Bayesian Change Points and Linear Filtering in Dynamic Linear Models using Shrinkage Priors

Jeffrey B. Arnold

March 22, 2016

Political and social processes are rarely, if ever, constant over time, and, as such, social scientists have a need to model that change (Büthe 2002; Lieberman 2002). Moreover, these political process are often marked by both periods of stability and periods of large changes (Pierson 2004). If the researcher has a strong prior about or wishes to test a specific location of a change, they may include indicator variables, an example in international relations is the ubiquitous Cold War dummy variable. Other approaches estimate locations of these change using change point or structural break models (Calderia and Zorn 1998; Western and Kleykamp 2004; Spirling 2007b; Spirling 2007a; Park 2010; Park 2011; Blackwell 2012).

This offers a different Bayesian approach to modeling change points. I combine a continuous latent state space approach, e.g. dynamic linear models, with recent advances in Bayesian shrinkage priors (Carvalho, Polson, and Scott 2009; Carvalho, Polson, and Scott 2010; Polson and Scott 2010). These sparse shrinkage priors are the Bayesian analog to regularization and penalized likelihood approaches such as the LASSO (Tibshirani 1996) in maximum likelihood. An example of such an approach, is a model of change points in the mean of normally distributed observations,

$$
\begin{aligned}
y_t &= \mu_t + \epsilon_t \qquad \epsilon \text{ iid}, \mathrm{E}(\epsilon) = 0 \\
\mu_t &= \mu_{t-1} + \omega_t
\end{aligned} \tag{1}
$$

Since this is a change point model, the change in the mean, ω_t, should be sparse, with most values at or near zero, and a few which can be large. To achieve this estimate, I model ω with a shrinkage prior distribution that has most of its mass concentrated near zero to shrink values of ω_t to zero, but has wide tails which will not shrink the non-zero ω_t at the change points. The estimated posterior distribution of the μ will resemble a step function, with the steps being the estimated change points. The particular shrinkage priors that will be used in this work are the horseshoe (Carvalho, Polson, and Scott 2009; Carvalho, Polson, and Scott 2010) and horseshoe+ distributions (Bhadra et al. 2015).

Modeling change points using sparse shrinkage prior distributions has several advantageous features. First, it does not require specifying the number of change points *ex ante*. The sparsity of the parameter changes can be estimated from the data. Second, although this method will not directly provide a posterior distribution of the locations of the change points, it is likely more representative of the data generating processes commonly observed in the social science domain. Traditional change point models have a data generating process in which there are periods of exactly no change, with a few periods of any change. But many social science processes are more akin parameter that is always always changing, but which changes by relatively small amounts in most periods, but changes by very large amounts in a few periods. Third, this method is flexible and extensible in that it can be adapted to a variety of models. This work considers the cases of change points in the level and both the level and trend of a single parameter. But, it can also be applied to linear regressions with time varying parameters, changes in seasonality, variance, and a variety of other models.

Fourth, this method is computationally efficient. Most shrinkage distributions, and all those used in this work, are representable as scale-mixtures of normal distributions. This allows the model to be expressed as a Gaussian dynamic linear models (GDLMs) (West and Harrison 1997). GDLMs are a class of models than incorporates many common time series models, including ARIMA, structural time series, and linear regression with time-varying parameters. and use specialized methods for sampling from GDLMs, *e.g.* the forward-filter backwards-sampling (FFBS) algorithm (Carter and Kohn 1994)Fruehwirth-Schnatter1994.

Fifth, this method is implementable and efficiently implemented in a popular Bayesian probabilistic programming language, Stan. Since Stan does not directly sample discrete variables, standard Bayesian change point methods based on a discrete state space such as (Chib 1998), are either difficult, requiring marginalizing over the discrete states, or impossible. Since in this approach all parameters are continuous, it can be directly implemented in Stan. However, since in many cases, the change point problem can be represented more efficient methods specific to GDLMs can be used. A complementary contribution of this work is that it provides a method to efficiently estimate Gaussian dynamic linear models (GDLMs) in Stan. These had previously been difficult to sample in general purpose Bayesian programming languages, such as JAGS (Jackman 2009, p. 477). This work provides a complete set of functions to perform Kalman filtering, smoothing, and backward sampling within Stan. This allows for efficiently estimating GDLMs in Stan using a partially collapsed Gibbs sampler in which Stan's standard algorithms are used to estimate parameters after marginalizing over the latent states, and the latent states of the GDLM are sampled using FFBS.

This work presents two examples of this approach to change points. The first calculates change points in the level of a time series, using the example of the annual flow of the Nile River, 1870-1970. The second calculates change points in both the level and trend of a time series, using the example of approval ratings for President George W. Bush.

1 Change points as a Variable Selection and Shrinkage Problem

For simplicity, I start with a model of change points in the level of a time-series, and later generalize to other cases. In this case, there are n ordered observations, $y_1, \dots, y_n$, with a time-varying mean, μ_t:

$$y_t = \mu_t + \epsilon_t \quad \epsilon_t \text{ are iid with } \mathrm{E}(\epsilon) = 0. \tag{2}$$

Suppose that there are M change points, with ordered change-point locations, $\tau_1, \dots, \tau_M$, and the convention that $\tau_0 = 0$ and $\tau_{M+1} = n$. This splits the mean into M segments, with values of the mean $\mu^*_1, \dots, \mu^*_M$, such that

$$\mu_t = \mu^*_m \quad \text{if } \tau_m \leq t < \tau_{m+1} \tag{3}$$

There are a variety of approaches to the change point problem in both classical (frequentist) (Page 1954; Hinkley 1970; Bai and Perron 2003; Olshen et al. 2004; Bai and Perron 1998; Killick, Fearnhead, and Eckley 2012) and Bayesian statistics (Yao 1984; Barry and Hartigan 1993; Chib 1998; Fearnhead 2006; Fearnhead and Liu 2007).

An alternative approach to the change point problem is to rewrite the problem in Equation (3) to focus on the changes in the mean (system errors), $\omega_t = \mu_t - \mu_{t-1}$, rather than the locations of the change points. In this case, replace Equation (3) with

$$\mu_t = \mu_{t-1} + \omega_t \tag{4}$$

In a change point model, the system errors, ω_t, are sparse, meaning that most values of ω_t are zero, and only a few are non-zero. In this formulation, the times of the change points are not directly estimated. Instead, the change points are those times at which ω_t is non-zero. This formulation turns the change-point problem from one of segmentation, to one of variable selection, in which the goal is to find the non-zero values of ω_t and estimate their values. The natural Bayesian approach to this variable selection problem is to explicitly model the values of ω as a discrete mixture between a point mass at zero for the non-change points, and an alternative distribution (Mitchell and Beauchamp 1988; Efron 2008):

$$\omega_t = \rho g(\omega) + (1 - \rho)\delta_0 \tag{5}$$

where g is a distribution of the non-zero ω, and δ_0 is a Dirac delta distribution (point mass) at 0. Equation (5) is a so-called “spike and slab” prior (Mitchell and Beauchamp 1988). Since Equation (5) explicitly models the two groups of zero and non-zero parameters, Efron (2008) calls this the two-group answer to the two-group problem. Spike and slab priors are convenient because they directly provide a posterior probability that a parameter is non-zero, or in this case that a time is a change point. Giordani and Kohn (2008) propose using the representation in Equation (4) with the a discrete mixture distribution for ω, as in Equation (5), to estimate change points.

Recent work in Bayesian computation has focused on one-group solutions to the variable selection problem. These combine shrinkage and variable selection through the use of continuous distributions with large spike at zero to shrink parameters towards zero and wide tails to avoid shrinking the non-zero parameters (Polson and Scott 2010). Numerous sparse sparse shrinkage priors have been proposed, but this paper will consider the Student’s *t* (Tipping 2001), the Laplace or double exponential (Park and Casella 2008)Hans2009, the horseshoe (Carvalho, Polson, and Scott 2010), and horseshoe+ distributions (Bhadra et al. 2015).

Student’s *t*: The Student’s *t* distribution for $x \in \mathbb{R}$ with scale $\tau \in \mathbb{R}^+$ and degrees of freedom $\nu \in \mathbb{R}^+$,

$$p(\omega_t|\tau, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\omega_t^2}{\nu}\right) \tag{6}$$

The Student’s *t* distribution can also be expressed as a scale mixture of normals,[1]

$$\begin{aligned} \omega_t|\tau, \nu, \lambda_t &\sim \mathcal{N}\left(0, \tau^2 \lambda_t^2\right) \\ \lambda_t^2|\nu &\sim \mathcal{IG}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \end{aligned} \tag{8}$$

(Tipping 2001) uses the Student’s *t* for sparse shrinkage by letting the degrees of freedom $\nu \to 0$.

Laplace: The Laplace (double exponential) distribution with scale τ,

$$p(\omega_t|\tau) = \frac{1}{2\tau} \exp\left(-\frac{|\omega_t|}{\tau}\right) \tag{9}$$

The Laplace distribution can also be expressed as a scale-mixture of normal distributions,[2]

$$\begin{aligned} \omega_t|\tau, \lambda_t &\sim \mathcal{N}\left(0, \tau^2 \lambda_t^2\right) \\ \lambda_t^2 &\sim \mathcal{E}\left(\frac{1}{2}\right) \end{aligned} \tag{11}$$

The Laplace distribution is the distribution that corresponds to the ℓ_1 penalty used in the LASSO estimator (Park and Casella 2008; Hans 2009). However, although an ℓ_1 penalty is able to produce sparse estimates in a maximum likelihood framework because, it does not produce sparse posterior means in Bayesian estimation (Park and Casella 2008). Another problem with using the Laplace distribution as a shrinkage prior is that its tails are narrow, so it tends to excessively shrink large signals (Carvalho, Polson, and Scott 2010).

[1] $\mathcal{IG}(x|\alpha, \beta)$ is an inverse-gamma distribution for $x \in \mathbb{R}^+$ with shape $\alpha \in \mathbb{R}^+$ and inverse-scale $\beta \in \mathbb{R}^+$,

$$\mathcal{IG}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha-1)} \exp\left(\beta \frac{1}{x}\right) \tag{7}$$

[2] $\mathcal{E}(x|\beta)$ is the exponential distribution for $x \in \mathbb{R}^+$ with inverse-scale (rate) parameter $\beta \in \mathbb{R}^+$,

$$\mathcal{E}(x|\beta) = \beta \exp(-\beta x) \tag{10}$$

Horseshoe: The horseshoe distribution (Carvalho, Polson, and Scott 2009; Carvalho, Polson, and Scott 2010) does not have an analytical form, but is defined hierarchically as a scale-mixture of normals

$$
\begin{aligned}
\omega_t | \lambda_t, \tau &\sim \mathcal{N}\left(0, \tau^2 \lambda_t^2\right) \qquad (12) \\
\lambda_t &\sim \mathcal{C}^{+}(0, 1)
\end{aligned}
$$

where $\mathcal{C}^{+}(x|0, s)$ denotes half-Cauchy distribution with a scale parameter s, and density

$$
p(x|s) = \frac{2}{\pi x \left(1 + \left(\frac{x}{s}\right)^2\right)} \qquad (13)
$$

The Horseshoe distribution has some theoretically attractive properties for shrinkage and variable selection (Carvalho, Polson, and Scott 2009; Carvalho, Polson, and Scott 2010; Datta and Ghosh 2012; Pas, Kleijn, and Vaart 2014).

Horseshoe+ The horseshoe+ distribution Bhadra et al. (2015) is similar to the Horseshoe distribution, but with an additional hyper-prior on λ_t,

$$
\begin{aligned}
\omega_t | \lambda_t, \eta_t, \tau &\sim \mathcal{N}\left(0, \tau^2 \lambda_t^2\right) \qquad (14) \\
\lambda_t &\sim \mathcal{C}^{+}\left(0, \eta_t\right) \\
\eta_t &\sim \mathcal{C}^{+}(0, 1)
\end{aligned}
$$

The shrinkage distributions considered here are global-local scale mixtures of normal distributions. These distributions contain a global variance component, τ, and local variance components, λ_t (Polson and Scott 2010). The global variance component, τ, concentrates the prior distribution around zero, while the local variance components, λ_t, allow individual parameters to be large without shrinking them towards zero. The choice of the prior distribution for the global variance component, τ, is particularly important in these shrinkage distributions as it effectively controls the sparsity of the estimates, which in this application is the number of change points. As per the suggestion in Bhadra et al. (2015) and Pas, Kleijn, and Vaart (2014), I use the prior distribution

$$
\tau \sim \mathcal{C}^{+}\left(0, \frac{1}{n}\right) \qquad (15)
$$

where n is the number of observations. [3] That these are distributions can all be expressed as normal distributions, conditional the values of λ_t and τ, is useful computationally. This property means that many change point problems can be expressed as Gaussian dynamic linear models, and can make use of computationally efficient methods as discussed in Section 2.

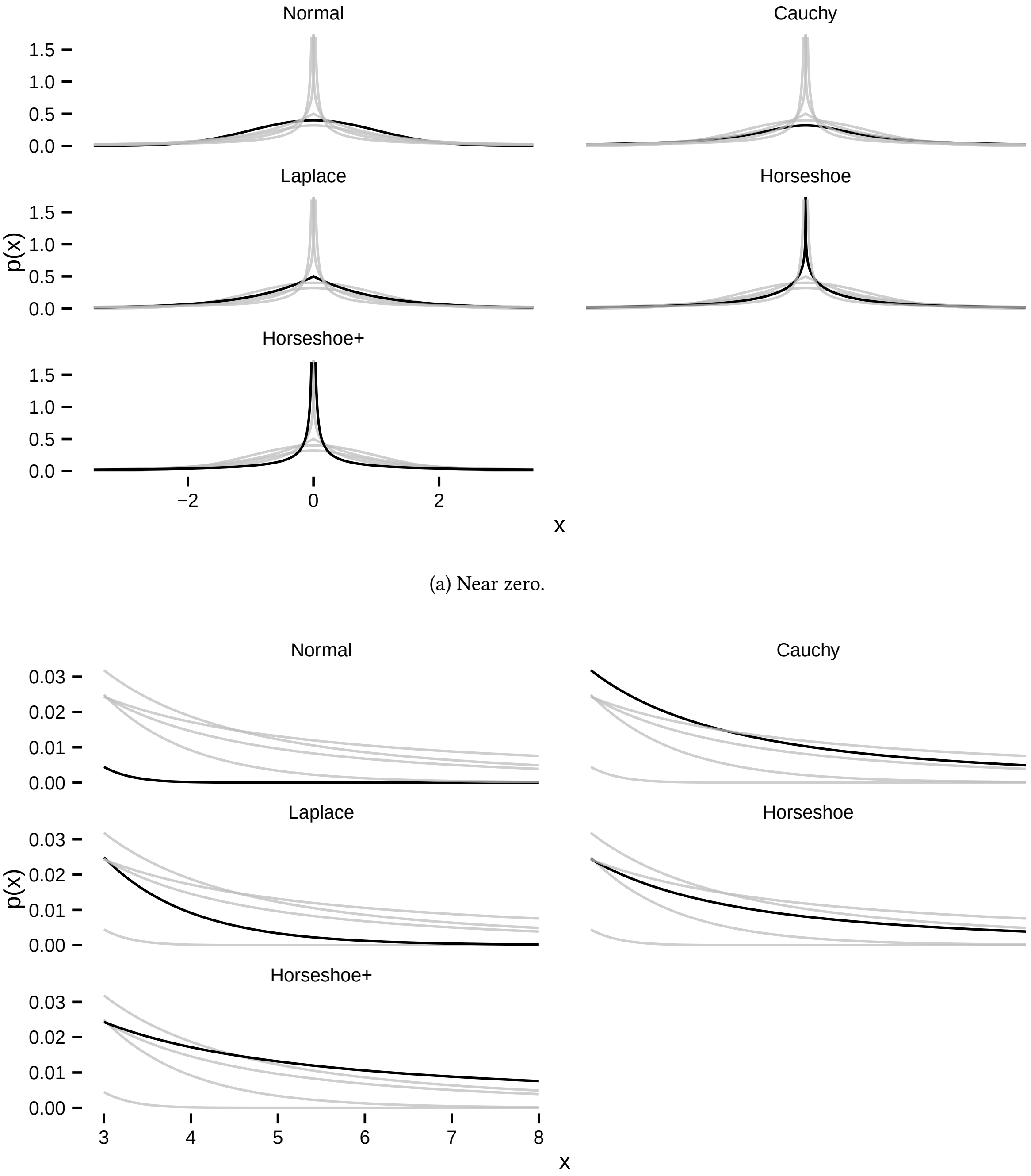
These Bayesian sparse shrinkage priors are analogous to the sparse regularization and penalized likelihood approaches in maximum likelihood, of which the LASSO estimator (Tibshirani 1996) and its the numerous variations are the most prominent and popular examples. Several papers have proposed using LASSO-like penalties and maximum likelihood to estimate change-points (Tibshirani et al. 2005; Harchaoui and Lévy-Leduc 2010; Chan, Yau, and Zhang 2014). This is a Bayesian extension of that approach.

To summarize, the proposed model for change points in the level of a time-series is,

$$
\begin{aligned}
y_t &\sim \epsilon_t & \epsilon_t \text{ iid}, \mathrm{E}(\epsilon) = 0, \mathrm{Var}(\epsilon) = \sigma^2 \qquad (16) \\
\mu_t &= \mu_{t-1} + \sigma \omega_t & \qquad (17)
\end{aligned}
$$

where ω_t is given a shrinkage prior distribution that induces sparsity. The values of ω_t is multiplied by the observation variance in order to avoid multi-modal posterior distributions (Polson and Scott 2010, p. 8). I propose using the horseshoe and horseshoe+ distributions as those shrinkage priors, and compare them the Student's t and Laplace distributions.

[3] Other suggestions include $\tau \sim \mathcal{C}^{+}(0, 1)$, $\tau \sim \mathcal{U}(0, 1)$ where $\mathcal{U}()$ is the uniform distribution, and a plug-in value of p/n, where p is the expected number of non-zero parameters. See Polson and Scott (2012), Pas, Kleijn, and Vaart (2014), and Bhadra et al. (2015).



(a) Near zero.

(b) Tail region.

Figure 1: Comparison of the density functions of normal, Cauchy, Laplace, horseshoe, and horseshoe+ distributions. All functions have location and scale parameters of 0. The Cauchy distribution is a special case of the Student's *t* distribution with degrees of freedom equal to 1.

The method proposed here is able to estimate the values of μ even when it is subject to large jumps or follows a step-function. However, unlike methods using discrete state space models, it does not directly a provide probability that a given location is a change point. Instead, the researcher can identify "change points" from the magnitudes of the posterior distribution of ω. When the posterior distribution of ω_t is far from zero, it is change point, and when it includes or is close to zero, it is not. This is similar to the auxiliary residual test of de Jong and Penzer (1998). For many practical purposes, visual inspection by the researcher of a plot of the posterior estimates of μ should be sufficient. As noted earlier, for many data generating processes, the hypothesis that $\omega_t = 0$ is implausible, so testing it is nonsensical. Rather, change point models are used to ease interpretation, so an informal visual method should suffice for interpretation. If the change in μ is not distinguishable in a plot, then it is unlikely to be of much substantive importance.

While this one-group approach does not provide clear posterior probabilities of the locations of change points, it is a reasonable model of change point processes for several reasons.[4] First, for many time-varying parameter processes typically modeled with a change-point model in the social sciences, the data generating process is probably more similar to one in which the parameter is changing in all periods, but most of those periods the changes are small relative to the magnitudes of changes in a few periods. In other words, the system errors, ω, are never zero, but they are relatively small in most periods. This seems plausible for many processes modeled by political scientists in which there is no physical reason to think there are actually discrete states. Thus the researcher is modeling the parameter to take on a few values for parsimony and interpretation, and not primarily because it matches the data generating process. Most of the data generating processes considered by political science papers using change points fall into this category: Supreme court dissent and consensus (Calderia and Zorn 1998), wage growth in OECD states (Western and Kleykamp 2004), casualties in the Iraq War (Spirling 2007b), presidential use of force (Park 2010), and campaign contributions (Blackwell 2012). Second, even in change point models where the model had discrete states, after marginalizing over the posterior distribution of the discrete states, the posterior distribution of the change in μ will never be exactly zero. The shrinkage prior effectively is approximating that posterior distribution of each μ_t after marginalizing over those states.

2 Estimation and Implementation in Stan

The model in the previous section is an example of a Gaussian dynamic linear model (GDLM), also known as linear Gaussian state space models.[5] This this model can be expressed as a GDLM is useful because there are efficient algorithms to calculate the likelihood and sample from these models, and since GDLMs are a particularly flexible class of models, it provides ways to generalize it. A GDLM is represented system of equations (Durbin and Koopman 2012; West and Harrison 1997; Petris, Petrone, and Campagnoli 2009; Shumway and Stoffer 2010, Ch 6),

$$y_t \sim \mathcal{N}\left(b_t + F_t \theta_t, V_t\right) \tag{18}$$

$$\theta_t \sim \mathcal{N}\left(g_t + G_t \theta_{t-1}, W_t\right) \tag{19}$$

In these equations, the observed data, y_t, is a linear function of the latent states, θ_t, which are a function of their previous values, θ_{t-1}. Equation (18) is the *observation equation*, where y_t (observation vector) is an $r \times 1$ vector of observed data, θ_t (state equation) is a $p \times 1$ vector of the latent states, b_t is an $r \times 1$ vector, F_t is a $r \times p$ matrix, and V_t (observation variance) is an $r \times r$ covariance matrix. Equation (19) is the *state equation* equation, which relates the current latent states to their previous values; g_t is a $p \times 1$ vector, G_t is a $p \times p$ matrix, and W_t (state variance) is an $p \times p$ covariance matrix. The vectors and matrices, $\Phi = \{b_t, g_t, F_t, G_t, V_t, W_t\}_{t \in 1:n}$, are *system matrices*. In applications, the system matrices, will often be functions

[4] These points are similar to those made by Polson and Scott (2012, pp. 2-3) with regard to the use of shrinkage priors in variable selection.

[5] See Beck (1989) and Martin and Quinn (2002) for examples of GDLMs in political science.

of parameters. GDLMs are a general class of models which includes many common time series models, including SARIMA, structural time series (Harvey 1990), dynamic factors, seemingly unrelated regression, and linear regression with time varying coefficients, among others (Durbin and Koopman 2012, Ch. 3).

The model in Equations (2) and (4) is a GDLM if $\epsilon_t \sim \mathcal{N}(0, V)$ and $\omega_t \sim \mathcal{N}\left(0, W_t\right)$. In this case, $\beta_t = g_t = 0$, $F_t = G_t = 1$. Even though the proposed distributions for ω_t are not normal, since they are all scale-mixtures of normal distributions, they are normal conditional on the values of λ_t.These models are similar to the *local level model* (Durbin and Koopman 2012, Ch 2.), except that they have time-varying state variances, $W_t = \tau^2 \lambda_t^2$. In other words, the change point model discussed here is simply a local level model with a sparse shrinkage prior on the state variance.

That these change point models can be represented as GDLMs is useful for two reasons. First, it suggests how these models can be extended beyond the simple change in level model in Section 1. Any sort of GDLM, which, as noted, includes many common models, can be adjusted to account for change points in in any of its states, by using a shrinkage prior distribution for its system variance, as long as the prior distribution is a scale-mixture of normals.[6]

Second, since in a GDLM both the observation and system equations are multivariate normal, there are analytical solutions that allow for efficiently computing its likelihood and sampling the latent states from their posterior distributions. The Kalman filter calculates the values of $p(\theta_t | y_{1:(t-1)})$ and $p(\theta_t | y_{1:t})$, and can be used to calculate the likelihood of $p(y|\Phi)$. Importantly, the Kalman filter can calculate the likelihood without needing the values of the latent states, θ. The derivations of the Kalman filter can be found in most time-series texts, including Durbin and Koopman (2012, Ch. 5–7) and West and Harrison (1997), and thus are not presented here. Given the results of the Kalman filter there are several methods to sample θ from $p(\theta | y, \Phi)$, a process called Forward-Filtering Backwards-Smoothing (FFBS) or simulation smoothing (Carter and Kohn 1994; Frühwirth-Schnatter 1994; De Jong and Shephard 1995; Durbin and Koopman 2002; Durbin and Koopman 2012, Ch 4.9). I make use of these methods to efficiently sample both the latent states, θ, and other parameters of the model in Stan.

Stan is a probabilistic programming language, with a BUGS-like modeling language, and interfaces to several programming languages, including R (Stan Development Team 2015; Carpenter et al. 2015). GDLMs can be directly estimated in Stan by translating the model described by Equations (18) and (19) into a Stan model. However, GDLMs can be estimated more efficiently in Stan by marginalizing over the latent states, θ. The sampling methods implemented in Stan, of which the default is HMC-NUTS (Hoffman and Gelman 2014), only require the calculation of a likelihood from the user.[7] The Kalman filter can be used to calculate the likelihood $p(y|\Phi)$, marginalizing out the latent states, θ_t. Marginalizing out parameters is required when estimating models with discrete parameters, such as mixture models, in Stan (Stan Development Team 2015, p. 104). Although it is not necessary to marginalize over parameters to sample from GDLMs in Stan, it helps the efficiency of sampling by reducing the correlation between the latent states and the other parameters of the model.The latent states can then be sampled using FFBS. [8] To summarize, an efficient method to sample GDLMs in Stan is:

1. Sample ϑ from $p(\vartheta | y)$ using HMC in Stan. This requires integrating out the latent states, θ, and calculating $p(y|\vartheta)$, which is done using a Kalman filter.

2. Sample $p(\theta | y, \vartheta)$ using a simulation smoother for a GDLM as in (Carter and Kohn 1994; Frühwirth-Schnatter 1994; De Jong and Shephard 1995; Durbin and Koopman 2002; Durbin and Koopman 2012, Ch 4.9).

This two-step process is an example of a partially collapsed Gibbs-sampler (van Dyk and Park 2008). I use this method to sample all the models in this work.

[6] That the shrinkage prior be a scale-mixture of normals is not required, but then the model would no longer be a GDLM, and the efficient methods discussed next cannot be used in estimation.

[7] Technically, it also requires derivatives of the likelihood, but these are generated automatically by Stan using its automatic differentiation engine.

[8] In Stan, the calculation of $p(y|.)$ is done in the `transformed parameter` or `model` blocks, while the sampling of the latent states is done in the `generated quantities` block.

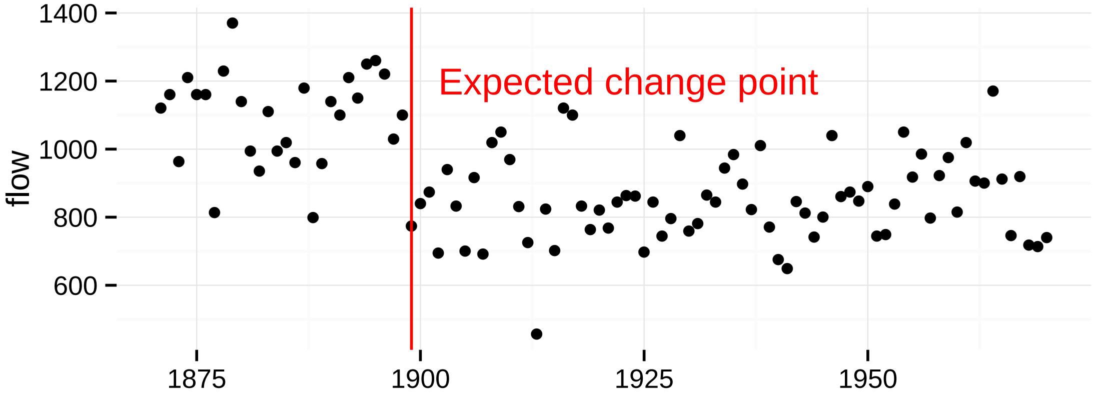


Figure 2: Annual flow of the Nile River, 1871–1970. Previous work has found a break point in this series near 1899.

This required implementing Kalman filter and simulation smoothing methods in Stan. Along with this work, I provide a full set of user-defined Stan functions that implement the Kalman filter, smoother, and simulation sampling Arnold (2015). Section 7 provides the code for one of the Stan models used in this paper which uses these functions.[9]

3 Example: Annual Nile River Flows

A classic dataset that has been analyzed in many works and texts on time series and structural breaks is the annual flow volume of the Nile River between 1871 and 1970 (Cobb 1978; Balke 1993; de Jong and Penzer 1998; Durbin and Koopman 2012; Commandeur, Koopman, and Ooms 2011).[10] Figure 2 plots this data. This series seems to show a single large shift in the average level of the annual flow around 1899. This level shift is attributed to construction of a dam at Aswan that started operation in 1902 or to climate changes that reduced rainfall in the area (Cobb 1978, p. 278).

I compare several models of this data. In all models, y_t is the annual flow of the Nile River,[11] and consists of $n = 100$ annual observations from 1871 to 1970. In all models, each observation y_t is distributed normal with mean μ_t and a common variance, σ^2.[12]

$$y_t \sim \mathcal{N}\left(\mu_t, \sigma^2\right) \tag{20}$$

The models differ in how they model the possibly time-varying mean, μ_t. I estimate the following models:

`Constant` In this model, the mean is a constant.

$$\mu_t = \mu_0 \quad \text{for all } t \tag{21}$$

`Intervention` This model includes an indicator variable for all years including and after 1899. This models a situation in which the researcher knows, or suspects they know, the change points, and is manually

[9] The full code for all models run in this work are available at https://github.com/jrnold/dlm-shrinkage.

[10] The dataset is included with `R` as `Nile` in the included package **datasets**.

[11] Discharge at Aswan in $10^8 m^3$.

[12] This may not be appropriate since that data contains possible outliers at 1879, 1913, and 1964. However, modeling outliers is not the purpose of this analysis, so I ignore that.

accounting for them.

$$\mu_t = \begin{cases} \mu_0 & t < 1899 \\ \mu_0 + \omega & t \geq 1899 \end{cases} \tag{22}$$

`Normal` In this model, the system errors are distributed normal. This corresponds to the a local level model (Durbin and Koopman 2012, Ch. 2; West and Harrison 1997, Ch. 2).[13]

$$\mu_t = \mu_{t-1} + \omega_t \quad \omega_t \sim N(0, \tau^2) \tag{23}$$

`StudentT` In this model, the system errors are distributed Student t with degrees of freedom ν,

$$\mu_t = \mu_{t-1} + \omega_t \quad \omega_t \sim \mathscr{T}_\nu(0, \tau) \tag{24}$$

The prior distribution of the degrees of freedom parameter ν is that suggested by Juárez and Steel (2010), a Gamma distribution with shape parameter 2 and rate parameter 0.1 (mode 10, mean 20, and variance 200). This places most of mass in the relevant region of the space—away from 0 but less than 30, after which the distribution is effectively indistinguishable from a normal distribution.

`Laplace` In this model, the system errors are distributed Laplace (double exponential):

$$\mu_t = \mu_{t-1} + \omega_t \quad \omega_t \sim \mathscr{L}(0, \tau) \tag{25}$$

`Horseshoe` In this model, the system errors are distributed horseshoe.

$$\begin{aligned} &\mu_t = \mu_{t-1} + \omega_t \quad \omega_t \sim \mathcal{N}\left(0, \tau^2 \lambda_t^2\right) \\ &\lambda_t \sim \mathscr{C}^+(0, 1) \end{aligned} \tag{26}$$

`Horseshoe+` In this model, the system errors are distributed horseshoe+.

$$\begin{aligned} &\mu_t = \mu_{t-1} + \omega_t \quad \omega_t \sim \mathcal{N}\left(0, \tau^2 \lambda_t^2\right) \\ &\lambda_t \sim \mathscr{C}^+\left(0, \eta_t\right) \\ &\eta_t \sim \mathscr{C}^+(0, 1) \end{aligned} \tag{27}$$

For the global scale parameter, τ, in the `Laplace`, `StudentT`, `Horseshoe`, and `Horseshoe+` models, I use the half-Cauchy prior $\tau \sim \mathscr{C}^+\left(0, \frac{1}{n}\right)$. In the `Normal` and `Intervention` models, τ is given a semi-informative half-Cauchy prior with a scale equal to a multiple of the standard deviation of the data.

Figure 3 plots the posterior distribution of μ for each model. The `Normal` model does not show a clean break at 1899, instead it estimates a change occurring over several years. The `Laplace` model looks similar to the `Normal` model. While the Laplace distribution achieves sparsity in maximum likelihood estimates because because they use the mode as the estimate, it does produce sparse posterior mean estimates. Since the distribution does not concentrate much mass near zero, it is not surprising that it does not perform much differently than the normal distribution (Park and Casella 2008). Both the `Horseshoe` and `Horseshoe+` models produce a posterior distribution of μ that appears similar to the step function in the `Intervention` model. Additionally, the `StudentT` also produces estimates similar to the `Intervention`, but with slightly wider posterior distributions than the horseshoe models. Figure 3 plots the posterior distribution of the system errors, ω, for each model. The `Horseshoe`, `Horseshoe+`, and `StudentT` models all estimate $\mathrm{E}(\omega_t | y)$ near zero for all years but 1899.

Table 1 compares the models using several statistics. First, I compare the models on their fit to the in-sample data using the root mean squared error (RMSE), RMSE(y), The RMSE is defined as $\sqrt{\frac{1}{n} \sum_i (y_i - \mathrm{E}(\mu_i | y))^2}$,

[13] For example, the local level model implemented in the `R` function `StructTS`.

where $\mathrm{E}(\mu_i|y)$ is the posterior mean of μ_i. I also compare the models based on their expected fit to out-of-sample data with the expected log predictive density calculated using two methods: the Widely Applicable Information Criterion (WAIC), elpd_{WAIC}, and leave-one-out (LOO) cross-validation, elpd_{loo}. The log probability density of a new observation is the expected value of the posterior density of a future observation, $\log \mathrm{E}(p(\tilde{y}|\theta))$. Since the value of the future observation, $\tilde{y}$, is unknown, the expected log probability density averages over the predictive distribution of $\tilde{y}$, $\mathrm{elpd} = \mathrm{E}_f(\log p(\tilde{y}|\theta, y_i))$, where f is the distribution of $\tilde{y}$. However, the distribution of future values is in general also unknown, which why two approximations are used. elpd_{WAIC} approximates the elpd using an information criteria similar to AIC, BIC or DIC, taking the in-sample log-likelihood and penalizing it for model complexity. elpd_{loo} approximates the elpd using leave-one-out cross validation. See Gelman, Carlin, et al. (2013), Gelman and Vehtari (2014), or Gelman, Hwang, and Vehtari (2014) for more thorough discussion of elpd and predictive measures for Bayesian model comparison.[14] In the RMSE, a lower value indicates a better fit, for elpd, a higher value indicates is a better fit. The `Horseshoe` and `Horseshoe+` models both have the best fit in terms of RMSE and elpd values of the shrinkage priors, though neither fits either the in-sample or out-of-sample data as well as the `Intervention` data. Surprisingly, the `StudentT` model is close in performance to the horseshoe models.

Second, I compare the fits of the model to the "true" values of μ_t. But, since this is real data, I do not know the true values of μ_t. Instead, I will will compare the other models to posterior mean estimate of `Intervention` model. The column RMSE(μ) of Table 1 is the root mean squared error of the models compared to the posterior mean of μ as estimated by the `Intervention` model, defined as $\sqrt{\frac{1}{n}\sum(\mathrm{E}(\mu_t|y) - \bar{\mu}_t)^2}$, where $\bar{\mu}_t$ is the posterior mean of μ_t in the `Intervention` model. As with comparisons of fit to the observed data, the `Horseshoe` and `Horseshoe+` models have the lowest RMSE, although the `StudentT` model is close.

model	RMSE(y)	elpd_{WAIC}	elpd_{loo}	RMSE(μ)
`Constant`	168.38	-656.49	-656.49	111.24
`Intervention`	126.39	-629.09	-629.10	
`Normal`	143.82	-641.87	-641.88	37.52
`StudentT`	137.46	-637.61	-637.66	13.50
`Laplace`	140.89	-640.43	-640.49	25.75
`Horseshoe`	136.48	-635.89	-635.98	9.44
`Horseshoe+`	136.42	-637.47	-638.26	12.59

Table 1: Model comparison statistics for models of the Nile Rive annual flow data.

4 Change Points in Levels and Trends

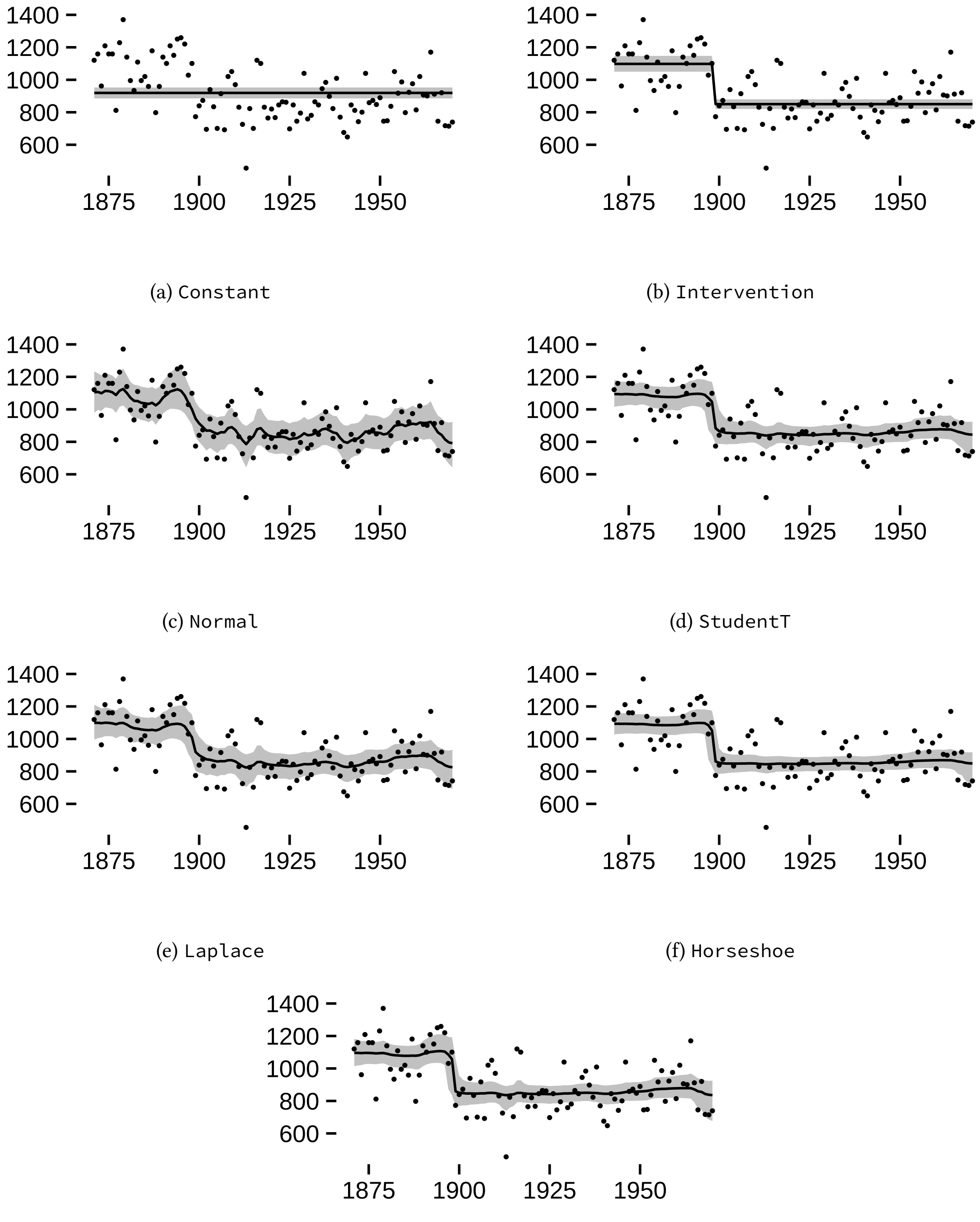
The use of sparsity inducing priors can be extended to model change points in trends in addition to the level.[15] The local level model considered in section 1 can be extended to a local trend model (Durbin and Koopman 2012, Ch 3.2; West and Harrison 1997, Ch 7),

$$\begin{aligned} y_t &= \mu_t + \epsilon_t & \epsilon_t &\sim \mathcal{N}\left(0, \sigma^2\right) \\ \mu_t &= \mu_{t-1} + \alpha_{t-1} + \omega_{1,t} \\ \alpha_t &= \alpha_{t-1} + \omega_{2,t} \end{aligned} \tag{28}$$

In Equation (28), there are two states: μ_t is the current level, and α_t is the current trend (change in the level). The level is changing over time both due to the current value of the trend, α_t, and the system errors, $\omega_{1,t}$.

[14] The way these are calculated does not fully account for the time-series nature of this data. The measures presented here should be seen as approximating the fit of the model to a previously missing value within the time-series. To calculate elpd_{WAIC} and elpd_{loo}, I use the **loo** `R` package (Vehtari, Gelman, and Gabry 2015), which implements the methods described in Gelman and Vehtari (2014).

[15] Kim et al. (2009) and Tibshirani (2014) consider similar problems in a maximum likelihood framework with ℓ_1 regularization.



(a) `Constant`

(b) `Intervention`

(c) `Normal`

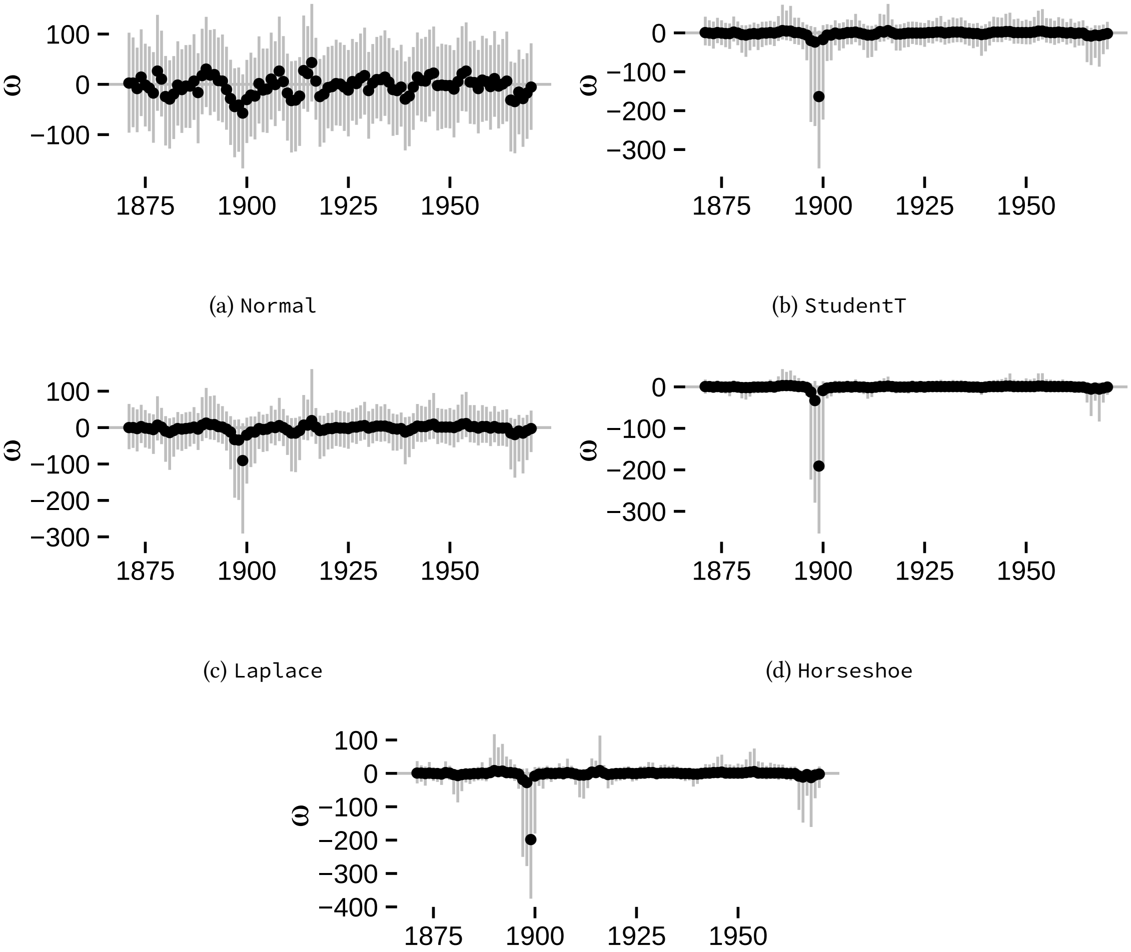
(d) `StudentT`

(e) `Laplace`

(f) `Horseshoe`

(g) `Horseshoe+`

Figure 3: Posterior distributions of μ_t for models of the Nile River annual flow data. The line is the posterior mean; the range of the ribbon the 2.5–97.5% percentiles of the posterior distribution.



(a) `Normal`

(b) `StudentT`

(c) `Laplace`

(d) `Horseshoe`

(e) `Horseshoe+`

Figure 4: Posterior distributions of ω_t for models of the Nile River annual flow data. The point is the posterior mean; the range of the line is the 2.5–97.5% percentiles of the posterior distribution.

While the trend is changing over time only due to the system error, $\omega_{2,t}$. This model will allow for change points in both the level and trend if sparsity inducing shrinkage priors are used for ω_1 and ω_2. The system errors in a local trend model could be modeled with an arbitrary covariance structure, but in this work, I follow the suggestion of West and Harrison (1997, Ch 7.),

$$\begin{bmatrix}\omega_{1,t} \\ \omega_{2,t}\end{bmatrix}' \sim \mathcal{N}\left(0, L\,\mathrm{diag}(W_1^2, W_2^2)L'\right); \quad L = \begin{bmatrix}1 & 1 \\ 0 & 1\end{bmatrix} \tag{29}$$

Since the sparsity inducing shrinkage priors discussed can be represented as scale-mixtures of normals, the previous equation can be expressed as:

$$\begin{bmatrix}\omega_{1,t} \\ \omega_{2,t}\end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix}\tau_1^2 \lambda_{1,t}^2 + \tau_2 \lambda_{2,t}^2 & \tau_2 \lambda_{2,t} \\ \tau_2^2 \lambda_{2,t} & \tau_2^2 \lambda_{2,t}\end{bmatrix}\right) \tag{30}$$

where τ_1 and $\lambda_{1,t}$ are the global and local variance components for the level, and τ_2 and $\lambda_{2,t}$ are the global and local variance components for the trend. Since this model is a GDLM, it can be efficiently sampled using the methods in 2.

5 Example: George W. Bush Approval Ratings

As an example of a time series that is a smooth curve with jumps Ratkovic and Eng (2010) use the approval ratings for George W. Bush, displayed in Figure 5. George W. Bush's approval ratings are difficult to fit with typical smoothing methods because it was subject to two large jumps, September 11th, 2001, and at the start of the Iraq War on March 20, 2003. The data used in this example consists of 270 polls between February 04, 2001 and January 11, 2009 from the Roper Center Public Opinion [16]

$$\begin{aligned} y_t &= \mu_t + \epsilon_t & \epsilon_t &\sim \mathcal{N}\left(0, \sigma^2\right) \\ \mu_t &= \alpha_t + \mu_{t-1} + \partial\mu_{t-1} + \omega_{1,t} \\ \partial\mu_t &= +\partial\mu_{t-1} + \omega_{2,t} \\ \begin{bmatrix}\omega_{1,t} \\ \omega_{2,t}\end{bmatrix} &\sim \mathcal{N}\left(0, \begin{bmatrix}\tau_1^2 \lambda_{1,t}^2 + \tau_2 \lambda_{2,t}^2 & \tau_2 \lambda_{2,t} \\ \tau_2^2 \lambda_{2,t} & \tau_2^2 \lambda_{2,t}\end{bmatrix}\right) \end{aligned} \tag{31}$$

`Normal` System errors are distributed normal. $\lambda_{i,t} = 1$ for all i for all t, $\alpha_t = 0$ for all t.

`Intervention` System errors are distributed normal, $\lambda_{i,t} = 1$ for all i for all t. There are manual interventions after 9/11 and the the Iraq War. The values of α_t for those two dates are non-zero and estimated, all other α_t are set to zero. This corresponds to a manual intervention for known change points.

`Horseshoe` The system errors, $\omega_{i,t}$, are distributed horseshoe, with $\alpha_t = 0$ for all t.

`Horseshoe+` The system errors, $\omega_{i,t}$, are distributed horseshoe+, with $\alpha_t = 0$ for all t.

Figures 6 and 7 plot the posterior distribution of μ_t for the models. The `Normal` model shows Bush's approval rating rising before 9/11 and is rough, with much small variation between the jumps. The `Horseshoe` and `Horseshoe+` models more closely resemble the `Intervention` model: Bush's approval rating are loping downward or steady until 9/11, and otherwise the approval ranting is mostly smooth. Table 2 shows the RMSE and expected log predictive densities of the these models. The horseshoe models fit the the data better than a normal distribution, but less well than the `Intervention` model.

[16] From http://webapps.ropercenter.uconn.edu/CFIDE/roper/presidential/webroot/presidential_rating_detail.cfm?allRate=True&presidentName=Bush#.UbeB8HUbyv8.

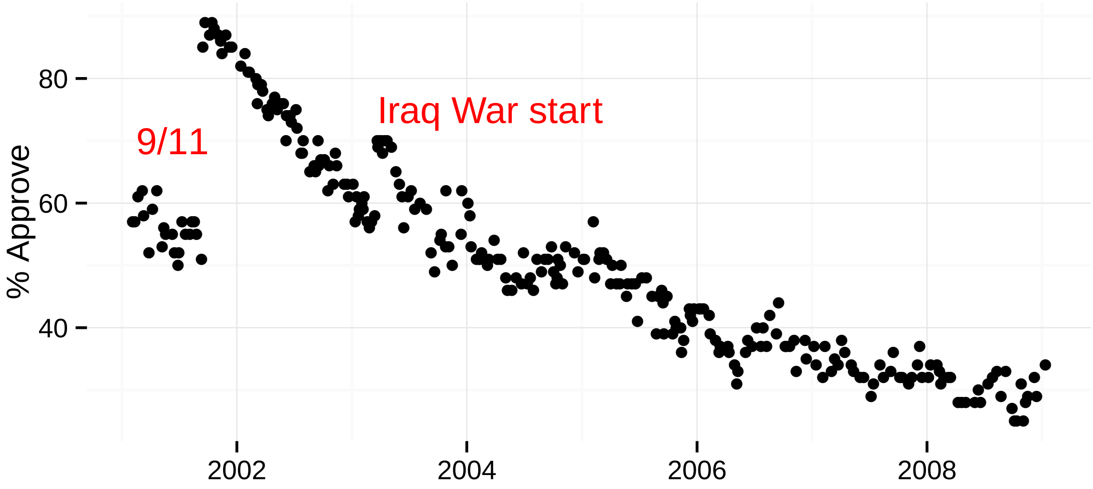
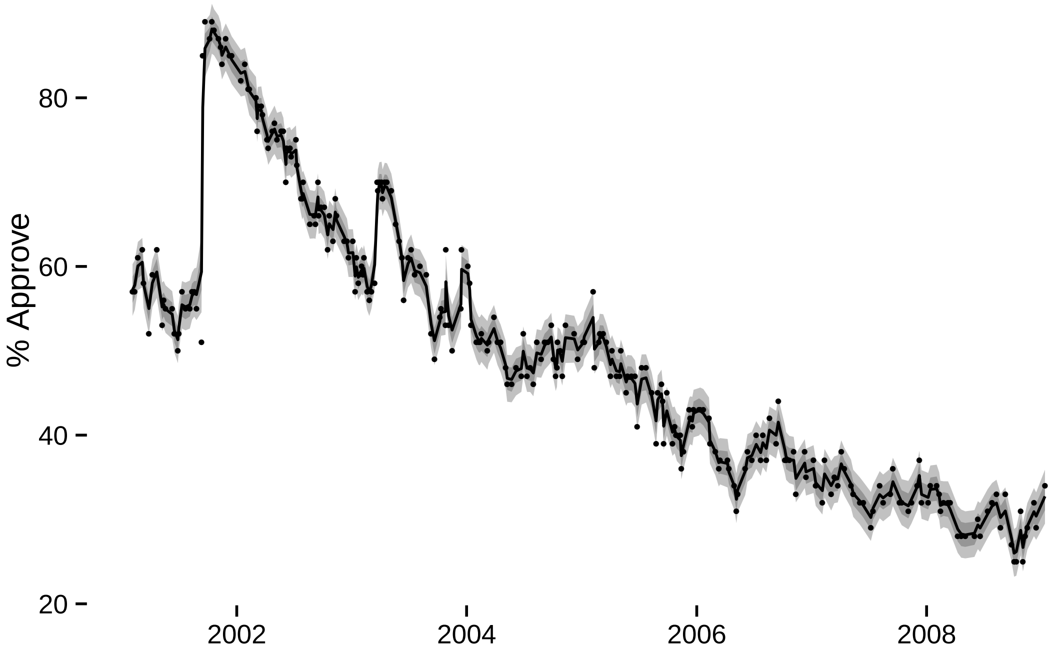


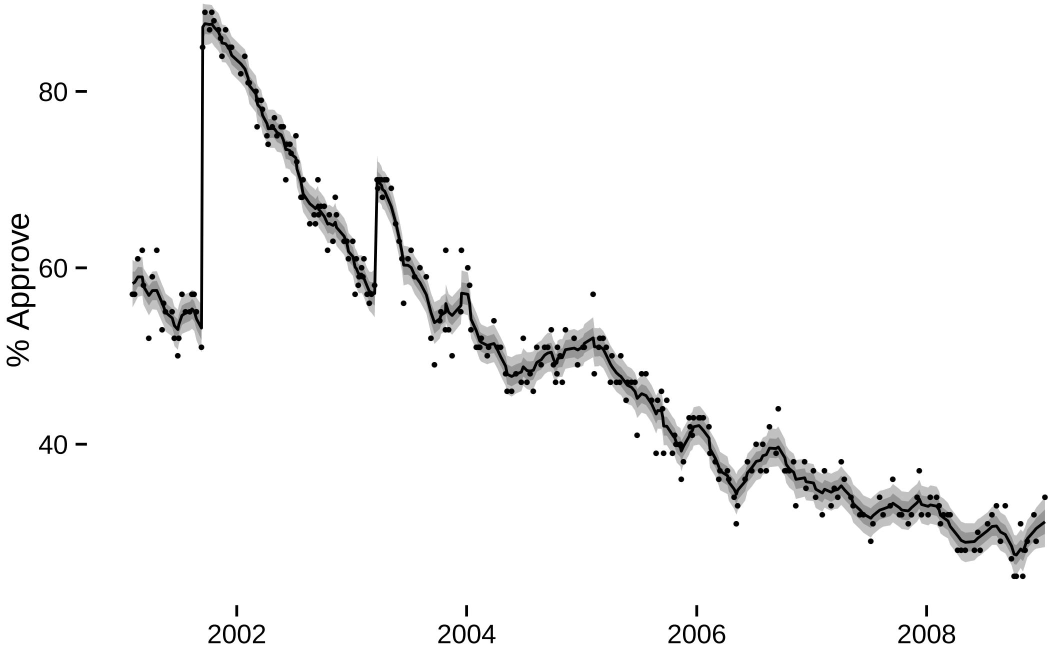
Figure 5: Approval ratings of President George W. Bush

model	RMSE	elpd$_{WAIC}$	elpd$_{loo}$
`Normal`	3.73	-748.51	-749.43
`Intervention`	2.80	-666.24	-666.84
`Horseshoe`	3.44	-680.65	-681.15
`Horseshoe+`	3.43	-682.56	-683.61

Table 2: Model comparison statistics for models of President George W. Bush’s approval rating.

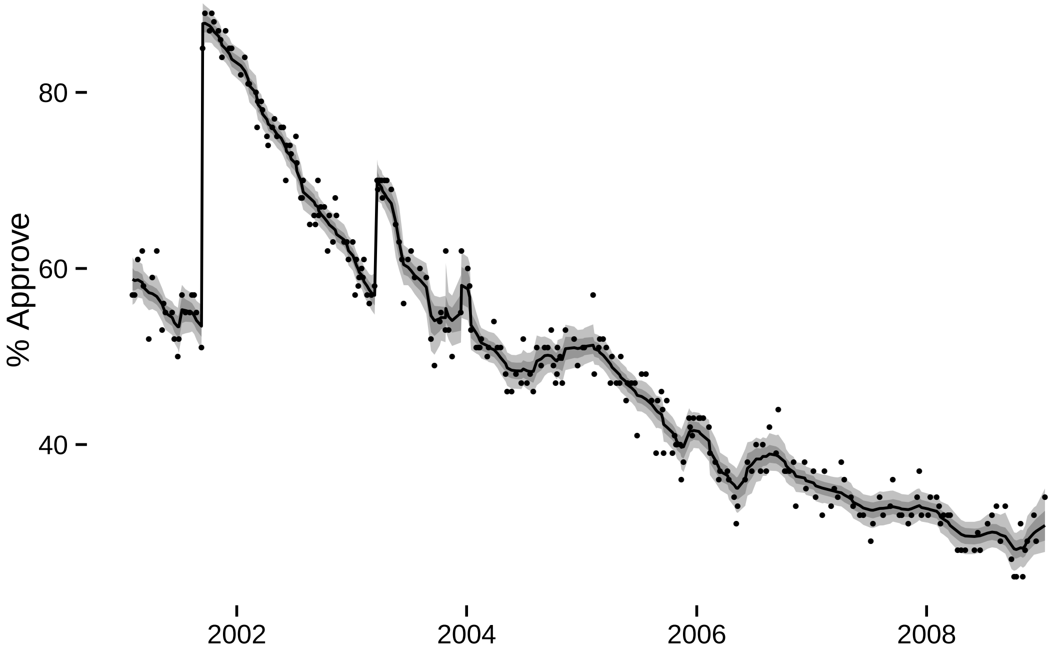


(a) `Normal`

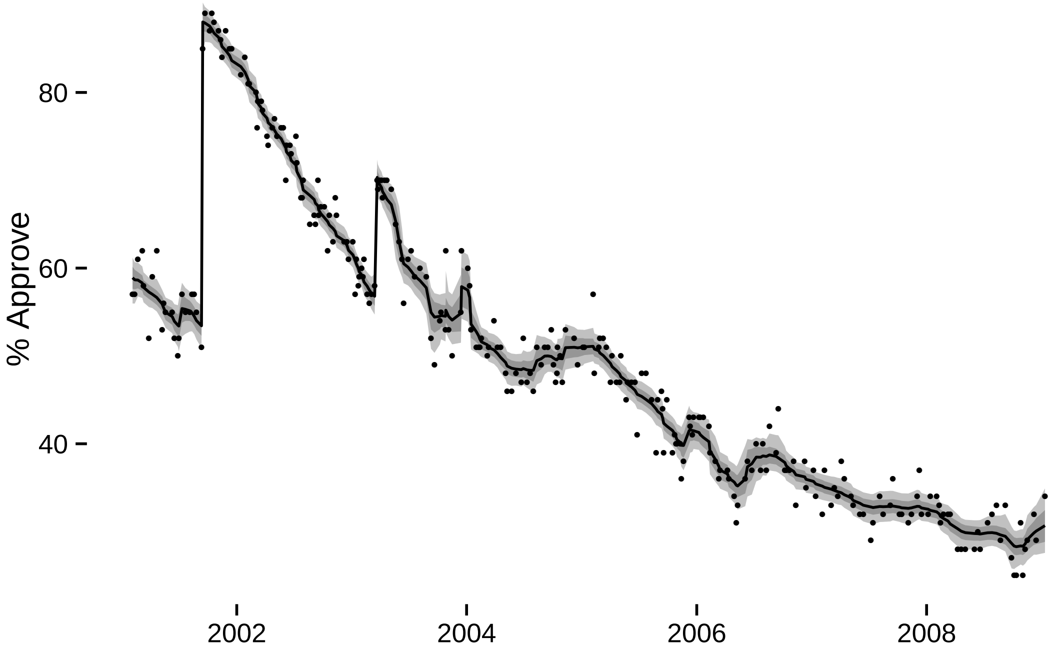


(b) `Intervention`

Figure 6: Posterior distribution of μ for `Normal` and `Intervention` models.



(a) `Horseshoe`



(b) `Horseshoe+`

Figure 7: Posterior distribution of μ for `Horseshoe` and `Horseshoe+` models.

6 Conclusion

This work proposes modeling change points through the use sparse shrinkage priors, such as the horseshoe, for the change in a parameter. This approach has several useful features. It does not require choosing a specific number of change points, and the sparsity of the changes can be estimated from the data. Although it does not directly estimate the probability of change points, it is closer to the data generating process of many political processes in which there is always change, but which are characterized by many periods of small changes, and only a few periods of large changes. Since these shrinkage priors are scale-mixtures of normal distributions, these models fall into the class of Gaussian dynamic linear models, and, thus, can be sampled using efficient algorithms specific to that class of models. This work provides a partially collapsed Gibbs sampler method to estimate GDLMs in Stan, as well Stan code that implements Kalman filtering and smoothing in Stan.

The most promising feature of this approach is that it is flexible and can be applied to a variety of models. For example, the following model is a linear regression with independent change points, and K variables,

$$\begin{aligned} y &\sim \mathcal{N}\left(\alpha_t + \beta X, \sigma^2\right) \\ \alpha_t &\sim \mathcal{N}\left(\alpha_{t-1}, \sigma^2 \tau_\alpha^2 \lambda_{\alpha,t}^2\right) \\ \beta_{t,k} &\sim \mathcal{N}\left(\beta_{t-1,k}, \sigma^2 \tau_{\beta,k}^2 \lambda_{\beta,t}^2\right) \quad \text{for } k \in 1, \dots, K \end{aligned} \tag{32}$$

where the local variance components, λ_α and λ_β are given prior distributions corresponding to a sparse shrinkage distribution such as the horseshoe or horseshoe+. This model corresponds to a GDLM with latent states α and β, and thus the partially collapsed Gibbs sampling method in Section 2 can be used to efficiently estimate it. As written, this would correspond to independent change points in the parameters, α and β. If the researcher wanted to impose a restriction that large changes occurred at the same time for all distributions, they could set $\tau_\alpha = \tau_{\beta,1} = \cdots = \tau_{\beta,K}$ and $\lambda_\alpha = \alpha_{\beta,1} = \cdots = \alpha_{\beta,K}$. This is one example, but these sparse shrinkage parameters can be applied to any model with a time-varying parameter with support $\mathbb{R}$. If that model is of the class of GDLMs, then there are efficient methods to sample it, if not, the model can still be estimated in Stan using its usual algorithms.

7 Example Stan Program

An example of a change point model implemnted in Stan. See the replication data for this work to see the code for all the Stan models estimated. The DLM related user-defined functions in the `functions` block are excluded. The code for them can be found in Arnold (2015) or https://raw.githubusercontent.com/jrnold/dlm-shrinkage/master/stan/includes/dlm.stan.

```
data {
  int<lower = 1> n;
  vector[n] y;
  int miss[n];
  real m0;
  real<lower = 0.0> C0;
  real<lower = 0.0> s;
  real<lower = 0.0> w;
}
parameters {
  real<lower = 0.0> sigma;
  real<lower = 0.0> tau;
  vector<lower = 0.0>[n] lambda;
}
transformed parameters {
  vector[n] log_lik;
  vector[6] dlm[n + 1];
  vector[n] W;

  for (i in 1:n) {
    W[i] <- pow(sigma * tau * lambda[i], 2);
  }
  {
    vector[n] V;
    V <- rep_vector(pow(sigma, 2), n);
    dlm <- dlm_local_level_filter(n, y, miss, V, W, m0, C0);
    log_lik <- dlm_local_level_filter_loglik(n, dlm, miss);
  }

}
model {
  real ll;

  sigma ~ cauchy(0.0, s);
  tau ~ cauchy(0.0, w);
  lambda ~ cauchy(0.0, 1);
  increment_log_prob(sum(log_lik));
}
generated quantities {
  vector[1] mu[n + 1];
  vector[1] omega[n];
  vector[1] kalman[n];

  {
    matrix[1, 1] G_tv[n];

    G_tv <- rep_array(rep_matrix(1.0, 1, 1), n);
    mu <- dlm_filter_bsample_rng(n, 1, 1, G_tv, dlm);
  }
  for (i in 1:n) {
    omega[i] <- mu[i + 1] - mu[i];
  }
  for (i in 1:n) {
    kalman[i] <- dlm_get_C(i, 1, 1, dlm) * dlm_get_Q_inv(i, 1, 1, dlm);
  }

}
```

References

Arnold, Jeffrey (2015). “Dynamic Linear Model Functions in Stan”. URL: http://dx.doi.org/10.6084/m9.figshare.1553209 (cit. on pp. 8, 17).

Bai, Jushan and Pierre Perron (1998). “Estimating and Testing Linear Models with Multiple Structural Changes”. *Econometrica* 66.1, pp. 47–78. ISSN: 00129682. URL: http://www.jstor.org/stable/2998540 (cit. on p. 2).

— (2003). “Computation and analysis of multiple structural change models”. *Journal of Applied Econometrics* 18.1, pp. 1–22. ISSN: 1099-1255. DOI: 10.1002/jae.659. URL: http://dx.doi.org/10.1002/jae.659 (cit. on p. 2).

Balke, Nathan S. (1993). “Detecting Level Shifts in Time Series”. *Journal of Business & Economic Statistics* 11.1, pp. 81–92. ISSN: 07350015. URL: http://www.jstor.org/stable/1391308 (cit. on p. 8).

Barry, Daniel and J. A. Hartigan (1993). “A Bayesian Analysis for Change Point Problems”. *Journal of the American Statistical Association* 88.421, pp. 309–319. ISSN: 01621459. URL: http://www.jstor.org/stable/2290726 (cit. on p. 2).

Beck, Nathaniel (1989). “Estimating Dynamic Models Using Kalman Filtering”. *Political Analysis* 1.1, pp. 121–156. DOI: 10.1093/pan/1.1.121. eprint: http://pan.oxfordjournals.org/content/1/1/121.full.pdf+html. URL: http://pan.oxfordjournals.org/content/1/1/121.abstract (cit. on p. 6).

Bhadra, Anindya, Jyotishka Datta, Nicholas G. Polson, and Brandon Willard (2015). “The Horseshoe+ Estimator of Ultra-Sparse Signals”. eprint: 1502.00560 (cit. on pp. 1, 3, 4).

Blackwell, Matthew (2012). “Game-changers: Detecting shifts in the Flow of campaign contributions”. URL: http://www.mattblackwell.org/files/papers/gamechangers.pdf (cit. on pp. 1, 6).

Büthe, Tim (2002). “Taking Temporality Seriously: Modeling History and the Use of Narratives as Evidence”. *American Political Science Review* null (03), pp. 481–493. ISSN: 1537-5943. DOI: 10.1017/S0003055402000278. URL: http://journals.cambridge.org/article_S0003055402000278 (cit. on p. 1).

Calderia, Gregory A. and Christopher J. W. Zorn (1998). “Of Time and Consensual Norms in the Supreme Court”. *American Journal of Political Science* 42.3, pp. 874–902. ISSN: 00925853. URL: http://www.jstor.org/stable/2991733 (cit. on pp. 1, 6).

Carpenter, Bob, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Micahel Betancourt, Marcus A. Brubaker, Jiquiang Guo, Peter Li, and Allend Riddell (2015). “Stan: A Probabilistic Programming Language”. *Journal of Statistical Software* ? (?). URL: http://www.stat.columbia.edu/~gelman/research/published/stan-paper-revision-feb2015.pdf (cit. on p. 7).

Carter, C. K. and R. Kohn (1994). “On Gibbs sampling for state space models”. *Biometrika* 81.3, pp. 541–553. DOI: 10.1093/biomet/81.3.541. eprint: http://biomet.oxfordjournals.org/content/81/3/541.full.pdf+html. URL: http://biomet.oxfordjournals.org/content/81/3/541.abstract (cit. on pp. 2, 7).

Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott (2009). “Handling Sparsity via the Horseshoe”. *Journal of Machine Learning and Research: Workshop and Conference Proceedings* 5, pp. 73–80 (cit. on pp. 1, 4).

— (2010). “The horseshoe estimator for sparse signals”. *Biometrika* 97.2, pp. 465–480. DOI: 10.1093/biomet/asq017. eprint: http://biomet.oxfordjournals.org/content/97/2/465.full.pdf+html. URL: http://biomet.oxfordjournals.org/content/97/2/465.abstract (cit. on pp. 1, 3, 4).

Chan, Ngai Hang, Chun Yip Yau, and Rong-Mao Zhang (2014). “Group LASSO for Structural Break Time Series”. *Journal of the American Statistical Association* 109.506, pp. 590–599. DOI: 10.1080/01621459.2013.866566. eprint: http://dx.doi.org/10.1080/01621459.2013.866566. URL: http://dx.doi.org/10.1080/01621459.2013.866566 (cit. on p. 4).

Chib, Siddhartha (1998). “Estimation and comparison of multiple change-point models”. *Journal of Econometrics* 86.2, pp. 221–241. ISSN: 0304-4076. DOI: DOI:10.1016/S0304-4076(97)00115-2. URL: http://www.sciencedirect.com/science/article/B6VC0-3VM1XM5-2/2/469ee3cba827365611dee3677f0babc6 (cit. on p. 2).

Cobb, George W. (1978). “The problem of the Nile: Conditional solution to a changepoint problem”. *Biometrika* 65.2, pp. 243–251. DOI: 10.1093/biomet/65.2.243. eprint: http://biomet.oxfordjournals.org/content/65/2/243.full.pdf+html. URL: http://biomet.oxfordjournals.org/content/65/2/243.abstract (cit. on p. 8).

Commandeur, Jacques J. F., Siem Jan Koopman, and Marius Ooms (2011). “Statistical Software for State Space Methods”. *Journal of Statistical Software* 41.1, pp. 1–18. issn: 1548-7660. url: http://www.jstatsoft.org/v41/i01 (cit. on p. 8).

Datta, Jyotishka and Jayanta. K. Ghosh (2012). “Asymptotic Properties of Bayes Risk for the Horseshoe Prior”. *Bayesian Analysis* 7 (4), pp. 771–792. doi: 10.1214/13-BA805. url: http://projecteuclid.org/euclid.ba/1362406654 (cit. on p. 4).

de Jong, Piet and Jeremy Penzer (1998). “Diagnosing Shocks in Time Series”. *Journal of the American Statistical Association* 93.442, pp. 796–806. issn: 01621459. url: http://www.jstor.org/stable/2670129 (cit. on pp. 6, 8).

De Jong, Piet and Neil Shephard (1995). “The simulation smoother for time series models”. *Biometrika* 82.2, pp. 339–350. doi: 10.1093/biomet/82.2.339. eprint: http://biomet.oxfordjournals.org/content/82/2/339.full.pdf+html. url: http://biomet.oxfordjournals.org/content/82/2/339.abstract (cit. on p. 7).

Durbin, J. and S. J. Koopman (2002). “A Simple and Efficient Simulation Smoother for State Space Time Series Analysis”. *Biometrika* 89.3, pp. 603–615. issn: 00063444. url: http://www.jstor.org/stable/4140605 (cit. on p. 7).

Durbin, J. and S.J. Koopman (2012). *Time Series Analysis by State Space Methods: Second Edition*. Oxford Statistical Science Series. OUP Oxford. isbn: 9780199641178. url: http://books.google.com/books?id=fOq39Zh0olQC (cit. on pp. 6–10).

Efron, Bradley (2008). “Microarrays, Empirical Bayes and the Two-Groups Model”. *Statist. Sci.* 23.1, pp. 1–22. doi: 10.1214/07-STS236. url: http://dx.doi.org/10.1214/07-STS236 (cit. on pp. 2, 3).

Fearnhead, Paul (2006). “Exact and efficient Bayesian inference for multiple changepoint problems”. *Statistics and Computing* 16.2, pp. 203–213. issn: 0960-3174. doi: 10.1007/s11222-006-8450-8. url: http://dx.doi.org/10.1007/s11222-006-8450-8 (cit. on p. 2).

Fearnhead, Paul and Zhen Liu (2007). “On-line inference for multiple changepoint problems”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.4, pp. 589–605. issn: 1467-9868. doi: 10.1111/j.1467-9868.2007.00601.x. url: http://dx.doi.org/10.1111/j.1467-9868.2007.00601.x (cit. on p. 2).

Frühwirth-Schnatter, Sylvia (1994). “Data Augmentation And Dynamic Linear Models”. *Journal of Time Series Analysis* 15.2, pp. 183–202. issn: 1467-9892. doi: 10.1111/j.1467-9892.1994.tb00184.x. url: http://dx.doi.org/10.1111/j.1467-9892.1994.tb00184.x (cit. on p. 7).

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin (2013). *Bayesian Data Analysis*. 3rd. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. isbn: 9781439840955. url: https://books.google.com/books?id=ZXL6AQAAQBAJ (cit. on p. 10).

Gelman, Andrew, Jessica Hwang, and Aki Vehtari (2014). “Understanding predictive information criteria for Bayesian models”. *Statistics and Computing* 24.6, pp. 997–1016. issn: 0960-3174. doi: 10.1007/s11222-013-9416-2. url: http://dx.doi.org/10.1007/s11222-013-9416-2 (cit. on p. 10).

Gelman, Andrew and Andrew Vehtari (2014). “WAIC and Cross-validation in STAN”. url: http://www.stat.columbia.edu/~gelman/research/unpublished/waic_stan.pdf (cit. on p. 10).

Giordani, Paolo and Robert Kohn (2008). “Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models”. *Journal of Business & Economic Statistics* 26.1, pp. 66–77. doi: 10.1198/073500107000000241. eprint: http://amstat.tandfonline.com/doi/pdf/10.1198/073500107000000241. url: http://amstat.tandfonline.com/doi/abs/10.1198/073500107000000241 (cit. on p. 3).

Hans, Chris (2009). “Bayesian lasso regression”. *Biometrika* 96.4, pp. 835–845. doi: 10.1093/biomet/asp047. eprint: http://biomet.oxfordjournals.org/content/96/4/835.full.pdf+html. url: http://biomet.oxfordjournals.org/content/96/4/835.abstract (cit. on p. 3).

Harchaoui, Z. and C. Lévy-Leduc (2010). “Multiple Change-Point Estimation With a Total Variation Penalty”. *Journal of the American Statistical Association* 105.492, pp. 1480–1493. doi: 10.1198/jasa.2010.tm09181. eprint: http://dx.doi.org/10.1198/jasa.2010.tm09181. url: http://dx.doi.org/10.1198/jasa.2010.tm09181 (cit. on p. 4).

Harvey, Andrew C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press. isbn: 9780521405737. url: http://books.google.com/books?id=Kc6tnRHBwLcC (cit. on p. 7).

Hinkley, David V. (1970). “Inference about the change-point in a sequence of random variables”. *Biometrika* 57.1, pp. 1–17. DOI: 10.1093/biomet/57.1.1. eprint: http://biomet.oxfordjournals.org/content/57/1/1.full.pdf+html. URL: http://biomet.oxfordjournals.org/content/57/1/1.abstract (cit. on p. 2).

Hoffman, Matthew D. and Andrew Gelman (2014). “The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo”. *Journal of Machine Learning Research* 15.1, pp. 1593–1623. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2627435.2638586 (cit. on p. 7).

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences.* Wiley Series in Probability and Statistics. Wiley. ISBN: 9780470686638. URL: http://books.google.com/books?id=QFqyrNL8yEkC (cit. on p. 2).

Juárez, Miguel A. and Mark F. J. Steel (2010). “Non-Gaussian dynamic Bayesian modelling for panel data”. *Journal of Applied Econometrics* 25.7, pp. 1128–1154. ISSN: 1099-1255. URL: http://dx.doi.org/10.1002/jae.1113 (cit. on p. 9).

Killick, R., P. Fearnhead, and I. A. Eckley (2012). “Optimal Detection of Changepoints With a Linear Computational Cost”. *Journal of the American Statistical Association* 107.500, pp. 1590–1598. DOI: 10.1080/01621459.2012.737745. eprint: http://dx.doi.org/10.1080/01621459.2012.737745. URL: http://dx.doi.org/10.1080/01621459.2012.737745 (cit. on p. 2).

Kim, S., K. Koh, S. Boyd, and D. Gorinevsky (2009). “ℓ_1 Trend Filtering”. *SIAM Review* 51.2, pp. 339–360. DOI: 10.1137/070690274. eprint: http://dx.doi.org/10.1137/070690274. URL: http://dx.doi.org/10.1137/070690274 (cit. on p. 10).

Lieberman, Robert C. (2002). “Ideas, Institutions, and Political Order: Explaining Political Change”. *American Political Science Review* null (04), pp. 697–712. ISSN: 1537-5943. DOI: 10.1017/S0003055402000394. URL: http://journals.cambridge.org/article_S0003055402000394 (cit. on p. 1).

Martin, Andrew D. and Kevin M. Quinn (2002). “Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999”. *Political Analysis* 10.2, pp. 134–153. DOI: 10.1093/pan/10.2.134. eprint: http://pan.oxfordjournals.org/cgi/reprint/10/2/134.pdf. URL: http://pan.oxfordjournals.org/cgi/content/abstract/10/2/134 (cit. on p. 6).

Mitchell, T. J. and J. J. Beauchamp (1988). “Bayesian Variable Selection in Linear Regression”. *Journal of the American Statistical Association* 83.404, pp. 1023–1032. DOI: 10.1080/01621459.1988.10478694. eprint: http://www.tandfonline.com/doi/pdf/10.1080/01621459.1988.10478694. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478694 (cit. on pp. 2, 3).

Olshen, Adam B., E. S. Venkatraman, Robert Lucito, and Michael Wigler (2004). “Circular binary segmentation for the analysis of array-based DNA copy number data”. *Biostatistics* 5.4, pp. 557–572. DOI: 10.1093/biostatistics/kxh008. eprint: http://biostatistics.oxfordjournals.org/content/5/4/557.full.pdf+html. URL: http://biostatistics.oxfordjournals.org/content/5/4/557.abstract (cit. on p. 2).

Page, E. S. (1954). “Continuous Inspection Schemes”. *Biometrika* 41.1/2, pp. 100–115. ISSN: 00063444. URL: http://www.jstor.org/stable/2333009 (cit. on p. 2).

Park, Jong Hee (2010). “Structural Change in U.S. Presidents’ Use of Force”. *American Journal of Political Science* 54.3, pp. 766–782. ISSN: 1540-5907. DOI: 10.1111/j.1540-5907.2010.00459.x. URL: http://dx.doi.org/10.1111/j.1540-5907.2010.00459.x (cit. on pp. 1, 6).

— (2011). “Changepoint Analysis of Binary and Ordinal Probit Models: An Application to Bank Rate Policy Under the Interwar Gold Standard”. *Political Analysis.* DOI: 10.1093/pan/mpr007. eprint: http://pan.oxfordjournals.org/content/early/2011/03/22/pan.mpr007.full.pdf+html. URL: http://pan.oxfordjournals.org/content/early/2011/03/22/pan.mpr007.abstract (cit. on p. 1).

Park, Trevor and George Casella (2008). “The Bayesian Lasso”. *Journal of the American Statistical Association* 103.482, pp. 681–686. DOI: 10.1198/016214508000000337. eprint: http://amstat.tandfonline.com/doi/pdf/10.1198/016214508000000337. URL: http://amstat.tandfonline.com/doi/abs/10.1198/016214508000000337 (cit. on pp. 3, 9).

Pas, S. L. van der, B. J. K. Kleijn, and A. W. van der Vaart (2014). “The Horseshoe Estimator: Posterior Concentration around Nearly Black Vectors”. *Electronic Journal of Statistics* 8, Number2, 2585–2618. DOI: doi:10.1214/14-EJS962. eprint: 1404.0202. URL: http://projecteuclid.org/euclid.ejs/1418134265 (cit. on p. 4).

Petris, G., S. Petrone, and P. Campagnoli (2009). *Dynamic Linear Models with R.* Use R! Springer. ISBN: 9780387772370. URL: http://books.google.com/books?id=VCt3zVq8TO8C (cit. on p. 6).

Pierson, Paul (2004). *Politics in Time: History, Institutions, and Social Analysis*. Princeton University Press. ISBN: 9780691117157. URL: http://books.google.com/books?id=nVtptUoWuO4C (cit. on p. 1).

Polson, Nicholas G. and James G. Scott (2010). "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction". *Bayesian Statistics* (cit. on pp. 1, 3, 4).

— (2012). "On the Half-Cauchy Prior for a Global Scale Parameter". *Bayesian Analysis* 7 (4), pp. 887–902. DOI: 0.1214/12-BA730 (cit. on pp. 4, 6).

Ratkovic, Marc T. and Kevin H. Eng (2010). "Finding Jumps in Otherwise Smooth Curves: Identifying Critical Events in Political Processes". *Political Analysis* 18.1, pp. 57–77. DOI: 10.1093/pan/mpp032. eprint: http://pan.oxfordjournals.org/content/18/1/57.full.pdf+html. URL: http://pan.oxfordjournals.org/content/18/1/57.abstract (cit. on p. 13).

Shumway, R.H. and D.S. Stoffer (2010). *Time Series Analysis and Its Applications*. Springer Texts in Statistics. Springer. ISBN: 9781441978653. URL: http://books.google.com/books?id=NIhXa6UeF2cC (cit. on p. 6).

Spirling, Arthur (2007a). "Bayesian Approaches for Limited Dependent Variable Change Point Problems". *Political Analysis* 15.4, pp. 387–405. DOI: 10.1093/pan/mpm022. eprint: http://pan.oxfordjournals.org/content/15/4/387.full.pdf+html. URL: http://pan.oxfordjournals.org/content/15/4/387.abstract (cit. on p. 1).

— (2007b). ""Turning Points" in the Iraq Conflict". *The American Statistician* 61.4, pp. 315–320. DOI: 10.1198/000313007X247076. eprint: http://pubs.amstat.org/doi/pdf/10.1198/000313007X247076. URL: http://pubs.amstat.org/doi/abs/10.1198/000313007X247076 (cit. on pp. 1, 6).

Stan Development Team (2015). *Stan Modeling Language Users Guide and Reference Manual, Version 2.7.0.* URL: http://mc-stan.org/ (cit. on p. 7).

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288. ISSN: 00359246. URL: http://www.jstor.org/stable/2346178 (cit. on pp. 1, 4).

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight (2005). "Sparsity and Smoothness via the Fused Lasso". *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.1, pp. 91–108. ISSN: 13697412. URL: http://www.jstor.org/stable/3647602 (cit. on p. 4).

Tibshirani, Ryan J. (2014). "Adaptive piecewise polynomial estimation via trend filtering". *The Annals of Statistics* 42.1. arXiv: 1304.2986, pp. 285–323. ISSN: 0090-5364. DOI: 10.1214/13-AOS1189. URL: http://arxiv.org/abs/1304.2986 (visited on 07/07/2014) (cit. on p. 10).

Tipping, Michael E. (2001). "Sparse bayesian learning and the relevance vector machine". *J. Mach. Learn. Res.* 1, pp. 211–244. ISSN: 1532-4435. DOI: 10.1162/15324430152748236. URL: http://dx.doi.org/10.1162/15324430152748236 (cit. on p. 3).

van Dyk, David A and Taeyoung Park (2008). "Partially Collapsed Gibbs Samplers". *Journal of the American Statistical Association* 103.482, pp. 790–796. DOI: 10.1198/016214508000000409. eprint: http://dx.doi.org/10.1198/016214508000000409. URL: http://dx.doi.org/10.1198/016214508000000409 (cit. on p. 7).

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2015). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 0.1. URL: https://github.com/jgabry/loo (cit. on p. 10).

West, M. and J. Harrison (1997). *Bayesian forecasting and dynamic models*. Springer series in statistics. Springer. ISBN: 9780387947259. URL: http://books.google.com/books?id=jcl8lD75fkYC (cit. on pp. 1, 6, 7, 9, 10, 13).

Western, Bruce and Meredith Kleykamp (2004). "A Bayesian Change Point Model for Historical Time Series Analysis". *Political Analysis* 12.4, pp. 354–374. DOI: 10.1093/pan/mph023. eprint: http://pan.oxfordjournals.org/content/12/4/354.full.pdf+html. URL: http://pan.oxfordjournals.org/content/12/4/354.abstract (cit. on pp. 1, 6).

Yao, Yi-Ching (1984). "Estimation of a Noisy Discrete-Time Step Function: Bayes and Empirical Bayes Approaches". *The Annals of Statistics* 12.4, pp. 1434–1447. ISSN: 00905364. URL: http://www.jstor.org/stable/2241012 (cit. on p. 2).