Quantile Trend Filtering via Spingarn's Method
Date: May 17, 2017

1 Introduction

We introduce a linear time algorithm for removing trends in time series.

Examples:

- Air quality
- ECG [9]
- Chromatogram baseline estimation [7, 5]
- Galaxy spectrum baseline estimation [5, 1]

Need to decide between detrending versus detrending + denoising. The BEADS [7] 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 [3, 4, 2]
- 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 $\tau \in (0, 1)$ as follows. 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 [6, 11], 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

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

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.

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

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 Spingarn's method of partial inverses

We first review Spingarn's method [10], which solves the following equality constrained convex problem:

$$\begin{aligned} &\text{minimize } \psi(\mathbf{x}) \\ &\text{subject to } \mathbf{x} \in V, \end{aligned} \tag{2}$$

where V is a subspace. The problem 2 can be expressed as the unconstrained optimization problem

$$\text{minimize } \psi(\mathbf{x}) + \iota_V(\mathbf{x}), \tag{3}$$

where ι_V is the indicator function of the set V. Spingarn's method applies Douglas-Rachford splitting to the problem (3) to give the following updates.

$$\begin{aligned} \mathbf{x}^+ &= \text{prox}_{t\psi}(\mathbf{z}) \\ \mathbf{y}^+ &= P_V(2\mathbf{x}^+ - \mathbf{z}) \\ \mathbf{z}^+ &= \mathbf{z} + \mathbf{y}^+ - \mathbf{x}^+. \end{aligned}$$

The parameter t is a step-size which can be fixed at 1 to guarantee convergence. The mapping P_V is the orthogonal projection onto the set V.

6 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 (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 should also consider trying the reparameterization for ADMM trend filtering proposed in [8].

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

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

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

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

The stationary condition is

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

Therefore, the proximal mapping is given by

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

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 $\text{prox}_{t\psi}(\boldsymbol{\theta}, \boldsymbol{\eta})$
- $\mathcal{O}((k+1)(n-k+1))$ 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 $\mathcal{O}(nk)$.

6.1 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 to reduce the number of Spingarin updates, since the Douglas-Rachford algorithm is a fixed point algorithm.

References

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

[2] Damek Davis. Convergence rate analysis of the forward-Douglas-Rachford splitting scheme. *SIAM Journal on Optimization*. in press.

[3] Bingsheng He and Xiaoming Yuan. On the $O(1/n)$ Convergence Rate of the Douglas-Rachford Alternating Direction Method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[4] Bingsheng He and Xiaoming Yuan. On the convergence rate of Douglas-Rachford operator splitting method. *Mathematical Programming*, 153(2):715–722, 2015.

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

[6] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009.

[7] Xiaoran Ning, Ivan W. Selesnick, and Laurent Duval. Chromatogram baseline estimation and denoising using sparsity (beads). *Chemometrics and Intelligent Laboratory Systems*, 139:156 – 167, 2014.

[8] Aaditya Ramdas and Ryan J. Tibshirani. Fast and Flexible ADMM Algorithms for Trend Filtering. *Journal of Computational and Graphical Statistics*, 0(ja):0–0, 2016.

[9] A. Sanyal, A. Baral, and A. Lahiri. Application of S-transform for removing baseline drift from ECG. In *2012 2nd National Conference on Computational Intelligence and Signal Processing (CISP)*, pages 153–157, March 2012.

[10] Jonathan E. Spingarn. Applications of the method of partial inverses to convex programming: Decomposition. *Mathematical Programming*, 32(2):199–223, 1985.

[11] Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 02 2014.