Bayesian trend filtering: adaptive temporal smoothing with shrinkage priors

James R. Faulkner

Quantitative Ecology and Resource Management

University of Washington, Seattle, Washington, U.S.A.

Vladimir N. Minin

Departments of Statistics and Biology

University of Washington, Seattle, Washington, U.S.A.

Abstract

We present a locally-adaptive nonparametric curve fitting method that we call Bayesian trend filtering. The method operates within a fully Bayesian framework and uses shrinkage priors to induce sparsity in order-k differences in the latent trend function, providing a combination of local adaptation and global control. Using a scale mixture of normals representation of shrinkage priors, we make explicit connections between our Bayesian trend filtering and kth order Gaussian Markov random field smoothing. We use Hamiltonian Monte Carlo to approximate the posterior distribution of model parameters because this method provides superior performance in the presence of the high dimensionality and strong parameter correlations exhibited by our models. We compare the performance of three prior formulations using simulated data and find the horseshoe prior provides the best compromise between bias and precision. We apply Bayesian trend filtering to two benchmark data examples frequently used to test nonparametric methods. We find that this method is flexible enough to accommodate a variety of data generating models and offers the adaptive properties and computational efficiency to make it a useful addition to the Bayesian nonparametric toolbox.

1 Introduction

Nonparametric curve fitting methods find extensive use in many aspects of statistical modeling such as nonparametric regression, spatial statistics, and survival models, to name a few. Although these methods form a mature area of statistics many computational and statistical challenges remain when such curve fitting needs to be incorporated into multi-level Bayesian models with complex data generating processes. This work is motivated by the need for a curve fitting method that could adapt to local changes in smoothness of a function, including abrupt changes or jumps, and would not be restricted by the nature of observations and/or their associated likelihood. Our desired method should offer measures of uncertainty for use in inference, should be relatively simple to implement and computationally efficient. There are many methods available for nonparametric curve fitting, but few which meet all of these criteria.

Gaussian process (GP) regression (Neal, 1998; Rasmussen and Williams, 2006) is a popular Bayesian nonparametric approach for functional estimation that places a GP prior on the function of interest. The covariance function must be specified for the GP prior, and the isotropic covariance functions typically used are not locally adaptive. Nonstationary covariance functions have been investigated to make GP regression locally adaptive (Brahim-Belhouari and

Bermak, 2004; Paciorek and Schervish, 2004, 2006). Any finite dimensional representation of GPs involves manipulations of, typically high dimensional, Gaussian vectors with mean vector and covariance matrix induced by the GP. Many GPs, including the ones with nonstationary covariance functions, suffer from high computational cost imposed by manipulations (e.g., Cholesky factorization) of the dense covariance matrix in the finite dimensional representation.

Sparsity can be imposed in the precision matrix (inverse covariance matrix) by constraining a finite dimensional representation of a GP to be a Gaussian Markov random field (GMRF), and then computational methods for sparse matrices can be employed to speed computations (Rue, 2001; Rue and Held, 2005). Fitting smooth functions with GMRFs has been practiced widely. These methods use difference equations as approximations to continuous function derivatives to induce smoothing, and have a direct relationship to smoothing splines (Speckman and Sun, 2003). GMRFs have also been used to develop Bayesian adaptive smoothing splines (Lang et al., 2002; Yue et al., 2012, 2014). A similar approach is the nested GP (Zhu and Dunson, 2013), which puts a GP prior on the order-k function derivative, which is in turn centered on another GP. This approach has good adaptive properties but has not been developed for non-Gaussian data.

Differencing has commonly been used as an approach to smoothing and trend estimation in time series analysis, signal processing, and spatial statistics. Its origins go back at least to Whittaker (1922), who suggested a need for a trade off between fidelity to the data and smoothness of the estimated function. This idea is the basis of some frequentist curve-fitting methods based on penalized least squares, such as the smoothing spline (Reinsch, 1967; Wahba, 1975) and the trend filter (Kim et al., 2009; Tibshirani, 2014). These penalized least-squares methods are closely related to regularization methods for high-dimensional regression such as ridge regression (Hoerl and Kennard, 1970) and the lasso (Tibshirani, 1996) due to the form of the penalties imposed.

Bayesian versions of methods like the lasso (Park and Casella, 2008) utilize shrinkage priors in place of penalties. Therefore, it is interesting to investigate how these shrinkage priors (Polson and Scott, 2010; Griffin et al., 2013; Bhattacharya et al., 2014) perform when applied to differencing-based time series smoothing. Although shrinkage priors have been used explicitly in the Bayesian nonparametric regression setting for regularization of wavelet coefficients (Abramovich et al., 1998; Johnstone and Silverman, 2005; Reményi and Vidakovic, 2015) and for shrinkage of order-k differences of basis spline coefficients in adaptive Bayesian P-splines (Scheipl and Kneib, 2009), Bayesian trend filtering with shrinkage priors was not thoroughly investigated. To our knowledge, only Roualdes (2015), independently from our work, looked at Laplace prior-based Bayesian trend filtering in the context of a normal response model. In this paper, we conduct a thorough investigation of Bayesian trend filtering with multiple shrinkage priors when such filtering is applied to Gaussian and non-Gaussian data.

We borrow the idea of shrinkage priors from the sparse regression setting and apply it to the problem of function estimation. We take the perspective that nonparametric curve fitting is essentially a regularization problem where estimation of an unknown function can be achieved by inducing sparsity in its order-k derivatives. We propose a few fully Bayesian variations of the trend filter (Kim et al., 2009; Tibshirani, 2014) which utilize shrinkage priors on the kth-order differences in values of the unknown target function. The shrinkage imposed by the priors induces an adaptive smoothing of the trend. The fully Bayesian implementation allows representation of parameter uncertainty through posterior distributions and eliminates the need to specify a single global smoothing parameter. In Section 2 we provide a derivation of the models starting from penalized frequentist methods and we show the relationship to GMRF models. In Section 2 we also describe our method of approximating the posterior distribution of

the Bayesian trend filter parameters using Hamiltonian Monte Carlo (HMC), which is efficient and straight forward to implement. In Section 3 we use simulations to investigate performance properties of the Bayesian trend filter under two different prior formulations and we compare results to those for a GMRF with constant precision. We show that the choice of shrinkage prior will affect the smoothness and local adaptive properties. In Section 4 we apply the method to two example data sets which are well known in the nonparametric regression setting.

2 Methods

2.1 Preliminaries

We start by reviewing a locally adaptive penalized least squares approach to nonparametric regression known as the trend filter (Kim et al., 2009; Tibshirani and Taylor, 2011; Tibshirani, 2014) and use that as a basis to motivate a general Bayesian approach that utilizes shrinkage priors in place of roughness penalties. We first consider the standard nonparametric regression problem to estimate the unknown function f. We let $\boldsymbol{\theta}$ represent a vector of values of f on a discrete uniform grid $t \in \{1, 2, \ldots, n\}$, and we assume $\boldsymbol{y} = \boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2)$, and $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ are vectors of length n. Here all vectors are column vectors. Following Tibshirani (2014) with slight modification, the least squares estimator of the kth order trend filtering estimate $\hat{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \|\mathbf{y} - \boldsymbol{\theta}\|_2^2 + \lambda \|\mathbf{D}^{(k)}\boldsymbol{\theta}\|_1 \,, \tag{1}$$

where $\|\cdot\|_q$ represents the L_q vector norm, and $\mathbf{D}^{(k)}$ is an $(n-k) \times n$ forward difference operator matrix of order k, such that the ith element of the vector $\Delta^k\boldsymbol{\theta} = \mathbf{D}^{(k)}\boldsymbol{\theta}$ is the forward difference $\Delta^k\theta_i = (-1)^k \sum_{j=0}^{k} (-1)^j \binom{k}{j} \theta_{i+j}$. Note that $\mathbf{D}^{(k)}$ has recursive properties such that $\mathbf{D}_n^{(k)} = \mathbf{D}_{n-k+1}^{(1)} \mathbf{D}_n^{(k-1)}$, where $\mathbf{D}_m^{(h)}$ has dimensions $(m-h) \times m$. The objective function in equation (1) balances the trade-off between minimizing the squared deviations from the data (the first term in the sum on the right) with minimizing the discretized roughness penalty of the function f (the second term in the sum on the right). The smoothing parameter $\lambda \geq 0$ controls the relative influence of the roughness penalty. Setting λ to 0 we get least squares estimation. As λ gets large, the roughness penalty dominates, resulting in a function with k-th order differences approaching 0 for all t. The trend filter produces a piecewise polynomial function of $t_1, \ldots, t_n$ with degree $k-1$ as an estimator of the unknown function f. Increasing the order of the difference operator will enforce a smoother function.

The L_1 penalty in equation (1) results in the trend filter having locally adaptive smoothing properties. Tibshirani (2014) shows that the trend filter is very similar in form and performance to smoothing splines and locally adaptive regression splines, but the trend filter has a finer level of local adaptivity than smoothing splines. A main difference between the trend filter and smoothing splines is that the latter uses a squared L_2 penalty, which is the same penalty used in ridge regression (Hoerl and Kennard, 1970). Note that the L_1 penalty used by the trend filter is also used by the lasso regression (Tibshirani, 1996), and the trend filter is a form of generalized lasso (Tibshirani and Taylor, 2011; Tibshirani, 2014). In the linear regression setting with regression coefficients β_js, the L_1 and L_2 penalties can be represented by the generalized ridge penalty $\lambda \sum_j |\beta_j|^q$ (Frank and Friedman, 1993), where $q = 2$ corresponds to the ridge regression penalty, $q = 1$ to the lasso penalty, and sending q to zero results in all subsets selection regression (Tibshirani, 2011). Based on what we know about lasso regression, subset selection regression, and ridge regression, we expect a penalty closer to subset selection to do better for

fitting functions with a small number of large jumps, a trend filter penalty (L_1) to do better for fitting functions with small to moderate deviations from polynomials of degree $k-1$, and a smoothing spline (squared L_2) penalty to do better for smooth polynomial-like functions with no jumps. This distinction will become important later when we assess the performance of different Bayesian formulations of the trend filter.

One can translate the penalized least squares formulation in equation (1) into either a penalized likelihood formulation or a Bayesian formulation. Penalized least squares can be interpreted as minimizing the penalized negative log-likelihood $-l_p(\boldsymbol{\theta} \mid \boldsymbol{y}) = -l(\boldsymbol{\theta} \mid \boldsymbol{y}) + p(\boldsymbol{\theta} \mid \lambda)$, where $l(\boldsymbol{\theta} \mid \boldsymbol{y})$ is the unpenalized log-likelihood and $p(\boldsymbol{\theta} \mid \lambda)$ is the penalty. It follows that maximization of the penalized log-likelihood is directly comparable to finding the mode of the log-posterior in the Bayesian formulation, where the penalty is represented as a prior. This implies independent Laplace (double-exponential) priors on the $\Delta^k\theta_j$, where $j = 1, \ldots, n-k$, for the trend filter formulation in equation (1). That is, $p(\Delta^k\theta_j \mid \lambda) = \frac{\lambda}{2}\exp\left(-\lambda\,|\Delta^k\theta_j\,|\right)$. This is a well-known result that has been used in deriving a Bayesian form of the lasso (Tibshirani, 1996; Figueiredo, 2003; Park and Casella, 2008). Note that putting independent priors on the kth order differences results in improper joint prior $p(\boldsymbol{\theta} \mid \lambda)$, which can be made proper by including a proper prior on the first k θ's.

The Laplace prior falls into a class of priors commonly known as shrinkage priors. An effective shrinkage prior has the ability to shrink noise to zero yet retain and accurately estimate signals (Polson and Scott, 2010). These properties translate into a prior density function that has a combination of high mass near zero and heavy tails. The high density near zero acts to shrink small values close to zero, while the heavy tails allow large signals to be maintained. A simple prior developed for subset selection in Bayesian setting is the spike-and-slab prior, which is a mixture distribution between a point mass at zero and a continuous distribution (Mitchell and Beauchamp, 1988). This prior works well for model selection, but some drawbacks are that it forces small signals to be exactly zero, and computational issues can make it difficult to use (Polson and Scott, 2010). There has been much interest in developing priors with continuous distributions (one group) that retain variable selection properties of the spike-and-slab (two-group) yet do so by introducing sparsity through shrinkage (Polson and Scott, 2010). This approach allows all of the coefficients to be nonzero, but most are small and only some are large. Many such shrinkage priors have been proposed, including the normal-gamma (Griffin et al., 2010), generalized double-Pareto (Armagan et al., 2013), and the horseshoe (Carvalho et al., 2010). The Laplace prior lies somewhere between the normal prior and the spike-and-slab in its shrinkage abilities, yet most shrinkage priors of current research interest have sparsity inducing properties closer to those of the spike-and-slab. Our main interest is in comparing the Laplace prior to other shrinkage priors in the context of Bayesian trend filtering.

2.2 Model Formulation

It is clear that shrinkage priors other than the lasso could represent different smoothing penalties and therefore could lead to more desirable smoothing properties. There is a large and growing number of shrinkage priors in the literature. It is not our goal to compare and characterize properties of Bayesian nonparametric function estimation under all of these priors. Instead, we wish to investigate a few well known shrinkage priors and demonstrate as proof of concept that adaptive functional estimation can be achieved with shrinkage priors. Further research can focus on improvements to these methods. What follows is a general description of our modeling approach and the specific prior formulations that will be investigated through the remainder of the paper.

We assume the n observations y_i, where $i = 1, \ldots, n$, are independent and follow some distribution dependent on the unknown function values θ_i and possibly other parameters $\boldsymbol{\xi}$ at discrete points t. We further assume that the order-k forward differences in the function parameters, $\Delta^k\theta_j$, where $j = 1, \ldots, n-k$, are independent and identically distributed conditional on a global scale parameter which is a function of the smoothing parameter λ. These assumptions result in the following general hierarchical form:

$$y_i \mid \theta_i, \boldsymbol{\xi} \sim p(y_i \mid \theta_i, \boldsymbol{\xi}), \quad \Delta^k\theta_j \mid \lambda \sim p(\Delta^k\theta_j \mid \lambda), \quad \lambda \sim p(\lambda), \quad \boldsymbol{\xi} \sim p(\boldsymbol{\xi}). \tag{2}$$

One convenient trait of many shrinkage priors, including the Laplace, the logistic, and the t-distribution, is that they can be represented as scale mixtures of normal distributions (Andrews and Mallows, 1974; West, 1987; Polson and Scott, 2010). The conditional form of scale mixture densities leads naturally to hierarchical representations. This can allow some otherwise intractable density functions to be represented hierarchically with standard distributions and can ease computation. To take advantage of this hierarchical structure, we restrict densities $p(\Delta^k\theta_j \mid \lambda)$ to be scale mixtures of normals, which allows us to induce a hierarchical form to our model formulation by introducing latent local scale parameters, τ_j. Here the order-k differences in the function parameters, $\Delta^k\theta_j$, are conditionally normally distributed with mean zero and variance τ_j^2, and the τ_j are independent and identically distributed with a global scale parameter which is a function of the smoothing parameter λ. The distribution statement for $\Delta^k\theta_j$ in Equation (2) can then be replaced with the following hierarchical representation:

$$\Delta^k\theta_j \mid \tau_j \sim \mathrm{N}(0, \tau_j^2), \quad \tau_j \mid \lambda \sim p(\tau_j \mid \lambda). \tag{3}$$

To complete the model specification, we place proper priors on $\theta_1, \ldots, \theta_k$. This maintains propriety and can improve computational performance for some Markov chain Monte Carlo (MCMC) samplers. We start by setting $\theta_1 \sim \mathrm{N}(\mu, \omega^2)$, where μ and ω can be constants or allowed to follow their own distributions. Then for $k \geq 2$ and $h = 1, \ldots, k-1$, we let $\Delta^h\theta_1 \mid \alpha_h \sim \mathrm{N}(0, \alpha_h^2)$ and $\alpha_h \mid \lambda \sim p(\alpha_h \mid \lambda)$, where $p(\alpha \mid \lambda)$ is the same form as $p(\tau \mid \lambda)$. That is, we assume the order-h differences are independent with scale parameters that follow the same distribution as the order-k differences. For most situations, the order of k will be less than 4, so issues of scale introduced by assuming the same distribution on the scale parameters for the lower and higher order differences will be minimal. One could alternatively adjust the scale parameter of each $p(\alpha_h \mid \lambda)$ to impose smaller variance for lower order differences.

For the remainder of the paper we investigate two specific forms of shrinkage priors: the Laplace and the horseshoe. We later compare the performance of these two priors to the case where the order-k differences follow identical normal distributions. The following provides specific descriptions of our shrinkage prior formulations.

Laplace. As we showed previously, this prior arises naturally from an L_1 penalty, making it the default prior for Bayesian versions of the lasso (Park and Casella, 2008) and trend filter. The Laplace distribution is leptokurtic and features high mass near zero and exponential tails (Figure 1). Various authors have investigated its shrinkage properties (Griffin et al., 2010; Kyung et al., 2010; Armagan et al., 2013). We allow the order-k differences $\Delta^k\theta_j$ to follow a Laplace distribution conditional on a global scale parameter $\gamma = 1/\lambda$, and we allow γ to follow a half-Cauchy distribution with scale parameter ζ. That is,

$$\Delta^k\theta_j \mid \gamma \sim \text{Laplace}(\gamma), \qquad \gamma \sim \mathrm{C}^+(0, \zeta). \tag{4}$$

The use of a half-Cauchy prior on γ is a departure from Park and Casella (2008), who make λ^2

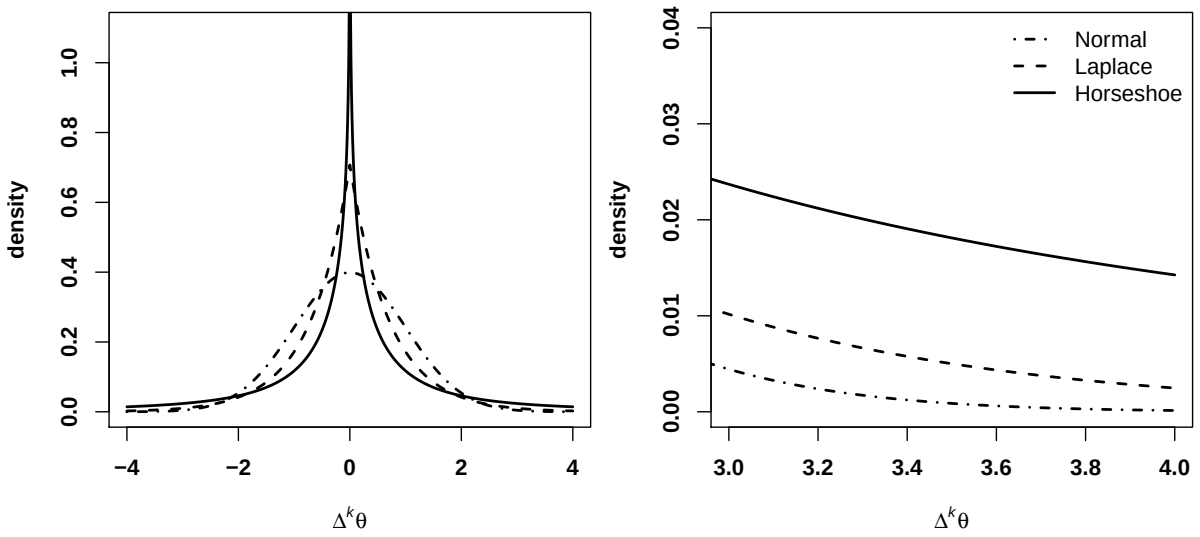

Figure 1: Shapes of prior distributions (left) and associated tail behavior (right) for priors used for $p(\Delta^k\theta \mid \lambda)$.

follow a gamma distribution to induce conjugacy in the Bayesian lasso. We chose to use the half-Cauchy prior on γ because its single parameter simplifies implementation, it has desirable properties as a prior on a scale parameter (Gelman et al., 2006; Polson and Scott, 2012b), and it allowed us to be consistent across methods (see horseshoe specification below). The hierarchical form of the Laplace prior arises when the mixing distribution on the square of the local scale parameter τ_j is an exponential distribution. Specifically, we specify $\tau_j^2 \mid \lambda \sim \text{Exp}(\lambda^2/2)$ and $\Delta^k\theta_j \mid \tau_j \sim \text{N}(0, \tau_j^2)$ in the hierarchical representation.

Horseshoe. The horseshoe prior (Carvalho et al., 2010) has an infinite spike in density at zero but also exhibits heavy tails (Figure 1). This combination results in excellent performance as a shrinkage prior (Polson and Scott, 2010), and gives the horseshoe shrinkage properties more similar to the spike-and-slab variable selection prior than those of the Laplace prior. We allow the order-k differences $\Delta^k\theta_j$ to follow a horseshoe distribution conditional on global scale parameter $\gamma = 1/\lambda$, and allow γ to follow a half-Cauchy distribution with scale parameter ζ. That is,

$$\Delta^k\theta_j \mid \gamma \sim \text{HS}(\gamma), \qquad \gamma \sim \text{C}^+(0, \zeta). \tag{5}$$

The horseshoe density function does not exist in closed form, but we have derived an approximate closed-form solution using the known function bounds (see Appendix A), which could be useful for application in some settings. Carvalho et al. (2010) represent the horseshoe density hierarchically as a scale mixture of normals where the local scale parameters τ_j are distributed half-Cauchy. In our hierarchical version, the latent scale parameter $\tau_j \mid \gamma \sim \text{C}^+(0, \gamma)$ and then conditional on τ_j the distribution on the order-k differences is $\Delta^k\theta_j \mid \tau_j \sim \text{N}(0, \tau_j^2)$.

The horseshoe prior arises when the mixing distribution on the local scale parameter τ_j is half-Cauchy, which is a special case of a half-t-distribution where degrees of freedom (df) equal 1. Setting $df > 1$ would result in a prior with lighter tails than the horseshoe, and setting $0 < df < 1$ would result in heavier tails. We tested half-t formulations with df between 1 and 5 in test scenarios, but did not find an appreciable difference in performance relative to the horseshoe. We also attempted to place a prior distribution on the df parameter, but found the data to be insufficient to gain information in the posterior for df in our test scenarios, so we did not pursue this further.

Normal. The normal distribution arises as a prior on the order-k differences when the penalty in the penalized likelihood formulation is a squared L_2 penalty. The normal prior is also the form of prior used in Bayesian smoothing splines. The normal is not considered a shrinkage prior and does not have the flexibility to allow locally-adaptive smoothing behavior. We use it for comparison to demonstrate the local adaptivity allowed by the shrinkage priors. For our investigations, the distribution on the order-k differences and associated scale parameter is:

$$\Delta^k \theta_j \mid \gamma \sim \mathrm{N}(0, \gamma^2), \qquad \gamma \sim \mathrm{C}^+(0, \zeta). \tag{6}$$

2.3 Connections to Markov Random Fields

Here we briefly show the models represented by (2) can be expressed with GMRF priors for $\boldsymbol{\theta}$ conditional on the local scale parameters $\boldsymbol{\tau}$. It is instructive to start with the normal increments model (6), which belongs to a class of time series models known as autoregressive models of order k. Rue and Held (2005) call this model a k-th order random walk and show that it is a GMRF with respect to a k-th order chain graph — a graph with nodes $\{1, 2, \ldots, n\}$, where the nodes $i \neq j$ are connected by an edge if and only if $|i - j| \leq k$. Since the normal model (6) does not fully specify the joint distribution of $\boldsymbol{\theta}$, it is an intrinsic (improper) GMRF. We make it a proper GMRF by specifying a prior density of the first k components of $\boldsymbol{\theta}$, $p(\theta_1, \ldots, \theta_k)$. The Markov property of the model manifests itself in the following factorization:

$$p(\boldsymbol{\theta}) = p(\theta_1, \ldots, \theta_k) p(\theta_{k+1} \mid \theta_1, \ldots, \theta_k) \cdots p(\theta_n \mid \theta_{n-1}, \ldots, \theta_{n-k}).$$

Equipped with initial distribution $p(\theta_1, \ldots, \theta_k)$, models (5) and (4) also admit this factorization, so they are k-th order Markov, albeit not Gaussian models. However, if we condition on the latent scale parameters $\boldsymbol{\tau}$, both the Laplace and horseshoe models become GMRFs, or more specifically k-th order normal random walks. One important feature of these random walks is that each step in the walk has its own precision. To recap, under prior specifications (5) and (4) $p(\boldsymbol{\theta} \mid \gamma)$ is a non-Gaussian Markov field, while $p(\boldsymbol{\theta} \mid \boldsymbol{\tau}, \gamma) = p(\boldsymbol{\theta} \mid \boldsymbol{\tau})$ is a GMRF.

Our GMRF point of view is useful in at least two respects. First, GMRFs with constant precision have been used for nonparametric smoothing in many settings (see Rue and Held (2005) for examples). GMRFs with nonconstant precision have been used much less frequently, but one important application is to the development of adaptive smoothing splines by allowing order-k increments to have nonconstant variances (Lang et al., 2002; Yue et al., 2012). The approach of these authors is very similar to our own but differs in at least two important ways. First, we specify the prior distribution on the latent local scale parameters τ_j with the resulting marginal distribution of $\Delta^k \theta_j$ in mind, such as the Laplace or horseshoe distributions which arise as scale mixtures of normals. This allows a better understanding of the adaptive properties of the resulting marginal prior in advance of implementation. In contrast, Lang et al. (2002) and Yue et al. (2012) appear to choose the distribution on local scale parameters based on conjugacy and do not consider the effect on the marginal distribution of $\Delta^k \theta_j$. Second, we allow the local scale parameters τ_j to be independent, whereas Lang et al. (2002) and Yue et al. (2012) impose dependence among the scale (precision) parameters by forcing them to follow another GMRF. Allowing the local scale parameters to be independent allows the model to be more flexible and able to adapt to jumps and sharp local features. We should also note that Rue and Held (2005) in section 4.3 show that the idea of scale mixtures of normal distributions can be used with GMRFs to generate order-k differences which marginally follow a t-distribution by introducing latent local scale parameters. Although they do not pursue this further, we mention it because

it bears similarity to our approach.

The second advantage of connecting our Bayesian trend filter models to GMRFs is that the GMRF representation allows us to connect our first oder Markov models to subordinated Brownian motion (Bochner, 1955; Clark, 1973), a type of Lévy process recently studied in the context of scale mixture of normal distributions (Polson and Scott, 2012a). Polson and Scott (2012a) use the theory of Lévy processes to develop shrinkage priors and penalty functions. Let us briefly consider a simple example of subordinated Brownian motion. Let W be a Weiner process, so that $W(t+s) - W(t) \sim \mathrm{N}(0, s\sigma^2)$, and W has independent increments. Let T be a subordinator, which is a Lévy process that is non-decreasing with probability 1, has independent increments, and is independent of W. The subordinated process Z results from observing W at locations $T(t)$. That is, $Z(t) = W[T(t)]$. The subordinator essentially generates a random set of irregular locations over which the Brownian motion is observed, which results in a new process. In our hierarchical representation of Laplace and horseshoe priors *for the first order differences*, we can define a subordinator process $T_j = \sum_{i=1}^{j} \tau_i^2$, so that the GMRF $p(\boldsymbol{\theta} \mid \boldsymbol{\tau})$ can be thought of as a subordinated Brownian motion or as a realization of a Brownian motion with unit variance on the random latent irregular grid $T_1, \ldots, T_n$. The subordinated Brownian motion interpretation is not so straight forward when applied to higher-order increments, but we think this interpretation will be fruitful for extending our Bayesian trend filter in the future.

2.4 Posterior Computation

Since we have two general model formulations, marginal and hierarchical, we could approximate the posterior distribution of heights of our piecewise step functions, $\boldsymbol{\theta}$, by approximating one of the two posterior distributions. The first one corresponds to the marginal model formulation:

$$p(\boldsymbol{\theta}, \gamma, \boldsymbol{\xi} \mid \mathbf{y}) \propto \prod_{i=1}^{n} p(y_i \mid \theta_i, \boldsymbol{\xi}) p(\boldsymbol{\theta} \mid \gamma) p(\boldsymbol{\xi}), \tag{7}$$

where $p(\boldsymbol{\theta} \mid \gamma)$ is a Markov field induced by the normal, or by the Laplace, or by the horseshoe densities. The second posterior corresponds to the hierarchical model with latent scale parameters $\boldsymbol{\tau}$:

$$p(\boldsymbol{\theta}, \boldsymbol{\tau}, \gamma, \boldsymbol{\xi} \mid \mathbf{y}) \propto \prod_{i=1}^{n} p(y_i \mid \theta_i, \boldsymbol{\xi}) p(\boldsymbol{\theta} \mid \boldsymbol{\tau}) \prod_{j=1}^{n-k} p(\tau_j \mid \gamma) p(\boldsymbol{\xi}), \tag{8}$$

where $p(\boldsymbol{\theta} \mid \boldsymbol{\tau})$ is a GMRF and the choice of $p(\tau_j \mid \gamma)$ makes the marginal prior specification for $\boldsymbol{\theta}$ correspond either to a Laplace or to a horseshoe Markov random field. Notice that the GMRF has only the marginal specification.

Both of the above model classes are highly parameterized with dependencies among parameters induced by differencing and the model hierarchy. It is well known that high-dimensional, hierarchical models with strong correlations among parameters can create challenges for standard MCMC samplers, such as random walk Metropolis or Gibbs. When faced with these challenges, random walk behavior can result in inefficient exploration of the parameter space, which can lead to poor mixing and prohibitively long convergence times. Many approaches have been proposed to deal with these issues, including block updating (Knorr-Held and Rue, 2002), elliptical slice sampling (Murray et al., 2010; Murray and Adams, 2010), the Metropolis adjusted Langevin algorithm (MALA) (Roberts and Stramer, 2002), and Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 1993, 2011). All of these approaches jointly update some or all of the parameters at each MCMC iteration, which usually improves mixing and speeds up convergence of MCMC. We used a modification of HMC proposed by Hoffman and

Gelman (2014) which automatically adjusts HMC tuning parameters. We used the open source package `rstan` (Stan Development Team, 2015a), which provides a platform for fitting models using HMC in the `R` computing environment (R Core Team, 2014). Even with HMC, we experienced severe problems with mixing of MCMC that targets the marginal posterior (7). In contrast, MCMC with stationary distribution equal to the posterior (8) worked well in nearly all of our numerical experiments. Therefore, in the rest of the manuscript we work with the hierarchical model posterior distribution (8).

We note that the performance of HMC for our models was greatly improved by applying the non-centered parameterization methods described by Papaspiliopoulos et al. (2003, 2007) and Betancourt and Girolami (2013). Non-centered parameterizations break the dependencies among parameters by introducing deterministic transformations of the parameters. The MCMC algorithm then operates directly on the independent parameters. Betancourt and Girolami (2013) discuss non-centered parameterizations in the context of HMC, and further examples are provided in the documentation for `stan` (Stan Development Team, 2015b). We developed an `R` package titled `bnps` which allows for easy implementation of our models via a wrapper to the `rstan` tools. The package code is publicly available at https://github.com/jrfaulkner/bnps.

3 Simulation Study

3.1 Simulation Protocol

We use simulations to investigate the performance of two Bayesian trend filter formulations using the Laplace and horseshoe shrinkage priors described in section (2.2) and compare results to those using a normal distribution on the order-k differences. We refer to the shrinkage prior methods as adaptive due to the local scale parameters, and the method with normal prior as non-adaptive due to the use of a single scale parameter. We constructed underlying trends with a variety of characteristics following approaches similar to those of other authors (Scheipl and Kneib, 2009; Yue et al., 2012; Zhu and Dunson, 2013). We investigated four different types of underlying trend (constant, piecewise constant, smooth function, and function with varying smoothness). The first row of Figure 2 shows examples of the trend functions, each illustrated with simulated normal observations centered at the function values over a regular grid. We used three observation types for each trend type where the observations were conditionally independent given the trend function values θ_i, where $i = 1, \ldots, n$. The observation distributions investigated were 1) normal: $y_i \,|\, \theta_i \sim \text{N}(\theta_i, \sigma^2)$, where $\sigma = 1.5$ or $\sigma = 4.5$; 2) Poisson: $y_i \,|\, \theta_i \sim \text{Pois}(\exp(\theta_i))$; and 3) binomial: $y_i \,|\, \theta_i \sim \text{Binom}(m, (1 + \exp(-\theta_i))^{-1})$, where $m = 20$ for all scenarios.

Note that we constructed the function values for the scenarios with normally distributed observations so that each function would have approximately the same mean and variance, where the mean and variance were calculated across the function values realized at the discrete time points. This allowed us to specify observation variances which resulted in the same signal-to-noise ratio for each function, where signal-to-noise ratio is defined as the standard deviation of function values divided by the standard deviation of observations. The signal-to-noise ratios for our scenarios with normal observations were 6 for $\sigma = 1.5$ and 2 for $\sigma = 4.5$. We describe the trend functions further in what follows.

Constant. This scenario uses a constant mean across all points. We use this scenario to investigate the ability of each method to find a straight horizontal line in the presence of noisy data. The values used for the constant mean were 20 for normal and Poisson observations, and 0.5 for binomial observations.

Piecewise constant. This type of function has been used by Tibshirani (2014) and others such as Scheipl and Kneib (2009) and Zhu and Dunson (2013). The horizontal trends combined with sharp breaks offer a difficult challenge for all methods. For the scenarios with normal or Poisson observations, the function values were 25, 10, 35, and 15 with break points at $t \in \{20, 40, 60\}$. For the binomial observations the function values on the probability scale were 0.65, 0.25, 0.85, and 0.45 with the same break points as the other observation types.

Smooth trend. We use this as an example to test the ability of the adaptive methods to handle a smoothly varying function. We generated the function $\boldsymbol{f}$ as a GP with squared exponential covariance function. That is, $\boldsymbol{f} \sim \text{GP}(\mu, \boldsymbol{\Sigma}), \Sigma_{i,j} = \sigma_f^2 \exp\left[-(t_j - t_i)^2/(2\rho^2)\right]$, where $\Sigma_{i,j}$ is the covariance between points i and j, $\sigma_f^2 > 0$ is the signal variance and $\rho > 0$ is the length scale. We set $\mu = 10$, $\sigma_f^2 = 430$, and $\rho = 10$ for the scenarios with normal or Poisson observations. For binomial observations, $\boldsymbol{f}$ was generated in logit space with $\mu = -0.5$, $\sigma_f^2 = 3$, and $\rho = 10$ and then back-transformed to probability space. For all scenarios the function was generated with the same random number seed.

Varying smoothness. This function with varying smoothness was initially presented by DiMatteo et al. (2001) and later used by others, including Yue et al. (2012). We adapted the function to a uniform grid, $t \in [1, n]$, where $n = 100$ in our case, resulting in the function

$$g(t) = \sin\left(\frac{4t}{n} - 2\right) + 2\exp\left(-30\left(\frac{4t}{n} - 2\right)^2\right).$$

For the normal and Poisson observations we made the transformation $f(t) = 20 + 10g(t)$. For binomial observations we used $f(t) = 1.25g(t)$ on the logit scale.

We generated 100 datasets for each combination of trend and observation type. Each dataset had 100 equally-spaced sample points over the interval $[1, 100]$. For each dataset we fit models representing three different prior formulations for the order-k differences, which were 1) normal, 2) Laplace, and 3) horseshoe. We used the hierarchical prior representations for these models given in Section 2.2. We selected the degree of k-th order differences for each model based on knowledge of the shape of the underlying function. We fit first-order models for the constant and piecewise constant functions, and we fit second-order models for the smooth and varying smooth functions. For the scenarios with normal observations, we set $\sigma \sim \text{C}^+(0, 5)$. In all cases, $\theta_1 \sim \text{N}(\mu, \omega^2)$, where μ is set to the sample mean and ω is two times the sample standard deviation of the observed data transformed to match the scale of θ. We also set $\gamma \sim \text{C}^+(0, 0.01)$ for all models.

We used HMC to approximate the posterior distributions. For each model we ran four independent chains with different randomly generated starting parameter values and initial burn-in of 500 iterations. For all scenarios except for normal observations with $\sigma = 1.5$, each chain had 2,500 posterior draws post-burn-in that were thinned to keep every 5th draw. For scenarios with normal observations with $\sigma = 1.5$, chains with 10,000 iterations post-burn-in were necessary, with additional thinning to every 20th draw. In all cases, these settings resulted in 2,000 posterior draws retained per model. We found that these settings consistently resulted in good convergence properties, where convergence and mixing were assessed with a combination of trace plots, autocorrelation values, effective sample sizes, and potential scale reduction statistics (Gelman and Rubin, 1992).

We assessed the relative performance of each model using three different summary statistics. We compared the posterior medians of the trend parameters ($\hat{\theta}_i$) to the true trend values

Table 1: Mean values of performance measures across 100 simulations for normal observations ($\sigma = 4.5$) for each model and trend function type.

Function	Model	SRE	MRW	Variation	True Var.
Constant	Normal	1.71	0.10	0.30	0.00
	Laplace	1.69	0.10	0.30	0.00
	Horseshoe	1.78	0.12	0.60	0.00
Piecewise Const.	Normal	12.78	0.66	157.06	60.00
	Laplace	11.01	0.60	142.66	60.00
	Horseshoe	5.35	0.36	68.18	60.00
Smooth	Normal	8.78	0.44	131.48	139.21
	Laplace	8.77	0.45	131.62	139.21
	Horseshoe	9.07	0.45	134.50	139.21
Varying Smooth	Normal	7.84	0.36	42.19	53.79
	Laplace	7.65	0.35	42.77	53.79
	Horseshoe	6.16	0.31	46.49	53.79

(θ_i) using the sum of relative errors (SRE):

$$\text{SRE} = \sum_{i=1}^{n} \frac{|\hat{\theta}_i - \theta_i|}{\theta_i}. \tag{9}$$

We assessed the width of the 95% Bayesian credible intervals (BCIs) using the mean relative width (MRW):

$$\text{MRW} = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{\theta}_{97.5,i} - \hat{\theta}_{2.5,i}|}{\theta_i}, \tag{10}$$

where $\hat{\theta}_{97.5,i}$ and $\hat{\theta}_{2.5,i}$ are the 97.5% and 2.5% quantiles of the posterior distribution for θ_i. We also computed the variation of $\hat{\boldsymbol{\theta}}$ as

$$\text{variation} = \sum_{i=1}^{n-1} |\hat{\theta}_{i+1} - \hat{\theta}_i|. \tag{11}$$

We compared the observed variation to the true variation in the underlying function, which is calculated by substituting true θ's into equation for variation.

3.2 Simulation Results

In the interest of space, we emphasize results for the scenarios with normally distributed observations with $\sigma = 4.5$ here. This level of observation variance was similar to that for Poisson and binomial observations and therefore offered results similar to those scenarios. We follow these results with a brief summary of results for the other observation types, and we provide further summary of other results in Appendix B.

Constant. The three models performed similarly in terms of absolute value of all the metrics (Table 1 and Figure 2), but the Laplace and normal models were slightly better at fitting straight lines than the horseshoe. This is evidenced by the fact that the horseshoe had larger MRW and larger variation than the other methods. The first column of plots in Figure 3 provides a visual example of the extra variation exhibited by the horseshoe.

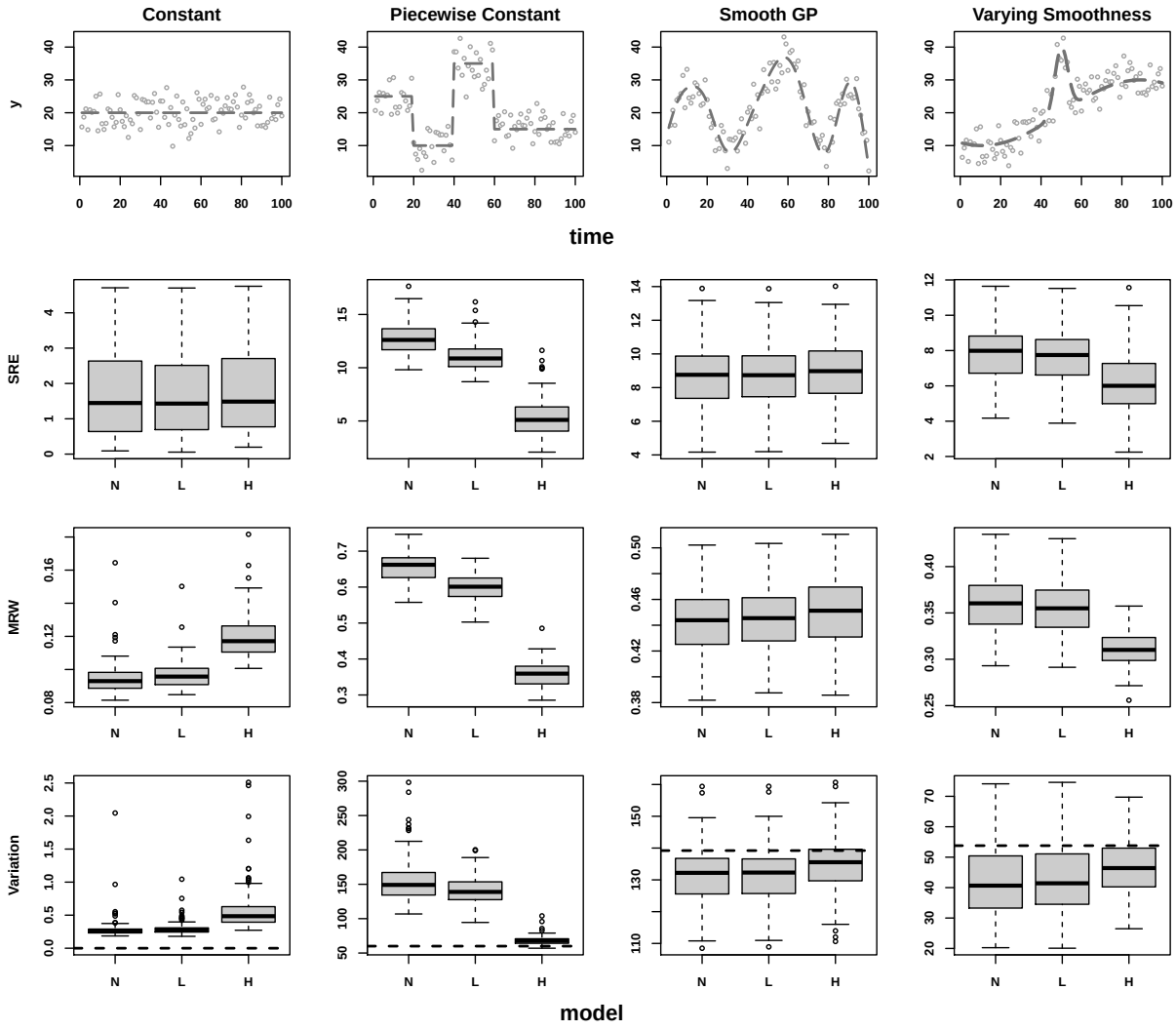

Figure 2: Functions used in simulations and simulation results by model (N=normal, L=Laplace, H=horseshoe) and function type for normally distributed data with $\sigma = 4.5$. Top row shows true functions (dashed lines) with example simulated data. Remaining rows show sum of relative errors (SRE), mean relative width (MRW), and sample variation. Horizontal dashed line in plots on bottom row is the true function variation.

Piecewise constant. The horseshoe model performed the best in all categories for this scenario and the normal model performed the worst (Table 1 and Figure 2). The Laplace model was closer to the normal model in performance. The horseshoe was flexible enough to account for the large function breaks yet still able to limit variation in the constant segments. Example fits for the piecewise constant function are shown in the second column of plots in Figure 3.

Smooth trend. The different models were all close in value of the performance metrics for the smooth trend scenario (Table 1 and Figure 2). The normal and Laplace models had smallest SRE, but the horseshoe had variation closer to the true variation. The fact that the values of the metrics were similar for all models suggests that not much performance is lost in fitting a smooth trend with the adaptive methods in comparison to non-adaptive.

Varying Smoothness. Again the models all performed similarly in terms of absolute value of the metrics, but there was a clear ordering among models in relative performance (Table 1 and

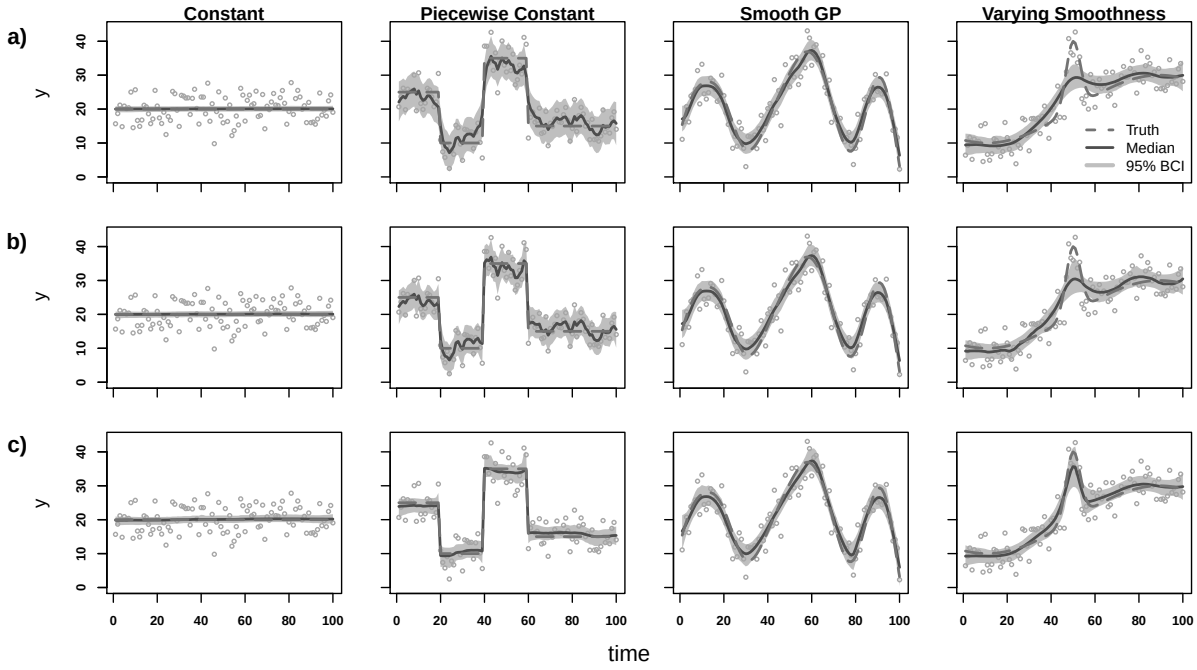

Figure 3: Example fits for models using a) normal, b) Laplace, and c) horseshoe priors where observations are drawn from normal distributions with SD = 4.5. Plots show true functions (dashed gray lines), posterior medians (solid dark gray lines), and associated Bayesian credible intervals (BCI; gray bands) for each θ. Values between observed locations are interpolated for plotting.

Figure 2). The horseshoe model performed the best relative to the other models on all metrics. This function forces a compromise between having large enough local variance to capture the spike and small enough local variance to remain smooth through the rest of the function. The horseshoe was more adaptive than the other two methods and therefore better able to meet the compromise. The plots in the last column of Figure 3 provide example fits for this function.

The results for the scenarios with normal observations with $\sigma = 1.5$ and Poisson and binomial observations (Appendix B) showed similar patterns to those with normal observations and $\sigma = 4.5$. For the constant function, the normal prior performed the best and the horseshoe prior the worst, although differences in terms of absolute values of the performance metrics were small. The relative differences were more pronounced with the scenarios with normal observations with $\sigma = 1.5$. For the piecewise constant function, the horseshoe prior performed the best for all scenarios and the normal prior the worst. All methods performed similarly for the smooth function, with the normal and Laplace generally performing a little better than the horseshoe. For the function with varying smoothness, the horseshoe performed the best and the normal the worst for all scenarios.

4 Data Examples

Here we provide two examples of fitting Bayesian trend filter models to real data. Each example uses a different probability distribution for the observations. The first example exhibits a change point, which makes it amenable to adaptive smoothing methods. The second example has a more uniformly smooth trend but also shows a period of rapid change, so represents a test for

all methods.

4.1 Coal Mining Disasters

This is an example of estimating the time-varying intensity of an inhomogeneous Poisson process that exhibits a relatively rapid period of change. The data are on the time intervals between successive coal-mining disasters, and were originally presented by Maguire et al. (1952), with later corrections given by Jarrett (1979) and Raftery and Akman (1986). We use the data format presented by Raftery and Akman (1986). A disaster is defined as an accident involving 10 or more deaths. The first disaster was recorded in March of 1851 and the last in March of 1962, with 191 total event times during the period 1 January, 1851 through 31 December, 1962. Visual inspection of the data suggests a decrease in rate of disasters over time, but it is unclear by eye alone whether this change is abrupt or gradual. The decrease in disasters is associated with a few changes in the coal industry at the time. A sharp decline in labor productivity at the end of the 1880's is thought to have decreased the opportunity for disasters, and the formation of the Miner's Federation, a labor union, in late 1889 brought added safety and protection to the workers (Raftery and Akman, 1986).

This data set has been of interest to various authors due to uncertainty in the timing and rate of decline in disasters and the computational challenge presented by the discrete nature of the observations. Some authors have fit smooth curves exhibiting gradual change (Adams et al., 2009; Teh and Rao, 2011) and others have fit change-point models with abrupt, instantaneous change (Raftery and Akman, 1986; Carlin et al., 1992; Green, 1995). An ideal model would provide the flexibility to automatically adapt to either scenario.

We assumed an inhomogeneous Poisson process for the disaster events and binned the event counts by year. We fit first-order models using the normal, Laplace, and horseshoe prior formulations. We assumed the event counts, y_i, were distributed Poisson conditional on the θ_i: $y_i \,|\, \theta_i \sim \text{Pois}\,(\exp(\theta_i))$. The marginal prior distributions for the first-order increments were $\Delta\theta_i \sim \text{N}(0, \gamma^2)$ for the Normal, $\Delta\theta_j \sim \text{Laplace}(\gamma)$ for the Laplace, and $\Delta\theta_j \sim \text{HS}(\gamma)$ for the horseshoe. We used the same prior specifications as those used in the simulations for the remaining parameters. We used HMC for approximating the posterior distributions. For each model we ran four independent chains, each with a burn-in of 500 followed by 6,250 iterations thinned at every 5. This resulted in a total of 5,000 posterior samples for each model. We were interested in finding the best representation of the process over time as well as finding the most likely set of years associated with the apparent change point. For this exercise we arbitrarily defined a change point as the maximum drop in rate between two consecutive time points.

Plots of the fitted trends (Figure 4) indicate that the horseshoe model picked up a sharper change in trend and had narrower BCIs than the other models. The normal and Laplace models did not have sufficient flexibility to allow large jumps and produced a gradual decline in accidents rate, which is less plausible than a sharp decline in light of the additional information about change in coal mining industry safety regulations. The relative qualitative performance of the normal, Laplace, and horseshoe densities is similar to that for the piecewise constant scenario from our simulation study. The posterior distributions of the change point times are shown in Figure 4. The horseshoe model clearly shows a more concentrated posterior for the break points, and that distribution is centered near the late 1880's, which corresponds to the period of change in the coal industry. Therefore, we think the Bayesian trend filter with the horseshoe prior is a better default model in cases where sharp change points are expected.

It is important to point out that we tried other values for the scale parameter (ζ) in the prior distribution for γ and found that the models were somewhat sensitive to that hyperparameter

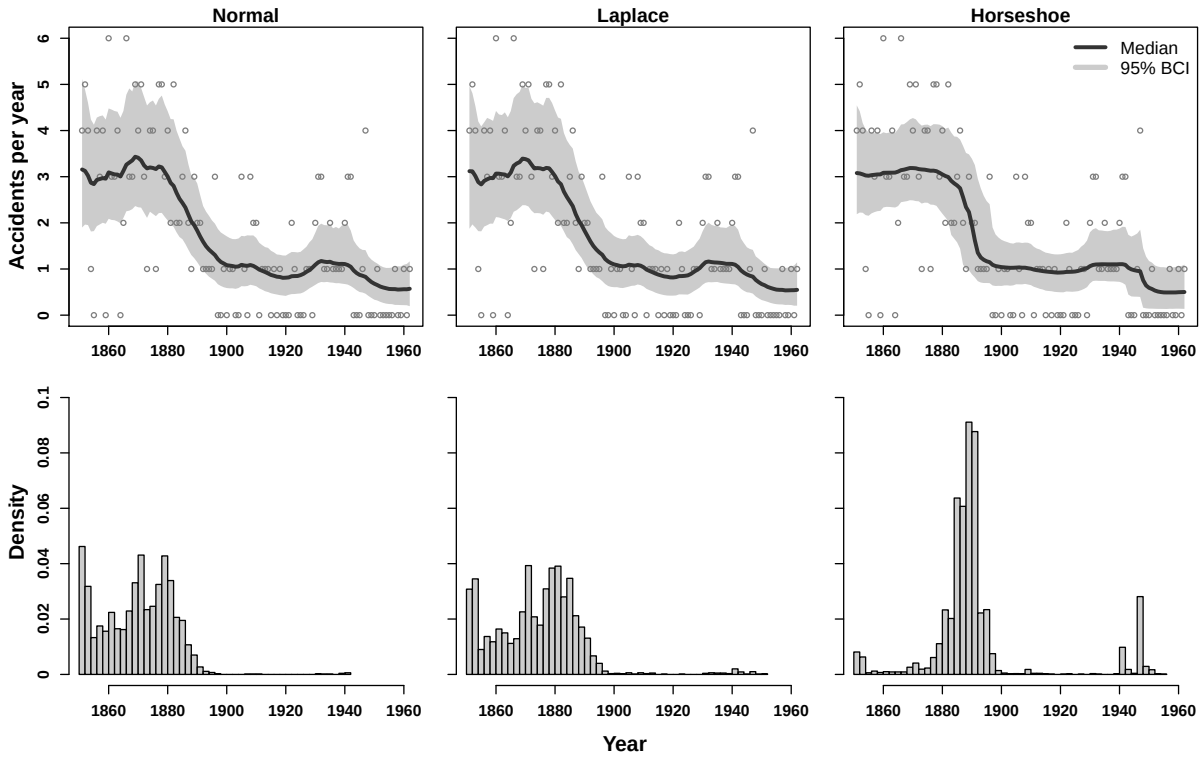

Figure 4: Top row: fits to coal mining disaster data for different prior distributions. Posterior medians (lines), 95 % credible intervals (shaded regions), and data points are shown. Bottom row: associated posterior distributions for changepoints.

for this data set. In particular, the horseshoe results for $\zeta = 1$ looked more like those for the other two models in Figure 4, but when $\zeta = 0.0001$, the horseshoe produced more defined break points and straighter lines with narrower BCIs compared to the results with $\zeta = 0.01$ (see Figure C.1 in Appendix C).

4.2 Tokyo Rainfall

This problem concerns the estimation of the time-varying mean of an inhomogeneous binomial process. We are interested in estimating the seasonal trend in daily probability of rainfall. The data are binary indicators of when daily rainfall exceeded 1 mm in Tokyo, Japan, over the course of 39 consecutive years (1951-1989). The indicators were combined by day of year across years, resulting in a sample size of $m = 39$ for each of 365 out of 366 possible days, and a size of $m = 10$ for the additional day that occurred in each of the 10 leap years. The observation variable y is therefore a count, where $y \in \{0, 1, \ldots, 39\}$. Data were obtained from the NOAA's National Center for Climate Information (https://www.ncdc.noaa.gov). A smaller subset of these data (1983-1984) was initially analyzed by Kitagawa (1987) and later by several others, including Rue and Held (2005).

We fit the Bayesian trend filter using the Laplace and horseshoe priors and a model using normal prior . All models were based on second order differences. The observation model was

$$y_i \mid \theta_i \sim \text{Bin}\left(m_i,\ \frac{1}{1 + \exp(-\theta_i)}\right),$$

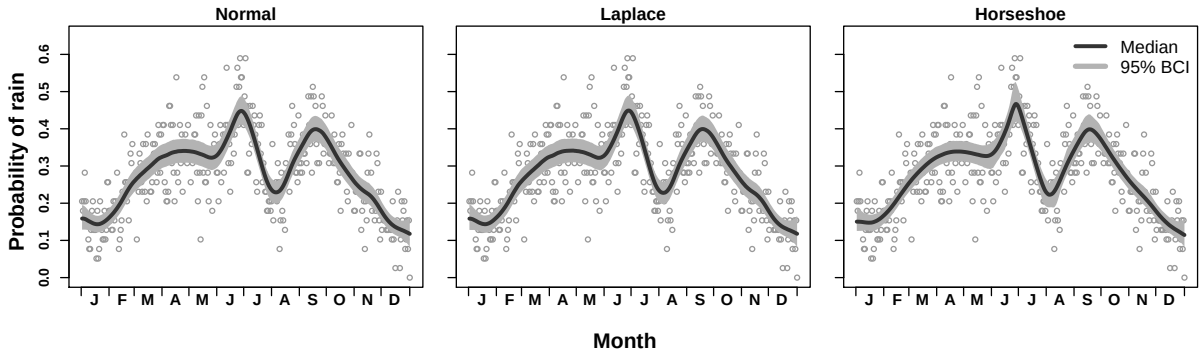

Figure 5: Fits to Tokyo rainfall data for different prior distributions. Posterior medians (lines), 95 % credible intervals (shaded regions), and estimated probabilities (y_i/n_i) are shown.

and the marginal prior distributions for the second-order differences were $\Delta^2\theta_j \sim \mathrm{N}(0, \gamma^2)$ for the normal prior, $\Delta^2\theta_j \sim \mathrm{Laplace}(\gamma)$ for the Laplace, and $\Delta^2\theta_j \sim \mathrm{HS}(\gamma)$ for the horseshoe. We used the same prior specifications as those used in the simulations for the remaining parameters. For each model we ran four independent chains, each with a burn-in of 500 followed by 6,250 draws thinned at every 5. This resulted in a total of 5,000 MCMC samples retained for each model.

The resulting function estimates for all models reveal a sharp increase in probability of rain in June followed by a sharp decrease through July and early August and a subsequent sharp increase in late August and September (Figure 5). Changes through the rest of the months were relatively smooth. The estimated function displays some variations in smoothness similar to the function with varying smoothness used in our simulations. All methods resulted in a similar estimated function, but the horseshoe prior resulted in a smoother function that displayed sharper features at transition points in late June and early August, yet also had narrower credible intervals over most of the function. The normal and Laplace models resulted in a little more variability in the trend in January-April and in November. In their analysis of a subset of these data, Rue and Held (2005) used a circular constraint to tie together the endpoints of the function at the beginning and end of the year. We did not use such a constraint here, but it is evident that the horseshoe model resulted in more similar function estimates at the endpoints than did the other two models.

5 Discussion

We presented a method for curve fitting in a Bayesian context that achieves locally adaptive smoothing by exploiting the sparsity-inducing properties of shrinkage priors and the smoothing properties of GMRFs. We compared the performance of the Laplace prior, which simply reformulates the frequentist trend filter to a Bayesian analog, to a more aggressive horseshoe shrinkage prior by using simulations and found that the horseshoe provided the best balance between bias and precision. The horseshoe prior has the greatest concentration of density near zero and the heaviest tails among the priors we investigated. This combination allows smooth functions to be fit in regions with weak signals or noisy data while still allowing for recovery of sharp functional changes when supported by informative data. The Laplace prior allowed more functional changes of moderate value to be retained and could not accommodate large changes without compromising the ability to shrink the noisy and smaller functional changes.

This resulted in greater variability in the estimated functions and wider associated credible intervals for the models with the Laplace prior in comparison to those with the horseshoe prior when the underlying true functions had jumps or varying smoothness. The Laplace prior did have adaptive ability not possessed by the normal prior, but the horseshoe prior clearly had the best adaptive properties among the priors we investigated.

The Laplace prior performed better than the horseshoe for the constant and smooth functions in our simulations, with results closer to those of the normal prior, although the differences in performance among the three methods were relatively small. These functions do not have large deviations in order-k differences, and so there are many small or medium sized values for the estimated $\Delta^k\theta$. This situation is reflective of cases described by Tibshirani (1996) where the lasso and ridge regression perform best, which helps explain why the analogous Bayesian trend filter models with Laplace or normal prior distributions do better here. We expect that non-adaptive or mildly adaptive methods will perform better when used on functions which do not exhibit jumps or varying smoothness. However, it is reassuring that an adaptive method does nearly as well as a non-adaptive method for these functions. This allows an adaptive model such as that using the horseshoe to be applied to a variety of functions with minimal risk of performance loss.

Our fully Bayesian implementation of the trend filter eliminates the need to explicitly select the global smoothing parameter λ. Some alternative methods include cross-validation in the frequentist setting (Tibshirani, 1996) and marginal maximum likelihood in the empirical Bayes setting (Park and Casella, 2008). However, the fully Bayesian approach does still require attention to the selection of the hyperparameter that controls the prior distribution on the smoothing parameter. A highly informative prior on the global smoothing parameter can result in over-smoothing if the prior overwhelms the information in the data, while a diffuse prior may result in a rougher function with insufficient smoothing. Noisier data are therefore more sensitive to choice of parameterization of the prior on the global smoothing parameter. We tested prior sensitivity in the coal mining example and found that the horseshoe prior was more responsive to changes in hyperparmeter values than the normal and Laplace priors (see Appendix C). Therefore, we recommend paying attention to prior sensitivity when analyzing noisy data with the horseshoe-based Bayesian trend filter.

We only addressed situations where observations occur on one-dimensional discrete grids with uniform spacing. However, the methods presented here can be extended to non-uniform grids or to continuously measured variables, as well as to higher-dimensional surfaces. Rue and Held (2005) and Lindgren and Rue (2008) provide methods for extending GMRF models to irregular grids, and Rue and Held (2005) provide extensions to higher dimensions. Our GMRF representation of adaptive smoothing should allow for porting these results to the Bayesian trend filter.

6 Acknowledgments

J.R.F, and V.N.M. were supported by the NIH grant R01 AI107034. V.N.M. was supported by the NIH grant U54 GM111274.

References

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodol-*

ogy) **60,** 725–749.

Adams, R. P., Murray, I., and MacKay, D. J. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16. ACM.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* **36,** 99–102.

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica* **23,** 119–143.

Betancourt, M. and Girolami, M. (2013). Hamiltonian Monte Carlo for hierarchical models. *arXiv preprint arXiv:1312.0906* .

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2014). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* .

Bochner, S. (1955). *Harmonic Analysis and the Theory of Probability*. University of California Press.

Brahim-Belhouari, S. and Bermak, A. (2004). Gaussian process for nonstationary time series prediction. *Computational Statistics & Data Analysis* **47,** 705–712.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics* **41,** 389–405.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97,** 465–480.

Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica* **41,** 135–155.

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88,** 1055–1071.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B* **195,** 216–222.

Figueiredo, M. A. (2003). Adaptive sparseness for supervised learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **25,** 1150–1159.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35,** 109–135.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1,** 515–534.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7,** 457–472.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82,** 711–732.

Griffin, J. E., Brown, P. J., et al. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5,** 171–188.

Griffin, J. E., Brown, P. J., et al. (2013). Some priors for sparse regression modelling. *Bayesian Analysis* **8,** 691–702.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12,** 55–67.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research* **15,** 1593–1623.

Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters. *Biometrika* **66,** 191–193.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics* pages 1700–1752.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). ℓ_1 trend filtering. *Siam Review* **51,** 339–360.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association* **82,** 1032–1041.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics* **29,** 597–614.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* **5,** 369–411.

Lang, S., Fronk, E.-M., and Fahrmeir, L. (2002). Function estimation with locally adaptive dynamic models. *Computational Statistics* **17,** 479–499.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics* **35,** 691–700.

Maguire, B. A., Pearson, E. S., and Wynn, A. H. A. (1952). The time intervals between industrial accidents. *Biometrika* **39,** 168–180.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83,** 1023–1032.

Murray, I. and Adams, R. P. (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems*, pages 1732–1740.

Murray, I., Adams, R. P., and Mackay, D. (2010). Elliptical slice sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 541–548.

Neal, R. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo* **2,**.

Neal, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Neal, R. M. (1998). Regression and classification using Gaussian process priors. *Bayesian statistics* **6,** 475–501.

Paciorek, C. and Schervish, M. (2004). Nonstationary covariance functions for Gaussian process regression. *Advances in Neural Information Processing Systems* **16,** 273–280.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* **17,** 483–506.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003). Non-centered parameterisations for hierarchical models and data augmentation. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, pages 307–326. Oxford University Press, USA.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science* **22,** 59–73.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103,** 681–686.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics* **9,** 501–538.

Polson, N. G. and Scott, J. G. (2012a). Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Methodological)* **74,** 287–311.

Polson, N. G. and Scott, J. G. (2012b). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis* **7,** 887–902.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E. and Akman, V. E. (1986). Bayesian analysis of a Poisson process with a changepoint. *Biometrika* **73,** 85–89.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik* **10,** 177–183.

Reményi, N. and Vidakovic, B. (2015). Wavelet shrinkage with double Weibull prior. *Communications in Statistics-Simulation and Computation* **44,** 88–104.

Roberts, G. O. and Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability* **4,** 337–357.

Roualdes, E. A. (2015). Bayesian trend filtering. *arXiv preprint arXiv:1505.07710* .

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63,** 325–338.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. CRC Press.

Scheipl, F. and Kneib, T. (2009). Locally adaptive Bayesian p-splines with a normal-exponential-gamma prior. *Computational Statistics & Data Analysis* **53,** 3533–3552.

Speckman, P. L. and Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika* **90,** 289–302.

Stan Development Team (2015a). RStan: the R interface to Stan, Version 2.6.2.

Stan Development Team (2015b). *Stan Modeling Language Users Guide and Reference Manual, Version 2.6.2*.

Teh, Y. W. and Rao, V. (2011). Gaussian process modulated renewal processes. In *Advances in Neural Information Processing Systems*, pages 2474–2482.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58,** 267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73,** 273–282.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics* **42,** 285–323.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics* **39,** 1335–1371.

Wahba, G. (1975). Smoothing noisy data with spline functions. *Numerische Mathematik* **24,** 383–393.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74,** 646–648.

Whittaker, E. T. (1922). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society* **41,** 63–75.

Yue, Y. R., Simpson, D., Lindgren, F., and Rue, H. (2014). Bayesian adaptive smoothing splines using stochastic differential equations. *Bayesian Analysis* **9,** 397–424.

Yue, Y. R., Speckman, P. L., and Sun, D. (2012). Priors for Bayesian adaptive spline smoothing. *Annals of the Institute of Statistical Mathematics* **64,** 577–613.

Zhu, B. and Dunson, D. B. (2013). Locally adaptive Bayes nonparametric regression via nested Gaussian processes. *Journal of the American Statistical Association* **108,** 1445–1456.

A Approximate Horseshoe Density

There is no exact closed-form expression available for the horseshoe density function. We present an approximation to the horseshoe density that can be used without the need for explicit specification of the nuisance local scale parameters. Following Carvalho et al. (2010), the marginal distribution of u given global scale parameter γ is found by integrating over possible values of the local scale parameter τ, where $u \,|\, \tau \sim \mathrm{N}(0, \tau^2)$ and $\tau \,|\, \gamma \sim \mathrm{C}^+(0, \gamma)$. This leads to

$$
\begin{aligned}
p(u \,|\, \gamma) &= \int_0^\infty p(u \,|\, \tau, \gamma) p(\tau \,|\, \gamma) \mathrm{d}\tau \\
&= \int_0^\infty \frac{1}{(2\pi)^{1/2}\tau} \exp\left(-\frac{u^2}{2\tau^2}\right) \frac{2\gamma}{\pi(\tau^2 + \gamma^2)} \mathrm{d}\tau \\
&= \left(\frac{1}{2\pi^3\gamma^2}\right)^{1/2} \exp\left(\frac{u^2}{2\gamma^2}\right) \mathrm{E}_1\left(\frac{u^2}{2\gamma^2}\right),
\end{aligned}
$$

where E_1 is the exponential integral function. Note that $\lim_{x\to 0^+} \mathrm{E}_1(x) = \infty$, but for $x > 0$, the function $\mathrm{E}_1(x)$ is bounded as follows:

$$
\frac{1}{2} e^{-x} \ln\left(1 + \frac{2}{x}\right) < \mathrm{E}_1(x) < e^{-x} \ln\left(1 + \frac{1}{x}\right).
$$

Then for $u \in \{\mathbb{R} : u \neq 0\}$ we have

$$
\frac{1}{2} \exp\left(\frac{-u^2}{2\gamma^2}\right) \ln\left(1 + \frac{4\gamma^2}{u^2}\right) < \mathrm{E}_1\left(\frac{-u^2}{2\gamma^2}\right) < \exp\left(\frac{-u^2}{2\gamma^2}\right) \ln\left(1 + \frac{2\gamma^2}{u^2}\right).
$$

It follows that the target density is bounded by

$$
\frac{1}{2}\left(\frac{1}{2\pi^3\gamma^2}\right)^{1/2} \ln\left(1 + \frac{4\gamma^2}{u^2}\right) < p(u \,|\, \gamma) < \left(\frac{1}{2\pi^3\gamma^2}\right)^{1/2} \ln\left(1 + \frac{2\gamma^2}{u^2}\right). \tag{A.1}
$$

Let the left bound in equation (A.1) be denoted $B_1(u)$ and the right bound $B_2(u)$. Note that as $u \to 0$, each of $B_1(u)$, $p(u \,|\, \gamma)$ and $B_2(u)$ approach ∞. It can be shown that $\int_{-\infty}^{\infty} B_1(u)\mathrm{d}u = \sqrt{2/\pi}$ and $\int_{-\infty}^{\infty} B_2(u)\mathrm{d}u = 2/\sqrt{\pi}$. Since $\sqrt{2/\pi} < 1 < 2/\sqrt{\pi}$, these bounds can be used to find an approximate expression for $p(u \,|\, \gamma)$ that integrates to 1 and still satisfies equation (A.1). We set

$$
\tilde{p}(u \,|\, \gamma) = wB_1(u) + (1 - w)B_2(u) \tag{A.2}
$$

with constraints $0 < w < 1$ and $\int_{-\infty}^{\infty} wB_1(u) + (1 - w)B_2(u)\mathrm{d}u = 1$. Using the values for the integrated bounds and solving gives $w = (\sqrt{\pi} - 2)/(\sqrt{2} - 2)$. Substituting this value for w into equation (A.2) and simplifying gives the following closed-form approximation to the horseshoe density function:

$$
\tilde{p}(u \,|\, \gamma) = \left(\frac{1}{2\pi^3\gamma^2}\right)^{1/2} \left[\frac{\sqrt{\pi} - 2}{2\sqrt{2} - 4} \ln\left(1 + \frac{4\gamma^2}{u^2}\right) + \frac{\sqrt{2} - \sqrt{\pi}}{\sqrt{2} - 2} \ln\left(1 + \frac{2\gamma^2}{u^2}\right)\right]. \tag{A.3}
$$

B Additional Simulation Results

Here we display plots with simulation results for normal data with $\sigma = 1.5$ (Figure B.1), Poisson data (Figure B.2), and binomial data (Figure B.3). Summary measures for all data types show similar patterns to each other and to those for normal data with $\sigma = 4.5$ (Figure 2).

Table B.1: Mean values of performance measures across 100 simulations for normal observations ($\sigma = 1.5$) for each model and trend function type.

Function	Model	SRE	MRW	Variation	True Var.
Constant	Normal	0.57	0.03	0.16	0.00
	Laplace	0.58	0.04	0.18	0.00
	Horseshoe	0.63	0.05	0.37	0.00
Piecewise Const.	Normal	6.32	0.33	162.36	60.00
	Laplace	5.47	0.20	154.01	60.00
	Horseshoe	1.69	0.12	63.17	60.00
Smooth	Normal	3.58	0.18	137.75	139.21
	Laplace	3.57	0.19	137.90	139.21
	Horseshoe	3.64	0.18	139.95	139.21
Varying Smooth	Normal	7.88	0.36	42.05	53.79
	Laplace	7.70	0.35	42.61	53.79
	Horseshoe	6.19	0.47	46.42	53.79

Table B.2: Mean values of performance measures across 100 simulations for Poisson observations for each model and trend function type.

Function	Model	SRE	MRW	Variation	True Var.
Constant	Normal	2.17	0.14	1.67	0.00
	Laplace	2.26	0.15	1.94	0.00
	Horseshoe	2.47	0.17	2.20	0.00
Piecewise Const.	Normal	11.11	0.57	153.95	60.00
	Laplace	9.28	0.54	126.32	60.00
	Horseshoe	5.09	0.34	70.57	60.00
Smooth	Normal	8.90	0.40	132.64	139.21
	Laplace	8.82	0.40	132.30	139.21
	Horseshoe	8.66	0.40	131.86	139.21
Varying Smooth	Normal	6.70	0.30	44.28	53.79
	Laplace	6.56	0.30	44.64	53.79
	Horseshoe	5.85	0.28	45.47	53.79

Table B.3: Mean values of performance measures across 100 simulations for binomial observations for each model and trend function type.

Function	Model	SRE	MRW	Variation	True Var.
Constant	Normal	2.09	0.12	0.03	0.00
	Laplace	2.14	0.13	0.03	0.00
	Horseshoe	2.36	0.16	0.04	0.00
Piecewise Const.	Normal	10.78	0.55	3.51	1.40
	Laplace	9.04	0.51	2.90	1.40
	Horseshoe	4.93	0.31	1.61	1.40
Smooth	Normal	5.59	0.29	2.42	2.60
	Laplace	5.59	0.29	2.42	2.60
	Horseshoe	5.76	0.29	2.45	2.60
Varying Smooth	Normal	7.09	0.32	1.10	1.20
	Laplace	6.93	0.32	1.10	1.20
	Horseshoe	5.74	0.28	1.09	1.20

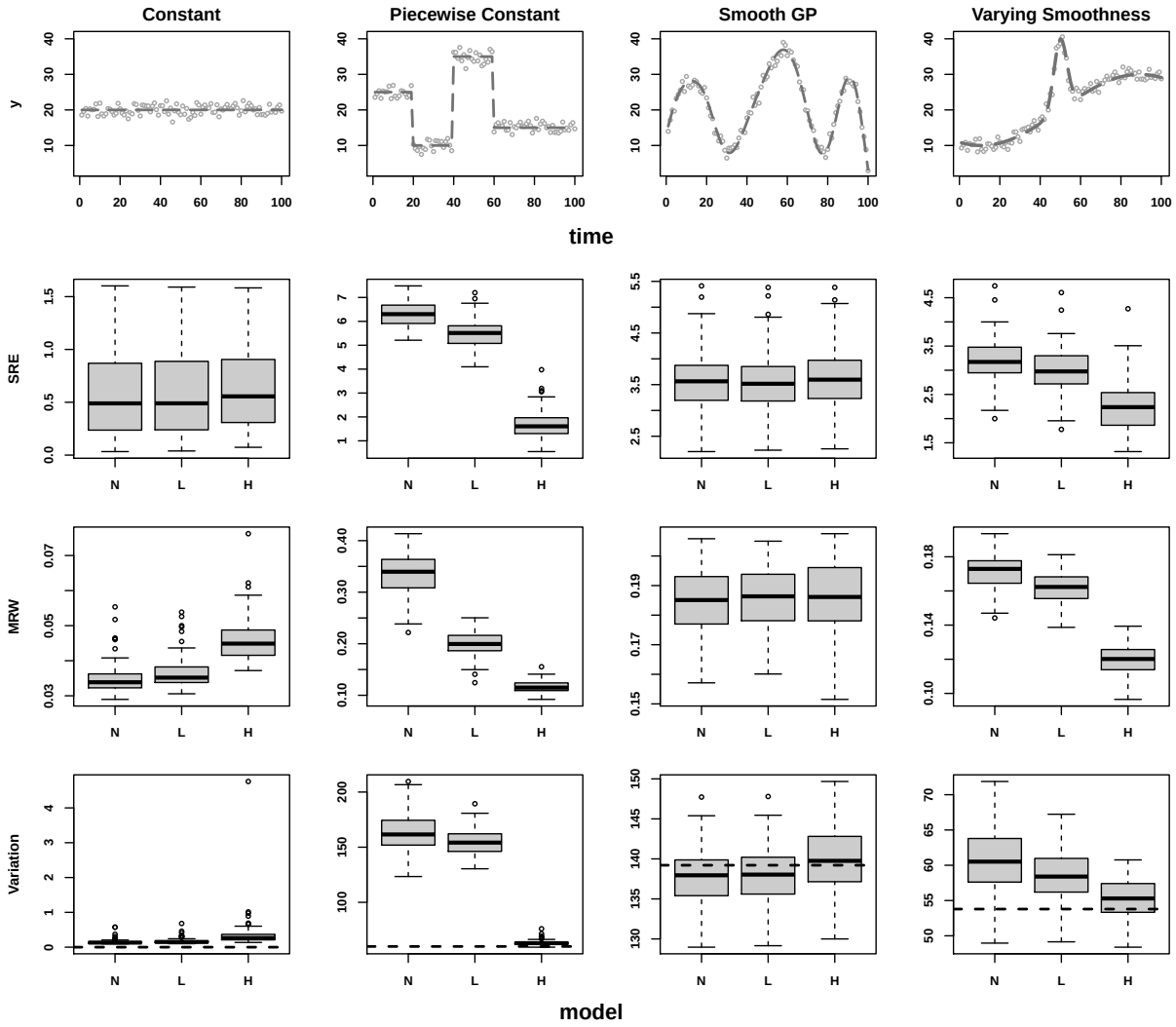

Figure B.1: Functions used in simulations and simulation results by model and function type for normally distributed data with $\sigma = 1.5$. Top row shows true functions (dashed lines) with example simulated data. Remaining rows show sum of relative errors (SRE), mean relative width (MRW), and sample variation. Dashed line in plots on bottom row is the true function variation.

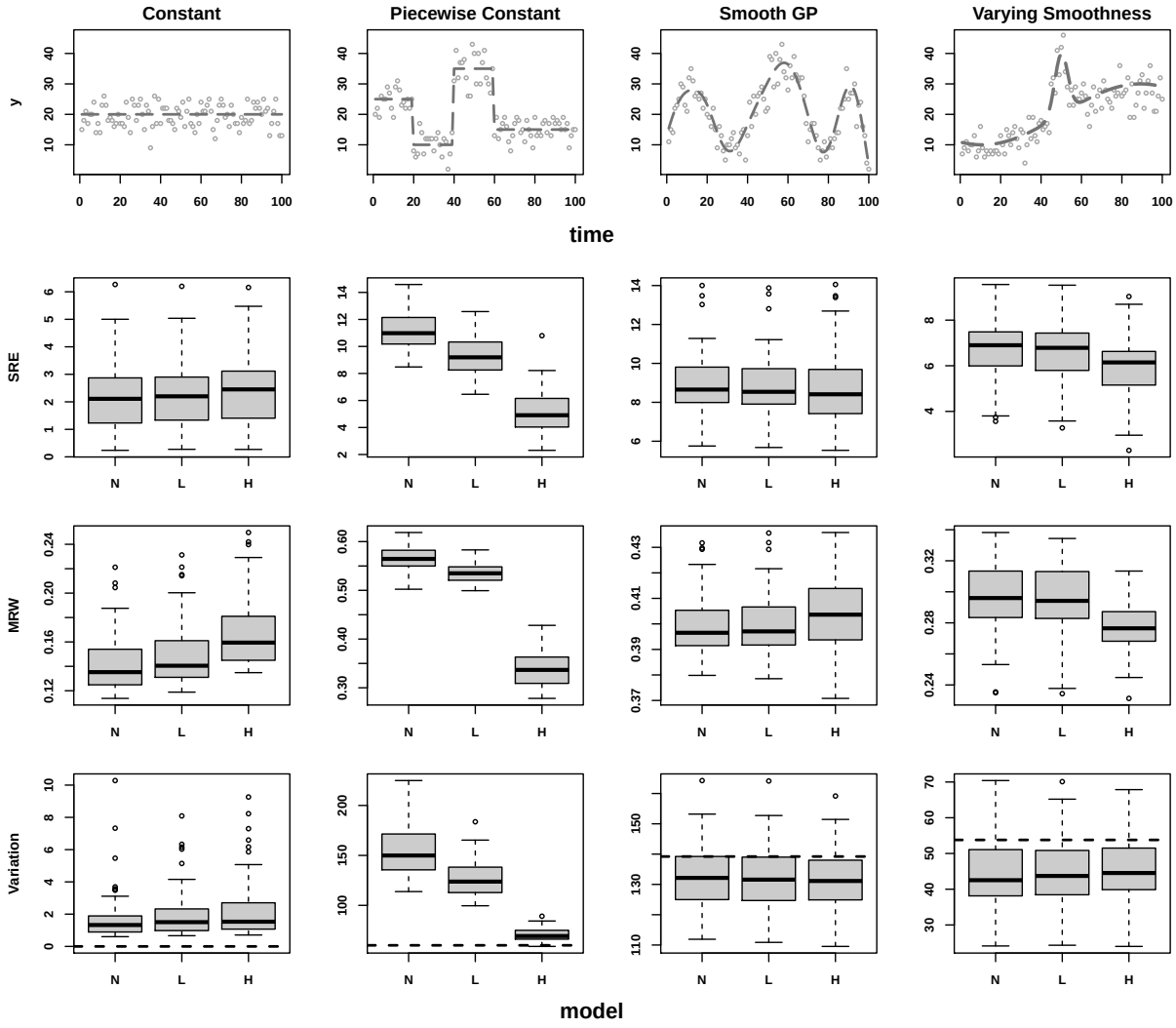

Figure B.2: Functions used in simulations and simulation results by model and function type for Poisson distributed data. Top row shows true functions (dashed lines) with example simulated data. Remaining rows show sum of relative errors (SRE), mean relative width (MRW), and sample variation. Dashed line in plots on bottome row is the true function variation.

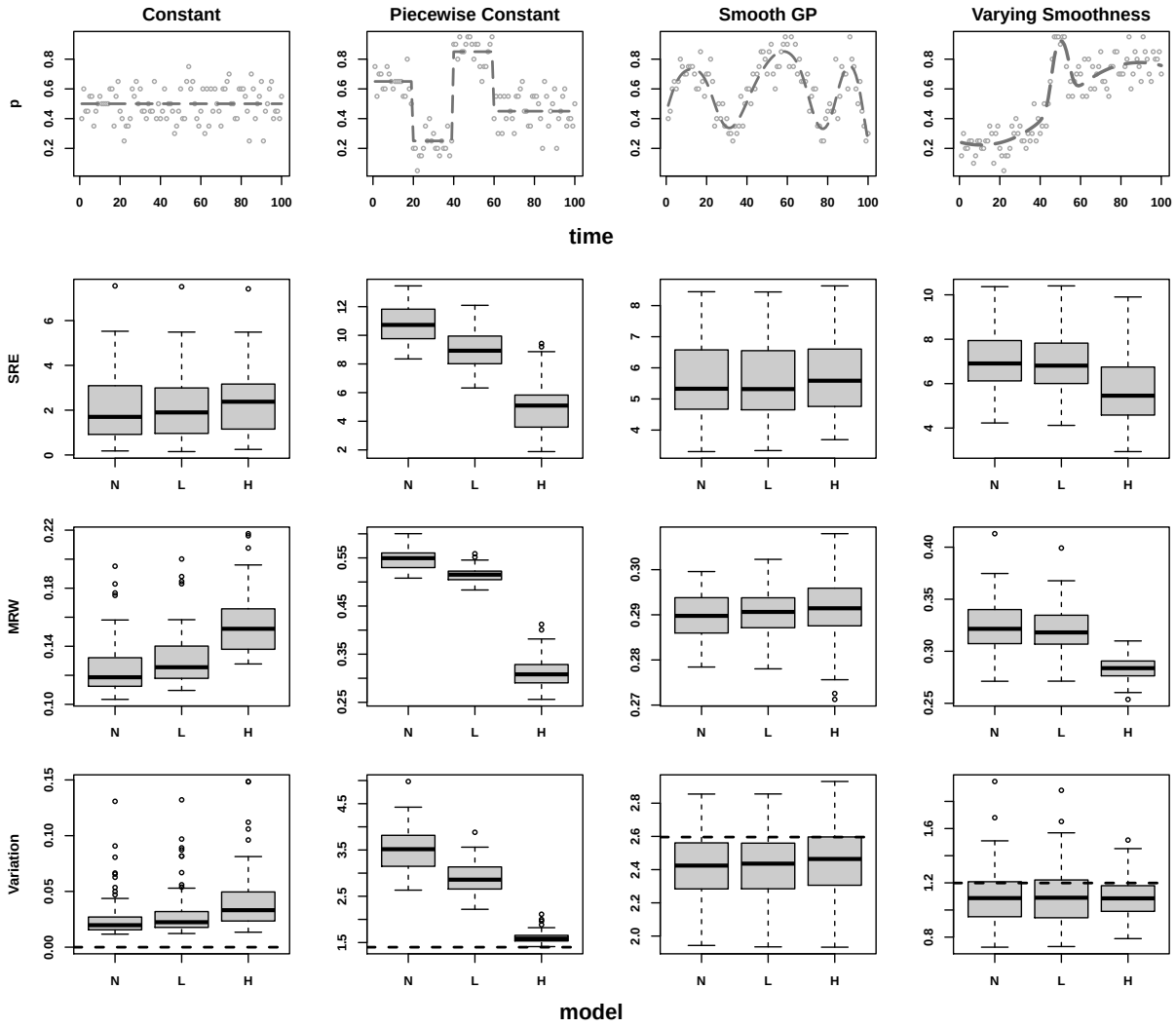

Figure B.3: Functions used in simulations and simulation results by model and function type for binomial distributed data. Top row shows true functions (dashed lines) with probability estimates from example simulated data. Remaining rows show sum of relative errors (SRE), mean relative width (MRW), and sample variation. Dashed line in plots on bottome row is the true function variation.

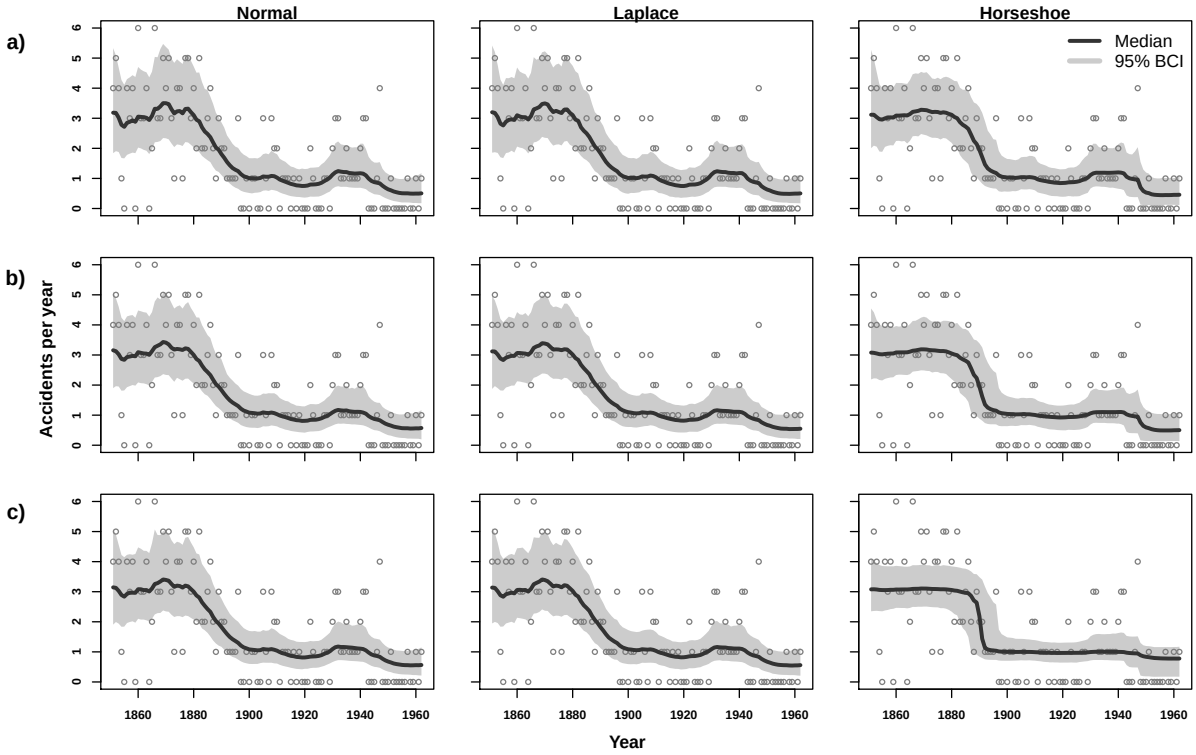

Figure C.1: Models fits to coal mining accidents data by model type and value of hyperparameter for global scale: a) $\zeta = 1$, b) $\zeta = 0.01$, and c) $\zeta = 0.0001$. Posterior medians and associated 95% Bayesian credible intervals are shown along with observed data.

C Prior Sensitivity

We tested the sensitivity of the three prior formulations (normal, Laplace, and horseshoe) to the value of the hyperparameter (ζ) which controls the scale of the distribution on the smoothing parameter γ, where $\gamma \sim \mathrm{C}^{+}(0, \zeta)$ A smaller value of ζ constricts γ to be closer to zero, which in turn constricts the scales of the priors on the order-k differences. We tested three levels for the hyperparameter: a) $\zeta = 1$, b) $\zeta = 0.01$, and c) $\zeta = 0.0001$. In general, we expect noisier data sets should be more sensitive to prior settings. The coal mine disaster data offered a good test set because the observations are relatively noisy.

Clearly the horseshoe prior was the most sensitive to the level of ζ (Figure C.1). In particular, the horseshoe results for $\zeta = 1$ looked more like those for the other two models in Figure 4, but when $\zeta = 0.0001$, the horseshoe produced more defined break points and straighter lines with narrower BCIs compared to the results with $\zeta = 0.01$.