Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Bayesian Trend Filtering via Proximal Markov Chain Monte Carlo

Qiang Heng, Hua Zhou & Eric C. Chi

To cite this article: Qiang Heng, Hua Zhou & Eric C. Chi (2023): Bayesian Trend Filtering via Proximal Markov Chain Monte Carlo, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2023.2170089

To link to this article: https://doi.org/10.1080/10618600.2023.2170089

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2023, VOL. 00, NO. 0, 1–12
https://doi.org/10.1080/10618600.2023.2170089

Bayesian Trend Filtering via Proximal Markov Chain Monte Carlo

Qiang Heng [a], Hua Zhou [b], and Eric C. Chi [c]

[a]Department of Statistics, North Carolina State University, Raleigh, NC; [b]Departments of Biostatistics and Computational Medicine, UCLA, Los Angeles, CA; [c]Department of Statistics, Rice University, Houston, TX

ABSTRACT

Proximal Markov chain Monte Carlo is a novel construct that lies at the intersection of Bayesian computation and convex optimization, which helped popularize the use of nondifferentiable priors in Bayesian statistics. Existing formulations of proximal MCMC, however, require hyperparameters and regularization parameters to be prespecified. In this article, we extend the paradigm of proximal MCMC through introducing a novel new class of nondifferentiable priors called epigraph priors. As a proof of concept, we place trend filtering, which was originally a nonparametric regression problem, in a parametric setting to provide a posterior median fit along with credible intervals as measures of uncertainty. The key idea is to replace the nonsmooth term in the posterior density with its Moreau-Yosida envelope, which enables the application of the gradient-based MCMC sampler Hamiltonian Monte Carlo. The proposed method identifies the appropriate amount of smoothing in a data-driven way, thereby automating regularization parameter selection. Compared with conventional proximal MCMC methods, our method is mostly tuning free, achieving simultaneous calibration of the mean, scale and regularization parameters in a fully Bayesian framework. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2021
Accepted January 2023

KEYWORDS
Convex optimization; Epigraphs; Moreau-Yosida envelope; Hamiltonian Monte Carlo; Trend filtering

1. Introduction

When analyzing time series data, we are often interested in estimating a slowly varying underlying trend with desired properties such as smoothness and shape restrictions. Smoothness can be achieved by constraining the underlying trend to be piecewise polynomial, while shape restrictions such as monotonicity and convexity can be enforced by linear inequality constraints. Let $\mathbf{y} \in \mathbb{R}^n$ denote an observed time series and $\boldsymbol{\beta} \in \mathbb{R}^n$ denote its underlying trend; then estimating $\boldsymbol{\beta}$ is commonly posed as the following constrained or penalized least squares problem

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}} \ \tfrac{1}{2}\|\mathbf{y}-\boldsymbol{\beta}\|_2^2 + g(\boldsymbol{\beta}), \tag{1}$$

where $g(\boldsymbol{\beta})$ is an indicator function encoding convex constraints or a nonsmooth penalty function inducing sparsity. Different choices of $g(\boldsymbol{\beta})$ induce a variety of sequence approximation problems. Representative examples include isotonic regression (Barlow 1972), univariate convex regression (Groeneboom, Jongbloed, and Wellner 2008), nearly-isotonic regression (Tibshirani, Hoefling, and Tibshirani 2011) and ℓ_1-trend filtering (Steidl, Didas, and Neumann 2006; Kim et al. 2009; Tibshirani 2014).

As a nonparametric regression problem, the solution to (1) only produces a point estimate. If we are interested in uncertainty quantification, data-resampling techniques like the bootstrap (Efron and Tibshirani 1994) can be adopted. The bootstrap, however, does not address the issue of regularization parameter selection. The bootstrap is only able to produce a confidence band with a given regularization parameter, which is often selected with cross-validation.

To quantify uncertainty and automate regularization parameter selection, many have placed (1) in a Bayesian framework. Inspired by the Bayesian Lasso (Park and Casella 2008), Roualdes (2015) first introduced Bayesian Trend Filtering (BTF), exploiting the Gaussian mixture representation of the Laplace prior. Independent from Rouadle's work, Faulkner and Minin (2018) proposed a closely related smoothing method, Shrinkage Prior Markov Random Fields (SPMRFs), which places sparsity inducing shrinkage priors on the adjacent differences of the elements of $\boldsymbol{\beta}$. In addition to the Laplace prior, Faulkner and Minin (2018) also investigated a more aggressive horseshoe prior (Carvalho, Polson, and Scott 2010), which demonstrated superior local adaptivity to abrupt changes or jumps. Recently, Kowal, Matteson, and Ruppert (2019) proposed dynamic shrinkage processes (DSP) for Bayesian trend filtering with even stronger localized adaptivity to irregular features through modeling dependence between the local scale parameters.

The literature of Bayesian shape-restricted regression is vast and diverse. Early works include Bayesian isotonic regression with piecewise linear models (Neelon and Dunson 2004), Bayesian P-splines (Brezger and Steiner 2008), Bayesian monotone regression with Bernstein polynomials (McKay Curtis and Ghosh 2011). Two more recent methods are Bayesian shape-restricted splines (Meyer, Hackstadt, and Hoeting 2011) and Bayesian shape-restricted regression using Gaussian process

CONTACT Qiang Heng qheng@ncsu.edu Department of Statistics, North Carolina State University, Raleigh, NC.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

priors (Lenk and Choi 2017), which can enforce both monotonicity and convexity.

Our approach to Bayesian trend filtering takes advantage of a relatively new Markov chain Monte Carlo (MCMC) sampling scheme in the Bayesian imaging literature, namely the proximal MCMC methods (Pereyra 2016; Durmus, Moulines, and Pereyra 2018; Pereyra, Mieles, and Zygalakis 2020). The current paradigm of proximal MCMC methods requires variance and regularization parameters to be fixed and predetermined. In this work, we incorporate those parameters into posterior inference, leveraging the data itself to automatically determine the appropriate amount of smoothing. We present two applications of our proposed methodology, namely Proximal Bayesian Trend Filtering (PBTF) and Proximal Bayesian Shape-Restricted Trend Filtering (PBSRTF).

2. Background

We first review the nonparameteric function estimation with ℓ_1-trend filtering as well as important concepts from convex optimization needed to develop our Bayesian trend filtering algorithms.

2.1. Nonparametric Estimation with ℓ_1-trend Filtering

Suppose that a time series $\mathbf{y} \in \mathbb{R}^n$ observed over a grid of time points $\mathbf{x} \in \mathbb{R}^n$ is the superposition of a smooth trend $\boldsymbol{\beta} \in \mathbb{R}^n$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$, namely

$$y_i = \beta_i + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

where the grid locations x_i are strictly increasing, that is, $x_1 < x_2 < \cdots < x_n$. For simplicity, we assume for now that a single measurement is observed at each grid point and the grid points are evenly spaced. We relax both assumptions later.

Kim et al. (2009) proposed ℓ_1-trend filtering to estimate $\boldsymbol{\beta}$ with piecewise polynomial structure, by solving the following regularized least squares problem

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\text{minimize}} \ \tfrac{1}{2} \|\mathbf{y} - \boldsymbol{\beta}\|_2^2 + \alpha \|\mathbf{D}_n^{(k+1)} \boldsymbol{\beta}\|_1, \tag{3}$$

where α is a positive regularization parameter, $\mathbf{D}_n^{(k+1)} \in \mathbb{R}^{(n-k-1) \times n}$ is the discrete difference operator or matrix of order $k+1$ and dimension n. To appreciate the effect of penalizing the ℓ_1-norm of $\mathbf{D}_n^{(k+1)} \boldsymbol{\beta}$, we explicitly write out the difference operator for $k = 0$,

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

When $k = 0$, the penalty term $\|\mathbf{D}_n^{(1)} \boldsymbol{\beta}\|_1 = \sum_{i=1}^{n-1} |\beta_{i+1} - \beta_i|$ is also known as the one-dimensional total variation denoising penalty (Rudin, Osher, and Fatemi 1992; Steidl, Didas, and Neumann 2006) in signal processing, or the fused lasso penalty (Tibshirani et al. 2005) in statistics. The penalty incentivizes recovery of piecewise constant solutions. Higher-order difference matrices are defined recursively as $\mathbf{D}_n^{(k+1)} = \mathbf{D}_{n-k}^{(1)} \mathbf{D}_n^{(k)}$. Choosing order $k = 1, 2$, and 3 incentivizes the recovery of piecewise linear, quadratic, and cubic solutions, respectively. Difference matrices of order higher than 4 are rarely of interest.

To handle irregular grids, namely when the time points $\mathbf{x} \in \mathbb{R}^n$ are strictly increasing but possibly unevenly spaced, Tibshirani (2014) proposed replacing $\mathbf{D}_n^{(k+1)}$ with the adjusted difference matrix $\mathbf{D}_n^{(\mathbf{x},k+1)}$. The first-order difference matrix remains the same, that is, $\mathbf{D}_n^{(\mathbf{x},1)} = \mathbf{D}_n^{(1)}$; for $k \geq 1$ the adjusted difference operators are now defined as

$$\mathbf{D}_n^{(\mathbf{x},k+1)} = \mathbf{D}_{n-k}^{(\mathbf{x},1)} \operatorname{diag}\left(\frac{k}{x_{k+1} - x_1}, \ldots, \frac{k}{x_n - x_{n-k}} \right) \mathbf{D}_n^{(\mathbf{x},k)} \quad \text{for } k = 1, 2, \ldots$$

Note when $x_1 = 1, x_2 = 2, \ldots, x_n = n$, the adjusted difference matrix $\mathbf{D}_n^{(\mathbf{x},k+1)}$ coincides with $\mathbf{D}_n^{(k+1)}$.

A variety of iterative and non-iterative algorithms have been proposed to compute a solution to (3). The ones that are relevant to this work are the dynamic programming algorithm by Johnson (2013) and the ADMM algorithm by Ramdas and Tibshirani (2016). Remarkably, the dynamic programming approach can solve (3) exactly in $O(n)$ steps for $k = 0$. Building on top of the dynamic programming algorithm, the ADMM algorithm solves (3) iteratively for $k = 1, 2$, and 3.

As discussed in Kim et al. (2009), adding additional shape restrictions to ℓ_1-trend filtering is straightforward. For example, one might require the underlying trend to be monotone-increasing. The isotonic ℓ_1-trend filtering problem is formulated as

$$\begin{aligned} &\underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2} \|\mathbf{y} - \boldsymbol{\beta}\|_2^2 + \alpha \|\mathbf{D}_n^{(\mathbf{x},k+1)} \boldsymbol{\beta}\|_1 \\ &\text{subject to} \quad \beta_1 \leq \beta_2 \leq \cdots \leq \beta_n. \end{aligned}$$

The monotonicity constraint $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$ can be written compactly as $\mathbf{D}_n^{(1)} \boldsymbol{\beta} \geq \mathbf{0}$, where $\geq$ represents elementwise inequality.

In addition to monotonicity, another common shape restriction is convexity. The underlying trend $\boldsymbol{\beta}$ is convex if

$$\frac{\beta_2 - \beta_1}{x_2 - x_1} \leq \frac{\beta_3 - \beta_2}{x_3 - x_2} \leq \cdots \leq \frac{\beta_n - \beta_{n-1}}{x_n - x_{n-1}}, \tag{4}$$

which can be written compactly as $\mathbf{D}_n^{(\mathbf{x},2)} \boldsymbol{\beta} \geq \mathbf{0}$.

For the rest of this article, we will work with the general case where we may have multiple observations per grid point. We assume that observations y_{ij} come from the model

$$\begin{aligned} y_{ij} = \beta(x_i) + \epsilon_{ij}, \ \epsilon_{ij} &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \\ i = 1, 2, \ldots, n, \ j &= 1, 2, \ldots, w_i, \end{aligned} \tag{5}$$

where $\beta(x)$ is the underlying trend function that we seek to estimate and w_i is the number of observations at a particular grid location x_i. We assume that the underlying function $\beta(x)$ has piecewise polynomial structure. Allowing multiple observations at a given grid location is useful as real data is often discrete.

2.2. Relevant Concepts from Convex Optimization

We next review concepts from convex optimization central to our proposed framework, specifically projection and proximal mappings which are the algorithmic primitives that we will use to build our Bayesian trend filtering methods.

In convex analysis, the indicator function $\iota_{\mathcal{A}}(\boldsymbol{\beta})$ of a set $\mathcal{A} \subset \mathbb{R}^n$ takes on the value of 0 when $\boldsymbol{\beta} \in \mathcal{A}$ and the value of $+\infty$ when $\boldsymbol{\beta} \notin \mathcal{A}$. The familiar 0–1 indicator function $\mathbb{1}_{\mathcal{A}}(\boldsymbol{\beta})$, which takes on the value of 1 when $\boldsymbol{\beta} \in \mathcal{A}$ and 0 when $\boldsymbol{\beta} \notin \mathcal{A}$ is an invertible transformation the indicator function from convex analysis, namely $\mathbb{1}_{\mathcal{A}}(\boldsymbol{\beta}) = \exp(-\iota_{\mathcal{A}}(\boldsymbol{\beta}))$. The projection of a point $\boldsymbol{\beta}$ onto a set $\mathcal{A}$, denoted by $P_{\mathcal{A}}(\boldsymbol{\beta})$, is a point in $\mathcal{A}$ that is closest in Euclidean distance to $\boldsymbol{\beta}$.

$$P_{\mathcal{A}}(\boldsymbol{\beta}) = \underset{\boldsymbol{\eta} \in \mathcal{A}}{\arg\min} \, \|\boldsymbol{\eta} - \boldsymbol{\beta}\|_2.$$

The projection $P_{\mathcal{A}}(\boldsymbol{\beta})$ exists and is unique when $\mathcal{A}$ is closed and convex,.

The proximal map of the function g is the following operator

$$\text{prox}_g(\boldsymbol{\beta}) = \underset{\boldsymbol{\eta} \in \mathbb{R}^n}{\text{argmin}} \left[g(\boldsymbol{\eta}) + \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\eta}\|_2^2 \right].$$

An additional positive parameter λ is often added to control proximity,

$$\text{prox}_{\lambda g}(\boldsymbol{\beta}) = \underset{\boldsymbol{\eta} \in \mathbb{R}^n}{\text{argmin}} \left[g(\boldsymbol{\eta}) + \frac{1}{2\lambda}\|\boldsymbol{\beta} - \boldsymbol{\eta}\|_2^2 \right].$$

Following the notation in prior proximal MCMC papers, we write $\text{prox}_{\lambda g}(\boldsymbol{\beta})$ as $\text{prox}_g^{\lambda}(\boldsymbol{\beta})$.

When g is an indicator function of a set $\mathcal{A}$, the proximal operator is the projection onto $\mathcal{A}$. Consequently, proximal maps generalize projection operations. Proximal maps play an important role in modern machine learning due to the fact that many nonsmooth penalties often have unique proximal mappings that either have explicit formulas or can be computed efficiently. In this article, we take advantage of two such proximal mappings, namely the proximal maps of $\|\boldsymbol{\beta}\|_1$ and $\|\mathbf{D}_n^{(1)}\boldsymbol{\beta}\|_1$. The proximal map of $\|\boldsymbol{\beta}\|_1$ is the celebrated soft-threshold operator

$$\left[\text{prox}_g^{\lambda}(\boldsymbol{\beta})\right]_i = \begin{cases} \beta_i & |\beta_i| \leq \lambda \\ \text{sgn}(\beta_i)(|\beta_i| - \lambda)_+ & \beta_i| > \lambda \end{cases}, \tag{6}$$

while the proximal map of $\|\mathbf{D}_n^{(1)}\boldsymbol{\beta}\|_1$ is the solution to the fused Lasso problem (Tibshirani et al. 2005):

$$\text{prox}_g^{\lambda}(\boldsymbol{\beta}) = \underset{\boldsymbol{\eta} \in \mathbb{R}^n}{\text{argmin}} \, \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\eta}\|_2^2 + \lambda\|\mathbf{D}_n^{(1)}\boldsymbol{\eta}\|_1, \tag{7}$$

which can be solved exactly in linear time via dynamic programming (Johnson 2013). We use these two proximal maps as a subroutine to perform a key computation, namely the epigraph projection, which we will describe later.

The λ-Moreau-Yosida envelope of a function g is given by

$$g^{\lambda}(\boldsymbol{\beta}) = \min_{\boldsymbol{\eta} \in \mathbb{R}^n} g(\boldsymbol{\eta}) + \frac{1}{2\lambda}\|\boldsymbol{\eta} - \boldsymbol{\beta}\|_2^2.$$

The envelope function g^{λ} has several important properties. First, g^{λ} is convex when g is convex. Second, g^{λ} is always differentiable even if g is not, and its gradient can be expressed in terms of the proximal map of λg, namely

$$\nabla g^{\lambda}(\boldsymbol{\beta}) = \frac{1}{\lambda}\left[\boldsymbol{\beta} - \text{prox}_g^{\lambda}(\boldsymbol{\beta})\right].$$

Moreover, ∇g^{λ} is λ^{-1}-Lipschitz since proximal operators are firmly nonexpansive (Combettes and Pesquet 2011). Finally and perhaps most importantly, g^{λ} converges pointwise to g as λ tends to 0 (Rockafellar and Wets 2009). In short, we see that the Moreau-Yosida envelope of a nonsmooth function g is a Lipschitz-differentiable, arbitrarily close approximation to g. In this work, we will rely on the Moreau-Yosida envelope of indicator functions. Since the proximal map of an indicator function $\iota_{\mathcal{E}}(\boldsymbol{\beta})$ is the projection $P_{\mathcal{E}}(\boldsymbol{\beta})$, its Moreau-Yosida envelope is $g^{\lambda}(\boldsymbol{\beta}) = \frac{1}{2\lambda}\|\boldsymbol{\beta} - P_{\mathcal{E}}(\boldsymbol{\beta})\|_2^2$, where $\|\boldsymbol{\beta} - P_{\mathcal{E}}(\boldsymbol{\beta})\|_2$ is also denoted as $d_{\mathcal{E}}(\boldsymbol{\beta})$, namely the distance of $\boldsymbol{\beta}$ to $\mathcal{E}$.

The Moreau-Yosida approximation is the key technical ingredient behind the proximal MCMC framework of Durmus, Moulines, and Pereyra (2018) which our algorithmic framework extends. We next review their prior formulation of the proximal MCMC method.

3. Proximal MCMC

Many modern machine learning applications employ log-concave models of the form

$$\pi(\boldsymbol{\beta}) \propto \exp\{-U(\boldsymbol{\beta})\} \quad \text{and} \quad U(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + g(\boldsymbol{\beta}), \tag{8}$$

where f is a Lipschitz-differentiable convex negative log-likelihood function and g is a lower-semicontinuous convex penalty function that shrinks the estimator toward some desired prior structure. The model in (2) that underlies the ℓ_1-trend-filtering problem is an example of such a log-concave model, where

$$f(\boldsymbol{\beta}) = \frac{1}{2\sigma^2}\|\mathbf{y} - \boldsymbol{\beta}\|_2^2 \quad \text{and} \quad g(\boldsymbol{\beta}) = \alpha\|\mathbf{D}_n^{(k+1)}\boldsymbol{\beta}\|_1.$$

Note that if we absorb σ^2 into the regularization parameter α, then computing the Maximum a Posteriori (MAP) estimate of $\boldsymbol{\beta}$ in this log-concave model is equivalent to solving the nonparameteric problem (3).

Given such a log-concave model, we may wish to facilitate uncertainty quantification and posterior inference by computing posterior samples. Unfortunately, while there are many scalable methods for computing the MAP estimate of $\boldsymbol{\beta}$, for example the Split-Bregman (Goldstein and Osher 2009) and Chambolle-Pock (Chambolle and Pock 2011) algorithms, sampling from the posterior distribution (8) is not as straightforward. Conventional high-dimensional MCMC algorithms, such as the Unadjusted Langevin Algorithm (ULA) (Roberts et al. 1996), Metropolis-Adjusted Langevin Algorithm (MALA) (Rossky, Doll, and Friedman 1978; Roberts et al. 1996), Hamiltonian Monte Carlo (HMC) (Neal 2011), rely on gradient mappings that in turn require U to be Lipschitz-differentiable or at least differentiable. These differentiability requirements can be extremely limiting, as they rule out many commonly used nonsmooth penalty functions g.

To make efficient high-dimensional MCMC algorithms applicable for nonsmooth U, Pereyra (2016) proposed replacing

U with a Lipschitz-differentiable approximation, namely the λ-Moreau-Yosida envelope of U, and then employing MALA to sample from the derived surrogate density (Px-MALA). Durmus, Moulines, and Pereyra (2018) proposed a slightly different strategy with the Moreau-Yosida regularized Unadjusted Langevin Algorithm (MYULA), by replacing g with its Moreau-Yosida approximation g^λ in (8) to obtain the surrogate density

$$\pi^\lambda(\boldsymbol{\beta}) \propto \exp\{-f(\boldsymbol{\beta}) - g^\lambda(\boldsymbol{\beta})\}. \tag{9}$$

Under additional assumptions on g, the surrogate density (9) is proper and converges to the original density (8) in total-variation norm (Durmus, Moulines, and Pereyra 2018). Moreover, if g is Lipschitz, then the total-variation norm of (8) and (9) is bounded linearly in λ. The MYULA algorithm simply applies ULA to the surrogate density (9):

$$\boldsymbol{\beta}_{l+1} = \left(1 - \frac{\gamma}{\lambda}\right)\boldsymbol{\beta}_l - \gamma\nabla f(\boldsymbol{\beta}_l) + \frac{\gamma}{\lambda}\operatorname{prox}_g^\lambda(\boldsymbol{\beta}_l) + \sqrt{2\gamma}\boldsymbol{\zeta}_{l+1}, \tag{10}$$

where $\boldsymbol{\zeta}_{l+1}$ is n-dimensional Brownian motion and γ is the step size of ULA. A Metropolis-Hastings correction step can be added to remove the asymptotic bias associated with Euler-Maruyama discretization that is common to Langevin algorithms. An extension of the MYULA algorithm is to combine several gradient evaluations to accelerate its convergence (SK-ROCK) (Pereyra, Mieles, and Zygalakis 2020). The recent review paper Durmus, Moulines, and Pereyra (2022) provides an overview for proximal MCMC methods and their applications in imaging inverse problems.

A hallmark application of proximal MCMC is Bayesian image deblurring, where $\boldsymbol{\beta}$ is a high-dimensional latent image, f is the negative log-likelihood that models blurring and additive Gaussian noise that together corrupt the latent image, and g is a total variation penalty that incentivizes the recovery of a latent image with sharp edges (Durmus, Moulines, and Pereyra 2018; Pereyra, Mieles, and Zygalakis 2020; Durmus, Moulines, and Pereyra 2022). In this context, the posterior of interest is

$$\pi(\boldsymbol{\beta} \mid \mathbf{y}) \propto \exp\left\{-\frac{\|\mathbf{y} - \mathbf{H}\boldsymbol{\beta}\|_2^2}{2\sigma^2} - \alpha\,\mathrm{TV}(\boldsymbol{\beta})\right\}, \tag{11}$$

where $\mathbf{H}$ is a blur operator, $\mathrm{TV}(\boldsymbol{\beta})$ is the total-variation semi-norm of $\boldsymbol{\beta}$ (Chambolle 2004), $\mathbf{y}$ is the corrupted image signal we observe, σ^2 is the noise variance, and α is a positive regularization parameter that trades off the emphasis between data fit and smoothness in the estimated image. In the framework of Durmus, Moulines, and Pereyra (2018) and Pereyra, Mieles, and Zygalakis (2020), the variance σ^2 and the regularization parameter α need to be manually selected by an expert or determined by an empirical Bayesian method (Vidal et al. 2020; De Bortoli et al. 2020). In this article, we propose to use a new construct that we refer to as epigraph priors and HMC sampling to incorporate σ^2 and α into posterior inference in the context of Bayesian trend filtering. Consequently, this work demonstrates how proximal MCMC can be applied as a statistical methodology in a unified and complete Bayesian framework. Figure 1 illustrates four examples of posterior fits using our fully Bayesian proximal MCMC method for trend filtering.

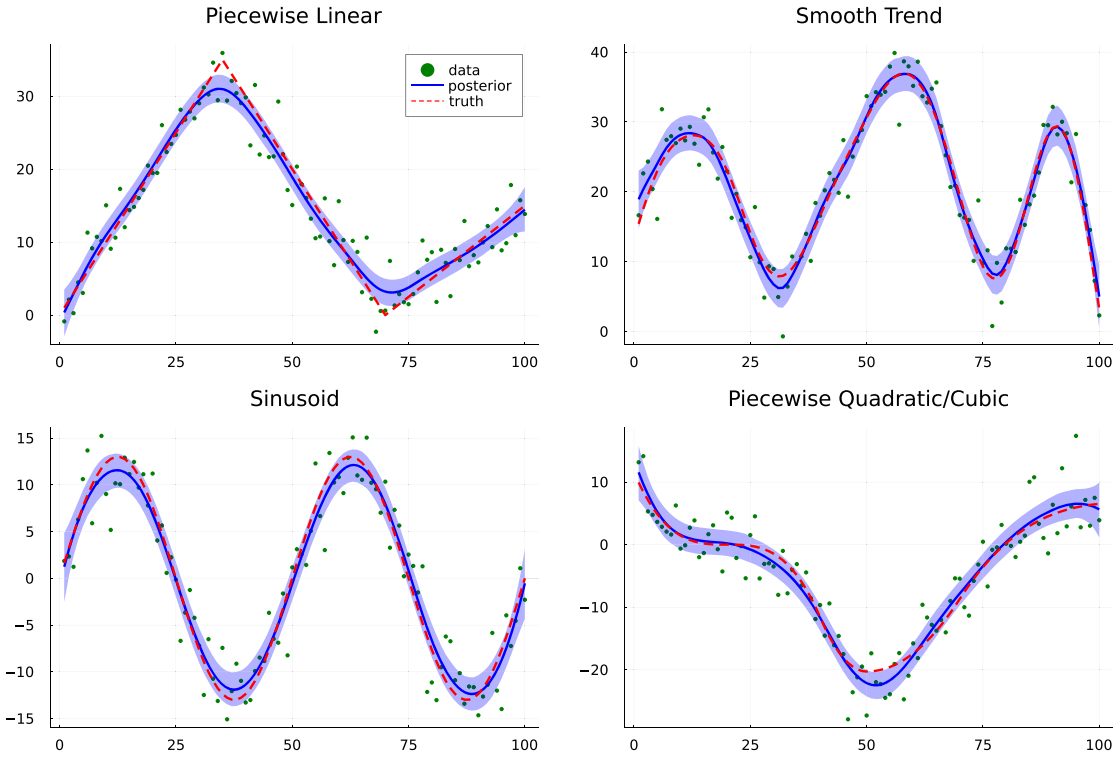

Figure 1. Example posterior fits for PBTF with noise level $\sigma = 3$. The standard deviation of the underlying trends is 9, thus, the signal-to-noise ratio is 3. Plots show data points (green dots), posterior median (blue solid lines), 95% Bayesian credible intervals (light blue bands) and true trends (red dashed lines).

4. Proximal Bayesian Trend Filtering

Our key methodological innovation that enables extending the proximal MCMC framework to a fully Bayesian one is the use of epigraph indicator functions to encode our structure-inducing prior. Prior proximal MCMC methods typically replace a non-smooth penalty $g(\boldsymbol{\beta}) = \alpha h(\boldsymbol{\beta})$ with its Moreau envelope in the posterior. The proximal operator is then evaluated as

$$\text{prox}_g^{\lambda}(\boldsymbol{\beta}) = \text{prox}_h^{\lambda\alpha}(\boldsymbol{\beta}),$$

where the proximal operator of h can be computed with an efficient off-the-shelf algorithm. The gradient of $g^{\lambda}(\boldsymbol{\beta})$ can then be computed as $(\boldsymbol{\beta} - \text{prox}_g^{\lambda}(\boldsymbol{\beta}))/\lambda$, which is a well-known fact about Moreau envelopes. However, the regularization parameter α is viewed as a hyperparameter in g^{λ} and needs to be determined prior to MCMC sampling. Although an empirical Bayesian method (Vidal et al. 2020; De Bortoli et al. 2020) can be used to estimate the appropriate α and σ^2, a fully Bayesian treatment is desirable since it may have better precision due to being able to account for the uncertainty of α and σ^2.

To incorporate α into posterior inference, an important concept in convex analysis, epigraph, comes in handy. The epigraph of a regularization function g is the set

$$\mathcal{E} = \{(\boldsymbol{\beta}, \alpha) \in \mathbb{R}^n \times \mathbb{R} : g(\boldsymbol{\beta}) \leq \alpha\}.$$

The Moreau-Yosida envelope of $\iota_{\mathcal{E}}(\boldsymbol{\beta}, \alpha)$ is $\frac{1}{2\lambda} d^2_{\mathcal{E}}(\boldsymbol{\beta}, \alpha)$, which is jointly differentiable in $\boldsymbol{\beta}$ and α. The gradient of $\frac{1}{2\lambda} d^2_{\mathcal{E}}(\boldsymbol{\beta}, \alpha)$ is simply $\frac{(\boldsymbol{\beta},\alpha) - P_{\mathcal{E}}(\boldsymbol{\beta},\alpha)}{\lambda}$, where $P_{\mathcal{E}}$ denotes projection on to $\mathcal{E}$. Figure 2 provides a visualization of the envelope function $\frac{1}{2\lambda} d^2_{\mathcal{E}}(\boldsymbol{\beta}, \alpha)$ when $\mathcal{E} = \{(\beta, \alpha) \in \mathbb{R}^2 : |\beta| \leq \alpha\}$ and $\lambda = 0.01$. Using $\frac{1}{2\lambda} d^2_{\mathcal{E}}(\boldsymbol{\beta}, \alpha)$ as our prior regularization term, we can further place hyperpriors on α, σ^2 and achieve fully Bayesian inference within the proximal MCMC framework. Computing with these priors relies on projection onto epigraphs which we describe next.

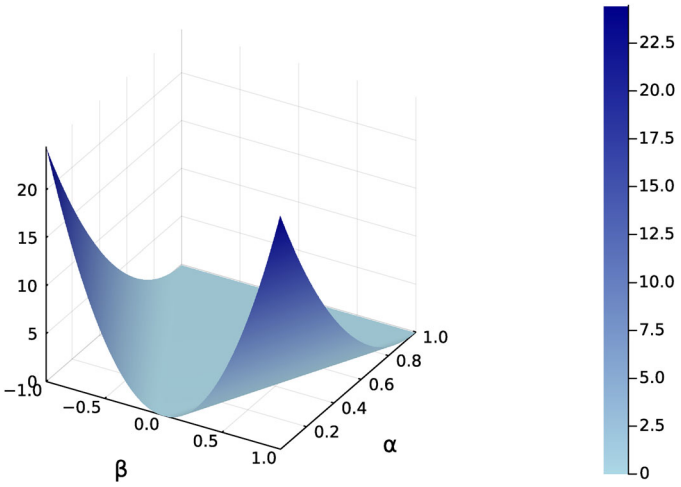

Figure 2. A visualization of the distance function $\frac{1}{2\lambda} d^2_{\mathcal{E}}(\boldsymbol{\beta}, \alpha)$ when $\mathcal{E} = \{(\beta, \alpha) \in \mathbb{R}^2 : |\beta| \leq \alpha\}$ and $\lambda = 0.01$.

4.1. Projection Onto Epigraph

Projection onto the epigraph of g depends on the proximal mapping of g (see Theorem 6.36 of Beck 2017), namely

$$P_{\text{epi}(g)}(\boldsymbol{\beta}, \alpha) = \begin{cases} (\boldsymbol{\beta}, \alpha) & g(\boldsymbol{\beta}) \leq \alpha \\ \left(\text{prox}_g^{\lambda^*}(\boldsymbol{\beta}), \alpha + \lambda^*\right) & g(\boldsymbol{\beta}) > \alpha \end{cases}, \quad (12)$$

where λ^* is root of the auxiliary function

$$F(\lambda) = g\left(\text{prox}_g^{\lambda}(\boldsymbol{\beta})\right) - \lambda - \alpha.$$

When $\text{prox}_g^{\lambda}(\boldsymbol{\beta})$ can be computed easily, we can compute the root λ^* of the function $F(\lambda)$ using a simple bisection procedure.

We will need to perform projections onto two sets: the epigraph of the ℓ_1-norm

$$\mathcal{E}_1 = \{(\boldsymbol{\beta}, \alpha) \in \mathbb{R}^n \times \mathbb{R}_{++} : \|\boldsymbol{\beta}\|_1 \leq \alpha\},$$

and the epigraph of $\|\mathbf{D}_n^{(1)}\boldsymbol{\beta}\|_1$

$$\mathcal{E}_2 = \left\{(\boldsymbol{\beta}, \alpha) \in \mathbb{R}^n \times \mathbb{R}_{++} : \|\mathbf{D}_n^{(1)}\boldsymbol{\beta}\|_1 \leq \alpha\right\}.$$

Since the proximal maps of $\|\boldsymbol{\beta}\|_1$ and $\|\mathbf{D}_n^{(1)}\boldsymbol{\beta}\|_1$ can be computed in linear time, projections onto $\mathcal{E}_1$ and $\mathcal{E}_2$ can be done efficiently. For projection onto $\mathcal{E}_1$, we set the initial bisection interval to be $(0, \lambda_{\max})$ where $\lambda_{\max} = \|\boldsymbol{\beta}\|_\infty$ is the smallest value of λ such that $\text{prox}_{\lambda\|\cdot\|_1}(\boldsymbol{\beta}) = \mathbf{0}$. For projection onto $\mathcal{E}_2$, we set the initial bisection interval to be $(0, \lambda_{\max})$ where

$$\lambda_{\max} = \left\| \left[\mathbf{D}_n^{(1)}(\mathbf{D}_n^{(1)})^{\mathsf{T}}\right]^{-1} \mathbf{D}_n^{(1)}\boldsymbol{\beta} \right\|_\infty,$$

is the smallest value of λ such that the solution to (7) is a multiple of the all ones vector. It is easy to verify that $F(0) > 0$ when $(\boldsymbol{\beta}, \alpha) \notin \text{epi}(g)$ and $F(\lambda_{\max}) < 0$ so that the root of the auxiliary function is guaranteed to lie within $(0, \lambda_{\max})$.

In a manner akin to Ramdas and Tibshirani (2016), projecting onto $\mathcal{E}_2$ instead of projecting onto $\mathcal{E}_1$ alleviates numerical issues associated with solving an ill-conditioned linear system, since it enables us to work with a transformation matrix that is one "order" lower. We will elaborate on this claim in Section 4.2.

4.2. Priors for Proximal Bayesian Trend Filtering

To obtain posterior trends with approximate piecewise polynomial structure, we place a constrained "flat" prior on $\boldsymbol{\beta}$ to induce sparsity and regularity, namely

$$\pi(\boldsymbol{\beta} \mid \alpha) = \alpha^{-(n-k-1)} \exp\left\{-\iota_{\mathcal{E}}(\boldsymbol{\beta}, \alpha)\right\}, \quad (13)$$

where

$$\mathcal{E} = \left\{(\boldsymbol{\beta}, \alpha) \in \mathbb{R}^n \times \mathbb{R}_{++} : \|\mathbf{D}_n^{(\mathbf{x},k+1)}\boldsymbol{\beta}\|_1 \leq \alpha\right\}.$$

Note that implicitly α must be positive in (13) and all our subsequent equations. The term $\alpha^{-(n-k-1)}$ reflects the fact that we are constraining $\mathbf{D}_n^{(\mathbf{x},k+1)}\boldsymbol{\beta}$ to an $(n-k-1)$-dimensional ℓ_1-norm ball, which has volume proportional to α^{n-k-1}. To complete the model specification, we need to place additional priors on σ^2 and α. For σ^2, the standard inverse Gamma prior $\text{IG}(s, r)$ suffices as the parameters s and r minimally influence the

posterior for small values. In contrast, some care is warranted for choosing the prior for α. Ideally, we seek a prior that cancels the term $\alpha^{-(n-k-1)}$ to ensure a proper surrogate posterior density.

A natural strategy is to use a Gamma prior, which achieves the goal of canceling out $\alpha^{-(n-k-1)}$. Placing a $\Gamma(n-k,\mu)$ prior on α, the joint prior on $(\boldsymbol{\beta},\alpha)$ becomes

$$\pi(\boldsymbol{\beta},\alpha) = \exp\{-\iota_{\mathcal{E}}(\boldsymbol{\beta},\boldsymbol{\alpha}) - \mu\alpha\}. \tag{14}$$

Choosing a Gamma prior, however, requires us to choose large μ values to impose a meaningful amount of shrinkage, which makes $\Gamma(n-k,\mu)$ an informative prior since its variance is $(n-k)/\mu^2$. In that case selecting an appropriate μ becomes challenging and stymies our goal of operating within a fully Bayesian framework.

Given these challenges with a Gamma prior, we propose using a beta-prime prior. A beta-prime distribution, denoted as $\beta'(s_1,s_2)$, has density

$$\pi(\alpha) \propto \alpha^{s_1-1}(1+\alpha)^{-s_1-s_2}.$$

If we place a $\beta'(n-k,s_2)$ prior on α, the joint prior for $(\boldsymbol{\beta},\alpha)$ becomes

$$\pi(\boldsymbol{\beta},\alpha) \propto \exp\{-\iota_{\mathcal{E}}(\boldsymbol{\beta},\boldsymbol{\alpha}) - (n-k+s_2)\log(1+\alpha)\}. \tag{15}$$

A $\beta'(s_1,s_2)$ distribution has mean $\frac{s_1}{s_2-1}$ and variance $\frac{s_1(s_1+s_2-1)}{(s_2-2)(s_2-1)^2}$. Consequently when s_2 is relatively small, the prior has high variance and becomes uninformative. What makes this prior setup preferred over the one induced by the Gamma prior in (14) is that even when s_2 is small, we still have $-(n-k+s_2)\log(1+\alpha)$ as a strong penalty to impose a useful measure of shrinkage. Therefore, the beta-prime prior is better than the Gamma prior in terms of hyperparameter sensitivity. Nonetheless, we will revisit using the Gamma prior later as it is better suited for our second application PBSRTF. Why that is the case will be discussed in Section 4.3.

Placing an IG(s,r) prior on σ^2 and a $\beta'(n-k,s_2)$ prior on α, our full posterior density reads

$$\begin{aligned}&\pi(\boldsymbol{\beta},\sigma^2,\alpha \mid y)\\ &\quad\propto (\sigma^2)^{-\frac{m}{2}-s-1}\exp\Bigg\{-\frac{\sum_{i=1}^n\sum_{j=1}^{w_i}(y_{ij}-\beta_i)^2+2r}{2\sigma^2}\\ &\qquad-\iota_{\mathcal{E}}(\boldsymbol{\beta},\boldsymbol{\alpha})-(n-k+s_2)\log(1+\alpha)\Bigg\},\end{aligned} \tag{16}$$

where $m=\sum_{i=1}^n w_i$ is the total number of observations. We can rewrite (16) in a vectorized format

$$\begin{aligned}&\pi(\boldsymbol{\beta},\sigma^2,\alpha \mid y)\\ &\quad\propto (\sigma^2)^{-\frac{m}{2}-s-1}\exp\Bigg\{-\frac{(\bar{\mathbf{y}}-\boldsymbol{\beta})^{\mathsf{T}}W(\bar{\mathbf{y}}-\boldsymbol{\beta})+\mathrm{SSE}+2r}{2\sigma^2}\\ &\qquad-\iota_{\mathcal{E}}(\boldsymbol{\beta},\alpha)-(n-k+s_2)\log(1+\alpha)\Bigg\},\end{aligned} \tag{17}$$

where

$$\begin{aligned}\bar{\mathbf{y}} &= (\bar{y}_{1\cdot},\bar{y}_{2\cdot},\ldots,\bar{y}_{n\cdot})^{\mathsf{T}},\\ \mathbf{W} &= \mathrm{diag}(w_1,w_2,\ldots,w_n),\end{aligned}$$

$$\mathrm{SSE} = \sum_{i=1}^n\sum_{j=1}^{w_i}(y_{ij}-\bar{y}_{i\cdot})^2.$$

There is no simple algorithm for projection onto $\mathcal{E}$ when $k\geq 1$. To take advantage of the epigraph projection algorithms described in Section 4.1, we consider the reparameterization $\boldsymbol{\theta}=\mathbf{T}_1\boldsymbol{\beta}$ where

$$\mathbf{T}_1 = \begin{bmatrix}\mathbf{I}_{(k+1)\times n}\\ \mathbf{D}_n^{(\mathbf{x},k+1)}\end{bmatrix}, \tag{18}$$

and $\mathbf{I}_{(k+1)\times n}$ is the matrix obtained by taking the first $k+1$ rows of a n-by-n identity matrix. In other words, we have $\boldsymbol{\theta}_{[1:(k+1)]}=\boldsymbol{\beta}_{[1:(k+1)]}$ and $\boldsymbol{\theta}_{[(k+2):n]}=\mathbf{D}_n^{(\mathbf{x},k+1)}\boldsymbol{\beta}$. To better visualize the reparameterization technique, we explicitly write out the reparameterization scheme for $x_i=i,\ i=1,2,\ldots,n$ and $k=1$,

$$\begin{bmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\\ \vdots\\ \theta_n\end{bmatrix} = \begin{bmatrix}1 & 0 & 0 & 0\cdots 0 & 0 & 0\\ 0 & 1 & 0 & 0\cdots 0 & 0 & 0\\ 1 & -2 & 1 & 0\cdots 0 & 0 & 0\\ 0 & 1 & -2 & 1\cdots 0 & 0 & 0\\ \vdots & & & & & \\ 0 & 0 & 0 & 0\cdots 1 & -2 & 1\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\\ \vdots\\ \beta_n\end{bmatrix}.$$

Note that the transformation matrix $\mathbf{T}_1$ is a lower-triangular banded matrix with $k+2$ nonzero diagonals. This means that given $\boldsymbol{\theta}$, we can retrieve $\boldsymbol{\beta}$ in $O(n(k+2))$ operations using a banded forward-solve step. The reparameterized posterior is

$$\begin{aligned}&\pi(\boldsymbol{\theta},\sigma^2,\alpha \mid \mathbf{y})\\ &\propto (\sigma^2)^{-\frac{m}{2}-s-1}\exp\Bigg\{-\frac{(\bar{\mathbf{y}}-\mathbf{T}_1^{-1}\boldsymbol{\theta})^{\mathsf{T}}\mathbf{W}(\bar{\mathbf{y}}-\mathbf{T}_1^{-1}\boldsymbol{\theta})+\mathrm{SSE}+2r}{2\sigma^2}\\ &\quad-\iota_{\mathcal{E}_1'}(\boldsymbol{\theta},\alpha)-(n-k+s_2)\log(1+\alpha)\Bigg\},\end{aligned} \tag{19}$$

where

$$\mathcal{E}_1' = \{(\boldsymbol{\theta},\alpha)\in\mathbb{R}^n\times\mathbb{R}_{++} : \|\boldsymbol{\theta}_{[(k+2):n]}\|_1\leq\alpha\}.$$

Replacing $\iota_{\mathcal{E}_1'}(\boldsymbol{\theta},\alpha)$ with its Moreau-Yosida envelope, we arrive at a smooth surrogate posterior

$$\begin{aligned}&\pi^{\lambda}(\boldsymbol{\theta},\sigma^2,\alpha \mid \mathbf{y})\\ &\propto (\sigma^2)^{-\frac{m}{2}-s-1}\exp\Bigg\{-\frac{(\bar{\mathbf{y}}-\mathbf{T}_1^{-1}\boldsymbol{\theta})^{\mathsf{T}}\mathbf{W}(\bar{\mathbf{y}}-\mathbf{T}_1^{-1}\boldsymbol{\theta})+\mathrm{SSE}+2r}{2\sigma^2}\\ &\quad-\frac{1}{2\lambda}d^2_{\mathcal{E}_1'}(\boldsymbol{\theta},\alpha)-(n-k+s_2)\log(1+\alpha)\Bigg\},\end{aligned} \tag{20}$$

Projection onto $\mathcal{E}_1'$ can be accomplished by applying the ℓ_1-norm epigraph projection process described in Section 4.1 to $\boldsymbol{\theta}_{[(k+2):n]}$. Working with this reparameterization raises some potential computational challenges, however. When evaluating the function value and calculating the gradient of (20), we need to solve two linear systems, namely $\mathbf{T}_1^{-1}\boldsymbol{\theta}$ and $\mathbf{T}_1^{-T}\mathbf{W}(\bar{\mathbf{y}}-\mathbf{T}_1^{-1}\boldsymbol{\theta})$. As n and k increases, the condition number of $\mathbf{T}_1$ increases, leading to numerical instability in the HMC sampler. To alleviate this numerical issue, we can use the projection onto the epigraph of $\|\mathbf{D}_n^{(1)}\boldsymbol{\beta}\|_1$, described in Section 4.1. Borrowing the idea of

Ramdas and Tibshirani (2016), we consider another reparameterization scheme $\boldsymbol{\theta} = \mathbf{T}_2\boldsymbol{\beta}$ where

$$\mathbf{T}_2 = \begin{bmatrix} \mathbf{I}_{k\times n} \\ \operatorname{diag}\left(\frac{k}{x_{k+1}-x_1}, \cdots, \frac{k}{x_n - x_{n-k}}\right)\mathbf{D}_n^{(\mathbf{x},k-1)} \end{bmatrix}. \quad (21)$$

The reparameterized density is now

$$\pi(\boldsymbol{\theta}, \sigma^2, \alpha \mid \mathbf{y}) \propto (\sigma^2)^{-\frac{m}{2}-s-1} \exp\left\{ -\frac{(\bar{\mathbf{y}} - \mathbf{T}_2^{-1}\boldsymbol{\theta})^\top \mathbf{W}(\bar{\mathbf{y}} - \mathbf{T}_2^{-1}\boldsymbol{\theta}) + \text{SSE} + 2r}{2\sigma^2} - \iota_{\mathcal{E}_2'}(\boldsymbol{\theta}, \alpha) - (n-k+s_2)\log(1+\alpha) \right\}, \quad (22)$$

where

$$\mathcal{E}_2' = \{(\boldsymbol{\theta}, \alpha) \in \mathbb{R}^n \times \mathbb{R}_{++} : \|\mathbf{D}_{n-k}^{(1)}\boldsymbol{\theta}_{[(k+1):n]}\|_1 \le \alpha\}.$$

Similarly, projection onto $\mathcal{E}'$ can be achieved by applying the $\|\mathbf{D}_n^{(1)}\boldsymbol{\beta}\|_1$ epigraph projection process to $\boldsymbol{\theta}_{[(k+1):n]}$. The advantage of using $\mathbf{T}_2$ as the reparameterization scheme is that the "order" of $\mathbf{T}_2$ is one below that of $\mathbf{T}_1$, so that solving the linear systems becomes more numerically stable. When n and k are relatively small, however, using $\mathbf{T}_2$ requires solving (7), which is more expensive than (6). Table 1 summarizes the approximate cutoffs of when to use $\mathbf{T}_1$ and when to use $\mathbf{T}_2$, based on our empirical studies. Table 1 does not include $k = 0$ and $k = 3$, since Faulkner and Minin (2018) demonstrated that the shrinkage property of the Laplace prior struggles to capture abrupt jumps of piecewise constant underlying trends, resulting in a posterior fit that is too wiggly. Our prior set up is analogous to the Laplace prior, so that our method runs into the same issue. Meanwhile, when $k = 3$, even $\mathbf{T}_2$ is extremely ill-conditioned and the HMC sampler is hampered from exploring the parameter space meaningfully. Therefore, we focus on the case where $k = 1$ (piecewise linear) and $k = 2$ (piecewise quadratic).

Using $\mathbf{T}_2$ as the reparameterization matrix mitigates but does not eliminate the ill-conditioning issue. As n increases, it becomes more difficult for the HMC sampler to sufficiently explore the parameter space due to numerical instability. We will introduce a data preprocessing technique called thinning in Section 5.2 as an alternative strategy to make PBTF applicable for long sequences with large n.

Replacing $\iota_{\mathcal{E}_2'}(\boldsymbol{\theta}, \alpha)$ with its Moreau-Yosida envelope, the surrogate posterior is now

$$\pi^\lambda(\boldsymbol{\theta}, \sigma^2, \alpha \mid \mathbf{y}) \propto (\sigma^2)^{-\frac{m}{2}-s-1} \exp\left\{ -\frac{(\bar{\mathbf{y}} - \mathbf{T}_2^{-1}\boldsymbol{\theta})^\top \mathbf{W}(\bar{\mathbf{y}} - \mathbf{T}_2^{-1}\boldsymbol{\theta}) + \text{SSE} + 2r}{2\sigma^2} - \frac{1}{2\lambda} d^2_{\mathcal{E}_2'}(\boldsymbol{\theta}, \alpha) - (n-k+s_2)\log(1+\alpha) \right\}. \quad (23)$$

Notice that (20) and (23) are now differentiable functions of $(\boldsymbol{\theta}, \sigma^2, \alpha)$ on $\mathbb{R}^n \times \mathbb{R}_{++} \times \mathbb{R}_{++}$.

Table 1. Choice of reparameterization scheme for different n and k.

	$n \le 200$	$200 < n \le 1000$	$n > 1000$
$k = 1$	$\mathbf{T}_1$	$\mathbf{T}_2$	Thinning needed
$k = 2$	$\mathbf{T}_2$	Thinning needed	Thinning needed

For notational simplicity, in the rest of the manuscript we will use $\pi(\boldsymbol{\theta}, \sigma^2, \alpha \mid \mathbf{y})$ to refer to both (19) and (22), $\pi^\lambda(\boldsymbol{\theta}, \sigma^2, \alpha \mid \mathbf{y})$ to refer to and (20) and (23), and $(\mathbf{T}, \mathcal{E}')$ to refer to $(\mathbf{T}_1, \mathcal{E}_1')$ and $(\mathbf{T}_2, \mathcal{E}_2')$. We overload notation in this way since proofs and statements about these two surrogate densities are essentially the same.

4.3. Adding Shape-Restrictions

Proximal MCMC presents a simple alternative framework to traditional Bayesian hierarchical models that can easily construct priors that encode multiple structural constraints. Similar to nonparameteric isotonic trend filtering (Kim et al. 2009; Ramdas and Tibshirani 2016), adding shape restrictions into our framework is as straightforward as imposing linear inequalities. For instance, if we believe that the underlying trend is monontone increasing, we can enforce monotonicity by refining the epigraph set $\mathcal{E}$ with a monotonicity constraint as follows

$$\mathcal{S} = \{(\boldsymbol{\beta}, \alpha) \in \mathbb{R} \times \mathbb{R}_{++} : \|\mathbf{D}_n^{(\mathbf{x},k+1)}\boldsymbol{\beta}\|_1 \le \alpha, \mathbf{D}_n^{(1)}\boldsymbol{\beta} \ge \mathbf{0}\}.$$

In addition to monotonicity, convexity can be encoded by the linear inequalities in (4). By replacing $\ge$ with $\le$, we get monotone decreasing and concave restrictions. Combining monotonicity and convexity is as simple as imposing two sets of linear inequalities. Therefore, our framework can model eight types of shape restrictions, namely increasing, decreasing, convex, concave, increasing-convex, increasing-concave, decreasing-convex and decreasing-concave. Lower or upper bounds on $\boldsymbol{\beta}$ can also be enforced if warranted or desired.

Figure 3 illustrates examples of posterior fits using both versions of our fully Bayesian proximal MCMC method for trend filtering with and without shape-restrictions. For proof of concept, projection onto $\mathcal{S}$ can be achieved by any quadratic programming solver. We report the results using the Gurobi solver and leave for future work developing customized algorithms for potentially greater scalability.

As alluded to earlier, for PBSRTF we consider a joint prior on $(\boldsymbol{\beta}, \alpha)$ that employs a Gamma prior on α

$$\pi(\boldsymbol{\beta}, \alpha) \propto \exp\{-\iota_{\mathcal{S}}(\boldsymbol{\beta}, \alpha) - \mu\alpha\}. \quad (24)$$

The joint prior in (24) is almost identical to the one in (14); we simply replaced $\mathcal{E}$ with $\mathcal{S}$, where shape restrictions are also present. There are several reasons to revisit a Gamma prior on α. First, we can no longer interpret $\mathcal{S}$ as an ℓ_1-norm ball so that it is unclear what the normalizing constant should be; contrast this to the nonshape-restricted case where the normalizing constant is $\alpha^{-(n-k-1)}$. In fact, using $\alpha^{-(n-k-1)}$ as the normalizing constant for PBSRTF results in too much shrinkage. Second, there are numerical challenges that make the sampler using the beta-prime prior typically slower overall. We discuss these challenges in the supplementary materials. Finally, issues of the posterior being sensitive to the choice of μ, as we highlighted in Section 4.2, are no longer prohibitively acute as in the non shape-restricted case. In the case of PBSRTF, shape restrictions impose a helpful dose of regularization on $\boldsymbol{\beta}$, therefore, blunting the influence of our choice of μ.

Using an inverse Gamma IG(s, r) as the prior for σ^2 and (24) as the prior for $(\boldsymbol{\beta}, \alpha)$, our full posterior density for PBSRTF is

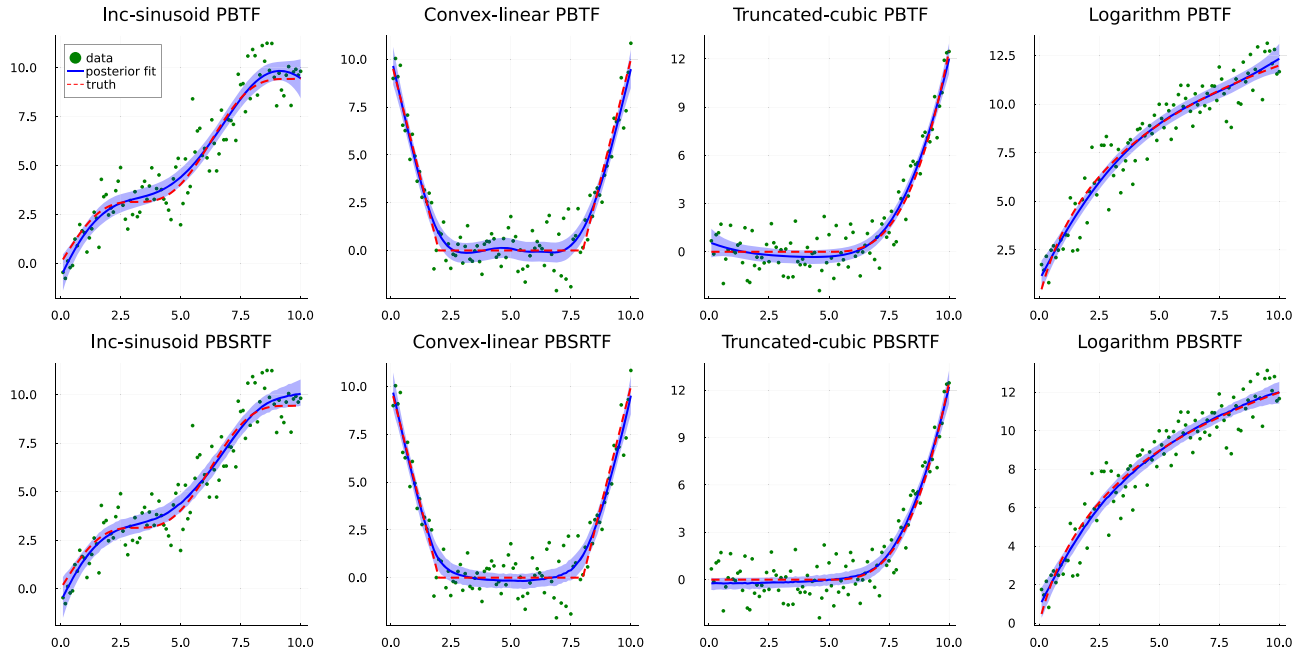

Figure 3. Example posterior fits for PBTF and PBSRTF with noise level $\sigma = 1$. The top row shows posterior fits of PBTF, and the bottom row shows posterior fits of PBSRTF. From left to right, the enforced shape restrictions are increasing, convex, increasing-convex, and increasing-concave.

$$\pi(\boldsymbol{\beta}, \sigma^2, \alpha \mid \mathbf{y}) \propto (\sigma^2)^{-\frac{m}{2}-s-1} \exp\left\{ -\frac{(\bar{\mathbf{y}}-\boldsymbol{\beta})^{\mathsf{T}}\mathbf{W}(\bar{\mathbf{y}}-\boldsymbol{\beta}) + \mathrm{SSE} + 2r}{2\sigma^2} - \iota_{\mathcal{S}}(\boldsymbol{\beta}, \alpha) - \mu\alpha \right\}. \tag{25}$$

Replacing $\iota_{\mathcal{S}}(\boldsymbol{\beta}, \alpha)$ with its Moreau-Yosida envelope, results in the surrogate posterior

$$\pi^{\lambda}(\boldsymbol{\beta}, \sigma^2, \alpha \mid \mathbf{y}) \propto (\sigma^2)^{-\frac{m}{2}-s-1} \exp\left\{ -\frac{(\bar{\mathbf{y}}-\boldsymbol{\beta})^{\mathsf{T}}\mathbf{W}(\bar{\mathbf{y}}-\boldsymbol{\beta}) + \mathrm{SSE} + 2r}{2\sigma^2} - \frac{1}{2\lambda} d^2_{\mathcal{S}}(\boldsymbol{\beta}, \alpha) - \mu\alpha \right\}. \tag{26}$$

Again, (26) is a differentiable function of $(\boldsymbol{\beta}, \sigma^2, \alpha)$ on $\mathbb{R}^n \times \mathbb{R}_{++} \times \mathbb{R}_{++}$. Neither (23) nor (26) is log-concave, however, so that Langevin algorithms are no longer suitable for MCMC sampling. Therefore, we turn to Hamiltonian Monte Carlo as our sampling engine.

4.4. Properties of the Surrogate Posteriors

We conclude this section, with two theorems that justify the practice of replacing the nonsmooth part of the posterior by its Moreau-Yosida envelope. The proofs are provided in the supplementary materials.

Theorem 4.1. The surrogate posterior densities (20), (23), and (26) are proper, that is,

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}_{++}} \int_{\mathbb{R}_{++}} \pi^{\lambda}(\boldsymbol{\theta}, \sigma^2, \alpha \mid \mathbf{y}) d\boldsymbol{\theta} d\sigma^2 d\alpha < +\infty,$$

and

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}_{++}} \int_{\mathbb{R}_{++}} \pi^{\lambda}(\boldsymbol{\beta}, \sigma^2, \alpha \mid \mathbf{y}) d\boldsymbol{\beta} d\sigma^2 d\alpha < +\infty.$$

Theorem 4.2. The surrogate posterior densities (20), (23), and (26) converges to the original nonsmooth densities (19), (22), and (25) in total-variation norm as $\lambda \downarrow 0$, that is,

$$\lim_{\lambda \downarrow 0} \int_{\mathbb{R}^n} \int_{\mathbb{R}_{++}} \int_{\mathbb{R}_{++}} \left| \pi^{\lambda}(\boldsymbol{\theta}, \sigma^2, \alpha \mid \mathbf{y}) - \pi(\boldsymbol{\theta}, \sigma^2, \alpha \mid \mathbf{y}) \right| d\boldsymbol{\theta} d\sigma^2 d\alpha = 0,$$

and

$$\lim_{\lambda \downarrow 0} \int_{\mathbb{R}^n} \int_{\mathbb{R}_{++}} \int_{\mathbb{R}_{++}} \left| \pi^{\lambda}(\boldsymbol{\beta}, \sigma^2, \alpha \mid \mathbf{y}) - \pi(\boldsymbol{\beta}, \sigma^2, \alpha \mid \mathbf{y}) \right| d\boldsymbol{\beta} d\sigma^2 d\alpha = 0.$$

Theorem 4.2 assures us that the surrogate density can approximate the original posterior density arbitrarily well by choosing a small enough λ. This is consistent with our experiments where we observe that the posterior fit is visually smooth once λ is sufficiently small. Note that λ should not be chosen to be too small, however, as doing so will lead to numerical instability since gradient evaluations involve division by λ. We discuss how to properly choose λ for the two different applications in in the supplementary materials. In practice, we recommend using the default parameters in our software.

5. Posterior Computation

5.1. HMC Sampling

We apply Hamiltonian Monte Carlo (HMC) to sample from the smoothed surrogate full posterior densities (20), (23), and

(26). Software for the proposed method is available at *https://github.com/qhengncsu/ProxBTF.jl*. We implement our method with DynamicHMC.jl package in the Julia computing environment. According to its documentation, the package implements a variant of the "No-U-Turn Sampler" (NUTS) of Hoffman and Gelman (2014), as described in Betancourt (2017). We direct readers to Betancourt (2017) for an accessible exposition on the algorithmic details of the sampling scheme. Since the NUTS algorithm operates on an unrestricted domain, we reparameterize σ^2 as $e^{\log\sigma^2}$ and α as $e^{\log\alpha}$ to model the two positive parameters.

For PBTF, evaluating the function-gradient pair at any given location requires $O(n)$ operations. While using Gurobi as a black box solver obscures the computational complexity of PBSRTF, we observe empirically that the computation time of PBSRTF also scales linearly with grid length n. This is likely due to the fact that Gurobi can effectively exploit the sparse matrices in our problem set up.

5.2. Thinning

As discussed in Section 4.2, PBTF may encounter numerical difficulties that accompany solving ill-conditioned linear systems. While the difference epigraph projection technique alleviates the ill-conditioning issue, it can not eliminate it; as n increases, eventually the condition number of $\mathbf{T}_2$ will eventually become problematic.

Another technique we propose to mitigate the ill-conditioning issue is thinning, which is similar to the thinning practice in R package glmgen Ramdas and Tibshirani (2016). We first split the range of $\mathbf{x}$ into intervals of equal length. Grid locations within the same interval are merged into a single new grid location, which is a weighted average of the original grid locations with weights being the numbers of observations. The data points (x_i, y_{ij}) are then horizontally shifted to the merged grid locations. After HMC sampling, if we are interested in the posterior median and confidence limits at the original grid locations, we can recover them through interpolation. We provide an illustration of thinning in the supplementary materials.

6. Numerical Experiments

We compare the empirical performance of PBTF with Shrinkage Prior Markov Random Fields (SPMRFs) by Faulkner and Minin (2018) and Dynamic Shrinkage Processes (DSP) by Kowal, Matteson, and Ruppert (2019). We note that DSP can be considered as an extention of SPMRFs and the software of DSP[1] in fact contains an implementation of the hierarchical models described in Faulkner and Minin (2018). Moreover, DSP uses customized Gibbs samplers which in practice are more efficient than the HMC sampler used by SPMRFs, thus, we primarily use the software of DSP in our experiments. In Table 2, BTF-BL (Bayesian Lasso prior or Laplace prior) and BTF-HS (horseshoe prior) correspond to the models presented in Faulkner and Minin (2018) while BTF-DHS (dynamic horseshoe prior) corresponds to the model presented in Kowal, Matteson, and Ruppert (2019). To investigate the relative strengths of different approaches, we selected four underlying trends, namely piecewise linear, smooth trend, sinusoid, and piecewise quadratic/cubic. We assess the precision of each method with mean absolute deviation (MAD), frequentist coverage probability (CP), and mean credible interval width (MCIW). We also include the total CPU time (TCPU), effective sample size of the slowest component (min. ESS) and multivariate effective sample size (MESS) (Vats, Flegal, and Jones 2019) as measures of sampling efficiency. The detailed definitions of the summary statistics are given in the supplementary materials.

Following Faulkner and Minin (2018) and Kowal, Matteson, and Ruppert (2019), we used evenly spaced grid locations of $\{1, 2, \ldots, 100\}$ and designed the underlying trends to have an approximate standard deviation of 9. We added two levels of Gaussian noise ($\sigma = 3.0$ and $\sigma = 4.5$) to the underlying trends, generating 50 noisy sequences for each combination of trend and noise level. For DSP, we used the default parameters, ran an initial burn-in of 1000 iterations followed by 2500 posterior draws. For PBTF, we set s_2 to be $\sqrt{n} = 10$, ran the default warm-up stage in DynamicHMC.jl and made another 2500 posterior draws. Table 2 shows the summary statistics

[1] *https://github.com/drkowal/dsp*

Table 2. Summary statistics for DSP and PBTF, averaged over 50 generated sequences at noise level $\sigma = 3$.

True trend	Method	MAD (s.d.)	MCIW	CP	TCPU(s)	min. ESS	MESS
Piece. Linear	BTF-BL	0.87 (0.18)	4.3	0.95	12	2271	4018
	BTF-HS	0.73 (0.19)	3.7	0.95	7	1368	3275
	BTF-DHS	0.70 (0.18)	3.7	0.95	17	880	3120
	PBTF ($k = 1$)	0.82 (0.17)	3.9	0.94	70	1902	2037
Smooth trend	BTF-BL	0.98 (0.16)	5.1	0.96	12	1674	2440
	BTF-HS	1.00 (0.15)	5.1	0.95	7	973	2491
	BTF-DHS	1.02 (0.15)	5.1	0.95	17	150	1893
	PBTF ($k = 2$)	0.87 (0.16)	4.3	0.95	896	857	2684
Sinusoid	BTF-BL	0.80 (0.14)	4.6	0.97	12	2080	4120
	BTF-HS	0.83 (0.14)	4.7	0.97	7	1203	2340
	BTF-DHS	0.86 (0.14)	4.8	0.97	17	260	1884
	PBTF ($k = 2$)	0.70 (0.14)	3.9	0.97	927	1207	3686
Piece. Quad./ Cubic	BTF-BL	0.77 (0.12)	4.3	0.97	12	845	4223
	BTF-HS	0.78 (0.15)	4.1	0.96	7	378	2585
	BTF-DHS	0.82 (0.15)	4.2	0.95	17	180	2190
	PBTF ($k = 2$)	0.70 (0.13)	3.8	0.96	931	1439	3326

for different methods averaged over 50 generated sequences with $\sigma = 3.0$. The results for noise level $\sigma = 4.5$ can be found in the supplementary materials, which exhibits a similar pattern.

The last three trends, namely smooth trend, sinuoid and piecewise quadratic/cubic are better approximated by piecewise quadratic functions. However, in Table 2 we only report the results of DSP using $k = 1$. This is partly because the software of DSP does not contain an option to fit models with $k = 2$. That being said, the software of SPMRFs[2] does offer an option to fit models with $k = 2$. Nevertheless, for the last three trends, when going from $k = 1$ to $k = 2$, SPMRFs overall suffers a decrease in MAD and CP in contrary to one's expectation. These additional simulation results can be found in the supplementary materials. SPMRFs' worse performance with $k = 2$, despite the underlying trends being better approximated by piecewise quadratic functions, may be attributed to the fact that it is inherently harder to sample from higher-order trend filtering models. In our framework, third-order PBTF alleviates part of that difficulty through leveraging the fused lasso subroutine, providing the best MAD and the narrowest confidence bands for the last three trends while maintaining ideal coverage probability.

BTF-HS achieves higher precision than BTF-BL and PBTF for piecewise linear trend, demonstrating stronger adaptivity to abrupt turns. This is attributed to the superior shrinkage properties of global-local priors like the horseshoe prior. Unfortunately, nonparametric analogues of the horseshoe prior are nonconvex, for example, Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li 2001) and Minimax concave Penalty (MCP) (Zhang 2010). Projection onto the epigraph of a nonconvex function is generally nontrivial. Therefore, it is not immediately obvious how to replicate the horseshoe prior's shrinkage property in our framework and presents an interesting avenue for future work. BTF-DHS achieved even better precision than BTF-HS for piecewise linear trend through modeling dependence between the local scale parameters. However, we also see that it will behave slightly worse than BTF-HS when modeling smooth underlying trends.

We note that DSP only applies to data on an evenly spaced grid. The framework of SPMRFs is extended to unevenly spaced grids for $k = 0$ and $k = 1$ in Faulkner and Minin (2018) using methods based on integrated Wiener processes. However, Faulkner and Minin (2018) did not further pursue the same for $k = 2$ due to its complexity. PBTF, on the other hand, naturally handles unevenly spaced grids for $k = 1, 2$ due to using the adjusted difference matrix $\mathbf{D}_n^{(\mathbf{x},k+1)}$ in its prior. In this section, we employed an evenly spaced grid $\{1, 2, \ldots, n\}$ in pursuit of simplicity and conformity. The real data analysis in Section 7 and the thinning example in the supplementary materials are both examples of third-order PBTF being applied to unevenly spaced grids.

7. Real Data Example

We apply PBTF and PBSRTF to the Munich dataset as a real data example. We focus on two variables in the dataset, with the response being rent per square meter in Munich, Germany, and the covariate being floor space in square meters. The dataset was first analyzed by Rue and Held (2005) using Gaussian Markov Random Fields. Faulkner and Minin (2018) analyzed this data as an illustration of SPMRFs being applied on an unevenly spaced grid. The dataset has 2035 observations in total and the covariate floor space has 134 distinct values. Other than second-order and third-order PBTF models, we also present second-order PBSRTF model fits with "decreasing" and "decreasing-convex" as shape restrictions. In the former case, we model the assumption that rent per square meter decreases as floor space increases. In the latter case, we model an additional diminishing returns effect.

We used $s_2 = 2 \times \sqrt{134}$ for PBTF and set $\mu = 4.0$ for PBSRTF to promote a bit more regularity. Figure 4 shows the posterior fits of the four different models. All four models captured an overall decreasing trend. It is notable that the confidence bands are narrower over intermediate values of floor space, which is expected as there are more data points over this range of floor spaces. Third-order PBTF produced a more variable posterior median and a wider confidence band than second-order PBTF, suggesting that third-order PBTF models exhibit more adaptivity but may be prone to overfitting. We notice that posterior fits with shape restrictions have much narrower confidence bands compared with their unconstrained counterparts. This is because the shape restrictions introduce additional regularization that further reduces variance.

8. Discussion

In this article, we introduced a new proximal MCMC methodology, which incorporates the variance parameter σ^2 and the regularization parameter α into posterior inference. The key to extending the conventional proximal MCMC paradigm to a fully Bayesian one is to use epigraph priors to induce sparsity and regularity. By substituting the nonsmooth components of the posterior with its Moreau-Yosida envelope, we can work with a differentiable surrogate density, on which HMC is be applied for efficient MCMC sampling.

As a proof of concept, we explored the application of the proposed methodology in Bayesian trend filtering. Compared with existing Bayesian trend filtering methods, our approach achieves higher precision for underlying trends that are better approximated by piecewise quadratic functions. To demonstrate the flexibility of our framework, we also explored incorporating shape restrictions like monotonicity and convexity.

Although we focused on Bayesian trend filtering in this work, the strategy of combining an epigraph prior with proximal MCMC readily applies to other types of nonsmooth estimation problems. For example, modern optimization extensively uses nuclear norms to induce low-rank structure, therefore, a Bayesian version of low-rank matrix completion based on projection onto the epigraph of nuclear norm is an interesting future venue. It is also of great appeal to venture beyond convex penalties and constraints for greater modeling power in structured regression problems.

[2] *https://github.com/jrfaulkner/spmrf*

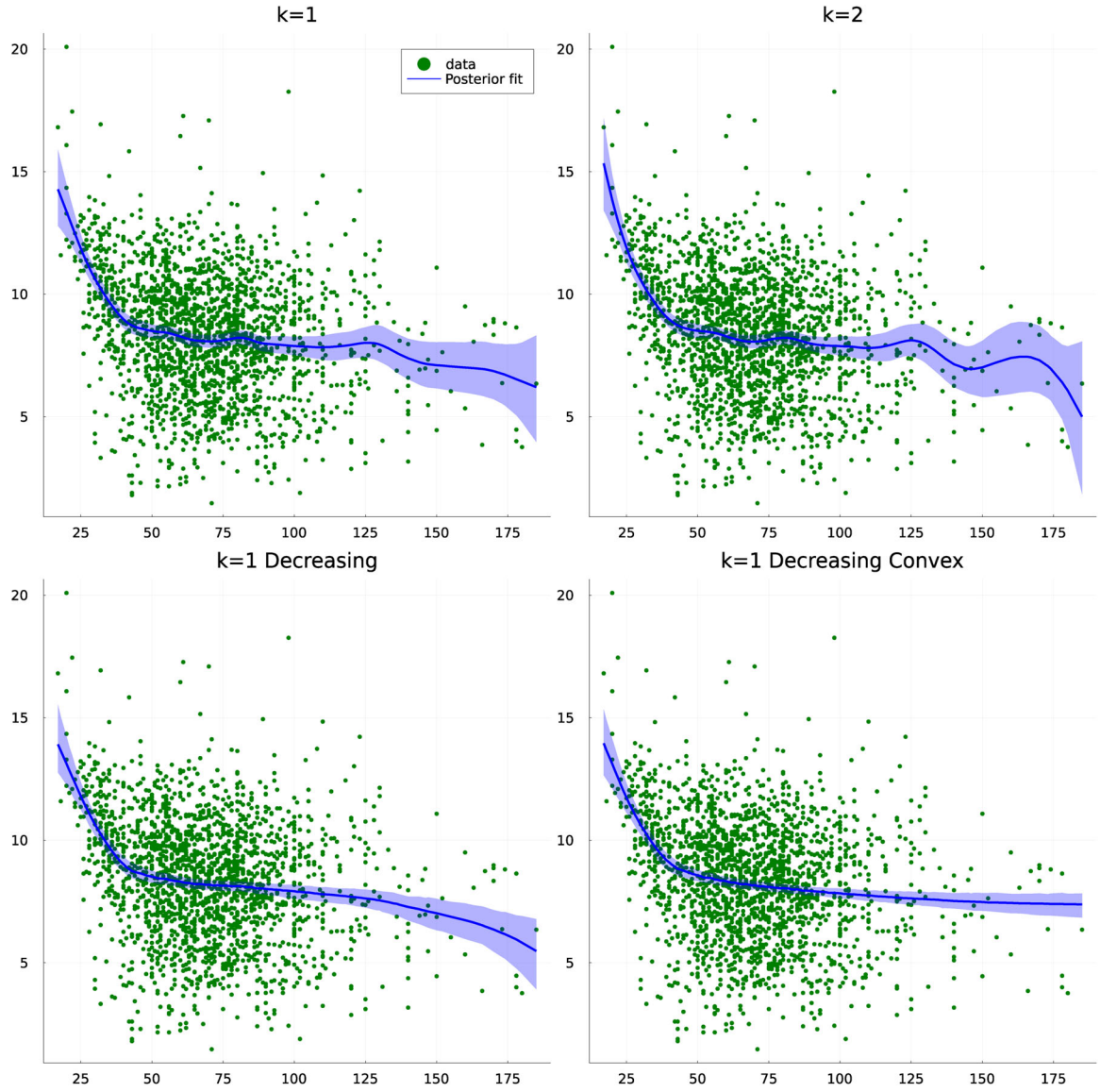

Figure 4. Posterior fits on Munich dataset. Plots show data points (green dots), posterior median (blue solid lines), and 95% Bayesian credible intervals (light blue bands).

Supplementary Materials

Title: Supplement to "Bayesian Trend Filtering via Proximal Markov Chain Monte Carlo". (.pdf file)

Software: Julia-package "ProxBTF.jl" containing code to perform the methods described in the article and scripts (R and Julia) to reproduce the numerical experiments. (.zipped file)

Acknowledgments

We thank the editor, the associate editor and the anonymous referee for inspiring comments and helpful suggestions, which greatly improved the presentation of this article. We thank James Faulkner for helpful discussions. We thank Marcelo for stimulating discussion that sparked interest in pursuing this work. We thank Patrick Combettes and Christian Müller for organizing a pair of workshops that introduced us to Marcelo's work as well as Ilse Ipsen for her oversight in planning the first workshop as the SAMSI Directorate Liaison.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This research was partially funded by grants from the National Institute of General Medical Sciences (R35GM141798: HZ, R01GM135928: EC), the National Human Genome Research Institute (R01HG006139: HZ), and the National Science Foundation (DMS-2054253 and IIS-2205441: HZ, DMS-2201136: EC). Much of the motivation for this work originated from talks on proximal MCMC methods by Marcelo Pereyra at a pair of workshops on operator splitting methods in data science which were funded by the National Science Foundation under Grant (DMS-1127914, the Statistical and Applied Mathematical Sciences Institute) and the Simons Foundation.

ORCID

Qiang Heng https://orcid.org/0000-0002-4042-6773
Hua Zhou https://orcid.org/0000-0003-1320-7118
Eric C. Chi https://orcid.org/0000-0003-4647-0895

References

Barlow, R. E. (1972), "Statistical Inference under Order Restrictions; the Theory and Application of Isotonic Regression," Technical Report. [1]

Beck, A. (2017), *First-Order Methods in Optimization*, Philadelphia: Society for Industrial and Applied Mathematics. [5]

Betancourt, M. (2017), "A Conceptual Introduction to Hamiltonian Monte Carlo," arXiv preprint arXiv:1701.02434. [9]

Brezger, A., and Steiner, W. J. (2008), "Monotonic Regression based on Bayesian p–splines: An Application to Estimating Price Response Functions from Store-Level Scanner Data," *Journal of Business & Economic Statistics*, 26, 90–104. [1]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [1]

Chambolle, A. (2004), "An Algorithm for Total Variation Minimization and Applications," *Journal of Mathematical Imaging and Vision*, 20, 89–97. [4]

Chambolle, A., and Pock, T. (2011), "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging," *Journal of Mathematical Imaging and Vision*, 40, 120–145. [3]

Combettes, P. L., and Pesquet, J.-C. (2011), "Proximal Splitting Methods in Signal Processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, eds. H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. Russell Luke, and H. Wolkowicz, pp. 185–212, New York: Springer. [3]

De Bortoli, V., Durmus, A., Pereyra, M., and Vidal, A. F. (2020), "Maximum likelihood estimation of regularization parameters in high-dimensional inverse problems: an Empirical Bayesian approach. Part II: Theoretical analysis," *SIAM Journal on Imaging Sciences*, 13, 1990–2028. [4,5]

Durmus, A., Moulines, E., and Pereyra, M. (2018), "Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau," *SIAM Journal on Imaging Sciences*, 11, 473–506. [2,3,4]

Durmus, A., Moulines, É., and Pereyra, M. (2022), "A Proximal Markov Chain Monte Carlo Method for Bayesian Inference in Imaging Inverse Problems: When Langevin Meets Moreau," *SIAM Review*, 64, 991–1028. [4]

Efron, B., and Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, Boca Raton, FL: CRC Press. [1]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [10]

Faulkner, J. R., and Minin, V. N. (2018), "Locally Adaptive Smoothing with Markov Random Fields and Shrinkage Priors," *Bayesian Analysis*, 13, 225–252. [1,7,9,10]

Goldstein, T., and Osher, S. (2009), "The Split Bregman Method for L1-Regularized Problems," *SIAM Journal on Imaging Sciences*, 2, 323–343. [3]

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (2008), "The Support Reduction Algorithm for Computing Non-parametric Function Estimates in Mixture Models," *Scandinavian Journal of Statistics*, 35, 385–399. [1]

Hoffman, M. D., and Gelman, A. (2014), "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, 15, 1593–1623. [9]

Johnson, N. A. (2013), "A Dynamic Programming Algorithm for the Fused Lasso and ℓ_0-segmentation," *Journal of Computational and Graphical Statistics*, 22, 246–260. [2,3]

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009), "ℓ_1 Trend Filtering," *SIAM Review*, 51, 339–360. [1,2,7]

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2019), "Dynamic Shrinkage Processes," *Journal of the Royal Statistical Society*, Series B, 81, 781–804. [1,9]

Lenk, P. J., and Choi, T. (2017), "Bayesian Analysis of Shape-Restricted Functions using Gaussian Process Priors," *Statistica Sinica*, 43–69. [2]

McKay Curtis, S., and Ghosh, S. K. (2011), "A Variable Selection Approach to Monotonic Regression with Bernstein Polynomials," *Journal of Applied Statistics*, 38, 961–976. [1]

Meyer, M. C., Hackstadt, A. J., and Hoeting, J. A. (2011), "Bayesian Estimation and Inference for Generalised Partial Linear Models using Shape-Restricted Splines," *Journal of Nonparametric Statistics*, 23, 867–884. [1]

Neal, R. M. (2011), "MCMC Using Hamiltonian Dynamics," in *Handbook of Markov Chain Monte Carlo*, eds. S. Brooks, A. Gelman, and G. Jones, X.-L. Meng, pp. 113–157, New York: Chapman and Hall/CRC. [3]

Neelon, B., and Dunson, D. B. (2004), "Bayesian Isotonic Regression and Trend Analysis," *Biometrics*, 60, 398–406. [1]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [1]

Pereyra, M. (2016), "Proximal Markov Chain Monte Carlo Algorithms," *Statistics and Computing*, 26, 745–760. [2,3]

Pereyra, M., Mieles, L. V., and Zygalakis, K. C. (2020), "Accelerating Proximal Markov Chain Monte Carlo by Using an Explicit Stabilized Method," *SIAM Journal on Imaging Sciences*, 13, 905–935. [2,4]

Ramdas, A., and Tibshirani, R. J. (2016), "Fast and Flexible ADMM Algorithms for Trend Filtering," *Journal of Computational and Graphical Statistics*, 25, 839–858. [2,5,7,9]

Roberts, G. O., and Tweedie, R. L. (1996), "Exponential Convergence of Langevin Distributions and their Discrete Approximations," *Bernoulli*, 2, 341–363. [3]

Rockafellar, R. T., and Wets, R. J.-B. (2009), *Variational Analysis* (Vol. 317), Berlin: Springer. [3]

Rossky, P. J., Doll, J., and Friedman, H. (1978), "Brownian Dynamics as Smart Monte Carlo Simulation," *The Journal of Chemical Physics*, 69, 4628–4633. [3]

Roualdes, E. A. (2015), "Bayesian Trend Filtering," arXiv preprint arXiv:1505.07710. [1]

Rudin, L., Osher, S., and Fatemi, E. (1992), "Non-linear Total Variation Noise Removal Algorithm," *Physica D: Nonlinear Phenomena*, 60, 259–268. [2]

Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton, FL: CRC press. [10]

Steidl, G., Didas, S., and Neumann, J. (2006), "Splines in Higher Order TV Regularization," *International Journal of Computer Vision*, 70, 241–255. [1,2]

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, 67, 91–108. [2,3]

Tibshirani, R. J., Hoefling, H., and Tibshirani, R. (2011), "Nearly-Isotonic Regression," *Technometrics*, 53, 54–61. [1]

Tibshirani, R. J. (2014), "Adaptive Piecewise Polynomial Estimation via Trend Filtering," *The Annals of Statistics*, 42, 285–323. [1,2]

Vats, D., Flegal, J. M., and Jones, G. L. (2019), "Multivariate Output Analysis for Markov Chain Monte Carlo," *Biometrika*, 106, 321–337. [9]

Vidal, A. F., De Bortoli, V., Pereyra, M., and Durmus, A. (2020), "Maximum Likelihood Estimation of Regularization Parameters in High-Dimensional Inverse Problems: An Empirical Bayesian Approach. Part I: Methodology and experiments," *SIAM Journal on Imaging Sciences*, 13, 1945–1989. [4,5]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [10]