Bagel : A Fast Bayesian Online Changepoint Detection Algorithm for Linear Models

October 13, 2025

Abstract

We consider Bayesian approaches to online detection of a change-point in linear regression models. Such methods are less common than frequentist methods due to their perceived higher computational cost, but they have advantages in terms of more naturally quantifying uncertainty and the ability to incorporate prior information about the type of change. If conjugate priors are used, we can compute the true posterior distribution for the presence and location of a changepoint with a computational cost that increases linearly over time as the number of possible locations for a change increases. We use the recent idea of merging change locations (Shamp et al., 2021) to reduce this to a linear cost, and propose a novel computationally efficient approach to choosing which locations to merge. Our method applies to a wide-range of problems, including detecting changes in mean or slope, and detecting changes in the presence of seasonal effects. In simulations, the algorithm has a similar speed but higher accuracy compared to a benchmark pruning approach, which only prunes the candidate with the lowest posterior probability. We also demonstrate our approach on the machine temperature failure data.

1 Introduction

Detecting changes in real time is a common problem. This has been well studied in the frequentist world, see for example (Page, 1954, 1955; Aue et al., 2012; Kirch and Weber, 2018; Romano et al., 2023; Ward et al., 2023, 2024; Yu et al., 2020). Whilst these approaches often work well, they do not quantify uncertainty or allow the inclusion of prior information about the type of change. To provide extra information about the change, it is intuitive to consider a Bayesian framework as it can incorporate prior knowledge and produces the posterior distribution which quantifies the uncertainty of the change.

Most Bayesian approaches to the changepoint detection problem consider the offline setting, where we have the whole data and have to infer the location of changepoints or the joint distribution of the number of changepoints and their locations. Existing approaches are either based on the Markov Chain Monte Carlo (Stephens, 1994; Green, 1995; Chib, 1998; Benson and Friel, 2018), or based on the direction simulation (Barry and Hartigan, 1993; Fearnhead, 2006; Fearnhead and Liu, 2011). These approaches have been used for modelling

disease outbreak (Verma et al., 2020), financial volatility (Ross, 2013; Thies and Molnár, 2018), environmental data (Kim and Cheon, 2010) amongst many other applications.

There has also been work on online formulations in models that allow multiple changes. In this case, under the assumption that the parameters specifying the model for data in one segment do not depend on the parameters for other segments, there are efficient algorithms for exactly calculating the posterior. These have a computational cost that increases linearly with time – so processing the observation at time T has an $O(T)$ cost. This cost can be reduced to $O(1)$ by introducing an appropriate approximation (Fearnhead and Liu, 2007).

In this paper, we consider the case of online detection of a single changepoint. We derive efficient algorithms for calculating the posterior distribution in this case under a wider range of models than previous works – for example, including the problem of detecting a change in slope, when the post-change parameters of the model depend on the pre-change parameters. Our general framework encompasses any model that can be written as a change in regression: which allows for detecting changes in mean, slope, slope with seasonality, or splines amongst many others.

The exact algorithm we have has a linear-increasing cost, which makes it unsuitable for long-term or real-time monitoring. To reduce the time complexity, existing approaches include pruning the least probable change-points (Adams and MacKay, 2007) or using resampling to prune the set of potential change-points that are being considered (Fearnhead and Liu, 2007). Instead, we proposed approximations based on the merging idea of Shamp et al. (2021) to reduce the cost per-iteration to constant. Our merging procedure takes account of the difference in the conditional distribution of the parameters given the change location into account, and is amenable to the wider range of change-point models we consider.

The outline of the paper is as follows. In Section 2 we define the Bayesian real-time change-point detection problem and introduce our proposed algorithm with two examples. Section 3 and Section 4 show the performance of our algorithm on simulated data and a real dataset. More examples, and all proofs, can be found in the Appendix. Throughout this paper, all vectors are assumed to be column vectors unless explicitly stated otherwise. We denote by $1:t$ the column vector $(1, \cdots, t)^\top$, $1_{1:t}$ and $0_{1:t}$ are column vectors of length t of ones and zeros respectively, and $y_{1:t}$ represents the column vector $(y_1, \cdots, y_t)^\top$. The probability density function (pdf) of the normal distribution with mean μ and variance σ^2 is denoted by $N(x; \mu, \sigma^2)$, the pdf of the Student-t distribution with ν degrees of freedom, location l, and scale e is denoted by $t_\nu(x; l, e)$, and the pdf of the Inverse-Gamma distribution with shape ν and scale ι is denoted by $IG(x; \nu, \iota)$. For a matrix A, we use $A_{i:j \times k:l}$ to denote the sub-matrix formed by rows i to j and columns k to l.

2 Univariate real-time Bayesian changepoint detection

2.1 The changepoint problem

2.1.1 The model

Assume an independent data stream $y_1, y_2, \ldots, y_t$ is observed in real-time, we wish to detect the presence, or not, of a changepoint at each time step. More precisely, at a given

time t, we wish to detect whether there has been a change prior to t under an appropriate model for the data. We will consider a class of linear models that encompasses the standard change-in-mean and change-in-slope problem, as well as more complex models that allow for incorporating seasonality or autocorrelation.

Let τ be the unknown time of the changepoint, where $\tau = \infty$ denotes no change. The observed data y_t at any time t is modeled as:

$$y_t = a_t^\top \beta + b_{t,\tau}^\top \gamma + \sigma \epsilon_t.$$

Here a_t and $b_{t,\tau}$ are known d_1 and d_2 dimensional vectors, while β and γ are unknown d_1 and d_2 dimensional parameters. The term $a_t^\top \beta$ gives the mean of the model if there has been no change prior to time t, while $b_{t,\tau}^\top \gamma$ specifies the change in this mean due to a changepoint. To be consistent with this we have $b_{t,\tau}^\top = 0_{1:d_2}$ if $\tau \geq t$. Finally, ϵ_t, is the realisation of a standard Gaussian random variable, and $\sigma > 0$ is the standard deviation of the noise in the model.

Consequently, at any time t, the model for the data up to time t, $y_{1:t}$, can be written as:

$$y_{1:t} = A_{1:t}\beta + B_{1:t,\tau}\gamma + \sigma \epsilon_{1:t},$$

where $\epsilon_{1:t}$ is a column vector of i.i.d Gaussian $N(0,1)$ realisations, and the tth rows of matrices $A_{1:t}$ and $B_{1:t,\tau}$ are $a_t^\top$ and $b_{t,\tau}^\top$ respectively. If $\tau \geq t$ then $B_{1:t,\tau}$ is the zero matrix. Since τ is unknown, at each time t we consider t potential models: one corresponding to no change and others where change occurs at $\tau = 1, 2, ..., t-1$. Our task is to determine whether there is significant evidence supporting the presence of a change and to infer the posterior distribution of the location of any change.

2.1.2 The prior for the model parameters

The likelihood of the observation depends on the parameters (β, γ), if the variance of the observations is known, or (β, γ, σ) otherwise. We take a Bayesian approach and assume a prior for these parameters. Furthermore, to enable efficient calculation of the posterior for the presence and location of a changepoint we will use conjugate priors.

For our model, the conjugate priors are Gaussian for (β, γ) conditional on σ, with, if it is unknown, an independent inverse gamma prior for σ. We will have fixed priors for β (and if unknown, σ), but we will allow the conditional prior for γ to potentially depend on both τ and β. Specifically, the joint prior for (β, γ) given σ and τ is

$$\beta, \gamma | \sigma^2, \tau \sim N(\mu^\tau, \sigma^2 \Sigma^\tau)$$

where mean $\mu^\tau \in \mathbb{R}^{d_1+d_2}$ and scaled covariance matrix $\Sigma^\tau \in \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}$. If σ is unknown, the prior is given by:

$$\sigma^2 \sim IG(\nu, \iota).$$

The constraint on the having a fixed prior for β given σ means that $\mu^\tau_{1:d_1}$ and $\Sigma^\tau_{1:d_1 \times 1:d_1}$, that is the marginal prior mean and variance of β, are constant as we vary τ.

While it is often possible to use improper priors for β and σ - the conditional prior for γ must be proper in order to have a well-defined posterior probability for whether there is a changepoint prior to any given time t. The choice of the prior for γ will affect the power of detecting different sizes and types of change.

Here we give two examples to illustrate our model framework:

Example 1. Change-in-mean model.
We first consider the change-in-mean model. This can be written as :

$$A_{1:t} = 1_{1:t}, B_{1:t,\tau} = \begin{bmatrix} 0_{1:\tau} \\ 1_{\tau+1:t} \end{bmatrix}$$

Here β is the mean before the change, γ represents the shift in the mean after the changepoint. There are two natural choices for the prior distribution in this setting: one where the pre-change mean and the change in mean are independent, and another where the pre-change and post-change means are independent.

For the former case we have $\mu = \begin{bmatrix} \mu_1 \\ 0 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix}$. For the latter case we have the same prior mean, but the covariance structure changes to

$$\Sigma = \begin{bmatrix} \delta_1 & -\delta_1 \\ -\delta_1 & 2\delta_1 \end{bmatrix}.$$

See Appendix A for the derivation. This framework can also be extended to the unknown variance case by incorporating an inverse Gamma prior to the variance.

Example 2. Change-in-slope models.
Next, we consider the change-in-slope with a known variance case. The model can be written as:

$$A_{1:t} = [1_{t\times 1} \quad 1:t], B_{1:t,\tau} = \begin{bmatrix} 0_{\tau\times 1} & 0_{\tau\times 1} \\ 1_{t\times 1} & (\tau+1):t \end{bmatrix}$$

Here β defines the linear trend before the change, and γ specifies the shift in the intercept or slope at the changepoint.

There are two natural types of change we can allow in this model, depending on whether or not we require continuity of the trend line at the change-point, as illustrated in Figure 1. If we wish to detect a continuous chance, also known as a change-in-slope (Fearnhead et al., 2019), then a natural model is that the change-in-slope, γ_2, has mean 0 and is independent of the pre-change trend. The distribution of γ_1 conditional on γ_2 is then deterministic due to the continuity constraint. This constraint implies for a change at τ that the trend is continuous at τ: $\gamma_1 + \tau\gamma_2 = 0$. This gives a prior for (β, γ) that has mean and covariance defined by the parameters:

$$\mu^\tau = (\mu_1, \mu_2, 0, 0)^\top, \text{ and } \quad \Sigma^\tau = \begin{bmatrix} \delta_1 & 0 & 0 & 0 \\ 0 & \delta_2 & 0 & 0 \\ 0 & 0 & \tau^2\delta_3 & -\tau\delta_3 \\ 0 & 0 & -\tau\delta_3 & \delta_3 \end{bmatrix}.$$

For a discontinuous change there are two natural models for the type of change. One is that

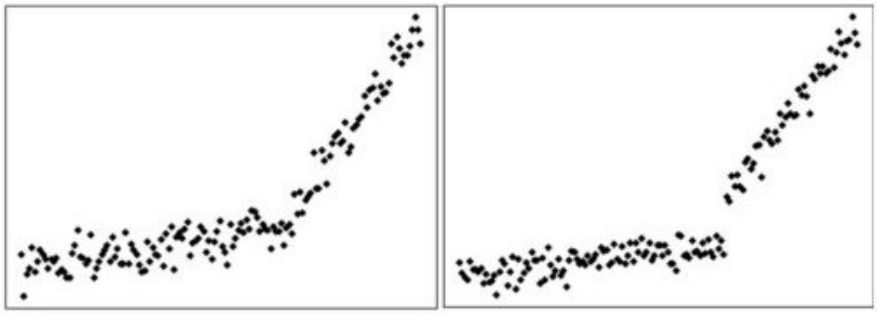


Figure 1: An example of continuous change (left) and discontinuous change (right) in linear trend model.

the change in the intercept and the slope are both mean 0 and independent of the pre-change model, in which case $\mu^\tau = \begin{bmatrix} \mu_1 \\ \mu_2 \\ 0 \\ 0 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \delta_1 & 0 & 0 & 0 \\ 0 & \delta_2 & 0 & 0 \\ 0 & 0 & \delta_3 & \frac{1}{2\tau}(\delta_1 - \delta_3 - \tau^2\delta_4) \\ 0 & 0 & \frac{1}{2\tau}(\delta_1 - \delta_3 - \tau^2\delta_4) & \delta_4 \end{bmatrix}$.
The other would be to assume that the distribution of the trend line starting at time 0 is independent of and has the same distribution of the trend line starting at time τ. This requires the value of the mean function at τ, $\beta_1 + \gamma_1 + \tau(\beta_2 + \gamma_2)$ has the same distribution as β_1, and the new slope $(\beta_2 + \gamma_2)$ has the same distribution as β_2. This implies a prior with

$$\mu^\tau = (\mu_1, \mu_2, -\mu_2\tau, 0)^\top, \text{ and } \quad \Sigma^\tau = \begin{bmatrix} \delta_1 & 0 & -\delta_1 & 0 \\ 0 & \delta_2 & 0 & -\delta_2 \\ -\delta_1 & 0 & 2\delta_1 + \tau^2\delta_2 & -\tau\delta_2 \\ 0 & -\delta_2 & -\tau\delta_2 & 2\delta_2 \end{bmatrix}.$$

For fuller details see Appendix A. Again, this framework can also be extended to the unknown variance case by incorporating an inverse Gamma prior to the variance.

2.1.3 Prior for the change-point

In order to calculate the posterior distribution of τ at any time t we will need a prior distribution. Denote the prior probabilities by $\Pr_t(\tau = i)$, for $i = 1, \ldots, t-1$ and $\Pr_t(\tau \geq t)$. The latter probability will encompass the event that a change has not occurred by time t.

One approach to defining the prior at time t is to construct a prior for τ taking values in $1, \ldots$ and $\tau = \infty$ (i.e. no change), and then define $\Pr_t(\tau)$ to be consistent with this. However, this is not necessarily appropriate in an online setting where we are monitoring a data stream and will intervene if we detect a change. In this case, there is information in the fact that we are still monitoring the data stream at time t that should affect our prior for τ. Intuitively, this information would mean that it is less likely for τ to take smaller values, for which there would be more power for detecting a change by t.

Thus we suggest the following approach. First let us define the Bayes factor for a change at

$\tau = i$ against no change $\tau > t$ as

$$BF_t(i) = \frac{p(y_{1:t}|\tau = i)}{p(y_{1:t}|\tau \geq t)},$$

where

$$p(y_{1:t}|\tau) = \int p(y_{1:t}|\tau, \beta, \gamma, \sigma)p(\beta, \gamma|\tau, \sigma)\mathrm{d}\beta\mathrm{d}\gamma,$$

in the case where σ is known, and where we would further integrate out with respect to the prior for σ if it were unknown. The Bayes factor for a change prior to t is obtained by averaging these Bayes factors with respect to the conditional prior for τ assuming a change by time t,

$$BF_t = \sum_{i=1}^{t-1} BF_t(i) \Pr_t(\tau = i|\tau < t).$$

So the posterior probability of a change before time t is

$$p_t(\tau < t|y_{1:t}) = \frac{(1 - \Pr_t(\tau \geq t)BF_t}{\Pr_t(\tau \geq t) + (1 - \Pr_t(\tau \geq t)BF_t}.$$

If we use a monitoring scheme which detects a change if $p_t(\tau < t|y_{1:t}) \geq c_t$ for some constant c_t, this will be equivalent to detecting a change based on the Bayes Factor being greater than a constant, i.e.

$$BF_t \geq \frac{c_t \Pr_t(\tau \geq t)}{(1 - c_t)(1 - \Pr_t(\tau \geq t))}.$$

In the absence of specific prior information, it would be natural to detect a change based on the Bayes Factor being above a threshold that is constant over time. This is easiest to implement with c_t and $\Pr_t(\tau \geq t)$ both being constant. Furthermore, if we take these to be constant, and tune the threshold for detecting a change based on properties of the test under data where there is no change, then this would make our test invariant to the choice of prior probability $\Pr_t(\tau \geq t)$: as choosing a different prior probability would lead to tuning a different threshold such that we were implementing exactly the same test. We take this approach in our simulation study.

The test does depend on our prior for τ conditional on $\tau < t$. It is natural to define this prior as

$$\Pr_t(\tau = t - k|\tau < t) \propto p^k, \text{ for } \quad k = 1, \ldots, t - 1.$$

This is a (truncated geometric) prior for the time since the change. A specific case, which we use in our simulations, is when $p = 1$ and we have a uniform distribution for this conditional distribution of the change location. In this case the Bayes Factor is similar to the Shiryaev-Roberts test statistics which is known to have good properties (Polunchenko and Tartakovsky, 2010).

2.2 Sequential Updating

Recall that we wish to calculate the posterior probability of a change, its location and the parameters of the model at each time t. To do this in a computationally efficient way, we need to update these posterior probabilities and distributions sequentially from time $t-1$ to time t given y_t.

The posterior probability for τ at time t satisfies

$$\Pr_t(\tau|y_{1:t}) \propto \Pr_t(\tau)p(y_{1:t}|\tau).$$

We will derive a recursion for the right-hand side, and define a set of weights $w_{i,t} = \Pr_t(\tau = i)p(y_{1:t}|\tau = i)$, for $i = 1, \ldots, t-1$. To simplify notation we have $w_{0,t} = \Pr_t(\tau \geq t)p(y_{1:t}|\tau \geq t)$. The following result gives the update for these weights in terms of the predictive density for y_t given $y_{1:t-1}$.

Theorem 1. *For $t > 1$,*

$$\begin{aligned} w_{i,t} &= w_{i,t-1}\frac{\Pr_t(\tau = i)}{\Pr_{t-1}(\tau = i)}p(y_t|y_{1:t-1}, \tau = i) \text{ for } i = 1, \ldots, t-2, \text{and } t > 2 \\ w_{0,t} &= w_{0,t-1}\frac{\Pr_t(\tau \geq t)}{\Pr_{t-1}(\tau \geq t-1)}p(y_t|y_{1:t-1}, \tau \geq t), \text{ and} \\ w_{t-1,t} &= w_{0,t-1}\frac{\Pr_t(\tau = t-1)}{Pr_{t-1}(\tau \geq t-1)}p(y_t|y_{1:t-1}, \tau = t-1). \end{aligned}$$

With the weights initialised at $t = 1$ by

$$w_{0,t} = \Pr_t(\tau \geq t)p(y_{1:t}|\tau \geq t).$$

Proof. See Appendix B. □

We can implement these recursions by recursively calculating the posterior of the model parameters, and hence the predictive distribution, for each possible value of τ. The following result gives updates for the posterior for the parameters in the case σ is unknown, but these can be applied in the case σ is known by ignoring the update for the posterior of σ.

As the prior for the post-change parameter can depend on the time of change, in the case of no changepoint we will update only the marginal posterior for β. To introduce notation, at time $t-1$ let the posterior for β and σ given there is no change be

$$p(\beta|y_{1:t-1}, \sigma, \tau \geq t-1) \sim N\left(\mu^{t-1,0}, \sigma^2\Sigma^{t-1,0}\right), \text{ and } \quad p(\sigma^2|\tau \geq t-1) \sim IG(\nu^{t-1,0}, \iota^{t-1,0}),$$

where $\mu^{t-1,0}$ is a d_1 dimensional vector and $\Sigma^{t-1,0}$ is a $d_1 \times d_1$ matrix. Similarly, we denote the posterior given a change at $i = 1, \ldots, t-2$ as

$$p(\beta, \gamma|y_{1:t-1}, \sigma, \tau = i) \sim N\left(\mu^{t-1,i}, \sigma^2\Sigma^{t-1,i}\right), \text{ and } \quad p(\sigma^2|\tau = i) \sim IG(\nu^{t-1,i}, \iota^{t-1,i}),$$

where now $\mu^{t-1,i}$ is a $d_1 + d_2$ dimensional vector and $\Sigma^{t-1,i}$ is a $(d_1 + d_2) \times (d_1 + d_2)$ matrix. Finally, we will need the posterior distribution for (β, γ) and σ assuming a change at time

$t-1$. Using the same notation as above, this can be obtained from the posterior for β and σ given no change prior to $t-1$.

Theorem 2. *Let the prior mean for a change at $t-1$ be $\mu^{t-1} = [\mu_\beta^\top, \mu_\gamma^\top]^\top$, and the prior covariance be*

$$\sigma^2\Sigma^{t-1} = \sigma^2 \begin{bmatrix} \Sigma_{\beta,\beta} & \Sigma_{\beta,\gamma} \\ \Sigma_{\gamma,\beta} & \Sigma_{\gamma,\gamma} \end{bmatrix},$$

so that μ_β denotes the prior mean of β, $\Sigma_{\beta,\beta}$ the prior variance of β, $\Sigma_{\beta,\gamma}$ the prior covariance between β and γ, and so on. Then the distribution of (β, γ, σ) given $y_{1:t-1}$ and a change at $\tau = t-1$ has parameters

$$\mu^{t-1,t-1} = \begin{bmatrix} \mu^{t-1,0} \\ \mu_\gamma + \Sigma_{\gamma,\beta}\Sigma_{\beta,\beta}^{-1}(\mu^{t-1,0} - \mu_\beta) \end{bmatrix},$$

$$\Sigma^{t-1,t-1} = \begin{bmatrix} \Sigma^{t-1,0} & \Sigma_{\gamma,\beta}\Sigma_{\beta,\beta}^{-1}\Sigma^{t-1,0} \\ \Sigma^{t-1,0}\Sigma_{\beta,\beta}^{-1}\Sigma_{\beta,\gamma} & \Sigma_{\gamma,\gamma} + \Sigma_{\gamma,\beta}\Sigma_{\beta,\beta}^{-1}(\Sigma^{t-1,0} - \Sigma_{\beta,\beta})\Sigma_{\beta,\beta}^{-1}\Sigma_{\beta,\gamma} \end{bmatrix}$$

with $\nu^{t-1,t-1} = \nu^{t-1,0}$ and $\iota^{t-1,t-1} = \iota^{t-1,0}$.

Proof. See Appendix B. □

We can update these parameters as follows.

Theorem 3. *Fix iteration $t-1$, define h_i as the following feature vector:*

$$h_i = \begin{cases} [a_{t-1}^\top, b_{t-1,i}^\top]^\top & 0 < i \leq t-2, \\ a_{t-1} & i = 0, \end{cases}$$

For $i = 0, 1, \ldots, t-2$, define $e_i = y_{t-1} - h_i\mu^{t-1,i}$,$Q = h_i^\top\Sigma^{t-1,i}h_i + 1$, $A = \Sigma^{t-1,i}h_i/Q$, The parameter updates for existing changepoint models after observing y_{t-1} follow:

$$\Sigma^{t,i} = \Sigma^{t-1,i} - A^\top AQ, \quad \mu^{t,i} = \mu^{t-1,i} + Ae,$$
$$\nu^{t,i} = \nu^{t-1,i} + \frac{1}{2}, \qquad \iota^{t,i} = \iota^{t-1,i} + \frac{1}{2}e_i^2/Q.$$

Finally, given the posterior for the parameters at time $t-1$, which is also the prior for the models at time t, we have the following calculations for the predictive density.

Theorem 4. *Using the same notation for the parameters of the posteriors as in Theorem 3, the predictive density can be calculated as*

$$P(y_t|y_{1:t-1}, \tau = i) = N\left(y_t; h_i\mu^{t,i}, \sigma^2(1 + h_i^\top\Sigma^{t,i}h_i)\right)$$

if σ is known, and

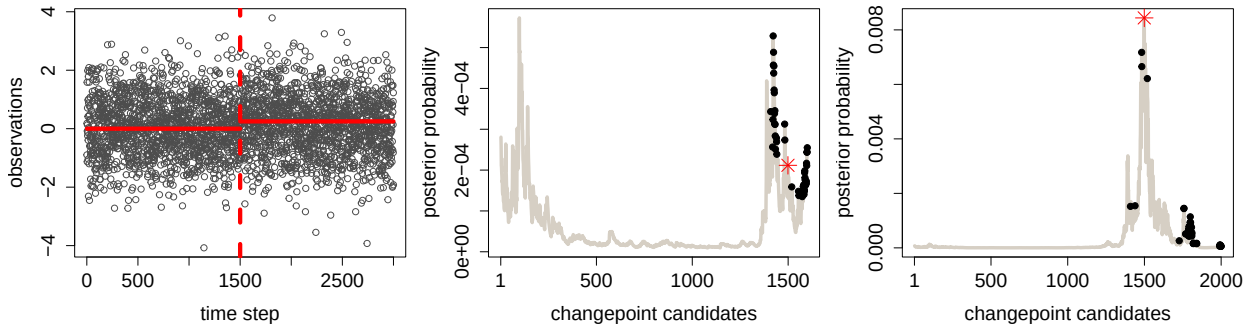
$$P(y_t|y_{1:t-1}, \tau = i) = t_{2\nu^{t,i}}\left(y_t; h_i\mu^{t,i}, \frac{\iota^{t,i}}{\nu^{t,i}}\left(1 + h_i^\top\Sigma^{t,i}h_i\right)\right)$$

if σ is unknown, where $t_\nu(y_t; l, e)$ is the location-scale t distribution with ν degrees of freedom, location l and scale parameter e.

In practice, we can sometimes simplify the updates by considering a suitable reparametrisation of the model to (β, θ) for some θ that is a linear function of β and γ. This can simplify the application of Theorem 2 if the prior for θ is independence of β – for example if we choose $\theta = \gamma + \beta$ in the chance-in-mean example with the independent mean prior. Moreover, if we choose a reparametrisation such that observations after the change depend on θ but not β, then we can simplify the updates for the changepoint cases by only updating the distribution of θ. Examples of such linear transformations are provided in Appendix A.

2.3 Reducing the computational complexity by merging

Implementing the univariate Bayesian changepoint detection methods gives a cost per iteration that increases with iteration t, so that the overall cost of analysing T data points is quadratic in T. To prevent this, we need to approximate the posterior distribution with fewer support points. A common and straightforward approach is to prune out the changepoint candidate with the posterior probability less than ϵ or the most M probable candidates (Adams and MacKay, 2007; Saatçi et al., 2010), then the computational complexity can be reduced to $O(\frac{T}{\epsilon})$ or $O(TM)$. However, such a method does not always work well. First, they reduce the support of the posterior for the change, which can be important if we wish to calculate credible intervals. If different changepoint values have substantially different posteriors for the post-change model then their relative probabilities can change substantially as we observe new data, as shown in Figure 2.



(a) Simulated data (b) At step 1600 (c) At step 2000

Figure 2: Figure (a) presents the simulated data with a change at 1500. Figures (b) and (c) show the posterior distribution obtained from the exact approach without pruning (gray line). The black dots represent the candidates not pruned by the benchmark approach with $M = 50$ at time steps 1600 and 2000, respectively. The red star indicates a candidate that was pruned at step 1600, but subsequently has the highest posterior probability after collecting the data.

2.3.1 Merging criteria

To overcome these issues, we first use the idea of merging the posterior distribution for the parameters for different change locations as a way to reduce the computational cost whilst keeping the full support of the posterior. Second, we take account of the similarity of the

posterior distributions of the parameters when deciding which distributions to merge. This is similar to the approach in Shamp et al. (2021). We will look at the similarity of the posteriors in terms of the parameters, or the function of parameters, that affect the distribution of the data after the change. Denote such parameters as $\theta = g(\beta, \gamma)$, where $g(\cdot)$ is a linear transformation that maps the pre-change parameters (β, γ) to the post-change parameters θ as Examples in Appendix A.

To explain our approach, it is simplest to first consider what happens when we first merge posteriors. Fix the time, t, when this happens, and denote the posterior density function for the post-change parameter as $f_i(\theta)$ when change occurs at i. We have $t-1$ possible values for the location of a change, assuming a change has occurred, and these will have weight and posterior distribution denoted by the pair $\{w_{i,t}, f_i(\theta)\}, 1 \leq i \leq t-1$.

Our merge procedure will replace $f_i(\theta)$ with $f_{i+1}(\theta)$ for some appropriately chosen i. We will choose $i \in \{1, \ldots, t-2\}$ to be the value for which the total variation distance between the posterior for θ and the resulting approximation is minimised. Let $TV(h_1(\theta), h_2(\theta)) = \frac{1}{2}\int |h_1(\theta) - h_2(\theta)| \mathrm{d}\theta$ be the total variation distance between two densities, $h_1(\theta)$ and $h_2(\theta)$. Then if the true posterior is $h_1(\theta) = \sum_{i=1}^{t-1} w_{i,t} f_i(\theta)$, then the approximation after we replace $f_i(\theta)$ with $f_{i+1}(\theta)$ is $h_2(\theta) = h_1(\theta) - w_{i,t}(f_i(\theta) - f_{i+1}(\theta))$, and

$$\begin{aligned} TV\left(h_1(\theta), h_2(\theta)\right) &= \frac{1}{2}\int \left|h_1(\theta) - (h_1(\theta) - w_{i,t}(f_i(\theta) - f_{i+1}(\theta)))\right| \mathrm{d}\theta \\ &= \frac{1}{2}\int \left|w_{i,t}(f_i(\theta) - f_{i+1}(\theta))\right| \mathrm{d}\theta \\ &= w_{i,t} TV(f_i(\theta), f_{i+1}(\theta)). \end{aligned}$$

Thus we will choose to merge component i which minimises

$$w_{i,t} TV(f_i(\theta), f_{i+1}(\theta)). \tag{2.1}$$

This will maintain the same support for the changepoint locations, but will reduce the computational cost. At any time, if we have M distinct posterior densities for θ, then we only need to update these posteriors when we get a new observation. To update the weights associated with a given posterior, we store these as the sum of the weights for that posterior and the relative weight for each change location. The latter will not change, so we only need to update the former weight.

So at any iteration t, after applying merging, we will have consecutive runs of changepoint locations that will have the same posterior for θ. At the current iteration, denote the largest value of such runs by an ordered set $\mathcal{T} = \{j_1, j_2, \ldots, j_M\}$. Let $w_{i,t}$ for $i = 0, \ldots, M$ denote the set of weights such that for $i \in \{1, \ldots, M\}$, $w_{i,t}$ is the sum of the weights associated with change locations $\{j_{i-1}+1, \ldots, j_i\}$, with $j_0 = 0$. Also we store a set of M ratio vectors $\mathcal{R} = \{r^{(1)}, \ldots, r^{(M)}\}$ which denote the proportion of the weight w_i associated with each changepoint location in the run from $j_{i-1}+1$ to j_i. That the vector of weights associated with these candidate locations will be $w_i r^{(i)}$, or equivalently the vector of posterior probabilities for change locations would be

$$(w_0, w_1 r^{(1)}, \ldots, w_M r^{(M)}),$$

with, after normalisation, the first component being the probability of no change, and the remaining components being the probability of a change at $1, \ldots, t = 1$.

The idea is that we can update this representation by just updating the weights, and the parameters associated with each distinct posterior – which has a computational cost proportional to M. At each iteration we will also add a new candidate location for a change, and to keep the number of distinct posteriors as M we will need to merge a pair of distinct posteriors for θ if $t > M$. As before let $f_i(\theta)$ denote the posterior associated with the ith run of changepoint locations. Using the same criteria on minimising the approximation based on the Total Variation distance, we will replace $f_i(\theta)$ with $f_{i+1}(\theta)$ for the value $i \geq 1$ which minimises $w_i TV(f_i(\theta), f_{i+1}(\theta)$. Algorithm 1 describes one iteration of the algorithm.

Algorithm 1: one iteration of Bagel

Input: existing candidates set $\mathcal{T} = \{0, i_1, \ldots, i_N\}$, set of ratio vectors $\{r^{(1)}, \ldots, r^{(N)}\}$, weights $\{w_{i,t-1}\}$ and parameters for the posterior distributions Θ_i for $i \in \{0, \ldots, N\}$.

Data: Observe y_t at time t

Introducing new changepoint candidate

Calculate Θ_{N+1}, the parameters of the posterior distribution given $\tau = t - 1$ using Theorem 2.

$\mathcal{T} \longleftarrow \mathcal{T} \cup \{t - 1\}$

Updating weights and posterior distributions

for $i \in \{0, \ldots, N + 1\}$ **do**

  Calculate $w_{i,t}$ based on Theorems 1 and 4.

Update parameters of posterior distribution, $\Theta_0, \ldots, \Theta_{N+1}$ using Theorem 3.

Normalize $w_{i,t}$ for $i \in \{0, \ldots, N + 1\}$.

Decide if there is a changepoint

if $1 - w_{0,t} > c$ **then**

  output: Change detected; $\mathcal{T}, \{r^{(1)}, \ldots, r^{(N+1)}\}, \{\Theta_0, \ldots, \Theta_{N+1}\}, w_{0:N+1,t}$

Merging step

if $N = M$ **then**

  Search the index i that gives the minimum error (2.1) for $i \in \{1, \ldots, N\}$.

  Update the ratio vector $\{r^{(i+1)}\} \longleftarrow (w_{i,t} + w_{i+1,t})^{-1}(w_{i,t}r^{(i)}, w_{i+1,t}r^{(i+1)})$.

  Combine $w_{i+1,t} \longleftarrow w_{i,t} + w_{i+1,t}$.

  Remove candidate i, Θ_i, $r^{(i)}$ and $w_{i,t}$.

  $N \longleftarrow N - 1$

output: $\mathcal{T}, \{r^{(1)}, \ldots, r^{(N+1)}\}, \{\Theta_0, \ldots, \Theta_{N+1}\}, w_{0:N+1,t}$

2.3.2 Calculating the total variation distance

Now we will show how to calculate the total variation, which is needed for our merging step. This can be calculated exactly if θ is 1-dimensional, but we need to resort to an approximation in higher dimensions.

One dimensional parameter

Calculating the total variation between two univariate distributions, such as in the change-in-mean case with known variance, is straightforward. In this case, θ will have a univariate Gaussian distribution, and we use the following result for the total variation distance between two univariate Gaussians.

Proposition 1. *The total variation between two univariate Gaussian distributions with*

means μ_i and μ_{i+1}, and variances $\sigma^2\Sigma_i$ and $\sigma^2\Sigma_{i+1}$ respectively is:

$$2\Big[\Phi(a\sqrt{\sigma^2\Sigma_i^2}+b\sqrt{\sigma^2\Sigma_{i+1}^2})-\Phi(a\sqrt{\sigma^2\Sigma_i^2}-b\sqrt{\sigma^2\Sigma_{i+1}^2}) \\ +\Phi(a\sqrt{\sigma^2\Sigma_{i+1}^2}-b\sqrt{\sigma^2\Sigma_i^2})-\Phi(a\sqrt{\sigma^2\Sigma_{i+1}^2}+b\sqrt{\sigma^2\Sigma_i^2})\Big].$$

where $a=\frac{\mu_i-\mu_{i+1}}{\sigma^2\Sigma_{i+1}^2-\sigma^2\Sigma_i^2}$ *and* $b=\frac{\sqrt{(\mu_i-\mu_{i+1})^2+(\sigma^2\Sigma_{i+1}^2-\sigma^2\Sigma_i^2)\log\frac{\sigma^2\Sigma_{i+1}^2}{\sigma^2\Sigma_i^2}}}{\sigma^2\Sigma_{i+1}^2-\sigma^2\Sigma_i^2}$.

Proof. See proof in Appendix C. □

Multi-dimensional parameters

In higher dimensions, it is often intractable to find the exact closed form for the total variation between two distributions. Instead, we use the following tractable bounds as approximations to the total variation distance.

For any pair of distributions for distributions θ over R^d with densities $f_i(\cdot)$ and $f_{i+1}(\cdot)$, Pinsker's inequality (Pinsker, 1964) states that

$$TV(f_i(\theta),f_{i+1}(\theta))\le\frac{1}{\sqrt{2}}\sqrt{\mathrm{KL}(f_i(\theta)\parallel f_{i+1}(\theta))}=\frac{1}{\sqrt{2}}\sqrt{\int f_i(\theta)\log\frac{f_i(\theta)}{f_{i+1}(\theta)}d\theta},$$

where KL denotes the Kullback–Leibler divergence. The KL distance for our posteriors, which will be either multivariate Gaussian or Normal Inverse Gamma, is given by the following results.

Theorem 5. *(Williams and Rasmussen, 2006) The KL between two multivariate Gaussian distributions with means μ_i and μ_{i+1}, and variances $\sigma^2\Sigma_i$ and $\sigma^2\Sigma_{i+1}$ respectively, is:*

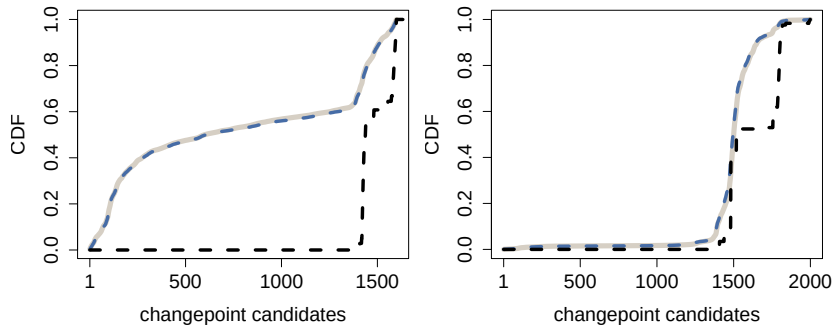
$$D_{KL}(N(\mu_i,\sigma^2\Sigma_i)||N(\mu_{i+1},\sigma^2\Sigma_{i+1}))= \\ \frac{1}{2}\left(tr(\Sigma_{i+1}^{-1}\Sigma_i-I)+\frac{1}{\sigma^2}(\mu_{i+1}-\mu_i)^\top\Sigma_{i+1}^{-1}(\mu_{i+1}-\mu_i)+\log\det(\Sigma_i^{-1}\Sigma_{i+1})\right)$$

Theorem 6. *(Soch and Allefeld, 2016) The KL distance between two normal inverse gamma distributions with means μ_i and μ_{i+1}, scaling variance Σ_i and Σ_{i+1}, shape ν_i and ν_{i+1}, scale ι_i and ι_{i+1} respectively is*

$$D_{KL}[NIG(\mu_i,\Sigma_i,\nu_i,\iota_i)||NIG(\mu_{i+1},\Sigma_{i+1},\nu_{i+1},\iota_{i+1})]= \\ \frac{1}{2}\left[(\mu_{i+1}-\mu_i)^\top\Sigma_j^{-1}(\mu_{i+1}-\mu_i)\frac{\nu_i}{\iota_i}+tr(\Sigma_{i+1}^{-1}\Sigma_i-I)+\log\det(\Sigma_i^{-1}\Sigma_{i+1})\right] \\ +\nu_{i+1}\log\frac{\iota_i}{\iota_{i+1}}-\log\frac{\Gamma(\nu_i)}{\Gamma(\nu_{i+1})}+(\nu_i-\nu_{i+1})\psi(\nu_i)-(\iota_i-\iota_{i+1})\frac{\nu_i}{\iota_i},$$

where $\Gamma(\cdot)$ is the gamma function and $\psi(\cdot)$ is the digamma function.

Using the KL bounds, we revisit the example in Figure 2. Figure 3 presents the cumulative density function (CDF) of the exact approach, our recovered posterior, and the benchmark. It demonstrates that the recovered posterior from Bagel closely matches the exact distribution.



(a) At step 1600 (b) At step 2000

Figure 3: The CDFs of the exact approach (grey), Bagel with recovered posterior distribution (blue) and benchmark (black) when $M = 50$.

3 Simulation Results

We conduct a simulation study to evaluate the performance of different pruning methods, focusing on detection power and computational efficiency on the above examples. For simplicity, we assume the changepoint locations are uniformly distributed for changepoint models. For the change-in-mean case, we reparametrize the model in terms of the pre-change parameter (β) and post-change parameters ($\theta = \beta + \gamma$), and place independent priors on them. A Normal prior is used when the variance is known, and a Normal-Inverse-Gamma prior when it is unknown. For detecting changes in slope, we allow for both continuous and discontinuous changes, assigning equal prior probabilities to the two types of change. Therefore we consider three possible models: no-change, continuous change, and discontinuous change, with bivariate normal prior on (β, γ) as specified in Section 2.1.2. We can then implement the recursion and pruning rules of Section 2 for the two change models separately, and get an approximation to the posterior of whether there is a change, and if so where and of what type it is. The estimated model is selected as the one with the highest marginal posterior probability. Further prior details are provided in Appendix D.1. Once the parameters are set, we simulate 2000 data points without change. The detection thresholds are chosen to control the false alarm rate, defined as the probability of falsely detecting a change, at 5%.

To evaluate the detection power of algorithms, we measure the following indices:

- Coverage rate: the proportion in which the true changepoint falls within the estimated 95% highest posterior density intervals.
- Error rate of the algorithm: calculated as $\frac{\text{missed alarms + false alarms}}{\text{replications}}$.
- Average detection delay and its standard deviation: the detection delay measures the difference between the estimated true positive stopping time and the true changepoint location.
- MAP estimator and its standard deviation: the *maximum a posteriori* estimation gives the estimation of the changepoint location with the largest posterior probability.
- Speed: the average running time at each time step.

Table 1 and Table 2 present the simulation results for the two examples, using KL bounds when total variation needs to be approximated. In Example 1, the detection power of Bagel is close to that of the exact approach, while the benchmark consistently shows the lowest

M	Algorithm	Known variance				Unknown variance			
		Coverage	Error	Delay	MAP	Coverage	Error	Delay	MAP
	Exact	0.95	0.03	293±196	991±168	0.95	0.06	295 ± 189	1025 ± 106
50	Bagel	0.95	0.04	283±176	1010±108	0.95	0.06	299 ± 192	1022 ± 105
	Benchmark	0.38	0.08	308±198	1009±116	0.36	0.09	326 ± 213	1025 ± 106
100	Bagel	0.95	0.05	282±175	1009±107	0.95	0.06	296 ± 190	1024 ± 106
	Benchmark	0.55	0.06	300±195	1012±102	0.56	0.08	317 ± 207	1023 ± 104
200	Bagel	0.95	0.05	281±175	1009±109	0.94	0.06	295 ± 190	1024 ± 106
	Benchmark	0.74	0.05	294±186	1011±109	0.74	0.07	306 ± 202	1026 ± 111

Table 1: Simulation results for the Example 1 change-in-mean case with 500 replicates. The simulated data follows $N(0, 1)$ before $\tau = 1000$ and $N(0.25, 1)$ after the change.

M	Algorithm	Example 2 with continuous change				Example 2 with discontinuous change			
		Coverage	Error	Delay	MAP	Coverage	Error	Delay	MAP
	Exact	0.87	0.05	329 ± 73	1037 ± 190	0.93	0.05	198 ± 68	959 ± 145
50	Bagel	0.86	0.05	336 ± 76	1038 ± 152	0.93	0.05	205 ± 70	949 ± 137
	Benchmark	0.00	0.05	399 ± 96	1199 ± 101	0.05	0.05	253 ± 89	1062 ± 75
100	Bagel	0.88	0.05	331 ± 74	1042 ± 150	0.92	0.05	199 ± 69	954 ± 106
	Benchmark	0.00	0.05	373 ± 90	1140 ± 106	0.22	0.05	231 ± 80	1017 ± 83
200	Bagel	0.87	0.05	329 ± 74	1036 ± 159	0.92	0.05	198 ± 68	949 ± 116
	Benchmark	0.05	0.05	353 ± 83	1082 ± 136	0.55	0.05	214 ± 75	978 ± 105

Table 2: Simulation results for Example 2 change-in-slope scenario with known variance with 500 replicates. The simulated data follow a normal distribution $N(0, 1)$ before time $t = 1000$, and follow $-1.75 + 0.002t + N(0, 1)$ after the change for continuous change and follow $-2.2 + 0.002t + N(0, 1)$ after the change for discontinuous change.

coverage rate, highest error rate, and longest detection delay. In Example 2, Bagel achieves a higher correct change type rate than the exact method, which is reasonable given its longer detection delay. The benchmark algorithm again performs the worst and fails to identify the type of change. In terms of computational speed, Bagel and the benchmark are comparable and significantly faster than the exact approach, as shown in Appendix D.3. Results with different prior settings and different bounds approximation are provided in Appendix D.2.

4 Real data example - Machine Temperature Failure

The Numenta Anomaly Benchmark (NAB) contains a collection of datasets for evaluating real-time anomaly detection algorithms (Ahmad et al., 2017). We evaluate our proposed approaches on one of the datasets - machine temperature failure dataset. This dataset records temperature readings from a heat sensor, and has been used in several algorithm evaluations, such as those in Fisch et al. (2022); Ahmad and Purdy (2016); Ahmad et al. (2017); Jesmeen et al. (2021), among others. It contains 22695 observations from 02/12/2013 to 19/02/2014, with data sampled every 5 minutes. Following the data preparation procedure in Lavin and Ahmad (2015); Fisch et al. (2022), we use the first 15% to be training data. We preprocess the data by removing autocorrelation components and normalising with the median and the median absolute deviation estimated from the training set. To reduce the impact of extreme values, we remove outliers with modified Z-scores exceeding an absolute value of 3.5, as suggested in Iglewicz and Hoaglin (1993). The threshold for detection is empirically

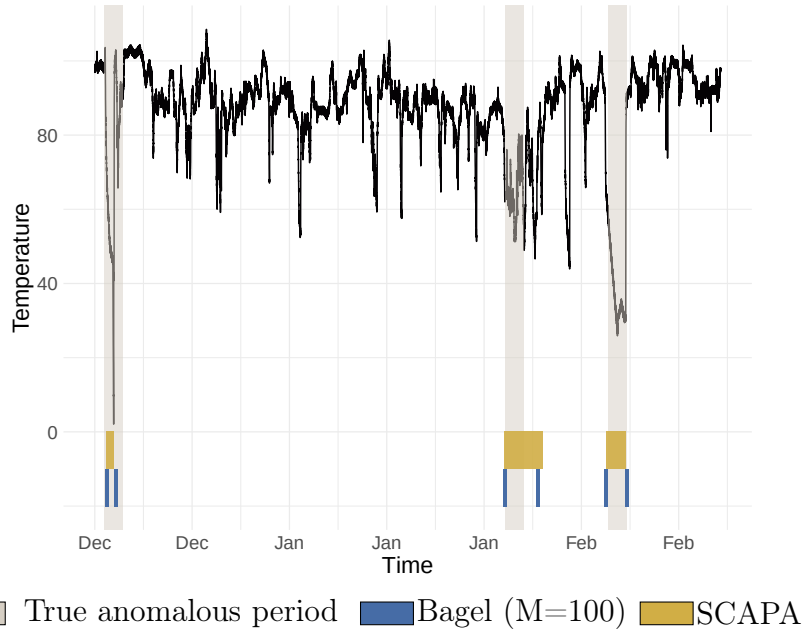


Figure 4: Machine temperature data with true anomalous period, MAP estimators of change points obtained from our proposed method, and the anomalous segment detected by SCAPA

Year	Start time	End time	SCAPA	Bagel (M=100)
2013	17:50 15/12	17:00 17/12	16:50 16/12	03:00 16/12
2014	14:20 27/01	13:30 29/01	21:25 28/01	07:15 28/01
2014	14:55 07/02	14:05 09/02	03:15 08/02	00:45 08/02

Table 3: Labelled anomalies along with the detection time of two approaches.

determined on simulated pseudo time series by resampling the block of observations (Politis and Romano, 1994). We perform 200 Monte Carlo replications to ensure that the false alarm rate remains 5% on 5000 observations.

We focus on the mean shift while also incorporating an additional autoregressive term into the model. When Bagel detects a change, the algorithm restarts and is initialised using the MAP estimators from the posterior of the previous segment as its prior. Figure 4 and Table 3 present the detection results of Bagel with $M = 100$, alongside the results of SCAPA as reported in their original paper Fisch et al. (2022). Although SCAPA is proposed to detect the anomalous segment, we can treat the start and the end of the segment as two changepoints for comparison. Two algorithms give similar changepoint estimates, while Bagel is faster in detection. Moreover, the estimators from both methods outperform most of the results reported in Ahmad and Purdy (2016); Jesmeen et al. (2021), where changepoints were either missed or falsely triggered.

5 Discussion

We proposed a fast online Bayesian changepoint detection algorithm for identifying changes in linear models. The resulting method provides uncertainty estimates for both the change and its location through the recovered posterior distribution. Moreover, the algorithm maintains constant time complexity per time step, empirically at a millisecond level, making it suitable for real-time applications. Future work includes extending the method to multivariate settings, where the dependencies exist across the dimensions, and changes may propagate across dimensions.

References

Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Ahmad, S., Lavin, A., Purdy, S., and Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147.

Ahmad, S. and Purdy, S. (2016). Real-time anomaly detection for streaming analytics. *arXiv preprint arXiv:1607.02480*.

Aue, A., Horváth, L., Kühn, M., and Steinebach, J. (2012). On the reaction time of moving sum detectors. *Journal of Statistical Planning and Inference*, 142(8):2271–2288.

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319.

Benson, A. and Friel, N. (2018). Adaptive MCMC for multiple changepoint analysis with applications to large datasets. *Electronic Journal of Statistics*, 12(2):3365 – 3396.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16:203–213.

Fearnhead, P. and Liu, Z. (2007). Online inference for multiple changepoint problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):589–605.

Fearnhead, P. and Liu, Z. (2011). Efficient Bayesian analysis of multiple changepoint models with dependence across segments. *Statistics and Computing*, 21:217–229.

Fearnhead, P., Maidstone, R., and Letchford, A. (2019). Detecting changes in slope with an l 0 penalty. *Journal of Computational and Graphical Statistics*, 28(2):265–275.

Fisch, A. T., Bardwell, L., and Eckley, I. A. (2022). Real time anomaly detection and categorisation. *Statistics and Computing*, 32(4):55.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Iglewicz, B. and Hoaglin, D. C. (1993). *Volume 16: how to detect and handle outliers*. Quality Press.

Jesmeen, M., Hossen, J., and Aziz, A. B. A. (2021). Unsupervised anomaly detection for energy consumption in time series using clustering approach. *Emerging Science Journal*, 5(6):840–854.

Kim, J. and Cheon, S. (2010). Bayesian multiple change-point estimation with annealing stochastic approximation monte carlo. *Computational Statistics*, 25(2):215–239.

Kirch, C. and Weber, S. (2018). Modified sequential change point procedures based on estimating functions. *Electronfic Journal of Statistics*, 12(1):1579–1613.

Lavin, A. and Ahmad, S. (2015). Evaluating real-time anomaly detection algorithms–the numenta anomaly benchmark. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*, pages 38–44. IEEE.

Mersmann, O. (2024). *microbenchmark: Accurate Timing Functions*. R package version 1.4.10.

Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

Pinsker, M. S. (1964). Information and information stability of random variables and processes. *Holden-Day*.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313.

Polunchenko, A. S. and Tartakovsky, A. G. (2010). On optimality of the shiryaev–roberts procedure for detecting a change in distribution. *The Annals of Statistics*, 38:3445–3457.

Romano, G., Eckley, I. A., Fearnhead, P., and Rigaill, G. (2023). Fast online changepoint detection via functional pruning cusum statistics. *Journal of Machine Learning Research*, 24:1–36.

Ross, G. J. (2013). Modelling financial volatility in the presence of abrupt changes. *Physica A: Statistical Mechanics and its Applications*, 392(2):350–360.

Saatçi, Y., Turner, R. D., and Rasmussen, C. E. (2010). Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 927–934.

Shamp, W., Varbanov, R., Chicken, E., Linero, A., and Yang, Y. (2021). Computationally efficient Bayesian sequential function monitoring. *Journal of Quality Technology*, 54(1):1–19.

Soch, J. and Allefeld, C. (2016). Kullback-leibler divergence for the normal-gamma distribution. *arXiv preprint arXiv:1611.01437*.

Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):159–178.

Thies, S. and Molnár, P. (2018). Bayesian change point analysis of bitcoin returns. *Finance Research Letters*, 27:223–227.

Verma, B. K., Verma, M., Verma, V. K., Abdullah, R. B., Nath, D. C., Khan, H. T., Verma, A., Vishwakarma, R. K., and Verma, V. (2020). Global lockdown: An effective safeguard in responding to the threat of covid-19. *Journal of evaluation in clinical practice*, 26(6):1592–1598.

Ward, K., Dilillo, G., Eckley, I., and Fearnhead, P. (2023). Poisson-FOCuS: An efficient online method for detecting count bursts with application to gamma ray burst detection. *Journal of the American Statistical Association*, pages 1–13.

Ward, K., Romano, G., Eckley, I., and Fearnhead, P. (2024). A constant-per-iteration likelihood ratio test for online changepoint detection for exponential family models. *Statistics and Computing*, 34(3):1–11.

West, M. and Harrison, J. (2006). *Bayesian forecasting and dynamic models*. Springer Science & Business Media.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Yu, Y., Padilla, O. H. M., Wang, D., and Rinaldo, A. (2020). A note on online change point detection. *arXiv preprint arXiv:2006.03283*.

A Properties of the Prior for Examples 1 and 2.

Example 1

To see that the second prior for Example 1 gives the same distribution for the pre-change and post-change mean, and these are independent, we apply a change of variable. We have $(\beta, \gamma)^\top$ has a Gaussian distribution with mean $\begin{bmatrix} \mu_1 \\ 0 \end{bmatrix}$ and variance $\sigma^2 \begin{bmatrix} \delta_1 & -\delta_1 \\ -\delta_1 & 2\delta_1 \end{bmatrix}$.

Let the pre-change and post-change means be $(\beta, \theta)^\top$ then

$$\begin{bmatrix} \beta \\ \theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix}.$$

This is a linear transformation, so $(\beta, \theta)^\top$ will also have a Gaussian distributions with mean

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_1 \end{bmatrix},$$

and variance

$$\sigma^2 \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \delta_1 & -\delta_1 \\ -\delta_1 & 2\delta_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^\top = \sigma^2 \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_1 \end{bmatrix},$$

as required.

Example 2

In the change-in-slope case, for a discontinuous change, denote the pre-change and post-change means of intercept and slope as $(\beta_1, \beta_2, \theta_1, \theta_2)^\top$, we have

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & \tau & 1 & \tau \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \gamma_1 \\ \gamma_2 \end{bmatrix}.$$

With this linear transformation, we have

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & \tau & 1 & \tau \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ -\mu_2\tau \\ 0 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_1 \\ \mu_2 \end{bmatrix},$$

and variance

$$\sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & \tau & 1 & \tau \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_1 & 0 & -\delta_1 & 0 \\ 0 & \delta_2 & 0 & -\delta_2 \\ -\delta_1 & 0 & 2\delta_1 + \tau^2\delta_2 & -\tau\delta_2 \\ 0 & -\delta_2 & -\tau\delta_2 & 2\delta_2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & \tau & 1 & \tau \\ 0 & 1 & 0 & 1 \end{bmatrix}^\top = \sigma^2 \begin{bmatrix} \delta_1 & 0 & 0 & 0 \\ 0 & \delta_2 & 0 & 0 \\ 0 & 0 & \delta_1 & 0 \\ 0 & 0 & 0 & \delta_2 \end{bmatrix},$$

as required. So the pre-change and post-change parameters are independent and follow the same distribution.

As for the continuous change, we have

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tau \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \gamma_1 \\ \gamma_2 \end{bmatrix}.$$

So the $(\theta_1, \theta_2, \theta_3, \theta_4)^\top$ follows the multivariate Gaussian distribution with mean

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tau \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ 0 \\ \mu_2 \end{bmatrix}$$

and variance

$$\sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tau \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_1 & 0 & 0 & 0 \\ 0 & \delta_2 & 0 & 0 \\ 0 & 0 & \tau^2\delta_3 & -\tau\delta_3 \\ 0 & 0 & -\tau\delta_3 & \delta_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tau \\ 0 & 1 & 0 & 1 \end{bmatrix}^\top = \sigma^2 \begin{bmatrix} \delta_1 & 0 & 0 & 0 \\ 0 & \delta_2 & 0 & \delta_2 \\ 0 & 0 & 0 & 0 \\ 0 & \delta_2 & 0 & \delta_2 + \delta_3 \end{bmatrix},$$

where the new intercept has mean 0 and variance 0, and the post-change slope is dependent on the pre-change slope.

B Proof of the sequential updating

Proof of Theorem 1

For $i = 1, \ldots, t-2$ the recursion comes from the definition of the weights.

$$\begin{aligned} w_{i,t} &= \Pr_t(\tau = i)p(y_{1:t}|\tau = i) \\ &= \Pr_{t-1}(\tau = i)p(y_{1:t-1}|\tau = i)\frac{\Pr_t(\tau = i)}{\Pr_{t-1}(\tau = i)}p(y_t|y_{1:t-1}, \tau = i) \\ &= w_{i,t-1}\frac{\Pr_t(\tau = i)}{\Pr_{t-1}(\tau = i)})p(y_t|y_{1:t-1}, \tau = i). \end{aligned}$$

A similar simple argument applies for $i = 0$. Finally for $i = t-1$ we have

$$\begin{aligned} w_{t-1,t} &= \Pr_t(\tau = t-1)p(y_{1:t}|\tau = t-1) \\ &= \Pr_{t-1}(\tau \geq t-1)p(y_{1:t-1}|\tau \geq t-1)\frac{\Pr_t(\tau = i)}{\Pr_{t-1}(\tau \geq t-1)}p(y_t|y_{1:t-1}, \tau = t-1) \\ &= w_{0,t-1}\frac{\Pr_t(\tau = i)}{\Pr_{t-1}(\tau \geq t-1)})p(y_t|y_{1:t-1}, \tau = t-1). \end{aligned}$$

Proof of Theorem 2

If we have a change at $t-1$ then the prior for (β, γ) has mean μ^{t-1} and variance $\sigma^2\Sigma^{t-1}$. Our assumption is that the marginal prior for β is the same as that assuming no change, but the conditional prior for γ given β can depend on the change location $t-1$.

The posterior distribution given $y_{1:t-1}$, conditional on a change at $t-1$ is

$$\begin{aligned} &p(\beta, \gamma, \sigma|y_{1:t-1}, \tau = t-1) \\ &\quad\propto p(\sigma)p(\beta|\sigma)p(\gamma|\beta, \sigma, \tau = t-1)p(y_{1:t-1}|\beta, \gamma, \sigma, \tau = t-1) \\ &\quad= p(\sigma)p(\beta|\sigma)p(\gamma|\beta, \sigma, \tau = t-1)p(y_{1:t-1}|\beta, \sigma, \tau = t-1) \\ &\quad\propto p(\beta, \sigma|y_{1:t-1}, \tau = t-1)p(\gamma|\beta, \sigma, \tau = t-1) \\ &\quad= p(\beta, \sigma|y_{1:t-1}, \tau \geq t-1)p(\gamma|\beta, \sigma, \tau = t-1). \end{aligned}$$

Here we have used the fact that the observations prior to the change do not depend on γ, and that as the prior for β and σ and the likelihood for $y_{1:t-1}$ are the same for $\tau = t-1$ and

$\tau \geq t-1$ so are the posteriors.

The posterior for (β, σ) given $y_{1:t-1}$ and $\tau \geq t-1$ is calculated at time $t-1$ for the case of no change by $t-1$ and conditional on σ the posterior for β is Gaussian with mean $\mu^{t-1,0}$ and variance $\Sigma^{t-1,0}$.

As $p(\beta, \gamma|\sigma, \tau = t-1)$ is multivariate Gaussian, the conditional prior $p(\gamma|\beta, \sigma, \tau = t-1)$ is also Gaussian, with a mean that is linear in β. Using standard results for the multivariate Gaussian gives that this conditional distribution is of the form

$$\gamma|\beta, \sigma, \tau = t-1 \sim \mathrm{N}(\mu_\gamma + \Sigma_{\gamma,\beta}\Sigma^{-1}_{\beta,\beta}(\beta - \mu_\beta), \sigma^2(\Sigma_{\gamma,\gamma} - \Sigma_{\gamma,\beta}\Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\gamma})),$$

Therefore, the marginal mean of γ after updating β is given by

$$E(\gamma) = E\left(E(\gamma \mid \beta)\right) = \mu_\gamma + \Sigma_{\gamma,\beta}\Sigma^{-1}_{\beta,\beta}(\mu^{t-1,0} - \mu_\beta).$$

The marginal scaled variance is

$$\begin{aligned}
\mathrm{Var}(\gamma) &= E\left(\mathrm{Var}(\gamma \mid \beta)\right) + \mathrm{Var}\left(E(\gamma \mid \beta)\right) \\
&= \sigma^2\left(\Sigma_{\gamma,\gamma} - \Sigma_{\gamma,\beta}\Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\gamma} + \Sigma_{\gamma,\beta}\Sigma^{-1}_{\beta,\beta}\Sigma^{t-1,0}\Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\gamma}\right) \\
&= \sigma^2\left(\Sigma_{\gamma,\gamma} + \Sigma_{\gamma,\beta}\Sigma^{-1}_{\beta,\beta}(\Sigma^{t-1,0} - \Sigma_{\beta,\beta})\Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\gamma}\right).
\end{aligned}$$

The covariance can be calculated as

$$\mathrm{Cov}(\beta, \gamma) = \mathrm{Cov}\left(\beta, \Sigma_{\gamma,\beta}\Sigma^{-1}_{\beta,\beta}\beta\right) = \Sigma_{\gamma,\beta}\Sigma^{-1}_{\beta,\beta}\Sigma^{t-1,0}.$$

Thus, the full updated joint distribution of (β, γ) is

$$N\left(\begin{bmatrix}\mu^{t-1,0} \\ \mu_\gamma + \Sigma_{\gamma,\beta}\Sigma^{-1}_{\beta,\beta}(\mu^{t-1,0} - \mu_\beta)\end{bmatrix}, \begin{bmatrix}\Sigma^{t-1,0} & \Sigma_{\gamma,\beta}\Sigma^{-1}_{\beta,\beta}\Sigma^{t-1,0} \\ \Sigma^{t-1,0}\Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\gamma} & \Sigma_{\gamma,\gamma} + \Sigma_{\gamma,\beta}\Sigma^{-1}_{\beta,\beta}(\Sigma^{t-1,0} - \Sigma_{\beta,\beta})\Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\gamma}\end{bmatrix}\right).$$

as required.

Proof of Theorems 3 and 4

The results in Theorems 3 and 4 follow from standard results. One way to see this is to view our model as a dynamic linear model (West and Harrison (2006)), where the hidden state is the parameter, but we define the dynamics such that the hidden state does not change over time.

For the case of no change, the dynamic linear model is

$$\begin{aligned}
\beta_t &= \beta_{t-1} \\
y_t &= h_0^T\beta_t + \sigma\epsilon_t,
\end{aligned}$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$, and h_0 is the known regression vector. At time $t-1$, given observations $y_{1:t-1}$, the posterior distribution is:

$$\beta_{t-1} \mid \sigma^2, y_{1:t-1}, \tau \geq t-1 \sim \mathcal{N}(\mu^{t-1,0}, \sigma^2\Sigma^{t-1,0}),$$

$$\sigma^2 \mid y_{1:t-1}, \tau \geq t-1 \sim IG(\nu^{t-1,0}, \iota^{t-1,0}).$$

After observing y_t, the posterior distribution of (β, σ^2) is updated using the following formulas as show in West and Harrison (2006); Fearnhead and Liu (2011):

$$\begin{aligned}
\Sigma^{t,0} &= \Sigma^{t-1,0} - A^\top A Q, \\
\mu^{t,0} &= \mu^{t-1,0} + A e_0, \\
\nu^{t,0} &= \nu^{t-1,0} + \frac{1}{2}, \\
\iota^{t,0} &= \iota^{t-1,0} + \frac{e_0^2}{2Q},
\end{aligned}$$

where $e_0 = y_t - h_0^\top \mu^{t-1,0}$ is the one-step-ahead forecast error, $Q = h_0^\top \Sigma^{t-1,0} h_0 + 1$ is the forecast error variance and $A = \Sigma^{t-1,0} h_0 / Q$ is the adaptive gain. Then the one-step-ahead predictive distribution of y_t given $y_{1:t-1}$ is:

$$y_t \mid y_{1:t-1}, \tau \geq t-1 \sim t_{2\nu^{t-1,0}} \left(h_0^\top \mu^{t-1,0}, \frac{\iota^{t-1,0}}{\nu^{t-1,0}} \left(h_0^\top \Sigma^{t-1,0} h_0 + 1 \right) \right),$$

where $t_{2\nu^{t-1,0}}$ denotes the Student-t distribution with $2\nu^{t-1,0}$ degrees of freedom.

A similar derivation works for the case where there is a change at time i, but now the dynamic linear model is

$$\begin{aligned}
(\beta_t^\top, \gamma_t^\top)^\top &= (\beta_{t-1}^\top, \gamma_{t-1}^\top)^\top \\
y_t &= h_i^T (\beta_t^\top, \gamma_t^\top)^\top + \sigma \epsilon_t.
\end{aligned}$$

The model follows the same dynamic linear structures, but with an augmented parameter vector. As such, the posterior and predictive quantities follow immediately by applying the standard DLM update formulas introduced earlier.

C The total variation between two univariate Gaussian with known variance

Corollary 1. *The posterior distributions for neighbouring $f_i(\mu, \sigma^2\Sigma^2)$ and $f_j(\mu, \sigma^2\Sigma^2)$ always intersects at two points as $\sigma^2\Sigma_i^2 > \sigma^2\Sigma_j^2$. Two points of intersection are:*

$$(c_1, c_2) = \frac{\mu_i \sigma^2 \Sigma_j^2 - \mu_j \sigma^2 \Sigma_i^2 \mp \sqrt{\sigma^2\Sigma_i^2 \sigma^2 \Sigma_j^2} \sqrt{(\mu_i - \mu_j)^2 + (\sigma^2\Sigma_j^2 - \sigma^2\Sigma_i^2) \log \frac{\sigma^2\Sigma_j^2}{\sigma^2\Sigma_i^2}}}{\sigma^2\Sigma_j^2 - \sigma^2\Sigma_i^2}.$$

Proposition 2. *The total variation between two normal distributions $f_i(\mu, \sigma^2\Sigma^2)$ and $f_j(\mu, \sigma^2\Sigma^2)$ is*

$$2 \left[\Phi(\frac{c_2 - \mu_i}{\sqrt{\sigma^2\Sigma_i^2}}) - \Phi(\frac{c_1 - \mu_i}{\sqrt{\sigma^2\Sigma_i^2}}) + \Phi(\frac{c_1 - \mu_j}{\sqrt{\sigma^2\Sigma_j^2}}) - \Phi(\frac{c_2 - \mu_j}{\sqrt{\sigma^2\Sigma_j^2}}) \right]. \tag{C.1}$$

Let $a = \frac{\mu_i - \mu_j}{\sigma^2\Sigma_j^2 - \sigma^2\Sigma_i^2}$ and $b = \frac{\sqrt{(\mu_i-\mu_j)^2 + (\sigma^2\Sigma_j^2 - \sigma^2\Sigma_i^2)\log\frac{\sigma^2\Sigma_j^2}{\sigma^2\Sigma_i^2}}}{\sigma^2\Sigma_j^2 - \sigma^2\Sigma_i^2}$, Formula C.1 could be rewritten as:

$$2\left[\Phi(a\sqrt{\sigma^2\Sigma_i^2} + b\sqrt{\sigma^2\Sigma_j^2}) - \Phi(a\sqrt{\sigma^2\Sigma_i^2} - b\sqrt{\sigma^2\Sigma_j^2}) + \Phi(a\sqrt{\sigma^2\Sigma_j^2} - b\sqrt{\sigma^2\Sigma_i^2}) - \Phi(a\sqrt{\sigma^2\Sigma_j^2} + b\sqrt{\sigma^2\Sigma_i^2})\right].$$

D Simulation results under different priors

D.1 Priors

For the change-in-mean case, we assume the probability of no change $Pr(\tau \geq 2000) = 0.9$, and the distribution of the changepoint location follows a uniform distribution $Pr(\tau = i | \tau < 2000) \propto 1$ for $i = 1, \cdots, 1999$. To simplify the updating procedure, we consider a model where the pre-change and post-change means are independent, through reparametrization. Specifically, we model the pre-change parameter (β) and post-change parameters ($\theta = \beta + \gamma$). If $\sigma^2 = 1$ is know, we have normal prior with parameters $(\mu, \sigma^2\Sigma)$; or normal-inverse-gamma prior with parameters $(\mu, \Sigma, \nu, \iota)$ if it is unknown. Table 4 and Table 5 in the Appendix present the detection power under different priors. In the main text, Table 1 shows the results under a Normal prior with parameters $(\mu = 0, \sigma^2\Sigma = 0.25^2)$ when the variance is known, and under a Normal-Inverse-Gamma prior with parameters $(\mu = 0, \Sigma = 10, \nu = 30, \iota = 30)$ when the variance is unknown.

For the change-in-slope case with known variance, we allow the model to detect the type of change, either continuous or discontinuous, assigning equal prior probabilities $\Pr^{\text{conti}}(\tau < 2000) = \Pr^{\text{dis}}(\tau < 2000) = 0.1$. The parameters before and after the change, (β, γ), are jointly modelled with bivariate normal priors with hyper-parameters (μ_1, μ_2) and $(\delta_1, \delta_2, \delta_3, \delta_4)$ as specified in Section 2.1.2. Different prior settings are considered as shown in Table 7 and 6. In the main text, Table 2 reports the results for $\mu_1 = \mu_2 = 0$ and $\delta_1 = \delta_2 = \delta_3 = \delta_4 = 1$.

The detection thresholds for each algorithm are chosen to control the false alarm rate at 5%, defined as the probability of detecting a change before time 2000 when no change.

D.2 Detection power

Table 4, Table 5, Table 7 and Table 6 present the simulation results for each example under different prior choices. Overall, the results are consistent with those reported in Section 3.

D.3 Speed

Here, we calculate the average computational speed at each step of each approach shown in Section 3. The main code is written in R, while the merging part is written in C++. The code was executed in a virtualised environment using VMware with full virtualisation, on a machine powered by an Intel(R) Xeon(R) Gold 6248R CPU, featuring 4 physical cores operating at 3.00 GHz. The time is recorded through the R package "microbenchmark" (Mersmann, 2024) which can accurately measure and compare the execution time of R expressions. From Figure 5, we can see that the exact approach is more expensive than the pruned approaches.

Prior	M	Algorithm	Coverage	Error	Delay	Map
$N(0, 0.125^2)$		Exact	0.97	0.02	276±168	968±144
	50	Bagel	0.97	0.03	267±155	974±116
		Benchmark	0.29	0.09	332±201	982±153
	100	Bagel	0.97	0.03	268±155	973±123
		Benchmark	0.47	0.06	305±183	982±127
	200	Bagel	0.97	0.03	269±155	975±118
		Benchmark	0.69	0.05	281±162	985±101
$N(0, 0.25^2)$		Exact	0.95	0.03	293±196	991±168
	50	Bagel	0.95	0.04	283±176	1010±108
		Benchmark	0.38	0.08	308±198	1009±116
	100	Bagel	0.95	0.05	282±175	1009±107
		Benchmark	0.55	0.06	300±195	1012±102
	200	Bagel	0.95	0.05	281±175	1009±109
		Benchmark	0.74	0.05	294±186	1011±109
$N(0, 0.5^2)$		Exact	0.90	0.04	337±231	985±221
	50	Bagel	0.91	0.08	314±196	1024±109
		Benchmark	0.41	0.10	332±209	1027±94
	100	Bagel	0.91	0.08	312±194	1023±105
		Benchmark	0.57	0.09	327±208	1029±109
	200	Bagel	0.91	0.08	311±194	1023±106
		Benchmark	0.74	0.09	319±200	1028±116
$N(0, 1^2)$		Exact	0.88	0.05	380±253	981±252
	50	Bagel	0.89	0.10	348±216	1034±113
		Benchmark	0.40	0.14	358±219	1031±97
	100	Bagel	0.89	0.10	347±214	1034±113
		Benchmark	0.58	0.13	346±210	1031±96
	200	Bagel	0.88	0.09	347±215	1033±114
		Benchmark	0.72	0.11	349±215	1031±101

Table 4: Simulation results for Example 1 (change-in-mean scenario with known variance) are based on 500 replicates when we vary the value of the variance. The simulated data follow a normal distribution $N(0, 1)$ before time $t = 1000$, and then follow $N(0.25, 1)$ after the change. The priors are specified as $p(\tau \geq 2000) = 0.9$ and $p = 0.1$. The prior column in the tables specifies that the distribution of pre-change and post-change, as they are independently and identically distributed.

Prior	M	Algorithm	Coverage	Error	Delay	MAP
$NIG(0, 0.1, 30, 30)$		Exact	0.95	0.06	295 ± 189	1025 ±106
	50	Bagel	0.95	0.06	299 ± 192	1022 ± 105
		Benchmark	0.36	0.09	326 ± 213	1025 ± 106
	100	Bagel	0.95	0.06	296 ± 190	1024 ± 106
		Benchmark	0.56	0.08	317 ± 207	1023 ± 104
	200	Bagel	0.94	0.06	295 ± 190	1024 ± 106
		Benchmark	0.74	0.07	306 ± 202	1026 ± 111
$NIG(0, 10, 5, 5)$		Exact	0.87	0.19	409 ± 232	1036 ± 102
	50	Bagel	0.87	0.19	411 ± 232	1037 ± 105
		Benchmark	0.32	0.24	441 ± 250	1040 ± 102
	100	Bagel	0.88	0.18	410 ± 232	1036 ± 103
		Benchmark	0.57	0.22	420 ± 238	1038 ± 105
	200	Bagel	0.88	0.18	411 ± 234	1035 ± 102
		Benchmark	0.75	0.20	413 ± 236	1040 ± 114
$NIG(0, 100, 2.1, 2.1)$		Exact	0.87	0.23	465 ± 246	1040 ± 119
	50	Bagel	0.88	0.24	463 ± 245	1041 ± 119
		Benchmark	0.30	0.32	469 ± 244	1047 ± 121
	100	Bagel	0.87	0.23	466 ± 246	1040 ± 119
		Benchmark	0.54	0.28	465 ± 243	1038 ± 108
	200	Bagel	0.87	0.23	467 ± 247	1040 ± 119
		Benchmark	0.75	0.25	466 ± 247	1040 ± 118

Table 5: Simulation results for Example 1 (change-in-mean scenario with unknown variance) are based on 500 replicates when we vary the value of the scaled covariance matrix and the prior on the variance. The simulated data follow a normal distribution $N(0, 1)$ before time $t = 1000$, and then follow $N(0.25, 1)$ after the change. The priors are specified as $p(\tau \geq 2000) = 0.9$ and $p = 0.1$. The prior column in the tables specifies that the distribution of pre-change and post-change, as they are independently and identically distributed.

Prior	M	Algorithm	Coverage	Error	Delay	MAP
$\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0.125^2$		Exact	0.86	0.05	304 ± 71	1045 ± 156
	50	Bagel	0.88	0.05	307 ± 73	1048 ± 139
		Benchmark	0.00	0.05	392 ± 104	1216 ± 111
	100	Bagel	0.87	0.05	305 ± 72	1047 ± 145
		Benchmark	0.01	0.05	350 ± 89	1146 ± 96
	200	Bagel	0.87	0.05	304 ± 71	1045 ± 144
		Benchmark	0.05	0.05	325 ± 79	1088 ± 121
$\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0.25^2$		Exact	0.86	0.05	318 ± 73	1033 ± 187
	50	Bagel	0.87	0.05	323 ± 73	1040 ± 155
		Benchmark	0.00	0.05	401 ± 105	1206 ± 114
	100	Bagel	0.86	0.05	319 ± 73	1039 ± 153
		Benchmark	0.01	0.05	364 ± 91	1139 ± 110
	200	Bagel	0.86	0.05	318 ± 73	1031 ± 163
		Benchmark	0.05	0.05	341 ± 82	1079 ± 131
$\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0.5^2$		Exact	0.87	0.05	329 ± 73	1037 ± 190
	50	Bagel	0.86	0.05	336 ± 76	1038 ± 152
		Benchmark	0.00	0.05	399 ± 96	1199 ± 101
	100	Bagel	0.88	0.05	331 ± 74	1042 ± 150
		Benchmark	0.00	0.05	373 ± 90	1140 ± 106
	200	Bagel	0.87	0.05	329 ± 74	1036 ± 159
		Benchmark	0.05	0.05	353 ± 83	1082 ± 136
$\delta_1 = \delta_2 = \delta_3 = \delta_4 = 1$		Exact	0.86	0.05	341 ± 75	1039 ± 167
	50	Bagel	0.75	0.05	350 ± 78	1042 ± 149
		Benchmark	0.00	0.05	414 ± 90	1203 ± 95
	100	Bagel	0.88	0.05	344 ± 75	1039 ± 140
		Benchmark	0.00	0.05	389 ± 89	1144 ± 106
	200	Bagel	0.88	0.05	342 ± 75	1039 ± 139
		Benchmark	0.05	0.05	367 ± 86	1080 ± 134

Table 6: Simulation results for Example 2 (continuous change-in-slope scenario with known variance) are based on 500 replicates when we vary the value of the scaled covariance matrix. The simulated data follow a normal distribution $N(0, 1)$ before time $t = 1000$, and then undergo a continuous change, gradually shifting to follow $0.002 \times t + N(0, 1)$ after the change. For the known variance setting, the priors are specified as $p(\tau \geq 2000) = 0.8$, $p_{\text{dis}} = 0.1$, $p_{\text{conti}} = 0.1$, $\sigma^2 = 1$ and $\mu_\beta = (0, 0)^\top$.

Prior	M	Algorithm	Coverage	Error	Delay	MAP
$\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0.125^2$		Exact	0.93	0.05	171 ± 63	965 ± 128
	50	Bagel	0.94	0.05	174 ± 64	964 ± 111
		Benchmark	0.07	0.05	246 ± 98	1076 ± 85
	100	Bagel	0.94	0.05	172 ± 64	967 ± 104
		Benchmark	0.29	0.05	206 ± 81	1022 ± 74
	200	Bagel	0.93	0.05	172 ± 63	962 ± 125
		Benchmark	0.63	0.05	186 ± 71	984 ± 92
$\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0.25^2$		Exact	0.94	0.05	186 ± 66	953 ± 144
	50	Bagel	0.95	0.05	191 ± 68	958 ± 109
		Benchmark	0.08	0.05	254 ± 99	1071 ± 89
	100	Bagel	0.94	0.05	188 ± 67	958 ± 109
		Benchmark	0.26	0.05	223 ± 80	1016 ± 79
	200	Bagel	0.94	0.05	187 ± 66	952 ± 123
		Benchmark	0.61	0.05	202 ± 75	971 ± 108
$\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0.5^2$		Exact	0.93	0.05	198 ± 68	959 ± 145
	50	Bagel	0.93	0.05	205 ± 70	949 ± 137
		Benchmark	0.05	0.05	253 ± 89	1062 ± 75
	100	Bagel	0.92	0.05	199 ± 69	954 ± 106
		Benchmark	0.22	0.05	231 ± 80	1017 ± 83
	200	Bagel	0.92	0.05	198 ± 68	949 ± 116
		Benchmark	0.55	0.05	214 ± 75	978 ± 105
$\delta_1 = \delta_2 = \delta_3 = \delta_4 = 1$		Exact	0.92	0.05	211 ± 69	959 ± 129
	50	Bagel	0.91	0.05	223 ± 75	953 ± 164
		Benchmark	0.04	0.05	267 ± 89	1065 ± 72
	100	Bagel	0.91	0.05	212 ± 70	952 ± 107
		Benchmark	0.21	0.05	247 ± 82	1021 ± 78
	200	Bagel	0.90	0.05	211 ± 69	952 ± 104
		Benchmark	0.52	0.05	228 ± 77	978 ± 103

Table 7: Simulation results for Example 2 (discontinuous change-in-slope scenario with known variance) are based on 500 replicates when we vary the value of the scaled covariance matrix. The simulated data follow a normal distribution $N(0, 1)$ before time $t = 1000$, and then undergo a continuous change, gradually shifting to follow $-1.75 + 0.002 \times t + N(0, 1)$ after the change. For the known variance setting, the priors are specified as $p(\tau \geq 2000) = 0.8$, $p_{\text{dis}} = 0.1$, $p_{\text{conti}} = 0.1$, $\sigma^2 = 1$ and $\mu_\beta = (0, 0)^\top$.

The running time of our proposed approach is close to that of the benchmark approach at each time step.

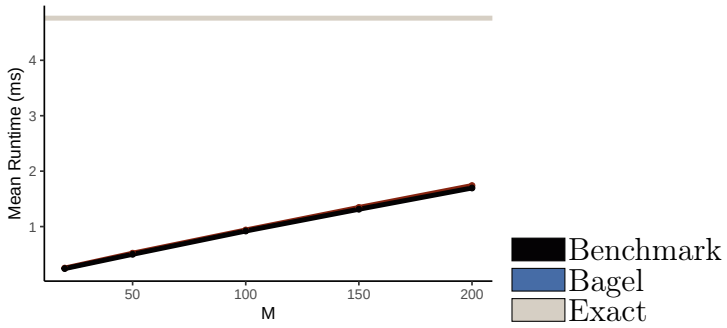


Figure 5: Average running speed per time step for exact, Bagel , and benchmark approaches against different values of M on data $n = 1000$.