



# Bayesian Change Point Detection with Spike-and-Slab Priors

Lorenzo Cappello<sup>a</sup>, Oscar Hernan Madrid Padilla<sup>b</sup>, and Julia A. Palacios<sup>c</sup>

<sup>a</sup>Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain; <sup>b</sup>Department of Statistics, University of California, Los Angeles, Los Angeles, CA; <sup>c</sup>Departments of Statistics and Biomedical Data Science, Stanford University, Stanford, CA

## ABSTRACT

We study the use of spike-and-slab priors for consistent estimation of the number of change points and their locations. Leveraging recent results in the variable selection literature, we show that an estimator based on spike-and-slab priors achieves optimal localization rate in the multiple offline change point detection problem. Based on this estimator, we propose a Bayesian change point detection method, which is one of the fastest Bayesian methodologies. We demonstrate through empirical work the good performance of our approach vis-a-vis some state-of-the-art benchmarks. Interestingly, despite having a Gaussian noise assumption, our approach is more robust to misspecification of the error terms than the competing methods in numerical experiments. Supplementary materials for this article are available online.

## ARTICLE HISTORY

Received December 2021  
Accepted February 2023

## KEYWORDS

Approximate inference;  
Optimality; Robust;  
Shrinkage

## 1. Introduction

Change point detection has received considerable attention in the statistical literature for several decades. Assume we observe a vector of independent random variables  $Y = (Y_1, \dots, Y_T)^\top$  according to the linear model

$$Y_t = f_t + \epsilon_t \quad \text{for } 1 \leq t \leq T, \quad (1)$$

where  $f_t$  is a right continuous function with an unknown number  $K$  of change points, and  $\epsilon_t$ 's are independent random variables with  $E(\epsilon_t) = 0$  for all  $t$ . The main goal of offline change-point detection is to simultaneously estimate  $K$  and the locations of the change points of  $f_t$ . A natural assumption is that  $f_t$  (i.e., the conditional mean) is sparse, in the sense that there are only a small number of change-points. We further assume that  $f_t$  is piecewise-constant (i.e., a right continuous step function). These modeling choices lead to a low-dimensional parametric model that is interpretable, can fit nonstationary time-series, and is suitable for prediction.

The literature on change point detection includes a large number of frequentist methods. Most of these methodologies rely on a test statistic to detect parametric changes in the distribution of the observables and model selection techniques to determine the number of parameters defining the signal  $f_t$ . Some examples of test statistics include the likelihood ratio and the CUSUM statistic (Page 1954). The model selection step typically relies on either an  $\ell_0$  or an  $\ell_1$  penalty. The penalty is included directly through a penalized likelihood or via an information criterion such as AIC and BIC. Consequently, several variable selection methodologies have a conceptual analogue in change point detection: for example, the Dantzig selector of Candès et al. (2007) has the same rationale as the multiscale SMUCE estimator of Frick, Munk, and Sieling (2014); the total

variation denoising (Rudin, Osher, and Fatemi 1992) and the fused LASSO procedure (Tibshirani et al. 2005) share the same penalty with the LASSO (Tibshirani 1996). Other frequentist methods for univariate change point detection include the wild binary segmentation of Fryzlewicz (2014) based on the CUSUM statistics, and various algorithms for  $\ell_0$  penalized change point detection (Friedrich et al. 2008; Rigai 2010; Killick, Fearnhead, and Eckley 2012; Maidstone et al. 2017).

From the Bayesian perspective, popular change point detection methods rely on product partition models (Barry and Hartigan 1992, 1993). However, the use of MCMC to approximate the posterior distributions of these models is challenging (Chib 1996, 1998), and much research has focused on alternatives to MCMC: Fearnhead (2006) proposed two algorithms to perform direct simulation from the posterior distribution (one to do exact simulation from the posterior and one using an approximate version); Rigai, Lebarbier, and Robin (2012) derived exact formulas for the posterior distribution. Recent works take an empirical Bayes approach to set the prior distributions (Du, Kao, and Kou 2016; Liu, Martin, and Shen 2017). Liu, Martin, and Shen (2017) is more general, allowing to recover piecewise polynomial signals. To the best of our knowledge, Bayesian variable selection procedures such as the horseshoe prior (Carvalho, Polson, and Scott 2010) and the spike-and-slab prior (Mitchell and Beauchamp 1988), have not been studied in this setting. Recent works used the horseshoe prior for trend filtering (Faulkner and Minin 2018; Kowal, Matteson, and Ruppert 2019) but not to explicitly infer  $K$  and the change points locations.

In this article, we study spike-and-slab priors for offline multiple change point detection. Starting from a baseline  $f_0$ , we model each increment through the operator  $\Delta f_i = f_i - f_{i-1}$  for  $1 \leq i \leq T$  and introduce latent binary variables  $(Z_1, \dots, Z_T)^\top$  to indicate whether  $\Delta f_i$  corresponds to a change point or not.

The prior distribution on the increment  $\Delta f_i$  under  $Z_i = 0$  is a distribution “concentrated” around 0 (or a point mass) called spike. The prior distribution for  $\Delta f_i$  under  $Z_i = 1$  is a diffused distribution called slab. The choice of which distributions to use for the spike, the slab, and the model space, has been a subject of extensive research; see Bhadra et al. (2019) for a recent review. We use the shrinking and diffusing prior of Narisetty and He (2014), which consists of Gaussian spike-and-slab priors with sample size dependent prior variances. The reasons for this choice are (i) Narisetty and He (2014) proved one of the strongest selection consistency results in the Bayesian variable selection literature, and (ii) Chen and Walker (2019) recently proposed a methodology for fast Bayesian variable selection employing this prior and not requiring MCMC.

Here, we make the following contributions to the change point detection literature. (i) We show how to employ spike-and-slab priors for change point detection and propose a fast algorithm capable of detecting change points from thousands of observations in under a second. (ii) We establish that a modified estimator based on the shrinking and diffusing prior is consistent and achieves optimal localization rates of multiple change points. (iii) Through simulations, we show that our procedure is competitive with state-of-the-art methodologies. A salient feature of our proposed method is that it does not rely on MCMC, avoiding the risk of MCMC chains failing to converge. Additionally, in numerical experiments, it is highly robust to misspecification of the noise term, a situation where many state-of-the-art benchmarks fail by substantially overestimating the number of change points. This is somewhat surprising given that the method is developed under the Gaussian noise assumption.

While in this article we study the univariate change point detection problem, methods for change point detection have also been studied for other types of data beyond univariate mean change point detection and settings more general than (1). Pein, Sieling, and Munk (2017) considered change point detection with heterogeneous noise. Carlstein (1988), Rizzo and Székely (2010), Zou et al. (2014), Matteson and James (2014), and Padilla et al. (2019a, 2019b, 2019c) developed nonparametric change point methods that can detect arbitrary changes in distribution. Cho and Fryzlewicz (2015), Cho (2016), and Wang and Samworth (2018) focused on high-dimensional change point estimators. Aue et al. (2009), Avanesov and Buzun (2018), and Wang, Yu, and Rinaldo (2021) studied covariance change point detection. Fearnhead and Rigai (2018) considered methods for change point detection combining a robust loss with the  $\ell_0$  penalty. Vanegas, Behr, and Munk (2022) proposed a multiscale method for quantile change point detection. Here, we focused on a simpler setting because spike-and-slab priors were not studied in the change point detection context, and, as stated by Wang, Yu, and Rinaldo (2020), the estimators built for (1) are often the building blocks for more complex settings.

The rest of the article is organized as follows. Section 2 describes the model, conditions on the parameters, and introduces the algorithm. Section 3 presents our main results on consistency. We present simulation studies in Section 4 to illustrate how our proposal fares with alternatives. Section 5 includes applications to microarray and ion channel data. Section 6 concludes with a discussion on the use of spike-and-slab priors for multiple change point detection.

## 2. Method

We assume that  $(\epsilon_t)_{1:T}$  are independent Gaussian random variables with mean zero and known variance  $\sigma^2$ . Let  $\mathcal{C}^* := \{\eta_1, \dots, \eta_K\} \subset \{1, \dots, T\}$  denote the set of change points of the piecewise-constant right continuous signal  $f_t$ , we can write this as

$$f_t = \begin{cases} \mu_0 & \text{if } t \leq \eta_1 - 1 \\ \mu_1 & \text{if } \eta_1 \leq t \leq \eta_2 - 1 \\ \vdots & \\ \mu_K & \text{if } \eta_K \leq t \leq T, \end{cases} \quad (2)$$

and assume that  $f_0 = \mu_0 = 0$ . Further technical conditions on  $f_t$  will be given in Section 3. We include a discussion on how to handle unknown  $\sigma$  and  $f_0 \neq 0$  later in the section.

*Basad.cp change point detection.* A working model employing spike-and-slab priors is as follows:

$$\begin{aligned} Y_t | f_t, \sigma^2 &\sim \mathcal{N}(f_t, \sigma^2) \\ \Delta f_t | \sigma^2, Z_t = 0, \tau_{0,T}^2 &\sim \mathcal{N}(0, \sigma^2 \tau_{0,T}^2), \\ \Delta f_t | \sigma^2, Z_t = 1, \tau_{1,T}^2 &\sim \mathcal{N}(0, \sigma^2 \tau_{1,T}^2), \\ \text{pr}(Z_t = 1) &= 1 - \text{pr}(Z_t = 0) = q_T, \end{aligned} \quad (3)$$

for  $t$  from 1 to  $T$ ,  $\Delta f_t = f_t - f_{t-1}$ , and  $\tau_{0,T}^2, \tau_{1,T}^2$ , and  $q_T$  are hyperparameters that depend on  $T$ , with  $\tau_{1,T}^2 \gg \tau_{0,T}^2$ . The rationale behind this set-up is that the posterior probability of  $Z_t = 1$  should be high for  $t \in \mathcal{C}^*$ ; vice versa, the posterior probability of  $Z_t = 0$  should be high for  $t \notin \mathcal{C}^*$ . A natural change point detection procedure is to employ the posterior probabilities of  $Z_t$  to determine if  $t$  is a change point or not, for example classifying  $t$  as a change point if the posterior exceeds a certain threshold. We elaborate on this selection rule later. The assignment of the same prior for  $\Delta f_1$  follows from the assumption that the mean process starts at zero.

A relevant difference with variable selection comes from the fact that while covariates are not ordered, in change point detection we generally want to avoid classifying consecutive time instances as change points. We expect this behavior because it is a common feature in the change point detection literature. Most of the procedures employ minimum spacing conditions, that is, the distance between consecutive change points is lower bounded by a quantity  $\Delta > 0$ , such that  $|\eta_j - \eta_{j+1}| > \Delta$ . Minimum spacing conditions are used both in the finite sample implementations of the estimators and in the proofs of consistency. We will introduce a procedure to avoid consecutive change points.

The model (3), which we will refer to as *basad.cp*, is the analogue of the *basad* variable selection procedure of Narisetty and He (2014) to change point detection. The sample size dependent hyperparameters  $\tau_{0,T}^2, \tau_{1,T}^2$  and  $q_T$ , are the salient feature of (3). We require that as  $T \rightarrow \infty$ ,  $\tau_{0,T}^2 \rightarrow 0$  and  $\tau_{1,T}^2 \rightarrow \infty$ . In variable selection, a shrinking  $\tau_{0,T}^2$  ensures that the marginal posterior probability of including (excluding) an active (inactive) covariate converges to one as sample size increases. Increasing  $\tau_{1,T}^2$  and  $q_T$  allows for the consistent estimation of the number of active covariates and consistent model selection. Narisetty and He (2014) proved that the penalization achieved through  $\tau_{1,T}^2$  and  $q_T$  is equivalent to an explicit  $\ell_0$  penalty. We will show in Section 3 that these parameters play a similar role for achieving

consistent estimation of  $K$  and  $\eta_1, \dots, \eta_K$ . A similar asymptotic result holds despite the settings being very different: in high-dimensional variable selection, the number of covariates grows at a rate faster than the number of samples; in change point detection, the number of piecewise increments is equal or smaller than the sample size.

Another key feature of the methodology of Narisetty and He (2014) is that they employ the marginal posterior probabilities  $\text{pr}(Z_t = 1|Y, \sigma^2)$  to select the active parameters in the finite sample implementation (to prove consistency they employ  $\text{pr}(Z|Y, \sigma^2)$ , with  $Z = (Z_1, \dots, Z_T)^\top$ ). The idea is motivated by computational reasons, given that one can sample from the marginals with a Gibbs sampler, which is not available for  $\text{pr}(Z|Y, \sigma^2)$ . Furthermore, the MCMC targeting  $\text{pr}(Z|Y, \sigma^2)$  has a much larger space of models and struggles to explore it, for example, Fearnhead (2006) and Rigai, Lebarbier, and Robin (2012). For this reason, we will employ marginals for change point detection.

We classify a time instance  $t$  as a change point if  $\text{pr}(Z_t = 1|Y, \sigma^2)$  is larger than a certain threshold. In this case, the estimated number of change points  $\hat{K}$  is the number of marginal posterior probabilities larger than the chosen threshold. The model selected using 0.5 as a threshold corresponds to the median probability model of Barbieri and Berger (2004), who also proved that it is the optimal predictive model. An alternative strategy to select the change points would be to first rank the  $\Delta f_t$  based on  $\text{pr}(Z_t = 1|Y, \sigma^2)$ , and then select the top  $\hat{K}$  (the model size) increments according to a given information criteria. We do not investigate this strategy and leave it for future work. We further stress that, under our selection rule, it is likely that there will be consecutive time instances such that  $\text{pr}(Z_t = 1|Y, \sigma^2) > 0.5$ , that is, consecutive points could be “classified” as change points. We deal with this issue after introducing an alternative methodology to compute  $\text{pr}(Z_t = 1|Y, \sigma^2)$ .

*Solo.sp change point detection.* Chen and Walker (2019) introduced a sequential procedure based on a misspecification of *basad* that admits marginal posterior probabilities in closed form. Their method, called *solo spike-and-slab*, has asymptotic properties and empirical accuracy similar to *basad*, while being substantially faster. While our setting can be seen as a particular instance of the linear regression framework of Chen and Walker (2019), there is an advantage in deriving the closed form marginal  $\text{pr}(Z_t = 1|Y, \sigma^2)$  for the setting considered in this article because we will be able to simplify certain calculations further. We do this following the same steps of Chen and Walker (2019). Suppose we are interested in testing whether a time instance  $j$  is a change point and consider the following model:

$$\begin{aligned} Y_t|f_t, \sigma^2 &\sim \mathcal{N}(f_t, \sigma^2), \quad t = 1 \dots, T, \\ \Delta f_j|\sigma^2, Z_j = 0, \tau_{0,T}^2 &\sim \mathcal{N}(0, \sigma^2 \tau_{0,T}^2), \\ \Delta f_j|\sigma^2, Z_j = 1, \tau_{1,T}^2 &\sim \mathcal{N}(0, \sigma^2 \tau_{1,T}^2), \\ \Delta f_i|\sigma^2, \tau_T^2 &\sim \mathcal{N}(0, \sigma^2 \tau_T^2), \quad i \in \{1 \dots, T\} \setminus \{j\}, \\ \text{pr}(Z_j = 1) &= 1 - \text{pr}(Z_j = 0) = q_T, \end{aligned} \quad (4)$$

where  $\tau_T^2$  is an additional sample size dependent hyperparameter. Thus, we place a spike-and-slab prior on a single change point at a time (in this case  $\Delta f_j$ ), and place conjugate Gaussian priors on the remaining terms. The tuning parameter  $\tau_T^2$  controls the shrinkage across time instances. The advantage

of model (4) is that it allows us to write the marginal posterior probabilities in closed-form. First, we can marginalize out  $\Delta f_{-j} := (\Delta f_1, \dots, \Delta f_{j-1}, \Delta f_{j+1}, \dots, \Delta f_T)^\top$  in the likelihood  $\mathcal{L}(Y|\Delta f_{-j}, \Delta f_j, \sigma^2)$  to compute the marginal likelihood

$$\mathcal{L}(Y|\Delta f_j, \sigma^2) \propto \exp \left\{ \frac{-\Delta f_j^2 (T - j - \sum_{i=j+1}^T n'_i) \gamma_{jj} + 2 \Delta f_j \bar{y}_{jj}'}{2\sigma^2} \right\} \quad (5)$$

where we have integrated out the random vector  $\Delta f_{-j}$  using the conjugate prior distribution given in (4). The parameters  $n'_{j+1}, \dots, n'_T, \gamma_{jj}$  and  $\bar{y}_{jj}'$  are function of the hyperparameters of the prior on  $\Delta f_{-j}$  and data  $Y_{1:T}$ , and are computed recursively as follows. Initialize  $n'_T = \tau_T^2 / (\tau_T^2 + \sigma^2)$ ,  $\bar{y}'_T = y_T$ , and compute recursively for  $i = T - 1$  to  $j + 1$

$$\begin{aligned} n'_i &= \frac{\tau_T^2 (T - i - 1 - \sum_{k=i+1}^T n'_k)^2}{\tau_T^2 (T - i - 1 - \sum_{k=i+1}^T n'_k) + \sigma^2}, \quad \text{and} \\ \bar{y}'_i &= \frac{\sum_{t=i}^T y_t - \sum_{k=i+1}^T n'_k \bar{y}'_k}{T - i - 1 - \sum_{k=i+1}^T n'_k}. \end{aligned} \quad (6)$$

Then, set  $\gamma_{1,j}$ , and for  $i = 1$  to  $j$  compute

$$\begin{aligned} n''_{i,j} &= \frac{\tau_T^2}{\tau_T^2 \gamma_{i,j} (T - i - 1 - \sum_{k=j+1}^T n'_k) + \sigma^2}, \\ \bar{y}''_{i,j} &= \sum_{t=i}^T y_t - \sum_{k=i+1}^T n'_k \bar{y}'_k - \left( T - i - 1 - \sum_{k=j+1}^T n'_k \right) \\ &\quad \sum_{k=1}^{i-1} n''_{k,j} \gamma_{k,j} \bar{y}''_{k,j}, \\ \gamma_{i+1,j} &= 1 - \left( T - i - 1 - \sum_{k=j+1}^T n'_k \right) \left( \sum_{k=1}^i n''_{k,j} \gamma_{k,j}^2 \right). \end{aligned} \quad (7)$$

Despite the involved notation, simple calculations lead to the definitions of parameters in (6) and (7). The basic idea is first to marginalize  $\Delta f_T$ , then  $\Delta f_{T-1}$ , then continue sequentially to  $\Delta f_{j+1}$ . This first step leads to the definition of the parameters in (6). In the second step, we first marginalize  $\Delta f_1$  and then recursively integrate out the remaining parameters until  $\Delta f_{j-1}$ . This second step leads to the definition of the parameters in (7). Equations (6)–(7), together with the priors on  $\Delta f_{-j}$ , define a vector of weights: observations recorded at time points closer to  $j$  have a higher weight than those further away from  $j$ . Essentially, what we have done is marginalizing two Gaussian Markov random fields (GMRF) of order 1 with parameter  $\tau_T^2$ , one to the left and one to the right of  $j$ ; see Rue and Held (2005) for an intro to GMRFs.

Notice that the quantities  $\{n'_i\}$  can be computed in  $O(T)$  as follows. First, we compute and store the quantities

$$\left( \bar{y}'_i, \sum_{k=i+1}^T n'_k, \sum_{k=i+1}^T n'_k \bar{y}'_k, \sum_{t=i}^T y_t \right), \quad i = T, \dots, 1,$$

which can be done in  $O(T)$ . Then a single pass using (6) allows us to obtain  $\{n'_i\}$  with  $O(T)$  cost. As for (7), we start by computing

$$\sum_{k=j+1}^T n'_k \quad \text{for } j = T - 1, \dots, 1.$$

Next, for a fix  $j$ , we calculate  $n''_{i,j}$  for  $i = 1, \dots, i$ . Then we store and compute

$$\left( \gamma_{i,j}, \sum_{k=1}^{i-1} n''_{k,j} \gamma_{k,j} \overline{y''_{k,j}}, \sum_{k=1}^i n''_{k,j} \gamma_{k,j}^2 \right), \quad i = 1, \dots, j.$$

Hence, the calculations in (7) can be done in  $O(T^2)$ .

Given the marginal likelihood  $\mathcal{L}(Y|\Delta f_j, \sigma^2)$ , we can compute the marginal posterior distribution of  $\Delta f_j$  through Bayes rule:

$$\Delta f_j | Y, \sigma^2, q_T, \tau_{0,T}^2, \tau_{1,T}^2 \sim (1 - q_T) \omega_{0,j} N(\mu_{0,j}, \xi_{0,j}^2) + q_T \omega_{1,j} N(\mu_{1,j}, \xi_{1,j}^2), \quad (8)$$

where the parameters are defined as follows, for  $k \in \{0, 1\}$ :

$$\begin{aligned} \mu_{k,j} &= \frac{\overline{y''_{j,j}}}{(T - j - 1 - \sum_{k=j+1}^T n'_k) \gamma_{j,j} + \tau_{k,T}^{-2} \sigma^2}, \\ \xi_{k,j} &= \frac{\sigma^2}{(T - j - 1 - \sum_{k=j+1}^T n'_k) \gamma_{j,j} + \tau_{k,T}^{-2} \sigma^2}, \\ \omega_{0,j} &= \exp \left\{ \frac{1}{2\sigma^2} \frac{\overline{y''_{j,j}}^2}{(T - j - 1 - \sum_{k=j+1}^T n'_k) \gamma_{j,j} + \tau_{k,T}^{-2} \sigma^2} \right\} \\ &\quad \left\{ \frac{\tau_{k,T}^{-2}}{(T - j - 1 - \sum_{k=j+1}^T n'_k) \gamma_{j,j} + \tau_{k,T}^{-2} \sigma^2} \right\}^{1/2}. \end{aligned} \quad (9)$$

The parameters above are all we need in order to compute  $\text{pr}(Z_j = 1|Y, \sigma^2)$ , which corresponds to

$$\text{pr}(Z_j = 1|Y, \sigma^2) = \frac{q_T \omega_{1,j}}{q_T \omega_{1,j} + (1 - q_T) \omega_{0,j}}. \quad (10)$$

Here,  $q_T$  is the sparsity inducing parameter. Given (10), we follow the same procedure described for *basad.cp*: a time instance is declared a change point if  $\text{pr}(Z_j = 1|Y, \sigma^2)$  exceeds the prespecified threshold. In practice, we would not be interested only in a single time instance  $j$ , so one needs to compute (10) for  $j$  in 1 to  $T$ . That is, we need to consider  $T$  models.

Equations (5), (8), and (10) are the analogues to (7), (11) and (19) in Chen and Walker (2019). Similarly, the definitions of parameters in (9) are the analogues of (12)–(14) from Chen and Walker (2019). The differences arise because their definitions rely on a matrix of covariates and require several matrix multiplications and inversions. Importantly here, we can write analytically all the formulas and bypass the need for these matrix operations.

*Solo.cp* notably differs from approaches based on generalized likelihood ratio, such as the CUSUM statistics. While CUSUM-based methods fit two constant means model (to the left and to the right of  $j$ ), *solo.cp* seeks for evidence for a jump in mean at time  $j$  after having accounted for all possible variations from 1 to  $j - 1$  and  $j + 1$  to  $T$ . This is done through the random vector  $\Delta f_{-j}$ , which we integrate out when computing the marginal likelihood. This difference is an essential feature of *solo.cp* and enables the detection of multiple change points. The idea is that a change in mean at time  $\eta_i$  will be marginalized out before getting to time  $\eta_{i+1}$ , as long as  $\eta_{i+1}$  and  $\eta_i$  are not “too close”. Such minimum spacing condition is standard in the literature.

In Section 4, we will show empirically that *solo.cp* is more robust to misspecification of the noise model and outliers than competitors. We conjecture that this property follows from the feature described above and the shrinkage implicit in Bayesian estimators. “Outliers” are shrunk toward multiple local means: An observation at time  $j$  contributes to estimating (or marginalizing) all increments from  $\Delta f_1$  to  $\Delta f_j$ . A potential drawback of this construction is that changes of very small magnitude are integrated out. Both hypotheses will be tested numerically.

**Implementation.** Regardless of whether we compute  $\text{pr}(Z_i = 1|Y, \sigma^2)$  through *basad.cp* or *solo.cp*, we propose the use of a post-processing step to avoid the detection of consecutive change points. This involves a rule that defines when two or more estimates will be considered as “consecutive”, and a selection rule to determine which estimates to keep.

In detail, let  $\widehat{\mathcal{C}}_0 := \{\widehat{\eta}_1, \dots, \widehat{\eta}_{\widehat{K}}\}$  be the set of points such that  $\text{pr}(Z_i = 1|Y, \sigma^2) > 0.5$ . Now, fix  $\Delta \in N$ , and partition  $\widehat{\mathcal{C}}_0$  into nonempty subsets  $\widehat{\mathcal{C}}_0^1, \dots, \widehat{\mathcal{C}}_0^{\widehat{K}}$  such that for all  $\eta \in \widehat{\mathcal{C}}_0^i$  there exists at least one  $\eta' \in \widehat{\mathcal{C}}_0^i$  (if  $\widehat{\mathcal{C}}_0^i \setminus \{\eta\} \neq \emptyset$ ) such that  $|\eta - \eta'| \leq \Delta$  and no  $\eta'' \in \widehat{\mathcal{C}}_0 \setminus \widehat{\mathcal{C}}_0^i$  such that  $|\eta - \eta''| \leq \Delta$ . Hence, the partition defines the notion of “consecutive change points”. Finally, within each subset  $\widehat{\mathcal{C}}_0^i$ , choose the point

$$\eta_i = \arg \max_{\eta \in \widehat{\mathcal{C}}_0^i} \text{pr}(Z_\eta = 1|Y, \sigma^2).$$

The estimated set of change points is  $\widehat{\mathcal{C}} := \{\eta_1, \dots, \eta_{\widehat{K}}\}$ .

A few remarks. First, the length of the partition determines the number of estimated change points  $\widehat{K}$ . Second, we need an extra parameter  $\Delta$  to define the partition of  $\widehat{\mathcal{C}}_0$ . The sensitivity of the two methods to  $\Delta$  is studied in the [supplementary material \(SM\)](#). Lastly, we pick the time instance having the maximum marginal posterior probability within each subset and classify it as the change point. Whereas this is an arbitrary criterion, choosing the point that maximizes a given test statistics is stan-

---

#### Algorithm 1 Spike-and-slab change point detection

---

**Inputs:**  $Y, T, \Delta, \sigma$

**Output:**  $\widehat{\mathcal{C}}, \widehat{K}$

---

1. Compute  $P(Z_1 = 1|Y, \sigma^2), \dots, P(Z_T = 1|Y, \sigma^2)$

**If** *basad.cp*

- Approximate  $P(Z_1 = 1|Y, \sigma^2), \dots, P(Z_T = 1|Y, \sigma^2)$  with the Gibbs sampler defined in Narisetty and He (2014).

**If** *solo.cp* **for**  $i = 1, \dots, T$  **do**

- Compute posterior parameters (9)
- Compute  $P(Z_i = 1|Y, \sigma^2)$  through (10).

2. Define  $\widehat{\mathcal{C}}_0 := \{i : P(Z_i = 1|Y, \sigma^2) > 0.5\}$

3. Partition  $\widehat{\mathcal{C}}_0$  into subsets of nonconsecutive change points  $\widehat{\mathcal{C}}_0^1, \dots, \widehat{\mathcal{C}}_0^{\widehat{K}}$  (see the main text)

4. Set  $\widehat{\mathcal{C}} := \{\eta_i = \arg \max_{\eta \in \widehat{\mathcal{C}}_0^i} P(Z_\eta = 1|Y, \sigma^2), \text{ for } i = 1, \dots, \widehat{K}\}$

---



dard in the change point detection literature; for example, Fryzlewicz (2014).

**Algorithm 1** summarizes our methodologies. An input of the algorithm is  $\sigma$ , as it is commonly done in the change-point literature; for example, Fryzlewicz (2014), Fang, Li, and Siegmund (2020), and Fearnhead and Rigai (2018). In the setting considered (piece-wise constant signal, Gaussian noise), there are many good estimators of the standard deviation, see Fang, Li, and Siegmund (2020). For example, *wbs* (Fryzlewicz 2014) and *r-fpop* (Fearnhead and Rigai 2018) employ the median absolute deviation estimator, *ebpice* of Liu, Martin, and Shen (2017) uses the fused LASSO residuals standard deviation computed through the “one standard error” rule (Tibshirani et al. 2005).

Both *basad.cp* and *solo.cp* can be easily extended to unknown  $\sigma$ . This instance is considered both by Narisetty and He (2014) and Chen and Walker (2019): one simply includes an inverse gamma prior in (3) and (4). *solo.cp* remains analytically tractable: the difference is that (8) is a mixture of Student’s  $t$ -distributions, whose parameters can be written in closed form. In this article, we do not study *solo.cp* and *basad.cp* with inverse-gamma priors, but with the fused LASSO residuals standard deviation computed through the “one standard error” rule.

In this section, we considered a mean-process starting at zero. In the case  $f_0 \neq 0$ , we can treat it as an unknown and assigning to it a Gaussian prior with mean  $\mu_0$  and variance  $\tau_0^2$ . These hyperparameters will generally be different to the prior on the increments  $(\Delta f_i)_{i \geq 1}$  because  $f_0$  is not an increment but a baseline mean. The choice of  $\mu_0$  and  $\tau_0^2$  can be based on prior information and how informative we want the prior to be. To prove the theory, we will assume that  $f_0 = 0$  holds and use the same prior in numerical experiments. If the assumption is violated, then  $\text{pr}(Z_1 = 1 | Y, \sigma^2) \rightarrow 1$  with high probability.

So far, we considered a setting where the number of observations collected at a given point  $t$  ( $n_t$ ) is equal to one (case  $n_t = 1$ ), which is the standard in the literature. In applications, this may not be the case (case  $n_t > 1$ ). This situation could arise if multiple observations are collected at once, or if observations are collected at distinct time points, but the reported data are binned into time intervals. To our knowledge, there are few methods in the literature dealing with this situation (Padilla et al. 2019b). The extension of *basad.cp* and *solo.cp* to the case  $n_t > 1$  is straightforward. Parameters (6), (7), and (9) can be written in closed form, including an adjustment done through  $n_t$  for all  $t$ . The explicit formulas are provided in the SM, where we also show numerically that the case  $n_t > 1$  is particularly beneficial in terms of empirical performance for the two methodologies discussed in this section.

### 3. Theory

In this section, we provide some theoretical support for the method that we study in this article. We show that for the task of multiple change point detection, a modified version of the estimator described in (3) (*basad.cp*) based on the spike-and-slab variable selection framework of Narisetty and He (2014) leads to optimal localization rates of the change points. Specifically, for the case of a bounded number of change points, under a slightly weaker signal-to-noise condition than the wild binary segmentation and  $\ell_0$  based methods, we attain optimal

localization rates. We also show in the SM that this optimality is preserved in the single change point detection framework if we consider a version of the *solo.cp* estimator. All the proofs can be found in the SM.

We consider a modified version of the *basad.cp* estimator defined as follows. Let  $m \in \mathbb{N}$  with  $m \leq T$  and consider  $\Lambda_1, \dots, \Lambda_m$  a partition of  $\{1, \dots, T\}$  such that  $|\Lambda_j| = T/m$  for all  $j \in \{1, \dots, m\}$ . Let  $\tilde{Y} \in \mathbb{R}^m$  satisfy

$$\tilde{Y}_j = \frac{1}{|\Lambda_j|^{1/2}} \sum_{i \in \Lambda_j} Y_i,$$

for  $j = 1, \dots, m$ . We also define

$$\tilde{f}_j = \frac{1}{|\Lambda_j|^{1/2}} \sum_{i \in \Lambda_j} f_i,$$

for  $j = 1, \dots, m$ . It is convenient to rewrite (3) for the data vector  $\tilde{Y}$  as

$$\begin{aligned} \tilde{Y} | \tilde{f}, \sigma^2 &\sim \mathcal{N}(\tilde{f}, \sigma^2 I_m), \\ \Delta \tilde{f}_j | \sigma^2, Z_j = 0, \tau_{0,m}^2 &\sim \mathcal{N}(0, \sigma^2 \tau_{0,m}^2), \\ \Delta \tilde{f}_j | \sigma^2, Z_j = 1, \tau_{1,m}^2 &\sim \mathcal{N}(0, \sigma^2 \tau_{1,m}^2), \\ \text{pr}(Z_j = 1) &= 1 - \text{pr}(Z_j = 0) = q_m, \\ &\quad j = 1, \dots, m, \end{aligned} \quad (11)$$

where  $q_m, \tau_{0,m}, \tau_{1,m} > 0$ . Furthermore,  $\Delta \tilde{f}_1 = \tilde{f}_1$ , and  $\Delta \tilde{f}_j = \tilde{f}_j - \tilde{f}_{j-1}$  for  $j = 2, \dots, m$ .

The goal is to define an estimator  $\hat{\mathcal{C}} \subset \{1, \dots, T\}$  of  $\mathcal{C}^*$ . We do this by first defining an estimator  $\tilde{\mathcal{C}}$  relying on the indexes of the partition  $1, \dots, m$ , then we use  $\tilde{\mathcal{C}}$  to construct our actual estimator. First, let  $\tilde{Z} \in \{0, 1\}^m$  be such that

$$\tilde{Z}_j := \begin{cases} 1 & \text{if } \text{pr}(Z_j = 1 | \tilde{Y}, \sigma^2) > 0.5, \\ 0 & \text{otherwise,} \end{cases}$$

and  $\tilde{\mathcal{C}} = \{j : \tilde{Z}_j = 1\}$ . The set  $\tilde{\mathcal{C}}$  is then used to construct  $\hat{\mathcal{C}} \subset \{1, \dots, T\}$  as follows:

- If  $i \in \hat{\mathcal{C}}$  then there exists a  $j \in \tilde{\mathcal{C}}$  with  $i \in \Lambda_j$ .
- If  $j \in \tilde{\mathcal{C}}$  then for a unique  $i \in \hat{\mathcal{C}}$  we have that  $i \in \Lambda_j$ .

The construction of  $\hat{\mathcal{C}}$  is used to map the estimates conditioned on the transformed data to the actual time indices we are trying to infer. Our results show that  $\hat{\mathcal{C}}$  defined by the modified estimator based on  $\tilde{Y}$  and  $\tilde{\mathcal{C}}$  attains optimal localization rates for estimating  $\mathcal{C}^*$ . This result exploits Theorem 4.1 in Narisetty and He (2014) which provides a consistency result for linear model estimation with the shrinking and diffusing prior. Before arriving at this result, we now make the following modeling assumption.

**Assumption 1.** Let  $\kappa$  be the minimum jump size, thus,

$$\kappa := \min_{j \in \mathcal{C}^*} |f_j - f_{j-1}|.$$

Then we require that  $\kappa^2 T / (\sigma^2 \log T) \rightarrow \infty$ , as  $T \rightarrow \infty$ . Furthermore, we impose the following minimum spacing condition

$$\Delta := \min_{j \neq j', j, j' \in \mathcal{C}^*} |j - j'| \geq \frac{c_1 \sigma^2 \log T}{\kappa^2},$$

for a large enough  $c_1 > 0$ , and require that  $K := |\mathcal{C}^*| = O(1)$ .

**Assumption 1** can be thought as a signal-to-noise-ratio condition. In fact, **Assumption 1** is a weaker condition than Assumption 2 from Wang, Yu, and Rinaldo (2020) which states that

$$\Delta \geq \frac{c\sigma^2 \log^{1+\xi} T}{\kappa^2},$$

for positive constants  $c$  and  $\xi$ . However, the framework in Wang, Yu, and Rinaldo (2020) allows the possibility that  $K$  diverges whereas here we require that  $K = O(1)$ . We are now ready to state the main result of this section.

**Theorem 1.** Suppose that **Assumption 1** holds. Then for a constant  $c_0 > 0$  the estimator  $\hat{C}$  satisfies

$$\Pr \left( |\hat{C}| = K, \max_{\eta \in C^*} \min_{\hat{\eta} \in \hat{C}} |\hat{\eta} - \eta| \leq \frac{c_0 \sigma^2 \log T}{\kappa^2} \right) \rightarrow 1,$$

as  $T \rightarrow \infty$ , provided that  $\tau_{0,m}^2 = o(1/m)$ ,  $q_m \asymp 1/m$ , and  $\tau_{1,m}^2 \asymp m^{1+\delta}$  for some  $\delta > 1$ , and  $m$  such that  $m \asymp \kappa^2 T / (\sigma^2 \log T)$ .

Notably, **Theorem 1** shows that the maximum a posteriori estimator constructed based on the model (11) attains a localization rate of order  $\log T / \kappa^2$ . As Wang, Yu, and Rinaldo (2020) showed, this localization rate is minimax optimal up to a logarithm factor. Importantly, our guarantee on the localization rate holds under the minimum signal-to-noise ratio condition possible; see Lemmas 1–2 in Wang, Yu, and Rinaldo (2020). Also, notice that provided that  $\log T / (\kappa^2 \Delta) \rightarrow 0$ , then in the language of change point detection, the estimator  $\hat{C}$  is consistent (Wang, Yu, and Rinaldo 2020).

We highlight that **Theorem 1** requires  $\sigma^2$  to be known. However, this is not restrictive. In fact, if  $\sigma^2$  is misspecified by some  $\hat{\sigma}^2 \neq \sigma^2$ , as in Remark 2 in Narisetty and He (2014), the conclusion of **Theorem 1** will still hold provided that  $c_1 \sigma^2 \leq \hat{\sigma}^2 \leq c_2 \sigma^2$  for some positive constants  $c_1$  and  $c_2$ . The latter can easily be attained in practice provided that  $\hat{\sigma}^2$  is a consistent estimator of  $\sigma^2$ . A simple consistent estimator can be found by splitting the data in two sets, one consisting of odd  $t$  and the other of even  $t$ . We mentioned in the previous section a few possible estimators. Finally, we can use the estimate as input to our method along with the other half of the data.

We stress that in practice we use the fast method described in the previous section based on a misspecification of (11). In the SM, we show that such surrogate procedure still enjoys a localization guarantee in the case of single change point detection.

One another note, we remark that the estimator studied in this section required specifying the parameter  $m$  that describes the number of elements in the partition used to construct  $\tilde{Y}$ . The intuitive reason for this construction is that in change point detection, unlike variable selection, no estimator can detect the exact locations of the change points with high probability, see Wang, Yu, and Rinaldo (2020). Hence, by using averages over small intervals and treating the resulting signal as the input to the *basad.cp*, we can expect to perform variable selection exactly with high probability and then translate this into optimal change point localization.

Finally, if the number of observations collected at a given point  $t$  is  $n_t$  and  $n_t \asymp n$  for all  $t$ , then it is possible to construct a

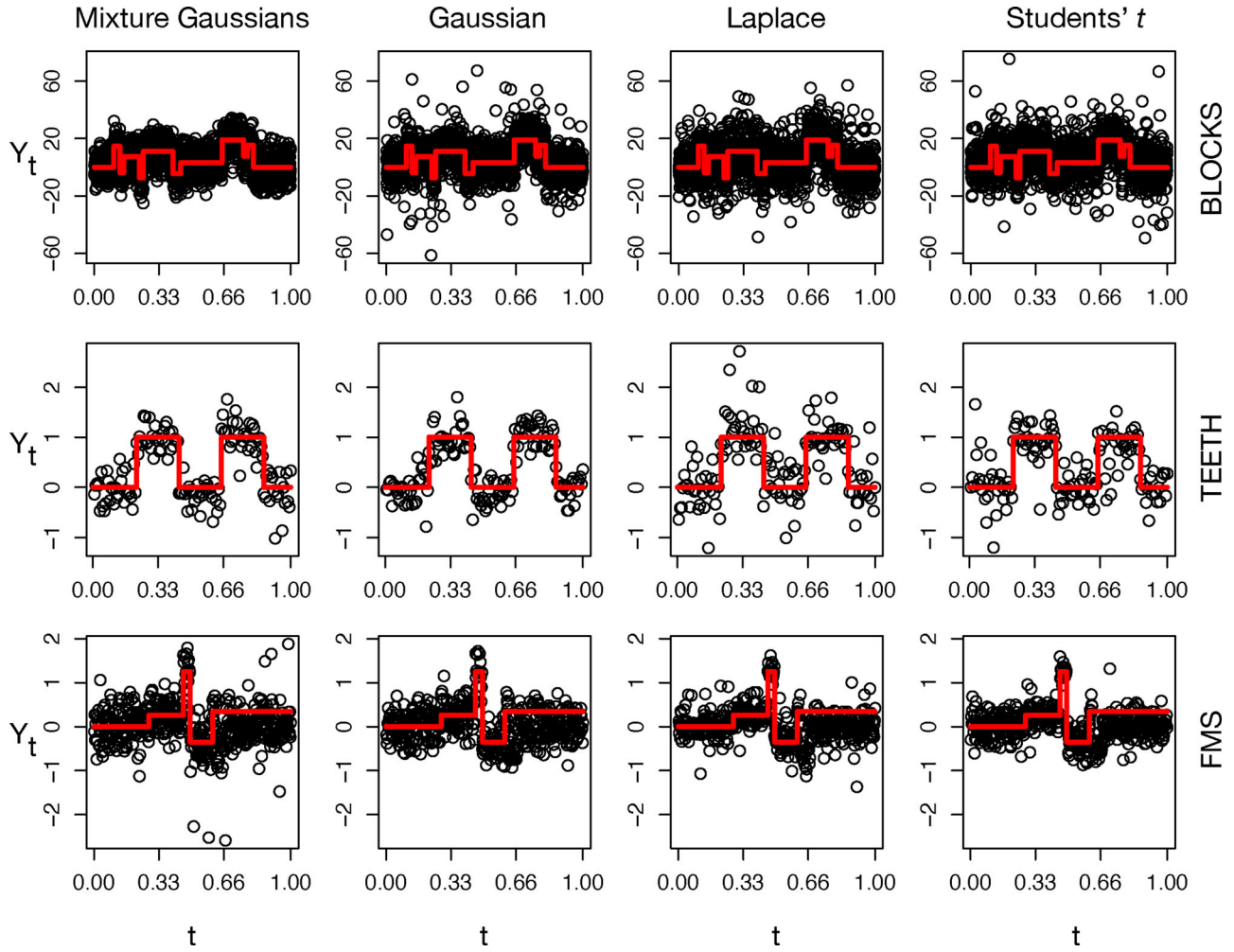
variant of  $\hat{C}$  that attains a localization rate of order  $\log T / (\kappa^2 n)$ . This agrees with the intuition that having multiple observations per time  $t$  can help with the estimation of the change points.

## 4. Simulations

We rely on simulations to explore the ability of the *solo.cp* and *basad.cp* estimators to accurately estimate  $K$  and change point locations  $\eta_1, \dots, \eta_K$ . We consider realistic scenarios designed to capture the variability encountered in applications, varying the conditional mean  $f_t$  and the distribution of the error terms  $(\epsilon_t)_{1:T}$ . We compare *basad.cp* and *solo.cp* with several state-of-the-art methods: *wbs* (Fryzlewicz 2014), *ebpiece* (Liu, Martin, and Shen 2017), *smuce* (Frick, Munk, and Sieling 2014), and *pelt* (Killick, Fearnhead, and Eckley 2012). Details of each implementation are given in the SM. All code to reproduce the results in this section is available at [https://github.com/lorenzocapp/solocp\\_experiments](https://github.com/lorenzocapp/solocp_experiments). The methodology is available as an R package available for download at <https://github.com/lorenzocapp/solocp>.

Our empirical comparisons assess the accuracy of the different estimators with the following criteria. We consider the statistic  $K - \hat{K}$  to measure how well each estimator recovers the true number of change points. We consider an order-invariant Hausdorff metric  $d(\hat{C}, C^*) = d(\hat{C}|C^*) + d(C^*|\hat{C})$ , where  $d(\hat{C}|C^*) = \max_{\eta \in C^*} \min_{x \in \hat{C}} |x - \eta|$  and  $d(C^*|\hat{C}) = \max_{\eta \in \hat{C}} \min_{x \in C^*} |x - \eta|$  are respectively the one-sided Hausdorff distances. We use  $d(\hat{C}, C^*)$  to assess the overall accuracy of the estimators in recovering the true change points locations  $\eta_1, \dots, \eta_K$ . We employ  $d(\hat{C}, C^*)$  in lieu of  $d(\hat{C}|C^*)$ , being the latter insensitive to overestimation. Lastly, for all  $\eta \in C^*$  we calculate  $\min_{x \in \hat{C}} |x - \eta|$ , and report the proportion of points that are at distance zero, one, two, and equal or greater than three. We refer to this criterion as the normalized empirical distribution and denote it by  $|\hat{\eta} - \eta|/K$ . It is a finer measure than the Hausdorff distance of the change point location estimation accuracy. Since this criterion is also insensitive to overestimation, we include the reciprocal  $|\eta - \hat{\eta}|/\hat{K}$ . The unnormalized version of this criterion is also considered by Fryzlewicz (2014).

Here, we focus on the setting where one observation is available per time point ( $n_t = 1$ ). Experiments for  $n_t > 1$  can be found in the SM. We consider three test signals and four error distributions. The first signal  $f_t$  is called BLOCKS ( $K = 11$ ,  $T = 2048$ ), a standard benchmark for change point detection procedures (e.g., used by Fryzlewicz 2014), the second test signal is called TEETH ( $K = 4$ ,  $T = 140$ ), the third signal is called FMS ( $K = 6$ ,  $T = 497$ ), used Fryzlewicz (2014) and Fearnhead and Rigaiil (2018) is characterized by some very narrow spacing and varying intensity. We consider four distributions for the error terms: Gaussian, Laplace, Student's  $t$ , and a mixture of Gaussians (to mimic the presence of outliers one of the two components has a larger variance). Change point locations of the test signals and the parameters of the error terms are fully specified in the SM. For each combination  $(f_t, \epsilon_t)$ , we sample 100 datasets and report the average value for each criterion considered. Figure 1 plots examples of datasets sampled for each scenario, along with the true test signals in red.



**Figure 1.** Examples of datasets for the 12 scenarios considered and true test signals. First row panels depict sample datasets along with the BLOCKS test signal (red), second row panels sample datasets along with the TEETH test signal (red), third row panels sample datasets along with the FMS test signal (red). Columns depict different noise models.

We consider the *solo.cp* algorithm with  $\tau_{0,T}^2 = T^{-1}$ ,  $\tau_{1,T}^2 = T$ . For the *basad.cp* algorithm, we employ the default choices of the parameters given by Narisetty and He (2014):  $\tau_{0,T}^2 = \hat{\sigma}^2(10T)^{-1}$ ,  $\tau_{1,T}^2 = \hat{\sigma}^2 \log T$ . For both *solo.cp* and *basad.cp*, we use  $q = 0.1$ , and  $\Delta = 2$ , that is, we are preventing the algorithm from detecting consecutive change points. *Pelt* also has the minimum spacing parameter  $\Delta$ , which we set equal to two as well. The parameters  $\tau_{0,T}^2$ ,  $\tau_{1,T}^2$  are set following the results of Section 3. The theory in this article does not provide guidance on the choice of  $\Delta$  and  $q$ . We study the robustness of the *solo.cp* algorithm to these parameters' choices in the SM.

The procedures *basad.cp*, *solo.cp*, and *ebpiece* have  $\hat{\sigma}$  as an input. Here, we computed it from the residuals of the fused LASSO (Tibshirani et al. 2005; implemented in the *genlasso* R package available on CRAN). We input the same  $\hat{\sigma}$  in *pelt* and *smuce* for a fair comparison. In the BLOCKS signal datasets, we initialize the *ebpiece* MCMC from the estimates of the fused LASSO (“one standard deviation rule”), otherwise it is not possible to achieve convergence in a reasonable time. This can be seen by the very poor performance of the method which can be due to the fact that the chains “get stucked” into local modes.

Tables 1 summarizes  $|\hat{\eta} - \eta|/K$ ,  $|\eta - \hat{\eta}|/\hat{K}$ ,  $K - \hat{K}$ ,  $\hat{K}$ ,  $d(\hat{C}, C^*)$ , and the mean computing time for the four scenarios considered

for the BLOCKS test signals, Table 2 for the TEETH test signals, and Table 3 for the FMS test signals. The *basad.cp* method is not included in Tables 1 and 3 because it was not computationally feasible to approximate the posterior distributions with MCMC in these datasets.

The procedures *wbs*, *smuce*, and *pelt* achieve the best overall performance according to  $|\eta - \hat{\eta}|/\hat{K}$ , suggesting that they recover very well  $C^*$ ; their performance under Gaussian noise scenarios is excellent. The criteria  $|\eta - \hat{\eta}|/\hat{K}$ ,  $d(\hat{C}, C)$ , and  $K - \hat{K}$  suggest that *wbs* and *smuce* tend to overestimate the number of change points. The problem is extremely severe in the presence of outliers and with Student's t-distributed errors. *pelt* is the fastest method and the best performing one in the FMS scenario.

*ebpiece* recovers well  $\hat{K}$  in the BLOCKS and TEETH scenarios but it does not seem to recover well the exact locations of the change points ( $|\hat{\eta} - \eta|/K$  and  $|\eta - \hat{\eta}|/\hat{K}$ ). It underestimates the number of change points in the FMS scenario. In the BLOCKS signals, the MCMC chain converges only when the fused LASSO is used as initialization. Essentially, we are starting the chain where we would want the posterior samples to concentrate. If we use the default initialization of the chain used in (Liu, Martin, and Shen 2017), the MCMC chain fails to converge within 30 min (computing time between brackets in Table 1).

**Table 1.** Hausdorff distance, empirical distributions, and estimation bias in  $K$  of the procedures considered for the BLOCKS test signals.

Noise	Method	$ \hat{\eta} - \eta /\hat{K}$				$ \eta - \hat{\eta} /\hat{K}$				$K - \hat{K}$	$d(\hat{C}, C^*)$	time
		0	1	2	$\geq 3$	0	1	2	$\geq 3$			
Mixt. Gaussians	ebpiece	0.12	0.1	0.07	0.71	0.15	0.12	0.08	0.64	-1.67	228.15	72.23 (1457.72)
	pelt	<b>0.48</b>	0.21	0.1	0.21	0.37	0.16	0.08	0.39	3.97	147.67	<b>0</b>
	smuce	0.38	0.16	0.07	0.39	0.16	0.07	0.03	0.73	15.73	301.44	0.02
	solo.cp	0.4	0.15	0.06	0.38	<b>0.44</b>	0.17	0.07	0.32	<b>-0.77</b>	<b>116.28</b>	0.45
	wbs	<b>0.5</b>	0.2	0.1	0.2	0.21	0.09	0.05	0.65	17.62	289.08	0.11
Gaussian	ebpiece	0.22	0.13	0.09	0.56	0.17	0.12	0.09	0.62	3.09	170.13	78.66 (1484.33)
	pelt	<b>0.59</b>	0.21	0.09	0.1	<b>0.6</b>	0.22	0.09	0.09	-0.02	<b>12.24</b>	<b>0</b>
	smuce	<b>0.55</b>	0.19	0.08	0.18	<b>0.6</b>	0.21	0.09	0.1	-0.93	44.12	0.02
	solo.cp	0.51	0.19	0.07	0.24	0.48	0.18	0.07	0.28	0.82	89.79	0.43
	wbs	<b>0.57</b>	0.21	0.1	0.12	<b>0.57</b>	0.21	0.1	0.12	<b>0</b>	17.22	0.11
Laplace	ebpiece	0.12	0.08	0.05	0.75	0.13	0.11	0.07	0.7	-1.38	264.18	85.95 (1648.14)
	pelt	<b>0.43</b>	0.2	0.1	0.27	<b>0.42</b>	0.19	0.1	0.28	<b>0.35</b>	<b>69.73</b>	<b>0</b>
	smuce	0.36	0.17	0.08	0.39	0.33	0.15	0.07	0.45	1.41	180.96	0.02
	solo.cp	0.35	0.14	0.06	0.46	<b>0.38</b>	0.16	0.06	0.4	-0.79	114.72	0.43
	wbs	0.42	0.19	0.09	0.3	<b>0.39</b>	0.18	0.09	0.34	1.17	124.53	0.1
Student's $t$	ebpiece	0.11	0.08	0.04	0.77	0.15	0.11	0.07	0.67	-2.91	277.52	77.74 (1537.1)
	pelt	<b>0.46</b>	0.21	0.09	0.24	<b>0.38</b>	0.18	0.08	0.37	2.62	141.76	<b>0</b>
	smuce	0.37	0.17	0.08	0.38	0.25	0.11	0.06	0.58	6.11	255.3	0.02
	solo.cp	0.36	0.15	0.06	0.44	<b>0.4</b>	0.17	0.07	0.36	<b>-0.95</b>	<b>113.22</b>	0.43
	wbs	<b>0.44</b>	0.2	0.1	0.26	0.3	0.14	0.07	0.49	6.21	215.14	0.1

Average statistics computed over 100 simulations for *solo.cp*, *ebpiece* (Liu, Martin, and Shen 2017), *smuce* (Frick, Munk, and Sieling 2014), *wbs* (Fryzlewicz 2014), and *pelt* (Killick, Fearnhead, and Eckley 2012). For  $|\hat{\eta} - \eta|/\hat{K}$  and  $|\eta - \hat{\eta}|/\hat{K}$  the higher the number in the zero column the better. Conversely, for  $d(\hat{C}, C^*)$  the lower the better. For  $K - \hat{K}$ , the closer to the zero the better. We report in bold the methods with best empirical performance and those within 10% of the best. The method *basad.cp* is not included since it required a computing time longer than 2 hr. We include between brackets the computing time for *ebpiece* if initialized with the in-built procedure (see text for a discussion). Average time is reported in seconds.

**Table 2.** Hausdorff distance, empirical distributions, and estimation bias in  $K$  of the procedures considered for the TEETH test signals.

Noise	Method	$ \hat{\eta} - \eta /\hat{K}$				$ \eta - \hat{\eta} /\hat{K}$				$K - \hat{K}$	$d(\hat{C}, C^*)$	time
		0	1	2	$\geq 3$	0	1	2	$\geq 3$			
Mixt. Gaussian	ebpiece	0.27	0.16	0.11	1.13	0.28	0.18	0.12	1.09	-0.91	18.03	12.13
	pelt	<b>0.84</b>	0.09	0.05	0.02	0.7	0.08	0.06	0.17	1.06	<b>7.53</b>	<b>0</b>
	smuce	<b>0.8</b>	0.1	0.06	0.03	0.5	0.08	0.05	0.37	2.94	13.77	0.01
	solo.cp	0.76	0.06	0.04	0.13	<b>0.8</b>	0.06	0.05	0.09	-0.12	17.36	0.01
	wbs	<b>0.88</b>	0.08	0.04	0.01	0.42	0.06	0.04	0.48	5.79	15.14	0.04
Gaussian	basad.cp	0.73	0.09	0.05	0.13	0.75	0.09	0.05	0.11	<b>-0.03</b>	16.68	55.82
	ebpiece	0.48	0.31	0.12	0.1	0.47	0.3	0.13	0.11	0.32	4.59	12.24
	pelt	<b>0.95</b>	0.04	0.01	0	<b>0.87</b>	0.04	0.01	0.08	0.92	3.34	<b>0</b>
	smuce	<b>0.96</b>	0.04	0	0	<b>0.95</b>	0.04	0	0	0.02	0.53	0.01
	solo.cp	<b>0.94</b>	0.04	0.01	0.01	<b>0.9</b>	0.03	0.01	0.06	0.31	3.41	0.01
Laplace	wbs	<b>0.96</b>	0.04	0	0	<b>0.93</b>	0.04	0	0.03	0.17	<b>1.5</b>	0.04
	basad.cp	0.86	0.09	0.01	0.04	<b>0.86</b>	0.09	0.01	0.04	<b>0.01</b>	4.38	67.99
	ebpiece	0.3	0.13	0.09	1.12	0.31	0.17	0.1	1.05	-0.78	23.03	9.66
	pelt	<b>0.78</b>	0.14	0.04	0.04	<b>0.69</b>	0.12	0.05	0.14	0.73	<b>6.55</b>	<b>0</b>
	smuce	<b>0.76</b>	0.14	0.04	0.05	0.69	0.13	0.04	0.13	0.56	<b>6.67</b>	0.01
Student's $t$	solo.cp	0.67	0.11	0.02	0.19	<b>0.76</b>	0.13	0.03	0.07	<b>-0.45</b>	20.48	0.01
	wbs	<b>0.77</b>	0.14	0.04	0.04	0.66	0.13	0.04	0.17	1.05	7.79	0.03
	basad.cp	0.65	0.13	0.04	0.18	<b>0.74</b>	0.15	0.04	0.07	<b>-0.44</b>	18.66	51.71
	ebpiece	0.31	0.16	0.07	1.17	0.31	0.18	0.08	1.13	-0.85	17.53	11.69
	pelt	<b>0.81</b>	0.12	0.04	0.03	0.66	0.1	0.05	0.19	1.14	<b>8.07</b>	<b>0</b>
	smuce	<b>0.8</b>	0.13	0.04	0.03	0.6	0.1	0.05	0.26	1.81	9.98	0.01
	solo.cp	0.73	0.09	0.03	0.15	<b>0.78</b>	0.09	0.04	0.09	-0.21	15.56	0.01
	wbs	<b>0.83</b>	0.12	0.04	0.01	0.55	0.09	0.04	0.32	2.97	10.98	0.04
	basad.cp	0.69	0.12	0.04	0.15	<b>0.7</b>	0.13	0.04	0.13	<b>-0.06</b>	15.5	52.98

Average statistics computed over 100 simulations for *solo.cp*, *basad.cp*, *ebpiece* (Liu, Martin, and Shen 2017), *smuce* (Frick, Munk, and Sieling 2014), *wbs* (Fryzlewicz 2014), and *pelt* (Killick, Fearnhead, and Eckley 2012). For  $|\hat{\eta} - \eta|/\hat{K}$  and  $|\eta - \hat{\eta}|/\hat{K}$  the higher the number in the zero column the better. Conversely, for  $d(\hat{C}, C^*)$  the lower the better. For  $K - \hat{K}$ , the closer to the zero the better. We report in bold the methods with best empirical performance and those within 10% of the best. Average time is reported in seconds.

*solo.cp* performs well in all scenarios, being consistently the best Bayesian method. It is not as accurate as the frequentist methods in recovering the exact location of the change points ( $|\hat{\eta} - \eta|/\hat{K}$ ). In particular in the BLOCKS and FMS scenarios, the reason seems to be that *solo.cp* underestimates  $K$ . On the other hand, it is the fastest Bayesian method and the algorithm is extremely robust to the misspecification of the error terms, being consistently among the best in terms of  $|\eta - \hat{\eta}|/\hat{K}$  and  $d(\hat{C}, C^*)$ .

In these TEETH scenarios, computation times are comparable to the state-of-the-art frequentist methods. The results of *solo.cp* are very robust to the choices of  $q$  and  $\Delta$ ; see SM.

Like *ebpiece*, *solo.cp* underestimates the number of change points in the FMS scenario across error distributions. In the FMS scenario, the high values of  $d(\hat{C}, C^*)$  and  $K - \hat{K}$  are due to the fact that both *ebpiece* and *solo.cp* do not detect the first change point, the one with the smallest magnitude. This effect may be linked to



**Table 3.** Hausdorff distance, empirical distributions, and estimation bias in  $K$  of the procedures considered for the FMS test signals.

Noise	Method	$ \hat{\eta} - \eta /\hat{K}$				$ \eta - \hat{\eta} /\hat{K}$				$K - \hat{K}$	$d(\hat{C}, C^*)$	time
		0	1	2	$\geq 3$	0	1	2	$\geq 3$			
Mixt. Gaussians	ebpiece	0.1	0.07	0.05	1.34	0.25	0.16	0.13	1.01	-3.77	95.78	18.12
	pelt	<b>0.59</b>	0.14	0.07	0.22	0.47	0.11	0.06	0.37	<b>1.88</b>	<b>71.33</b>	<b>0</b>
	smuce	0.41	0.14	0.06	0.4	0.25	0.08	0.05	0.62	4.42	118.72	0.08
	solo.cp	0.36	0.06	0.03	0.56	<b>0.74</b>	0.12	0.05	0.09	-3.05	93.87	0.05
	wbs	<b>0.57</b>	0.14	0.06	0.23	0.29	0.07	0.04	0.6	7.22	117.14	0.04
Gaussian	ebpiece	0.22	0.09	0.06	0.68	0.38	0.17	0.12	0.37	-2.25	88.62	20.15
	pelt	<b>0.68</b>	0.13	0.06	0.12	0.68	0.13	0.06	0.13	0.06	<b>16.45</b>	<b>0</b>
	smuce	0.52	0.12	0.07	0.29	0.59	0.13	0.08	0.2	-0.71	17.86	0.08
	solo.cp	0.52	0.08	0.03	0.38	<b>0.81</b>	0.11	0.04	0.05	-2.04	86.99	0.05
	wbs	<b>0.68</b>	0.13	0.07	0.12	0.68	0.13	0.07	0.12	<b>0.01</b>	<b>14.3</b>	<b>0.04</b>
Laplace	ebpiece	0.23	0.15	0.06	0.56	0.32	0.24	0.12	0.32	-1.58	75.25	19.94
	pelt	<b>0.76</b>	0.11	0.05	0.08	0.7	0.1	0.05	0.15	<b>0.68</b>	<b>28.63</b>	<b>0</b>
	smuce	<b>0.75</b>	0.11	0.05	0.09	0.63	0.09	0.04	0.23	1.48	59.13	0.09
	solo.cp	0.6	0.1	0.02	0.29	<b>0.8</b>	0.13	0.02	0.05	-1.45	85.08	0.04
	wbs	<b>0.76</b>	0.1	0.04	0.09	0.65	0.09	0.04	0.22	1.38	43.43	0.04
Student's t	ebpiece	0.23	0.17	0.06	0.53	0.35	0.25	0.1	0.3	-1.49	75.79	19.08
	pelt	<b>0.76</b>	0.12	0.05	0.08	0.7	0.11	0.04	0.14	<b>0.56</b>	<b>23.07</b>	<b>0</b>
	smuce	<b>0.72</b>	0.13	0.04	0.12	0.62	0.11	0.03	0.24	1.26	47.9	0.08
	solo.cp	0.6	0.11	0.02	0.28	<b>0.79</b>	0.14	0.02	0.06	-1.36	87.35	0.05
	wbs	<b>0.72</b>	0.13	0.05	0.1	0.59	0.11	0.04	0.25	1.71	42.8	0.04

Average statistics computed over 100 simulations for *solo.cp*, *ebpiece* (Liu, Martin, and Shen 2017), *smuce* (Frick, Munk, and Sieling 2014), *wbs* (Fryzlewicz 2014), and *pelt* (Killick, Fearnhead, and Eckley 2012). For  $|\hat{\eta} - \eta|/K$  and  $|\eta - \hat{\eta}|/\hat{K}$  the higher the number in the zero column the better. Conversely, for  $d(\hat{C}, C^*)$  the lower the better. For  $K - \hat{K}$ , the closer to the zero the better. We report in bold the methods with best empirical performance and those within 10% of the best. Average time is reported in seconds.

the better robustness to outliers of the two methodologies. Very small effects are shared among multiple random means, which are integrated out, hence, even more challenging to detect. In this scenario, *solo.cp* outperforms *ebpiece* and its performance is comparable to *smuce* in terms of  $|\hat{\eta} - \eta|/K$ .

*basad.cp* achieves a performance comparable to *solo.cp* in the TEETH scenarios. This is expected given that both methods are based on the shrinking and diffusing priors of Narisetty and He (2014). A similar performance is achieved at a much higher computational cost.

In Section 2, we attributed the robustness of *solo.cp* and *basad.cp* to their peculiar construction that involves multiple random increments and shrinkage estimators. A competing hypothesis is that the spike prior on a candidate change point is more flexible than a sharp  $l_0$  penalty, allowing for small fluctuations in the signal. This would not be desirable because, under the default spike prior hyperparameter  $\tau_{0,T}^2 = T^{-1}$ , the spike prior would degenerate into a point mass at 0, that is, allowing for no fluctuations. We investigate this hypothesis repeating the analysis for the BLOCKS signals for a range of prior hyperparameters  $\tau_{0,T}^2(T^{-1}, T^{-2}, T^{-3}, T^{-4})$  (Table S4, supplementary material). Results are robust to a varying  $\tau_{0,T}^2$ , essentially ruling out this hypothesis. Another hypothesis is that the robustness is due to the postprocessing (steps 3 and 4 in Algorithm 1). We similarly repeat the BLOCKS signals simulation study for  $\Delta = 1$  (Table S2 SM). We observe a small increase in the number of false positives, leaving *solo.cp* as the most accurate method in the misspecified scenarios. We stress that the effect of the post processing is minimal: the results presented in the section were obtained with  $\Delta = 2$ , which simply prevents consecutive change points from being detected. Most of the methodologies included here have such a minimum spacing constraint.

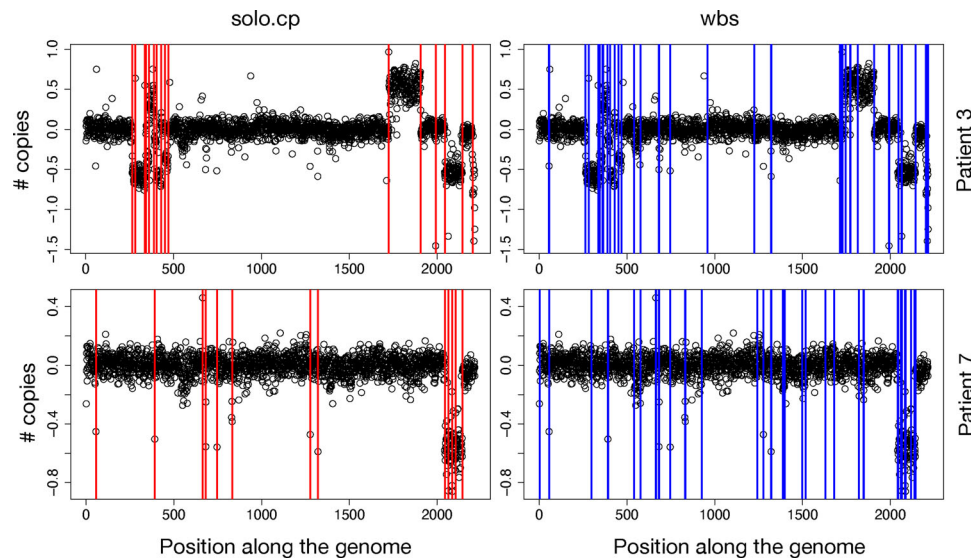
## 5. Applications

We analyze two real datasets (aCGH and Ion channel data) to compare *solo.cp* to alternative methodologies built under the same assumptions (Gaussian iid noise, piecewise constant mean). In particular, we compare *solo.cp* to *wbs*. We will acknowledge the limitations of our analyses and possible solutions at the end of each subsection.

### 5.1. Array Comparative Genomic Hybridization (aCGH) Data

Genomic alternations happen in the development of tumors. Studying these alternations, for example determining the copy-number variations, is important for understanding cancer and also used for its diagnosis. Array Comparative Genomic Hybridization (aCGH) is a popular method that generates this type of data (Skena et al. 1995). We analyze an aCGH dataset of individuals with a bladder tumor collected by Stransky et al. (2006). The dataset is publicly available in the R package *ecp* (James and Matteson 2014), and includes 43 individuals and 2215 locations. The goal of the analysis is to detect changes in the copy-number. The underlying assumption is that alternations are constant within a segment. Segments involved in the tumor should be equally affected across patients.

While we could repeat the analysis for all the patients, we include only two in this manuscript for parsimony. The number of samples is approximately identical to the BLOCKS test signal, hence we use the same parameters for *solo.cp* ( $\tau_{0,T}^2 = T^{-1}$ ,  $\tau_{1,T}^2 = T$ ,  $\tau_T^2 = T^{-1/2}$ ,  $q = 0.1$ ,  $\Delta = 5$ ) and  $\hat{\sigma}^2$  equal to the variance of the residuals of the fused LASSO (tuning parameter chosen by one-standard-error rule). We compare the results of *solo.cp* with *wbs* (default implementation). The left column of



**Figure 2.** aCGH data. Copy number variations of patients having a bladder tumor recorded at  $n = 2215$  sites by Stransky et al. (2006). First row depicts the observations for Patient 3, second row depicts the observations for Patient 7. First column includes the change points estimated by *solo.cp* (red), second column the ones estimated by *wbs* (blue).

Figure 2 depicts the estimates of the *solo.cp* change points, the right column depicts the one obtained with *wbs*. The two rows refer to the two different patients.

Both methods seem to recover more change points than the number of blocks identified through a visual inspection of the data. There are a few points where the change point corresponds to a single observation, not an entire segment along the genome. We would need further research to determine if these points can be classified as outliers. However, *solo.cp* appears more parsimonious:  $\hat{K} = 19$  for Patient 3 and  $\hat{K} = 13$  for Patient 7, while *wbs* estimates  $\hat{K} = 49$  for Patient 3 and  $\hat{K} = 36$  for Patient 7. The results are consistent with what we observed in the simulation section. We also note that both methods ignore autocorrelation. Sequences may exhibit some serial dependence because nearby genomic segments are known to be associated. The autocorrelation in the noise process could lead to overestimation. Recently, methods robust to this serial dependence has been proposed (Cho and Fryzlewicz 2020; Romano et al. 2021). The method of Cho and Fryzlewicz (2020) returns  $\hat{K} = 14$  for Patient 3 and  $\hat{K} = 2$  for Patient 7.

## 5.2. Ion Channels Data

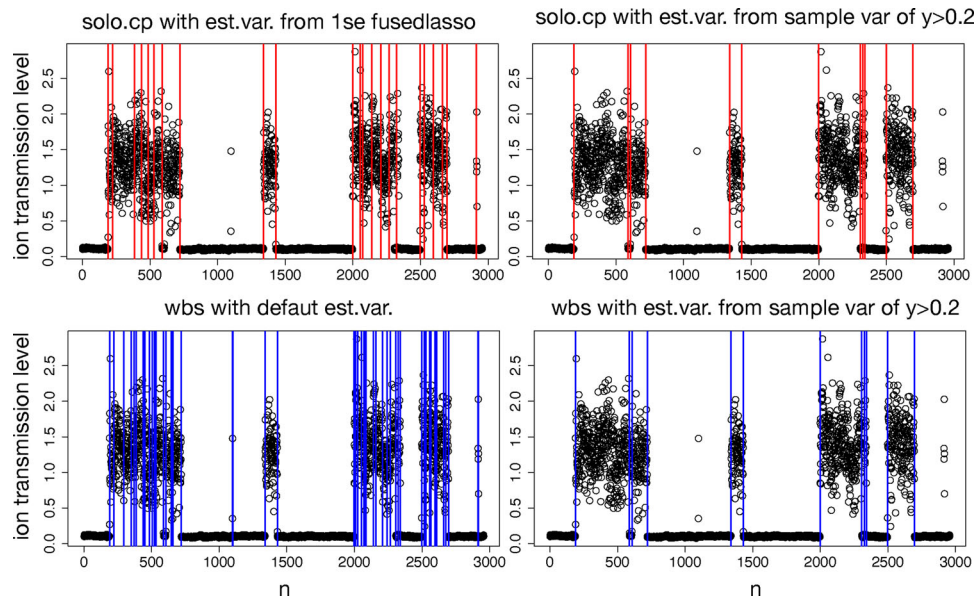
Ion channels are a class of proteins expressed by all cells that create pathways for ions (charged particles) to pass through the otherwise impermeable cell membrane. The opening of these pathways is essential for cell operations in the nervous system, in the muscles, and in the pancreas. The study of ion channels plays a fundamental role in the development of new drugs (Alexander, Mathie, and Peters 2008). The patch clamp technique is an electrophysiological tool for understanding ion channel behavior. It is used to measure ionic currents from single living cells or tissues (Neher and Sakmann 1995). Electrophysiologists use glass microelectrodes to gain access to cells expressing ion channels. Through the microelectrode, a voltage is applied, forming a voltage clamp, and the current passing across the cell membrane through the ion channels is measured.

We consider a dataset produced by the Steinem Lab (Institute of Organic and Biomolecular Chemistry, University of Göttingen), recently analyzed by Vanegas, Behr, and Munk (2022), measuring a single ion channel of the bacterial porin PorB, a bacterium that plays a role in the pathogenicity of *Neisseria gonorrhoeae*. The experiment design includes a technique that induces local dependencies of the error terms (Pein, Sieling, and Munk 2017). To remove these dependencies, we follow the same approach of Vanegas, Behr, and Munk (2022), subsampling every 11th observation. The original dataset includes 600,000 time instances. We analyze a portion of the dataset of length 32,511. After subsampling, the dataset is composed of 2956 time points. Figure 3 depicts the dataset.

Figure 3 suggests that the noise variance when the channels are open is much higher than when they are closed. This feature of ion channel data is known as *open channel noise* (Neher and Sakmann 1995). The methods studied in this article, and considered in Section 4, do not assume error heterogeneity. The first column of Figure 3 depicts a naive application of *solo.cp* ( $\tau_{0,T}^2 = T^{-1}$ ,  $\tau_{1,T}^2 = T$ ,  $\tau_T^2 = T^{-1/2}$ ,  $q = 0.1$ ,  $\Delta = 5$ ) and *wbs* where we estimate a single global variance ignoring variance heterogeneity. This leads to overestimation on the change points.

The second column depicts an attempt to account for the heteroscedasticity due to the *open channel noise* phenomena. We estimate the sample variance when the ion channels are open (we approximate it considering the observations above 0.2). We input to both methods this sample variance and reestimate the change points (for *solo.cp* same parameters as before; for *wbs*, we set  $\theta_0$  set to 3). Now,  $\hat{K}$  is 12 for both methods and the locations of the change points seem reasonable by visual inspection. A few isolated points are not detected as change points (approximately around 1100 and 2900). We note though that the result of *wbs* largely depends on other tuning parameters (e.g.  $\theta_0$ ), while the estimates of *solo.cp* are very robust to the choices of all the parameters that are not  $\hat{\sigma}^2$ .

There are methods in the literature that are robust to heterogeneous variance (Killick and Eckley 2014; Pein, Sieling,



**Figure 3.** Ion data. Ion channel data recorded at the Steinem lab (Insitute of Organic and Biomolecular Chemistry, University of Göttingen) at  $n = 2956$  time instances. First row depicts the change points estimated through *solo.cp*, second row depicts the change points estimated through *wbs*. Estimates in the first column are obtained using “default” estimates of the sample standard deviation, which means the standard deviation of the residuals obtained from the fused LASSO (one standard error rule) for *solo.cp*, and the median absolute deviation estimates for *wbs*. Estimates in the second column are obtained using the sample standard deviation of observations taking values larger than 0.2 (to approximate when ion channels are open).

and Munk 2017). These can be directly deployed and do not require the extra step of understanding how to estimate the input variance. An application of HSMUCE (Pein, Sieling, and Munk 2017) to this dataset gives  $\hat{K}$  equal to 25. Further study and domain knowledge is required to determine whether extra change points are justified.

## 6. Discussion

We studied spike-and-slab priors for change point detection leveraging recent results in the variable selection literature. We chose to work with a prior having both the spike-and-slab component defined by Gaussian distributions and sample size-dependent hyperparameters. We established that an estimator based on this prior distribution is consistent and achieves optimal localization rates of multiple change points. Furthermore, the use of this prior allowed us to propose a fast Bayesian change point estimator based on a slightly misspecified model. A version of the fast estimator achieves the optimal rate in the single change point problem. In simulations, its empirical accuracy is comparable to state-of-the-art benchmarks. Its salient features are being one of the fastest Bayesian methods available (no MCMC required) and being robust to misspecification of the error model in numerical experiments. The latter property is somewhat surprising, given that we only considered Gaussian noise in the method development and in the theory. We showed these features in simulation studies, displaying situations where our estimator performs well while many competing methods severely overestimate the number of change points.

There is a rich literature on change point detection for settings more general than the one considered in this article. Nevertheless, our results are promising and suggest that it is worth investigating the use of spike-and-slab priors in change

point detection for more general settings, such as settings with unknown variance, heterogeneous errors, and different types of dependence.

The first area of future work is to further improve the computational performance of the *solo.cp* algorithm. The main bottleneck of the algorithm is the computation of the parameters in (9). The computing time of *solo.cp* is comparable to those of frequentist estimators for small sample sizes up to thousands of observations. A possible strategy to tackle the large  $T$  scenario is to design a more efficient (possibly approximate) version of *solo.cp* that provides a preliminary data partition, then apply the regular algorithm to the different chunks. For larger sample sizes, we expect numerical and memory issues to arise, such as floating point collapse and lack of memory to compute all quantities necessary for the procedure.

The second area of research is the detection of higher-order changes, such as in piecewise-linear signals. A version of the *solo.cp* algorithm for piecewise-linear change point detection is readily available (as well as higher-order changes). However, our preliminary results suggest that a vanilla version of this estimator does not work well in this setting. Liu, Martin, and Shen (2017) suggest that a possible explanation is that one cannot fix the prior means at zero in this setting.

## Supplementary Materials

The supplementary material file contains the following sections:

- (S1) Localization rate of the fast Bayesian estimator in single change point setting;
- (S2) Simulations with  $n_t > 1$ ;
- (S3) Extension of *solo.cp* to the case  $n_t > 1$ ;
- (S4) Notation for proofs;
- (S5) Auxiliary lemmas for proof of Theorem 1;
- (S6) Proof of Theorem 1;



- (S7) Proof of Theorem S1;  
 (S8) Details of the simulation scenarios and the implementations;  
 (S9) Sensitivity of *solo.cp* to the choices of  $\Delta$ ,  $q$  and  $\tau_{0,T}^2$ .

## Acknowledgments

We thank an associate editor and two anonymous reviewers for constructive comments.

## Disclosure Statement

The authors report there are no competing interests to declare.

## Funding

Computing resources were provided by Stanford University research computing center (Sherlock cluster) and Pompeu Fabra Marvin cluster. Oscar Padilla acknowledges support from NSF DMS-2015489.

## References

- Alexander, S. P., Mathie, A., and Peters, J. A. (2008), "Guide to Receptors and Channels (grac)," *British Journal of Pharmacology*, 153, S1–S1. [1497]
- Aue, A., Hörmann, S., Horváth, L., and Reimherr, M. (2009), "Break Detection in the Covariance Structure of Multivariate Time Series Models," *Annals of Statistics*, 37, 4046–4087. [1489]
- Avanesov, V., and Buzun, N. (2018), "Change-Point Detection in High-Dimensional Covariance Structure," *Electronic Journal of Statistics*, 12, 3254–3294. [1489]
- Barbieri, M. M., and Berger, J. O. (2004), "Optimal Predictive Model Selection," *Annals of Statistics*, 32, 870–897. [1490]
- Barry, D., and Hartigan, J. A. (1992), "Product Partition Models for Change Point Problems," *Annals of Statistics*, 20, 260–279. [1488]
- (1993), "A Bayesian Analysis for Change Point Problems," *Journal of the American Statistical Association*, 88, 309–319. [1488]
- Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019), "Lasso Meets Horseshoe: A Survey," *Statistical Science*, 34, 405–427. [1489]
- Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation when  $p$  is much larger than  $n$ ," *Annals of Statistics*, 35, 2313–2351. [1488]
- Carlstein, E. (1988), "Nonparametric Change-Point Estimation," *Annals of Statistics*, 16, 188–197. [1489]
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [1488]
- Chen, S., and Walker, S. G. (2019), "Fast Bayesian Variable Selection for High Dimensional Linear Models: Marginal Solo Spike and Slab Priors," *Electronic Journal of Statistics*, 13, 284–309. [1489,1490,1491,1492]
- Chib, S. (1996), "Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models," *Journal of Econometrics*, 75, 79–97. [1488]
- (1998), "Estimation and Comparison of Multiple Change-Point Models," *Journal of Econometrics*, 86, 221–241. [1488]
- Cho, H. (2016), "Change-Point Detection in Panel Data via Double CUSUM Statistic," *Electronic Journal of Statistics*, 10, 2000–2038. [1489]
- Cho, H., and Fryzlewicz, P. (2015), "Multiple-Change-Point Detection for High Dimensional Time Series via Sparsified Binary Segmentation," *Journal of the Royal Statistical Society, Series B*, 77, 475–507. [1489]
- Cho, H., and Fryzlewicz, P. (2020), "Multiple Change Point Detection Under Serial Dependence: Wild Energy Maximisation and Gappy Schwarz Criterion," arXiv preprint arXiv:2011.13884. [1497]
- Du, C., Kao, C.-L. M., and Kou, S. C. (2016), "Stepwise Signal Extraction via Marginal Likelihood," *Journal of the American Statistical Association*, 111, 314–330. [1488]
- Fang, X., Li, J., and Siegmund, D. (2020), "Segmentation and Estimation of Change-Point Models: False Positive Control and Confidence Regions," *The Annals of Statistics*, 48, 1615–1647. [1492]
- Faulkner, J. R., and Minin, V. N. (2018), "Locally Adaptive Smoothing with Markov Random Fields and Shrinkage Priors," *Bayesian Analysis*, 13, 225–252. [1488]
- Fearnhead, P. (2006), "Exact and Efficient Bayesian Inference for Multiple Changepoint Problems," *Statistics and Computing*, 16, 203–213. [1488,1490]
- Fearnhead, P., and Rigai, G. (2018), "Changepoint Detection in the Presence of Outliers," *Journal of the American Statistical Association*, 114, 169–183. [1489,1492,1493]
- Frick, K., Munk, A., and Sieling, H. (2014), "Multiscale Change Point Inference," *Journal of the Royal Statistical Society, Series B*, 76, 495–580. [1488,1493,1495,1496]
- Friedrich, F., Kempe, A., Liebscher, V., and Winkler, G. (2008), "Complexity Penalized  $m$ -estimation: Fast Computation," *Journal of Computational and Graphical Statistics*, 17, 201–224. [1488]
- Fryzlewicz, P. (2014), "Wild Binary Segmentation for Multiple Change-Point Detection," *Annals of Statistics*, 42, 2243–2281. [1488,1492,1493,1495,1496]
- James, N. A., and Matteson, D. S. (2014), "ecp: An R Package for Nonparametric Multiple Change Point Analysis of Multivariate Data," *Journal of Statistical Software*, 62, 1–25. [1496]
- Killick, R., and Eckley, I. (2014), "changeoint: An R Package for Change-point Analysis," *Journal of Statistical Software*, 58, 1–19. [1497]
- Killick, R., Fearnhead, P., and Eckley, I. A. (2012), "Optimal Detection of Changeoints with a Linear Computational Cost," *Journal of the American Statistical Association*, 107, 1590–1598. [1488,1493,1495,1496]
- Kowal, D. R., Matteson, D. S., and Ruppert, D. (2019), "Dynamic Shrinkage Processes," *Journal of the Royal Statistical Society, Series B*, 81, 781–804. [1488]
- Liu, C., Martin, R., and Shen, W. (2017), "Empirical Priors and Posterior Concentration in a Piecewise Polynomial Sequence Model," arXiv preprint arXiv:1712.03848. [1488,1492,1493,1494,1495,1496,1498]
- Maidstone, R., Hocking, T., Rigai, G., and Fearnhead, P. (2017), "On Optimal Multiple Changeoint Algorithms for Large Data," *Statistics and Computing*, 27, 519–533. [1488]
- 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. [1489]
- Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032. [1488]
- Narisetty, N. N., and He, X. (2014), "Bayesian Variable Selection with Shrinking and Diffusing Priors," *Annals of Statistics*, 42, 789–817. [1489,1490,1491,1492,1493,1494,1496]
- Neher, E., and Sakmann, B. (1995), *Single-Channel Recording*, New York: Plenum Press. [1497]
- Padilla, O. H. M., Athey, A., Reinhart, A., and Scott, J. G. (2019a), "Sequential Nonparametric Tests for a Change in Distribution: An Application to Detecting Radiological Anomalies," *Journal of the American Statistical Association*, 114, 514–528. [1489]
- Padilla, O. H. M., Yu, Y., Wang, D., and Rinaldo, A. (2019b), "Optimal Nonparametric Change Point Detection and Localization," arXiv preprint arXiv:1905.10019. [1489,1492]
- (2019b), "Optimal Nonparametric Multivariate Change Point Detection and Localization," arXiv preprint arXiv:1910.13289. [1489]
- Page, E. S. (1954), "Continuous Inspection Schemes," *Biometrika*, 41, 100–115. [1488]
- Pein, F., Sieling, H., and Munk, A. (2017), "Heterogeneous Change Point Inference," *Journal of Royal Statistical Society, Series B*, 79, 1207–1227. [1489,1497,1498]
- Rigai, G. (2010), "Pruned Dynamic Programming for Optimal Multiple Change-Point Detection," arXiv preprint arXiv:1004.0887, 17. [1488]
- Rigai, G., Lebarbier, E., and Robin, S. (2012), "Exact Posterior Distributions and Model Selection Criteria for Multiple Change-Point Detection Problems," *Statistics and computing*, 22, 917–929. [1488,1490]
- Rizzo, M. L., and Székely, G. J. (2010), "Disco Analysis: A Nonparametric Extension of Analysis of Variance," *Annals of Applied Statistics*, 4, 1034–1055. [1489]
- Romano, G., Rigai, G., Runge, V., and Fearnhead, P. (2021), "Detecting Abrupt Changes in the Presence of Local Fluctuations and Autocorrelated Noise," *Journal of the American Statistical Association*, 117, 2147–2162. [1497]
- Rudin, L. I., Osher, S., and Fatemi, E. (1992), "Nonlinear Total Variation based Noise Removal Algorithms," *Physica D: nonlinear phenomena*, 60, 259–268. [1488]



- Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton, FL: Chapman and Hall/CRC. [1490]
- Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995), “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray,” *Science*, 270, 467–470. [1496]
- Stransky, N., Vallot, C., Rey, F., Bernard-Pierrot, I., De Medina, S. G. D., Segre, R., De Ryck, Y., Elvin, P., Cassidy, A., Spraggon, C. et al. (2006), “Regional Copy Number–Independent Deregulation of Transcription in Cancer,” *Nature Genetics*, 38, 1386–1396. [1496,1497]
- Tibshirani, R. (1996), “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [1488]
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), “Sparsity and Smoothness via the Fused Lasso,” *Journal of the Royal Statistical Society, Series B*, 67, 91–108. [1488,1492,1494]
- Vanegas, L. J., Behr, M., and Munk, A. (2022), “Multiscale Quantile Regression,” *Journal of the American Statistical Association*, 117, 1384–1397. [1489,1497]
- Wang, D., Yu, Y., and Rinaldo, A. (2020), “Univariate Mean Change Point Detection: Penalization, CUSUM and Optimality,” *Electronic Journal of Statistics*, 14, 1917–1961. [1489,1493]
- (2021), “Optimal Covariance Change Point Detection in High Dimension,” *Bernoulli*, 27, 554–575. [1489]
- Wang, T., and Samworth, R. J. (2018), “High-Dimensional Changepoint Estimation via Sparse Projection,” *Journal of the Royal Statistical Society, Series B*, 80, 57–83. [1489]
- Zou, C., Yin, G., Feng, L., and Wang, Z. (2014), “Nonparametric Maximum Likelihood Approach to Multiple Change-Point Problems,” *Annals of Statistics*, 42, 970–1002. [1489]