Overview

1. (The $k^{th}$ Order) Trend Filtering
2. Comparison to smoothing spline
3. Comparison to locally adaptive regression spline
4. Rate of Convergence
5. Extensions
(The $k^{th}$ Order) Trend Filtering
Motivation

$l_1$ Trend Filtering (Kim, 2009):

- An $l_1$ filtering or smoothing method for trend estimation in time series data.
- Suited to analyze time series with an underlying piecewise linear trend.
- A special type of basis pursuit problem.
Prior Works in Nonparametric Regression:

- **Smoothing Splines:** Not locally adaptive.
- **Locally Adaptive Regression Spline:** Computational expensive ($O(n^3)$)
Usual Setup in Nonparametric Regression:
Assume $n$ observations $y_1, \ldots, y_n \in R$ and $n$ input points $x_1, x_2, \ldots, x_n \in R$ from the model:

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where $f_0$ is the underlying function and $\epsilon_1, \ldots, \epsilon_n$ are independent.

Further Setup Here:
Assume the $n$ input points are ordered and evenly spaced over $[0,1]$, i.e., $x_i = i/n$ for $i = 1, \ldots, n$.
The $k^{th}$ order trend filtering estimate $\hat{\beta} = (\hat{f}_0(x_1), \cdots, \hat{f}_0(x_n))$ is defined as the following:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^n} \frac{1}{2} \| y - \beta \|_2^2 + \frac{n^k}{k!} \lambda \| D^{(k+1)} \beta \|_1,$$

(2)

where $y = (y_1, \cdots, y_n)^T$ and $D^{(k+1)} \in \mathbb{R}^{(n-k) \times n}$ is the discrete difference operator of order $k + 1$ defined in the next slide.

**Notice:** Trend filtering estimators are ONLY defined over the discrete set of inputs.
The discrete difference operator $D^{(k+1)}$ is defined recursively as:

$$D^{(k+1)} = D^{(1)} \cdot D^{(k)} \in \mathbb{R}^{(n-k) \times n} \quad (3)$$

where $D^{(1)}$ is defined as:

$$D^{(1)} = 
\begin{bmatrix}
-1 & 1 & 0 & \cdots & 0 & 0 \\
0 & -1 & 1 & \cdots & 0 & 0 \\
\vdots & & & & & \\
0 & 0 & 0 & \cdots & -1 & 1
\end{bmatrix} \in \mathbb{R}^{(n-k-1) \times (n-k)} \quad (4)$$
More discrete difference operators:

\[ D^{(2)} = \begin{bmatrix}
1 & -2 & 1 & 0 & \cdots & 0 \\
0 & 1 & -2 & 1 & \cdots & 0 \\
0 & 0 & 1 & -2 & \cdots & 0 \\
\vdots & & & & & \\
\end{bmatrix} \]  \quad (5)

\[ D^{(3)} = \begin{bmatrix}
-1 & 3 & -3 & 1 & \cdots & 0 \\
0 & -1 & 3 & -3 & \cdots & 0 \\
0 & 0 & -1 & 3 & \cdots & 0 \\
\vdots & & & & & \\
\end{bmatrix} \]  \quad (6)
Linear interpolated trend filtering examples for constant, linear and quadratic orders ($k=0, 1, 2$, respectively)
Inference

The inference in continuous domain for trend filtering lies in its equivalence at the input points to the lasso problem:

\[
\hat{\alpha} = \text{arg} \min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \| y - H\alpha \|_2^2 + \lambda \sum_{j=k+2}^{n} |\alpha_j| \quad (7)
\]

The solutions satisfy \( \hat{\beta} = H\hat{\alpha} \), where \( H \in \mathbb{R}^{n \times n} \) is a basis matrix, \( H_{ij} = h_j(x_i), i, j = 1, \cdots, n \),

\[
h_j(x) = \prod_{l=1}^{j-1} (x - x_l), j = 1, \cdots, k + 1, \quad (8)
\]

\[
h_{k+1+j}(x) = \prod_{l=1}^{k} (x - x_{j+l}) \cdot 1\{x \geq x_{j+k}\}, j = 1, \cdots, n - k - 1.
\]
Properties

- **Recursive Decomposition:** For $k \geq 1$,

\[
H^{(k)} = H^{(k-1)} \cdot \begin{bmatrix} I_k & 0 \\ 0 & \frac{k}{n} L_{n-k} \end{bmatrix}
\]

(9)

where $L_{n-k}$ denotes the $(n - k) \times (n - k)$ lower triangular matrix of 1s.

- **Inverse Basis:**

\[
(H^{(k)})^{-1} = \begin{bmatrix} C \\ \frac{1}{k!} \cdot D^{(k+1)} \end{bmatrix}
\]

(10)

It shows that the last $n-k-1$ rows of $(H^{(k)})^{-1}$ are given exactly by $D^{(k+1)}/k!$
Other Properties:

- Efficient Computation – $O(n^{1.5})$
- Locally Adaptive Polynomial Approximation
- Minimax Convergence Rate
Comparison to smoothing spline
The kth (k is an odd number) order smoothing spline estimate is defined as

\[
\hat{f} = \arg\min_{f \in W_{(k+1)/2}} \sum_{i=1}^{n} \|y_i - f(x_i)\|^2_2 + \lambda \int_{0}^{1} (f^{(k+1)/2}(t))^2 dt,
\]

(11)

where \(f^{(k+1)/2}(t)\) is the derivative of \(f\) of order \((k + 1)/2\), \(\lambda \geq 0\) is a tuning parameter, and the domain of minimization here is Sobolev space \(W_{(k+1)/2} = \{ f : [0, 1] \rightarrow R : f \text{ is } (k+1)/2 \text{ times differentiable, and } \int_{0}^{1} (f^{(k+1)/2}(t))^2 dt < \infty \}\).
It can be shown that the infinite-dimensional problem (11) has a unique minimizer [see, e.g., Wahba (1990)] and the minimizer is linear combination of \( n \) basis function. Hence to solve problem (11), we can solve for coefficients \( \theta \in \mathbb{R}^n \) in this basis expansion:

\[
\hat{\theta} = \text{argmin}_{\theta \in \mathbb{R}^n} \| y - N\theta \|_2^2 + \lambda \theta^T \Omega \theta,
\]  

(12)

If \( \eta_1, \cdots, \eta_n \) denotes a collection of basis functions for the set of \( k \)th degrees natural splines with knots \( x_1, \cdots, x_n \), then

\[
N_{ij} = \eta_j(x_i) \text{ and } \Omega_{ij} = \int_{0}^{1} \eta_i^{(k+1)/2}(t) \eta_j^{(k+1)/2}(t) dt \text{ for all } i, j
\]  

(13)
The solution to problem (11) at given input points $x_1, \cdots, x_n$ and the solution to problem (12) are connected by

$$\left( \hat{f}(x_1), \cdots, \hat{f}(x_n) \right) = N\hat{\theta}$$

(14)

More generally,

$$\hat{f}(x) = \sum_{j=1}^{n} \hat{\theta}_j \eta_j(x).$$

(15)
Generalized ridge representation

To compare smoothing spline with trend filtering, we rewrite the smoothing spline fitted values as:

\[ N\hat{\theta} = N(N^T + \lambda \Omega)^{-1}N^T y \]
\[ = N(N^T(I + \lambda N^{-T} \Omega N^{-1})N)^{-1}N^T y \]
\[ = (I + \lambda K)^{-1} y \]  \hspace{1cm} (16)

where \( K = N^{-T} \Omega N^{-1} \). Then \( \hat{u} = N\hat{\theta} \) is solution of the minimization problem

\[ \hat{u} = \arg\min_{u \in \mathbb{R}^n} \|y - u\|_2^2 + \lambda u^T Ku \]
\[ = \arg\min_{u \in \mathbb{R}^n} \|y - u\|_2^2 + \lambda \|K^{1/2}u\|_2^2 \] \hspace{1cm} (17)
Empirical comparison

The form of problem (17) is similar to trend filtering and there are two differences:

- $K^{1/2} u$ is similar to $D^{(k)} u$ but strictly different. For example, for $k = 3$ and input points $x_i = \frac{i}{n}$, it can be shown that $K^{1/2} u = C^{-1/2} D^{(2)} u$ where $D^{(2)}$ is second order derivative operator, can $C \in R^{n \times n}$ is a tridiagonal matrix.

- Smoothing spline utilizes $l_2$ penalty while trend filtering uses $l_1$ penalty. Thus later one shrinks some components of $D\hat{u}$ to zero, which therefore exhibits a finer degree of local adaptivity.
Empirical comparison

- **True function**

- **Trend filtering, df=19**

- **Smoothing spline, df=19**

- **Smoothing spline, df=30**
Empirical comparison

**True function**

**Trend filtering, df=50**

**Smoothing spline, df=50**

**Smoothing spline, df=90**
Empirical comparison

Hills example

Doppler example

Squared error

Trend filtering

Smoothing splines

Degrees of freedom

Degrees of freedom
By choosing B-spline basis functions, the matrix $N^T N + \lambda \Omega$ is banded, and so the smoothing spline fitted values can be computed in $O(n)$ operations.

Primal-dual interior point method is one option to solve trend filtering problem with fixed value of $\lambda$. This algorithm solves a sequence of banded linear system and the worst number of iterations scales as $O(n^{1/2})$. Hence interior point method is in $O(n^{3/2})$ worst-case complexity.

The dual path algorithm of Tibshirani & Taylor (2011) constructs solution path as $\lambda$ varies from $\infty$ to 0. The computation requires $O(n)$ operations.
Comparison to locally adaptive regression spline
Locally adaptive regression spline

Given arbitrary integer $k$, we first define the knot superset

$$
T = \begin{cases} 
\{x_{k/2+2}, \cdots , x_{n-k/2}\} & \text{if } k \text{ is even}, \\
\{x_{(k+1)/2+1}, \cdots , x_{n-(k+1)/2}\} & \text{if } k \text{ is odd}. 
\end{cases} \quad (18)
$$

which excludes the points near boundaries of inputs $\{x_1, \cdots , x_n\}$. We then define the $k$th order locally adaptive regression spline estimate as

$$
\hat{f} = \arg\min_{f \in \mathcal{G}_k} \frac{1}{2} \sum_{i=1}^{n} \|y_i - f(x_i)\|_2^2 + \lambda \text{TV}(f^{(k)}) \quad (19)
$$

where $f^{(k)}$ is now the $k$th weak derivative of $f$, $\text{TV}(\cdot)$ denotes the total variation operator.
$G_k$ is the set

$$G_k = \{ f : [0,1] \to R : f \text{ is kth degree spline with knots contained in } T \}$$

Total variation of a function $f : [0,1] \to R$ is defined as:

$$TV(f) = \sup \left\{ \sum_{i=1}^{p} |f(z_{i+1}) - f(z_i)| : z_1 < \cdots < z_p \text{ is partition of } [0,1] \right\},$$

and this reduces to $TV(f) = \int_0^1 |f'(t)| dt$ if $f$ is (strongly) differentiable.
$G_k$ is spanned by $n$ basis function $\{g_1, \cdots, g_n\}$. Each $g_j$ is $k$th degree spline with knots in $T$, we know that its $k$th weak derivative is piecewise constant and right-continuous, with jump point contained in $T$; therefore, writing $t_0 = 0$ and $T = \{t_1, \cdots, t_{n-k-1}\}$, we have

$$TV(g_j) = \sum_{i=1}^{n-k-1} |g_j^{(k)}(t_i) - g_j^{(k)}(t_{i-1})|. \quad (22)$$

Similarly, any linear combination of $g_1, \cdots, g_n$ has total variation:

$$TV\left(\sum_{j=1}^{n} \theta_j g_j\right) = \sum_{i=1}^{n-k-1} \left| \sum_{i=1}^{n-k-1} \left[ g_j^{(k)}(t_i) - g_j^{(k)}(t_{i-1}) \right] \cdot \theta_j \right|. \quad (23)$$
Hence problem (19) can be expressed in terms of $\theta \in \mathbb{R}^n$, 

$$
\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^n} \frac{1}{2} \| y - G\theta \|^2_2 + \lambda \| C\theta \|_1,
$$

(24)

where

$$
G_{ij} = g_j(t_i) \quad \text{for } i, j = 1, \ldots, n,
$$

(25)

$$
C_{ij} = g_j^{(k)}(t_i) - g_j^{(k)}(t_{i-1}) \quad \text{for } i = 1, \ldots, n - k - 1, j = 1, \ldots, n
$$

(26)
Given $\hat{\theta}$, the estimates of the locally adaptive spline over the input points are given by:

$$(\hat{f}(x_1), \cdots, \hat{f}(x_n)) = G\hat{\theta}$$

(27)

or, at an arbitrary point $x \in [0, 1]$ by

$$\hat{f}(x) = \sum_{j=1}^{n} \hat{\theta}_j g_j(x).$$

(28)

By taking $g_1, \cdots, g_n$ to be truncated power basis, we can turn (a block of) $C$ into identity, and problem (24) into a lasso problem.
When introducing trend filtering, we showed that trend filtering problem can be written as a lasso problem with design matrix $H$. $H = G$ for $k < 2$.

Although $G \neq H$ for $k \geq 2$, the estimates of two methods are practically similar and difficult to distinguish by eyes.
The difference of the basis functions:
The kth order truncated power basis is given by:

\[ g_1(x) = 1, \quad g_2(x), \ldots, g_{k+1}(x) = x^k, \]
\[ g_{k+1+j} = (x - t_j)^k \cdot 1\{x \geq t_j\}, j = 1, \ldots, n - k - 1. \] (29)
Empirical comparison

Truncated power basis

Trend filtering basis

Zoomed in TF basis
There is no specialized method for the locally adaptive regression spline.

Choosing either B-spline or truncated power basis, we are more or less stuck with solving a generalized lasso problem with dense design matrix.
Rate of Convergence
It has been shown that Locally adaptive regression splines converges at the minimax rate (Mammen & van de Geer 1997).

As $n \to \infty$, trend filtering estimates lies close enough to locally adaptive regression spline estimates, thus sharing their favorable asymptotic properties.
Extensions
Extensions

- Unevenly spaced inputs

\[
D^{(x,k+1)} \cdot \text{diag}(\frac{k}{x_{k+1} - x_1}, \frac{k}{x_{k+2} - x_2}, \cdots, \frac{k}{x_n - x_{n-k}}) \cdot D^{(x,k)}
\]

\(D^{(x,k+1)}\) can still be thought of as a difference operator of order \(k + 1\), but adjusted to account for the unevenly spaced inputs \(x_1, \cdots, x_n\).

- Sparse trend filtering

\[
\hat{\beta} = \min_{\beta \in \mathbb{R}^n} \frac{1}{2} \| y - \beta \|_2^2 + \lambda_1 \| D^{(k+1)} \beta \|_1 + \lambda_2 \| \beta \|_1
\]

- Mixed trend filtering

\[
\hat{\beta} = \min_{\beta \in \mathbb{R}^n} \frac{1}{2} \| y - \beta \|_2^2 + \lambda_1 \| D^{(k_1+1)} \beta \|_1 + \lambda_2 \| D^{(k_2+1)} \beta \|_1
\]
Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky.
\( l_1 \) trend filtering.  

Ryan J Tibshirani et al.
Adaptive piecewise polynomial estimation via trend filtering.  

Yu-Xiang Wang, Alex Smola, and Ryan J Tibshirani.
The falling factorial basis and its statistical applications.  