Adaptive Piecewise Polynomial Estimation via Trend Filtering

Presenter: Yandi Shen

April 4, 2017

Introduction

- Regression problems:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) + \boldsymbol{\epsilon} \qquad E(\boldsymbol{\epsilon}) = \mathbf{0}, \mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbb{I}_n$$

  How to estimate $\mathbf{f}$?
- **Nonparametric** regression v.s. **Parametric** regression
- Parametric: **fixed** form of $\mathbf{f}$,
  e.g. linear regression, GLM, regularized regression
- Nonparametric: No fixed form of $\mathbf{f}$, constrained to some function class $\mathcal{F}$

Motivation

- Parametric: easy to compute and interpret, require less data, but prior information
- Nonparametric: more adaptive, larger sample size
- Trend filtering (TF) falls into the category of nonparametric regression
- Nonparametric regression has a long history and a well-developed toolbox

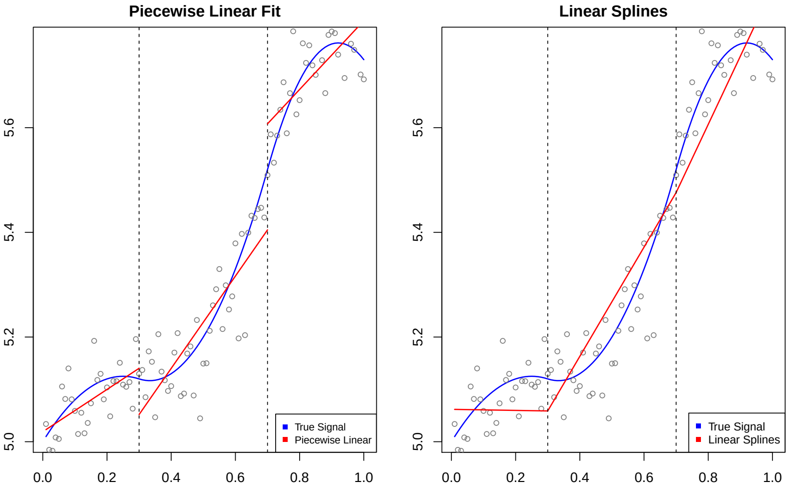
Why do we want to add TF into this toolbox?

A Quick Review of Nonparametric Toolbox

1. **Piecewise Polynomials/Splines**([De Boor et al., 1978])

   - Fit piecewise polynomials instead of a global polynomial, more adaptivity
   - In practice, we tend to use splines more often because splines are more "smooth" than piecewise polynomials
   - **Pros**: easy to compute and interpret (using basis)
     Cons: Pre-specified knots, less adaptivity

A Quick Review of Nonparametric Toolbox



Splines are smoother than piecewise polynomials!

A Quick Review of Nonparametric Toolbox

2. **Smoothing Splines**([Wahba, 1990, Green and Silverman, 1993])

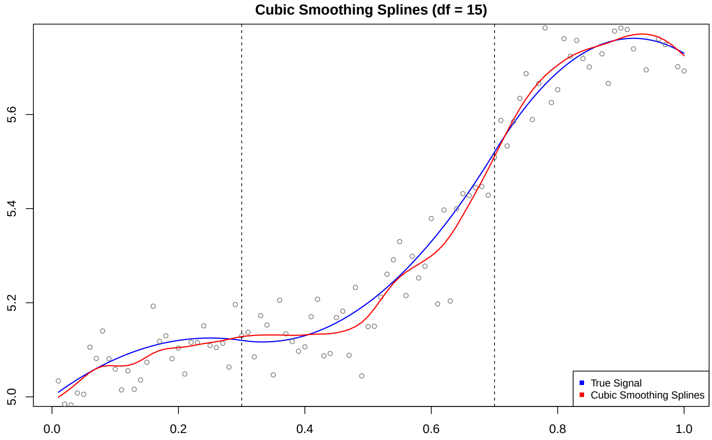
$$\underset{f \in \mathcal{W}_{(k+1)/2,2}}{\text{minimize}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_0^1 (f^{((k+1)/2)}(t))^2 dt$$

where $\mathcal{W}_{(k+1)/2,2}$ is the Sobolev space

- Can be recast into a generalized ridge regression (ℓ_2^2), solution is piecewise polynomials with knots at distinct values of $\{x_1, \ldots, x_n\}$
- **Pros**: Better flexibility, nice computation cost $\mathcal{O}(n)$
  Cons: Still not flexible enough, global shrinkage

A Quick Review of Nonparametric Toolbox

Cubic smoothing splines ($k = 3$) with 15 degrees of freedom



A Quick Review of Nonparametric Toolbox

3. **Locally Adaptive Splines**([Mammen et al., 1997])

$$\underset{f \in \mathcal{G}_K}{\text{minimize}} \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \cdot \mathrm{TV}(f^{(k)})$$

where $\mathcal{G}_k$ is the set of splines with knots as a subset of $\{x_1, \ldots, x_n\}$, $TV(\cdot)$ is the total variation penalty

- Less well-known, can be recast as a (generalized) Lasso
- Extremely similar to TF (solution & convergence rate)
- **Pros**: Nice adaptivity, minimax convergence rate under mild assumptions
  Cons: Computationally intensive (slow at 10,000 points)

Trend Filtering

- **Trend Filtering**(TF) is a method that is adaptive, computationally easy, and also enjoys nice theoretical properties (minimax rate under mild assumptions)!
- Originally proposed in [Kim et al., 2009]

$$\underset{u \in \mathbb{R}^n}{\text{minimize}} \frac{1}{2}\|y - u\|_2^2 + \lambda \sum_{i=2}^{n-1} |u_{i-1} - 2u_i + u_{i+1}|$$

- It's important in the setup that $\{x_1, \ldots, x_n\}$ are **evenly spaced** (e.g. $x_i = i/n$ on $[0, 1]$)
- Because of the sparsity of ℓ_1 norm, the solution is piecewise linear (actually linear splines)

Trend Filtering

- A more general form:

$$\underset{u \in \mathbb{R}^n}{\text{minimize}} \frac{1}{2}\|y - u\|_2^2 + \lambda \|D^{(k+1)} u\|_1$$

  where $D^{(k+1)}$ is the $(k+1)$-st discrete difference operator.

- For $k = 0$,

$$D^{(1)} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}$$

- For higher orders, $D^{(k+1)} \equiv D^{(1)} \cdot D^{(k)}$

Trend Filtering

- For $k = 1$ (same problem in [Kim et al., 2009]),

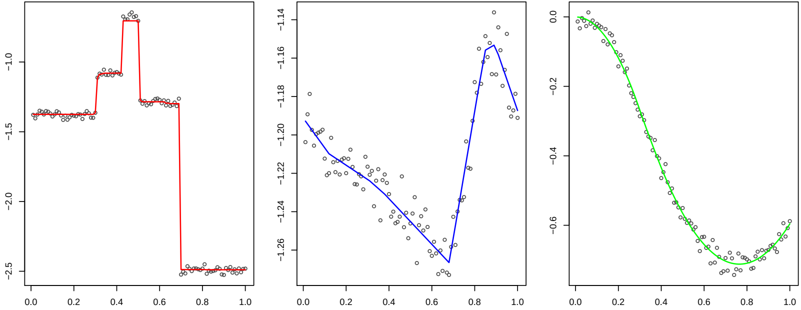
$$D^{(2)} = \begin{bmatrix} 1 & -2 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(n-2)\times n}$$

- Under the setup of evenly-spaced $\{x_1, \dots, x_n\}$, $D^{(k+1)}u$ is a **discrete** version of the derivative of order $(k+1)$
- Compare with smoothing splines, we may ask:

Will the solution of TF be piecewise polynomials/splines?

Trend Filtering

TF fit for $k = 0$, $k = 1$ and $k = 2$:



- Remark: the original problem is **discrete**, we need a **continuous-time** representation of TF (if such exists) to actually make the claim

One Application of TF

- For $k = 0$,

$$\underset{u \in \mathbb{R}^n}{\text{minimize}} \frac{1}{2}\|y - u\|_2^2 + \lambda \sum_{i=1}^{n-1} |u_{i+1} - u_i|$$

- This is total-variation denoising([Rudin et al., 1992])
- Equivalently, 1-d fused Lasso problem with only the fuse penalty term([Tibshirani et al., 2005])
- Total Variation denoising has wide applications in signal processing

Computation

$$\underset{u \in \mathbb{R}^n}{\text{minimize}} \frac{1}{2}\|y - u\|_2^2 + \lambda \|D^{(k+1)} u\|_1$$

Two algorithms:

- Single tuning λ: use the primal-dual interior point algorithm introduced in [Kim et al., 2009], worst $\mathcal{O}(n^{3/2})$
- A path algorithm for all λ: consider the generalized Lasso problem with general penalty matrix D

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \|D\beta\|_1$$

  use the algorithm proposed in [Tibshirani et al., 2011], order $\mathcal{O}(n)$ for each critical point of the solution path

Summary of Results

Nice Adaptivity

- A continuous-time representation for TF does exist
- For $k = 0, 1$, solution is constant/linear splines, and **exactly the same as locally adaptive splines**
- For $k \geq 2$, solution is piecewise polynomials

Nice Computation

- Two efficient algorithms to solve single λ or all λ
- Not much worse than smoothing splines ($\mathcal{O}(n^{3/2})$ v.s. $\mathcal{O}(n)$), much faster than locally adaptive splines

Nice Convergence Rate

- Under mild conditions, TF has same convergence rate as locally adaptive splines, thus achieves minimax rate

References

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., and De Boor, C. (1978).
A practical guide to splines, volume 27.
Springer-Verlag New York.

Green, P. J. and Silverman, B. W. (1993).
Nonparametric regression and generalized linear models: a roughness penalty approach.
CRC Press.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009).
ℓ_1 trend filtering.
SIAM review, 51(2):339–360.

Mammen, E., van de Geer, S., et al. (1997).
Locally adaptive regression splines.
The Annals of Statistics, 25(1):387–413.

References

Rudin, L. I., Osher, S., and Fatemi, E. (1992).
Nonlinear total variation based noise removal algorithms.
Physica D: Nonlinear Phenomena, 60(1-4):259–268.

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(1):91–108.

Tibshirani, R. J., Taylor, J. E., Candes, E. J., and Hastie, T. (2011).
The solution path of the generalized lasso.
Stanford University.

Wahba, G. (1990).
Spline models for observational data.
SIAM.

Thank you for listening!

Questions?