Decoupling Trends and Changepoint Analysis

Haoxuan Wu, Sean Ryan and David S. Matteson *
Department of Statistics and Data Science, Cornell University
January 19, 2022

Abstract

This paper introduces a new Bayesian changepoint approach called the decoupled approach that separates the process of modeling and changepoint analysis. The approach utilizes a Bayesian dynamic linear model (DLM) for the modeling step and a weighted penalized likelihood estimator on the posterior of the Bayesian DLM to identify changepoints. A Bayesian DLM, with shrinkage priors, can provide smooth estimates of the underlying trend in presence of complex noise components; however, the inability to shrink exactly to zero make changepoint analysis difficult. Penalized likelihood estimators can be effective in estimating location of changepoints; however, they require a relatively smooth estimate of the data. The decoupled approach combines the flexibility of the Bayesian DLM along with the hard thresholding property of penalized likelihood estimator to extend application of changepoint analysis. The approach provides a robust framework that allows for identification of changepoints in highly complex Bayesian models. The approach can identify changes in mean, higher order trends and regression coefficients. We illustrate the approach’s flexibility and robustness by comparing against several alternative methods in a wide range of simulations and two real world examples.

1 Introduction

In time series analysis, Bayesian dynamic linear models (DLMs) track the observed variable as function of latent variables that evolve linear over-time (Chan and Eisenstat, 2018). As the parameters in this type of model are allowed to change over time, the model provides great flexibility for dealing with many complexities of time series data and can be utilized in a wide variety of settings. For examples, Bayesian dynamic linear models have seen applications in economics (Nakajima et al., 2011), climate analysis (Pichuka and Maity, 2018) and financial markets (Guidolin et al., 2019).

*The authors gratefully acknowledge financial support from the National Science Foundation 1455172, 1934985, 1940124, 1940276, 2114143, and USAID 7200AA18CA00014.
To ensure sufficient flexibility, Bayesian DLMs tend to be heavily overparameterised models (with at least one parameter per data point). As a result, they are typically applied alongside shrinkage priors which produce smoother and more reliable estimates for the dynamic features by “shrinking” small changes closer to zero. Various forms of shrinkage priors such as the horseshoe prior (Carvalho et al., 2009), double-gamma prior (Bitto and Fruhwirth-Schnatter, 2019), double exponential prior (Park and Casella, 2008) and dynamic horseshoe (Kowal et al., 2019) have shown to be effective. When used in conjunction with shrinkage priors, DLMs are excellent for capturing changes that occur smoothly over time. However, since the priors do not shrink parameters exactly to zero, they often struggle to capture sudden breaks, such as changepoints.

Within the changepoint literature, many different approaches have been taken for identifying these sudden breaks (Aminikhanghahi and Cook, 2017). Likelihood ratio tests using various forms cumulative statistics have been utilized to select the optimal partitioning (Jeske et al., 2009; Fryzlewicz, 2014). However, these methods typically rely on the data fitting a pre-specified parametric model and lack flexibility to adapt to more complex datasets. Penalized likelihood approaches have been utilized to optimize a cost function balancing trade-off between goodness-of-fit and number of changepoints (Killick et al., 2012; Bai and Perron, 2002; Maidstone et al., 2017). Other approaches such as non-parametric distance based metrics (Matteson and James, 2014; Harchaoui et al., 2008), Bayesian probabilistic modeling (Erdman and Emerson, 2008; Adams and MacKay, 2007) and subspace identification (Liu et al., 2013) have also been utilized to identify changepoints. While these methods have shown to be effective on well-behaved time series, they tend to struggle when we combine stationary and non-stationary models.

Due to the discrete nature of the changepoint problem, estimating changepoints in a Bayesian setting is very challenging. A number of authors develop changepoint models for which it is possible to generate I.I.D. samples from the posterior (Fearnhead, 2006; Fearnhead and Liu, 2011); however, these methods require relatively simple models and struggle to take account of stochastic volatility or heavy tailed errors. More complex models can be estimated via Markov chain based sampling such as reversible jump MCMC (Green, 1995) and Metropolis Hastings (Green, 2003); however, these methods are known to struggle to converge in practice (Eckley et al., 2011). More recently, Ko et al. (2015) investigates the use of hidden markov models with Dirichlet priors for changepoint analysis; however, the authors only consider settings with 1 or 2 clear changepoints. Furthermore even in settings where the posterior can be easily sampled, it can be challenging to interpret the resulting uncertainty estimates since the uncertainty applies to both the changepoint estimates and the number of changepoints (Siems et al., 2019). Thus inference is only possible if we condition on the number of changepoints.

Given the complementary strengths of Bayesian DLMs and changepoint models, it is natural to ask whether these methods can be combined in some way. In this paper, we propose a Bayesian method called the decoupled approach that allows us to identify changepoints in any Bayesian dynamic linear model.
The decoupled approach consists of two steps. First, a Bayesian dynamic linear model will be utilized to separate the true signal of the data from the noise components. We do not specify a particular structure for the dynamic linear model but rather will show the approach works for a wide range of structures. Second, a regularized loss will be imposed on the posterior of the model to estimate the changepoints.

The decoupled approach follows from the method developed by Hahn and Carvalho (2015) which separates the processes of modeling and inference. In their work, the authors focused on the problem of variable selection in linear regression. For the modeling step, a Bayesian linear regression model is fit to estimate the coefficients associated with a set of predictors. For the inference step, a penalized loss function is placed on posterior mean of the coefficient estimates to select a sparse set of predictors that can be predictive of the response variable. In essence, the paper introduces a two step Bayesian framework for sparse model summary. Since then, this approach has been shown to be effective in a variety of other contexts such as sparse graph estimation (Bashir et al., 2019), portfolio optimization (Puelz et al., 2015) and Bayesian function-on-scalar regression (Kowal and Bourgeois, 2020). In this paper, we extend the decoupled approach to Bayesian dynamic linear models in order to select the optimal changepoints.

The decoupled approach provides two key advantages. First, the decoupled approach separates the process of Bayesian modeling from changepoint analysis. As we will illustrate in Section 2, only the trend component of the Bayesian model will be used to identify changepoints. As a result, we can fit a highly complex Bayesian model to deal with the intricacies of the data such as outliers, heterogeneity and seasonality. Most existing changepoint algorithms struggle to deal with these components as they tend to significant skew the distribution of the data and violate the Gaussian assumption. By separating the processes of trend-filtering and changepoint selection, we can create a highly robust framework that can deal with all these issues. Second, by using a regularized loss on the posterior of the Bayesian model, the decoupled approach is able to provide uncertainty estimates for the number of changepoints selected. The loss is marginalized over all values of the penalty parameter and MCMC draws from the posterior can be used to estimate uncertainty. In turn, the decoupled approach can provide more insights into the selection process and the trade-off between goodness-of-fit and number of changepoints.

The paper proceeds as follows. In Section 2, we introduce the decoupled approach for identifying changepoints in Bayesian dynamic linear models, define diagnostic tools for selection of optimal number of changepoints and detail extensions to covariates/multiple predictors. We also discuss the benefits of using shrinkage priors in Bayesian DLMs. Section 3 illustrates the effectiveness of the decoupled approach in diverse sets of simulation scenarios. We will compare the decoupled approach with a variety of other changepoint methods for both change in mean and change in regression coefficients. The robustness of the decoupled approach will be shown through examples with outliers and heteroskedasticity. Section 4 further illustrates the effectiveness of the decoupled approach through two world datasets. We conclude

3
with a discussion of the overall framework and its key benefits.

2 Methodology

2.1 Decoupled Modeling

To introduce the decoupled approach, we start by utilizing a standard Bayesian state-space model. The model later will be extended to incorporate shrinkage priors to produce smoother estimates of the underlying trend. Suppose we are given a time series \( Y = (y_1, \ldots, y_n)' \) and with a predictor series \( X = (x_1, \ldots, x_n)' \) a Bayesian state-space model can be formulated as follows:

\[
\begin{align*}
    y_t &= x_t \beta_t + \epsilon_t, \\
    \Delta^D \beta_t &= \omega_t
\end{align*}
\]

\[\epsilon_t \sim N(0, \sigma^2_{\epsilon,t}) \quad \omega_t \sim N(0, \sigma^2_{\omega}) \quad (1)\]

In this setup, \{\beta_t\} models with the time-varying relationship between the predictor series \{x_t\} and the response series \{y_t\}. As an example, to estimate changes in mean for \( Y \), we can set \( x_t = 1 \) for \( t = 1, \ldots, n \). The process \{\epsilon_t\} deals with the noise; \{\epsilon_t\} is modeled as a heteroskedastic process to give additional flexibility. For now, we will assume only one predictor series. Later on, we will extend the framework to deal with multiple predictors.

The primary goal of the Bayesian DLM in this step is to estimate the process \{\beta_t\} to understand the time-varying relationship between the predictor and the response. We assume the process \{\beta_t\} is a relatively smooth process with large jumps which will be denoted as changepoints. For well-behaved series, the random walk process allows us to estimate smooth functions for \{\beta_t\}. However, the random walk process has no way of estimating discrete processes such as changepoints. As a result, the decoupled approach utilizes penalized loss as a secondary step to estimate the changepoint. The requirement for the penalized loss in the second step is a smooth estimation of \{\beta_t\}. As we will shown later on, for more noisy series, shrinkage priors may be necessary to induce this smooth estimation.

We fit the above model to the observed data via Gibbs sampling and retrieve posterior estimates for the latent parameters. Let \( \beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(M)} \) denote the M MCMC draws of \{\beta_t\} from the Bayesian model where \( \beta^{(i)} = (\beta_1^{(i)}, \ldots, \beta_n^{(i)})' \) and let \( \bar{\beta} \) denote the posterior mean of the M draws. Since the Bayesian DLM models heteroskedastic variance, we will start with a general weighted least squares loss function as follows:

\[
L_\lambda(\bar{\beta}) = \|W^{1/2}(Y - X \circ \bar{\beta})\|_2^2 + q_\lambda(\bar{\beta})
\]

where \( \circ \) denote element-wise product, \( q_\lambda(\cdot) \) is a penalty function to induce sparsity in the estimator \( \bar{\beta} \) and \( W = \text{diag}(w_1, \ldots, w_n) \) is a diagonal matrix with weights for each measurement. Since the penalty term \( q_\lambda(\bar{\beta}) \) and \( W \) won’t change throughout the derivation process, we will discuss their details after...
the derivation. Next, following derivation seen in Hahn and Carvalho (2015), we take integral of the loss with respect to the conditional distribution of \( Y \) on the parameters \( \beta, \sigma^2_{\epsilon}, \sigma^2_{\omega} \) and drop constant terms w.r.t. \( \tilde{\beta} \).

\[
E[L_\lambda(\tilde{\beta}, Y)] = ||W^{1/2}(X \circ \beta) - W^{1/2}(X \circ \tilde{\beta})||_2^2 - 2(X \circ \tilde{\beta})'W\epsilon + q_\lambda(\tilde{\beta})
\]

Next, an integration over the posterior of \( p(\beta, \sigma^2_{\epsilon}, \sigma^2_{\omega} | Y) \) results in:

\[
E[L_\lambda(\tilde{\beta}, \beta, \sigma^2_{\epsilon}, \sigma^2_{\omega})] = ||W^{1/2}(X \circ \bar{\beta} - X \circ \tilde{\beta})||_2^2 + (X \circ \tilde{\beta})'W(X \circ \tilde{\beta}) + q_\lambda(\tilde{\beta})
\]

Dropping all constant terms once again in the above equation and overloading notation for \( L_\lambda(\tilde{\beta}) \) results in the decoupled loss:

\[
L_\lambda(\tilde{\beta}) = ||W^{1/2}(X \circ \bar{\beta} - X \circ \tilde{\beta})||_2^2 + q_\lambda(\tilde{\beta}) \tag{2}
\]

The full details of the above derivation are shown in the Online Appendix. Equation 2 can be thought of as a second level shrinkage on the underlying coefficients to induce hard thresholding. The loss function, parametricized by the sparsity parameter \( \lambda \), will be utilized to select changepoints from the posterior estimates of a Bayesian DLM.

### 2.2 Weights and Penalty Function

Next, we want to discuss our choice for the weight matrix \( W \) and the penalty function \( q_\lambda(\tilde{\beta}) \). For the weight matrix \( W = \text{diag}(w_1, ..., w_n) \), the classic choice for weights for weighted least square algorithm is inverse of the variance (Kiers, 1997). In our case, since we have estimate of posterior variance at each time-step through the Bayesian DLM, we will set our weights to be

\[
w_i = \frac{1}{\bar{\sigma}^2_{\epsilon,i}} \quad \text{for } i = 1, ..., n \tag{3}
\]

where \( \bar{\sigma}^2_{\epsilon,i} \) is the posterior mean for the variance of time \( i \) which will be utilized as our point estimate for the variance at time \( i \). This weight will induce a smaller \( l_2 \)-norm for time-steps with a larger variance and a larger \( l_2 \)-norm for time-steps with a smaller variance. This makes sense as we expect the process \( \{\beta_t\} \) to have more variability in regions of high volatility. The weights induces additional robustness for heteroskedastic data.

The penalty term \( q_\lambda(\tilde{\beta}) \) should penalize the number time-steps for which the \( D \)th difference in \( \tilde{\beta} \) is non-zero. Ideally, the \( l_0 \) norm will be used to identify the optimal subset of time-steps in which the \( D \)th difference are non-zero. However, since the \( l_0 \) norm is difficult to estimate efficiently, we start by relaxing
the $l_0$ norm to the $l_1$ norm as follows:

$$q_\lambda(\tilde{\beta}) = \lambda \sum_t |\triangle^D \beta_t|_1$$

One issue with this penalty is that it will bias the result toward changepoints of low magnitudes. This is due to the fact the penalty term increases linearly in relation to the change in the $\{\beta_t\}$ process. Without a methodology to normalize for the magnitude of change in the above penalty function, this penalized loss will prefer changepoints with smaller changes in $\{\beta_t\}$ over changepoints with larger changes. As a result, we propose the following penalty function for the decoupled approach:

$$q_\lambda(\tilde{\beta}) = \lambda \sum_t \frac{1}{|\psi_t|} |\triangle^D \beta_t|$$

(4)

where $\psi_t = \frac{1}{M} \sum_{i=1}^M \triangle^D \beta^{(i)}_t$. Motivated by similar correction in Hahn and Carvalho (2015), $\psi_t$ at time index $t$ is the posterior mean of the $D$th degree difference of $\beta_t$. This term can function as a normalizer which levels the impact of each changepoint regardless of the magnitude of the change at that time-step. In time-step with a larger change in $\{\beta_t\}$, $\psi_t$ will tend to be higher in magnitude, leading a changepoint to be penalized less. As a result, this weight term pushes the $l_1$ norm closer to the $l_0$ norm and gives better results for changepoint estimation.

2.3 Selecting the Optimal Number of Changepoints

As seen in equation 2 and 4, we have a loss functions indexed by the parameter $\lambda$. As $\lambda$ approaches 0, there would be no enforcement of sparsity and every point will be treated as a changepoint. As $\lambda$ approaches $\infty$, all $\{\triangle^D \tilde{\beta}_t\}$ will be 0 and no changepoint will be detected. The value of $\lambda$ plays a critical role in the final selection of the number of changepoints. Typically, with a penalized loss function, cross-validation is used to select the penalty parameter. However, in the case for the decoupled approach, since the loss is taken over the posterior estimate for the latent parameter $\{\beta_t\}$, the MCMC samples can be utilized in identifying the optimal set of changepoints. We propose using the following methodology to select the optimal number of changepoints.

First, the loss in equation 2 can be solved via coordinate descent (implemented using R package glmnet Friedman et al. (2010)) to produce a path of $\lambda$ values corresponding to different number of changepoints. This path of solutions will express a direct trade-off between goodness-of-fit and the number of changepoints. As the number of changepoints increases, the estimated solution $\tilde{\beta}$ will be fit closer to the posterior mean $\bar{\beta}$.

Second, for each $\lambda$ in the corresponding solution path, we will compute the “projected posterior” (Woody et al., 2020) to quantify its uncertainty. The key idea behind the “projected posterior” is to
project each MCMC draw from the Bayesian model onto the summary space defined by location of changepoints. For a given value \( \lambda \), let \( \eta_\lambda \) denote the time indices which \( \{ \Delta^D \hat{\beta}_t \neq 0 \} \) (the changepoint locations). Index 1 to \( D \) are automatically included in every \( \eta_\lambda \) as they are unpenalized. Let \( \hat{\beta}^{(i)} \) denote the \( i \)th MCMC draw from the Bayesian model, \( Z^{(D)} \) denote the inverse of the \( D \)th difference matrix, \( Z^{(D)}_{\eta} \) denote the subset of columns of \( Z^{(D)} \) indexed by a given \( \eta \). The \( i \)th projected posterior is then given by:

\[
\hat{\beta}^{(i)}_{\eta} = (t(Z^{(D)}_{\eta} Z^{(D)}_{\eta}^{-1} t(Z^{(D)}_{\eta} \beta^{(i)})
\]

where \( t() \) is the transpose operator. This projects \( \{ \beta^{(i)}, i = 1, \ldots, M \} \) from each MCMC draw onto the best fitted model given the changepoints. In summary, the “projected posterior” takes a set of changepoint locations and produces the best estimate of \( \beta \) for each of the MCMC draws given the changepoint locations. This, in turn, allows us to visualize a trade-off between the number of changepoints and the corresponding fit for the posterior estimates.

Third, after deriving the “projected posteriors”, we need a diagnostic tool to calculate goodness-of-fit. A common criterion for model selection is the amount of variation explained. Since we accounted for heteroskedasticity in the error term of the Bayesian DLM, we propose using the following metric as an estimate for amount of variation explained by the changepoints for the \( i \)th MCMC draw:

\[
R^2_{\lambda}^{(i)} = 1 - \frac{\| W^{1/2} (X \circ \beta^{(i)} - X \circ \hat{\beta}^{(i)}_{\eta}) \|^2}{\| W^{1/2} (X \circ \beta^{(i)} - X \circ \mu_{\beta^{(i)}}) \|^2}
\]

where \( \mu_{\beta^{(i)}} \) is the mean of \( \beta^{(i)} \). This metric is a similar R-squared in that it measures the amount of variation explained by the projected posterior \( \hat{\beta}_{\eta} \) for each of the MCMC draws. However, error for each time-step is multiplied by the corresponding weight matrix value, giving time-steps with higher variances lower weights. This makes sense as we expect more uncertainty in regions of high volatility. In turn, this metric provides an estimate of variation explained for each of the MCMC draws. The higher the value of \( R^2_{\lambda}^{(i)} \), the better the fit of the “projected posterior” for the \( i \)th MCMC draw. For selecting the optimal value of \( \lambda \), we will select the lowest number of changepoints which the upper 90% credible interval for \( R^2_{\lambda} \) exceeds a certain threshold. Details on the threshold selection will be given in Section 3.

2.4 Locally Adaptive with Shrinkage Priors

In the current model, we assume the coefficients \( \{ \beta_t \} \) follows a random walk with constant variance \( \sigma^2_\omega \). While this setup can be sufficient for data with strong signal, this model tends over-fit in datasets with low signal-to-noise ratios. Previous work have found shrinkage priors reduces over-fitting and improve the performance of Bayesian dynamic linear models (Kowal et al., 2019; Belmonte et al., 2013; Banbura et al., 2010). Shrinkage priors present a trade-off between goodness-of-fit and smoothness of the underlying process; a stronger shrinkage prior will typically result in a smoother underlying fit for the \( \{ \beta_t \} \) process.
In this section, we will introduce the use shrinkage priors for the decoupled approach.

As previously discussed in Section 1, various forms of shrinkage priors have shown to be effective in Bayesian DLMs. For this section, we will focus on the class of so-called “global-local shrinkage priors” which have shown to be effective for Bayesian modeling (Bhadra et al., 2016). Equation 1 will be modified to be as follows:

\[ y_t = x_t \beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_{\epsilon,t}^2) \]

\[ \Delta^D \beta_t = \omega_t \]

\[ \omega_t \sim N(0, \tau^2 \lambda^2_{\omega,t}) \] (5)

This extension induces global-local shrinkage on the \( D \)th difference of the coefficients for the predictor. The parameter \( \tau^2 \) induces global shrinkage across all time-steps and the process \( \{\lambda^2_{\omega,t}\} \) induce time-specific shrinkage for the coefficients at each time-step. The two parameters combined together shrink small deviations toward zero while allowing large signals to remain unchanged. This provides localized adaptivity while maintaining strong global shrinkage. One example of such shrinkage prior is the dynamic shrinkage process detailed in Kowal and Bourgeois (2020). The shrinkage process is detailed as follows:

\[ \Delta^D \beta_t = \omega_t, \quad h_t = \log(\tau^2 \lambda^2_{\omega,t}), \]

\[ \omega_t \sim N(0, \tau^2 \lambda^2_{\omega,t}), \quad h_t = \mu + \phi_1(h_{t-1} - \mu) + \eta_t, \] (6)

where \( \tau^2 \) tracks global shrinkage and the process \( \{\lambda^2_{\omega,t}\} \) tracks local shrinkage. Due to the effectiveness of the model shown in the paper, we will utilize this model for all simulations and call this version of the decoupled approach: decoupled dynamic shrinkage (DC-DS). Note that the decoupled approach is not restricted to the choice of shrinkage priors, other global-local shrinkage priors such as the double-gamma, double-exponential priors will achieve similar performances.

One key note for the decoupled approach is that the choice of shrinkage priors on the \( \{\beta_t\} \) process does not affect the second step of the decoupled analysis. One of the major advantages of the decoupled approach is that a highly complex shrinkage framework can be fit onto the existing data. The decoupled approach only utilizes the posteriors estimates of \( \{\beta_t\} \) in the second step for changepoint selection. As we will shown in the simulations, the resulting smoother estimate of the \( \{\beta_t\} \) process from incorporation of shrinkage priors will reduce over-estimation and improve the selection of changepoints.

### 2.5 Multiple Linear Predictors

In this section, we will extend the decoupled approach to deal with multiple predictors. Changepoint detection in multiple linear regression can be a challenging problem as different predictors may have changepoints at different time-steps. Most existing work in this area focus on identifying changepoint
jointly across series and lack the ability to identify predictors associated with each changepoint. To accommodate for additional flexibility, we allow the users to specify any type of group structure among predictor series. All predictor series in the same group are assumed to have the same changepoints. For example, if we wish to identify changepoint separately for each individual predictor, then we can place all predictors in its own group and identify changes separately. On the other hand, if we wish to identify joint changepoints across all predictor series, then we can place all predictors in one group. By separating the process of modeling and changepoint selection, the decoupled approach allows for creation of any subset of groupings for selection without affecting the modeling of the Bayesian DLM. This additional flexibility can provide great insights on which predictor series is changing over time.

The details for the updated model can be seen as follows. Suppose now we have p predictor series \( X = \text{blockdiag}(x_1', ..., x_p') \) where \( X \) is a block-diagonal matrix with row \( i \) containing \( x_i' \) on the block-diagonal. Let \( x_i = (x_{i,1}, ..., x_{i,p})' \), we can rewrite Equation 1 as follows:

\[
y_t = x_i' \beta_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2_{\epsilon,t}) \\
\Delta^D \beta_t = \omega_t, \quad \omega_t \sim N(0, \Sigma_{\omega,t})
\]

In comparison to Equation 1, the model now involves \( p \) predictor series at each time-step. The variance is modeled by a heteroskedastic process \( \{\sigma^2_{\epsilon,t}\} \) and the \( D \)th difference for \( \{\beta_t\} \) is assumed to be normal with mean 0 and variance \( \Sigma_{\omega,t} \). Equation 7 provides a general framework which works well for a wide range of Bayesian DLMs. For the simulations and real world analysis, we assume \( \Sigma_{\omega,t} \) is diagonal with a dynamic shrinkage priors on each diagonal term. However, the decoupled approach will also work with other structures for this variance matrix. Once the above model is fit and MCMC samples are estimated, the decoupled loss for multiple predictor series can be written as follows:

\[
L_\lambda(\tilde{\beta}) = ||W^{1/2}(X\tilde{\beta} - X\tilde{\beta})||^2_2 + q_\lambda(\tilde{\beta})
\]

\[
q_\lambda(\tilde{\beta}) = \lambda \sum_{t=1}^n \sum_{g=1}^G \frac{1}{|\psi_{g,t}|} |\Delta^D \beta_{g,t}|
\]

where \( \tilde{\beta} = \text{vec}(\tilde{\beta}_1, ..., \tilde{\beta}_n)' \) and \( W \) is the same as Equation 3. The penalty function \( q_\lambda(\tilde{\beta}) \) is a variation of grouped lasso (Yuan and Lin, 2007) where \( G \) is the number of groups, \( |\Delta^D \beta_{g,t}| \) is the sum of absolute value of \( D \)th difference of all series in the particular group and \( |\psi_{g,t}| \) is the average sum of weight for data in each group. Similar to Equation 4, the term \( |\psi_{g,t}| \) functions as a normalizer across different groups to adjust for the number of predictor series and the magnitude of change.
2.6 Incorporating Static Parameters in Changepoint Analysis

Another important element to consider for Bayesian DLMs is the existence of static variables aka co-
variates. Covariates typically have non-time-varying relation with the response variable. As a result, we
are not interested in identifying changepoints regarding the relationship between the covariates and the
response. Rather, we wish to fit a model for our predictors with covariates included and identify changes
in relationship between the predictors and the response. Suppose we have a set of $l$ covariates denoted
by $Z = (z_1, ..., z_n)$ where $z_i = (z_{i,1}, ..., z_{i,l})'$ for $i = 1, ..., n$. The Bayesian DLM, modified from Equation
7 can be written as follows:

$$y_t = x_t' \beta_t + \alpha' z_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_{\epsilon,t}^2)$$

$$\Delta^D \beta_t = \omega_t, \quad \omega_t \sim N(0, \Sigma_{\omega,t})$$  (9)

where $\alpha = (\alpha_1, ..., \alpha_l)'$ are the estimated coefficients for the covariates. This model can be estimated
using standard Bayesian methodology for DLMs. Once the posterior estimates are derived, we combine
$X$ and $\alpha$ for the decoupled loss as follows:

$$L_\lambda(\tilde{\beta}, \tilde{\alpha}) = ||W^{1/2}([X, Z][\tilde{\beta}', \tilde{\alpha}'])' - [X, Z][\tilde{\beta}', \tilde{\alpha}'])||^2_2 + q_\lambda(\tilde{\beta})$$  (10)

where $[\cdot, \cdot]$ denote concatenation of two matrices among columns. Using same notation as previous, $X$
is a block-diagonal matrix where row $i$ contains $x'_i$ on the block-diagonal and $\tilde{\beta} = vec(\tilde{\beta}_1, ..., \tilde{\beta}_T)^T$. In
Equation 10, $X \in R^{T \times Tp}$, $Z \in R^{T \times l}$ and $[X, Z]$ results in a matrix of dimension $T \times (Tp + l)$ and $[\tilde{\beta}', \tilde{\alpha}']'$
results in a matrix of dimension $(Tp + l) \times 1$. $q_\lambda(\tilde{\beta})$ is the same as Equation 8.

One key note is that the penalty function $q_\lambda(\tilde{\beta})$ does not change with incorporation of covariates. This
is due to the fact that covariates relationship with the response do not change over-time and therefore
are not penalized. If there exist a large number of covariates, variable selection can be done by penalizing
the $l_1$ norm of the coefficients $\alpha$. However, as the focus of this paper is on changepoint analysis, we will
not investigate into this further and will leave the covariates unpenalized in the decoupled loss for the
rest of the paper. The decoupled loss can be estimated using package for solving grouped lasso such as
R package gglasso.

The existence of time-varying and non-time-varying predictors have traditionally posed a significant
challenge for changepoint analysis in regression. The decoupled approach’s ability to deal with time-
varying and non-time-varying predictors grants the algorithm additional flexibility and robustness to deal
with various types of real world problems. This extension further highlights the benefit of separating
the processes of modeling and changepoint selection. During the modeling step, a complex Bayesian
model can be fit to the existing data that incorporates a variety of complexities such as covariates,
heteroskedastic noise and non-stationary inputs. As long as the model can produce fairly smooth estimate of the coefficients for the predictors of interest, the decoupled approach can adapt the inference part to identify key changepoints. This level of flexibility grants the decoupled approach the ability to work with more applications than previous existing changepoint algorithms. We will illustrate the decoupled approach’s effectiveness in comparison to other algorithms in the following sections.

3 Simulated Experiments

3.1 Change in Mean in Noisy Environment

In this section, we illustrate the effectiveness and flexibility of the decoupled approach by comparing our method against two other changepoint methods based on penalized likelihood functions. The competing methods are the Pruned Exact Linear Time method (PELT, Killick et al., 2012) and Robust FPOP algorithm (R-FPOP, Fearnhead and Rigaill, 2019). PELT identifies changepoint based on penalized cost function using a goodness-of-fit metric based on maximum negative likelihood for each segment and a penalty parameter on the number of changepoints. R-FPOP adapts the penalty function using a biweight-loss in order to deal with outliers by establishing a maximum threshold for the impact of each time-step.

We chose these two methods as they are the similar to the decoupled approach in their utilization of penalized cost function for identification of changepoints. However, unlike the decoupled approach, they utilize the data rather than posterior from a Bayesian DLM in their penalized loss function. By comparing the decoupled approach against these methods, we will illustrate the benefits of using Bayesian modeling as first step before penalized loss. The comparisons will start on simple cases of change in mean with Gaussian noise, then extend to more complicated scenarios adding in outliers and heterogeneity. For both competing methods, we will use the default parameters as utilized in the original papers. For the decoupled approach, we use a cutoff threshold for 0.9 for lowest number of changepoints which the upper 90% credible interval for $R^2$ exceeds. Full detail of the parameters used for the Bayesian DLM is shown in the Online Appendix.

Five metrics are used to evaluate the results: Rand index, adjusted Rand index, precision, recall and F1-score. Rand index calculates a similarity score between the predicted partition and the true partition; the score ranges between 0 and 1 with 1 being a perfect match. Adjusted Rand index provides an additional correction step to the Rand Index by accounting for random chance of a correct partition. Precision measures proportion of true changepoints in the number of predicted changepoints while recall measures proportion of all true changepoints detected by the models. F1-score calculates the harmonic mean between precision and recall. Since changepoints occur very rarely in the data, the F1-score is a good indicator for the accuracy of predictions (van den Burg and Williams, 2020). We consider a
### Table 1: Single Change in Mean

| Mag. of Change | Algorithms   | Rand Avg. | Adj. Rand Avg. | Precision | Recall | F1-score |
|---------------|--------------|-----------|----------------|-----------|--------|----------|
| 1             | DC-RW        | 0.795(0.013) | 0.590(0.027) | 0.18      | 0.83   | 0.29     |
|               | DC-DS        | 0.964(0.004) | 0.929(0.009) | 0.78      | 0.83   | 0.80     |
|               | PELT         | 0.968(0.004) | 0.936(0.007) | 0.82      | 0.87   | 0.84     |
| 0.75          | DC-RW        | 0.689(0.016) | 0.379(0.031) | 0.14      | 0.53   | 0.22     |
|               | DC-DS        | 0.926(0.009) | 0.851(0.018) | 0.68      | 0.56   | 0.61     |
|               | PELT         | 0.930(0.009) | 0.860(0.021) | 0.62      | 0.67   | 0.64     |
| 0.5           | DC-RW        | 0.579(0.012) | 0.156(0.024) | 0.12      | 0.25   | 0.16     |
|               | DC-DS        | 0.791(0.017) | 0.582(0.035) | 0.23      | 0.38   | 0.29     |
|               | PELT         | 0.755(0.021) | 0.512(0.041) | 0.38      | 0.30   | 0.33     |
| 0.25          | DC-RW        | 0.498(0.000) | 0.000(0.000) | 0.00      | 0.00   | 0.00     |
|               | DC-DS        | 0.510(0.006) | 0.025(0.013) | 0.02      | 0.01   | 0.01     |
|               | PELT         | 0.498(0.000) | 0.000(0.000) | 0.00      | 0.00   | 0.00     |

Table details results of the decoupled approach with random walk, decoupled approach with dynamic shrinkage, and PELT on simulated data with one change in mean of varying magnitudes and standard Gaussian noise. Rand average and adjusted Rand average measures the similarity between predicted partition and true partition. Standard error for Rand average and adjusted Rand average are given in subscripts. F1-score measures accuracy of changepoint detection through a comparison of precision and recall.

predicted changepoint to be a true positive if it is within ±5 of a true changepoint, with the caveat that each true changepoint can only match to at most one predicted changepoint.

### 3.1.1 Change in Mean with Gaussian Noise

For the first set of simulations, we start with a simple change in mean with standard Gaussian noise. We simulate data of length 200, with a changepoint in the middle of the data at location 100. We adjust different levels for the magnitude of change to understand the effectiveness of the algorithms with varying signal-to-noise ratios. We test 4 different magnitudes of change values of \{1, 0.75, 0.5, 0.25\}.

We will compare the decoupled approach with random walk (DC-RW, model introduced in Section 2.1) and the decoupled approach with dynamic shrinkage (DC-DS, model introduced in Section 2.4) against PELT. PELT is a penalized likelihood changepoint algorithm which is designed to identify changes in this setting, making it a good baseline for comparison. We expect the decoupled approach to perform slightly worse than PELT in this setting as a trade-off for increased flexibility. As we will show in the later simulations, the flexibility of the decoupled approach allows it to perform much better when the assumptions of homoskedasticity and Gaussian noise are violated.

As seen in Table 1, the decoupled approach with dynamic shrinkage performs slightly worse than PELT. With a signal-to-noise ratio of 1 to 1 (magnitude of change 1), both DC-DS and PELT perform similarly well with Rand average above 0.95, adjusted Rand average above 0.925 and F1-score above 0.8. As the signal-to-noise ratio reaches a low of 1 to 4 (magnitude of change 0.25), both changepoint algorithms can no longer distinguish the correct changepoint. For the magnitude change of 0.75 and 0.5, PELT performs slightly better in terms of F1-score. However, we still see a trade-off of precision and recall.
| Mag. of Change | Algorithms | Rand Avg. | Adj. Rand Avg. | Precision | Recall | F1-score |
|---------------|------------|-----------|----------------|-----------|--------|---------|
| 2             | DC-DS      | 0.977(0.003) | 0.954(0.007) | 0.88      | 0.91   | 0.89    |
|               | PELT       | 0.681(0.006) | 0.361(0.012) | 0.06      | 0.82   | 0.11    |
|               | R-FPOP     | 0.952(0.011) | 0.904(0.022) | 0.86      | 0.83   | 0.84    |
| 1.5           | DC-DS      | 0.967(0.005) | 0.934(0.009) | 0.80      | 0.82   | 0.81    |
|               | PELT       | 0.689(0.007) | 0.376(0.014) | 0.05      | 0.74   | 0.10    |
|               | R-FPOP     | 0.860(0.020) | 0.721(0.040) | 0.80      | 0.63   | 0.70    |
| 1             | DC-DS      | 0.935(0.007) | 0.870(0.015) | 0.68      | 0.60   | 0.63    |
|               | PELT       | 0.680(0.007) | 0.358(0.013) | 0.04      | 0.59   | 0.08    |
|               | R-FPOP     | 0.804(0.009) | 0.638(0.019) | 0.35      | 0.62   | 0.44    |
| 0.5           | DC-DS      | 0.742(0.018) | 0.486(0.036) | 0.18      | 0.32   | 0.24    |
|               | PELT       | 0.676(0.006) | 0.350(0.013) | 0.04      | 0.48   | 0.07    |
|               | R-FPOP     | 0.741(0.017) | 0.484(0.035) | 0.15      | 0.25   | 0.19    |

Table 2: Change in Mean with Outliers

Table details results of DC-DS, PELT and R-FPOP on simulated data with one change in mean of varying magnitude of change and outliers. Outliers are simulated using t-distributed noise with 2 degrees of freedom.

between the two algorithms. For magnitude of change of 0.5, PELT has a higher precision while DC-DS has a higher recall. This shows that DC-DS has a tendency to slightly over-predict in low signal-to-noise ratio while PELT has a tendency to under-predict. A key note is that DC-DS maintains the highest Rand and adjusted Rand average in this settings, showing that DC-DS produces the partition closest to the true partition. Comparing the two metrics, adjusted Rand penalizes more for under-prediction while F1-score penalizes more over-prediction. As a trade-off, DC-DS produces more accurate partition, but as a cost, has slightly more number of false positives.

Comparing DC-RW against DC-DS, we can clearly see that using shrinkage priors in the Bayesian DLM significantly improves the performance of the decoupled approach. This is due to the fact that shrinkage priors induce a smoother estimate of the underlying trend; as a result, changepoint inference on the posterior becomes much easier. This further supports the discussion in Section 2.4 of the advantages of the decoupled framework by allowing for fitting of a highly complex shrinkage framework. Due to the significant improvements of using shrinkage priors in the baseline case, we will use DC-DS as our main method from this point onward.

3.1.2 Change in Mean with Outliers

For this set of simulations, we added outliers onto the same problem as Section 3.1.1 to illustrate how the performance of these algorithms change in the presence of outliers. We will add the R-FPOP method to the comparison as this method is designed to identify changepoints in presence of outliers. The simulation setting will be kept the same as Section 3.1.1; 100 simulation will be generated of data of length 200 with various magnitude of changes. As this is a significantly more difficult problem, we increase the magnitude of change to be selected from \{2, 1.5, 1, 0.5\}. Instead of Gaussian noise, we will utilize t-distributed noise with 2 degrees to simulate data with extreme outliers.
The top-left, mid-left and bottom-left plots show posterior mean of Bayesian DLMs fit on data with Gaussian Noise, t-distributed noise and stochastic volatility respectively. The blue line is the posterior mean for fit with dynamic shrinkage (DC-DS) while the red line is posterior mean for fit with random walk (DC-RW). As seen, the addition of shrinkage priors induce smoother estimates for the $\{\beta_t\}$ process. The top-right, mid-right and bottom-right plot shows posterior mean (blue line) of the optimal decoupled result for DC-DS on the same data as left-side with 95 percent credible bands in dark gray. The correct changepoint is shown by the brown vertical bar.

The results of the simulation can be seen in Table 2. With the addition of outliers, the decoupled approach is able to achieve the best performance across all settings. By utilizing a Bayesian DLM with dynamic shrinkage, the decoupled approach is robust to the presence of extreme outliers. PELT, with no mechanism to deal with extreme values, are significantly influenced by outliers. This lead to PELT to significantly over-predict the number of changepoints. To be fair to PELT, the algorithm is not intended
The left plot shows F1-score of DC-DS (blue line) against PELT (red line) in simulated data with Gaussian noise from Section 3.1.1. The middle plot shows F1-score of DC-DS (blue line) against R-FPOP (orange line) in simulated data with outliers from Section 3.1.2. The right plot shows F1-score of DC-DS (blue line) against R-FPOP (orange line) in simulated data with stochastic volatility from Section 3.1.3.

to work on datasets with outliers. In the setting of magnitude of change of 2, DC-DS achieves an F1-score of 0.89 in comparison to 0.84 for R-FPOP. As the signal-to-noise ratio decreases, we can see an increasing gap in adjusted Rand average and F1-score between DC-DS and R-FPOP. This indicates that DC-DS can produce more precise changepoint estimations and more accurate partitions. The advantage becomes more significant in the setting of magnitude of change of 1. DC-DS achieves a F1-score of 0.63 in comparison to 0.44 of R-FPOP. As magnitude of change approaches 0.5, the problem becomes too difficult for all algorithms and F1-scores are very low.

Figure 1 illustrates the results of the decoupled approach in three example series generated with Gaussian noise, t-distributed noise and stochastic volatility. The blue line in the middle-left plot shows the posterior mean of the $\{\beta_t\}$ process from a Bayesian dynamic linear model with dynamic shrinkage on a series generated with outliers. The middle-right plot then shows the optimal fit from the decoupled loss and the resulting projected posterior. This highlights one major advantage of separating shrinkage and changepoint selection. Despite the presence of outliers, a Bayesian DLM with dynamic shrinkage is able to predict a relatively smooth underlying trend with one jump around the changepoint location. The decoupled loss then identifies the optimal piece-wise constant fit to the smooth posterior and flags the correct changepoint location.

3.1.3 Change in Mean with Heterogeneity

For the next set of simulations, we evaluate these algorithms in presence of heterogeneity. We simulate 100 series of length 200, with a changepoint in the middle of the data at location 100. However, instead of standard Gaussian noise or t-distributed noise, we generate noise using stochastic volatility of order 1.
Table 3: Change in Mean with Heterogeneity

| Mag. of Change | Algorithms | Rand Avg. | Adj. Rand Avg. | Precision | Recall | F1-score |
|----------------|------------|-----------|----------------|-----------|--------|----------|
| 2              | DC-DS      | 0.982(0.004) | 0.964(0.008) | 0.89      | 0.90   | 0.89     |
|                | PELT       | 0.834(0.010) | 0.667(0.019) | 0.11      | 0.96   | 0.20     |
|                | R-FPOP     | 0.949(0.009) | 0.889(0.018) | 0.73      | 0.85   | 0.79     |
| 1.5            | DC-DS      | 0.977(0.004) | 0.955(0.008) | 0.87      | 0.90   | 0.88     |
|                | PELT       | 0.827(0.010) | 0.654(0.020) | 0.10      | 0.89   | 0.19     |
|                | R-FPOP     | 0.926(0.011) | 0.852(0.022) | 0.64      | 0.81   | 0.72     |
| 1              | DC-DS      | 0.931(0.009) | 0.862(0.020) | 0.60      | 0.64   | 0.62     |
|                | PELT       | 0.800(0.010) | 0.599(0.020) | 0.07      | 0.61   | 0.13     |
|                | R-FPOP     | 0.837(0.019) | 0.671(0.037) | 0.39      | 0.50   | 0.44     |
| 0.5            | DC-DS      | 0.836(0.015) | 0.673(0.030) | 0.29      | 0.38   | 0.33     |
|                | PELT       | 0.686(0.013) | 0.373(0.026) | 0.03      | 0.25   | 0.05     |
|                | R-FPOP     | 0.626(0.018) | 0.254(0.036) | 0.12      | 0.11   | 0.12     |

Table details results of DC-DS, PELT and R-FPOP on simulated data with one change in varying magnitude and stochastic volatility. Stochastic volatility is simulated using highly autocorrelated SV(1) model. (Kim et al., 1998) as follows:

$$\log(\sigma^2_{\epsilon,t}) = \mu_\epsilon + \phi_\epsilon [\log(\sigma^2_{\epsilon,t-1}) - \mu_\epsilon] + \eta_{\epsilon,t}, \quad \eta_{\epsilon,t} \sim N(0, \sigma^2_\eta).$$  \hspace{2cm} (11)

We set the following values: \(\mu_\epsilon = 0, \phi_\epsilon = 0.9,\) and \(\sigma^2_{\epsilon,t} = 0.5.\) This creates high auto-correlation which causes regions of high/low volatility which can occur frequently in real world data. We utilized 4 magnitude of change values of \(\{0.5, 1, 1.5, 2\}\) to evaluate the algorithms’ effectiveness in varying signal-to-noise ratios. The results are reported in Table 3. To be fair to PELT and R-FPOP, the algorithms are not intended to work in this setting. As a result, the performances are not reflective of the effectiveness of the algorithms. Rather, this illustrates the flexibility of the decoupled approach in being able to utilize Bayesian DLMs to locally adapt to heterogeneity inherent in datasets.

Comparing results in Table 3 to Table 1 for magnitude of change 1, we see that the performance of all changepoint algorithms decreased in presence of stochastic volatility. This is to be expected as stochastic volatility makes detection of changepoints much more difficult. With the addition of stochastic volatility, DC-DS outperformed other competing changepoint methods in all settings. DC-DS achieves a F1-score of 0.89 for setting of magnitude of change of 2, a F1-score of 0.62 for setting of magnitude of change of 1 and a F1-score of 0.33 for setting of magnitude of change of 0.5. DC-DS achieves the most accurate partitions by having the highest adjusted Rand average and the best trade-offs of precision/recall. This illustrates the robustness of the decoupled approach in dealing with heterogeneity.

Figure 2 summarizes the results in term of F1-score for Section 3.1. As seen in the plots, the decoupled approach is slightly worse in standard Gaussian noise setting but performs significantly better when outliers or stochastic volatility are added. This illustrates the trade-off of the decoupled approach. By fitting a Bayesian DLM to the data first, the decoupled approach can be more locally adaptive to
the complexities inherent in time series data. Outliers and heterogeneity are just two examples of the challenges that the decoupled approach can deal with. As long as the posterior estimates for the \(\{\beta_t\}\) process remains relatively smooth, the decoupled loss can identify correct changepoint locations in a variety of complex scenarios.

### 3.2 Change in Regression

For the next sets of experiments, we will examine the effectiveness of the decoupled approach in various regression situations. In particular, we will look at three different situations associated with changes in relationship between predictors and response. First, we will look at the case of multiple predictors with two changepoints and constant variance. Next, we will extend the analysis for data with stochastic volatility. Finally, we will examine the case with added covariates.

A tremendous amount of work has been done using variations of the lasso penalty for identifying changepoints in linear regression (Ciuperca and Maciak, 2019; Jin et al., 2016; Qian and Su, 2016; Ciuperca, 2014). We will compare against the algorithm SGL based sparse group lasso penalty (Zhang et al., 2015) as it closest resembles the decoupled loss. In addition, we will compare against two other regression changepoint algorithms: variance projected wild binary segmentation denoted by VPWBS (Wang et al., 2021) and a method called computationally efficient change point detection which we will denote by CECPD (Leonardi and Bühlmann, 2016). VPWBS projects the predictors onto a univariate series and then utilizes wild binary segmentation to identify the changepoints. CECPD uses a combination of binary segmentation and dynamic programming to estimate the optimal set of changepoints that minimizes a \(l_1\) loss function. We use default parameters for CECPD as recommended in the original paper. For VPWBS and SGL, we utilize cross-validation to select the optimal parameters.

Out of the three methods, SGL is the closest related to our method as it utilizes a sparse group lasso penalty that penalizes for number of changepoints as well as goodness of fit within each segment. However, there are two key differences between the decoupled loss and SGL. First, the decoupled approach does not use cross-validation to select the optimal penalty parameter. Instead, the penalty selection criterion introduced in Section 2.3 is used to select optimal changepoint locations. Second, a normalizer (\(\{\psi_{g,t}\}\)) based on posterior average magnitude of change is utilized in the penalty term for the decoupled approach. As we will show throughout this section, these differences induce more robustness and better performance for the decoupled approach. The same metrics introduced in Section 3.1 will be used to evaluate the performance of the methods.

#### 3.2.1 Change in Regression for Multiple Predictors

For the first set of simulations, we examine the cases of multiple predictors with two joint changepoints. This is a standard case for changepoint analysis in regression setting and can be used as a baseline...
for comparison. We perform 100 simulations to generate data of length 300 with two changepoints at location 75 and 225. The response series \( \{y_t\} \) and the predictor series \( \{x_t\} \) where \( x_t = (x_{t,1}, x_{t,2}, x_{t,3}) \) are generated with the following relationship:

\[
y_t = \begin{cases} 
\beta_1'x_t + \epsilon_t & t \leq 74 \\
\beta_2'x_t + \epsilon_t & 75 \leq t \leq 224 \\
\beta_3'x_t + \epsilon_t & t \geq 225 
\end{cases}
\]

where \( \{\epsilon_t\} \) is a standard Gaussian noise, \((\beta_1, \beta_2) = (0, C)\) where \( C \) denotes the magnitude of change. As change in regression tends to be more difficult than change in mean, the magnitude of change will be higher in the coefficients. We will select \( M \) from the following set: \( \{2, 1, 0.5\} \). The predictors are generated from \( x_t \sim_{i.i.d.} N(0, I) \) for \( t = 1, ..., 300 \) where \( I \in R^{3 \times 3} \) is the identity matrix.

Table 4: Change in Regression with Multiple Predictors

| Mag. of Change | Algorithms | Rand Avg. | Adj. Rand Avg. | Precision | Recall | F1-score |
|----------------|------------|-----------|----------------|-----------|--------|----------|
| 2              | DC-DS      | 0.991(0.001) | 0.981(0.002) | 0.99      | 0.99   | 0.99     |
|                | SGL        | 0.902(0.004) | 0.776(0.010) | 0.13      | 1.00   | 0.23     |
|                | CECPD      | 0.987(0.001) | 0.971(0.001) | 0.99      | 0.99   | 0.99     |
|                | VPWBS      | 0.971(0.004) | 0.937(0.008) | 0.92      | 0.98   | 0.95     |
| 1              | DC-DS      | 0.979(0.002) | 0.935(0.005) | 0.85      | 0.85   | 0.85     |
|                | SGL        | 0.886(0.004) | 0.736(0.011) | 0.13      | 1.00   | 0.23     |
|                | CECPD      | 0.826(0.027) | 0.713(0.042) | 0.97      | 0.73   | 0.83     |
|                | VPWBS      | 0.957(0.005) | 0.902(0.011) | 0.81      | 0.88   | 0.84     |
| 0.5            | DC-DS      | 0.898(0.008) | 0.782(0.016) | 0.38      | 0.55   | 0.45     |
|                | SGL        | 0.678(0.015) | 0.406(0.023) | 0.14      | 0.40   | 0.21     |
|                | CECPD      | 0.558(0.027) | 0.285(0.042) | 0.92      | 0.28   | 0.42     |
|                | VPWBS      | 0.782(0.019) | 0.575(0.028) | 0.40      | 0.48   | 0.43     |

Table details results of the decoupled approach, SGL, CECPD and VPWBS on simulated data with standard Gaussian noise and three predictors with two changepoints at location 75 and 225. Three settings for magnitude of change \((C)\) are simulated in this experiment.

The results can be seen in Table 4. Starting from setting of \( C = 2 \), most algorithms are able to perform well. DC-DS, CECPD and VPWBS achieve high F1-scores above 0.95, showing their effectiveness in the baseline setting. DC-DS is able to achieve the highest adjusted Rand of 0.991, showing that the partition it finds is the most accurate to the true partition. DC-DS is the most precise at finding the exact changepoint locations. For the setting of \( C = 1 \) and \( C = 0.5 \), DC-DS once again achieves the best Rand average and adjusted Rand average. DC-DS, CECPD, VPWBS all achieve similar level of F1-scores. In the precision-recall trade-off for the setting \( C = 0.5 \), DC-DS achieves best recall of 0.55 while CECPD achieves the best precision of 0.92. This supports findings in Section 3.1 for change in mean. Overall, this illustrates DC-DS is competitive, albeit a bit better, than other changepoint algorithms in the baseline setting.

Comparing DC-DS with its closest competitor SGL, we see two major issues with SGL for flagging
changepoints. First, as SGL identifies changepoints using a penalized loss function with \( l_1 \) penalty based on the data, it has a tendency to over-predict the number of changepoints. Without a normalizer (\( \psi_{g,t} \)), SGL tends to prefer predicting multiple small changes rather than one significant change, leading to high precision, very low recall and low F1-score. Second, cross-validation struggles to identify the correct tuning parameter for the loss. As seen in the results, the tuning parameter value selected lead to significant over-prediction. Due to its poor performance in the baseline setting, we will drop SGL from the rest of the simulations.

### 3.2.2 Change in Regression with Stochastic Volatility

For the next set of simulations, we examine how the performance of these changepoint algorithms deteriorate in the presence of heterogeneity. For the most part, we use the same setting as Section 3.2.1. 100 series of length 300 with two changepoints at location 75 and 225 will be generated. However, instead of Gaussian noise, the noise will follow stochastic volatility of order 1 seen in Equation 11. We set \( \mu_\epsilon \) to be 0, \( \phi_\epsilon \) to be 0.9, and \( \sigma^2_{\epsilon,t} \) to be 1, simulating high auto-correlation which lead to regions of high/low volatility.

Table 5: Change in Regression with Stochastic Volatility

| Mag. of Change | Algorithms | Rand Avg. | Adj. Rand Avg. | Precision | Recall | F1-score |
|----------------|------------|-----------|----------------|-----------|--------|----------|
| 2              | DC-DS      | 0.960 (0.007) | 0.922 (0.014) | 0.89      | 0.86   | 0.88     |
|                | CECPD      | 0.883 (0.015) | 0.724 (0.013) | 0.63      | 0.89   | 0.73     |
|                | VPWBS      | 0.881 (0.014) | 0.752 (0.025) | 0.54      | 0.65   | 0.59     |
| 1              | DC-DS      | 0.898 (0.017) | 0.814 (0.028) | 0.75      | 0.70   | 0.72     |
|                | CECPD      | 0.869 (0.016) | 0.700 (0.015) | 0.52      | 0.60   | 0.56     |
|                | VPWBS      | 0.758 (0.019) | 0.545 (0.030) | 0.39      | 0.35   | 0.37     |
| 0.5            | DC-DS      | 0.777 (0.019) | 0.587 (0.031) | 0.42      | 0.43   | 0.42     |
|                | CECPD      | 0.726 (0.019) | 0.480 (0.028) | 0.20      | 0.18   | 0.19     |
|                | VPWBS      | 0.512 (0.019) | 0.185 (0.025) | 0.17      | 0.05   | 0.08     |

Table details results of DC-DS, CECPD and VPWBS on simulated data with stochastic volatility and three predictors with two changepoints at location 75 and 225. Three settings for magnitude of change (C) are simulated in this experiment.

The results are shown in Table 5. In comparison to the results in Table 4, we see that the performances of all 3 algorithms decreased in this setting. This is to be expected as switching from Gaussian noise to stochastic volatility makes changepoint identification much more difficult. DC-DS performs the best in terms of F1-score and average adjusted Rand in all three magnitude of change settings. DC-DS achieves a F1-score of 0.88 in magnitude of change 2 setting and 0.72 in magnitude of change 1 setting. CECPD comes in second place with a F1-score of 0.73 in magnitude of change 2 setting and 0.56 in magnitude of change 1 setting. VPWBS struggles more in this setting as the algorithm is not intended to work in presence of heterogeneity.

In presence of stochastic volatility, the gap between DC-DS widened in terms of both F1-score and adjusted Rand. average. This shows that DC-DS is more precise at finding the correct changepoint.
Figure 3: Change in Regression: Comparison of F1-Scores

The left plot shows F1-score of DC-DS (blue line) against CECPD (orange line) and VPWBS (red line) in simulated regression with multiple predictors and Gaussian noise from Section 3.2.1. The middle plot shows F1-score comparisons in simulated regression with stochastic volatility from Section 3.2.2. The right plot shows F1-score comparisons in simulated regression with covariates from Section 3.2.3.

locations. The results illustrate further support for one of the major advantages of the decoupled approach: the robustness and flexibility of Bayesian DLM in dealing with heterogeneity make the decoupled approach more adaptive to complex datasets.

3.2.3 Change in Regression with Covariates

For the final set of simulations, we evaluate the performance of the decoupled approach against other changepoint algorithms when covariates are added to the model. We simulate 100 series of length 300 with one changepoint at location 150. For each series, we generate one predictor series \( \{x_t\}_{t=1}^{300} \) and two covariate series \( \{z_{t,1}\}_{t=1}^{300}, \{z_{t,2}\}_{t=1}^{300} \). The relationship between the response, the predictor and the covariates are as follows:

\[
y_t = \begin{cases} 
\beta_1 x_t + \alpha_1 z_{t,1} + \alpha_2 z_{t,2} + \epsilon_t & t \leq 149 \\
\beta_2 x_t + \alpha_1 z_{t,1} + \alpha_2 z_{t,2} + \epsilon_t & t \geq 150
\end{cases}
\]

where \( \{\epsilon_t\} \) is a standard Gaussian noise, \((\beta_1, \beta_2) = (0, M)\), and \((\alpha_1, \alpha_2) = (0.3, 0.1)\). The predictor series is generated from \( x_t \sim_i.i.d. N(0, 1) \) and the covariate series are generated from \( z_{t,1}, z_{t,2} \sim_i.i.d. \) Bern(0.5) where Bern(·) are samples from the Bernoulli distribution. As seen, the covariates relationship with \( y \) does not change over-time while one changepoint exist for the predictor series.

Table 6 details the results for this set of simulations. As seen, DC-DS outperforms the competing methods and the advantage becomes more significant as the signal-to-noise ratio decreases. In the highest setting for magnitude of change of 2, DC-DS, CECPD and VPWBS all perform well in terms of adjusted Rand average and F1-score. This is due to the fact that the change in predictor is significantly greater than the impact of the covariates. However, as the magnitude of change decreases to 1, we can see competing methods’ performances begin to deteriorate. This is to be expected as CECPD and VPWBS
Table 6: Change in Regression with Covariates

| Mag. of Change | Algorithms | Rand Avg. | Adj. Rand Avg. | Precision | Recall | F1-score |
|---------------|-----------|-----------|----------------|-----------|--------|----------|
| 2             | DC-DS     | 0.987 (0.002) | 0.974 (0.004) | 0.96 | 0.96 | 0.96 |
|               | CECPD     | 0.981 (0.002) | 0.961 (0.003) | 0.95 | 0.95 | 0.95 |
|               | VPWBS     | 0.961 (0.005) | 0.922 (0.010) | 0.78 | 0.90 | 0.84 |
| 1             | DC-DS     | 0.949 (0.007) | 0.899 (0.014) | 0.72 | 0.73 | 0.72 |
|               | CECPD     | 0.908 (0.009) | 0.816 (0.018) | 0.48 | 0.72 | 0.56 |
|               | VPWBS     | 0.850 (0.014) | 0.701 (0.028) | 0.38 | 0.55 | 0.45 |
| 0.5           | DC-DS     | 0.767 (0.021) | 0.535 (0.041) | 0.44 | 0.31 | 0.37 |
|               | CECPD     | 0.565 (0.016) | 0.133 (0.033) | 0.67 | 0.10 | 0.17 |
|               | VPWBS     | 0.681 (0.020) | 0.365 (0.040) | 0.15 | 0.10 | 0.12 |

Table details results of the decoupled approach, CECPD and VPWBS on simulated data with standard Gaussian noise, one predictor with one changepoint at location \{150\} and 2 covariates. Three settings for magnitude of change (C) are simulated in this experiment.

have no way to distinguish thee covariates from the predictor. DC-DS, on the other hand, maintains a high F1-score of 0.72 and adjusted Rand average of 0.899 in this setting. With a Bayesian DLM separating the impact of covariates from predictor, DC-DS is able to produce correct coefficient estimates. This advantage becomes more apparent as magnitude of change decreases to 0.5. In this setting, the predictor series has a lower impact on the response than the covariates. As a result, no other changepoint algorithm can reliably identify the correct changepoint location (the highest competing F1-score being CECPD at 0.17). DC-DS, in this setting, achieves a F1-score of 0.37, showing its robustness.

Figure 3 summarizes the F1-score results of DC-DS against the competing methods for all 3 regression simulation settings. As seen in the plot, all three algorithms perform similarly well in the baseline setting with Gaussian noise. However, as the scenarios become more complicated such as when stochastic volatility is added or covariates are included, the advantages of DC-DS become more apparent. This further highlights the advantage of the decoupled approach in its ability to fit a highly complex Bayesian DLM to the underlying data to deal with complications inherent in series.

4 Real World Examples

4.1 Global Land Surface Air Temperature Example

For an illustrative application we consider monthly global land surface air temperature from 1880 to 2018. The data is made available on National Aeronautics and Space Administration website and has been previously analyzed (Yang and Song, 2014). Identifying changes in climate patterns has become an urgent area of analysis (Alley et al., 2003). Understanding when the underlying dynamics of global temperature pattern shifts can be very important in policy decision making. The top panel in Figure 4 shows the recorded global land surface air temperature. As shown, there are clear long term linear time trends underlying local annual trends in the data. Overall global temperature appear increasing over
time; however, three features make standard changepoint analysis difficult. First, there exists seasonal fluctuation in the data, and these seem somewhat irregular. This implies the local trend is not flat but rather a smooth curve fluctuating through the months. Second, there is differing levels of variability over time. Third, there may be anomalies throughout the data as a result of certain global events.

Figure 4: Monthly Global Land Surface Air Temperature

As seen in the top-left plot of Figure 4, the underlying signal fluctuates over time as a result of irregular cyclical patterns over the years; these patterns have less variability than the longer term approximately linear time trends. This results in a very wiggly fit from the Bayesian dynamic linear model. Using the decoupled approach, we can visualize different fits of the projected posterior. The top-right plot of Figure 4 illustrates the mean of the “projected posterior” for \{0, 1, 2\} number of changepoints. As seen in the figure, as the number of changepoints increase, the fit becomes increasingly better. This is because increasing the number of changepoints essentially increases the degrees of freedom for the “projected posterior”. We select 1 as the optimal number of changepoints as it’s the simplest fit in which the upper 90% credible interval exceeds the 0.9 threshold. We see from bottom plot of Figure 4 that additional changepoints do not significantly increase $R^2_\lambda$ and seem to overfit the data. We find the single changepoint around the year 1963, after which there is a steeper long term slope.
4.2 Financial Data

For the next application, we apply the decoupled approach for identifying changepoint in financial stock returns. Stock returns represent a significant challenge for changepoint analysis due to its low signal-to-noise ratio. Many methods have been previously utilized for identifying changepoint in this area such as regime switching model (Hardy, 2001), penalized likelihood via dynamic programming (Bai and Perron, 2002), and Bayesian optimal stopping time model (Shiryaev et al., 2014). We choose to focus on Apple stocks as it's one of the largest companies in the market place and represent a significant portion of the market indices. Apple stock returns have been analyzed for changepoints in previous works (van den Burg and Williams, 2020; Lleo et al., 2020). Lleo et al. (2020) applied many change-point algorithms for an in-depth analysis for apple stock return; they primarily found breaks in variance. In this application, we will focus on the regression setting, analyzing changes in relationship between Apple and the market across the last 5 years. Building of foundations of Capital Asset Pricing Model (Fama and French, 2014), identifying changes in coefficient between a particular asset and the market can provide key information for portfolio optimization. Note that while we chose Apple return for this example, the same analysis can be performed for any other assets.

We collected daily closing price of Apple stock and S&P 500 over the last 4 years and calculate daily log return for both assets. We regressed daily Apple return on S&P 500 return to calculate the beta for the stock. Figure 5 illustrate the results. As seen, the return data for Apple and S&P 500 are quite noisy and exhibit regions of high and low volatility. This makes changepoint detection quite difficult. Competing methods such as CECPD and VPWBS identify no changepoints in this regression context. However, the decoupled approach identifies two changepoints at location 03/02/2020 and 07/22/2020. These dates mark the beginning of the pandemic where significant volatility occurred within the stock market. During this time period, the market movement became the dominated trend which led to coefficient of beta to be close to 1. These changepoints can be supported through a simple correlation analysis. The correlation of daily return between Apple and S&P 500 before 03/02/2020 is around 0.73; correlation from 03/02/2020 to 07/22/2020 is around 0.93 and correlation after 07/22/2020 is around 0.69. As seen, the returns become significantly more correlated during this time period which support beta estimated to be around 1. Unlike other time periods where changes in beta is gradual, we see drastic change in beta as a result of the pandemic which induces the estimation of changepoints. The decoupled approach is the only method to estimate changepoints in this period, further illustrating its robustness.

5 Conclusion

In conclusion, the paper proposes a decoupled approach for changepoint analysis that separates the processing of modeling and inference. First, a Bayesian dynamic linear model is fit to the underlying
The top-left plot shows the distribution of daily return over the past 4 years for Apple. The top-right plot shows the distribution of daily return over the past 4 years for S&P 500. The bottom-left plot shows the Bayesian DLM estimate (black) in comparison with a 22 day rolling ordinary least squares (red) for coefficient beta. The estimated credible bands for Bayesian DLM in shown in dark gray. As seen, the Bayesian DLM produces much a smoother estimate for \( \beta_t \). The bottom-right plot shows the resulting decoupled fit in blue line with two changepoints in blue vertical bars. The 95% credible bands for “projected posterior” is shown in dark gray and the 95% credible bands from Bayesian DLM is shown in light gray.

Bayesian DLM is a popular tool for time series analysis and can deal with many complexities inherent in the data such as outliers and heterogeneity. With incorporation of shrinkage priors, Bayesian DLM can fit a complex model to the underlying data while producing smooth estimates of the underlying coefficients. Second, utilizing the posterior draws of a Bayesian DLM, we derive a decoupled weighted least square loss to estimate the number of changepoints. The loss is composed of a goodness-of-fit term for posterior mean of the Bayesian DLM and a penalty term for the number of changepoints. The weights for each time-step is chosen to be the inverse of the estimated variance; making the approach more robust regions of high/low volatility.

As seen throughout the simulations and real world examples, the decoupled approach offers two key advantages over the competing method. First, by separating the process of modeling and inference, the decoupled approach allows for fitting of a highly complex Bayesian model to the underlying while still allowing for reasonable inference of changepoints. This allows the decoupled approach to deal with
many complexities inherent in time series. As the data becomes increasing complex, the decoupled approach can adapt the Bayesian DLM to deal with these issues while maintaining the same inference process for changepoints. Second, the decoupled approach is flexible in its ability to identify different types of changepoints. From the examples, we have shown that the decoupled approach’s ability to identify changes in mean, changes in relationship between predictors and changes in higher order trends. In addition, the decoupled approach can effectively identify changepoints in predictors in presence of covariates. Most other changepoint can only be utilized for one specific scenario. The flexibility of the decoupled approach allow the algorithm to be more adaptive to a wide variety of datasets.

References

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

Alley, R. B., Marotzke, J., Nordhaus, W. D., Overpeck, J. T., Peteet, D. M., Pielke, R., and et al. (2003). “Abrupt climate change,” Science 299, 2005–2010.

Aminikhanghahi, S., and Cook, D. J. (2017). “A survey of methods for time series change point detection,” Knowledge Information System 51, 339–367.

Bai, J., and Perron, P. (2002). “Computation and analysis of multiple structural change models,” Journal of Applied Econometrics 18, 1–22.

Banbura, M., Giannone, D., and Reichlin, L. (2010). “Large bayesian vector auto regressions,” Journal of Applied Econometrics 25, 71–92.

Bashir, A., Carvalho, C. M., Hahn, P. R., and Jones, M. B. (2019). “Post-processing posteriors over precision matrices to produce sparse graph estimates.,” Bayesian Analysis 14, 1075–1090.

Belmonte, M. A., Koop, G., and Korobilis, D. (2013). “Hierarchical shrinkage in time-varying parameter models,” Journal of Forecasting 33, 80–94.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). “Default bayesian analysis with global-local shrinkage priors,” Biometrika 103, 955–969.

Bitto, A., and Fruhwirth-Schnatter, S. (2019). “Achieving shrinkage in a time-varying parameter model framework,” Journal of Econometrics 210, 75–97.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). “Handling sparsity via the horseshoe,” AISTATS 5, 73–80.

Chan, J. C. C., and Eisenstat, E. (2018). “Bayesian model comparison for time-varying parameter vars with stochastic volatility,” Journal of Applied Econometrics 33, 509–532.
Ciuperca, G. (2014). “Model selection by lasso methods in a change-point model,” Stat. Papers 55, 349–374.

Ciuperca, G., and Maciak, M. (2019). “Change-point detection in a linear model by adaptive fused quantile method,” Scandinavian Journal of Statistics 47, 425–463.

Eckley, I. A., Fearnhead, P., and Killick, R. (2011). “Analysis of changepoint models,” Bayesian time series models 205–224.

Erdman, C., and Emerson, J. W. (2008). “A fast bayesian change point analysis for the segmentation of microarray data,” Bioinformatics 24, 2143–2148.

Fama, E. F., and French, K. R. (2014). “The capital asset pricing model: Theory and evidence,” Journal of Economic Perspectives 18, 25–46.

Fearnhead, P. (2006). “Exact and efficient bayesian inference for multiple changepoint problems,” Statistics and computing 16(2), 203–213.

Fearnhead, P., and Liu, Z. (2011). “Efficient bayesian analysis of multiple changepoint models with dependence across segments,” Statistics and Computing 21(2), 217–229.

Fearnhead, P., and Rigaill, G. (2019). “Changepoint detection in the presence of outliers,” Journal of the American Statistical Association 114, 169–183.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). “Regularization paths for generalized linear models via coordinate descent,” Journal of Statistical Software 33(1), 1–22.

Fryzlewicz, P. (2014). “Wild binary segmentation for multiple change-point detection,” Annals of Statistics 42, 2243–2281.

Green, P. J. (1995). “Reversible jump markov chain monte carlo computation and bayesian model determination,” Biometrika 82(4), 711–732.

Green, P. J. (2003). “Trans-dimensional markov chain monte carlo,” Oxford Statistical Science Series 179–198.

Guidolin, M., Hansen, E., and Pedio, M. (2019). “Cross-asset contagion in the financial crisis: A bayesian time-varying parameter approach,” Journal of Financial Markets 45, 83–114.

Hahn, P. R., and Carvalho, C. M. (2015). “Decoupling shrinkage and selection in bayesian linear models: A posterior summary perspective,” Journal of American Statistical Association 110, 435–448.

Harchaoui, Z., Moulines, E., and Bach, F. (2008). “Kernel change-point analysis,” Advances in Neural Information Processing Systems 27, 519–533.
Hardy, M. R. (2001). “A regime-switching model of long-term stock,” North American Actuarial Journal 5, 41–53.

Jeske, D. R., Montes De Oca, V., Bischoff, W., and Marvasti, M. (2009). “Cumsum techniques for timeslot sequences with applications to network surveillance,” Computational Statistics and Data Analysis 53, 4332–4344.

Jin, B., Wu, Y., and Shi, X. (2016). “Consistent two-stage multiple change-point detection in linear models,” Canadian Journal of Statistics 44, 161–179.

Kiers, H. A. L. (1997). “Weighted least squares fitting using ordinary least squares algorithm,” Psychometrika 62, 251–266.

Killick, R., Fearnhead, P., and Eckley, A. (2012). “Optimal detection of changepoints with a linear computational cost,” Journal of American Statistical Association 107, 1590–1598.

Kim, S., Shephard, N., and Chib, S. (1998). “Stochastic volatility: Likelihood inference and comparison with arch models,” Review of Economic Studies 65, 361–393.

Ko, S. I., Chong, T. T., and Ghosh, P. (2015). “Dirichlet process hidden markov multiple change-point model,” Bayesian Analysis 10(2), 275–296.

Kowal, D., Matteson, D., and Ruppert, D. (2019). “Dynamic shrinkage process,” J. R. Statist. Soc. B 81.

Kowal, D. R., and Bourgeois, D. C. (2020). “Bayesian function-on-scalars regression for high dimensional data,” .

Leonardi, F., and Bühlmann, P. (2016). “Computationally efficient change point detection for high-dimensional regression,” arXiv:1601.03704.

Liu, S., Yamada, M., Collier, N., and Sugiyama, M. (2013). “Change-point detection in time-series data by relative density-ratio estimation,” Neural Networks 43, 72–83.

Lleo, S., Ziemba, W. T., and Li, J. (2020). “Exploring breaks in the distribution of stock returns: Empirical evidence from apple inc,” Available at SSRN: https://ssrn.com/abstract=3700419.

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). “On optimal multiple changepoint algorithms for large data,” Statistics and Computings 27, 519–533.

Matteson, D. S., and James, N. A. (2014). “A nonparametric approach for multiple change point analysis of multivariate data,” Journal of the American Statistical Association 109, 334–345.
Nakajima, J., Kasuya, M., and Watanabe, T. (2011). “Bayesian analysis of time-varying parameter vector autoregressive model for the japanese economy and monetary policy,” Journal of The Japanese and International Economies 25, 225–245.

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

Pichuka, S., and Maity, R. (2018). “Development of a time-varying downscaling model considering non-stationarity using a bayesian approach,” International Journal of Climatology 38, 3157–3176.

Puelz, D., Carvalho, C. M., and Hahn, P. R. (2015). “Optimal etf selection for passive investing,” .

Qian, J., and Su, L. (2016). “Shrinkage estimation of regression models with multiple structural changes,” Econometric Theory 32, 1376–1433.

Shiryaev, A. N., Zhitlukhin, M. V., and Ziemba, W. T. (2014). “Land and stock bubbles, crashes and exit strategies in japan circa 1990 and in 2013,” Quantitative Finance 15, 1449–1469.

Siems, T., Hellmuth, M., and Liebscher, V. (2019). “Simultaneous credible regions for multiple change-point locations,” Journal of Computational and Graphical Statistics 28(2), 290–298.

van den Burg, G. J. J., and Williams, C. K. I. (2020). “An evaluation of change point detection algorithms,” .

Wang, D., Zhao, Z., Lin, K. Z., and Willett, R. (2021). “Statistically and computationally efficient change point localization in regression settings,” Journal of Machine Learning Research 22, 1–46.

Woody, S., Carvalho, C. M., and Murray, J. S. (2020). “Model interpretation through lower-dimensional posterior summarization,” .

Yang, Y., and Song, Q. (2014). “Jump detection in time series nonparametric regression models: a polynomial spline approach,” Annals of the Institute of Statistical Mathematics 66, 325–344.

Yuan, M., and Lin, Y. (2007). “Model selection and estimation in regression with grouped variables,” Journal of the Royal Statistical Society, Series B 68, 49–67.

Zhang, B., Geng, J., and Lai, L. (2015). “Multiple change-points estimation in linear regression models via sparse group lasso,” IEEE Transactions on Signal Processing 63.