Drift Removal for Time Series Data Using Quantile Trend Filtering

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

Abstract

The text of your abstract. 200 or fewer words.

Keywords: 3 to 6 keywords, that do not appear in the title

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

†Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853 (E-mail: guinness@cornell.edu)

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

1 Introduction

We propose to use the trend filtering penalty with the check loss function to produce a non-parametric quantile regression estimate 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 ensure non-crossing while modeling multiple quantiles. We also implement a parallel ADMM algorithm for series that are too large to be computed simultaneously and proposed a modified criteria for choosing the smoothing parameter. We demonstrate through simulation studies that our proposed model provides better or comparable estimates of non-parametric quantile trends than existing methods and is a more effective method of drift removal for low-cost air quality sensors.

1.1 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 $y \in \mathbb{R}^n$, we estimate a regression vector $\theta \in \mathbb{R}^p$ so that $\mathbf{X}\theta$ is a good approximation of y. The LAD estimator is a solution to the problem

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

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

$$f(u) \;=\; \frac{1}{n} \sum_{i=1}^{n} |y_i - \theta|. \tag{2}$$

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"

$$\rho_\tau(\Delta) \;=\; \begin{cases} \tau\Delta & \Delta \geq 0 \\ -(1-\tau)\Delta & \Delta < 0 \end{cases} \tag{3}$$

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

$$f_\tau(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \theta). \tag{4}$$

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

$$f_\tau(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \langle x_i \mid \theta \rangle), \tag{5}$$

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

1.2 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 $y \in \mathbb{R}^n$ by solving the following convex problem.

$$\arg\min_{\theta} \frac{1}{2n}\|y - \theta\|_2^2 + \lambda\|\mathbf{D}^{(k+1)}\theta\|_1, \tag{6}$$

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 $k = 0$.

$$\mathbf{D}^{(1)} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \tag{7}$$

Thus, $\|\mathbf{D}^{(1)}\theta\|_1 = \sum_{i=1}^{n-1}|\theta_i - \theta_{i+1}|$ 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

$$\mathbf{D}^{(k+1)} \;=\; \mathbf{D}^{(1)}\mathbf{D}^{(k)}. \tag{8}$$

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

2 Methods

2.1 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.

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) + \lambda \|\mathbf{D}^{(k)}\theta\|_1, \tag{9}$$

where λ is a non-negative tuning parameter. As with the classic quantile regression, the quantile trend filtering problem can be solved by a linear program.

2.2 Smoothing parameter choice

Our method can easily handle missing data by changing the check loss function to output 0 for missing values. This allows us to leave out validation observations that can be used to select the tuning parameter λ and to compare method performance on real data. A number of methods have been proposed for selecting the quantile regression smoothing spline tuning parameter Yuan (2006). Koenker et al. (1994) relate λ to the number of interpolated points $p_\lambda = \sum I(y_i = \widehat{g}_i(x_i))$, which can be thought of as active knots, they propose the Schwarz criterion for the selection of λ

$$SIC(p_\lambda) = log[n^{-1} \sum \rho_\tau(y_i - \widehat{g}(x_i))] + \frac{1}{2n} p_\lambda log n \tag{10}$$

The traditional Bayesian Information Criterion (BIC) is given by

$$\text{BIC}(s) = -2\log(L\{\hat{\theta}(s)\}) + \nu(s)\log n \tag{11}$$

where $\theta(s)$ is the parameter θ with those components outside s being set to 0, and $\nu(s)$ is the number of components in s. If we assume an asymmetric Laplace likelihood $L(y|\theta) = \left(\frac{\tau^n(1-\tau)}{\sigma}\right)^n \exp\left\{-\sum_i \rho_\tau(\frac{y_i-\theta_i}{\sigma})\right\}$ and the number of non-zero elements of $D\theta$ as df

$$\text{BIC}(df) = 2\sum_i \frac{1}{\sigma}\rho_\tau(y_i - \theta_i) + df \log n \tag{12}$$

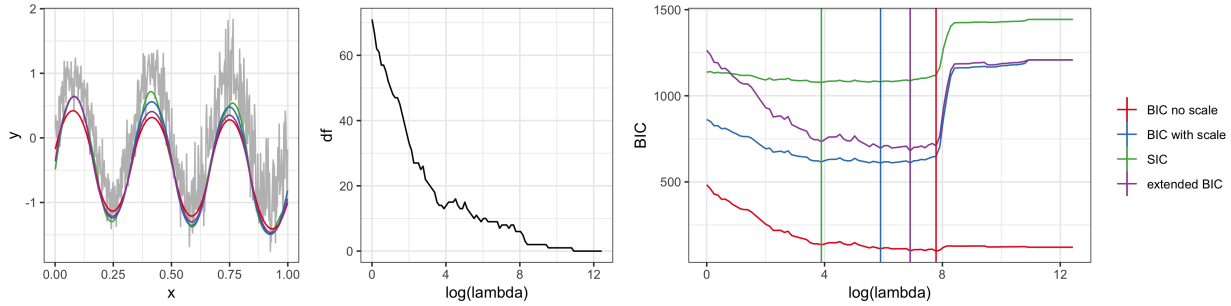
We can choose and $\sigma > 0$ and have found empirically that $\sigma = \frac{1-|1-2\tau|}{2}$ produces stable estimates. Chen and Chen (2008) proposed the extended BIC for large parameter spaces

$$BIC_\gamma(s) = -2\log(L\{\hat{\theta}(s)\}) + \nu(s)\log n + 2\gamma \log \binom{P}{j,} \quad \gamma \in [0,1] \tag{13}$$

where P is the total number of possible parameters and j is the number of parameters included in given model. We used this criteria with $\gamma = 1$, $P = n - k$ where k is the order of the differencing matrix and $j = \nu(s)$ is the number of non-zero entries in $D^{(k)}\theta$.

We used a single dataset to illustrate the difference between the scaled, unscaled and extended BIC criteria.

Figure 1: Degrees of freedom (number of non-zero elements of $D\theta$) by $\log(\lambda)$.



2.3 ADMM for Smoothing Windows

To our knowledge, no one has addressed the problem of finding smooth quantile trends of series that are too large to be processed simultaneously. We propose an alternating direction method of multipliers (ADMM) algorithm for solving large problems in a piecewise fashion. The ADMM algorithm, described by Boyd et al. (2011), relies on the idea of dual ascent.

If we consider the equality optimization problem

$$\text{minimize } f(x) \tag{14}$$

$$\text{subject to } Ax = b \tag{15}$$

with variable $x \in \mathbf{R}^n$, where $A \in \mathbf{R}^{mxn}$ and $f : \mathbf{R}^n \to \mathbf{R}$ is convex. The corresponding Lagrangian function is

$$L(x, \omega) = f(x) + \omega^T(Ax - b) \tag{16}$$

and the corresponding dual problem is

$$\text{maximize } \inf_x L(x, \omega) \tag{17}$$

Assuming that strong duality holds (Slater's constraint: there exists a solution in the interior of the domain), the optimal values of the primal and dual variables are the same. The dual ascent method, uses gradient ascent to solve the dual problem and thus the primal problem using the iterates

$$x^{k+1} = \arg\min L(x, \omega^k) \tag{18}$$

$$\omega^{k+1} = \omega^k + \alpha^k(Ax^{k+1} - b) \tag{19}$$

where $\alpha^k > 0$ is a step size and k is the iteration counter. In many cases (including our problem), the x update fails because the problem is unbounded below. To remedy this issue the augmented Lagrangian was introduced:

$$L_\gamma(x, \omega) = f(x) + \omega^T(Ax - b) + \frac{\gamma}{2}||Ax - b||_2^2 \tag{20}$$

where $\gamma > 0$ is called the penalty parameter. Applying dual ascent to the modified problem results in the method of multipliers algorithm which converges under more general conditions:

$$x^{k+1} = \arg\min L_\gamma(x, \omega^k) \tag{21}$$

$$\omega^{k+1} = \omega^k + \gamma(Ax^{k+1} - b) \tag{22}$$

Consider a new optimization problem with a separable objective function.

$$\text{minimize } f_1(x_1) + f_2(x_2) \tag{23}$$

$$\text{subject to } A_1x_1 + A_2x_2 = b \tag{24}$$

where $x = (x_1, x_2)$, $i \in \{1, 2\}$, and $x_i \in \mathbf{R}^{n_i}$. The dual ascent updates can be expressed as

$$x_i^{k+1} = \arg\min L_i(x_i, \omega^k) \tag{25}$$

$$\omega^{k+1} = \omega^k + \alpha^k(Ax^{k+1} - b) \tag{26}$$

however we lose the separability if the augmented Lagrangian is used instead, i.e. the augmented Lagrangian can not be separated into functions of x_i. The Alternating Direction Method of Multipliers (ADMM) addresses this problem and maintains separability while using the augmented Lagrangian to improve convergence through the following updates

$$x_1^{k+1} = \arg\min_{x_1} L_\gamma(x_1, x_2^k, \omega^k) \tag{27}$$

$$x_2^{k+1} = \arg\min_{x_2} L_\gamma(x_1^{k+1}, x_2, \omega^k) \tag{28}$$

$$\omega^{k+1} = \omega^k + \alpha^k(A_1x_1^{k+1} + A_2x_2^{k+1} - b) \tag{29}$$

We now consider the quantile regression trend filtering problem,

$$\text{minimize } \rho_\tau(y - \theta) + \lambda||D^{(k)}\theta||_1 \tag{30}$$

where $y, \theta \in \mathbf{R}^n$, y is the observed data, $\rho_\tau(u) = \sum_i(\tau - I(u_i < 0))u_i$ is the check loss function corresponding to quantile level τ and $D^{(k)}$ is the discrete differencing matrix of order k and λ is a tuning parameter that controls smoothness. We first re-parameterize $\phi = y - \theta$ so the problem is

$$\text{minimize } \rho_\tau(\phi) + \lambda||D^{(k)}(y - \phi)||_1 \tag{31}$$

We further divide ϕ order to solve smaller problems: Defining

$$\phi_1 = (\phi_{11}, \phi_{12}) \tag{32}$$

$$\phi_2 = (\phi_{21}, \phi_{22}, \phi_{23}) \tag{33}$$

$$\phi_3 = (\phi_{31}, \phi_{32}) \tag{34}$$

$$\phi = (\phi_{11}, \phi_{12} = \phi_{21}, \phi_{22}, \phi_{23} = \phi_{31}, \phi_{32}) \tag{35}$$

$$\tag{36}$$

Dividing y similarly, the problem then becomes

$$\text{minimize } \sum_{i=1}^{3} \rho_\tau(\phi_i) + \lambda||D^{(k)}(y_i - \phi_i)||_1 \tag{37}$$

$$\text{subject to: } \phi_{12} = \phi_{21}, \quad \phi_{23} = \phi_{31} \tag{38}$$

$$\tag{39}$$

We can further simplify by defining

$$\overline{\phi} = (\phi_{11}, \frac{\phi_{12} + \phi_{21}}{2}, \phi_{22}, \frac{\phi_{23} + \phi_{31}}{2}, \phi_{32}) \tag{40}$$

$$\overline{\phi_1} = (\phi_{11}, \frac{\phi_{12} + \phi_{21}}{2}) \tag{41}$$

$$\overline{\phi_2} = (\frac{\phi_{12} + \phi_{21}}{2}, \phi_{22}, \frac{\phi_{23} + \phi_{31}}{2}) \tag{42}$$

$$\overline{\phi_3} = (\frac{\phi_{23} + \phi_{31}}{2}, \phi_{32}) \tag{43}$$

so the problem becomes

$$\text{minimize } \sum_{i=1}^{3} \rho_\tau(\phi_i) + \lambda||D^{(k)}(y_i - \phi_i)||_1 \tag{44}$$

$$\text{subject to: } \phi_i = \overline{\phi_i} \tag{45}$$

$$\tag{46}$$

The augmented Lagrangian for this problem is

$$L_\gamma(\phi_1, \phi_2, \phi_3, \overline{\phi_1}, \overline{\phi_2}, \overline{\phi_3}, \omega) = \tag{47}$$

$$\sum_{i=1}^{3} \rho_\tau(\phi_i) + \lambda||D^{(k)}(y_i - \phi_i)||_1 + \omega_i^T(\phi_i - \overline{\phi_i}) + \frac{\gamma}{2}||\phi_i - \overline{\phi_i}||_2^2 \tag{48}$$

The ADMM updates are then given by

$$\phi_i^{k+1} = \arg\min_{\phi_i} \rho_\tau(\phi_i) + \lambda||D^{(k)}(y_i - \phi_i)||_1 + \omega_i^{kT}(\phi_i - \overline{\phi_i}^k) + \frac{\gamma}{2}||\phi_i - \overline{\phi_i}^k||_2^2 \tag{49}$$

$$\omega_i^{k+1} = \omega_i^k + \gamma(\phi_i^{k+1} - \overline{\phi_i}^{k+1}) \tag{50}$$

The ϕ_i updates can be obtained using a quadratic program solver such as Gurobi and can be obtained in parallel.

Figure 2: Windows fit separately compared to simultaneous fit, no signal present.

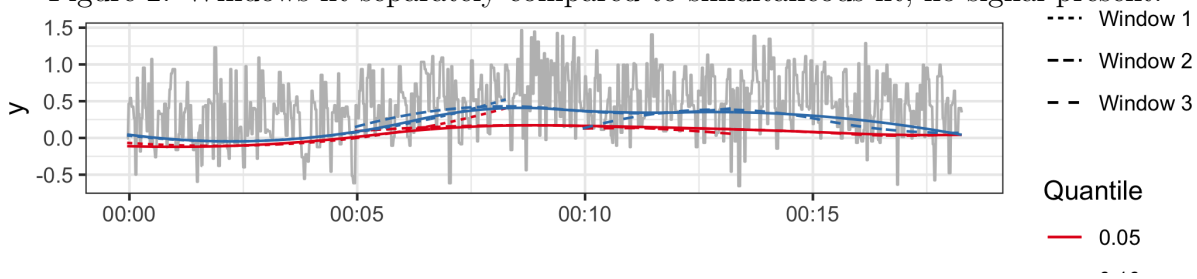


Figure 3: Windows fit with 150 iterations of ADMM, no signal present.

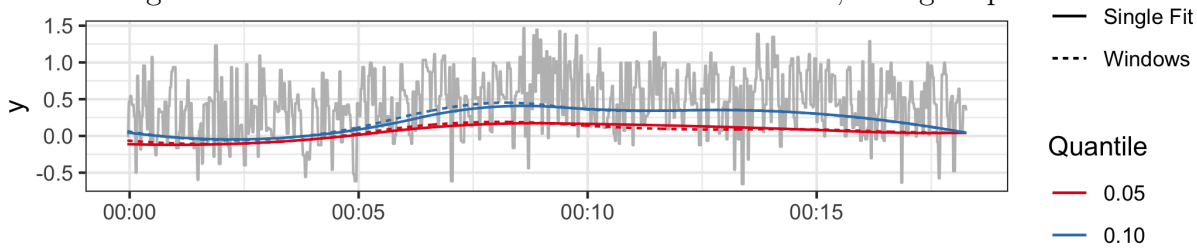


Figure 4: Windows fit separately compared to simultaneous fit, signal present.

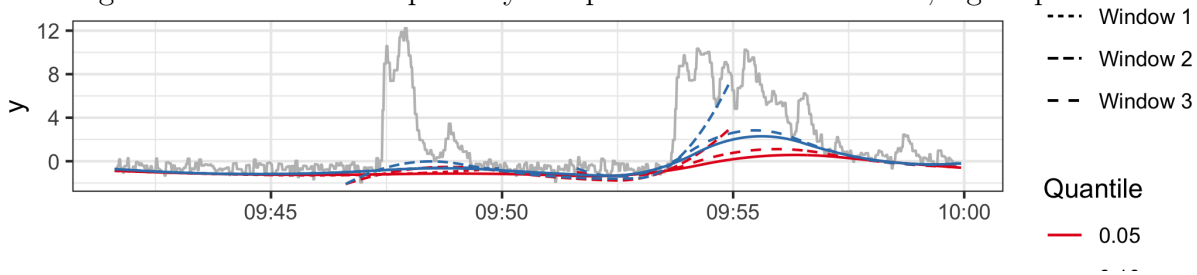
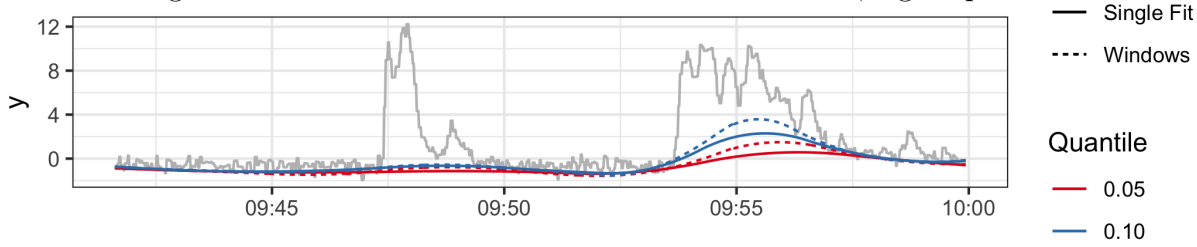


Figure 5: Windows fit with 300 iterations of ADMM, signal present.



3 Simulation Study

We compare the performance of our quantile trend filtering method with the three previously published methods using designs proposed by Racine and Li (2017). The methods compared are: `npqw` which is the quantile-ll method described in Racine and Li (2017), code was obtained from the author; `qsreg` in the `fields` R package and described in Oh et al. (2011); `rqss` available in the `quantreg` package and described in Koenker et al. (1994). The smoothing parameter λ for the `rqss` method is chosen using a grid search and minimizing the SIC criteria as described in Koenker et al. (1994). We further compare three criteria for choosing the smoothing parameter for our detrend method: `detrendr_SIC`: Our method where we minimize $\sum_i \rho_\tau(y_i - \theta_i) + \lambda||D\theta||_1$ and λ is chosen using SIC (Koenker et al., 1994). `detrendr_valid`: Our method where lambda is chosen by leaving out every 5th observation as a validation data set and minimizing the evaluating the check loss function evaluated at the validation data. `detrendr_eBIC`: the new criteria we have proposed based on the extended BIC proposed by Chen and Chen (2008).

Three simulation designs from Racine and Li (2017) were considered. For all designs X_i was generated as a uniformly spaced sequence in $[0, 1]$ and the response Y was generated as

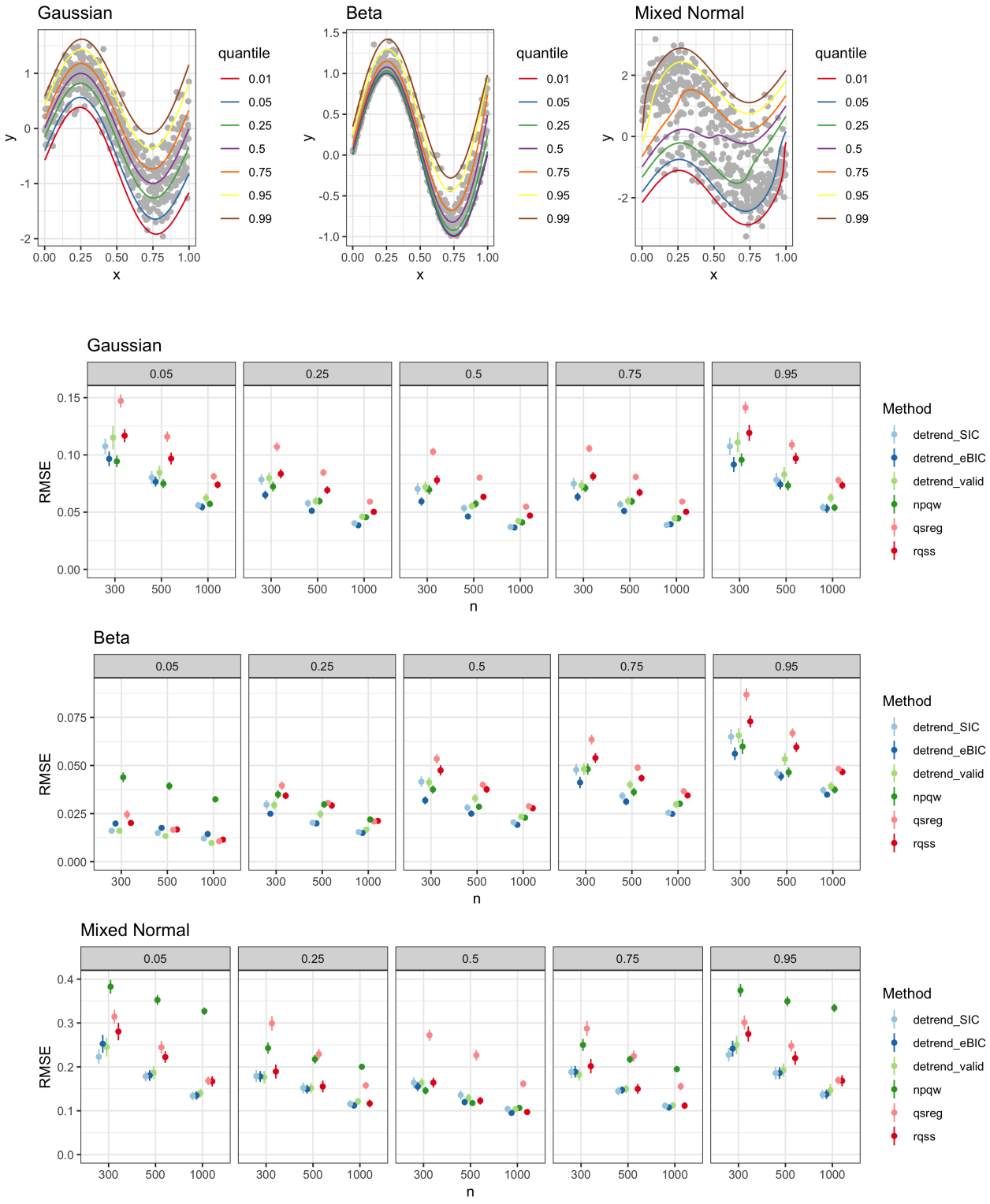
$$Y_i = sin(2\pi x_i) + \epsilon_i(x_i)$$

The three error distributions considered were

- Gaussian: $\epsilon_i(x_i) \sim N\left(0, \left(\frac{1+x_i^2}{4}\right)^2\right)$
- Beta: $\epsilon_i \sim Beta(1, 11 - 10x_i)$
- Mixed normal: ϵ_i is simulated from a mixture of $N(-1, 1)$ and $N(1, 1)$ with mixing probability x_i.

100 datasets were generated of sizes 300, 500 and 1000. The MSE was calculated as $\frac{1}{n}\sum_i(\hat{q}_\tau(x_i) - q_\tau(x_i))^2$. The plots below show the mean MSE $\pm$ twice the standard error by method, quantile level and sample size.

Figure 6: Simulated data with true quantiles $\tau \in \{0.01, 0.05, 0.25, 0.5, .75, 0.95, 0.99\}$



In all of the designs the proposed detrend methods are either better than or comparable

to existing methods. The npqw method performs particularly poorly in the mixed normal design, due to the fact that it assumes the data comes from a scale-location model which is violated in this case.

4 Peaks Simulation

We use another simulation design based on the applied problem we aim to solve. We assume that the measured data can be represented by

$$Y(t) = s(t) + b(t) + \epsilon \tag{51}$$

where $s(t)$ is the true signal at time t, $b(t)$ is the drift component that varies smoothly over time and $\epsilon \sim N(0, \sigma^2)$ is an error component. We assume t is a uniformly spaced sequence between 0 and 1. We generate $b(t)$ using a cubic natural spline basis function with degrees of freedom sampled from $n/50$ to $n/25$ with equal probability, and coefficients drawn from an exponential distribution with rate 1. The true signal function is assumed to be zero with Gaussian peaks. The number of peaks is sampled from $n/100$ to $n/50$ with equal probability with centers uniformly distributed between 0.1 and 0.9 and bandwidths uniformly distributed between $1/n$ and $5/n$ and areas uniformly distributed between 0 and $20/n$. One hundred datasets were generated for $n = \{300, 500, 1000, 5000\}$. We compare the methods ability to estimate the true quantiles of $b(t) + \epsilon$ for $\tau \in \{0.01, 0.05, 0.1\}$ and calculate the RMSE.

In addition to the RMSE we also calculated the signal miss-classification rate. We first classified the "true" peaks function into signal or not based on a threshold of 0.1. We then classified the detrended data using three different thresholds and calculated the fraction miss-classified. An illustration of the observations classified as signal after detrending compared to the "true signal" is shown in Fig. 9.

Our detrend_BIC method performs the best overall in terms of both RMSE and miss-classification rate. The lowest miss-classification rates were obtained using the detrend_eBIC method and a threshold of 1 for all data sizes. While qsreg was competitive with our method in some cases, both the RMSE and miss-classification rate increased substantially with the

Figure 7: Example of simulated peaks, baseline, and observed measurements.

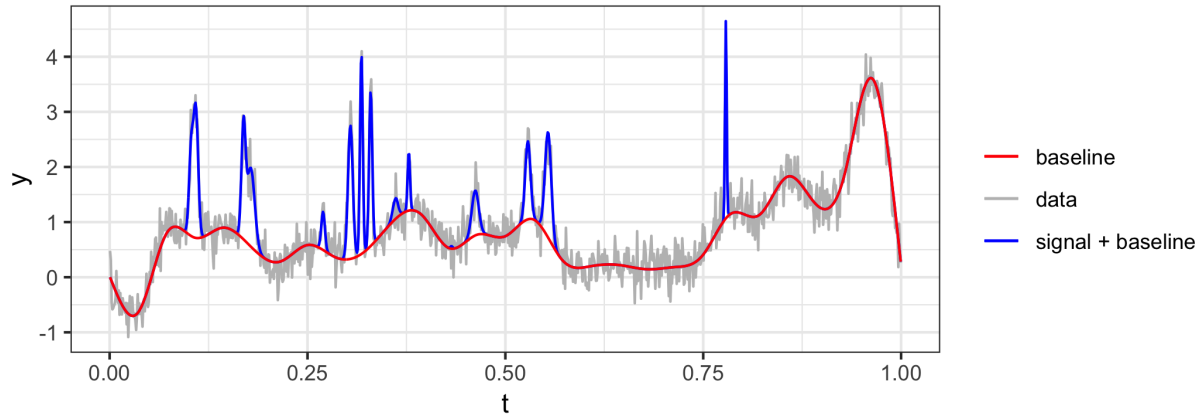
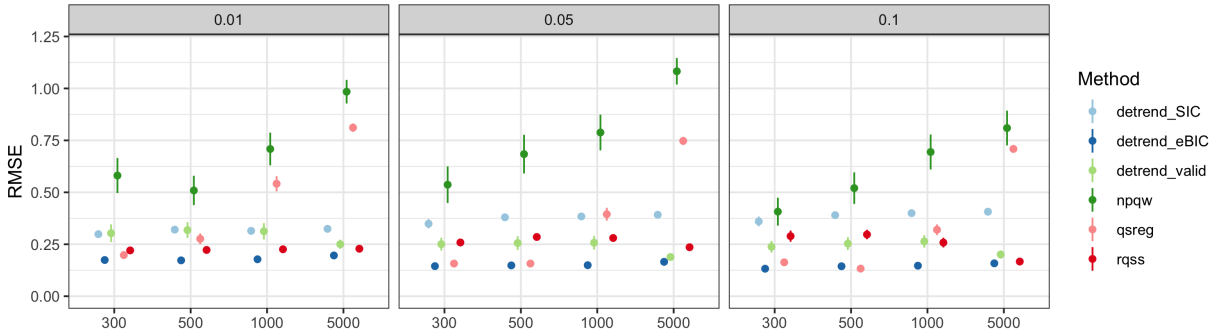


Figure 8: RMSEs compared to the simulated baseline function.



size of the dataset.

Figure 9: Example signal classification using threshold. Red indicates true signal > 0.1, blue indicates classified as signal after baseline removal using eBIC detrendr and a threshold of 1.

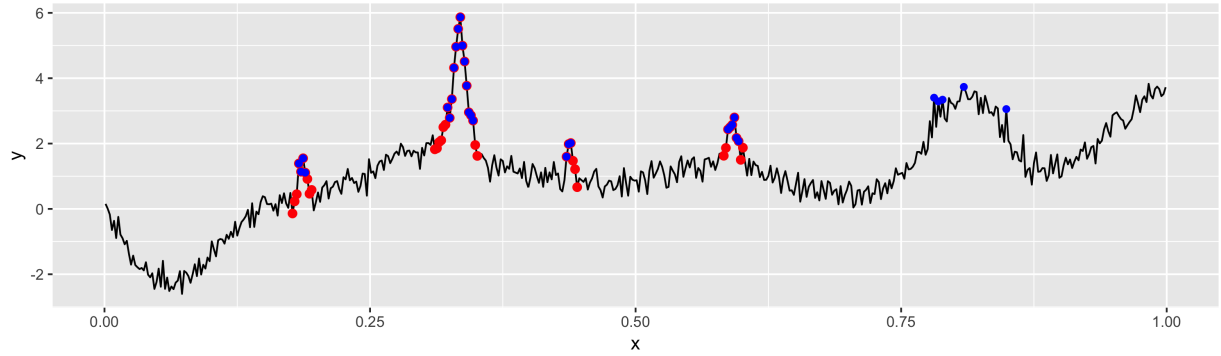


Figure 10: Miss-classification rates by threshold, data size, and method, values above the upper limit (npqw) not shown.

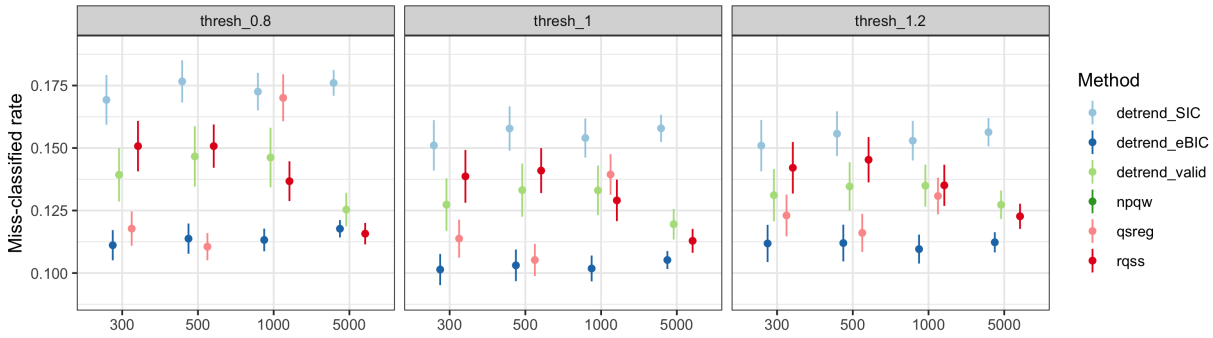


Table 1: Fraction of seconds with different signal (0/1) classifications.

Comparison	tau 0.1	tau 0.2	tau 0.3
fg	0.018	0.016	0.015
fh	0.015	0.013	0.011
gh	0.018	0.018	0.017

5 Application

Our windowed detrend method was used to removed the baseline drift from low cost air quality sensors so that the signal could be categorized using a simple threshold. The measurements were first standardized to have mean zero and variance 1. Three quantile levels for estimating the baseline trend were compared. The total dataset consisted of 52,322 observations per node. The signal thresholds were set using the first 15,000 observations where it was known no signal was present. The thresholds were set as 3 times the standard deviation plus the mean of observations in this time period. The total number of seconds of signal for each node as well as the number of seconds where multiple nodes both reported signal is shown in Table 1. Table 2 shows the fraction of observations with different signal classifications by node combination.

Figure 11: Low cost sensor data before drift removal (top), after drift removal using windowed detrend with eBIC (middle), and zoomed in on signal area (bottom). Horizontal lines represent signal thresholds.

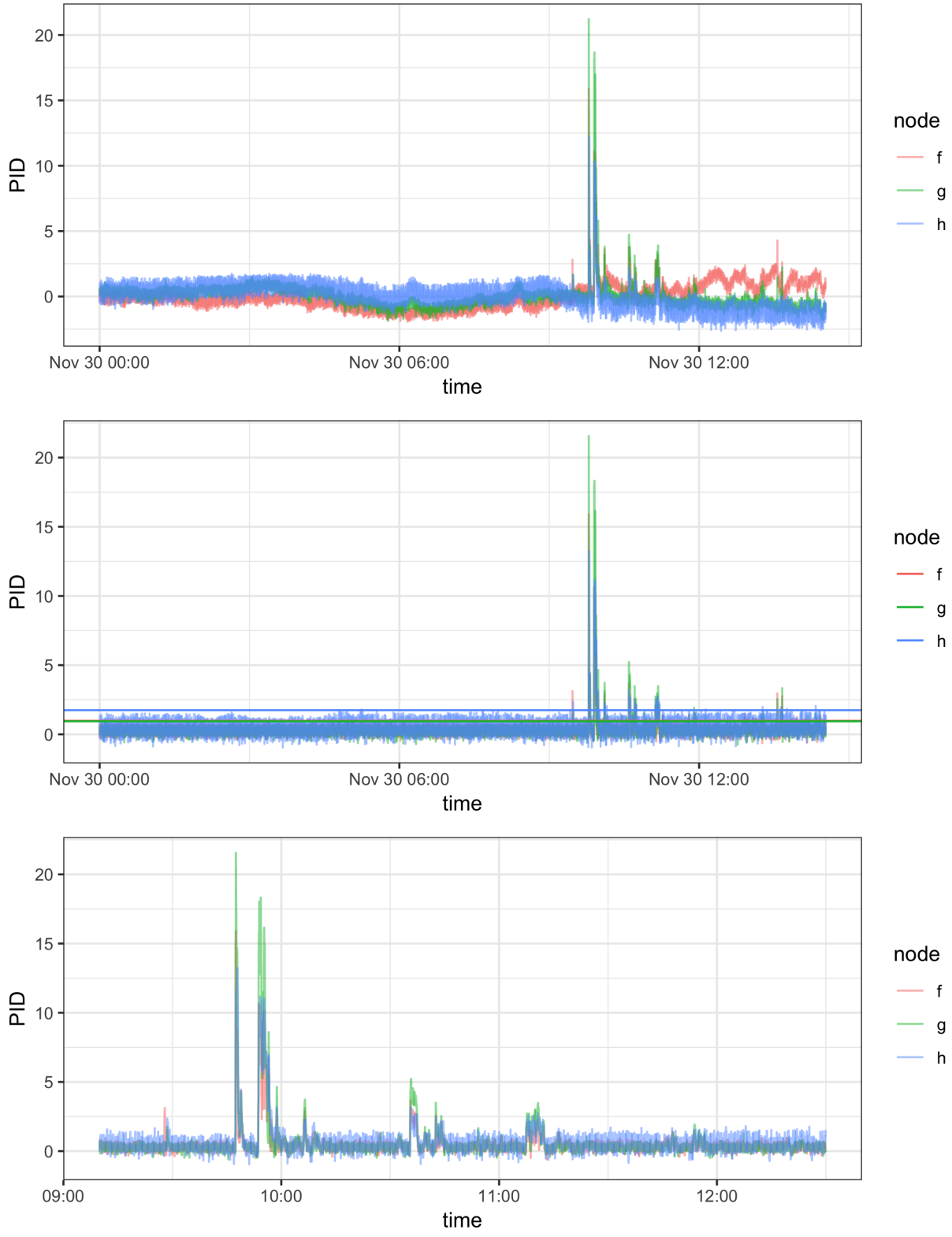


Table 2: Seconds of signal by node combination and quantile level.

Node	tau 0.1	tau 0.2	tau 0.3
f	1219	1079	933
g	1500	1410	1363
h	703	594	605
fh	561	491	487
gh	621	529	540
fg	891	832	764
fgh	542	469	469

6 Conclusion

SUPPLEMENTARY MATERIAL

R-package for detrend routine: R-package detrendr containing code to perform the diagnostic methods described in the article. The package also contains all datasets used as examples in the article. (GNU zipped tar file)

7 References

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011), "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, 3, 1–122.

Chen, J. and Chen, Z. (2008), "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, 95, 759–771.

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.

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.

Racine, J. S. and Li, K. (2017), "Nonparametric conditional quantile estimation: A locally weighted quantile kernel approach," *Journal of Econometrics*, 201, 72–94.

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

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