Robust Bayesian Non-segmental Detection of Multiple Change-points

Chong Zhong *, Zhihua Ma †, Xu Zhang ‡, and Catherine C. Liu §

Abstract

Change-points detection has long been important and active research areas with broad application in economics, medical research, genetics, social science, etc. It includes themes of the number, locations, and jump sizes of change-points. However, most existing methods focus on segment parameters or features, regardless of Bayesian or frequentist. We propose an innovative non-segmental approach to detect multiple change-points by concentrating the abrupt changes into a global distributional parameter, which is characterized by a function of states of the system. We construct a class of discrete spike and Cauchy-slab variate prior, which distinguishes from existing such two-group mixture priors by a dynamic rate of a latent indicator. A 3-sigma discrimination criterion of change-points is built with sigma being the standard deviation of the sequence of differences between marginal maximum a posteriori estimates on two adjacent discretized states. It intrinsically guarantees reasonable false positive rate to prevent over-detection. The proposed method is powerful and robust to unveil structure changes of the autoregression model for house prices in London and the linear regression model for age-specific fertility in US, not to mention consistent detection results of shifts of mean or scale on known data sets. Abundant simulations demonstrate that the proposed method outperforms state-of-the-art others in finite-sample performance. An R package BaRDCP is developed for detecting changes in mean, scale, and the regression or autocorrelation coefficient.

Keywords: Completely random measure; Discrete spike-and-slab prior; Maximum a posteriori; Sparsity; Three-sigma rule.

*The author is a PhD student at Department of Applied Mathematics, The Hong Kong Polytechnic University, HKSAR.
†The author is an assistant professor at School of Economics, Shenzhen University, Shenzhen, China
‡The author is a PostDoc at School of Mathematical Sciences, South China Normal University, Guangzhou, China.
§The author is an associate professor at Department of Applied Mathematics, The Hong Kong Polytechnic University, HKSAR (macliu@polyu.edu.hk).
1 Introduction

Detection of multiple change-points is a classical problem in a broad range of fields not limited to medicine, biology, econometrics, and social science. It has wealth of studies to detect change-point issues covering topics of number, location, and jump size of change-points when data of interest is univariate. We may categorize methods of multiple change-points detection into two main streams, penalization methods and binary-segmentation variants.

Penalization methods optimize an objective function in the sum of segment-specific costs and a penalty. The cost is versatile and chosen based on types of changes (mean, scale, or autocorrelation for instance) while the penalty term is deterministic to the methodology. For the penalty term, linear $l_0$ penalization to the vector of segment parameters/features to control the number of change-points might be the most popular choice (Yao (1984); Boysen et al. (2009); Killick et al. (2012); Romano et al. (2021); among others); alternatives include $l_0$ penalization to the discrete derivative of the stepwise signal (Frick et al. (2014); Jula Vanegas et al. (2021); among others), and $l_1$ penalization to the vector of segment parameters/features and their jump sizes (Tibshirani et al. (2005); Chernozhukov et al. (2017); among others).

We note that Bayesian approaches can be attributed to penalization methods in the sense that one models priors to automatically penalize the number of change-points (Fearnhead (2006); Wyse et al. (2011); Ko et al. (2015); among others), or even cover ratios between observations in segments and total sample size (Du et al., 2016).

Binary segmentation variants examine all possible choices of change-point through some user-specified criteria. Heuristic methods are usually employed to simplify exhaustive search, including the binary segment method and its variants (Vostrikova (1981); Olshen et al. (2004); Fryzlewicz (2014); Baranowski et al. (2019); among others), and a bottom-up searching algorithm to overcome configurations of change-points that are not suitable for binary division (Fryzlewicz, 2018). Meanwhile, testing-based methods are widely used under high-dimensional time series data settings (Chen et al. (2021); Zhang et al. (2021); among others).
These methods avoid the unstable tuning procedure, while the criteria are case specific, limiting their applications to general scenarios.

Summarizing the approaches, one may find all methods aforementioned are segment-based in the sense that, one used to represent the abrupt change into some stepwise functions, where $K$ change-points partition $K+1$ segments (Chib (1998); Frick et al. (2014); Fryzlewicz (2018); among others). Then one’s inference to change-points fully fully depends on functions (Killick et al. (2012); Du et al. (2016); Romano et al. (2021); among others) or statistics (Fryzlewicz (2014); Baranowski et al. (2019); Jula Vanegas et al. (2021); among others) based on these segmental parameters or features. The length of the segments or within-segment variation will be sensitive and incurs over- or under- detection of change-points (see our numerical results). In addition, algorithms may be sensitive to the initial value (Ko et al., 2015), not to mention burden by tuning parameter(s) (Fearnhead and Rigaill, 2019). Thus we still need detection approach that is accurate, robust, and expedient in implementation.

In this paper, we propose an innovative approach to detect change-points efficiently and robustly within a Bayesian paradigm, to be called BaRDCP, which takes three steps to complete the change-point detection, a) We model the change mechanism by an infinite dimensional global parameter that comes out of a class of distribution family; b) we construct a novel two-group mixture distribution to model the global curve function parameter which has dynamic rate of indicator weight to characterize a kind of sparsity that may be modeled by the change-point; 3) based on the maximum a posteriori estimates, we raise an acute test statistic, which can discriminate change-points decently without over-or under- detection. Our methods outperforms other state-of-the-art methods in all numerical studies.

2 Methodology

Let $Y_i \in \mathbb{R}$ be the data observed at state $t_i \in \mathcal{T}$ for $i \in [N]$, where $\mathcal{T}$ is a state space representing time, spatial locations, or just natural indices. We assume that the distributions
of $Y_i$ come from a class of parametric distribution family $f(y|\theta, \gamma)$, where $\theta \in \Theta$ is the underlying global parameter that fully determines all possible variations of the distribution, and $\gamma$ is the nuisance parameter vector. We model $\theta \equiv \theta(t)$, a curve function defined on the state space $\mathcal{T}$ to characterize the change mechanism, which has been expressed as a stepwise function in literature (Du et al. (2016); Baranowski et al. (2019); among others). Let $\tau_{1:K} = (\tau_1, \ldots, \tau_K)$ be $K$ change-points of $\theta$, where $K$ is unknown. Rather than looking into segment parameters, we propose a non-segmental Bayesian approach to detect change-points $\tau_{1:K}$.

### 2.1 Shrinkage prior for $\theta(t)$

We model $\theta(t)$ through a completely random measure (Kingman, 1967) on $\mathcal{T}$ with the atomic expression

$$
\theta(t) \sim \mathcal{H} \equiv \mathcal{H}(t) = \sum_{\ell=1}^{\infty} h_{\ell} I(\xi_{\ell} \leq t),
$$

where $\xi_{\ell}$ are atoms at which random jumps take place, and $h_{\ell}$ are corresponding uncertain jump sizes or heights. According to Cauchy criterion of convergence, to guarantee the absolute convergence, there exists some large integer $M$ such that the sum of $|h_{\ell}|$ is negligible for $\ell > M$. Hence, we may approximate $\theta(t)$ through the following truncated form

$$
\theta(t) \sim \mathcal{H}_L \equiv \mathcal{H}_L(t) = \sum_{\ell=1}^{L} h_{\ell} I(\xi_{\ell} \leq t).
$$

In practice, one may assume the number of change-points $K$ is bounded by some large number $L$, say, $L = [N/d_0]$, the integer part of the ratio between the number of observations $N$ and $d_0$, one’s prior belief on minimum distance between any two adjacent change-points. Note that both the atomic representation $\mathcal{H}$ and its truncated approximation $\mathcal{H}_L$ are independent increment processes. As a common practice in independent increment processes (Ghahramani and Griffiths (2005); Paisley et al. (2010); among others), the random atoms $\xi = (\xi_1, \ldots, \xi_L)^T$ can be modeled by a noninformative uniform prior $\xi_{\ell} \sim U(t_1, t_N)$, yielding
equally likely change-point among all data points and a dense posterior distribution. In the following context, we focus on modeling the vector of random heights \( \mathbf{h} = (h_1, \ldots, h_L)^T \) in that \( h_\ell \) are all possible increments of \( \theta(t) \).

**L-sparsity under change-point problems**

Obviously, \( \mathbf{h} \) is \( L \)-sparse in the sense that number of the non-zero entries of \( \mathbf{h} \), i.e. the \( l_0 \) norm or the cardinality, is at most \( L \) (Rigollet and Hütter, 2015). The sparse solution of \( \mathbf{h} \) determines both the number and locations of change-points *simultaneously* since the pairs \((h_\ell, \xi_\ell)\) within the atomic expressions jointly characterize the change-point mechanism from the following thought that, an atom \( \xi_\ell \) is identified as a change-point if the jump on its location is significant enough, or equivalently, the companion height \( h_\ell \) is apparently above the noise level. That is, our \( L \)-sparsity is sort of “vertical” since \( \sum_{\ell=1}^L |h_\ell| = \sum_{i=1}^{N-1} |\theta(t_{i+1}) - \theta(t_i)| \), which indicates that the sparsity of \( \mathbf{h} \) is equivalent to the sparsity of the vector of difference of \( \theta(t) \) on states of observations. In contrast, sparsity in existing literature is characterized “horizontally” through vectors of segment parameters/features such as discrete derivatives of the signal \( \vartheta \) (Frick et al., 2014) and segment-wise mean (Du et al. (2016); Romano et al. (2021)). The dimensionality of such horizontally segmental vectors determines the number of change-points, while their locations are still uncertain. Though perspectives are rotated a right angle, our \( L \)-sparsity is consistent with that in Frick et al. (2014) which is an analogue to the high-dimensional regression setting. Next we assign a shrinkage prior for the sparse \( \mathbf{h} \).

**Dynamic spike-and-slab-Cauchy prior**

Based on the above \( L \)-sparsity, we construct a class of two-group discrete mixture priors for \( h_\ell \), formulated as

\[
\begin{align*}
h_\ell|Z_\ell & \sim (1 - Z_\ell)I(h_\ell = 0) + Z_\ell\text{Cauchy}(h_\ell|0, 1), \quad \text{(spike + Cauchy slab)} \\
Z_\ell|\eta_\ell & \sim \text{Bernoulli}(\eta_\ell), \quad (\eta_1, \ldots, \eta_L) = \text{sort}(p_1, \ldots, p_L) \\
p_\ell|\alpha & \sim \text{Beta}(\frac{\alpha}{L}, 1), \quad \alpha \sim U(0, L),
\end{align*}
\]

(3)
where \(\text{sort}\) is a function that sort \((p_1, \ldots, p_L)\) into a decreasing series such that \(\eta_1 > \eta_2 > \cdots > \eta_L\). The formulation (3) is a variant of the classic spike-and-slab (SaSL) priors, in which the “spike” part is a point mass at zero, the “slab” part is a standard Cauchy density, and particularly the occurrence rates \(\eta_\ell\) of the latent indicators \(Z_\ell\) are dynamic rather than an invariant constant that is widely used in literature. We call the proposed prior (3) DySaSC prior hereafter because of the two characteristics of dynamic rates and the Cauchy slab.

The exponentially decaying \(\eta_\ell\) are employed for several concerns. The foremost is to guarantee that prior process \(\mathcal{H}\) in (1) is convergent so that the difference between priors \(\mathcal{H}\) and \(\mathcal{H}_L\) in (2) is negligible for sufficiently large \(L\). Convergence of the prior process \(\mathcal{H}\) in (1) is supported by the following result.

**Theorem 1** (Uniform convergence). With the DySaSC prior (3), given \(\xi\) and \(\alpha\), the atomic expression of \(\mathcal{H}_L(t) = \sum_{\ell=1}^L h_\ell I(\xi_\ell \leq t)\) in (2) converges to \(\mathcal{H}(t)\) in (1) uniformly almost surely in event for all \(t \in \mathcal{T}\).

The uniform convergence guarantees that the approximation in (2) is good enough as \(L\) is sufficiently large. Meanwhile, it indicates that the approximation \(\mathcal{H}_L\) in (2) is robust to choices of large \(L\), demonstrated by sensitivity analysis in supplemental materials.

**Sparseness of posterior solution to \(h\)**

The proposed DySaSC prior can provide a sparse solution to \(h\) in the sense that there are many exact zero entries as \(L \to \infty\). First of all, we derive the joint posterior density to elucidate the sparseness of the posterior estimate of \(h\). Next, we stress that the posterior estimate of \(h\) is more suitable to correctly specify change-points under our Bayesian paradigm than an \(l_0\)-regularized solution which may be provided by the classic SaSL prior. Finally, we illustrate the use of the Cauchy slab in controlling the sparseness of the posterior estimate of \(h\).

To illustrate their difference, we derive the joint posterior density of all parameters in (2) with the DySaSC prior here. Let \(\pi(\cdot)\) denote a (joint) prior density and \(p\) denote the
posterior. Let $\mathbf{Z} = (Z_1, \ldots, Z_L)^T$, and $\eta = (\eta_1, \ldots, \eta_L)^T$. We rewrite the prior density of $h_\ell$ as $\pi(h_\ell) = \text{Cauchy}(0, 1) \times I(h_\ell = 0)^{1-Z_\ell}$ for simplicity. Since $\pi(\alpha) \propto 1$ and $\pi(\xi) \propto 1$, the joint posterior of $(\mathbf{h}, \mathbf{Z}, \eta, \alpha, \xi)$ is expressed as

$$p(\mathbf{h}, \mathbf{Z}, \eta, \alpha, \xi | \mathbf{Y}, t_{1:N}) \propto \prod_{i=1}^{N} \prod_{\ell=1}^{L} \pi(h_\ell | Z_\ell) \prod_{\ell=1}^{L} \pi(Z_\ell | \eta_\ell) p(\eta)$$

(4)

where $\theta(t_i)$ is expressed in the form of model (2). The solution of maximizing the posterior (4) is equivalent to minimizing the following constrained penalized likelihood

$$- \log p(\mathbf{h}, \mathbf{Z}, \eta, \alpha, \xi | \mathbf{Y}, t_{1:N}) = - \sum_{i=1}^{N} \log f\{Y_i | \theta(t_i)\} + \sum_{\ell=1}^{L} Z_\ell \log(1 + h_\ell^2) + \sum_{\ell=1}^{L} Z_\ell \log \left(\frac{1 - \eta_\ell}{\eta_\ell}\right)$$

$$+ \sum_{\ell=1}^{L} \log \left(\frac{1}{1 - \eta_\ell}\right) + (1 - \alpha/L) \sum_{\ell=1}^{L} \log(\eta_\ell) + C,$$

(5)

$$\text{s.t. } 1 > \eta_1 > \cdots > \eta_L > 0,$$

where $C$ is a constant.

Note that the function $\log\{(1 - x)/x\}$ is decreasing on $(0, 1)$. Hence, given $|\mathbf{h}|_0 = \sum_{\ell=1}^{L} I(Z_\ell = 1)$, the third term of the RHS of (5) is minimized at $\mathbf{Z} = (1_{|\mathbf{h}|_0}, 0_{L-|\mathbf{h}|_0})^T$. Consequently, (5) is further simplified as

$$- \log p(\mathbf{h}, \eta, \alpha, \xi | \mathbf{Y}, t_{1:N})$$

$$= - \sum_{i=1}^{N} \log f\{Y_i | \theta(t_i)\} + \sum_{\ell=|\mathbf{h}|_0+1}^{L} \log(1 + h_\ell^2) + \sum_{\ell=|\mathbf{h}|_0+1}^{L} \log \left(\frac{\eta_\ell}{1 - \eta_\ell}\right) + \frac{\alpha}{L} \sum_{\ell=1}^{L} \log \left(\frac{1}{\eta_\ell}\right) + C$$

$$= - \sum_{i=1}^{N} \log f\{Y_i | \theta(t_i)\} + \sum_{\ell=|\mathbf{h}|_0+1}^{L} \log(1 + h_\ell^2) + \sum_{\ell=|\mathbf{h}|_0+1}^{L} \log \left(\frac{\eta_\ell^{(1-\alpha)/L}}{1 - \eta_\ell}\right) + \frac{\alpha}{L} \sum_{\ell=1}^{L} \log \left(\frac{1}{\eta_\ell}\right) + C,$$

(6)

$$\text{s.t. } 1 > \eta_1 > \cdots > \eta_L > 0.$$

Note that the third term of the RHS of the second equation in (6) is increasing as $\alpha < L$, and the fourth term is decreasing. Hence the posterior solution to $\eta_\ell$ is extreme in the sense that the first $|\mathbf{h}|_0$ entries of $\eta$ will quickly converge to $\eta_1$, while the remaining smaller $\eta_\ell$ will converge to $\eta_L$, the minimum entry of $\eta$ instead. Hence, the optimization solution to (6) is
equivalent to the optimization solution to the following constrained penalized likelihood
\[- \log p(h, \eta_1, \alpha, \xi|Y, t_{1:N}, \eta_L)\]
\[- \sum_{i=1}^{N} \log f(Y_i|\theta(t_i)) + \sum_{\ell=1}^{\|h\|_0} \log(1 + h_{\ell}^2) + \log \left( \frac{1 - \eta_L}{\eta_1^{1-\alpha/L} \eta_L^{\alpha/L}} \right) \|h\|_0 + L \log \left( \frac{\eta_1^{1-\alpha/L}}{1 - \eta_L} \right) + C, \quad (7)\]
\[s.t. \quad 1 > \eta_1 > \eta_L > 0.\]

Note that here we fix \(\eta_L\) to avoid the loss being infinite. In (7) we still penalize the \(l_0\) norm and hence obtain a sparse solution to \(h\).

However, the optimization solution to (7) differs from that the normal \(l_0\) regularized solution provided by traditional SaSL prior (Yen, 2011) since here we take much more complicated multiplicity correction. Actually, the dynamic rate here plays a role of measuring the importance of \(h_{\ell}\) rather than controls the prior probability of sparsity. In our numerical studies, we find that the commonly used Beta hyper-pair for for multiplicity correction (Scott and Berger, 2010) leads to under detection of change-points. The is owing to the uniform prior \(\xi_{\ell}\) i.e. the locations of \(h_{\ell}\) without the prior belief on the minimum distance \(d_0\) between the atoms. As a result, the posterior estimates of \(\xi_{\ell}\) are “dense” but not “sparse” in the sense that, many posterior estimates \(\xi_{\ell}\) will fall into a small neighbour of change-points and come into being a cluster, but not isolate with minimum distance \(d_0\). The widths of such clusters may be interpreted as lengths of credible intervals of \(\theta(t)\) on change-points. Under some conditions, the center of posterior paths of \(\theta(t)\) will converge to the true \(\theta(t)\) and hence the number of clusters of \(\xi_{\ell}\) will also converge to \(K\), the number of change-points. Consequently, \(\|h\|_0\) will surely exceed \(K\) since one cluster consists of several significant jumps. If one adopts the SaSL prior, some important clusters of \(\xi_{\ell}\) consisting of change-points are possible to be penalized, leading to under detection of change-points. This drawback is remedied by the complicated multiplicity correction given by DySaSC in (7).

**Cauchy slab works!**

The above issues caused by the dense \(\xi_{\ell}\) also suggests the use of the Cauchy slab. We employ the Cauchy slab as a kind of heavy tail distribution to avoid overshrinkage of the
non-negligible entries. This is motivated by our strategy of change-point detection that identifies a change-point by significant jumps. Therefore, we expect that the jumps on change-points are as significant as possible. For any real number $x$, the penalty function $\log(1 + x^2)$ induced by the Cauchy slab is smaller than those penalties given by $x^2$ and $|x|$, i.e. penalties induced by Gaussian and Laplace slabs, respectively. This result enables the prior in (3) to avoid non-zero but under noise level solutions of $h_\ell$.

2.2 Maximum a posteriori (MAP) estimates of $\theta(t)$

Even though we are unable to get the analytic form of the posterior, we are able to obtain the marginal maximum a posteriori (MAP) estimates of $u(t_i)$ through Markov Chain Monte Carlo (MCMC) sampling. The MAP estimate is defined as the point with highest posterior density, representing the most possible value of a parameter given the data. Then we conduct a discrimination procedure to detect change-points based on the MAP estimates of $u(t_i)$, denoted as $\hat{\theta}_\text{MAP}(t_i)$. We employ the so-called three-sigma rule of random variables with unimodal density (Pukelsheim, 1994) to build the criterion for discrimination between change-points and stationary points. The logic is naive since one would expect that if a point $t_j$ for $j \geq 2$ is a stationary point, then $\{\hat{\theta}_\text{MAP}(t_j) - \hat{\theta}_\text{MAP}(t_{j-1})\}$, the difference between the MAP estimates on $t_j$ and $t_{j-1}$, is close enough to zero; otherwise for change-points, the difference $\{\hat{\theta}(t_j) - \hat{\theta}(t_{j-1})\}$ will not fall around zero. Then we may treat the difference of two neighbour MAP estimates at stationary points as a population with zero mean, and those difference at change-points as outliers.

2.3 Discrimination of change-points

Let $\{\zeta_j\}_{j=2}^n$ be the diffed series $\{\hat{\theta}(t_j) - \hat{\theta}(t_{j-1})\}_{j=2}^n$, and let $\hat{\sigma} \equiv [\text{Var}\{\zeta_j\}_{j=2}^n]^{1/2}$ be their sample standard deviation. We have the following criterion to discriminate change-point locations $\tau_k, k \in 1, \ldots, K$. 
**3-sigma rule.** If at $t_j$, $|\zeta_j - 0| > 3\hat{\sigma}$, then $t_j$ is discriminated as a change-point; otherwise, a stationary point.

The three-sigma rule is often applied to outlier detection under Gaussian populations. As it is easy to estimate the parameter $\alpha_0$ in $g(t; \alpha_0)$ by plugging the known form of $g$ into the likelihood, we assume that $g(t; \alpha_0) \equiv \alpha_0$ is a single scalar parameter to simplify the theoretical difficulty. Although the observed data may be non-Gaussian, we show the consistency and asymptotic normality of $\zeta_j$ at stationary points under following assumptions. The proof is deferred to the supplementary material.

We impose following assumptions.

**(A1)** The density $f_Y(\cdot|\theta)$ is twice differentiable on $\theta$.

**(A2)** For any $\theta \in \Theta$, the Fisher information $J(\theta) < \infty$ always exists.

**(A3)** The density $f_Y(\cdot|\theta)$ is a log concave function of $\theta$.

Assumptions (A1) and (A2) are general assumptions about continuous distributions; assumption (A3) requires a log-concave distribution family, covering Gaussian, Gamma, Beta, Laplace, logistic, and many other continuous distribution families. We further assume that $g(t; \alpha_0) \equiv \alpha_0$ is a constant function, we may simply specify that $\alpha_0 = 0$ such that $\theta(t) = \theta(t)$. In case there is no change-point, we will have $\theta(t) \equiv \theta_0$ for all $t$.

**Proposition 1** (Stationary points). For any fixed $n > 1$, the time or locations of observations $(t_1, \ldots, t_n)$ are fixed. Let $Y_1, \ldots, Y_n \sim f(\cdot|\theta_0)$ be $n$ i.i.d. samples on $t = t_1, \ldots, t_n$, where no change-point exists. Let $\mathcal{D}$ be the domain of $f$ and $\theta_0$ is an inner point of the parameter space $\Theta$. When $n$ varies to large enough, under assumptions (A3) to (A5), for $j = 2, \ldots, n$, we have

$$\zeta_j \overset{d}{\to} N \left[0, f_j \{J(\theta_0^{-1})\}\right].$$

(The form of $f_j$ remains to be further explored.)
Note that proposition 1 not only holds for the case without change-points, but also holds under the change-point model at stationary points (the proof is similar). Meanwhile, proposition 1 also shows the asymptotic normality of the marginal posterior distribution of diffed process $d_i$ at stationary points. Then a natural question comes that what will happen at change-points.

We shall empirically show that the asymptotic normality of marginal posterior distributions of $\theta(t)$ will not hold at a neighborhood change-points. We need an additional assumption about the variation of $\theta(t)$.

(A4) Suppose $\theta(t)$ takes $K + 1$ distinct values on the $K + 1$ segments $D_1, \ldots, D_{K+1}$, denoted as $\theta_0, \ldots, \theta_K$. There exists $\delta > 0$ such that $\inf_{2 \leq k \leq K+1} \{|\theta_{k-1} - \theta_k|\} > \delta$.

Assumption (A4) is similar to the fourth condition in Du et al. (2016), indicating the existence of the lower bound of the variation of pure jump component $\theta(t)$.

Proposition 2 (Mode-shifting). Let $A_k$ be the neighborhood of the $k$th change-point $\tau_k$ such that $\forall s \in A_k, |s - \tau_k| < \delta$ for some $\delta > 0$. Then the marginal posterior distributions of $u(s)$ are bimodal; then for $s \in A_k$,

\[
\hat{\theta}(s_1)^{MAP} \rightarrow \theta(\tau_{k-1}), \text{ for } s_1 < \tau_k; \quad \hat{\theta}(s_2)^{MAP} \rightarrow \theta(\tau_k), \text{ for } s_2 > \tau_k.
\]

That is, the mode of the marginal posterior distribution of $\theta(s)$ shifts from $\theta(\tau_{k-1})$ to $\theta(\tau_k)$ when crossing change-point $\tau_k$. Therefore, $\hat{\theta}(s_2)^{MAP} - \hat{\theta}(s_2)^{MAP} \neq 0$.

We validate the result in Proposition 2 empirically, referred to supplementary materials. Similar numerical evidence of Proposition 1 is also deferred to supplementary materials. Numerically we find that the marginal posterior distribution at change-points is clearly distinct from that at stationary points.

The above two propositions reveal the facts the marginal posterior distribution of $\theta(t)$ is unimodal at stationary points and is bimodal in neighborhoods of the change-points. Then
the three-sigma discrimination is similar to a kind of hypothesis testing, where the type I error i.e. the false positive rate (FPR) is no more than 0.05 if the density of the population is unimodal (Savage, 1961); and is 0.001 if the population is Gaussian. By the unimodal nature of marginal posterior distributions at stationary points, we directly have the following corollary about the theoretical bound of false positive rate of overestimation of change-points through the three-sigma discrimination.

**Corollary 1.** *The false positive rate of over detection of change-points falls into* $[0.001, 0.05]$.

This corollary is a straightforward result of the three-sigma rule discrimination. It actually provides a theoretical bound of the proportion of overestimation of change-points.

### 2.4 Application scenarios

We illustrate some application scenarios of the proposed method here. The method is not only limited to general parametric families, but also can be applied to detect structural changes in regression/autoregression models.

(i) Changes on mean of Gaussian variable. We have a series of real observations $Y_i \sim N\{\theta(t_i), \sigma^2\}$, for $i = 1, \ldots, N$. The global parameter $\theta(t)$ represents the location parameter.

(ii) Changes on the parameter of Poisson variables. We have a series of integer observations $Y_i \sim \text{Poisson}\{\theta(t_i)\}$, for $i = 1, \ldots, N$. The global parameter $\theta(t)$ characterize the changes on mean and variance simultaneously.

(iii) Changes on variance of Gaussian variables. We have a series of real observations $Y_i \sim N\{\mu, \exp[\theta(y)]\}$, for $i = 1, \ldots, N$. The global parameter $\theta(t)$ represents the scale parameter through an exponential transformation to guarantee the non-negativity.
(iv) Structural changes of an AR(1) model. Data are generated from the model $Y_t = \phi_0 + \theta(t)Y_{t-1} + \epsilon_t$, where, $\phi_0$ is the fixed intercept, $E(\epsilon_t) = 0$ and $E(\epsilon_t\epsilon_s) = \sigma^2 I(t = s)$. The global parameter $\theta(t)$ represents the autocorrelation coefficient.

(v) Structural changes of a linear regression model. Data are recorded as independent pairs of $(y_{tj}, X_{tj})$, for $j = 1, \ldots, n_t$, $t = 1, \ldots, T$. The association between $y$ and $X$ is characterized by

$$y_{tj} = \beta_0 + \theta(t)X_{tj} + \epsilon_{tj},$$

where $\beta_0$ is a fixed intercept, $E(\epsilon_t) = 0$ and $E(\epsilon_t\epsilon_s) = \sigma^2 I(t = s)$. The global parameter $\theta(t)$ represents the regression coefficient at time $t$. By taking $n_t = 1$ for all $t$ and $X_t = Y_{t-1}$, this scenario reduces to scenario (iv).

3 Implementation

We ease some numerical challenges that employing the DySaSC prior may encounter. We first represent the ordered rates $\eta_{\ell}$ into it limiting approximation, and then introduce relates MCMC sampler here.

**Limiting expression of the DySaSC prior**

Recall that the dynamic rates $\eta_{\ell}$ of latent indicators $Z_{\ell}$ are decreasing order statistics of Beta($\alpha/L, 1$) distribution. However, such kind of order statistics brings much difficulties to MCMC sampling since they are not conjugate. Here we adopt the following construction rule for $\eta_{\ell}$,

$$\eta_{\ell} = \prod_{j=1}^{\ell} p_j, \quad p_j | \alpha \sim \text{Beta}(\alpha, 1), \quad \ell = 1, \ldots, L. \quad (8)$$

Such kind of construction of decreasing order statistics of the Beta distribution is proposed by Teh et al. (2007) for the stick-breaking construction of the Indian buffet process. They show that as $K \to \infty$, the joint distribution of the $\eta_{\ell}$ in (8) converges to the $\eta_{\ell}$ in (3) in
distribution. This limiting expression of the DySaSC prior significantly ease the sampling since all parameters are conjugate.

**MCMC sampling in NIMBLE**

We carry out the posterior inference through MCMC sampling algorithm to sample from the posterior distribution. Computation is facilitated by the *nimble* (de Valpine et al., 2017) package in R, which uses **BUGS** type syntax (Lunn et al., 2000) and complies the code into **C++**. **NIMBLE** automatically select suitable samplers for different parameters. For conjugate parameters, say, $p_\ell$, **NIMBLE** assigns Gibbs samplers; for parameters $\xi_\ell$ and $\alpha$, **NIMBLE** assigns the default Metropolis-Hasting sampler; for $h_\ell$ and the corresponding indicator $Z_\ell$, a reverse jump MCMC sampler is assign to speed up the sampling. We build an R package **BaRDCP** based on **nimble** and include several R functions applied to application scenarios mentioned in subsection 2.4.

As proposition 2 shows, the marginal posterior distributions at change-points are bi-modal, leading to lower effective sample sizes (ESS) compared with unimodal posterior distributions. We recommend to run multiple independent chains and aggregate their posterior samples to obtain larger ESS. We find that in all numerical studies the chains are convergent and mixed well.

### 4 Simulations

Comprehensive simulations are conducted to evaluate performance of **BaRDCP**, our proposed method named hereafter. We compare **BaRDCP** with other state-of-the-art methods available in R Archive Network on three tasks, detecting multiple changes on mean, scale, and regression coefficient in linear regression. We adopt an AR(1) model setting with changes on auto-correlation coefficient as an alternative of the scenario of changes on regression coefficient for the sake of comparison, since there is rare work available to linear regression model with an explanatory variable. Results of additional simulations under settings of
changes on Gaussian mean on dependent segments, changes on mean of heavy tail distributions and changes of autocorrelation coefficient with model misspecification are deferred to supplemental materials.

**Settings**

We consider following settings. Under each setting, 300 Monte Carlo replicate datasets are generated.

**S.1** Changes of normal mean on short segments. We have $N = 400$ independent Gaussian observations with $K = 7$ change-points at $t = (50, 100, 150, 200, 250, 300, 350)$, leading to 8 segments with segment mean $\mu = (0, 1.5, 3, 1.5, 3, 0.5, 2, 0)$. The common scale parameter is set to be $\sigma = \sqrt{2}$.

**S.2** Changes of normal mean on unequal-length segments with large variations. We have $N = 916$ independent Gaussian observations with $K = 11$ change-points at $t = (81, 134, 178, 267, 346, 413, 528, 577, 636, 741, 822)$, leading to 12 segments with segment mean $\mu = (0, 1.23, -0.248, 0.861, -0.534, 1.057, 0.369, 1.331, 0.483, 1.105, -1.101, 0)$. The common scale parameter is set to be $\sigma = 1$. Some jump sizes are smaller than the within segment variation, leading to much difficulties to correctly identify change-points.

**S.3** Changes of Poisson parameter. We have $N = 400$ independent Poisson variables with $K = 7$ change-points at $t = (50, 100, 150, 200, 250, 300, 350)$, leading to 8 segments with segment parameter $\lambda = (1, 0.25, 2, 1, 3, 1.5, 2.5, 1)$.

**S.4** Changes of normal scale with small variations on mean. We have $N = 756$ independent Gaussian observations with $K = 7$ change-points at $t = (150, 250, 300, 450, 550, 650, 700)$, leading to 8 segments with segment scales $\sigma = (1, 1.68, 0.57, 0.20, 2.18, 3.09, 1.83, 1)$. Meanwhile, we allow small variations on the mean such that the segment mean is $\mu = (0.056, 0.047, -0.034, -0.017, 0.032, 0.068, -0.042, 0.017)$. 

15
Changes of autocorrelation coefficient in an AR(1) model. The data generating process is $Y_t = \phi Y_{t-1} + \epsilon_t$. We have $N = 450$ observations with 5 change-points at $t = (50, 100, 200, 300, 400)$, leading to 6 segments with segment autocorrelation coefficient $\phi = (0.5, -0.5, 0.65, -0.25, -0.85, 0.45)$. The model error $\epsilon_t \sim N(0, 1)$.

Estimators

In all simulations, we set $L = 25$ in BaRDCP unifiedly. We run 4 independent MCMC chains and obtain 1000 scans in each chains thinned from total 20000 after a burn-in period of 8000 iterations. Finally we get 4000 posterior samples for change-point discrimination. The prior belief on minimum distance between change-points is set to be $d_0 = 15$.

Competitors vary among different settings since none of them can be applied to all the above simulation settings. For settings S.1, S.2 and S.3, where the mean parameter changes, we compare with the NOT method by Baranowski et al. (2019) in package not, the FDRSeg method by Li et al. (2016) in package FDRSeg, the SMUCE method by Frick et al. (2014) in package StepR, the WBS method by Fryzlewicz (2014) in package wbs, and the PELT method by Killick et al. (2012) in package changepoint (Killick and Eckley, 2014); for setting S.4, where the scale parameter changes, we compare with NOT, SMUCE and PELT methods; for setting S.5, where data are autocorrelated, we compare with the WBSTS method by Korkas and Pryzlewicz (2017) in package wbsts and the B-P method by Bai and Perron (2003) in package struchange (Zeileis et al., 2002). The tuning parameters for the competing methods are set to the values recommended by the corresponding R packages.

Assessments and results

Several assessments are employed to measure the accuracy of detected number of change-points and the accuracy of locations of estimated change-points. We report the a frequency table for $\hat{K} - K$, the difference between number of detected change-points and the true number of change-points to evaluate the accuracy in detected number of change-points. To measure the accuracy in locations, three assessments are considered, precision, recall, and the
scaled Hausdorff distance (Hausdorff). For all true change-points, we count one true positive (TP) if there is at least one change-point identified within a window of 10 data-points and compute the number of false positive (FP) as the number of predicted changes minus TP. Let $K$ be the true number of change-points. Then precision is computed as $TP / (TP + FP)$, and recall is computed as $TP / K$. The scaled Hausdorff distance is computed as

$$d_H = N^{-1} E \left[ \max \left\{ \max_{j=0, \ldots, K} \min_{k=0, \ldots, K+1} |\tau_j - \hat{\tau}_k|, \min_{k=0, \ldots, K+1} \min_{j=0, \ldots, K+1} |\hat{\tau}_k - \tau_j| \right\} \right],$$

where $t_0 = \tau_0 < \cdots < \tau_K < \tau_{K+1} = t_N$ and $t_0 = \hat{\tau}_0 < \hat{\tau}_1 < \cdots < \hat{\tau}_K < \hat{\tau}_{K+1} = t_N$ denotes true and estimated change-points, respectively. The scaled Hausdorff distance takes values in $[0, 1]$ and is the smaller the better.

From Table 1 we find that BaRDCP outperforms in frequency of correctly specifying the number of change-points in all settings. In contrast, other competitors tend to under detect the number of change-points except the setting S.3, where changes take place on both mean and variance of data. This phenomenon may be an evidence that performances of frequentist approaches for change-point detection seem to rely heavier on large data sizes and long segments than our Bayesian approach. Although the jump sizes under these simulation settings (especially setting S.2) are not significant enough to make the changes be indentified by eyes, BaRDCP still enjoys the highest recall in all settings, demonstrating its capability to correctly identify change-points. The precision and Hausdorff distance given by BaRDCP outperform under settings S.1 and S.3, and are competitive under others settings. Note that other winners on precision and scaled Hausdorff distance actually under estimate the number of change-points, while a most parsimonious estimator usually brings higher precision and lower Hausdorff distance. In summary, BaRDCP performs to be the most competitive and robust to correctly specify the number of change-points and estimate their locations accurately.
Table 1: Results of change-points detection under settings S.1 to S.5 among 300 Monte Carlo replicates. The best results are bold.

| Setting | Method   | Frequency of $\hat{K} - K$ | Precision | Recall | $d_H \times 10^2$ |
|---------|----------|-----------------------------|-----------|--------|------------------|
|         |          | $\leq -3$ | -2 | -1 | 0 | +1 | +2 | $\geq +3$ |       |
| S.1     | BaRDCP   | 1          | 1 | 33 | 252 | 13 | 0 | 0 | 0.95 | 0.94 | 2.1 |
|         | NOT      | 9          | 12 | 31 | 227 | 19 | 2 | 0 | 0.93 | 0.91 | 2.4 |
|         | SMUCE    | 47         | 68 | 130 | 55  | 0  | 0 | 0 | 0.85 | 0.7  | 3.1 |
|         | WBS      | 16         | 35 | 95  | 138 | 14 | 2 | 0 | 0.93 | 0.84 | 2.5 |
|         | FDRSeg   | 6          | 16 | 63  | 171 | 29 | 10 | 5 | 0.90 | 0.88 | 3.0 |
|         | PELT     | 1          | 6  | 12  | 210 | 52 | 16 | 3 | 0.91 | 0.93 | 2.8 |
| S.2     | BaRDCP   | 15         | 48 | 77  | 144 | 15 | 1 | 0 | 0.93 | 0.87 | 1.5 |
|         | NOT      | 52         | 91 | 49  | 101 | 7  | 0 | 0 | 0.94 | 0.82 | 1.4 |
|         | SMUCE    | 136        | 113| 50  | 1   | 0  | 0 | 0 | 0.86 | 0.67 | 2.1 |
|         | WBS      | 68         | 120| 74  | 38  | 0  | 0 | 0 | 0.95 | 0.79 | 1.2 |
|         | FDRSeg   | 28         | 71 | 74  | 100 | 23 | 2 | 2 | 0.88 | 0.81 | 2.2 |
|         | PELT     | 38         | 101| 42  | 107 | 12 | 0 | 0 | 0.83 | 0.83 | 1.4 |
| S.3     | BaRDCP   | 4          | 28 | 113 | 148 | 6  | 1 | 0 | 0.90 | 0.82 | 2.9 |
|         | NOT      | 37         | 71 | 77  | 90  | 23 | 1 | 1 | 0.87 | 0.74 | 3.2 |
|         | SMUCE    | 10         | 68 | 151 | 69  | 2  | 0 | 0 | 0.89 | 0.76 | 3.0 |
|         | WBS      | 1          | 5  | 34  | 41  | 65 | 63 | 85| 0.64 | 0.76 | 4.8 |
|         | FDRSeg   | 0          | 3  | 6   | 8   | 20 | 22 | 241| 0.47 | 0.83 | 5.7 |
|         | PELT     | 25         | 50 | 102 | 61  | 38 | 15 | 9 | 0.77 | 0.69 | 3.5 |
| S.4     | BaRDCP   | 0          | 75 | 71  | 150 | 4  | 0 | 0 | 0.84 | 0.75 | 2.3 |
|         | NOT      | 25         | 221| 39  | 14  | 0  | 0 | 1 | 0.91 | 0.67 | 1.5 |
|         | SMUCE    | 40         | 211| 49  | 0   | 0  | 0 | 0 | 0.64 | 0.64 | 1.2 |
|         | PELT     | 1          | 153| 58  | 83  | 5  | 0 | 0 | 0.88 | 0.72 | 2.0 |
| S.5     | BaRDCP   | 0          | 0  | 98  | 154 | 46 | 2 | 0 | 0.85 | 0.82 | 2.6 |
|         | WBSTS    | 4          | 36 | 74  | 122 | 48 | 14 | 2 | 0.61 | 0.47 | 2.8 |
|         | B-P      | 102        | 68 | 128 | 2   | 0  | 0 | 0 | 0.89 | 0.38 | 1.8 |
5 Applications

5.1 Array Comparative Genomic Hybridization (ACGH) data

We analyse the dataset of DNA copy number using array comparative genomic hybridization (ACGH) for 43 different individuals with a bladder tumor, which is available in R package ecp. For each individual, copy number is recorded on 2215 locations. Interest usually falls to the changes in the mean copy number. Hence we employ BaRDCP for Gaussian mean changes. As the number of change-points is usually considered to be quite large, we set $L = 65$ to incorporate sufficiently many change-points. The prior belief on the minimum distance between change-points are set as $d_0 = 15$. We display the analysis result of the 37th individual in this article.

Figure 1: The ACGH data with locations of estimated change-points given by BaRDCP (blue) and NOT (red) methods.

As shown by Figure 1, BaRDCP detects 14 change-points at locations (73, 124, 263, 342, 524, 583, 657, 719, 744, 1724, 1906, 1965, 2041, 2143). For comparison, not detects 16 change-points, 14 among which are close to those detected by BaRDCP. A major difference is that not detect two more change-points at 1271, leading to a very short segment of length 7. We conje-
ture that these two additional change-points are caused by the violent shock in the interval (1271, 1277), where the range is (−0.538, 0.411), greatly larger than the standard deviation of the dataset (0.227). In BaRDCP, these two change-points are merged since their distance is smaller than $d_0$. Such conjecture may be demonstrated by the fact that if one simply removes the data points from 1260 to 1275, not detects exact 14 change-points with similar locations as that of BaRDCP. Other methods tend to be easy to over detect the change-points, detecting much more change-points than these two methods.

5.2 Agriculture industry daily return data

Portfolio allocation is an important topic in finance research, and the key is daily returns of the portfolio. Guo et al. (2017) studied the portfolio allocation of 49 industries using a Fama-French three factor model with the portfolio returns as the response. All data used can be freely downloaded from the website http://mba.tuck.dartmouth.edu. The daily returns of 49 industry portfolios from 1995 to 2014 are available. We select the data of the agriculture industry observed from Jan 2007 to Dec 2019. The data are centered around zero while the variance seems to change along with the time. Hence we apply BaRDCP to detect changes of scale within this dataset. We set $L = 25$ and $d_0 = 15$ in this case.

As shown by Figure 2, BaRDCP detects 7 change-points at $t = (37, 137, 206, 333, 426, 513, 617)$, with 8 estimated segment scale parameters (1, 1.31, 1.81, 3.27, 2.67, 5.55, 2.41, 1.59). The sample standard deviations on each segments are (1.17, 1.369, 1.873, 3.529, 2.550, 5.748, 2.459, 1.653). The estimated scale parameters and sample standard deviations are quite close, and both suggest a significant shift on the estimated change-points, supporting the detection result by BaRDCP. For other competitors, changepoint and not detect five and four change-points respectively, which may be considered as under detection of the change-points. As an evidence, we conduct another simulation with five change-points in supplementary materials, where BaRDCP still outperforms.
5.3 US age-specific fertility rate (ASFR) data

The declining birth rates in many developed countries arouses much interest to analysis of the annual Age-Specific Fertility Rate (ASFR). Given the year \( t \), let \( B_t(s) \) be the number of births during the year to females of a specified age \( s \), and \( N_t(s) \) be the number of females of the age \( s \) in that reference year. In year \( t \), the ASFR \( X_t(s) \) is defined as the ratio between \( B_t(s) \) and \( N_t(s) \). We collect ASFR data in US from 1940 to 2021 at ages from 22 to 35, the age period which covers the age with the highest ASFR. Then totally we obtain 1134 responses \( X_t(s) \), for \( t = 1, \ldots, 81 \) and \( s = 22, \ldots, 35 \). The visualization of the ASFR versus age grouped by year is shown in Figure 3.

Figure 3: Visualization of the ASFR data in US from 1940 to 2021 for women aged from 22 to 35.
The relationship between the ASFR and specific ages from 22 to 35 seems to be linear. Hence, we consider a linear model with changes on the regression coefficient to characterize their association. For each \( t \), we consider a linear model \( X_t(s) = \beta_0 + \theta(t)s + \epsilon_{ts} \), where the regression coefficient \( \theta(t) \) may change along with time \( t \), \( \beta_0 \) is a fixed intercept and \( \epsilon_{ts} \sim N(0, \sigma^2) \) are i.i.d. model errors. We apply \text{BaRDCP} to detect changes of \( \theta(t) \), where the state of data is set to be the year \( t \). We set \( L = 25 \) and \( d_0 = 15 \).

![Graph showing data relationship]

(a) Relationship between age and ASFR before year 1992.

(b) Relationship between age and ASFR after year 1992.

Figure 4: Visualization of the pre- and post-change-points ASFR data in US.

Only one change-point is detected by \text{BaRDCP} at \( t = 1992 \). To understand the effect of the change-point, we plot the curves of ASFR versus ages before and after 1992 in Figure 4. From the figure we clearly find that before the change-point, the ASFR is almost linearly decreasing as the age increases so that the ASFR at age 22 is the highest. However, after the change-point, the association between the ASFR and ages is non-linear and even non-monotone,
where the ASFR increases first and reaches the peak at age 29, and then decreases. This might be the phenomenon that people with higher income often share lower fertility. This application provides an evidence that even though BaRDCP meets model misspecification, it is still powerful enough to correctly identify the change-points.

5.4 House prices in London Borough of Newham

We further explore a real dataset, the average monthly property price $P_t$ in the London Borough of Newham. We take the average of all properties and select the data recorded from January 2010 to November 2020 and we totally have 131 observations. This dataset was once analyzed by Fryzlewicz (2021) to identify the shortest interval of change-points under an AR(1) model. We adopt the AR(1) model $P_t = \theta(t)P_{t-1} + \theta_0 + \epsilon_t$, where the autocorrelation coefficient $\theta(t)$ is treated as the global parameter that may change, the intercept $\theta_0$ is fixed, and $\epsilon_t \sim N(0, \sigma^2)$ are independent model errors. We set $L = 25$ and $d_0 = 15$.

As shown in Figure 5, BaRDCP detects 1 change-point locating at Oct 2016 (location 82). The date of change-point is close to the beginning of vote of Britain’s EU membership referendum, indicating that the structural change may be caused by the event. The B-P method provides a consistent result of change-point detection, where the estimated location is 79. Meanwhile, the estimated confidence interval given by R package nsp (Fryzlewicz, 2021) is (24, 97), which covers the change-point estimated by BaRDCP.

References

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, 18(1):1–22.

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2019). Narrowest-over-threshold detection
Figure 5: House prices in London Borough of Newham and locations of estimated change-points given by BaRDCP (the red line).

of multiple change points and change-point-like features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):649–672.

Boysen, L., Kempe, A., Liebscher, V., Munk, A., and Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics*, 37(1):157–183.

Chen, L., Wang, W., and Wu, W. B. (2021). Inference of breakpoints in high-dimensional time series. *Journal of the American Statistical Association*, pages 1–13.

Chernozhukov, V., Hansen, C., and Liao, Y. (2017). A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1):39–76.

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

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2):403–413.
Du, C., Kao, C.-L. M., and Kou, S. (2016). Stepwise signal extraction via marginal likelihood. 
*Journal of the American Statistical Association*, 111(513):314–330.

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213.

Fearnhead, P. and Rigaill, G. (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, 114(525):169–183.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281.

Fryzlewicz, P. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *The Annals of Statistics*, 46(6B):3390–3421.

Fryzlewicz, P. (2021). Narrowest significance pursuit: inference for multiple change-points in linear models. *arXiv preprint arXiv:2009.05431*.

Ghahramani, Z. and Griffiths, T. (2005). Infinite latent feature models and the indian buffet process. *Advances in neural information processing systems*, 18.

Guo, S., Box, J. L., and Zhang, W. (2017). A dynamic structure for high-dimensional covariance matrices and its application in portfolio allocation. *Journal of the American Statistical Association*, 112(517):235–253.

Jula Vanegas, L., Behr, M., and Munk, A. (2021). Multiscale quantile segmentation. *Journal of the American Statistical Association*, pages 1–14.

Killick, R. and Eckley, I. (2014). changepoint: An r package for changepoint analysis. *Journal of statistical software*, 58(3):1–19.
Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Kingman, J. (1967). Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78.

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

Korkas, K. K. and Pryzlewiczv, P. (2017). Multiple change-point detection for non-stationary time series using wild binary segmentation. *Statistica Sinica*, pages 287–311.

Li, H., Munk, A., and Sieling, H. (2016). Fdr-control in multiscale change-point segmentation. *Electronic Journal of Statistics*, 10(1):918–959.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs—a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572.

Paisley, J. W., Zaas, A. K., Woods, C. W., Ginsburg, G. S., and Carin, L. (2010). A stick-breaking construction of the beta process. In *ICML*.

Pukelsheim, F. (1994). The three sigma rule. *The American Statistician*, 48(2):88–91.

Rigollet, P. and Hütter, J.-C. (2015). High dimensional statistics. *Lecture notes for course 18S997*, 813(814):46.
Romano, G., Rigail, 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*, pages 1–16.

Savage, I. R. (1961). Probability inequalities of the tchebycheff type. *Journal of Research of the National Bureau of Standards-B. Mathematics and Mathematical Physics B*, 65(3):211–222.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619.

Teh, Y. W., Grür, D., and Ghahramani, Z. (2007). Stick-breaking construction for the indian buffet process. In *Artificial intelligence and statistics*, pages 556–563. PMLR.

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 (Statistical Methodology)*, 67(1):91–108.

Vostrikova, L. Y. (1981). Detecting “disorder” in multidimensional random processes. In *Doklady akademii nauk*, volume 259, pages 270–274. Russian Academy of Sciences.

Wyse, J., Friel, N., Rue, H., et al. (2011). Approximate simulation-free bayesian inference for multiple changepoint models with dependence within segments. *Bayesian Analysis*, 6(4):501–528.

Yao, Y.-C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical bayes approaches. *The Annals of Statistics*, pages 1434–1447.

Yen, T.-J. (2011). A majorization–minimization approach to variable selection using spike and slab priors. *The Annals of Statistics*, 39(3):1748–1775.
Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). strucchange: An r package for testing for structural change in linear regression models. *Journal of statistical software*, 7:1–38.

Zhang, Y., Wang, R., and Shao, X. (2021). Adaptive inference for change points in high-dimensional data. *Journal of the American Statistical Association*, pages 1–12.