Trend filtering – I. A modern statistical tool for time-domain astronomy and astronomical spectroscopy

Article *in* Monthly Notices of the Royal Astronomical Society · March 2020

DOI: 10.1093/mnras/staa106

Collin A. Politsch
Carnegie Mellon University

Jessi Cisewski
Yale University

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **492,** 4005–4018 (2020)
Advance Access publication 2020 January 14

doi:10.1093/mnras/staa106

Trend filtering – I. A modern statistical tool for time-domain astronomy and astronomical spectroscopy

Collin A. Politsch,[1,2,3]⋆ Jessi Cisewski-Kehe,[4] Rupert A. C. Croft[3,5,6]⋆ and Larry Wasserman[1,2,3]

[1] *Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[2] *Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[3] *McWilliams Center for Cosmology, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[4] *Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA*
[5] *Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[6] *School of Physics, University of Melbourne, Parkville, VIC 3010, Australia*

Accepted 2020 January 9. Received 2020 January 7; in original form 2019 August 21

ABSTRACT

The problem of denoising a 1D signal possessing varying degrees of smoothness is ubiquitous in time-domain astronomy and astronomical spectroscopy. For example, in the time domain, an astronomical object may exhibit a smoothly varying intensity that is occasionally interrupted by abrupt dips or spikes. Likewise, in the spectroscopic setting, a noiseless spectrum typically contains intervals of relative smoothness mixed with localized higher frequency components such as emission peaks and absorption lines. In this work, we present trend filtering, a modern non-parametric statistical tool that yields significant improvements in this broad problem space of denoising *spatially heterogeneous* signals. When the underlying signal is spatially heterogeneous, trend filtering is superior to any statistical estimator that is a linear combination of the observed data – including kernel smoothers, LOESS, smoothing splines, Gaussian process regression, and many other popular methods. Furthermore, the trend filtering estimate can be computed with practical and scalable efficiency via a specialized convex optimization algorithm, e.g. handling sample sizes of $n \gtrsim 10^7$ within a few minutes. In a companion paper, we explicitly demonstrate the broad utility of trend filtering to observational astronomy by carrying out a diverse set of spectroscopic and time-domain analyses.

Key words: methods: statistical – techniques: photometric – techniques: spectroscopic.

1 INTRODUCTION

Many astronomical observations produce 1D data with varying (or unknown) degrees of smoothness. These include data from time-domain astronomy, where transient events such as supernovae can show light-curve variations on time-scales ranging from seconds to years (e.g. Dimitriadis et al. 2017; Tolstov et al. 2019). Similarly, in astronomical spectroscopy, with wavelength (or frequency) as the input variable, sharp absorption or emission-line features can be present alongside smoothly varying blackbody or other continuum radiation (see e.g. Tennyson 2019). In each of these general settings, we observe a signal plus noise and would like to denoise the signal as accurately as possible. Indeed the set of statistical tools available for addressing this general problem is quite vast. Commonly used non-parametric regression methods include kernel smoothers (e.g. Hall et al. 2002; Croft et al. 2002), local polynomial regression (LOESS; e.g. Maron & Howes 2003; Persson et al. 2004), splines (e.g. Peiris & Verde 2010; Contreras et al. 2010; Dhawan et al. 2015), Gaussian process regression (e.g. Gibson et al. 2012; Aigrain, Parviainen & Pope 2016; Gómez-Valent & Amendola 2018), and wavelet decompositions (e.g. Fligge & Solanki 1997; Theuns & Zaroubi 2000; Golkhou & Butler 2014). A rich and elegant statistical literature exists on the theoretical and practical achievements of these methods (see e.g. Györfi et al. 2002; Wasserman 2006; Hastie, Tibshirani & Friedman 2009 for general references). However, when the underlying signal is *spatially heterogeneous*, i.e. exhibits varying degrees of smoothness, the power of classical statistical literature is quite limited. Kernels, LOESS, smoothing splines, and Gaussian process regression belong to a broad family of non-parametric methods called *linear smoothers*, which has been shown to be uniformly suboptimal for estimating spatially heterogeneous signals (Nemirovskii, Polyak & Tsybakov 1985; Nemirovskii 1985; Donoho & Johnstone 1998). The common limitation of these methods is that they are not locally adaptive; i.e. by construction, they do not adapt to local degrees of smoothness in a signal. In particular,

⋆ E-mail: capolitsch@cmu.edu (CAP); rcroft@cmu.edu (RAC)

Published by Oxford University Press on behalf of the Royal Astronomical Society

continuing with the example of a smoothly varying signal with occasional sharp features, a linear smoother will tend to oversmooth the sharp features and/or overfit the smooth regions in its effort to optimally balance statistical bias and variance. Considerable effort has been made to address this problem by locally varying the hyperparameter(s) of a linear smoother. For example, locally varying the kernel bandwidth (e.g. Muller & Stadtmuller 1987; Fan & Gijbels 1992, 1995; Lepski, Mammen & Spokoiny 1997; Gijbels & Mammen 1998) irregularly varying spline knot locations (e.g. De Boor 1974; Jupp 1978; Dimatteo, Genovese & Kass 2001), and constructing non-stationary covariance functions for Gaussian process regression (e.g. Schmidt & O'Hagan 2003; Paciorek & Schervish 2004, 2006). However, since hyperparameters typically need to be estimated from the data, such exponential increases in the hyperparameter complexity severely limit the practicality of choosing the hyperparameters in a fully data-driven, generalizable, and computationally efficient fashion. Wavelet decompositions offer an elegant solution to the problem of estimating spatially heterogeneous signals, providing both statistical optimality (e.g. Donoho & Johnstone 1994, 1998) and only requiring data-driven tuning of a single (scalar) hyperparameter. Wavelets, however, possess the practical limitation of requiring a stringent analysis setting, e.g. equally spaced inputs and sample size equal to a power of two, among other provisions; and when these conditions are violated, the optimality guarantees are void. So, seemingly at an impasse, the motivating question for this work is *can we have the best of both worlds?* More precisely, is there a statistical tool that simultaneously possesses the following properties:

P1. Statistical optimality for estimating spatially heterogeneous signals
P2. Practical analysis assumptions; for example, not limited to equally spaced inputs
P3. Practical and scalable computational speed
P4. A 1D hyperparameter space, with automatic data-driven methods for selection

In this paper, we introduce trend filtering (Tibshirani 2014), a statistical method that is new to the astronomical literature and provides a strong affirmative answer to this question.

The layout of this paper is as follows. In Section 2, we provide both theoretical and empirical evidence of the superiority of trend filtering for estimating spatially heterogeneous signals compared to classical statistical methods. In Section 3, we introduce trend filtering, including a general overview of the estimator's machinery, its connection to spline methods, automatic methods for choosing the hyperparameter, uncertainty quantification, generalizations, and recommended software implementations in various programming languages. In Politsch et al. (2020) – hereafter referred to as Paper II – we directly illustrate the broad utility of trend filtering to astronomy by conducting various analyses of spectra and light curves.

2 CLASSICAL STATISTICAL METHODS AND THEIR LIMITATIONS

We begin this section by providing background and motivation for the non-parametric approach to estimating (or denoising) signals. We then discuss statistical optimality for estimating spatially heterogeneous signals, with an emphasis on providing evidence for the claim that trend filtering is superior to classical statistical methods in this highly general setting. Finally, we end this section by illustrating this superiority with a direct empirical comparison of trend filtering and several popular classical methods on simulated observations of a spatially heterogeneous signal.

2.1 Non-parametric regression

Suppose we observe noisy measurements of a response variable of interest (e.g. flux, magnitude, and photon counts) according to the data generating process (DGP)

$$f(t_i) = f_0(t_i) + \epsilon_i, \; i = 1, \ldots, n \tag{1}$$

where $f_0(t_i)$ is the signal at input t_i (e.g. a time or wavelength) and ϵ_i is the noise at t_i that contaminates the signal, giving rise to the observation $f(t_i)$. Let $t_1, \ldots, t_n \in (a, b)$ denote the observed input interval and $\mathbb{E}[\epsilon_i] = 0$ (where we use $\mathbb{E}[\cdot]$ to denote mathematical expectation). Here, the general statistical problem is to estimate (or *denoise*) the underlying signal f_0 from the observations as accurately as possible. In the non-parametric setting, we refrain from making strong a priori assumptions about f_0 that could lead to significant modelling bias, e.g. assuming a power law or a light-curve/spectral template fit. Mathematically, a non-parametric approach is defined through the deliberately weak assumption $f_0 \in \mathcal{F}$ (i.e. the signal belongs to the function class $\mathcal{F}$) where $\mathcal{F}$ is *infinite-dimensional*. In other words, the assumed class of all possible signals $\mathcal{F}$ cannot be spanned by a finite number of parameters. Contrast this to the assumption that the signal follows a pth degree power law, i.e. $f_0 \in \mathcal{F}_{\mathrm{PL}}$ where

$$\mathcal{F}_{\mathrm{PL}} = \left\{ f_0 : f_0(t) = \beta_0 + \sum_{j=1}^{p} \beta_j t^j \right\}, \tag{2}$$

a class that is spanned by $p + 1$ parameters. Similarly, given a set of p spectral/light-curve templates $b_1(t), \ldots, b_p(t)$, the usual template-fitting assumption is that $f_0 \in \mathcal{F}_{\mathrm{TEMP}}$ where

$$\mathcal{F}_{\mathrm{TEMP}} = \left\{ f_0 : f_0(t) = \beta_0 + \sum_{j=1}^{p} \beta_j b_j((t - s)/v) \right\} \tag{3}$$

and s and v are horizontal shift and scale hyperparameters, respectively. Both equations (2) and (3) represent very stringent assumptions about the underlying signal f_0. If the signal is anything other than exactly a power law in t – a highly unlikely occurrence – non-trivial statistical bias will arise by modelling it as such. Likewise, if a class of signals has a rich physical diversity (e.g. Type Ia supernova light curves; Woosley et al. 2007) that is not sufficiently spanned by the library of templates used in modelling, then statistical biases will arise. Depending on the size of the imbalance between class diversity and the completeness of the template basis, the biases could be significant. Moreover, these biases are rarely tracked by uncertainty quantification. To be clear, this is not a uniform criticism of template-fitting. For example, templates are exceptionally powerful tools for object classification and redshift estimation (e.g. Howell et al. 2005; Bolton et al. 2012). Furthermore, much of our discussion in Paper II centres around utilizing the flexible non-parametric nature of trend filtering to construct more complete spectral/light-curve template libraries for various observational objects and transient events.

Let $\widehat{f}_0$ be any statistical estimator for the signal f_0, derived from the noisy observations in equation (1). Further, let $p_t(t)$ denote the probability density function that specifies the sampling distribution of the inputs on the interval (a, b), and let $\sigma^2(t) = \mathrm{Var}(\epsilon(t))$ denote the noise level at input t. In order to assess the accuracy of the estimator it is common to consider the mean-squared prediction error (MSPE):

$$R(\widehat{f}_0) = \mathbb{E}\left[(\widehat{f}_0 - f)^2\right] \tag{4}$$

$$= \mathbb{E}\left[(\widehat{f}_0 - f_0)^2\right] + \overline{\sigma}^2 \tag{5}$$

$$= \int_a^b \left(\mathrm{Bias}^2(\widehat{f}_0(t)) + \mathrm{Var}(\widehat{f}_0(t))\right) \cdot p_{\mathrm{t}}(t)\mathrm{d}t + \overline{\sigma}^2, \tag{6}$$

where

$$\mathrm{Bias}(\widehat{f}_0(t)) = \mathbb{E}[\widehat{f}_0(t)] - f_0(t) \tag{7}$$

$$\mathrm{Var}(\widehat{f}_0(t)) = \mathbb{E}\left(\widehat{f}_0(t) - \mathbb{E}[\widehat{f}_0(t)]\right)^2 \tag{8}$$

$$\overline{\sigma}^2 = \int_a^b \sigma^2(t) \cdot p_{\mathrm{t}}(t)\mathrm{d}t. \tag{9}$$

The equality in equation (6) is commonly referred to as the bias-variance decomposition. The first term is the squared bias of the estimator $\widehat{f}_0$ (integrated over the input interval) and intuitively measures how appropriate the chosen statistical estimator is for modelling the observed phenomenon. The second term is the variance of the estimator that measures how stable or sensitive the estimator is to the observed data. And the third term is the irreducible error – the minimum prediction error we cannot hope to improve upon. The bias-variance decomposition therefore illustrates that an optimal estimator is one that combines appropriate modelling assumptions (low bias) with high stability (low variance).

2.1.1 *Statistical optimality (minimax theory)*

In this section, we briefly discuss a mathematical framework for evaluating the performance of statistical methods over non-parametric signal classes in order to demonstrate that the superiority of trend filtering is a highly general result. Ignoring the irreducible error, the problem of minimizing the MSPE of a statistical estimator can be equivalently stated as a minimization of the first term in (5) – the mean-squared estimation error (MSEE). In practice, low bias is attained by only making very weak assumptions about what the underlying signal may look like, e.g. f_0 has k continuous derivatives. An ideal statistical estimator for estimating signals in such a class (call it $\mathcal{F}$) may then be defined as

$$\inf_{\widehat{f}_0} \left(\sup_{f_0 \in \mathcal{F}} \mathbb{E}\left[(\widehat{f}_0 - f_0)^2\right]\right). \tag{10}$$

That is, we would like our statistical estimator to be the minimizer (infimum) of the worst-case (supremum) MSEE over the signal class $\mathcal{F}$. This is rarely a mathematically tractable problem for any practical signal class $\mathcal{F}$. A more tractable approach is to consider how the worst-case MSEE behaves as a function of the sample size n. A reasonable baseline metric for a statistical estimator is to require that it satisfies

$$\sup_{f_0 \in \mathcal{F}} \mathbb{E}\left[(\widehat{f}_0 - f_0)^2\right] \to 0 \tag{11}$$

as $n \to \infty$. That is, for any signal $f_0 \in \mathcal{F}$, when a large amount of data is available, $\widehat{f}_0$ gets arbitrarily close to the true signal. In any practical situation, this is not true for parametric models because the bias component of the decomposition never vanishes. This, however, is a widely held – perhaps, defining – property of non-parametric methods. Therefore, in order to distinguish optimality among non-parametric estimators, we require a stronger metric. In particular, we study *how quickly* the worst-case error goes to zero as more data are observed. This is the core idea of a rich area of statistical literature called *minimax theory* (see e.g. Van der Vaart 1998; Wasserman 2006; Tsybakov 2008). For many infinite-dimensional classes of signals, theoretical lower bounds exist on the rate at which the MSEE of *any* statistical estimator can approach zero. Therefore, if a statistical estimator is shown to achieve that rate, it can be considered optimal for estimating that class of signals. Formally, letting $g(n)$ be the rate at which the MSEE of the theoretically optimal estimator (10) goes to zero (a monotonically decreasing function in n), we would like our estimator $\widehat{f}_0$ to satisfy

$$\sup_{f_0 \in \mathcal{F}} \mathbb{E}\left[(\widehat{f}_0 - f_0)^2\right] = \mathcal{O}(g(n)), \tag{12}$$

where we use $\mathcal{O}(\cdot)$ to denote big O notation. If this is shown to be true, we say the estimator achieves the minimax rate over the signal class $\mathcal{F}$. Loosely speaking, we are stating that a *minimax optimal* estimator is an estimator that learns the signal from the data just as quickly as the theoretical gold standard estimator (10).

2.1.2 *Spatially heterogeneous signals*

Thus far we have only specified that the signal underlying most 1D astronomical observations should be assumed to belong to a class $\mathcal{F}$ that is infinite-dimensional (i.e. non-parametric). Further, in Section 2.1.1 we introduced the standard metric used to measure the performance of a statistical estimator over an infinite-dimensional class of signals. Recalling the discussion in the abstract and Introduction, trend filtering provides significant advances for estimating signals that exhibit varying degrees of smoothness across the input domain. We restate this definition below.

> **Definition.** A *spatially heterogeneous* signal is a signal that exhibits varying degrees of smoothness in different regions of its input domain.
> **Example.** A smooth light curve with abrupt transient events.
> **Example.** An electromagnetic spectrum with smooth continuum radiation and sharp absorption/emission-line features.

To complement the above definition we may also loosely define a *spatially homogeneous signal* as a signal that is *either* smooth *or* wiggly[1] across its input domain, but not both. As 'smoothness' can be quantified in various ways these definitions are intentionally mathematically imprecise. A class that is commonly considered in the statistical literature is the L_2 Sobolev class:

$$\mathcal{F}_{2,\mathrm{k}}(C_1) := \left\{ f_0 : \int_a^b f_0^{(k)}(t)^2 \mathrm{d}t < C_1 \right\}, \quad C_1 > 0,\ k \in \mathbb{N}. \tag{13}$$

That is, an L_2 Sobolev class is a class of all signals such that the integral of the square (the 'L_2 norm') of the kth derivative of each signal is less than some constant C_1. Statistical optimality in the sense of Section 2.1.1 for estimating signals in these classes (and some other closely related ones) is widely held among statistical methods in the classical toolkit; for example, kernel smoothers (Ibragimov & Hasminiskii 1980; Stone 1982), LOESS (Fan 1993; Fan et al. 1997), and smoothing splines (Nussbaum 1985). However, a seminal result by Nemirovskii et al. (1985) and Nemirovskii (1985) showed that a statistical estimator can be minimax optimal over signal classes of the form (13) and still perform quite poorly on other signals. In particular, the authors showed that, when considering the broader L_1 Sobolev class

$$\mathcal{F}_{1,\mathrm{k}}(C_2) := \left\{ f_0 : \int_a^b \left|f_0^{(k)}(t)\right| \mathrm{d}t < C_2 \right\}, \quad C_2 > 0,\ k \in \mathbb{N}, \tag{14}$$

[1] This is, in fact, a technical term used in the statistical literature.

all linear smoothers[2] – including kernels, LOESS, smoothing splines, Gaussian process regression, and many other methods – are strictly suboptimal. The key difference between these two types of classes is that L_2 Sobolev classes are rich in spatially homogeneous signals but not spatially heterogeneous signals, while L_1 Sobolev classes[3] are rich in both (see e.g. Donoho & Johnstone 1998).

The intuition of this result is that linear smoothers cannot optimally recover signals that exhibit varying degrees of smoothness across their input domain because they operate as if the signal possesses a fixed degree of smoothness. For example, this intuition is perhaps most clear when considering a kernel smoother with a fixed bandwidth. The result of Nemirovskii et al. (1985) and Nemirovskii (1985) therefore implies that, in order to achieve statistical optimality for estimating spatially heterogeneous signals, a statistical estimator must be non-linear (more specifically, it must be locally adaptive). Tibshirani (2014) showed that trend filtering is minimax optimal for estimating signals in L_1 Sobolev classes. Since L_2 Sobolev classes are contained within L_1 Sobolev classes, this result also guarantees that trend filtering is also minimax optimal for estimating signals in L_2 Sobolev classes. Wavelets share this property, but require restrictive assumptions on the sampling of the data (Donoho & Johnstone 1994).

How large is this performance gap? The collective results of Nemirovskii et al. (1985), Nemirovskii (1985), and Tibshirani (2014) reveal that the performance gap between trend filtering and linear smoothers when estimating spatially heterogeneous signals is significant. For example, when $k = 0$, the minimax rate over L_1 Sobolev classes (which trend filtering achieves) is $n^{-2/3}$, but linear smoothers cannot achieve better than $n^{-1/2}$. To put this in perspective, this result says that the trend filtering estimator, training on n data points, learns these signals with varying smoothness as quickly as a linear smoother training on $n^{4/3}$ data points. As we demonstrate in the next section, this gap in theoretical optimality has clear practical consequences.

In order to minimize the pervasion of technical statistical jargon throughout the paper, henceforth we simply refer to a statistical estimator that achieves the minimax rate over L_2 Sobolev classes as statistically optimal for estimating spatially homogeneous signals, and we refer to a statistical estimator that achieves the minimax rate over L_1 Sobolev classes as statistically optimal for estimating spatially heterogeneous signals. As previously mentioned, the latter implies the former, but not vice versa.

2.2 Empirical comparison

In this section, we analyse noisy observations of a simulated spatially heterogeneous signal in order to compare the empirical performance of trend filtering and several classical statistical methods – namely, LOESS, smoothing splines, and Gaussian process regression. The mock observations are simulated on an unequally spaced grid $t_1, \ldots, t_n \sim \mathrm{Unif}(0, 1)$ according to the DGP

$$f(t_i) = f_0(t_i) + \epsilon_i \tag{15}$$

with

$$f_0(t_i) = 6\sum_{k=1}^{3}(t_i - 0.5)^k + 2.5\sum_{j=1}^{4}(-1)^j\phi_j(t_i), \tag{16}$$

where $\phi_j(t)$, $j = 1, \ldots, 4$ are compactly supported radial basis functions distributed throughout the input space and $\epsilon_i \sim N(0, 0.125^2)$. We therefore construct the signal f_0 to have a smoothly varying global trend with four sharp localized features – two dips and two spikes. The signal and noisy observations are shown in the top panel of Fig. 1.

In order to facilitate the comparison of methods, we utilize a metric for the total statistical complexity (i.e. total wiggliness) of an estimator known as the effective degrees of freedom (see e.g. Tibshirani 2015). Formally, the effective degrees of freedom of an estimator $\widehat{f}_0$ is defined as

$$\mathrm{df}(\widehat{f}_0) = \overline{\sigma}^{\,-2}\sum_{i=1}^{n}\mathrm{Cov}(\widehat{f}_0(t_i), f(t_i)) \tag{17}$$

where $\overline{\sigma}^2$ is defined in (9). In Fig. 1, we fix all estimators to have 55 effective degrees of freedom. This exercise provides insight into how each estimator relatively distributes its complexity across the input domain. In the second panel, we see that the trend filtering estimate has sufficiently recovered the underlying signal, including both the smoothness of the global trend and the abruptness of the localized features. All three of the linear smoothers, on the other hand, severely oversmooth the localized peaks and dips. Gaussian process regression also exhibits some undesirable oscillatory features that do not correspond to any real trend in the signal. In order to better recover the localized features the linear smoothers require a more complex fit, i.e. smaller LOESS kernel bandwidth, smaller smoothing spline penalization, and smaller Gaussian process noise-signal variance. In Fig. 2, we show the same comparison, but we grant the linear smoothers more complexity. Specifically, in order to recover the sharp features comparably with the trend filtering estimator with 55 effective degrees of freedom, the linear smoothers require 192 effective degrees of freedom – approximately 3.5 times the complexity. As a result, although they now adequately recover the peaks and dips, each linear smoother severely overfits the data in the other regions of the input domain, resulting in many spurious fluctuations.

As discussed in Section 2.1.2, the suboptimality of LOESS, smoothing splines, and Gaussian process regression illustrated in this example is an inherent limitation of the broad *linear smoother* family of statistical estimators. Linear smoothers are adequate tools for estimating signals that exhibit approximately the same degree of smoothness throughout their input domain. However, when a signal is expected to exhibit varying degrees of smoothness across its domain, a locally adaptive statistical estimator is needed.

3 TREND FILTERING

Trend filtering, in its original form, was independently proposed in the computer vision literature (Steidl, Didas & Neumann 2006) and the applied mathematics literature (Kim et al. 2009), and has recently been further developed in the statistical and machine learning literature, most notably with Tibshirani & Taylor (2011), Tibshirani (2014), Wang et al. (2016), and Ramdas & Tibshirani (2016). This work is in no way related to the work of Kovács, Bakos & Noyes (2005), which goes by a similar name. At a high level, trend filtering is closely related to two familiar non-parametric

[2] A *linear smoother* is a statistical estimator that is a linear combination of the observed data. Many popular statistical estimators, although often motivated from seemingly disparate premises, can be shown to fall under this definition. See e.g. Wasserman (2006) for more details.

[3] The L_1 Sobolev class is often generalized to a nearly equivalent but slightly larger class – namely, signals with derivatives of bounded variation. See Tibshirani (2014) for the generalized definition.

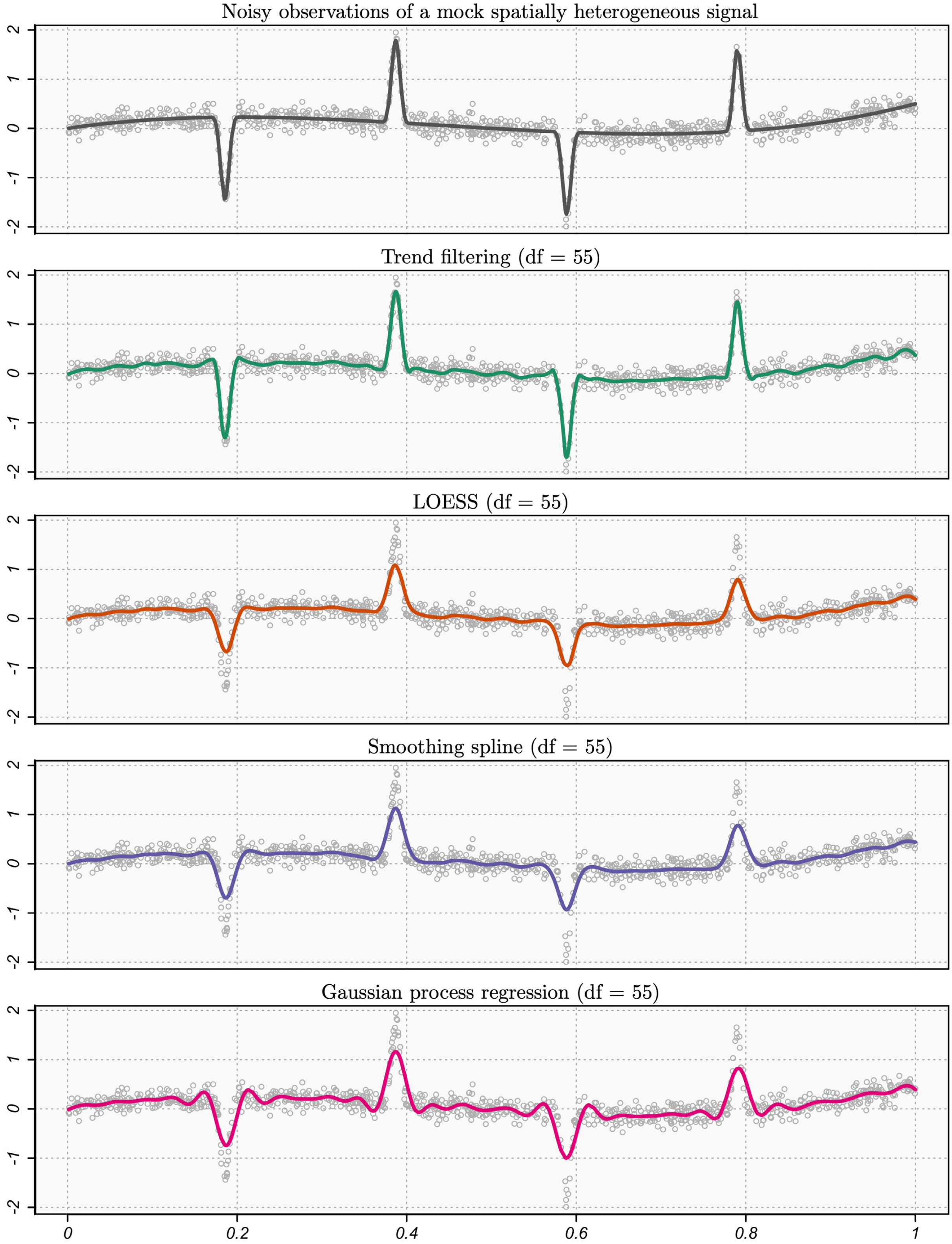

Figure 1. Comparison of statistical methods on data simulated from a spatially heterogeneous signal. Each statistical estimator is fixed to have 55 effective degrees of freedom in order to facilitate a direct comparison. The trend filtering estimator is able to sufficiently distribute its effective degrees of freedom such that it simultaneously recovers the smoothness of the global trend, as well as the abrupt localized features. The LOESS, smoothing spline, and Gaussian process regression each estimates the smooth global trend reasonably well here, but significantly oversmooths the sharp peaks and dips. Here, we utilize quadratic trend filtering (see Section 3.2).

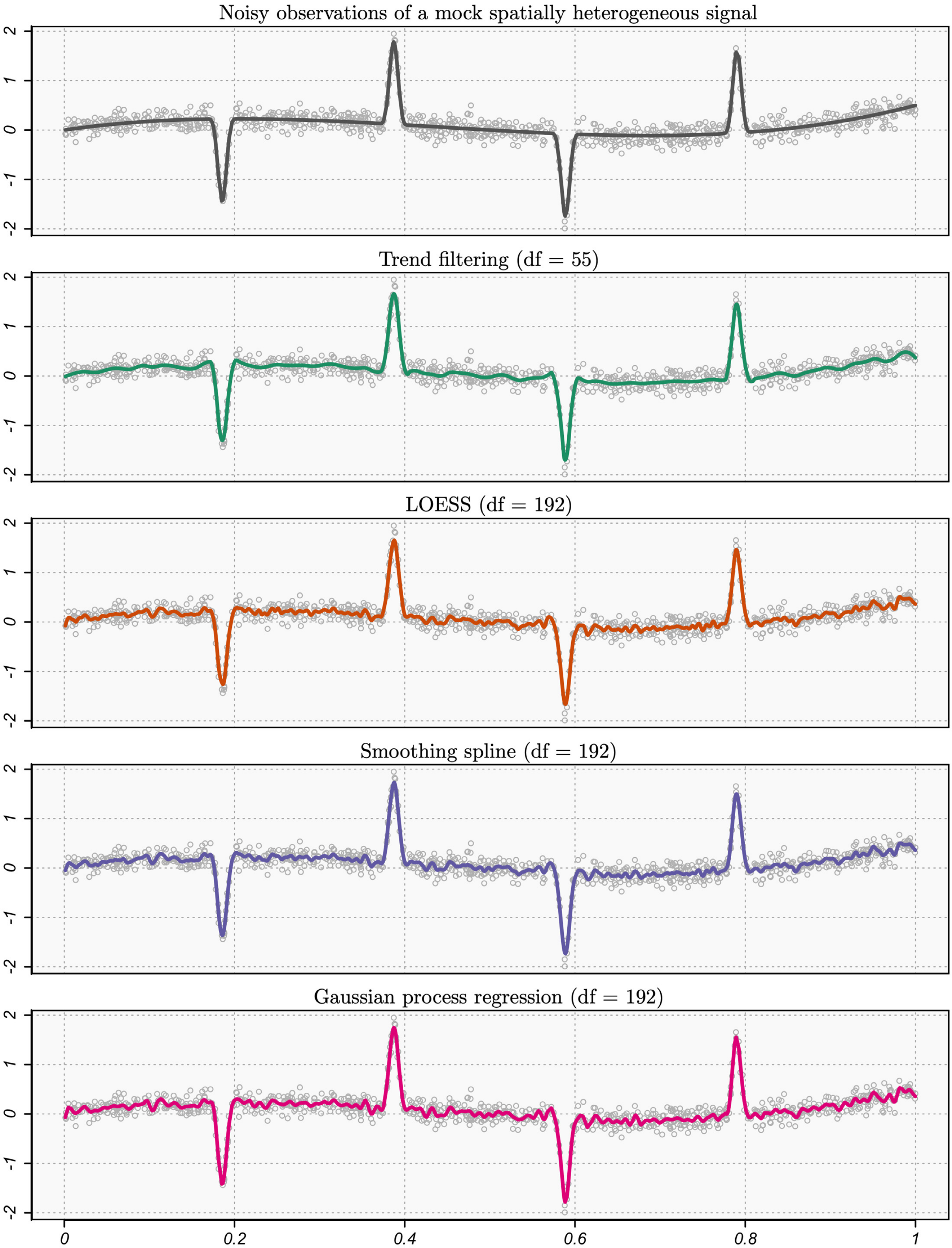

Figure 2. (Continued): Comparison of statistical methods on data simulated from a spatially heterogeneous signal. Here, each of the linear smoothers (i.e. the LOESS, smoothing spline, and Gaussian process regression) is fixed at 192 effective degrees of freedom – the complexity necessary for each estimator to recover the sharp localized features approximately as well as the trend filtering estimator with 55 effective degrees of freedom. While the linear smoothers now estimate the four abrupt features well, each severely overfits the data in the other regions of the input domain.

regression methods: variable-knot regression splines and smoothing splines. We elaborate on these relationships below.

3.1 Closely related methods

Splines have long played a central role in estimating complex signals (see e.g. De Boor 1978 and Wahba 1990 for general references). Formally, a kth-order spline is a piecewise polynomial (i.e. piecewise power law) of degree k that is continuous and has $k - 1$ continuous derivatives at the knots. As their names suggest, variable-knot regression splines and smoothing splines centre around fitting splines to observational data. Recall from equation (1), the observational DGP

$$f(t_\mathrm{i}) = f_0(t_\mathrm{i}) + \epsilon_\mathrm{i}, \quad t_1, \dots, t_\mathrm{n} \in (a, b), \tag{18}$$

where $f(t_\mathrm{i})$ is a noisy measurement of the signal $f_0(t_\mathrm{i})$, and $\mathbb{E}[\epsilon_\mathrm{i}] = 0$. Given a set of knots $\kappa_1, \dots, \kappa_\mathrm{p} \in (a, b)$, the space of all kth-order splines on the interval (a, b) with knots at $\kappa_1, \dots, \kappa_\mathrm{p}$ can be parametrized via a basis representation

$$m(t) = \sum_j \beta_\mathrm{j}\eta_\mathrm{j}(t), \tag{19}$$

where $\{\eta_j\}$ is typically the truncated power basis or B-spline basis. A suitable estimator for the signal f_0 may then be

$$\widehat{f}_0(t) = \sum_j \widehat{\beta}_\mathrm{j}\eta_\mathrm{j}(t), \tag{20}$$

where the $\widehat{\beta}_\mathrm{j}$ are the ordinary least-squares (OLS) estimates of the basis coefficients. This is called a *regression spline*. The question of course remains where to place the knots.

3.1.1 Variable-knot regression splines

The variable-knot (or free-knot) regression spline approach is to consider all regression spline estimators with knots at a subset of the observed inputs, i.e. $\{\kappa_1, \dots, \kappa_\mathrm{p}\} \subset \{t_1, \dots, t_\mathrm{n}\}$ for all possible p. Formally, the variable-knot regression spline estimator is the solution to the following constrained least-squares minimization problem:

$$\begin{aligned} \min_{\{\beta_j\}} \quad & \sum_{i=1}^{n}\Big(f(t_\mathrm{i}) - \sum_j \beta_\mathrm{j}\eta_\mathrm{j}(t_\mathrm{i})\Big)^2 \\ \text{s.t.} \quad & \sum_{j\geq k+2} \mathbb{1}\{\beta_\mathrm{j} \neq 0\} = p \\ & p \geq 0 \end{aligned} \tag{21}$$

where $p \geq 0$ is the number of knots in the spline and $\mathbb{1}(\cdot)$ is the indicator function satisfying

$$\mathbb{1}\{\beta_\mathrm{j} \neq 0\} = \begin{cases} 1 & \beta_\mathrm{j} \neq 0, \\ 0 & \beta_\mathrm{j} = 0. \end{cases} \tag{22}$$

Furthermore, note that the equality constraint on the basis coefficients excludes those of the 'first' $k + 1$ basis functions that span the space of global polynomials and only counts the number of active basis functions that produce knots. The variable-knot regression spline optimization is therefore a problem of finding the *best subset* of knots for the regression spline estimator. Due to the sparsity of the coefficient constraint, the variable-knot regression spline estimator allows for highly locally adaptive behaviour for estimating signals that exhibit varying degrees of smoothness. However, the problem itself cannot be solved in polynomial time, requiring an exhaustive combinatorial search over all $\sim 2^n$ feasible models. It is common to utilize stepwise procedures based on iterative addition and deletion of knots in the active set, but these partial searches over the feasible set inherently provide no guarantee of finding the optimal global solution to equation (21).

In order to make the connection to trend filtering more explicit it is helpful to reformulate the constrained minimization (21) into the following penalized unconstrained minimization problem:

$$\min_{\{\beta_j\}} \quad \sum_{i=1}^{n}\Big(f(t_\mathrm{i}) - \sum_j \beta_\mathrm{j}\eta_\mathrm{j}(t_\mathrm{i})\Big)^2 + \gamma \sum_{j\geq k+2} \mathbb{1}\{\beta_\mathrm{j} \neq 0\}, \tag{23}$$

where $\gamma > 0$ is a hyperparameter that determines the number of knots in the spline and the sum of indicator functions serves as a smoothness 'penalty' on the OLS minimization. Penalized regression is a popular area of statistical methodology (see e.g. Hastie et al. 2009), in which the cost functional (i.e. the quantity to be minimized) quantifies a trade-off between the training error of the estimator (here, the sum of squared residuals) and the statistical complexity of the estimator (here, the number of knots in the spline). In particular, equation (23) is known as an ℓ_0-penalized least-squares regression because of the penalty's connection to the mathematical ℓ_0 vector quasi-norm.

3.1.2 Smoothing splines

Smoothing splines counteract the computational issue faced by variable-knot regression splines by simply placing knots at all of the observed inputs $t_1, \dots, t_\mathrm{n}$ and regularizing the smoothness of the fitted spline. For example, letting $\mathcal{G}$ be the space of all cubic natural splines with knots at $t_1, \dots, t_\mathrm{n}$, the cubic smoothing spline estimator is the solution to the optimization problem

$$\min_{m\in\mathcal{G}} \quad \sum_{i=1}^{n}\big(f(t_\mathrm{i}) - m(t_\mathrm{i})\big)^2 + \gamma \int_a^b \big(m''(t)\big)^2 \mathrm{d}t, \tag{24}$$

where m'' is the second derivative of m and $\gamma > 0$ tunes the amount of regularization. Letting $\eta_1, \dots, \eta_\mathrm{n}$ be a basis for cubic natural splines with knots at the observed inputs, equation (24) can be equivalently stated as a minimization over the basis coefficients:

$$\min_{\{\beta_j\}} \quad \sum_{i=1}^{n}\Big(f(t_\mathrm{i}) - \sum_j \beta_\mathrm{j}\eta_\mathrm{j}(t_\mathrm{i})\Big)^2 + \gamma \sum_{j,k=1}^{n} \beta_\mathrm{j}\beta_\mathrm{k}\omega_\mathrm{jk} \tag{25}$$

where

$$\omega_\mathrm{jk} = \int_a^b \eta_\mathrm{j}''(t)\eta_\mathrm{k}''(t)\mathrm{d}t. \tag{26}$$

The cost functional (25) is differentiable and leads to a linear system with a special sparse structure (i.e. bandedness), which yields a solution that can both be found in closed-form and computed very quickly – in $O(n)$ elementary operations. This particular choice of cost functional, however, produces an estimator that is a linear combination of the observations – a *linear smoother*. Therefore, as discussed and demonstrated in Section 2, smoothing splines are suboptimal for estimating spatially heterogeneous signals. Equation (25) is known as an ℓ_2-penalized least-squares regression because of the penalty's connection to the mathematical ℓ_2 vector norm.

3.2 Definition

Trend filtering can be viewed as a blending of the strengths of variable-knot regression splines (local adaptivity and interpretability) and the strengths of smoothing splines (simplicity and speed). Mathematically, this is achieved by choosing an appropriate set of basis functions and penalizing the least-squares problem with an ℓ_1 norm on the basis coefficients (sum of absolute values), instead of the ℓ_0 norm of variable-knot regression splines (sum of indicator functions) or the ℓ_2 norm of smoothing splines (sum of squares).

This section is primarily summarized from Tibshirani (2014) and Wang, Smola & Tibshirani (2014). Let the inputs be ordered with respect to the index, i.e. $t_1 < \cdots < t_{\rm n}$. For the sake of simplicity, we consider the case when the inputs $t_1, \ldots, t_{\rm n} \in (a, b)$ are equally spaced with $\Delta t = t_{\rm i+1} - t_{\rm i}$. See the aforementioned papers for the generalized definition of trend filtering to unequally spaced inputs.

For any given integer $k \geq 0$, the kth-order trend filtering estimate is a piecewise polynomial of degree k with knots *automatically selected* at a sparse subset of the observed inputs $t_1, \ldots, t_{\rm n}$. In Fig. 3, we provide an example of a trend-filtered data set for orders $k = 0$, 1, 2, and 3. Specifically, the panels of the figure respectively display piecewise constant, piecewise linear, piecewise quadratic, and piecewise cubic fits to the data with the automatically selected knots indicated by the tick marks on the horizontal axes. Constant trend filtering is equivalent to total variation denoising (Rudin, Osher & Faterni 1992), as well as special forms of the fused LASSO of Tibshirani et al. (2005) and the variable fusion estimator of Land & Friedman (1996). Linear trend filtering was independently proposed by Steidl et al. (2006) and Kim et al. (2009). Higher order polynomial trend filtering ($k \geq 2$) was developed by Tibshirani & Taylor (2011) and Tibshirani (2014). In the Fig. 3 example, the quadratic and cubic trend filtering estimates are nearly visually indistinguishable, and this is true in general. Although, as we see here, trend filtering estimates of different orders typically select different sets of knots.

Like the spline methods discussed in Section 3.1, for any order $k \geq 0$, the trend filtering estimator has a basis representation

$$m(t) = \sum_{j=1}^{n} \beta_{\rm j} h_{\rm j}(t), \tag{27}$$

but, here, the trend filtering basis $\{h_1, \ldots, h_{\rm n}\}$ is the *falling factorial* basis, which is defined as

$$h_{\rm j}(t) = \begin{cases} \prod_{i=1}^{j-1}(t - t_{\rm i}) & j \leq k+1, \\ \prod_{i=1}^{j-1}(t - t_{\rm j-k-1+i}) \cdot \mathbb{1}\{t \geq t_{\rm j-1}\} & j \geq k+2. \end{cases} \tag{28}$$

Like the truncated power basis, the first $k + 1$ basis functions span the space of global kth-order polynomials and the rest of the basis adds the piecewise polynomial structure. However, the knot-producing basis functions of the falling factorial basis $h_{\rm j}$, $j \geq k + 2$ have small discontinuities in their jth-order derivatives at the knots for all $j = 1, \ldots, k - 1$, and therefore for orders $k \geq 2$ the trend filtering estimate is *close to*, but not quite a spline. The discontinuities are small enough, however, that the trend filtering estimate defined through the falling factorial basis representation is visually indistinguishable from the analogous spline produced by the truncated power basis (see Tibshirani 2014; Wang et al. 2014). The advantage of utilizing the falling factorial basis in this context instead of the truncated power basis (or the B-spline basis) comes in the form of significant computational speedups, as we detail below.

Analogous to the continuous smoothing spline problem (24), we let $\mathcal{H}_{\rm k}$ be the space of all functions spanned by the kth-order falling factorial basis, and pose the trend filtering problem as a least-squares minimization with a derivative-based penalty on the fitted function. In particular, the kth-order trend filtering estimator is the solution to the problem

$$\min_{m \in \mathcal{H}_k} \sum_{i=1}^{n} \big(f(t_{\rm i}) - m(t_{\rm i})\big)^2 + \gamma \cdot \mathrm{TV}(m^{(k)}), \tag{29}$$

where $m^{(k)}$ is the kth derivative of m, $\mathrm{TV}(m^{(k)})$ is the *total variation* of $m^{(k)}$, and $\gamma > 0$ is the model hyperparameter that controls the smoothness of the fit. When $m^{(k)}$ is differentiable everywhere in its domain, the penalty term simplifies to

$$\mathrm{TV}(m^{(k)}) = \int_a^b |m^{(k+1)}(t)| \mathrm{d}t. \tag{30}$$

Avoiding the technical generalized definition of total variation (see e.g. Tibshirani 2014), we can simply think of $\mathrm{TV}(\cdot)$ as a generalized L_1 norm[4] for our piecewise polynomials that possess small discontinuities in the derivatives. Again referring back to the smoothing spline problem (24), definitions (29) and (30) reveal that trend filtering can be thought of as an L_1 analogue of the (L_2-penalized) smoothing spline problem. Moreover, note that unlike smoothing splines, trend filtering can produce piecewise polynomials of all orders $k \geq 0$.

Replacing m with its basis representation, i.e. $m(t) = \sum_{\rm j} \beta_{\rm j} h_{\rm j}(t)$, yields the equivalent finite-dimensional trend filtering minimization problem:[5]

$$\min_{\{\beta_j\}} \sum_{i=1}^{n} \left(f(t_{\rm i}) - \sum_{j=1}^{n} \beta_{\rm j} h_{\rm j}(t_{\rm i}) \right)^2 + \gamma \cdot k! \cdot \Delta t^k \sum_{j=k+2}^{n} |\beta_{\rm j}|. \tag{31}$$

The terms $k!$ and Δt^k are constants and can therefore be ignored by absorbing them into the hyperparameter γ. Visual inspection of equation (31) reveals that trend filtering is also analogous to the variable-knot regression spline problem (21) – namely, by replacing the ℓ_0 norm on the basis coefficients with an ℓ_1 norm. The advantage here is that the problem is now strictly convex and can be efficiently solved by various convex optimization algorithms. Furthermore, the ℓ_1 penalty still yields a sparse solution (i.e. many $\beta_{\rm j} = 0$), which provides the automatic knot-selection property. Letting $\widehat{\beta}_1, \ldots, \widehat{\beta}_{\rm n}$ denote the solution to equation (31) for a particular choice of $\gamma > 0$, the trend filtering estimate is then given by

$$\widehat{f}_0(t; \gamma) = \sum_{j=1}^{n} \widehat{\beta}_{\rm j} h_{\rm j}(t), \tag{32}$$

with the automatically selected knots corresponding to the basis functions with $\widehat{\beta}_{\rm j} \neq 0, j \geq k + 1$.

The advantage of utilizing the falling factorial basis is found by reparametrizing the problem (31) into an optimization over the fitted values $m(t_1), \ldots, m(t_{\rm n})$. The problem then reduces to

$$\min_{\{m(t_{\rm i})\}} \sum_{i=1}^{n} \big(f(t_{\rm i}) - m(t_{\rm i})\big)^2 + \gamma \sum_{i=1}^{n-k-1} |\Delta^{(k+1)} m(t_{\rm i})| \cdot \Delta t \tag{33}$$

where $\Delta^{(k+1)} m(t_{\rm i})$ can be viewed as a discrete approximation of the $(k + 1)$st derivative of m at $t_{\rm i}$. For $k = 0$, the discrete derivatives are

$$\Delta^{(1)} m(t_{\rm i}) = \frac{m(t_{\rm i+1}) - m(t_{\rm i})}{\Delta t}, \tag{34}$$

[4] We use the upper case notation $L_{\rm p}$, $p = 1, 2$ for the p-norm of a continuous function, and $\ell_{\rm p}$, $p = 0, 1, 2$ for the p-norm of a vector.

[5] This may be recognized as a LASSO regression (Tibshirani 1996), with the features being the falling factorial basis functions.

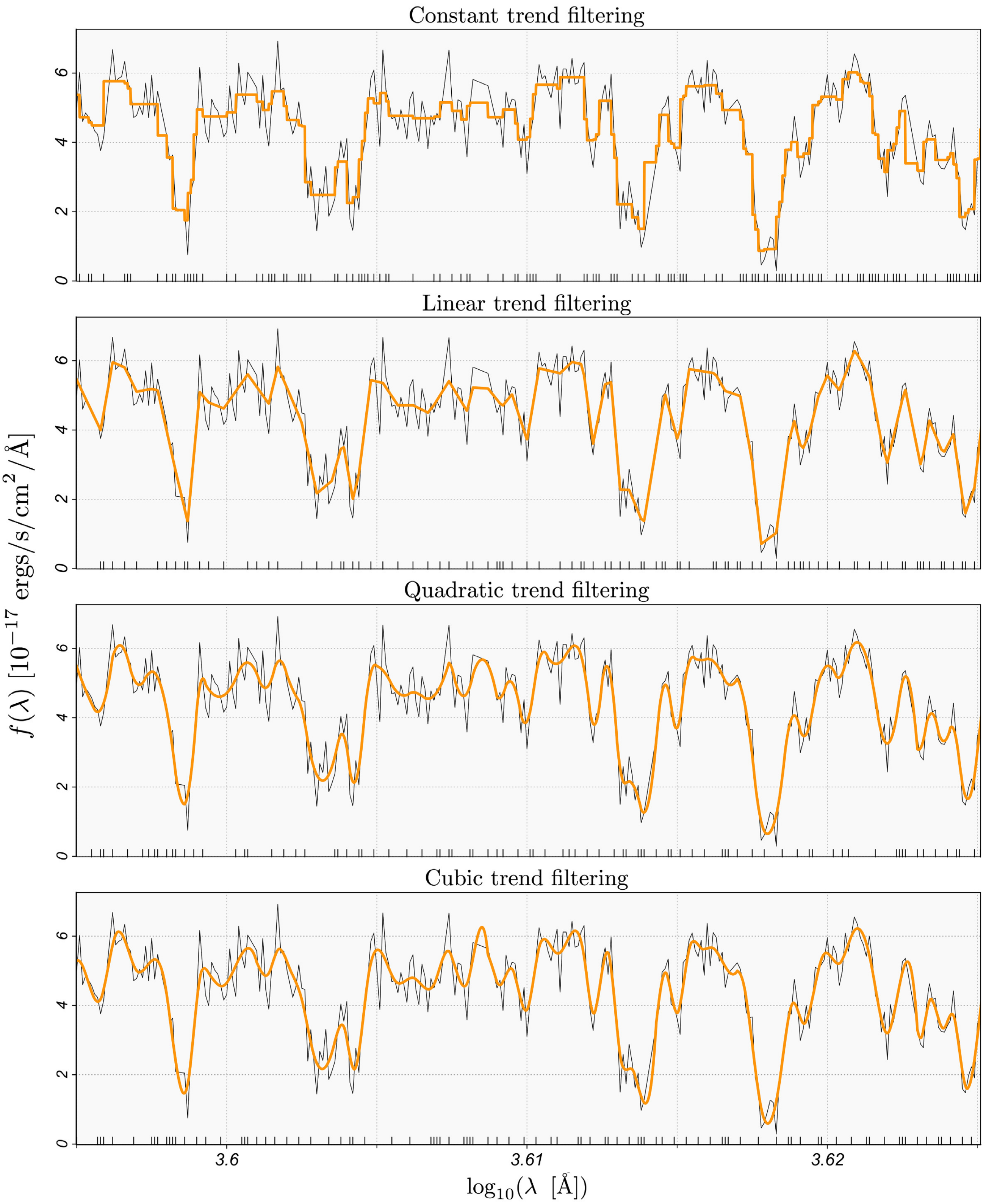

Figure 3. Piecewise polynomials with adaptively chosen knots produced by trend filtering. From top to bottom, we show trend filtering estimates of orders $k =$ 0, 1, 2, and 3, which take the form of piecewise constant, piecewise linear, piecewise quadratic, and piecewise cubic polynomials, respectively. The adaptively chosen knots of each piecewise polynomial are indicated by the tick marks along the horizontal axes. The constant trend filtering estimate is discontinuous at the knots, but we interpolate here for visual purposes. The data set is taken from the Lyman-α forest of a mock quasar spectrum (Bautista et al. 2015), sampled in logarithmic-angstrom space. We study this phenomenon in detail in Paper II.

and then can be defined recursively for $k \geq 1$:

$$\Delta^{(k+1)}m(t_\mathrm{i}) = \frac{\Delta^{(k)}m(t_{\mathrm{i}+1}) - \Delta^{(k)}m(t_\mathrm{i})}{\Delta t}. \tag{35}$$

The penalty term in equation (33) can be viewed as a Riemann-like discrete approximation of the integral in equation (30). Because of the choice of basis, the problem has reduced to a simple generalized LASSO problem (Tibshirani & Taylor 2011; Arnold & Tibshirani 2016) with an identity predictor matrix and a banded[6] penalty matrix. This special structure allows the solution to be computed very efficiently and with a nearly linear time-scaling – i.e. $\mathcal{O}(n)$ elementary operations – via the specialized alternating direction method of multipliers (ADMM) algorithm of Ramdas & Tibshirani (2016). This algorithm has a linear complexity per iteration, so the overall complexity is $\mathcal{O}(nr)$ where r is the number of iterations necessary to converge to the solution. In the worst case scenario $r \sim n^{1/2}$, so the worst-case overall complexity is $\mathcal{O}(n^{1.5})$. In practice, the computations of the specialized trend filtering optimization algorithm are highly efficient and scale to massive data sets, e.g. handling data sets with $n \gtrsim 10^7$ within a few minutes. See Ramdas & Tibshirani (2016) for more rigorous timing results. The practical and scalable computational speed further illustrates the value of trend filtering to astronomy, as it is readily compatible with the large-scale analysis of 1D data sets that has become increasingly ubiquitous in large sky surveys. We show a comparison in Table 1 of the computational costs associated with trend filtering and other popular 1D non-parametric methods.

Given the trend filtering fitted values obtained by the optimization (33) the full continuous-time representation of the trend filtering estimate follows by inverting the parametrization back to the basis function coefficients and plugging them into the basis representation (32).

3.3 Extension to heteroskedastic weighting

Thus far we have considered the simple case where the observations are treated as equally weighted in the cost functional (33). Recall from equation (18), the observational DGP and define $\sigma_\mathrm{i}^2 = \mathrm{Var}(\epsilon_\mathrm{i})$ to be the noise level – the (typically heteroskedastic) uncertainty in the measurements that arises from instrumental errors and removal of systematic effects. When estimates for σ_i^2, $i = 1, \ldots, n$ accompany the observations, as they often do, they can be used to weight the observations to yield a more efficient statistical estimator (i.e. smaller mean-squared error). The error-weighted trend filtering estimator is the solution to the following minimization problem:

$$\min_{\{m(t_\mathrm{i})\}} \sum_{i=1}^{n} \big(f(t_\mathrm{i}) - m(t_\mathrm{i})\big)^2 w_\mathrm{i} + \gamma \sum_{i=1}^{n-k-1} |\Delta^{(k+1)}m(t_\mathrm{i})| \cdot \Delta t, \tag{36}$$

where the optimal choice of weights is $w_\mathrm{i} = \sigma_\mathrm{i}^{-2}$, $i = 1, \ldots, n$. Much of the publically available software for trend filtering allows for a heteroskedastic weighting scheme (see Section 3.4).

3.4 Software

Trend filtering software is available online across various platforms. For the specialized ADMM algorithm of Ramdas & Tibshirani (2016) that we utilize in this work, implementations are available in R and C (Arnold, Sadhanala & Tibshirani 2014), as well as JULIA (Kornblith 2014). MATLAB and PYTHON implementations are available for the primal-dual interior point method of Kim et al. (2009), but only for equally weighted linear trend filtering (Koh, Kim & Boyd 2008; Diamond & Boyd 2016). We provide links to our recommended implementations in Table 2. Note that in all software implementations the trend filtering hyperparameter is called λ instead of γ, which we use here to avoid ambiguity with the notation for wavelength in our spectroscopic analyses in Paper II.

3.5 Choosing the hyperparameter

The choice of the piecewise polynomial order k generally has minimal effect on the performance of the trend filtering estimator in terms of mean-squared error and therefore can be treated as an a priori aesthetic choice based on how much smoothness is desired or believed to be present. For example, we use $k = 2$ (quadratic trend filtering) throughout our analyses in Paper II so that the fitted curves are smooth, i.e. differentiable everywhere.

Given the choice of k, the hyperparameter $\gamma > 0$ is used to tune the complexity (i.e. the wiggliness) of the trend filtering estimate by weighting the trade-off between the complexity of the estimate and the size of the squared residuals. Obtaining an accurate estimate is therefore intrinsically tied to finding an optimal choice of γ. The selection of γ is typically done by minimizing an estimate of the MSPE of the trend filtering estimator. Here, there are two different notions of error to consider, namely, *fixed-input* error and *random-input* error. As the names suggest, the distinction between which type of error to consider is made based on how the inputs are sampled. As a general rule-of-thumb, we recommend optimizing with respect to fixed-input error when the inputs are regularly sampled and optimizing with respect to random-input error on irregularly sampled data.

Recall the DGP stated in equation (18) and let it be denoted by Q so that $\mathbb{E}_\mathrm{Q}[\cdot]$ is the mathematical expectation with respect to the randomness of the DGP. Further, let $\sigma_i^2 = \mathrm{Var}(\epsilon_i)$. The fixed-input MSPE is given by

$$R(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\mathrm{Q}\Big[\big(f(t_\mathrm{i}) - \widehat{f}_0(t_\mathrm{i};\gamma)\big)^2 \,\Big|\, t_1, \ldots, t_\mathrm{n}\Big] \tag{37}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \Big(\mathbb{E}_\mathrm{Q}\Big[\big(f_0(t_\mathrm{i}) - \widehat{f}_0(t_\mathrm{i};\gamma)\big)^2 \,\Big|\, t_1, \ldots, t_\mathrm{n}\Big] + \sigma_\mathrm{i}^2\Big) \tag{38}$$

and the random-input MSPE is given by

$$\widetilde{R}(\gamma) = \mathbb{E}_\mathrm{Q}\Big[\big(f(t) - \widehat{f}_0(t;\gamma)\big)^2\Big], \tag{39}$$

where, in the latter, t is considered to be a random component of the DGP with a marginal probability density $p_\mathrm{t}(t)$ supported on the observed input interval. In each case, the theoretically optimal choice of γ is defined as the minimizer of the respective choice of error. Empirically, we estimate the theoretically optimal choice of γ by minimizing an estimate of equation (37) or (39). For fixed-input error we recommend Stein's unbiased risk estimate (SURE; Stein 1981; Efron 1986) and for random-input error we recommend K-fold cross validation with $K = 10$. We elaborate on SURE here and refer the reader to Wasserman (2003) for K-fold cross-validation.

The SURE formula provides an unbiased estimate of the fixed-input MSPE of a statistical estimator:

$$\widehat{R}_0(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \big(f(t_\mathrm{i}) - \widehat{f}_0(t_\mathrm{i};\gamma)\big)^2 + \frac{2\overline{\sigma}^2 \mathrm{df}(\widehat{f}_0)}{n}, \tag{40}$$

[6] A banded matrix only contains non-zero elements in the main diagonal and zero or more diagonals on either side.

Table 1. Comparison of computational costs associated with popular 1D non-parametric regression methods. The computational complexity column states how the number of elementary operations necessary to obtain the fitted values of each estimator (i.e. the estimator evaluated at the observed inputs) scales with the sample size n. For trend filtering, the $\mathcal{O}(n^{1.5})$ complexity represents the worst-case complexity of the Ramdas & Tibshirani (2016) convex optimization algorithm. In most practical settings, the actual complexity of this algorithm is close to $\mathcal{O}(n)$. Variable-knot regression splines require a (non-convex) exhaustive combinatorial search over the set of possible knots and the complexity therefore includes a binomial coefficient term $\binom{n}{p} = n!/(n!(n-p)!)$, where p is the number of knots in the spline. The remaining methods are explicitly solvable and the stated complexity represents the cost of an exact calculation. The $\mathcal{O}(n)$ complexity of wavelets relies on restrictive sampling assumptions (e.g. equally spaced inputs, sample size equal to a power of two). The stated computational complexity of all methods represents the cost of a single model fit and does not include the cost of hyperparameter tuning. Gaussian process regression suffers from the most additional overhead in this regard because of the (often) large number of hyperparameters used to parametrize the covariance function (e.g. shape, range, marginal variance, and noise variance). Each of the non-adaptive methods (linear smoothers) can be made to be locally adaptive (e.g. by locally varying the hyperparameters of the model), but at the expense of greatly increasing the dimensionality of the hyperparameter space to be searched over.

	Method	Computational complexity	Hyperparameters to estimate
Locally adaptive	Wavelets	$\mathcal{O}(n)$	1
	Trend filtering	$\boldsymbol{\mathcal{O}(n^{1.5})}$	**1**
	Variable-knot regression splines	$\mathcal{O}(n \cdot \binom{n}{p})$	1
Non-adaptive	Uniform-knot regression splines	$\mathcal{O}(n)$	1
	Smoothing splines	$\mathcal{O}(n)$	1
	Kernel smoothers	$\mathcal{O}(n^2)$	1
	LOESS	$\mathcal{O}(n^2)$	1
	Gaussian process regression	$\mathcal{O}(n^3)$	3 +

Table 2. Recommended implementations for trend filtering in various programming languages. See Section 3.4 for details. We provide supplementary R code at `github.com/capolitsch/trendfilteringSupp` for selecting the hyperparameter via minimization of SURE (see Section 3.5) and various bootstrap methods for uncertainty quantification (see Section 3.6). Our implementations are built on top of the GLMGEN R package of Arnold et al. (2014).

Language	Recommended implementation
R	github.com/glmgen
C	github.com/glmgen
PYTHON	cvxpy.org
MATLAB	https://stanford.edu/∼boyd/l1_tf/
JULIA	github.com/JuliaStats/Lasso.jl

where $\overline{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\sigma_i^2$ and $\mathrm{df}(\widehat{f}_0)$ is defined in equation (17). A formula for the effective degrees of freedom of the trend filtering estimator is available via the generalized LASSO results of Tibshirani & Taylor (2012); namely,

$$\mathrm{df}(\widehat{f}_0) = \mathbb{E}[\text{number of knots in } \widehat{f}_0] + k + 1. \tag{41}$$

We then obtain our hyperparameter estimate $\widehat{\gamma}$ by minimizing the following plug-in estimate for equation (40):

$$\widehat{R}(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\big(f(t_i) - \widehat{f}_0(t_i;\gamma)\big)^2 + \frac{2\widehat{\overline{\sigma}}^2\,\widehat{\mathrm{df}}(\widehat{f}_0)}{n}, \tag{42}$$

where $\widehat{\mathrm{df}}$ is the estimate for the effective degrees of freedom that is obtained by replacing the expectation in equation (41) with the observed number of knots, and $\widehat{\overline{\sigma}}^2$ is an estimate of $\overline{\sigma}^2$. If a reliable estimate of $\overline{\sigma}^2$ is not available a priori, a data-driven estimate can be constructed (see e.g. Wasserman 2006). We provide a supplementary R package on the corresponding author's GitHub page[7] for implementing SURE with trend filtering. The package is built on top of the GLMGEN R package of Arnold et al. (2014), which already includes an implementation of K-fold cross-validation.

Because of the existence of the degrees of freedom expression (41), trend filtering is also compatible with reduced chi-squared model assessment and comparison procedures under a Gaussian noise assumption (Pearson 1900; Cochran 1952).

3.6 Uncertainty quantification

3.6.1 Frequentist

Frequentist uncertainty quantification for trend filtering follows by studying the sampling distribution of the estimator that arises from the randomness of the observational DGP. In particular, most of the uncertainty in the estimates is captured by studying the variability of the estimator with respect to the DGP. We advise three different bootstrap methods (Efron 1979) for estimating the variability of the trend filtering estimator, with each method corresponding to a distinct analysis setting. Here, we emphasize the terminology *variability* – as opposed to the variance of the trend filtering estimator – since, by construction, as a non-linear function of the observed data, the trend filtering estimator has a non-Gaussian sampling distribution even when the observational noise is Gaussian. For that reason, each of our recommended bootstrap approaches is based on computing

[7] https://github.com/capolitsch/trendfilteringSupp

sample quantiles (instead of pairing standard errors with Gaussian quantiles).

We restate the assumed DGP here for clarity:

$$f(t_i) = f_0(t_i) + \epsilon_i, \quad t_1, \ldots, t_n \in (a, b) \tag{43}$$

where $\mathbb{E}[\epsilon_i] = 0$. We make the further assumption that the errors $\epsilon_1, \ldots, \epsilon_n$ are independent.[8] The three distinct settings we consider are:

S1. The inputs are irregularly sampled

S2. The inputs are regularly sampled and the noise distribution is known

S3. The inputs are regularly sampled and the noise distribution is unknown

The corresponding bootstrap methods are detailed in Algorithm 1 (non-parametric bootstrap; Efron 1979), Algorithm 2 (parametric bootstrap; Efron & Tibshirani 1986), and Algorithm 3 (wild bootstrap; Wu 1986; Liu 1988; Mammen 1993), respectively. We include implementations of each of these algorithms in the R package on our GitHub page.

Algorithm 1 NONPARAMETRIC BOOTSTRAP FOR RANDOM-INPUT UNCERTAINTY QUANTIFICATION

```
Require: Training Data $(t_1, f(t_1)), \ldots, (t_n, f(t_n))$, hyperparameters $\gamma$ and $k$, prediction input grid $t'_1, \ldots, t'_m$
for all $b$ in $1 : B$ do
    Define a bootstrap sample of size $n$ by resampling the observed pairs with replacement:
    $(t^*_1, f^*_b(t^*_1)), \ldots, (t^*_n, f^*_b(t^*_n))$
    Let $\widehat{f}^*_b(t'_1), \ldots, \widehat{f}^*_b(t'_m)$ denote the trend filtering estimate fit on the bootstrap sample and evaluated on the prediction grid $t'_1, \ldots, t'_m$
end for
Output: The full trend filtering bootstrap ensemble
    $\{\widehat{f}^*_b(t'_i)\}_{\substack{i=1,\ldots,m \\ b=1,\ldots,B}}$
```

Given the full trend filtering bootstrap ensemble provided by the relevant bootstrap algorithm, for any $\alpha \in (0, 1)$, a $(1 - \alpha) \cdot 100$ per cent quantile-based pointwise variability band is given by

$$V_{1-\alpha}(t'_i) = \left(\widehat{f}^*_{\alpha/2}(t'_i),\ \widehat{f}^*_{1-\alpha/2}(t'_i)\right), \quad i = 1, \ldots, m \tag{44}$$

where

$$\widehat{f}^*_{\beta}(t'_i) = \inf_g \left\{ g : \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left\{ \widehat{f}^*_b(t'_i) \leq g \right\} \geq \beta \right\}, \quad \beta \in (0, 1). \tag{45}$$

Analogously, bootstrap sampling distributions and variability intervals for observable parameters of the signal may be studied by deriving a bootstrap parameter estimate from each trend filtering estimate within the bootstrap ensemble. For example, in Paper II, we examine the bootstrap sampling distributions of several observable light-curve parameters of exoplanet transits and supernovae.

Algorithm 2 PARAMETRIC BOOTSTRAP FOR FIXED-INPUT UNCERTAINTY QUANTIFICATION (WHEN NOISE DISTRIBUTION $\epsilon_i \sim Q_i$ IS KNOWN *a priori*)

```
Require: Training Data $(t_1, f(t_1)), \ldots, (t_n, f(t_n))$, hyperparameters $\gamma$ and $k$, assumed noise distribution $\epsilon_i \sim Q_i$, prediction input grid $t'_1, \ldots, t'_m$
Compute the trend filtering point estimate at the observed inputs:
$(t_1, \widehat{f}_0(t_1)), \ldots, (t_n, \widehat{f}_0(t_n))$
for all $b$ in $1 : B$ do
    Define a bootstrap sample by sampling from the assumed noise distribution:
    $f^*_b(t_i) = \widehat{f}_0(t_i) + \epsilon^*_i \quad$ where $\epsilon^*_i \sim Q_i, \quad i = 1, \ldots, n$
    Let $f^*_b(t'_1), \ldots, f^*_b(t'_m)$ denote the trend filtering estimate fit on the bootstrap sample and evaluated on the prediction grid $t'_1, \ldots, t'_m$
end for
Output: The full trend filtering bootstrap ensemble
    $\{\widehat{f}^*_b(t'_i)\}_{\substack{i=1,\ldots,m \\ b=1,\ldots,B}}$
```

Algorithm 3 WILD BOOTSTRAP FOR FIXED-INPUT UNCERTAINTY QUANTIFICATION (WHEN NOISE DISTRIBUTION IS NOT KNOWN *a priori*)

```
Require: Training Data $(t_1, f(t_1)), \ldots, (t_n, f(t_n))$, hyperparameters $\gamma$ and $k$, prediction input grid $t'_1, \ldots, t'_m$
Compute the trend filtering point estimate at the observed inputs:
$(t_1, \widehat{f}_0(t_1)), \ldots, (t_n, \widehat{f}_0(t_n))$
Let $\widehat{\epsilon}_i = f(t_i) - \widehat{f}_0(t_i)$, $i = 1, \ldots, n$ denote the residuals
for all $i$ do
    Define a bootstrap sample by sampling from the following distribution:
    $f^*_b(t_i) = \widehat{f}_0(t_i) + u^*_i \quad i = 1, \ldots, n$
    where
    $u^*_i = \begin{cases} \widehat{\epsilon}_i(1 + \sqrt{5})/2 & \text{with probability } (1 + \sqrt{5})/(2\sqrt{5}) \\ \widehat{\epsilon}_i(1 - \sqrt{5})/2 & \text{with probability } (\sqrt{5} - 1)/(2\sqrt{5}) \end{cases}$
    Let $f^*_b(t'_1), \ldots, f^*_b(t'_m)$ denote the trend filtering estimate fit on the bootstrap sample and evaluated on the prediction grid $t'_1, \ldots, t'_m$
end for
Output: The full trend filtering bootstrap ensemble
    $\{\widehat{f}^*_b(t'_i)\}_{\substack{i=1,\ldots,m \\ b=1,\ldots,B}}$
```

3.6.2 *Bayesian*

There is a well-studied connection between ℓ_1-penalized least-squares regression and a Bayesian framework (see e.g. Tibshirani 1996; Figueiredo 2003; Park & Casella 2008). A discussion specific to trend filtering can be found in Faulkner & Minin (2018).

3.7 Relaxed trend filtering

We are indebted to Ryan Tibshirani for a private conversation that motivated the discussion in this section. Trend filtering can be generalized to allow for greater flexibility through a technique

[8] If non-trivial autocorrelation exists in the noise then a block bootstrap (Kunsch 1989) will yield a better approximation of the trend filtering variability than the bootstrap implementations we discuss.

that we call *relaxed trend filtering*.[9] Although the traditional trend filtering estimator is already highly flexible, there are certain settings in which the relaxed trend filtering estimator provides non-trivial improvements. In our experience, these typically correspond to settings where the optimally tuned trend filtering estimator selects very few knots. For example, we use relaxed trend filtering in Paper II to model the detrended, phase-folded light curve of a Kepler star with a planetary transit event.

The relaxed trend filtering estimate is defined through a two-stage sequential procedure in which the first stage amounts to computing the traditional trend filtering estimate discussed in Section 3.2. Recall the trend filtering minimization problem (31). For any given order $k \in \{0, 1, 2, \dots\}$ and hyperparameter $\gamma > 0$, let us amend our notation so that

$$\widehat{f}_0^{\,\mathrm{TF}}(t) = \sum_{j=1}^{n} \widehat{\beta}_{\mathrm{j}}^{\,\mathrm{TF}} h_{\mathrm{j}}(t) \tag{46}$$

denotes the basis representation of the traditional trend filtering estimate. Further, define the index set

$$\mathcal{K}_\gamma = \left\{ 1 \le j \le n \mid \widehat{\beta}_{\mathrm{j}}^{\,\mathrm{TF}} \neq 0 \right\} \tag{47}$$

that includes the indices of the non-zero falling factorial basis coefficients for the given choice of γ. Now let $\widehat{\beta}_{\mathrm{j}}^{\,\mathrm{OLS}}$, $j \in \mathcal{K}_\gamma$, denote the solution to the OLS minimization problem

$$\min_{\{\beta_j\}} \sum_{i=1}^{n} \left(f(t_{\mathrm{i}}) - \sum_{j \in \mathcal{K}_\gamma} \beta_{\mathrm{j}} h_{\mathrm{j}}(t_{\mathrm{i}}) \right)^2, \tag{48}$$

and define the corresponding OLS estimate as

$$\widehat{f}_0^{\,\mathrm{OLS}}(t) = \sum_{j \in \mathcal{K}_\gamma} \widehat{\beta}_{\mathrm{j}}^{\,\mathrm{OLS}} h_{\mathrm{j}}(t). \tag{49}$$

That is, the OLS estimate (49) uses trend filtering to find the knots in the piecewise polynomial, but then uses OLS to estimate the reduced set of basis coefficients. The relaxed trend filtering estimate is then defined as a weighted average of the traditional trend filtering estimate and the corresponding OLS estimate:

$$\widehat{f}_0^{\,\mathrm{RTF}}(t) = \phi \widehat{f}_0^{\,\mathrm{TF}}(t) + (1 - \phi) \widehat{f}_0^{\,\mathrm{OLS}}(t), \tag{50}$$

for some choice of relaxation hyperparameter $\phi \in [0, 1]$. Relaxed trend filtering is therefore a generalization of trend filtering in the sense that the case $\phi = 1$ returns the traditional trend filtering estimate.

In principle, it is preferable to jointly optimize the trend filtering hyperparameter γ and the relaxation hyperparameter ϕ, e.g. via cross-validation. However, it often suffices to choose γ and ϕ sequentially, which in turn adds minimal computational cost on top of the traditional trend filtering procedure. Because of the trivial proximity of the falling factorial basis to the truncated power basis (established in Tibshirani 2014 and Wang et al. 2014), it is sufficient to let $\widehat{f}_0^{\,\mathrm{OLS}}$ be the kth-order regression spline with knots at the input locations selected by the trend filtering estimator. In heteroskedastic settings, as discussed in Section 2.1.2, a piecewise polynomial or regression spline fit by weighted least-squares should be used in place of the OLS estimate (49).

4 CONCLUDING REMARKS

The analysis of 1D data arising from signals possessing varying degrees of smoothness is central to a wide variety of problems in time-domain astronomy and astronomical spectroscopy. Trend filtering is a modern statistical tool that provides a unique combination of (1) statistical optimality for estimating signals with varying degrees of smoothness; (2) natural flexibility for handling practical analysis settings (general sampling designs, heteroskedastic noise distributions, etc.); (3) practical computational speed that scales to massive data sets; and (4) a single model hyperparameter that can be chosen via automatic data-driven methods. Software for trend filtering is freely available online across various platforms and we provide links to our recommendations in Table 2. Additionally, we make supplementary R code available on the corresponding author's GitHub page[10] for: (1) selecting the trend filtering hyperparameter by minimizing SURE (see Section 3.5); and (2) various bootstrap methods for trend filtering uncertainty quantification (see Section 3.6).

ACKNOWLEDGEMENTS

We gratefully thank Ryan Tibshirani for his inspiration and generous feedback on this topic. This work was partially supported by the National Aeronautics and Space Administration (NASA) and the National Science Foundation (NSF), under NASA ATP grant NNX17AK56G, NASA ATP grant 80NSSC18K1015, and NSF grant AST1615940.

REFERENCES

Aigrain S., Parviainen H., Pope B. J. S., 2016, MNRAS, 459, 2408
Arnold T. B., Tibshirani R. J., 2016, J. Comput. Graph. Stat., 25, 1
Arnold T. B., Sadhanala V., Tibshirani R. J., 2014, Fast Algorithms for Generalized Lasso Problems. https://github.com/glmgen (accessed 2020 January 9).
Bautista J. E. et al., 2015, J. Cosmol. Astropart. Phys., 1505, 060
Bolton A. S. et al., 2012, AJ, 144, 144
Cochran W. G., 1952, Ann. Math. Stat., 23, 315
Contreras C. et al., 2010, AJ, 139, 519
Croft R. A. C. et al., 2002, ApJ, 581, 20
De Boor C., 1974, in Conf. Numerical Solution of Differential Equations, Springer, Berlin, Heidelberg, p. 12
De Boor C., 1978, in Applied Mathematical Sciences, Springer-Verlag.
Dhawan S., Leibundgut B., Spyromilio J., Maguire K., 2015, MNRAS, 448, 1345
Diamond S., Boyd S., 2016, J. Mach. Learn. Res., 17, 1
Dimatteo I., Genovese C. R., Kass R. E., 2001, Biometrika, 88, 1055
Dimitriadis G. et al., 2017, MNRAS, 468, 3798
Donoho D. L., Johnstone I. M., 1994, Probab. Theor. Relat. Fields, 99, 277
Donoho D. L., Johnstone I. M., 1998, Ann. Stat., 26, 879
Efron B., 1979, Ann. Stat., 7, 1
Efron B., 1986, J. Am. Stat. Assoc., 81, 461
Efron B., Tibshirani R., 1986, Stat. Sci., 1, 54
Fan J., 1993, Ann. Stat., 21, 196
Fan J., Gijbels I., 1992, Ann. Stat., 20, 2008
Fan J., Gijbels I., 1995, J. R. Stat. Soc. Ser. B (Methodol.), 57, 371
Fan J., Gasser T., Gijbels I., Brockmann M., Engel J., 1997, Anna. Inst. Stat. Math., 49, 79
Faulkner J. R., Minin V. N., 2018, Bayesian Anal., 13, 225
Figueiredo M. A. T., 2003, IEEE Trans. Pattern Anal. Mach. Intell., 25, 1150
Fligge M., Solanki S. K., 1997, A&AS, 124, 579

[9] We choose this term because the generalization of trend filtering to relaxed trend filtering is analogous to the generalization of the LASSO (Tibshirani 1996) to the relaxed LASSO (Meinshausen 2007).

[10] https://github.com/capolitsch/trendfilteringSupp

Gibson N. P., Aigrain S., Roberts S., Evans T. M., Osborne M., Pont F., 2012, MNRAS, 419, 2683

Gijbels I., Mammen E., 1998, Scand. J. Stat., 25, 503

Golkhou V. Z., Butler N. R., 2014, ApJ, 787, 90

Gómez-Valent A., Amendola L., 2018, J. Cosmol. Astropart. Phys., 2018, 051

Györfi L., Kohler M., Krzyzak A., Walk H., 2002, A Distribution-Free Theory of Nonparametric Regression. Springer Series in Statistics, New York

Hall P. B. et al., 2002, ApJS, 141, 267

Hastie T., Tibshirani R., Friedman J., 2009, The Elements of Statistical Learning: Data Mining, Inference and Prediction, second edn. Springer, New York

Howell D. A. et al., 2005, ApJ, 634, 1190

Ibragimov I. A., Hasminiskii R. Z., 1980, Zap. Nauch. Seminar. LOMI (in Russian), 97, 88

Jupp D. L., 1978, SIAM J. Numer. Anal., 15, 328

Kim S.-J. et al., 2009, SIAM Rev., 51, 339

Koh K., Kim S.-J., Boyd S., 2008, l1_tf: Software for l1 Trend Filtering. https://stanford.edu/∼boyd/l1_tf/ (accessed 2019 August 14).

Kornblith S., 2014, Lasso/Elastic Net Linear and Generalized Linear Models. https://github.com/JuliaStats/Lasso.jl (accessed 2019 August 14).

Kovács G., Bakos G., Noyes R. W., 2005, MNRAS, 356, 557

Kunsch H. R., 1989, Ann. Stat., 17, 1217

Land S., Friedman J., 1996, Technical Report, Variable Fusion: A New Method of Adaptive Signal Regression. Department of Statistics, Stanford University, Stanford, CA

Lepski O. V., Mammen E., Spokoiny V. G., 1997, Ann. Stat., 25, 929

Liu R., 1988, Ann. Stat., 16, 1696

Mammen E., 1993, Ann. Stat., 21, 255

Maron J. L., Howes G. G., 2003, ApJ, 595, 564

Meinshausen N., 2007, Comput. Stat. Data Anal., 52, 374

Muller H.-G., Stadtmuller U., 1987, Ann. Stat., 15, 610

Nemirovskii A., 1985, Izv. Akad. Nauk. SSSR Tekhn. Kibernet. (in Russian), 3, 50

Nemirovskii A., Polyak B., Tsybakov A., 1985, Probl. Inform. Transm., 21

Nussbaum M., 1985, Ann. Stat., 13, 984

Paciorek C. J., Schervish M. J., 2004, Adv. Neural Inform. Process. Syst., 16, 273

Paciorek C. J., Schervish M. J., 2006, Environmetrics, 17, 483

Park T., Casella G., 2008, J. Am. Stat. Assoc., 103, 681

Pearson K., 1900, London, Edinburgh, Dublin Philos. Mag. J. Sci., 50, 157

Peiris H. V., Verde L., 2010, Phys. Rev. D, 81, 021302

Persson S. E., Madore B. F., Krzemiński W., Freedman W. L., Roth M., Murphy D. C., 2004, AJ, 128, 2239

Politsch C. A., Cisewski-Kehe J., Croft R. A. C., Wasserman L., 2020, MNRAS, 492, 4019

Ramdas A., Tibshirani R. J., 2016, J. Comput. Graph. Stat., 25, 839

Rudin L. I., Osher S., Faterni E., 1992, Phys. D: Nonlinear Phenom., 60, 259

Schmidt A. M., O'Hagan A., 2003, J. R. Stat. Soc. Ser. B (Stat. Methodol.), 65, 743

Steidl G., Didas S., Neumann J., 2006, Int. J. Comput. Vis., 70, 241

Stein C. M., 1981, Ann. Stat., 9, 1135

Stone C. J., 1982, Ann. Stat., 10, 1040

Tennyson J., 2019, Astronomical Spectroscopy: An Introduction to the Atomic and Molecular Physics of Astronomical Spectroscopy. World Scientific Publishing, University College London, UK

Theuns T., Zaroubi S., 2000, MNRAS, 317, 989

Tibshirani R., 1996, J. R. Stat. Soc. Ser. B (Methodol.), 58, 267

Tibshirani R. J., 2014, Ann. Stat., 42, 285

Tibshirani R. J., 2015, Stat. Sinica, 25. p. 1265

Tibshirani R. J., Taylor J., 2011, Ann. Stat., 39, 1335

Tibshirani R. J., Taylor J., 2012, Ann. Stat., 40, 1198

Tibshirani R., Saunders M., Rosset S., Zhu J., Knight K., 2005, J. R. Stat. Soc. Ser. B, 67, 91

Tolstov A., Nomoto K., Sorokina E., Blinnikov S., Tominaga N., Taniguchi Y., 2019, ApJ, 881, 35

Tsybakov A. B., 2008, Introduction to Nonparametric Estimation, 1st edn. Springer Publishing Company, Incorporated, New York

Van der Vaart A. W., 1998, Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge Univ. Press, Cambridge

Wahba G., 1990, Spline Models for Observational Data (CBMS-NSF Regional Conference Series in Applied Mathematics). Society for Industrial and Applied Mathematics, Philadelphia, PA

Wang Y.-X. et al., 2016, J. Mach. Learn. Res., 17, 1

Wang Y.-X., Smola A., Tibshirani R., 2014, in Xing E. P., Jebara T., eds, Proceedings of Machine Learning Research Vol. 32, Proceedings of the 31st International Conference on Machine Learning. PMLR, Bejing, China, p. 730

Wasserman L., 2003, All of Statistics: A Concise Course in Statistical Inference. Springer Publishing Company, Incorporated, New York

Wasserman L., 2006, All of Nonparametric Statistics. Springer Texts in Statistics, Springer-Verlag, Berlin.

Woosley S. E., Kasen D., Blinnikov S., Sorokina E., 2007, ApJ, 662, 487

Wu C., 1986, Ann. Stat., 14, 1261

This paper has been typeset from a TEX/LATEX file prepared by the author.