Adaptive Piecewise Polynomial Estimation via Trend Filtering

Yandi Shen

April 15, 2017

1 Introduction

Nonparametric statistics has a long history and various applications such as density estimation, nonparametric regression, bootstrapping, etc. In the regression setup, we will generally assume the following additive noise model

$$y_i = f_0(x_i) + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{1}$$

where $\{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^p$ are the p-dimensional observed covariates, $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$ are i.i.d. errors with $E(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$, $\{y_1, y_2, \ldots, y_n\}$ are the observed response, and $f_0(\cdot)$ is the true regression function to be estimated. In the parametric setting, we will assume some known form of f_0 through parametrization. For example, if we assume that f_0 is a linear function, then we perform linear regression and end up learning the linear coefficients of the covariates. On the contrary, in nonparametric setting, we don't assume a fixed form of f_0, but rather restrict our search in some function class $\mathcal{H}$. If the function class $\mathcal{H}$ is spanned by some possibly uncountably infinite set of atoms $\mathcal{A} = \{f_1, f_2, \ldots\}$, then our goal is to find the best linear combination that minimizes certain criterion. Generally speaking, nonparametric methods usually require a larger training set to fit the model since we don't specify the exact form of our estimator, yet still in a lot of cases, nonparametric tools are more preferable compared to their parametric counterparts because of their superior flexibility.

The nonparametric regression toolbox is highly-developed and already offers plenty of good methods. Some of the most well-known methods include piecewise polynomials/splines[1], smoothing splines [1, 18, 2], locally adaptively splines [13], etc. Each of these methods have distinct advantages and disadvantages. Piecewise polynomials/splines have nice interpretation and offer a direct generalization of the global polynomial estimation without incurring much extra computation cost. However, they are less data-driven because we need some prior information (which we usually don't have) to pre-specify the number and the location of the segmenting knots. For smoothing splines, after recasting the original minimization problem into a generalized ridge regression problem, computation cost is reduced to the order of $\mathcal{O}(n)$ but still this method is not flexible enough because the solution places knots at every distinct values in the observed covariates $\{x_1, x_2, \ldots, x_n\}$. Locally adaptive regression has nice theoretical properties but is empirically slow if we have more than $10,000$ data points. More details about these methods and their comparison with the main method of this paper will be presented in later sections.

In this paper, we will focus on the broad class of estimators: trend filtering. Trend filtering, as its name suggests, was originally proposed to estimate the underlying trend in time series data, and had since then been applied in a variety of disciplines, such as economics (e.g. [5, 17]), astronomy (e.g. [8]), social science (e.g. [15, 10]), medical science (e.g. [3, 11]), etc. Many methods for trend filtering have been proposed, including moving averaging smoothing (e.g. [9, 6, 12]), Hodrick-Prescott(H-P) filtering [5], smoothing splines (e.g. [1, 18, 2]). Both H-P filtering and smoothing splines are ℓ_2 based techniques: H-P filtering directly penalizes the square of the second order difference of the fitted value in a set of time series data; smoothing splines, after transformation, could be written as a generalized ridge regression problem (see Section ? for details). A relatively new trend filtering technique based on ℓ_1 norm was proposed in [7], where the method solves the following optimization problem

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

where $y \in \mathbb{R}^n$ is the observation signal of length n and λ is the tuning parameter. (2) penalizes the discrete second-order difference, and since ℓ_1 will reduce some of its components to 0, the solution to (2) will be piecewise linear. As a natural extension of (2), a more general version of trend filtering is proposed by modifying (2) to penalize all orders of discrete difference. Throughout the paper, we will work through and explore this more general version of trend filtering, and compare it with other nonparametric methods in terms of both computation of statistical consistency.

The rest of the paper will be organized as follows. In Section 2, we will introduce the trend filtering model and discuss the corresponding algorithms to solve the minimization problem. We then compare trend filtering with smoothing splines and locally adaptive regression splines respectively in Section 3 and Section 4. In Section 5, we establish a continuous-time representation of trend filtering, which demonstrates the continuous version of the solution of trend filtering is piecewise polynomial. In Section 6, we derive the minimax convergence rate of trend filtering by linking it to locally adaptive regression splines. A real-data study is done in Section 7, followed by some discussion in Section 8 and a conclusion in Section 9.

2 Trend Filtering(Still working...)

we could generalize trend filtering by substituting the second-order difference operator with a more general discrete difference operator for different orders. The resultant k-th order trend filtering is defined via the following optimization problem

$$\underset{u \in \mathbb{R}^n}{\text{minimize}} \frac{1}{2}\|y-u\|_2^2 + \lambda \|D^{(k+1)}u\|_1 \tag{3}$$

where λ is the tuning parameter, $D^{(k+1)} \in \mathbb{R}^{(n-k-1)\times n}$ is the $(k+1)$-st order discrete derivative operator. For order $k=0$, the difference operator $D^{(1)}$ in (3) becomes

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

and therefore the penalty term in (3) becomes $\lambda \sum_{i=1}^{n-1} |u_{i+1} - u_i|$. This coincides with 1-dimensional total-variation denoising [14, 4] and also the fused-lasso problem with solely the fused penalty term [16], and it is well-known that the solution is piece-wise constant. When $k=1$, the difference operator becomes

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

and (3) is equivalent to (2), with the solution being piecewise linear. When $k \geq 2$, the difference operator is defined recursively by

$$D^{(k+1)} \equiv D^{(1)} \cdot D^{(k)} \tag{4}$$

thus the $(k+1)$-st difference operator penalizes the change in the k-th order difference. The definition in (4) mimics the definition of derivatives and could be seen as a discrete version of derivatives. One might therefore expect the k-th order trend filtering estimator to behave somewhat similar to the piecewise k-th order polynomials. Notice that piecewise polynomials are defined continuous on the input variables, so a natural question arises: is there a continuous representation of trend filtering? And if so, is the continuous version piecewise polynomials, or even more ambitiously, splines? (Recall that k-th order splines are piecewise polynomials with continuous derivatives up to order $k-1$). We give an empirically affirmative answer in Figure ?, where we plot the trend filtering fit of order $k=0$, $k=1$ and $k=2$. As expected, the fits are piecewise constant, piecewise linear and piecewise quadratic respectively.

References

[1] Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

[2] Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press, 1993.

[3] Sander Greenland and Matthew P Longnecker. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American journal of epidemiology*, 135(11):1301–1309, 1992.

[4] Zaıd Harchaoui and Céline Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010.

[5] Robert J Hodrick and Edward C Prescott. Postwar us business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, pages 1–16, 1997.

[6] Maurice George Kendall et al. The advanced theory of statistics. *The advanced theory of statistics.*, (2nd Ed), 1946.

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

[8] Géza Kovacs and GA Bakos. Application of the trend filtering algorithm in the search for multiperiodic signals. *arXiv preprint arXiv:0812.2824*, 2008.

[9] CEV Leser. A simple method of trend construction. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 91–107, 1961.

[10] Steven D Levitt. Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. *The Journal of Economic Perspectives*, 18(1):163–190, 2004.

[11] WILLIAM A Link and JOHN R Sauer. Estimating equations estimates of trends. *Bird Populations*, 2:23–32, 1994.

[12] Robert E Lucas. Two illustrations of the quantity theory of money. *The American Economic Review*, 70(5):1005–1014, 1980.

[13] Enno Mammen, Sara van de Geer, et al. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

[14] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[15] Ankan Saha and Vikas Sindhwani. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 693–702. ACM, 2012.

[16] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[17] Ruey S Tsay. *Analysis of financial time series*, volume 543. John Wiley & Sons, 2005.

[18] Grace Wahba. *Spline models for observational data.* SIAM, 1990.