Adaptive Piecewise Polynomial Estimation via Trend Filtering

Presenter: Yandi Shen

May 9, 2017

Introduction

- Univariate nonparametric estimation:

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

  How to estimate f_0?

- Fixed design (x_i are nonrandom),
  i.i.d. noise ϵ_i with $\mathbb{E}(\epsilon_i) = 0, \mathrm{Var}(\epsilon_i) = \sigma^2$
- Linear smoothers & nonlinear smoothers

Linear Smoothers

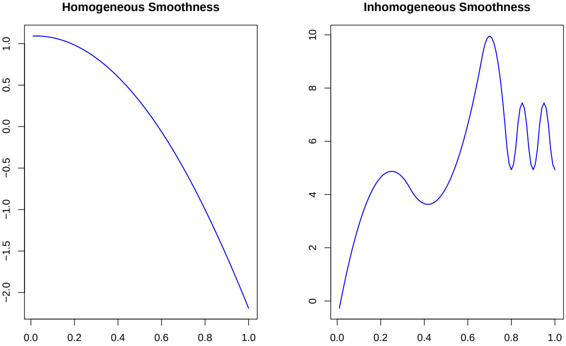
- Many well-known methods are linear smoothers:
  - Regression splines/smoothing splines [De Boor et al., 1978, Wahba, 1990, Green and Silverman, 1993]
  - K nearest neighbor (KNN) smoother[Györfi et al., 2006]
  - Kernel smoothers[Friedman et al., 2001, Loader, 2006]
  - RKHS[Smola and Schölkopf, 1998, Wahba, 1990]
  - Sieves[Shen and Wong, 1994, Wong and Shen, 1995]
- Linear smoothers have fit linear in y,

$$\hat{u} = (\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n)) = S_\lambda y$$

  where S_λ is smoothing matrix, λ is the smoothness index, e.g. k in KNN or bandwidth h in kernel smoothing
- **Nice properties**: Closed form $df = \mathrm{tr}(S_\lambda)$, efficient CV

Linear Smoothers

- **Defect:** linear smoothers are not **locally adaptive**, i.e. they cannot represent heterogeneous signals well



- Left panel: **homogeneous** smoothness; Right panel: **heterogeneous** smoothness

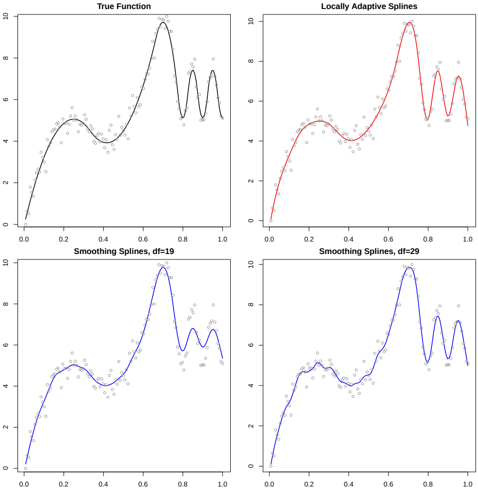
Smoothing Spline

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

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

- Most commonly used: cubic smoothing spline ($k = 3$)
- $\mathcal{W}_2$ is infinite-dimensional, but solution is natural cubic spline (finite-dimensional)

Smoothing Spline



- Small df: **over-smooth**; large df: **under-smooth**

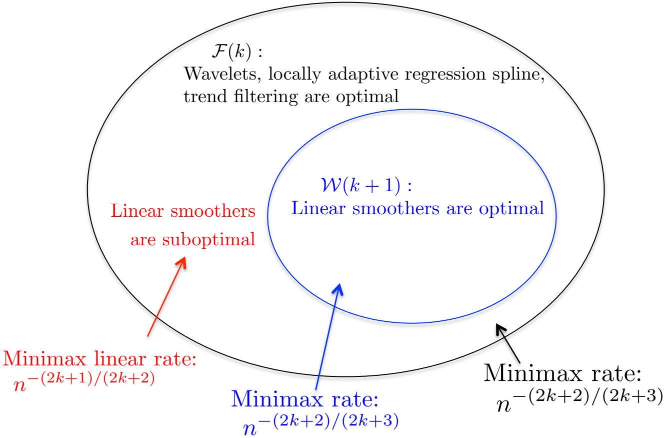
Minimax Convergence Rate

- In nonparametric statistics, we usually consider the **minimax** convergence rate:

$$\min_{\hat{f}} \max_{f_0 \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f_0(x_i) - \hat{f}(x_i))^2$$

  where $\mathcal{F}$ is some smooth function class, e.g. Sobolev $\mathcal{W}(k)$, bounded TV $\mathcal{F}(k)$, Hölder $H(k)$, etc.
- Minimax rate is the best achievable rate over $\mathcal{F}$
- Estimators that achieve minimax rate are called **(minimax) optimal**

Minimax Convergence Rate

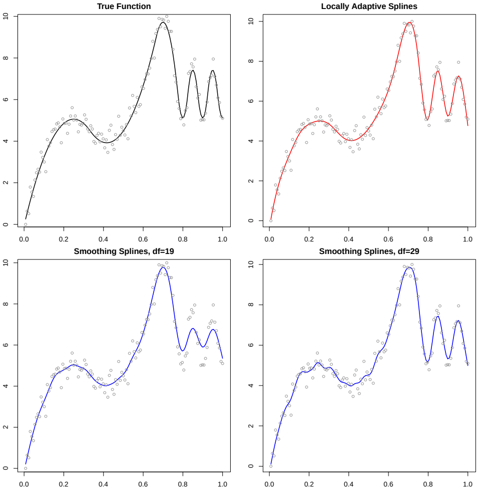


- Linear smoothers are **sub-optimal** in bounded TV class $\mathcal{F}(k)$! (e.g. for $k=0$, linear $n^{-1/2}$ v.s. nonlinear $n^{-2/3}$)

Nonlinear Smoothers

- Three methods:
  - Wavelet smoother[Johnstone, 2011, Mallat, 2008, Donoho and Johnstone, 1994]
  - Locally adaptive spline[Mammen et al., 1997]
  - Trend filtering[Kim et al., 2009]
- In 90's, wavelet smoother and locally adaptive spline were proved to be minimax optimal ($n^{-(2k+2)/(2k+3)}$) over bounded TV class $\mathcal{F}(k)$
- 20 years later, this paper proves trend filtering is also minimax-optimal

Nonlinear Smoothers



- Nonlinear smoothers are **locally adaptive**!

Nonlinear Smoothers

Why another nonlinear smoother???

- Wavelet method has stringent conditions: inputs $(x_1, \ldots, x_n)$ are evenly spaced, sample size n power of 2, boundary issues, etc.
- Locally adaptive spline is hard to compute for order $k \geq 2$, also hard to choose the location of the knots
- Trend filtering is **minimax optimal, locally adaptive**, and **computationally efficient**!

Trend Filtering

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

where $D^{(k+1)}$ is the $(k+1)$-st discrete difference operator, inputs $(x_1, \ldots, x_n)$ assumed to be evenly-spaced (**for simplicity, not must**).

- For $k = 0$,

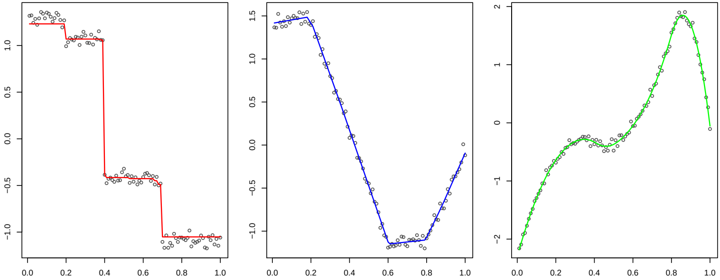
$$D^{(1)} = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}$$

- This recovers the 1d TV denoising[Rudin et al., 1992], fused lasso problem[Tibshirani et al., 2005]

Trend Filtering

- For higher orders, $D^{(k+1)} \equiv D^{(1)} \cdot D^{(k)}$, can be seen as a discrete analogue of $(k+1)$th order **derivative**

Will the solution of TF be piecewise polynomials/splines?



- Need a **continuous representation** of trend filtering!

(Generalized) Lasso Type Problem

- Generalized Lasso problem:

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

  thus TF is generalized Lasso with $X = \mathbb{I}_n, D = D^{(k+1)}$.

- Can be recast as an ordinary Lasso problem

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^n}{\arg\min} \frac{1}{2}\|y - H\alpha\|_2^2 + \lambda \sum_{j=k+2}^{n} |\alpha_j|$$

  with some sparse design H, and the TF fit is $\hat{u} = H\hat{\alpha}$

Algorithms

1. Lasso solver:

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^n}{\arg\min} \frac{1}{2}\|y - H\alpha\|_2^2 + \lambda \sum_{j=k+2}^{n} |\alpha_j|$$

2. Generalized Lasso path algorithm [Harchaoui and Lévy-Leduc, 2010, Tibshirani et al., 2011]:

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

3. Primal-dual interior point algorithm[Kim et al., 2009]

Continuous Representation of TF

- Lasso form: $\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{2}\|y - H\alpha\|_2^2 + \lambda \sum_{j=k+2}^{n} |\alpha_j|$
- Goal is to find a set of basis functions $(h_1, \ldots, h_n)$ such that $H_{ij} = h_j(x_i)$
- **Falling factorial basis**[Wang et al., 2014]:

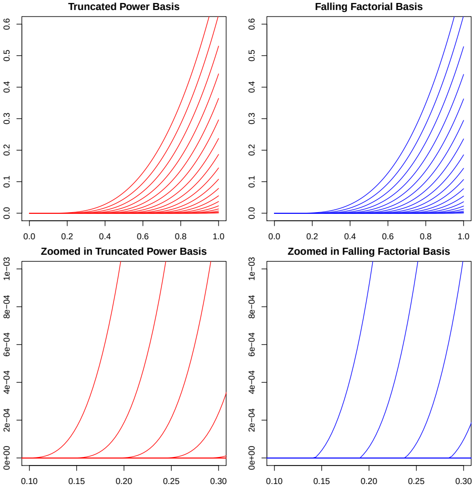
$$h_1(x) = 1, \quad h_2(x) = x, \ldots, h_{k+1}(x) = x^k,$$

$$h_{k+1+j}(x) = \prod_{l=1}^{k}(x - x_{j+l})\mathbb{1}_{[x \geq x_{j+k}]}, \quad \text{for } j = 1, \ldots, n-k-1$$

- Truncated power basis:

$$g_1(x) = 1, \quad g_2(x) = x, \ldots, g_{k+1}(x) = x^k,$$
$$g_{k+1+j}(x) = (x - t_j)^k \mathbb{1}_{[x \geq t_j]}, \quad \text{for } j = 1, \ldots, n-k-1$$

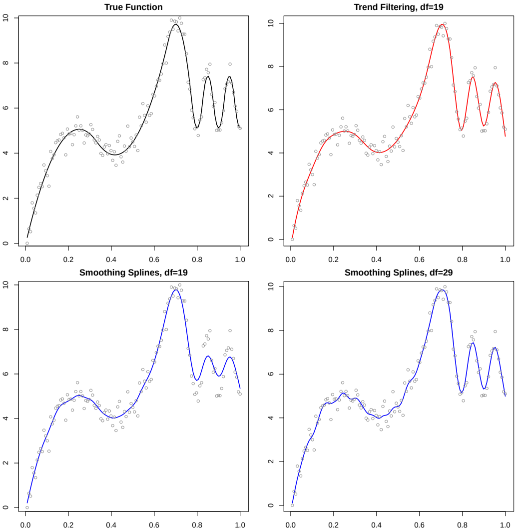
Fall Factorial Basis v.s. Truncated Power Basis



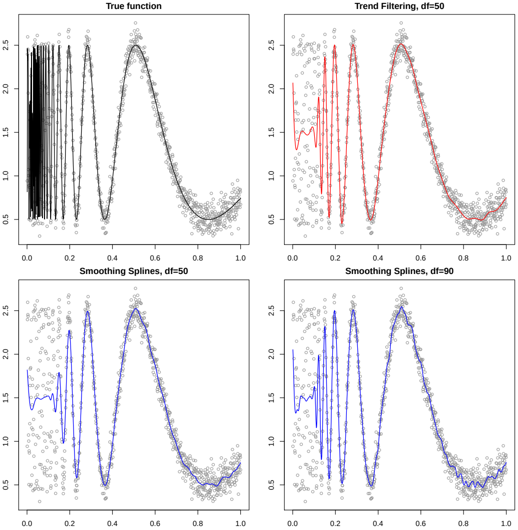
- Conclusion: $k = 0, 1$, TF is constant/linear spline;
  $k \geq 2$, TF is piecewise polynomial

Empirical Comparison with Smoothing Spline

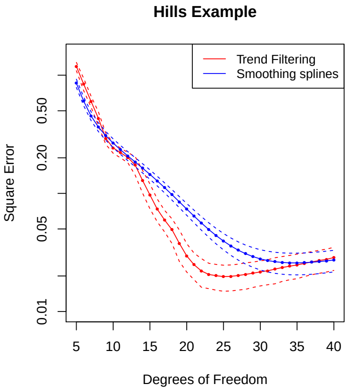
- Heterogeneous "hills" signal

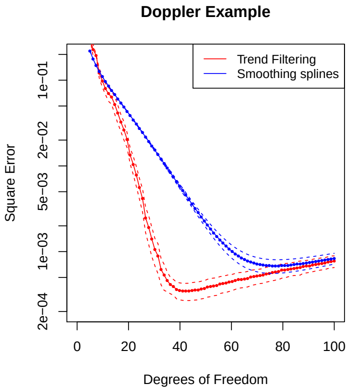


- Heterogeneous Doppler signal



Quantitative Comparison with Smoothing Spline





- In terms of MSE, trend filtering is superior to smoothing spline.

Comparison with Locally Adaptive Spline

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

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

- Equivalent to the following Lasso problem:

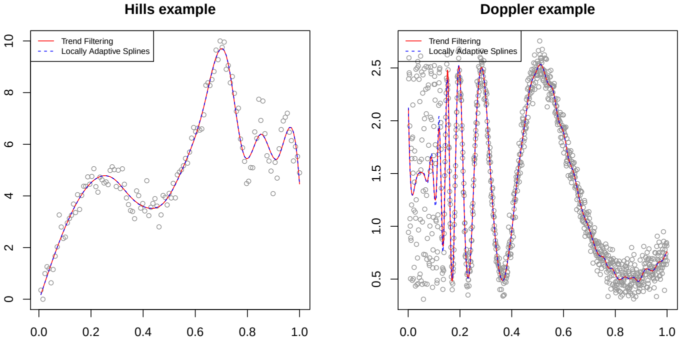
$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\arg\min} \frac{1}{2} \|y - \boldsymbol{G}\theta\|_2^2 + \lambda \sum_{j=k+2}^{n} |\theta_j|$$

- Compare with TF:

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^n}{\arg\min} \frac{1}{2} \|y - \boldsymbol{H}\alpha\|_2^2 + \lambda \sum_{j=k+2}^{n} |\alpha_j|$$

Comparison with Locally Adaptive Spline

- For $k = 0, 1$, $G = H$, thus trend filtering and locally adaptive spline give the same fit
- For $k \geq 2$, $G \neq H$, trend filtering and locally adaptive spline give different but extremely similar fit



Real Data Example

Still working on this slide

Convergence Rate of Trend Filtering

Theorem

For a fixed order k and constant $C > 0$, then over the kth order bounded TV class $\mathcal{F}(k, C) \equiv \{f_0 : TV(f_0^{(k)}) \leq C\}$, with tuning parameter chosen as $\lambda \asymp n^{1/(2k+3)}$, the fit of kth order trend filtering $\hat{u}$ satisfies

$$\sup_{f_0 \in \mathcal{F}(k,C)} \mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} (\hat{u}_i - f_0(x_i))^2] \asymp n^{-(2k+2)/(2k+3)}$$

Convergence Rate of Trend Filtering

Idea is simple:

- Trend filtering and locally adaptive spline converge to each other at minimax rate
- Locally adaptive spline converges to the true function at minimax rate[Mammen et al., 1997]
- Hence trend filtering also converges at minimax rate

More Recent Results on Trend Filtering

- [Wang et al., 2016] extends the notion of univariate trend filtering onto trend filtering on graphs
- [Wang et al., 2014] explores in depth the falling factorial basis used by the continuous representation of trend filtering.
- [Sadhanala and Tibshirani, 2017] and [Petersen et al., 2014] both extend trend filtering to high dimensions under the framework of additive model
- [Maidstone et al., 2017] explores linear trend filtering ($k = 1$) with ℓ_0 penalty

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.

Donoho, D. L. and Johnstone, I. M. (1994).
Ideal spatial adaptation by wavelet shrinkage.
biometrika, pages 425–455.

Friedman, J., Hastie, T., and Tibshirani, R. (2001).
The elements of statistical learning, volume 1.
Springer series in statistics Springer, Berlin.

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

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006).
A distribution-free theory of nonparametric regression.
Springer Science & Business Media.

Harchaoui, Z. and Lévy-Leduc, C. (2010).
Multiple change-point estimation with a total variation penalty.
Journal of the American Statistical Association, 105(492):1480–1493.

Johnstone, I. M. (2011).
Gaussian estimation: Sequence and wavelet models.
Manuscript, December.

References

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

Loader, C. (2006).
Local regression and likelihood.
Springer Science & Business Media.

Maidstone, R., Fearnhead, P., and Letchford, A. (2017).
Detecting changes in slope with an l_0 penalty.
arXiv preprint arXiv:1701.01672.

Mallat, S. (2008).
A wavelet tour of signal processing: the sparse way.
Academic press.

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

Petersen, A., Witten, D., and Simon, N. (2014).
Fused lasso additive model.
arXiv preprint arXiv:1409.5391.

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.

Sadhanala, V. and Tibshirani, R. J. (2017).
Additive models with trend filtering.
arXiv preprint arXiv:1702.05037.

References

Shen, X. and Wong, W. H. (1994).
Convergence rate of sieve estimates.
The Annals of Statistics, pages 580–615.

Smola, A. J. and Schölkopf, B. (1998).
Learning with kernels.
Citeseer.

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.

Wang, Y.-X., Sharpnack, J., Smola, A., and Tibshirani, R. J. (2016).
Trend filtering on graphs.
Journal of Machine Learning Research, 17(105):1–41.

Wang, Y.-X., Smola, A., and Tibshirani, R. (2014).
The falling factorial basis and its statistical applications.
In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 730–738.

Wong, W. H. and Shen, X. (1995).
Probability inequalities for likelihood ratios and convergence rates of sieve mles.
The Annals of Statistics, pages 339–362.

Thank you for listening!

Questions?