$a(t) + b(t)\epsilon_i$ and estimate the τ_{th} conditional quantile given $t = t_0$ using a kernel estimator and local linear approach.

We propose to use the trend filtering penalty combined with the check loss function to produce quantile trends that are piecewise quadratic. 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 propose 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.

2 Methods

2.1 Quantile Trend Filtering

We combine the ideas of quantile regression and trend filtering. For a single quantile level τ the estimation of the quantile trend filtering model can be posed as the following optimization problem.

maybe this will go in the intro?

$$\min_{\theta} \rho_\tau(y - \theta) + \lambda \|\mathbf{D}^{(k+1)}\theta\|_1, \tag{5}$$

where λ is a non-negative regularization parameter. We address the problem of choosing λ in Section 2.3. As with classic quantile regression, the quantile trend filtering problem is a linear program which can be solved by a number of methods ~~free or commercial solvers~~. 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 $Q(\tau_2) \geq Q(\tau_1)$. Given quantiles $\{\tau_1, ..., \tau_J\}$ such that $\tau_1 < \tau_2 < ... < \tau_J$, the optimization problem becomes

$$\min_{\theta_1,\ldots,\theta_J} \sum_{j=1}^{J} \left[\rho_{\tau_j}(y-\theta_j) + \lambda_j \|\mathbf{D}^{(k+1)}\theta_j\|_1\right] \tag{6}$$

$$\text{subject to: } \theta_1(t) \le \theta_2(t) \le \ldots \le \theta_J(t) \text{ for all } t, \tag{7}$$

← where $\theta_j \in \mathcal{R}^n$. 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 rely on the commercial solver Gurobi (Gurobi Optimization, 2018) and its R package implementation. However, we could ~~but it can~~ easily ~~be~~ substituted for a free solver such as the Rglpk package by Theussl and Hornik (2017). 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, all solvers will eventually break.

maybe move here

2.2 ADMM for Big Data

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 (Gabay and Mercier, 1975; Glowinski and Marroco, 1975) is fully described by Boyd et al. (2011). We apply the consensus ADMM algorithm to the ~~the~~ quantile regression trend filtering problem given in Eq. (5) by dividing our observed series $y(t)$ with $t = \{1, ..., n\}$ into overlapping windows

$$\begin{aligned} y_1(t) &= y(t) \quad \text{if} \quad 1 \le t \le u_1 \\ y_2(t) &= y(t) \quad \text{if} \quad l_2 \le t \le u_2 \\ &\cdots \\ y_M(t) &= y(t) \quad \text{if} \quad l_M \le t \le n \end{aligned}$$

← with boundaries $1 < l_2 < u_1 < l_3 < u_2 < l_4 < u_3 < ... < n$. An illustration is given in Fig. 3. Given quantiles $\tau_1 < ... < \tau_J$, ~~to be estimated~~ we define $\theta_{j,m}(t)$ as the value of the ${\tau_j}^{\text{th}}$ quantile trend in window M m at time ~~point~~ t. In order to write out the constraint that the overlapping sections must be equal we define a consensus variable

$$\cancel{\bar{\theta}_{j,m} = g(\theta_{j,m-1}, \theta_{j,m}, \theta_{j,m+1})}$$

$$\bar{\theta}_{j,m}(t) = \begin{cases} \frac{\theta_{j,m-1}(t)+\theta_{j,m}(t)}{2} & \text{if } \; l_m \leq t \leq u_{m-1} \\ \theta_{j,m}(t) & \text{if } \; u_{m-1} \leq t \leq l_{m+1} \\ \frac{\theta_{j,m}(t)+\theta_{j,m+1}(t)}{2} & \text{if } \; l_{m+1} \leq t \leq u_m \end{cases},$$

Is there only one $\bar{\theta}$ for each j and t? In other words, can we write this as $\bar{\theta}_j(t)$?

defining $\theta_{j,M+1} = \theta_{j,M}$ and $\theta_{j,0} = \theta_{j,1}$. Our windowed quantile trend optimization problem can then be written as

$$\sum_{m=1}^{M} \min_{\theta_{1,m},\ldots,\theta_{J,m}} \sum_{j=1}^{J} \left[\rho_{\tau_j}(y_m - \theta_{j,m}) + \lambda_j \|\mathbf{D}^{(k+1)}\theta_{j,m}\|_1\right]$$

subject to:

$$\theta_{1,m}(t) \leq \theta_{2,m}(t) \leq ... \leq \theta_{J,m}(t) \text{ for all } m, t$$

$$\theta_{j,m}(t) = \bar{\theta}_{j,m}(t) \text{ for all } j, m, t$$

Defining Lagrange multipliers $\omega_{j,m}$, and penalty parameter $\gamma > 0$, ~~we can write~~ the augmented Lagrangian for finding the trends in window m is

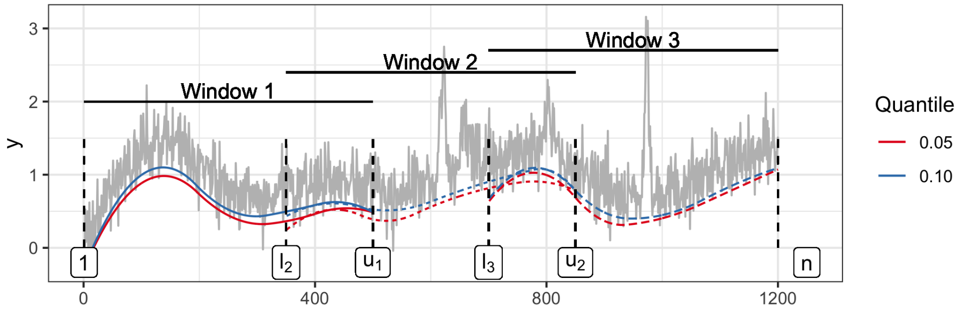
$$\mathcal{L}(\theta_{j,m}, \bar{\theta}_{j,m}, \omega_{j,m}) = \sum_{j=1}^{J} \rho_{\tau_j}(y_m - \theta_{j,m}) + \lambda \|\mathbf{D}^{(k+1)}\theta_{j,m}\|_1 + \omega_{j,m}^T(\theta_{j,m} - \bar{\theta}_{j,m}) + \frac{\gamma}{2}||\theta_{j,m} - \bar{\theta}_{j,m}||_2^2.$$

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.

We measure convergence 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 the primal and dual feasibility, respectively. The primal residual,

$$r_p^{(q)} = \sqrt{\sum_{m=1}^{M} \sum_{j=1}^{J} ||\theta_{j,m}^{(q)} - \bar{\theta}_{j,m}^{(q)}||_2^2}, \tag{8}$$

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



Algorithm 1 ADMM algorithm for quantile trend filtering with windows

```
Define $D = D^{(k+1)}$.
**initialize:**
$\theta^{(0)}_{j,m} = \arg\min \sum_{j=1}^{J} \rho_{\tau_j}(y_m - \theta_{j,m}) + \lambda \|D\theta_{j,m}\|_1$ subject to $\theta_{1,m}(t) < ... < \theta_{J,m}(t)$ for all $t$.
$\omega^{(0)}_{j,m} = \mathbf{0}$
**repeat**
    $\bar{\theta}^{(q)}_{j,m} = g(\theta^{(q-1)}_{j,m-1}, \theta^{(q-1)}_{j,m}, \theta^{(q-1)}_{j,m+1})$
    $\omega^{(q)}_{j,m} = \omega^{(q-1)}_{j,m} + \gamma(\theta^{(q-1)}_{j,m} - \bar{\theta}^{(q)}_{j,m})$
    $\theta^{(q)}_{j,m} = \arg\min \mathcal{L}(\theta_{j,m}, \bar{\theta}^{(q-1)}_{j,m}, \omega^{(q-1)}_{j,m})$ subject to $\theta_{1,m}(t) < ... < \theta_{J,m}(t)$ for all $t$.
**until** convergence
**return** Non-overlapping sequence of $\bar{\theta}^{(q)}_{j,m}$ for all $j$, $m$.
```

represents the difference between the trend values in the windows and the consensus trend value while the dual residual,

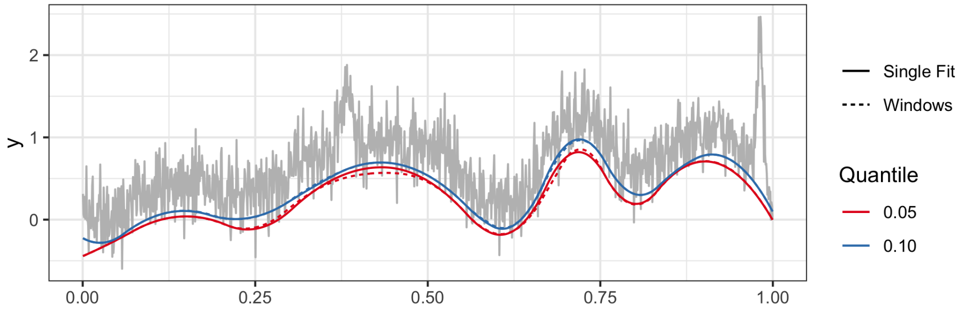
$$r_d^{(q)} = \gamma \sqrt{\sum_{m=1}^{M} \sum_{j=1}^{J} \|\bar{\theta}^{(q)}_{j,m} - \bar{\theta}^{(q-1)}_{j,m}\|_2^2},$$

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

$$r_p^{(q)} < \epsilon_{abs}\sqrt{nJ} + \epsilon_{rel} \max_m \left[\max \left(\sqrt{\sum_{j=1}^{J} \|\theta^{(q)}_{j,m}\|_2^2}, \sqrt{\sum_{j=1}^{J} \|\bar{\theta}^{(q)}_{j,m}\|_2^2} \right) \right] \quad (9)$$

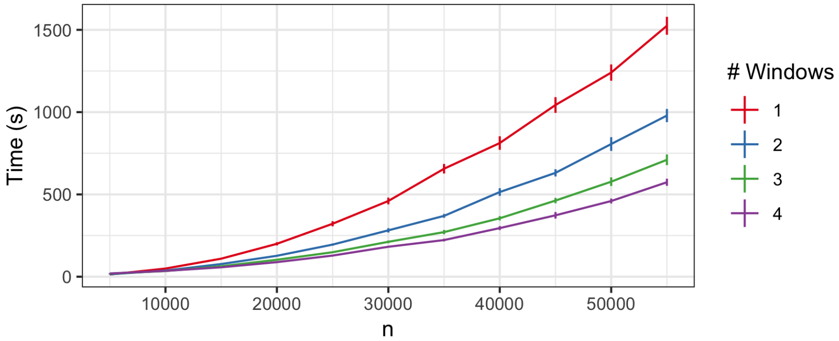
$$r_d^{(q)} < \epsilon_{abs}\sqrt{nJ} + \epsilon_{rel} \sqrt{\sum_{m=1}^{M} \sum_{j=1}^{J} \|\omega^{(q)}_{j,m}\|_2^2} \quad (10)$$

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



Timing experiments illustrate the advantages of using our ADMM algorithm even on datasets where solving the problem simultaneously is possible. For each data size, n, 25 datasets were simulated using the peaks simulation design described below. ~~and~~ T trends for three quantiles were fit simultaneously: 0.05, 0.1, and 0.15 using a $\lambda = n/5$. 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 $\epsilon_{rel} = 0.001$. As is shown in Fig. 5, 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 ... *

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



Information criteria and Cross-validation are commonly used procedures for the selection of smoothing parameters.

2.3 Regularization Parameter Choice

An important problem in trend estimation is the choice of regularization parameter, or degree of smoothness. Our method can easily handle missing data by defining the check loss function to output 0 for missing values, ~~This~~ allow~~s~~ing us to implement cross-validation, ~~leave out validation observations that can be used to select the tuning parameter λ~~. ~~However, the use of an information criteria metric can result in a better choice of regularization parameter than the validation method~~. Koenker et al. (1994) addressed the choice of regularization parameter by proposing the Schwarz criterion for the selection of λ

We also explore two (?) information criteria.

$$\mathrm{SIC}(p_\lambda) = \log\left[\frac{1}{n}\rho_\tau(y-\theta)\right] + \frac{1}{2n}p_\lambda \log n. \tag{11}$$

where $p_\lambda = \sum_t I(y(t) = \widehat{\theta}(t))$ 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

$$\mathrm{BIC}(s) = -2\log(L\{\hat{\theta}\}) + \nu \log n \tag{12}$$

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

$$L(y|\theta) = \left(\frac{\tau^n(1-\tau)}{\sigma}\right)^n \exp\left\{-\sum_t \rho_\tau\left(\frac{y(t)-\theta(t)}{\sigma}\right)\right\}, \tag{13}$$

we can compute the BIC as

be consistent

bold?

$$\mathrm{BIC}(\boxed{df}) = 2\frac{1}{\sigma}\rho_\tau(y-\hat{\theta}) + \boxed{\mathrm{df}}\log n \tag{14}$$

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

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

$$\mathrm{eBIC}_\gamma(s) = -2\log(L\{\hat{\theta}\}) + \nu\log n + 2\gamma\log\binom{P}{\nu}, \quad \gamma \in [0,1] \tag{15}$$