Baseline Drift Estimation for Time Series Data Using Quantile Trend Filtering

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

Abstract

We address the problem of estimating smoothly varying baseline trends in time series data using quantile trend filtering. We first extend the basic framework to ensure non-crossing while estimating multiple quantile trends simultaneously. We also implement a parallelizable alternating direction method of moments (ADMM) algorithm for estimating trends in longer series. The ADMM algorthim enables the estimation of trends in a piecewise manner, both reducing the computation time and extending the limits of the method to larger data sizes. We also address smoothing parameter selection and propose a modified criterion based on the extended Bayesian Information Criterion. Through simulation studies and an application to low cost air quality sensor data, we demonstrate that our model provides better quantile trend estimates than existing methods and improves signal classification of low-cost air quality sensor output.

Keywords: quantile regression, non-parametric regression, trend estimation, smoothing splines

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

†Department of Statistics and Data Science, 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

In a wide range of applications that spans chemistry (Ning et al., 2014), macroeconomics (Yamada, 2017), environmental science (Brantley et al., 2014), and medical sciences (Pettersson et al., 2013; Marandi and Sabzpoushan, 2015), scalar functions of time $y(t)$ are observed and assumed to be a superposition of an underlying slowly varying baseline trend $\theta(t)$, other more rapidly varying components $s(t)$, and noise. In practice, $y(t)$ is observed at discrete time points $t_1, \ldots, t_n$, and we model the vector of signal samples $y_i = y(t_i)$ as

$$\mathbf{y} \;=\; \boldsymbol{\theta} + \mathbf{s} + \boldsymbol{\varepsilon},$$

where $\theta_i = \theta(t_i)$, $s_i = s(t_i)$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of uncorrelated noise. For notational simplicity, for the rest of the paper, we assume that the time points take on the values $t_i = i$, but it is straightforward to generalize to an arbitrary grid of time points.

In some applications, the slowly varying component $\boldsymbol{\theta}$ is the signal of interest and the transient component $\mathbf{s}$ is a vector of nuisance parameters. In other applications, the roles of $\boldsymbol{\theta}$ and $\mathbf{s}$ are reversed. The latter applies in the air quality applications motivating this work, as $\mathbf{s}$ represents the signal of interest and $\boldsymbol{\theta}$ represents a baseline drift that obscures the identification of the important transient events encoded in $\mathbf{s}$. We describe this motivating problem further to illustrate the critical importance of being able to separate $\mathbf{s}$ from $\boldsymbol{\theta}$.

In the last decade, low cost and portable air quality sensors have enjoyed dramatically increased usage. These sensors can provide an un-calibrated measure of a variety of pollutants in near real time, but deriving meaningful information from sensor data remains a challenge (Snyder et al., 2013). The "SPod" is a low-cost sensor currently being investigated by researchers at the U.S. Environmental Protection Agency to detect volatile organic compound (VOC) emissions from industrial facilities (Thoma et al., 2016). To reduce cost and power consumption of the SPod, the temperature and relative humidity of the air presented to the photoionization detectors (PIDs) is not controlled and as a result the output signal exhibits a slowly varying baseline drift on the order of minutes to hours. Figure 1 provides an example of measurements from three SPod sensors co-located at the border of an industrial facility. All of the sensors respond to the pollutant signal, the three

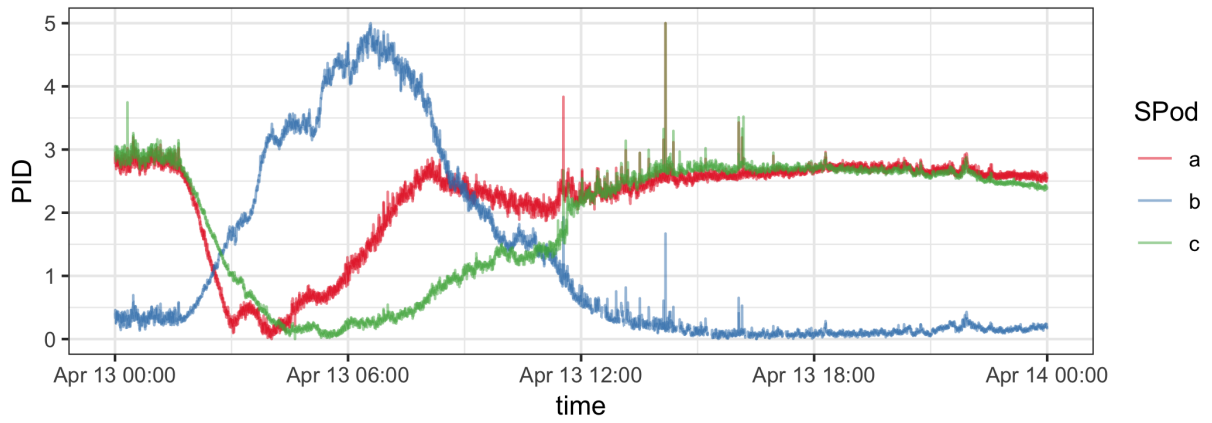


Figure 1: Example of 3 co-located SPod PID sensor readings.

sharp transient spikes at 11:32, 14:10, and 16:10, but the baseline drift varies from one sensor to another obscuring the detection of the peaks that alert the intrusion of pollutants. We show later that by estimating the baseline drift in each sensor and removing it from the observed signals, peaks can be reliably detected from concordant residual signals from a collection of SPods using a simple data-driven thresholding strategy. Thus, accurately demixing a noisy observed vector $\mathbf{y}$ into a slowly varying component $\boldsymbol{\theta}$ and a transient component $\mathbf{s}$ can lead to greatly improved and simplified downstream analysis. This is true for not just air quality applications, but universally across the many domains listed earlier.

To tackle demixing problems, we introduce a scalable baseline estimation framework by building on ℓ_1-*trend filtering*, a relatively new nonparametric estimation framework. Our contributions are three-fold.

- Kim et al. (2009) proposed using the check function as a possible extension of ℓ_1-trend filtering but did not investigate it further. Here, we develop the basic ℓ_1-quantile-trend-filtering framework and extend it to model multiple quantiles simultaneously with non-crossing constraints to ensure valid trend estimates.

- To reduce computation time and extend the method to long time series, we develop a parallelizable ADMM algorithm.

- Finally, we propose a modified criterion for performing model selection.

In the rest of the paper, we detail our quantile trend filtering algorithms as well as how to choose the smoother parameter λ (Section 2). We demonstrate through simulation studies that our proposed model provides better or comparable estimates of non-parametric quantile trends than existing methods (Section 4). We further show that quantile trend filtering is a more effective method of drift removal for low-cost air quality sensors and results in improved signal classification compared to quantile smoothing splines (Section 5). Finally, we discuss potential extensions of quantile trend filtering (Section 6).

2 Baseline Trend Estimation

2.1 Background

Kim et al. (2009) originally proposed ℓ_1-*trend filtering* to estimate trends with piecewise polynomial functions, assuming that the observed time series $\mathbf{y}$ consists of a trend $\boldsymbol{\theta}$ plus uncorrelated noise $\boldsymbol{\varepsilon}$, namely $\mathbf{y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}$. The estimated trend is the solution to the following convex optimization problem

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \; \frac{1}{2}\|\mathbf{y} - \boldsymbol{\theta}\|_2^2 + \lambda \|\mathbf{D}^{(k+1)}\boldsymbol{\theta}\|_1,$$

where λ is a nonnegative regularization parameter, 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 the size of $\mathbf{D}^{(k+1)}\boldsymbol{\theta}$ consider the difference operator when $k = 0$.

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

Thus, $\|\mathbf{D}^{(1)}\boldsymbol{\theta}\|_1 = \sum_{i=1}^{n-1} |\theta_i - \theta_{i+1}|$, which is known as total variation denoising penalty in one dimension in the signal processing literature (Rudin et al., 1992) or the fused lasso penalty in the statistics literature (Tibshirani et al., 2005). The penalty term incentivizes solutions which are piecewise 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

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

Penalizing the 1-norm of the vector $\mathbf{D}^{(k+1)}\boldsymbol{\theta}$ produces estimates of $\boldsymbol{\theta}$ that are piecewise polynomials of order k.

Tibshirani (2014) proved that with a judicious choice of λ the trend filtering estimate converges to the true underlying function at the minimax rate for functions whose kth derivative is of bounded variation and showed that trend filtering is both fast and locally adaptive when the time series consists of only the trend and random noise, which is illustrated in Figure 2a. As noted earlier though in some applications, such as the air quality monitoring problem considered in this paper, the data contain a rapidly varying signal in addition to the slowly varying trend and noise. Figure 2b shows that standard trend filtering is not designed to distinguish between the slowly varying trend and the rapidly-varying signal, as the smooth component estimate $\boldsymbol{\theta}$ is biased towards the peak of the transient component.

To account for the presence of a transient component in the observed time series $\mathbf{y}$, we propose quantile trend filtering Figure 2b. To estimate the trend in the τ^{th} quantile, we solve the convex optimization problem

$$\underset{\boldsymbol{\theta}}{\text{minimize}}\ \rho_\tau(\mathbf{y}-\boldsymbol{\theta}) + \lambda\|\mathbf{D}^{(k+1)}\boldsymbol{\theta}\|_1, \tag{1}$$

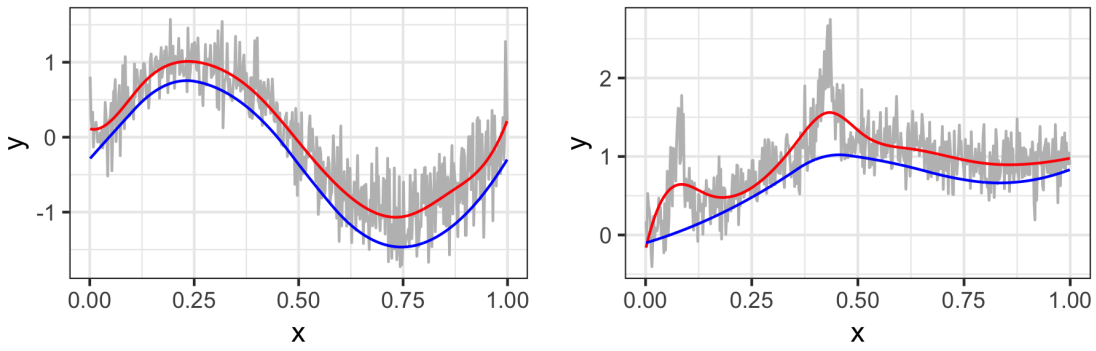
where $\rho_\tau(\mathbf{r})$ is the check function

$$\rho_\tau(\mathbf{r}) \quad = \quad \sum_{i=1}^{n} r_i(1-\mathbb{1}(r_i<0)), \tag{2}$$

and $\mathbb{1}(A)$ is 1 if its input A is true and 0 otherwise.

Note that we do not explicitly model $\mathbf{s}$. Rather, we focus on estimating $\boldsymbol{\theta}$. We then estimate $\mathbf{s}+\boldsymbol{\varepsilon}$ as the difference $\mathbf{y}-\boldsymbol{\theta}$.

Before elaborating on how we compute our proposed ℓ_1-quantile trend filtering estimator, we discuss similarities and differences between our proposed estimator and existing baseline estimators.



(a) Slow Trend + Noise

(b) Slow Trend + Spikes + Noise

Figure 2: Examples of trend filtering solutions (red) and 15^{th} quantile trending filtering solution (blue). Standard trend filtering performs well in the no-signal case (right) but struggles to distinguish between the slowly varying trend and the rapidly-varying signal (left). The quantile trend is not affected by the signal and provides an estimate of the baseline.

2.1.1 Relationship to Prior Work

In this application, as well as those described in Ning et al. (2014), Marandi and Sabzpoushan (2015), and Pettersson et al. (2013), the goal is to estimate the trend in the baseline not the mean. We can define the trend in the baseline as the trend in a low quantile of the data. A variety of methods for estimating quantile trends have already been proposed. Koenker and Bassett (1978) were the first to propose substituting the sum-of-squares term with the check function (2) to estimate a conditional quantile instead of the conditional mean. Later, Koenker et al. (1994) proposed quantile trend filtering with $k = 2$ producing quantile trends that are piecewise linear, but they did not consider extensions to higher order differences. Rather than using the ℓ_1-norm to penalize the discrete differences, Nychka et al. (1995) used the smoothing spline penalty based on the square of the ℓ_2-norm:

$$\sum_{i=1}^{n} \rho_\tau(y(t_i) - \theta(t_i)) + \lambda \int \theta''(t)^2 dt,$$

where $\theta(t)$ is a smooth function of time and λ is a tuning parameter that controls the degree of smoothing. Oh et al. (2011) proposed an algorithm for solving the quantile smoothing

spline problem by approximating the check function with a differentiable function. Racine and Li (2017) take a different approach and constrain the response to follow a smooth location scale model of the form $y(t) = a(t) + b(t)\epsilon_i$ and estimate the τ^{th} conditional quantile using a kernel estimator and local linear approach. **@Halley from Eric: What are $a(t), b(t)$, and ϵ_i?**

2.2 Quantile Trend Filtering

We combine the ideas of quantile regression and trend filtering. For a single quantile level τ, the quantile trend filtering problem is given in (1). As with classic quantile regression, the quantile trend filtering problem is a linear program which can be solved by a number of methods. In many cases, including ours, we are interested in estimating multiple quantiles simultaneously. We also want to ensure that our quantile estimates are valid by enforcing the constraint that if $\tau_2 > \tau_1$ then $Q(\tau_2) \geq Q(\tau_1)$ where Q is the quantile function of $\mathbf{y}$. Given quantiles $\tau_1 < \tau_2 < \ldots < \tau_J$, the optimization problem becomes

$$\underset{\boldsymbol{\Theta} \in \mathcal{C}}{\text{minimize}} \sum_{j=1}^{J} \left[\rho_{\tau_j}(\mathbf{y} - \boldsymbol{\theta}_j) + \lambda \|\mathbf{D}^{(k+1)} \boldsymbol{\theta}_j\|_1 \right]$$

where $\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 & \cdots & \boldsymbol{\theta}_J \end{pmatrix} \in \mathbb{R}^{n \times J}$ is a matrix whose jth column corresponds to the jth quantile signal $\boldsymbol{\theta}_j$ and the set $\mathcal{C} = \{\boldsymbol{\Theta} \in \mathbb{R}^{n \times J} : \theta_{ij} \leq \theta_{ij'} \text{ for } j \leq j'\}$ encodes the non-crossing quantile constraints. The additional constraints are linear in the parameters, so the non-crossing quantile trends can still be estimated by a number of available solvers. In the rest of this paper, we report numerical results using the commercial solver Gurobi (Gurobi Optimization, 2018) and its R package implementation. However, we could easily substitute a free solver such as the Rglpk package by Theussl and Hornik (2017).

2.3 ADMM for Big Data

The number of parameters to be estimated in this problem is equal to the number of observations multiplied by the number of quantiles of interest. As the size of the data and the number of quantiles grows, eventually the matrix of constraints will no longer fit into

memory. 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 a divide-and-conquer approach via an ADMM algorithm for solving large problems in a piecewise fashion. The ADMM algorithm (Gabay and Mercier, 1975; Glowinski and Marroco, 1975) is described in greater detail by Boyd et al. (2011), but we briefly review how it can be used to iteratively solve the following equality constrained optimization problem.

$$\begin{aligned} &\text{minimize } f(\boldsymbol{\theta}) + g(\tilde{\boldsymbol{\theta}}) \\ &\text{subject to } \mathbf{A}\boldsymbol{\theta} + \mathbf{B}\tilde{\boldsymbol{\theta}} = \mathbf{c}. \end{aligned} \tag{3}$$

Recall that finding the minimizer to an equality constrained optimization problem is equivalent to the identifying the saddle point of the Lagrangian function associated with the problem (3). ADMM seeks the saddle point of a related function called the augmented Lagrangian,

$$\mathcal{L}_\gamma(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}, \boldsymbol{\omega}) \;=\; f(\boldsymbol{\theta}) + g(\tilde{\boldsymbol{\theta}}) + \langle \boldsymbol{\omega}, \mathbf{c} - \mathbf{A}\boldsymbol{\theta} - \mathbf{B}\tilde{\boldsymbol{\theta}} \rangle + \frac{\gamma}{2} \|\mathbf{c} - \mathbf{A}\boldsymbol{\theta} - \mathbf{B}\tilde{\boldsymbol{\theta}}\|_2^2,$$

where the dual variable $\boldsymbol{\omega}$ is a vector of Lagrange multipliers and γ is a nonnegative tuning parameter. When $\gamma = 0$, the augmented Lagrangian coincides with the ordinary Lagrangian.

ADMM minimizes the augmented Lagrangian one block of variables at a time before updating the dual variable $\boldsymbol{\omega}$. This yields the following sequence of updates at the $m+1$th ADMM iteration

$$\begin{aligned} \boldsymbol{\theta}_{m+1} &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}_\gamma(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}_m, \boldsymbol{\omega}_m) \\ \tilde{\boldsymbol{\theta}}_{m+1} &= \arg\min_{\tilde{\boldsymbol{\theta}}} \mathcal{L}_\gamma(\boldsymbol{\theta}_{m+1}, \tilde{\boldsymbol{\theta}}, \boldsymbol{\omega}_m) \\ \boldsymbol{\omega}_{m+1} &= \boldsymbol{\omega}_m + \gamma(\mathbf{c} - \mathbf{A}\boldsymbol{\theta}_{m+1} - \mathbf{B}\tilde{\boldsymbol{\theta}}_{m+1}). \end{aligned} \tag{4}$$

Let $\mathcal{I}$ denote a subset of the indices $\{1, \ldots, n\}$ and $\mathbf{y}_\mathcal{I} \in \mathbb{R}^{|I|}$ denote the subvector obtained by stacking on top of each other the elements y_i for $i \in \mathcal{I}$ in their canonical ordering. We apply the consensus ADMM algorithm to the quantile regression trend filtering problem given in (1), by dividing our observed series y_i with $i = 1, \ldots, n$ into overlapping W

windows of observations $\mathbf{y}^{(w)} = \mathbf{y}_{\mathcal{I}_w}$ where $\mathcal{I}_w$ is the set of integers from l_w to u_w inclusive, namely $\mathcal{I}_w = \{l_w, l_w + 1, \ldots, u_w - 1, u_w\}$ where

$$1 = l_1 < l_2 < u_1 < l_3 < u_2 < l_4 < u_3 < \cdots < u_W = n.$$

Figure 3 shows an example of 1200 observations being mapped into three equally sized overlapping windows of observations.

Given quantiles $\tau_1 < \cdots < \tau_J$, we introduce dummy variables $\boldsymbol{\theta}_j^{(w)} \in \mathbb{R}^{|\mathcal{I}_w|}$ as the value of the $\tau_j{}^{\text{th}}$ quantile trend in window w. We then "stitch" together the W quantile trend estimates into consensus over the overlapping regions by introducing the constraint $\theta_{ij}^{(w)} = \theta_{i+l_w-1,j}$ for $i = 1, \ldots, u_w - l_w + 1$ and for all j. Let $\boldsymbol{\Theta}^{(w)}$ be the matrix whose jth column is $\boldsymbol{\theta}^{(w)}$. Then we can write these constraints more compactly as $\boldsymbol{\Theta}^{(w)} = \mathbf{U}^{(w)}\boldsymbol{\Theta}$, where $\mathbf{U}^{(w)} \in \{0,1\}^{|\mathcal{I}_w| \times n}$ is a matrix that selects rows of $\boldsymbol{\Theta}$ corresponding to the wth window, namely

$$\mathbf{U}^{(w)} \quad = \quad \begin{pmatrix} \mathbf{e}_{l_w}^\mathsf{T} \\ \vdots \\ \mathbf{e}_{u_w}^\mathsf{T} \end{pmatrix},$$

where $\mathbf{e}_i \in \mathbb{R}^n$ denotes the ith standard basis vector. Furthermore, let $\iota_{\mathcal{C}}$ denote the indicator function of the non-crossing quantile constraint, namely $\iota_{\mathcal{C}}(\boldsymbol{\Theta})$ is zero if $\boldsymbol{\Theta} \in \mathcal{C}$ and infinity otherwise.

Our windowed quantile trend optimization problem can then be written as

$$\begin{aligned} &\text{minimize} \sum_{w=1}^{W} \left\{ \sum_{j=1}^{J} \left[\rho_{\tau_j}(\mathbf{y}^{(w)} - \boldsymbol{\theta}_j^{(w)}) + \lambda \|\mathbf{D}^{(k+1)} \boldsymbol{\theta}_j^{(w)}\|_1 \right] + \iota_{\mathcal{C}}(\boldsymbol{\Theta}^{(w)}) \right\} \\ &\text{subject to} \qquad \boldsymbol{\Theta}^{(w)} = \mathbf{U}^{(w)}\boldsymbol{\Theta} \quad \text{for } w = 1, \ldots, W. \end{aligned} \tag{5}$$

Let $\boldsymbol{\Omega}^{(w)}$ denote the Lagrange multiplier matrix for the wth consensus constraint, namely $\boldsymbol{\theta}^{(w)} = \mathbf{U}^{(w)}\boldsymbol{\theta}$, and let $\boldsymbol{\omega}_j^{(w)}$ denote its jth column.

The augmented Lagrangian is given by

$$\mathcal{L}(\boldsymbol{\Theta}, \{\boldsymbol{\Theta}^{(w)}\}, \{\boldsymbol{\Omega}^{(w)}\}) \quad = \quad \sum_{w=1}^{W} \mathcal{L}_w(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(w)}, \boldsymbol{\Omega}^{(w)}),$$

where

$$\begin{aligned}\mathcal{L}_w(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(w)}, \boldsymbol{\Omega}^{(w)}) &= \sum_{j=1}^{J}\Big[\rho_{\tau_j}(\mathbf{y}^{(w)} - \boldsymbol{\theta}_j^{(w)}) + \lambda\|\mathbf{D}^{(k+1)}\boldsymbol{\theta}_j^{(w)}\|_1 \\ &\quad + (\boldsymbol{\theta}_j^{(w)} - \mathbf{U}^{(w)}\boldsymbol{\theta}_j)^{\mathsf{T}}\boldsymbol{\omega}_j^{(w)} + \frac{\gamma}{2}\|\boldsymbol{\theta}_j^{(w)} - \mathbf{U}^{(w)}\boldsymbol{\theta}_j\|_2^2\Big] + \iota_{\mathcal{C}}(\boldsymbol{\Theta}^{(w)}),\end{aligned}$$

where γ is a positive penalty parameter.

The ADMM algorithm alternates between updating the consensus variable $\boldsymbol{\Theta}$, the window variables $\{\boldsymbol{\Theta}^{(w)}\}$, and the Lagrange multipliers $\{\boldsymbol{\Omega}^{(w)}\}$. At the $m+1$th iteration, we perform the following sequence of updates

$$\begin{aligned}\boldsymbol{\Theta}_{m+1} &= \underset{\boldsymbol{\Theta}}{\arg\min}\ \mathcal{L}(\boldsymbol{\Theta}, \{\boldsymbol{\Theta}_m^{(w)}\}, \{\boldsymbol{\Omega}_m^{(w)}\}) \\ \boldsymbol{\Theta}_{m+1}^{(w)} &= \underset{\{\boldsymbol{\Theta}^{(w)}\}}{\arg\min}\ \mathcal{L}(\boldsymbol{\Theta}_{m+1}, \{\boldsymbol{\Theta}^{(w)}\}, \{\boldsymbol{\Omega}_m^{(w)}\})\end{aligned}$$

Updating $\boldsymbol{\Theta}$: Some algebra shows that updating the consensus variable step is computed as follows.

$$\theta_{ij} = \begin{cases} \frac{1}{2}\left(\theta_{ij}^{(w-1)} - \gamma^{-1}\omega_{ij}^{(w-1)}\right) + \frac{1}{2}\left(\theta_{ij}^{(w)} - \gamma^{-1}\omega_{ij}^{(w)}\right) & \text{if } l_w \le i \le u_{w-1}, \\ \theta_{ij}^{(w)} & \text{if } u_{w-1} < i \le l_{w+1}\,, \\ \frac{1}{2}\left(\theta_{ij}^{(w)} - \gamma^{-1}\omega_{ij}^{(w)}\right) + \frac{1}{2}\left(\theta_{ij}^{(w+1)} - \gamma^{-1}\omega_{ij}^{(w+1)}\right) & \text{if } l_{m+1} < i \le u_m \end{cases} \tag{6}$$

The consensus update (6) is rather intuitive. We essentially average the trend estimates in overlapping sections of the windows, subject to some adjustment by the Lagrange multipliers, and leave the trend estimates in non-overlapping sections of the windows untouched. For notational ease, we write the consensus update (6) compactly as $\boldsymbol{\Theta} = \psi(\{\boldsymbol{\Theta}^{(w)}\}, \{\boldsymbol{\Omega}^{(w)}\})$.

Updating $\{\boldsymbol{\Theta}^{(w)}\}$: We then estimate the trend separately in each window, which can be done in parallel, while constraining the overlapping pieces of the trends to be equal as outlined in Algorithm 1. The use of the Augmented Lagrangian converts the problem from a linear program into a quadratic program. The `gurobi` R package (Gurobi Optimization,

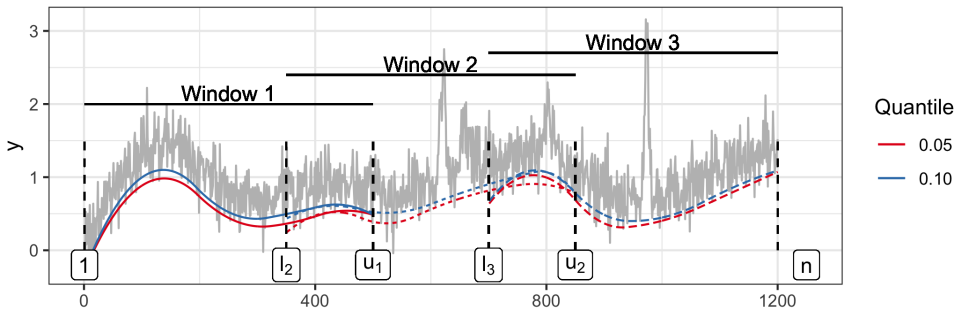


Figure 3: Window boundaries and trends fit separately in each window.

2018) can solve quadratic programs in addition to linear programs, but we can also use the free R package `quadprog` (Weingessel and Turlach, 2013).

To terminate our algorithm, we use the stopping criteria described by Boyd et al. (2011). The criteria are based on the primal and dual residuals, which represent the residuals for primal and dual feasibility, respectively. The primal residual at the mth iteration,

$$r^m_{\text{primal}} = \sqrt{\sum_{w=1}^{W} \|\boldsymbol{\Theta}^{(w)}_m - \boldsymbol{\Theta}_m\|^2_{\text{F}}},$$

represents the difference between the trend values in the windows and the consensus trend value. The dual residual at the mth iteration,

$$r^m_{\text{dual}} = \gamma\sqrt{\sum_{w=1}^{W} \|\boldsymbol{\Theta}_m - \boldsymbol{\Theta}_{m-1}\|^2_{\text{F}}},$$

represents the change in the consensus variable from one iterate to the next. The algorithm is stopped when

$$\begin{aligned} r^m_{\text{primal}} &< \epsilon_{\text{abs}}\sqrt{nJ} + \epsilon_{\text{rel}} \max_w \left[\max\left\{\|\boldsymbol{\Theta}^{(w)}_m\|_{\text{F}}, \|\boldsymbol{\Theta}_m\|_{\text{F}}\right\}\right] \\ r^m_{\text{dual}} &< \epsilon_{\text{abs}}\sqrt{nJ} + \epsilon_{\text{rel}} \sqrt{\sum_{w=1}^{W} \|\boldsymbol{\Omega}^{(w)}_m\|^2_{\text{F}}}. \end{aligned}$$

Timing experiments illustrate the advantages of using our ADMM algorithm when the trend can be estimated with a single window. For each data size, n, 25 datasets were simulated using the peaks simulation design described below. Trends for three quantiles

Algorithm 1 ADMM algorithm for quantile trend filtering with windows

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

initialize:

for $w = 1, \ldots, W$ **do**

$\boldsymbol{\Theta}_0^{(w)} \leftarrow \underset{\boldsymbol{\Theta}^{(w)} \in \mathcal{C}}{\arg\min} \sum_{j=1}^{J} \rho_{\tau_j}(\mathbf{y}^{(w)} - \boldsymbol{\theta}_j^{(w)}) + \lambda \|\mathbf{D}\boldsymbol{\theta}_j^{(w)}\|_1$

$\boldsymbol{\Omega}_0^{(w)} \leftarrow \mathbf{0}$

end for

$m \leftarrow 0$

repeat

$\boldsymbol{\Theta}_{m+1} \leftarrow \psi(\{\boldsymbol{\Theta}_m^{(w)}\}, \{\boldsymbol{\Omega}_m^{(w)}\})$

for $w = 1, \ldots, W$ **do**

$\boldsymbol{\Theta}_{m+1}^{(w)} \leftarrow \underset{\boldsymbol{\Theta}^{(w)}}{\arg\min}\, \mathcal{L}_w(\boldsymbol{\Theta}_{m+1}, \boldsymbol{\Theta}^{(w)}, \boldsymbol{\Omega}_m^{(w)})$

$\boldsymbol{\Omega}_{m+1}^{(w)} \leftarrow \boldsymbol{\Omega}_m^{(w)} + \gamma(\boldsymbol{\Theta}_{m+1}^{(w)} - \mathbf{U}^{(w)}\boldsymbol{\Theta}_{m+1})$

end for

$m \leftarrow m + 1$

until convergence

return $\boldsymbol{\Theta}_m$

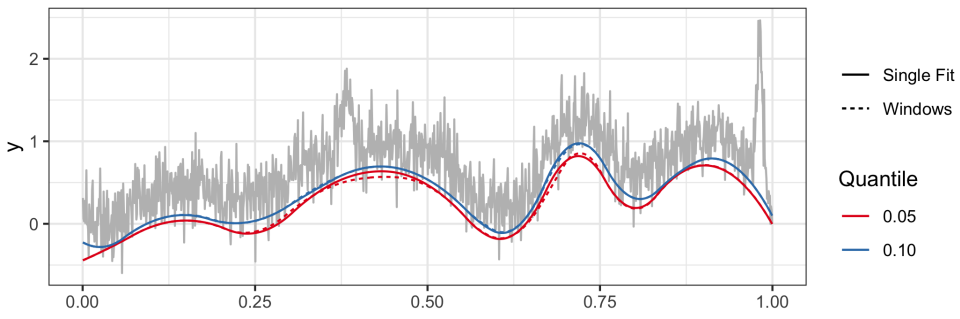


Figure 4: Trend fit with our ADMM algorithm with 3 windows which converged in 7 iterations compared to trend from simultaneous fit.

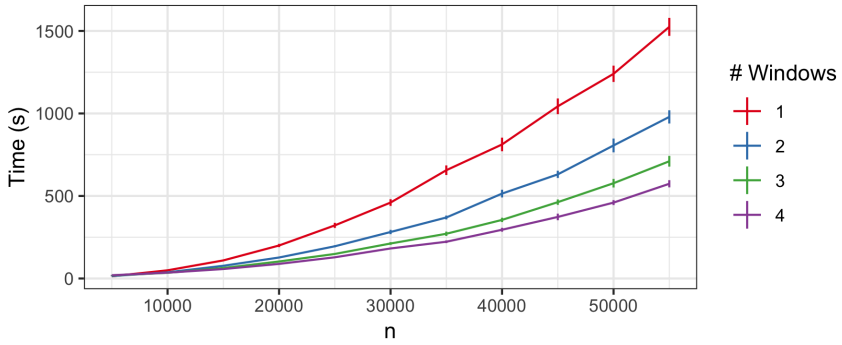


Figure 5: Timing experiments comparing quantile trend filtering with varying numbers of windows by data size.

were fit simultaneously: 0.05, 0.1, and 0.15 using $\lambda = n/5$. We use from one to four windows for each data size with an overlap of 500. The windows algorithm was run until the stopping criteria were met using $\epsilon_{abs} = 0.01$ and $\epsilon_{rel} = 0.001$. Figure 5 shows that using 4 windows instead of one on data sizes of 55000 provides a factor of 3 decrease in computation time. The timing experiments were conducted on an Intel Xeon based Linux cluster using two processor cores.

2.4 Computational Complexity and Convergence

@Eric from Eric: Add computational complexity of quadratic program and discuss the savings compared with solving a single large LP.

Algorithm 1 has the following convergence guarantees.

Proposition 2.1. *Let* $\{\{\boldsymbol{\Theta}_m^{(w)}\}, \boldsymbol{\Theta}_m\}$ *denote the* m*th collection of iterates generated by Algorithm 1. Then (i)* $\|\boldsymbol{\Theta}_m^{(w)} - \mathbf{U}^{(w)}\boldsymbol{\Theta}_m\|_F \to 0$ *and (ii)* $p_m \to p^\star$*, where* $p^\star$ *is the optimal objective function value of (5) and* p_m *is the objective function value of (5) evaluated at* $\{\{\boldsymbol{\Theta}_m^{(w)}\}, \boldsymbol{\Theta}_m\}$.

The proof of Proposition 2.1 is a straightforward application of the convergence result presented in Section 3.2 of Boyd et al. (2011).

3 Model Selection

An important practical issue in baseline estimation is the choice of regularization parameter λ, which controls the degree of smoothness in $\boldsymbol{\theta}$. In this section, we introduce four methods for choosing λ. The first is a validation based approach; the latter three are based on information criteria.

3.1 Validation

Our method can easily handle missing data by defining the check loss function to output 0 for missing values. Specifically, we use the following modified function $\tilde{\rho}_\tau$ in place of the ρ_τ function given in (2)

$$\tilde{\rho}_\tau(\mathbf{r}) \quad = \quad \sum_{i \notin \mathcal{V}} r_i(1 - \mathbb{1}(r_i < 0)), \tag{7}$$

where $\mathcal{V}$ is a held-out validation subset of $\{1, \ldots, n\}$ and solve the problem

$$\underset{\boldsymbol{\Theta} \in \mathcal{C}}{\text{minimize}} \sum_{j=1}^{J} \left[\tilde{\rho}_{\tau_j}(\mathbf{y} - \boldsymbol{\theta}_j) + \lambda \|\mathbf{D}^{(k+1)} \boldsymbol{\theta}_j\|_1 \right], \tag{8}$$

which can be solved via Algorithm 1 with trivial modification to the quadratic program subproblems. We then select the λ that minimizes the hold-out prediction error quantified by $\breve{\rho}(\mathbf{y} - \hat{\boldsymbol{\theta}}(\lambda))$ where $\hat{\boldsymbol{\theta}}(\lambda)$ is the solution to (8) and

$$\breve{\rho}_\tau(\mathbf{r}) \quad = \quad \sum_{i \in \mathcal{V}} r_i(1 - \mathbb{1}(r_i < 0)).$$

3.2 Information Criteria

Koenker et al. (1994) addressed the choice of regularization parameter by proposing the Schwarz criterion for the selection of λ

$$\text{SIC}(p_\lambda) \quad = \quad \log \left[\frac{1}{n} \rho_\tau(\mathbf{y} - \boldsymbol{\theta}) \right] + \frac{1}{2n} p_\lambda \log n.$$

where $p_\lambda = \sum_i \mathbb{1}(y_i = \widehat{\theta}_i)$ is the number of interpolated points, which can be thought of as active knots. The SIC is based on the traditional Bayesian Information Criterion (BIC)

which is given by

$$\text{BIC}(\nu) = -2\log(L\{\hat{\boldsymbol{\theta}}\}) + \nu \log n \tag{9}$$

where L is the likelihood function and ν is the number of non-zero components in $\hat{\boldsymbol{\theta}}$. If we take the approach used in Bayesian quantile regression (Yu and Moyeed, 2001), and view minimizing the check function as maximizing the asymmetric Laplace likelihood,

$$L(\mathbf{y} \mid \boldsymbol{\theta}) \quad = \quad \left[\frac{\tau^n(1-\tau)}{\sigma}\right]^n \exp\left\{-\sum_i \rho_\tau\left(\frac{y_i - \theta_i}{\sigma}\right)\right\},$$

we can compute the BIC as

$$\text{BIC}(\nu) = 2\frac{1}{\sigma}\rho_\tau(\mathbf{y} - \hat{\boldsymbol{\theta}}) + \nu \log n$$

where $\hat{\boldsymbol{\theta}}$ is the estimated trend, and ν is the number of non-zero elements of $\mathbf{D}^{(k+1)}\hat{\boldsymbol{\theta}}$. We can choose any $\sigma > 0$ and have found empirically that $\sigma = \frac{1-|1-2\tau|}{2}$ produces stable estimates.

Chen and Chen (2008) proposed the extended Bayesian Information Criteria (eBIC), specifically designed for large parameter spaces.

$$\text{eBIC}_\gamma(\nu) \quad = \quad -2\log(L\{\hat{\boldsymbol{\theta}}\}) + \nu \log n + 2\gamma \log \binom{P}{\nu}, \quad \gamma \in [0,1]$$

where P is the total number of possible parameters and ν is the number of non-zero parameters included in given mod. We used this criteria with $\gamma = 1$, and $P = n - k - 1$. Figure 6 illustrates the difference among the scaled, unscaled ($\sigma = 1$), and scaled extended BIC criteria when applied to a single dataset replicate from our simulation study. In the simulation study, we compare the performance of the SIC, scaled eBIC, and validation methods.

4 Simulation Studies

We conduct two simulation studies to compare the performance of our quantile trend filtering method and regularization parameter selection criteria with previously published

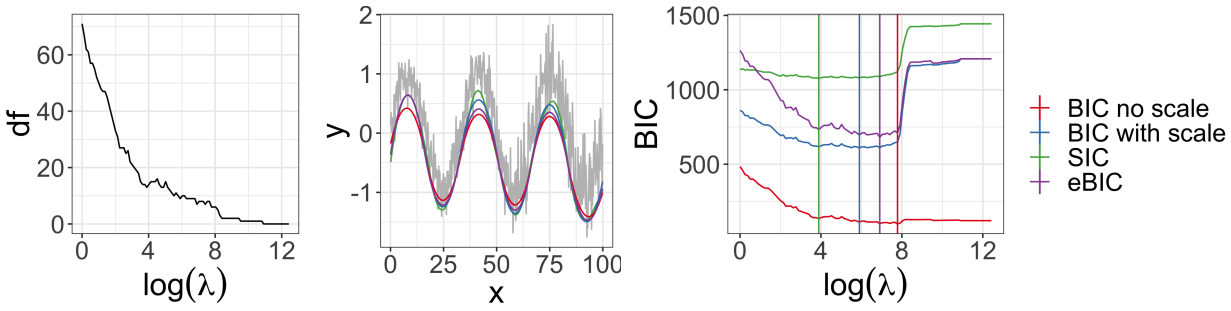


Figure 6: (Left) Estimated degrees of freedom (number of non-zero elements of $\mathbf{D}\boldsymbol{\theta}$) as a function of $\log(\lambda)$. (Middle) Estimated 10th quantile trend with regularization parameter chosen using various criteria. (Right) Criteria values as a function of $\log(\lambda)$, vertical lines indicate locations of minima.

methods. The first study compares the method's ability to estimate quantiles when the only components of the observed series are a smooth trend and a random component. The second study is based on our application and compares the method's ability to estimate baseline trends and enable peak detection when the time series contains a non-negative signal component in addition to the trend and random component.

We compare three criteria for choosing the smoothing parameter with our quantile trend filtering method with a single window: `detrendr_SIC` λ chosen using SIC (9) (Koenker et al., 1994); `detrendr_valid`: λ is chosen by leaving out every 5th observation as a validation data set and minimizing the check loss function evaluated at the validation data; `detrendr_eBIC`: the proposed scaled eBIC criteria (10).

We also compare the performance of our quantile trend filtering method with three previously published methods: `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); Nychka et al. (1995); `rqss` available in the `quantreg` package and described in Koenker et al. (1994). The regularization parameter λ for the `rqss` method is chosen using a grid search and minimizing the SIC criteria as described in Koenker et al. (1994), the regularization parameter for `qsreg` was chosen using generalized cross-validation based on the quantile criterion Oh et al. (2011).

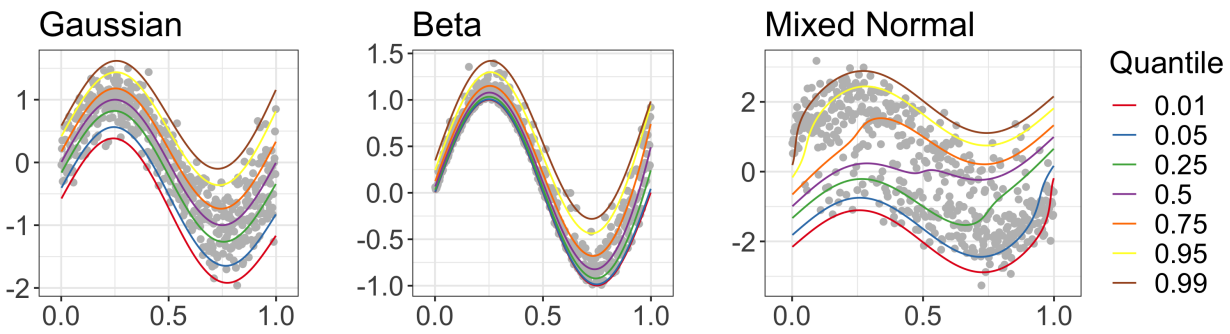


Figure 7: Simulated data with true quantile trends.

4.1 Estimating Quantiles

To compare performance in estimating quantile trends, three simulation designs from Racine and Li (2017) were considered. For all designs $t = 1, \ldots, n$, $x(t) = t/n$, and the response $\mathbf{y}$ was generated as

$$y(t) \quad = \quad \sin(2\pi x(t)) + \epsilon(x(t))$$

The three error distributions considered were

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

One hundred datasets were generated of sizes 300, 500 and 1000. For each method quantile trends were estimated for $\tau = \{0.05, 0.25, 0.5, 0.75, 0.95\}$. Only our detrend methods guarantee non-crossing quantiles. For each quantile trend and method the root mean squared error was calculated as RMSE $= \sqrt{\frac{1}{n}\sum_i (\hat{q}_\tau(t_i) - q_\tau(t_i))^2}$. Figure 8 shows the mean RMSE $\pm$ twice the standard error for each method, quantile level, and sample size. In all three designs the proposed detrend methods are either better than or comparable to existing methods. Overall the `detrend_eBIC` performs best, and especially in the mixed normal design our methods have lower RMSEs for the 5^{th} and 95^{th} quantiles. The `npqw`

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

4.2 Peak Detection

We use another simulation design based on our motivating application. We assume that the measured data can be represented by

$$y(t_i) \quad = \quad \theta(t_i) + s(t_i) + \varepsilon_i,$$

where $t_i = i$ for $i = 1, ..., n$, $\theta(t)$ is the drift component that varies smoothly over time, $s(t)$ is the true signal at time t, and ε_i are i.i.d. errors distributed as $N(0, 0.25^2)$. We generate $\theta(t)$ using cubic natural spline basis functions with degrees of freedom sampled from a Poisson distribution with mean parameter equal to $n/100$, and coefficients drawn from an exponential distribution with rate 1. The true signal function $s(t)$ is assumed to be zero with peaks generated using the Gaussian density function. The number of peaks is sampled from a binomial distribution with size equal to n and probability equal to 0.005 with location parameters uniformly distributed between 1 and $n-1$ and bandwidths uniformly distributed between 2 and 12. The simulated peaks were multiplied by a factor that was randomly drawn from a normal distribution with mean 20 and standard deviation of 4. One hundred datasets were generated for each $n = \{500, 1000, 2000, 4000\}$. We compare the ability of the methods to estimate the true quantiles of $y(t_i) - s(t_i)$ for $\tau \in \{0.01, 0.05, 0.1\}$ and calculate the RMSE (Figure 9). In this simulation study, our proposed method `detrend_eBIC` method substantially outperforms the others. The `qsreg` method is comparable to the `detrend_eBIC` method on the smaller datasets, but its performance deteriorates as the data size grows. The `npqw` and `detrend_valid` methods both perform poorly on this design.

While minimizing RMSE is desirable in general, in our application, the primary metric of success is accurately classifying observations y_i as signal or no signal. To evaluate the accuracy of our method compared to other methods we define true signal as any time point when the simulated peak value is greater than 0.5. We compare three different quantiles for the baseline estimation and four different thresholds for classifying the signal after

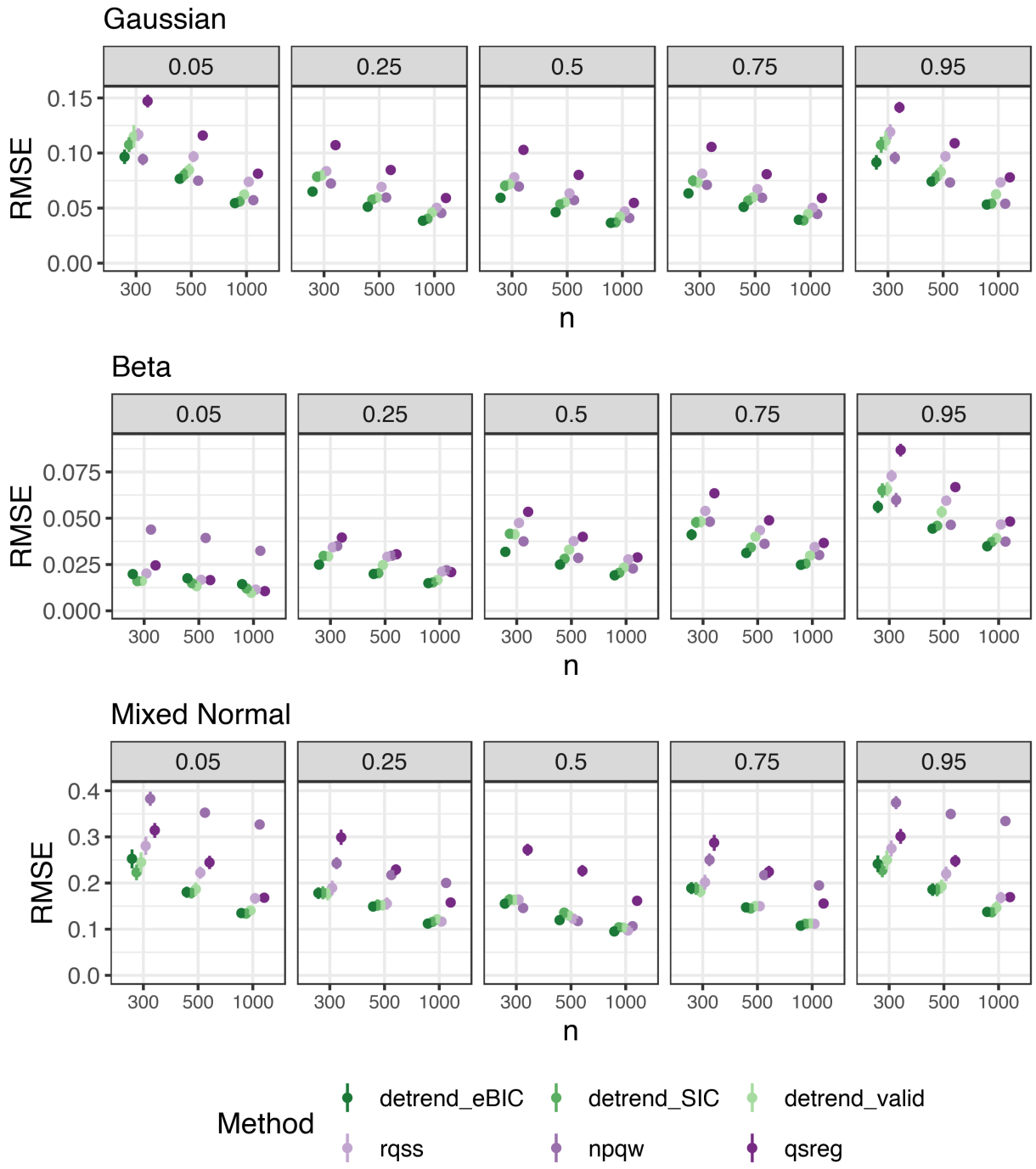


Figure 8: RMSE by design, method, quantile and data size. Points and error bars represent mean RMSE ± twice the standard error.

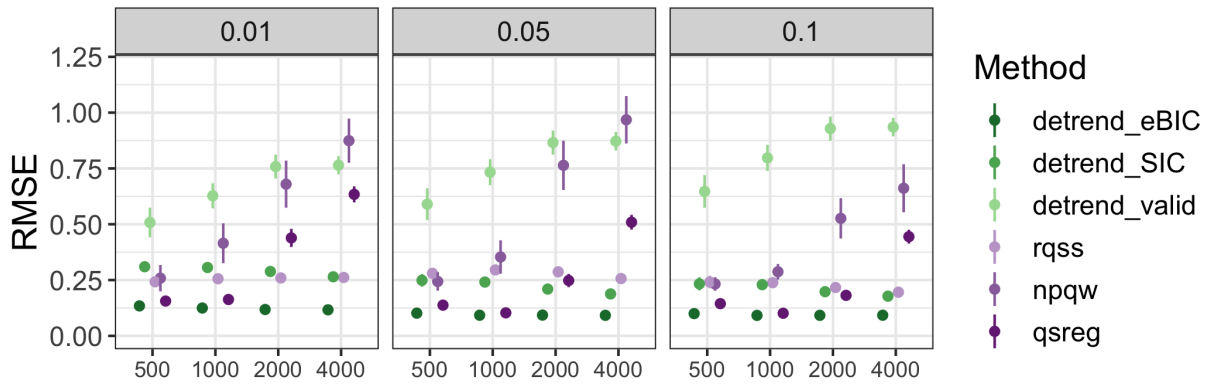


Figure 9: RMSE by method, quantile, and data size for peaks design.

subtracting the estimated baseline from the observations. Figure 10 shows an illustration of the observations classified as signal after subtracting the baseline trend compared to the "true signal." Let $\delta_i \in \{0, 1\}$ be the vector of true signal classifications and $\hat{\delta}_i \in \{0, 1\}$ be the vector of estimated signal classifications, namely $\hat{\delta}_i = \mathbb{1}(y_i - \hat{\theta}_i > 0.5)$.

To compare the resulting signal classifications, we calculate the class averaged accuracy (CAA), which is defined as

$$\text{CAA} \quad = \quad \frac{1}{2}\left[\frac{\sum_{i=1}^{n} \mathbb{1}[\delta_i = 1 \cap \hat{\delta}_i = 1]}{\sum_{i=1}^{n} \mathbb{1}[\delta_i = 1]} + \frac{\sum_{i=1}^{n} \mathbb{1}[\delta_i = 0 \cap \hat{\delta}_i = 0]}{\sum_{t=i=1}^{n} \mathbb{1}[\delta_i = 0]}\right].$$

We use this metric because our classes tend to be very imbalanced with many more zeros than ones. The CAA metric will give a score of 0.5 for random guessing and also for trivial classifiers such as $\hat{\delta}_i = 0$ for all i.

Our `detrend_BIC` method performs the best overall in terms of both RMSE and CAA. While `qsreg` was competitive with our method in some cases, in the majority of cases the largest CAA values for each threshold were produced using the `detrend_eBIC` method with the 1^{st} or 5^{th} quantiles.

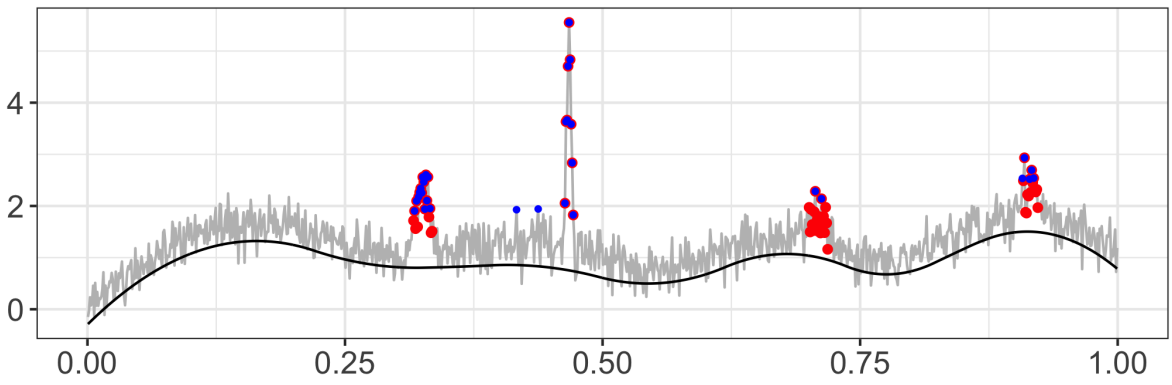


Figure 10: Example signal classification using threshold. Red indicates true signal > 0.5, blue indicates observations classified as signal after baseline removal using `detrend_eBIC` and a threshold of 1.2.

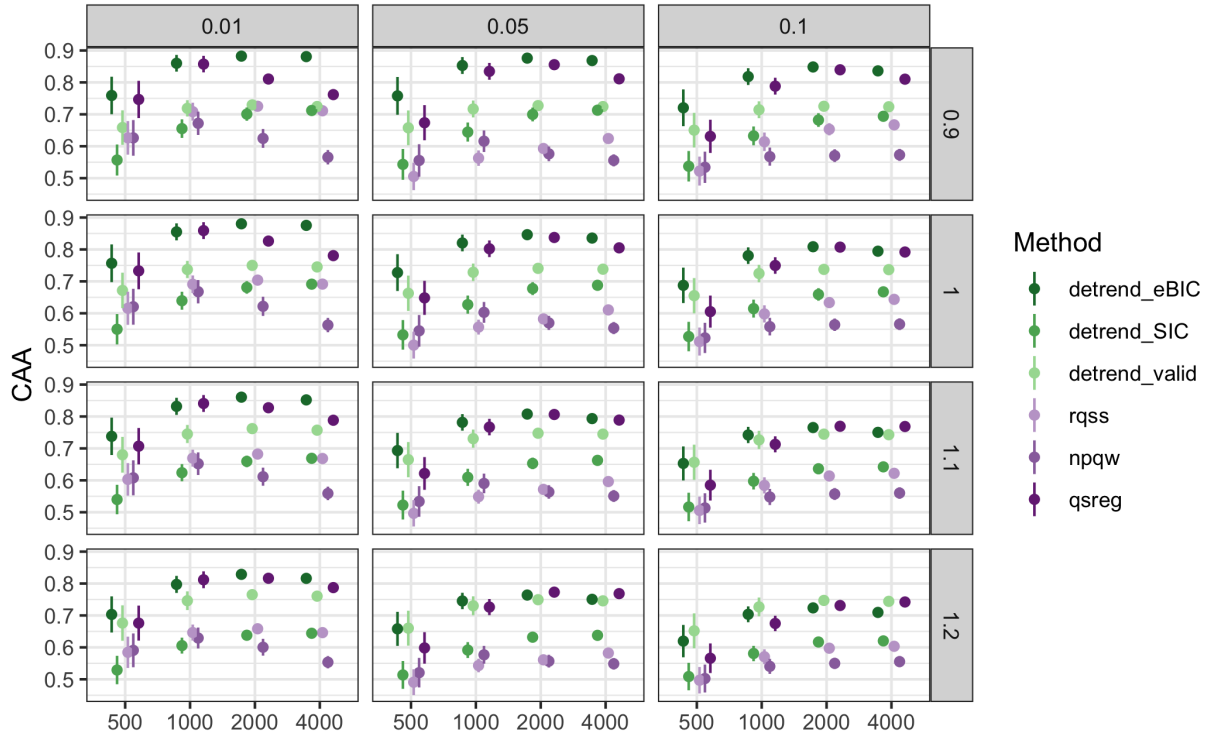


Figure 11: Class averaged accuracy by threshold, data size, and method (1 is best 0.5 is worst).

5 Application in Air Quality Monitoring

The low-cost "SPod" air quality sensors output a time series that includes a slowly varying baseline, the sensor response to pollutants, and high frequency random noise. These sensors are used to monitor pollutant concentrations at the fence lines of industrial facilities and detect time points when high concentrations are present. Ideally, three co-located and time aligned sensors (as shown in Figure 1) responding to a pollutant plume would result in the same signal classification after baseline trend removal and proper threshold choice.

We compare our `detrend_eBIC` method with the `qsreg` method on a subset of the SPod data (n=6000) since the `qsreg` method cannot handle all 24 hours simultaneously. We estimate the baseline trend using the 10^{th} and 15^{th} quantiles and compare three thresholds for classifying signal. The thresholds are calculated using the median plus a multiple of the median absolute deviation (10) of the detrended series.

$$\text{MAD} = \text{median}\|y - \tilde{y}\|, \tag{10}$$

where $\tilde{y}$ is the median of y. Given a method, quantile level, and MAD multiple, we estimate the quantile trend for each of the sensors and subtract it from the observations. We then calculate the threshold using the median plus the MAD multiple of the corrected series and classify the corrected series based on the threshold. Figure 12 shows an example of the estimated baseline fit for each method, while Figure 13 shows the series after subtracting the `detrend_eBIC` estimate of the 15^{th} quantile and classifying the signal using a MAD multiple of 3.

Given the signal classifications for SPods a and b, $\delta_i^{(a)} \in \{0,1\}$ and $\delta_i^{(b)} \in \{0,1\}$, we want to compare the similarity between the two classifications. One metric for evaluating the distance between two classifications is the variation of information (VI):

$$\begin{aligned} r_{ij} &= \frac{1}{n}\sum_t \mathbb{1}\left(\delta_i^{(a)} = i \cap \delta_i^{(b)} = j\right) \\ \text{VI}(s_a, s_b) &= -\sum_{i,j} r_{ij}\left[\log\left(\frac{r_{ij}}{\frac{1}{n}\sum_t \mathbb{1}(\delta_i^{(a)} = i)}\right) + \log\left(\frac{r_{ij}}{\frac{1}{n}\sum_t \mathbb{1}(\delta_i^{(b)} = j)}\right)\right] \end{aligned}$$

The VI is a distance metric for measuring similarity of classifications and will be 0 if the classifications are identical and increase as the classifications become more different.

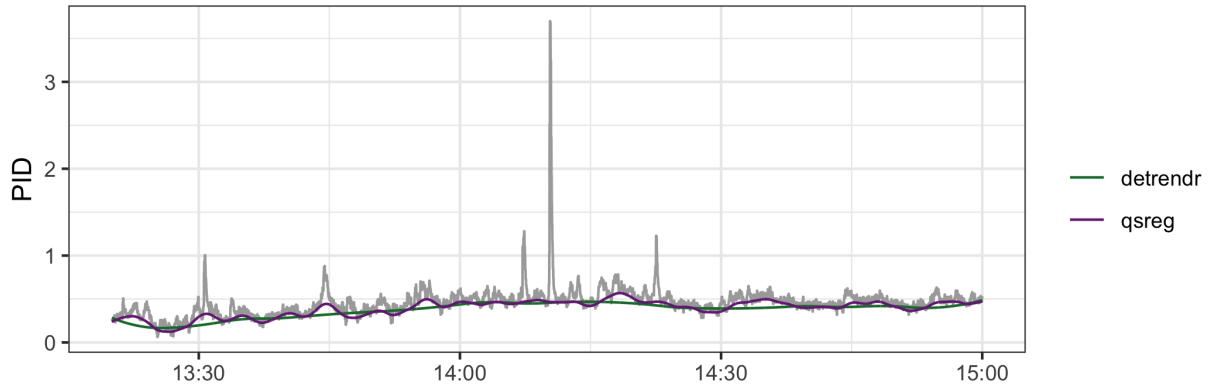


Figure 12: Estimated 15^{th} quantile trends on subset of the data using `qsreg` and `detrend_eBIC`.

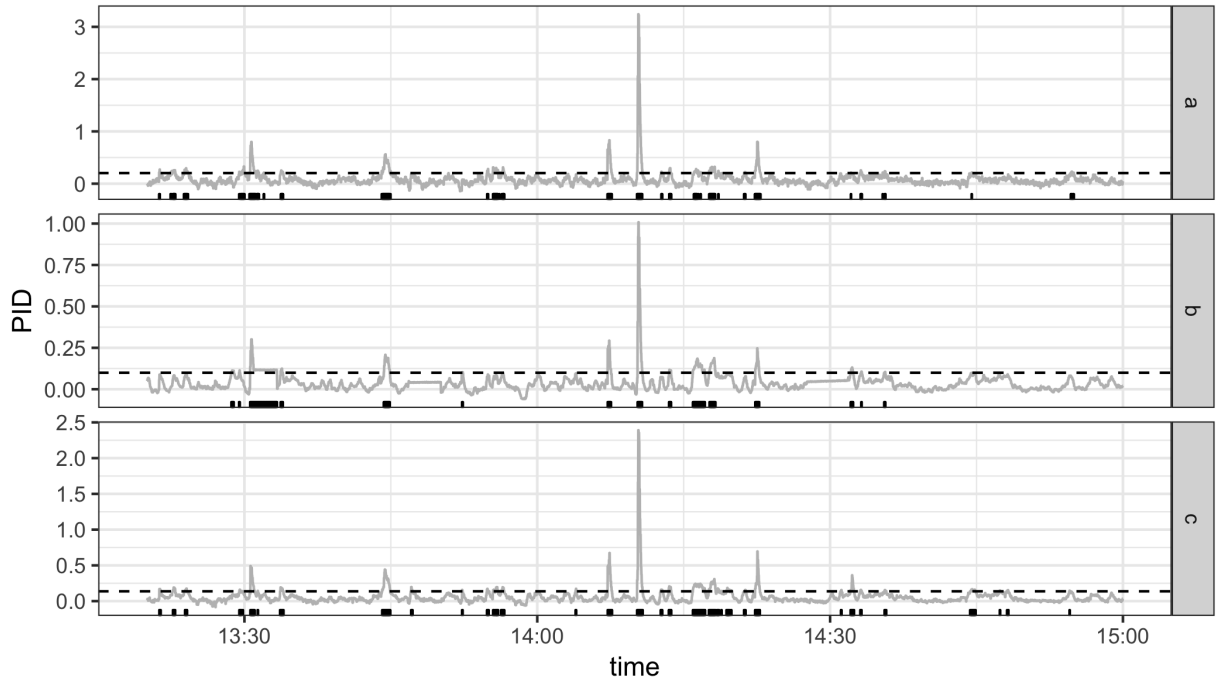


Figure 13: Rugplot showing locations of signal after baseline removal using detrendr estimate of 15th quantile. Horizontal dashed lines represent the thresholds.

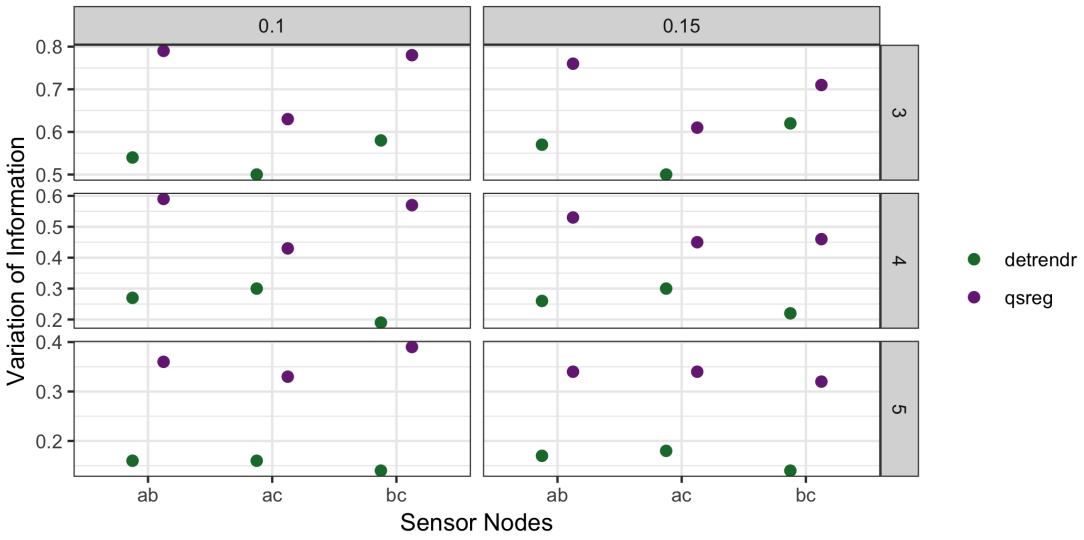


Figure 14: Variation of Information between sensors by method (color), quantile (columns) and threshold MAD multiple (rows).

	a = 0		**a = 1**	
	c = 0	c = 1	c = 0	c = 1
b = 0	80383	578	2551	145
b = 1	1298	737	111	598

Table 1: Confusion matrices for 3 SPods after baseline removal using 15th quantile and threshold of 5*MAD (n=86,401).

Figure 14 shows plots the VIs by method, quantile, and trend. In all cases, our `detrend` method results in classifications that are more similar than those from the `qsreg` method. Figure 12 and Figure 13 illustrate that our method results in a smoother baseline estimate which improves signal classification.

Our windowed `detrend_eBIC` method was used to removed the baseline drift from the total dataset consisting of 86,401 observations per node. The VI scores for the full dataset were 0.36, 0.24, and 0.41 for sensors a and b, a and c, and b and c, respectively. The complete confusion matrices for the nodes using a MAD multiple of 5 for the threshold is given in Table 1.

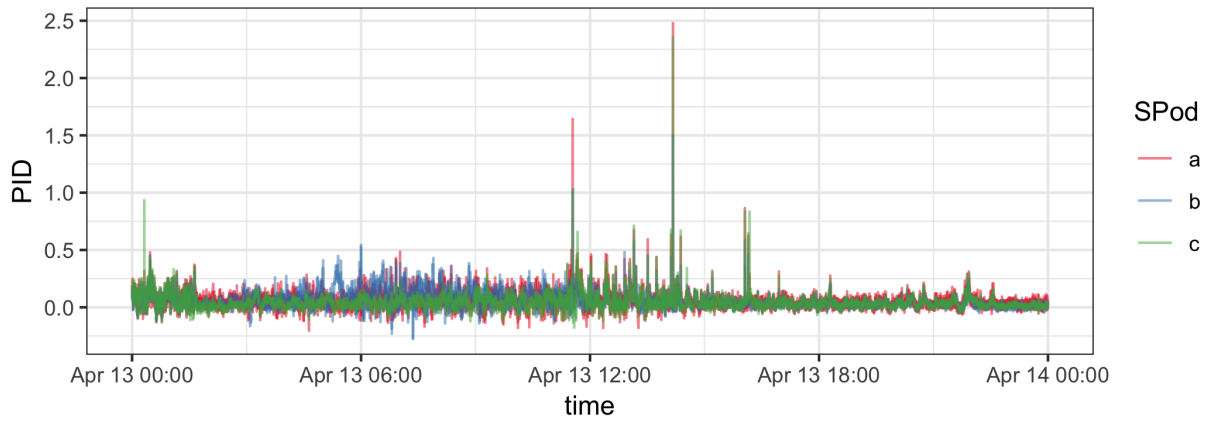


Figure 15: Low cost sensor data after drift removal using windowed detrend with eBIC.

6 Conclusion and Discussion

We have expanded the quantile trend filtering method by implementing a non-crossing constraint and a new algorithm for processing large series, and proposing a modified criteria for smoothing parameter selection. Furthermore we have demonstrated the utility of quantile trend filtering in both simulations and applied settings. Our ADMM algorithm for large series both reduces the computing time and allows trends to be estimated on series that cannot be estimated simultaneously while our scaled extended BIC criteria was shown to provided better estimated of quantile trends in series with and without a signal component. We have also shown that the baseline drift in low cost air quality sensors can be removed through estimating quantile trends.

In the future, quantile trend filtering could be extended to observations measured at non-uniform spacing by incorporating the distance in covariate spacing into the differencing matrix. It could also be extended to estimate smooth spatial trends by a similar adjustment to the differencing matrix based on spatial distances between observations.

SUPPLEMENTARY MATERIAL

R-package for detrend routine: R-package detrendr containing code to perform the diagnostic methods described in the article. (GNU zipped tar file)

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.

Brantley, H., Hagler, G., Kimbrough, E., Williams, R., Mukerjee, S., and Neas, L. (2014), "Mobile air monitoring data-processing strategies and effects on spatial air pollution trends," *Atmospheric measurement techniques*, 7, 2169–2183.

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

Gabay, D. and Mercier, B. (1975), *A dual algorithm for the solution of non linear variational problems via finite element approximation*, Institut de recherche d'informatique et d'automatique.

Glowinski, R. and Marroco, A. (1975), "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires," *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9, 41–76.

Gurobi Optimization, L. (2018), "Gurobi Optimizer Reference Manual," .

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.

Marandi, R. Z. and Sabzpoushan, S. (2015), "Qualitative modeling of the decision-making process using electrooculography," *Behavior research methods*, 47, 1404–1412.

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.

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

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.

Pettersson, K., Jagadeesan, S., Lukander, K., Henelius, A., Hæggström, E., and Müller, K. (2013), "Algorithm for automatic analysis of electro-oculographic data," *Biomedical engineering online*, 12, 110.

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

Rudin, L. I., Osher, S., and Fatemi, E. (1992), "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, 60, 259 – 268.

Snyder, E., Watkins, T., Solomon, P., Thoma, E., Williams, R., Hagler, G., Shelow, D., Hindin, D., Kilaru, V., and Preuss, P. (2013), "The changing paradigm of air pollution monitoring." *Environmental science & technology*, 47, 11369.

Theussl, S. and Hornik, K. (2017), *Rglpk: R/GNU Linear Programming Kit Interface*, r package version 0.6-3.

Thoma, E. D., Brantley, H. L., Oliver, K. D., Whitaker, D. A., Mukerjee, S., Mitchell, B., Wu, T., Squier, B., Escobar, E., Cousett, T. A., et al. (2016), "South Philadelphia passive sampler and sensor study," *Journal of the Air & Waste Management Association*, 66, 959–970.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 91–108.

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

Weingessel, A. and Turlach, B. A. (2013), "quadprog: Functions to solve Quadratic Programming Problems." R package version 1.5-5.

Yamada, H. (2017), "Estimating the trend in US real GDP using the ℓ_1 trend filtering," *Applied Economics Letters*, 24, 713–716.

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