of solution with existing methods
- Do comparisons on real data examples

2 Quantile Regression

The classic least squares regression is notoriously sensitive to outliers. One remedy to blunt the influence of outliers is to compute the least absolute deviations (LAD) solution in place of the least squares one. Given a design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and continuous responses $\mathbf{y} \in \mathbb{R}^n$, we estimate a regression vector $\boldsymbol{\theta} \in \mathbb{R}^p$ so that $\mathbf{X}\boldsymbol{\theta}$ is a good approximation of $\mathbf{y}$. The LAD estimator is a solution to the problem

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_1.$$

The above optimization problem generalizes the notion of the median of a collection of numbers. A median μ of n reals $y_1, \ldots, y_n$ is the minimizer of the function

$$f(u) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} |y_i - \theta|.$$

Recall that the median is the 50th percentile or 0.5-quantile, namely half of the y_i are less than or equal to μ and the other half is greater than or equal to μ. The median can be generalized to arbitrary τ-quantiles for $\tau \in (0,1)$ to give us quantile regression (Koenker and Bassett, 1978).

First define the so-called "check function"

$$\rho_\tau(\Delta) \quad = \quad \begin{cases} \tau\Delta & \Delta \geq 0 \\ -(1-\tau)\Delta & \Delta < 0. \end{cases}$$

Then the τth quantile of the y_i is a minimizer of the function

$$f_\tau(\theta) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \theta).$$

Returning to the regression context, we can generalize LAD regression to quantile regression, namely computing the minimizer of the function

$$f_\tau(\boldsymbol{\theta}) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \langle \mathbf{x}_i \mid \boldsymbol{\theta} \rangle),$$

where $\mathbf{x}_i \in \mathbb{R}^p$ denotes the ith row of $\mathbf{X}$.

3 Trend Filtering

In the trend filtering problem (Kim et al., 2009; Tibshirani, 2014), one is interested in finding an adaptive polynomial approximation to noisy data $\mathbf{y} \in \mathbb{R}^n$ by solving the following convex problem.

$$\arg\min_{\boldsymbol{\theta}} \frac{1}{2n}\|\mathbf{y} - \boldsymbol{\theta}\|_2^2 + \lambda\|\mathbf{D}^{(k+1)}\boldsymbol{\theta}\|_1,$$

where $\lambda \geq 0$ is a regularization parameter that trades off the emphasis on the data fidelity term and the matrix $\mathbf{D}^{(k+1)} \in \mathbb{R}^{(n-k-1)\times n}$ is the discrete difference operator of order $k+1$. To understand the purpose of penalizing $\mathbf{D}^{(k+1)}$ consider the difference operator when $k=0$.

$$\mathbf{D}^{(1)} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

Thus, $\|\mathbf{D}^{(1)}\boldsymbol{\theta}\|_1 = \sum_{i=1}^{n-1}|\theta_i - \theta_{i+1}|$ which is just total variation denoising in one dimension. The penalty incentivizes solutions which are piece-wise constant. For $k \geq 1$, the difference operator $\mathbf{D}^{(k+1)} \in \mathbb{R}^{(n-k-1)\times n}$ is defined recursively as follows

$$\mathbf{D}^{(k+1)} = \mathbf{D}^{(1)}\mathbf{D}^{(k)}.$$

By penalizing the $k+1$ fold composition of the discrete difference operator, we obtain solutions which are piecewise polynomials of order k.

4 Quantile Trend Filtering

We combine the ideas of quantile regression and trend filtering, namely consider the signal approximation problem, where the design $\mathbf{X}$ is the identity matrix.

The estimation of the quantile trend filtering model can be posed as the following optimization problem.

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) + \lambda\|\mathbf{D}^{(k)}\boldsymbol{\theta}\|_1, \tag{4.1}$$

where λ is a nonnegative tuning parameter. As with the classic quantile regression, the quantile trend filtering problem can be solved by a linear program. We argue that it is better solved by Spingarn's method of partial inverses.

5 Related Work

- Quantile splines (Oh et al., 2011)
- BEADS

6 Spingarn's method of partial inverses

We first review Spingarn's method (Spingarn, 1985), which solves the following equality constrained convex problem:

$$\begin{aligned} &\text{minimize } \psi(\mathbf{x}) \\ &\text{subject to } \mathbf{x} \in V, \end{aligned} \tag{6.1}$$

where V is a subspace. The problem (6.1) can be expressed as the unconstrained optimization problem

$$\text{minimize } \psi(\mathbf{x}) + \iota_V(\mathbf{x}), \tag{6.2}$$

where ι_V is the indicator function of the set V. Spingarn's method applies Douglas-Rachford splitting to the problem (6.2) to give the following updates.

$$\begin{aligned}
\mathbf{x}^{(k+1)} &= \operatorname{prox}_{t\psi}(\mathbf{z}^{(k)}) \\
\mathbf{y}^{(k+1)} &= P_V\left(2\mathbf{x}^{(k+1)} - \mathbf{z}^{(k)}\right) \\
\mathbf{z}^{(k+1)} &= \mathbf{z}^{(k)} + \lambda^{(k)}\left(\mathbf{y}^{(k+1)} - \mathbf{x}^{(k+1)}\right).
\end{aligned}$$

The parameter t is a step-size and $\lambda^{(k)}$ is Krasnosel'skiĭ-Mann iteration (Need citation). We require that $\lambda^{(k)} \in]0,2[$ and that $\sum_n \lambda^{(k)}(2-\lambda^{(k)}) = \infty$. The mapping P_V is the orthogonal projection onto the set V. Note that the algorithm iterates $\mathbf{x}^{(k)}$ will converge to a solution to problem (6.1) (Combettes and Wajs, 2005). We need to experiment with different over / under-relaxation parameters $\lambda^{(k)}$ and step sizes t. It will converge if we take $\lambda^{(k)} = 1$ and $t = 1$, but we may be able to converge faster in practice by taking non-trivial values.

7 Applying Spingarn's Method to Quantile Trend Filtering

To simplify the notation we suppress the order k and write $\mathbf{D}^{(k)}$ as $\mathbf{D}$. We can reformulate our optimization problem (4.1) as the following equality constrained convex optimization problem.

$$\begin{aligned}
\text{minimize} \quad & f_1(\boldsymbol{\theta}) + f_2(\boldsymbol{\eta}) \\
\text{subject to} \quad & \boldsymbol{\eta} = \mathbf{D}\boldsymbol{\theta}
\end{aligned}$$

where

$$f_1(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) \qquad \text{and} \qquad f_2(\boldsymbol{\eta}) = \lambda\|\boldsymbol{\eta}\|_1.$$

If we set $\psi(\boldsymbol{\theta}, \boldsymbol{\eta}) = f_1(\boldsymbol{\theta}) + f_2(\boldsymbol{\eta})$ and $V = \{\mathbf{z}^\mathsf{T} = (\boldsymbol{\theta}^\mathsf{T}, \boldsymbol{\eta}^\mathsf{T}) : \boldsymbol{\eta} = \mathbf{D}\boldsymbol{\theta}\}$, then we can apply Spingarn's method. Note that

$$\begin{aligned}
\text{prox}_{th}(\boldsymbol{\theta}, \boldsymbol{\eta}) &= (\text{prox}_{tf_1}(\boldsymbol{\theta}), \text{prox}_{tf_2}(\boldsymbol{\eta})) \\
P_V(\boldsymbol{\theta}, \boldsymbol{\eta}) &= \begin{pmatrix} \mathbf{I} \\ \mathbf{D} \end{pmatrix} \left(\mathbf{I} + \mathbf{D}^\mathsf{T}\mathbf{D}\right)^{-1} \left(\boldsymbol{\theta} + \mathbf{D}^\mathsf{T}\boldsymbol{\eta}\right),
\end{aligned}$$

where the projection P_V requires a banded linear system solve, with bandwidth $k+1$. This linear solve can be accomplished in $\mathcal{O}(n(k+1)^2)$. The first solve using a banded Cholesky decomposition requires $\mathcal{O}(n(k+1)^2)$. Subsequent solves require $\mathcal{O}(n(k+1))$. We can use RccpArmadillo to do banded Cholesky (Eddelbuettel and Sanderson, 2014).

Proximal mappings

We need the proximal mappings for tf_1 and tf_2.

$$\begin{aligned}
\left[\text{prox}_{tf_1}(\boldsymbol{\theta})\right]_i &= y_i - \text{prox}_{(t/n)\rho_\tau}(y_i - \theta_i), \\
\left[\text{prox}_{tf_2}(\boldsymbol{\eta})\right]_j &= S(\eta_j, t\lambda).
\end{aligned}$$

The proximal mapping for tf_2 is the element-wise softhresholding operator. We now derive the proximal mapping of $\rho_\tau(\Delta)$, which can be evaluated in closed form. We need to find the minimizer of the following univariate function

$$g_\tau(\Delta) = \Delta[\tau - I(\Delta < 0)] + \frac{n}{2t}(\Delta - w)^2,$$

where $w \in \mathbb{R}$ is given and $I(\Delta < 0)$ is 0 when $\Delta < 0$ and 1 otherwise.

The subgradient of $\rho_\tau(\Delta) = \Delta[\tau - I(\Delta < 0)]$ is given by

$$\partial\rho_\tau(\Delta) = \begin{cases} \tau & \text{if } \Delta > 0 \\ \tau - 1 & \text{if } \Delta < 0 \\ [\tau - 1, \tau] & \text{if } \Delta = 0. \end{cases}$$

The stationary condition is

$$\frac{n}{t}[w - \Delta] \;\in\; \partial\rho_\tau(\Delta).$$

Therefore, the proximal mapping is given by

$$\operatorname{prox}_{(t/n)\rho_\tau}(w) \;=\; \begin{cases} w - \tau\frac{t}{n} & \text{if } w > \tau\frac{t}{n} \\ w + (1-\tau)\frac{t}{n} & \text{if } w < -(1-\tau)\frac{t}{n} \\ 0 & \text{if } -(1-\tau)\frac{t}{n} \le w \le \tau\frac{t}{n}. \end{cases}$$

Computational Costs

Precomputation

The following calculations need only be done once.

- $\mathcal{O}(n(k+1)^2)$ to compute the banded Cholesky factorization of $\mathbf{I} + [\mathbf{D}^{(k)}]^\mathsf{T}[\mathbf{D}^{(k)}]$

Per-Iteration

The following calculations will be done every iteration.

- $\mathcal{O}(n)$ to compute $\operatorname{prox}_{t\psi}(\boldsymbol{\theta}, \boldsymbol{\eta})$
- $\mathcal{O}((k+1)(n-k+1))$ to compute $\boldsymbol{\theta} + [\mathbf{D}^{(k)}]^\mathsf{T}\boldsymbol{\eta}$
- $\mathcal{O}(n(k+1))$ to compute $\boldsymbol{\phi} = (\mathbf{I} + [\mathbf{D}^{(k)}]^\mathsf{T}[\mathbf{D}^{(k)}])^{-1}(\boldsymbol{\theta} + \mathbf{D}^\mathsf{T}\boldsymbol{\eta})$
- $\mathcal{O}((k+1)(n-k+1))$ to compute $\begin{pmatrix} \mathbf{I} \\ \mathbf{D}^{(k)} \end{pmatrix} \boldsymbol{\phi}$

  The total cost is $\mathcal{O}(nk)$.

7.1 Summary

- The overall computational complexity is essentially linear $\mathcal{O}(nk^2)$ for the initial banded Cholesky decomposition and the per-iteration complexity is $\mathcal{O}(nk)$.
- One could also apply Anderson acceleration (Walker and Ni, 2011) to reduce the number of Spingarn updates, since the Douglas-Rachford algorithm is a fixed point algorithm.

8 Applying Spingarn's Method to Quantile Trend Filtering version 2

Following the specialized ADMM algorithm for trend filtering (Ramdas and Tibshirani, 2016), we can also take advantage of fast exact solvers of the one-dimensional fussed lasso problem (Davies and Kovac, 2001; Johnson, 2013). The first method is based on taut strings, and the second is based on dynamic programming. Both of these methods run in linear time. A third exact method with worst case quadratic penalty but linear time in typical cases was proposed by Condat (2013). C code is available for all three methods. We should compare them.

What is the reparameterization?

$$
\begin{aligned}
\text{minimize} \quad & f_1(\boldsymbol{\theta}) + f_2(\boldsymbol{\eta}) \\
\text{subject to} \quad & \boldsymbol{\eta} = \mathbf{D}\boldsymbol{\theta}
\end{aligned}
$$

where

$$
f_1(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) \qquad \text{and} \qquad f_2(\boldsymbol{\eta}) = \lambda \|\mathbf{D}^{(1)}\boldsymbol{\eta}\|_1.
$$

Difference with version 1: If we are interested in a 3rd order penalization, then $\mathbf{D} = \mathbf{D}^{(2)}$ in version 2 as opposed to $\mathbf{D} = \mathbf{D}^{(3)}$ in version 1. The proximal mapping for f_2 is solved using one of the exact solvers (Davies and Kovac, 2001; Johnson, 2013; Condat, 2013). Show some timing results between the different versions. Point readers to Ramdas and Tibshirani (2016) for a discussion of why this minor reparameterization may lead to speed up.

9 Other Practical Issues

- Read compressed data
- May want to read data directly in C and not pull into R; make R just an interface
- Use historical data to choose τ and λ

10 To Do

- Find out if we can use the EPA data. Add co-authors?
- Make version 2 with the three fast fused lasso solvers
- Add homotopy / warm start
- Add model selection
- Do experiments to evaluate different choices of t and $\lambda^{(k)}$
- See if Anderson acceleration helps. Talk to Tim Kelley?
- Write vignette
- Do comparisons with BEAD
- Do wind polar plots with and without removing trends
- Compare with Splines

11 Numerical Studies

We compare our detrending method with BEADS and the more general nonparametric quantile method introduced by Oh et al. (2011).

References

Apte, J. S., Messier, K. P., Gani, S., Brauer, M., Kirchstetter, T. W., Lunden, M. M., Marshall, J. D., Portier, C. J., Vermeulen, R. C., and Hamburg, S. P. (2017), "High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data," *Environmental Science & Technology*, in press.

Bacher, R., Chatelain, F., and Michel, O. (2016), "An adaptive robust regression method: Application to galaxy spectrum baseline estimation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4423–4427.

Combettes, P. L. and Wajs, V. R. (2005), "Signal Recovery by Proximal Forward-Backward Splitting," *Multiscale Modeling & Simulation*, 4, 1168–1200.

Condat, L. (2013), "A Direct Algorithm for 1-D Total Variation Denoising," *IEEE Signal Processing Letters*, 20, 1054–1057.

Davies, P. L. and Kovac, A. (2001), "Local Extremes, Runs, Strings and Multiresolution," *Annals of Statistics*, 29, 1–65.

Davis, D. (2017), "Convergence rate analysis of the forward-Douglas-Rachford splitting scheme," *SIAM Journal on Optimization*, in press.

Eddelbuettel, D. and Sanderson, C. (2014), "RcppArmadillo: Accelerating R with high-performance C++ linear algebra," *Computational Statistics and Data Analysis*, 71, 1054–1063.

He, B. and Yuan, X. (2012), "On the $O(1/n)$ Convergence Rate of the Douglas-Rachford Alternating Direction Method," *SIAM Journal on Numerical Analysis*, 50, 700–709.

— (2015), "On the convergence rate of Douglas-Rachford operator splitting method," *Mathematical Programming*, 153, 715–722.

Ilewicz, W., Kowalczuk, M., Niezabitowski, M., Buchczik, D., and Głuszka, A. (2015), "Comparison of baseline estimation algorithms for chromatographic signals," in *2015 20th International Conference on Methods and Models in Automation and Robotics (MMAR)*, pp. 925–930.

Johnson, N. A. (2013), "A Dynamic Programming Algorithm for the Fused Lasso and L0-Segmentation," *Journal of Computational and Graphical Statistics*, 22, 246–260.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009), "ℓ_1 Trend Filtering," *SIAM Review*, 51, 339–360.

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.

Ning, X., Selesnick, I. W., and Duval, L. (2014), "Chromatogram baseline estimation and denoising using sparsity (BEADS)," *Chemometrics and Intelligent Laboratory Systems*, 139, 156 – 167.

Oh, H.-S., Lee, T. C. M., and Nychka, D. W. (2011), "Fast Nonparametric Quantile Regression With Arbitrary Smoothing Methods," *Journal of Computational and Graphical Statistics*, 20, 510–526.

Ramdas, A. and Tibshirani, R. J. (2016), "Fast and Flexible ADMM Algorithms for Trend Filtering," *Journal of Computational and Graphical Statistics*, 0, 0–0.

Sanyal, A., Baral, A., and Lahiri, A. (2012), "Application of S-transform for removing baseline drift from ECG," in *2012 2nd National Conference on Computational Intelligence and Signal Processing (CISP)*, pp. 153–157.

Spingarn, J. E. (1985), "Applications of the method of partial inverses to convex programming: Decomposition," *Mathematical Programming*, 32, 199–223.

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. (2014), "Adaptive piecewise polynomial estimation via trend filtering," *The Annals of Statistics*, 42, 285–323.

Walker, H. F. and Ni, P. (2011), "Anderson Acceleration for Fixed-Point Iterations," *SIAM Journal on Numerical Analysis*, 49, 1715–1735.

Yuan, M. (2006), "GACV for quantile smoothing splines," *Computational statistics & data analysis*, 50, 813–829.