Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

CONTINUOUS OPTIMIZATION FOR OFFLINE CHANGE POINT DETECTION AND ESTIMATION

Hans Reimann[1], Sarat Moka[2], and Georgy Sofronov[3]

[1]Department of Mathematics, University of Potsdam, Potsdam, Brandenburg, GERMANY
[2]School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, AUSTRALIA
[3]School of Mathematical and Physical Sciences, Macquarie University, Sydney, NSW, AUSTRALIA

ABSTRACT

This work explores use of novel advances in best subset selection for regression modelling via continuous optimization for offline change point detection and estimation in univariate Gaussian data sequences. The approach exploits reformulating the normal mean multiple change point model into a regularized statistical inverse problem enforcing sparsity. After introducing the problem statement, criteria and previous investigations via Lasso-regularization, the recently developed framework of continuous optimization for best subset selection (COMBSS) is briefly introduced and related to the problem at hand. Supervised and unsupervised perspectives are explored with the latter testing different approaches for the choice of regularization penalty parameters via the discrepancy principle and a confidence bound. The main result is an adaptation and evaluation of the COMBSS approach for offline normal mean multiple change-point detection via experimental results on simulated data for different choices of regularisation parameters. Results and future directions are discussed.

1 INTRODUCTION

Change point detection and estimation are an incredibly diverse and widely scattered field in applied and mathematical statistics, with a large variety of applications. To provide a high-level intuition, change point detection may be understood as a signal processing tool for identifying abrupt changes in the generative parameters of a data sequence. While a strong line of work in change point detection is well established with Page's pioneering work (see Page (1954)) and rigorous results by Chernoff and Zacks (1964), Lorden (1971) and Sen and Srivastava (1975), many aspects of this problem are open and the general understanding of good solutions depends strongly on the problem at hand (Niu et al. 2016; Truong et al. 2020; Ma et al. 2020). Among the open research questions, the simultaneous detection of multiple change points in large data sets is of major interest.

Taking a machine learning and data scientific motivated approach, in this paper, we explore the applicability of recent advances in best subset selection of covariates in linear regression proposed by Moka et al. (2024). This method, a continuous optimization approach for best subset selection, claims to offer faster performance compared to existing exhaustive search methods, while maintaining comparable accuracy. Furthermore, it achieves superior accuracy relative to methods of similar speed, thus embodying highly desirable properties for addressing the subset selection problem in offline change point detection and estimation. Some of the existing methods including Huang et al. (2005), Rinaldo (2009) and Qian and Jia (2016) utilize Lasso-regularisation which provides convex relaxation of the underlying sparsity problem. However, such Lasso-regularization approaches do not directly focus on optimizing the actual discrete constrained best subset selection problem at hand.

In this paper, we focus on the popular normal mean multiple change point model assuming univariate, independent Gaussian random variable with a constant variance. This simple yet challenging model has applications in several fields including genomics and econometrics. Further, detection of change points

can be reduced to detecting change in mean in adapted sequences Niu et al. (2016) such as detection of change in slope reducing to detection of mean in the difference sequence. The challenge in detection and estimation of an unknown number of change points in a piece-wise identical distributed sequence of observations resembles a problem in choice of regularization penalty parameter in practice. We will explore two different methods and compare their performance assuming knowledge on variance of the noise.

This work formulates offline change point detection as a best subset selection problem and it allows us to explore recent advances in this domain. Our goal is to provide a proof of concept as well as a stepping stone for future investigations. In particular, we employ the novel advances in continuous optimization for best subset selection in the given context to provide a new and modern perspective in approaching the change point detection problem via modern machine learning methods. We believe that the employed framework of simultaneous offline multiple change point detection as a sparse linear regression problem has high potential in utilizing the broad established toolbox of both areas.

The remaining paper is organized as follows. In Section 2, we formulate the normal mean multiple change point model as a best subset selection problem in linear regression and then discuss the key results of Moka et al. (2024). Our method for change point detection problem is presented in Section 3. In Section 4, linear time running time complexity of the proposed method is presented. In Section 5 the simulation experimental designs and the corresponding results are presented. Finally, a discussion on the results and future directions for further studies are presented in Section 6.

2 THEORETICAL BACKGROUND

In this section, we first provide formulation of offline multiple change point detection as a subset selection problem in linear regression via boundary conditions on sparsity of covariates (see Rinaldo (2009) and Niu et al. (2016)). Then, we provide the gradient based continuous optimization introduced of Moka et al. (2024) towards solving the subset selection problem.

2.1 Change Point Detection via Subset Selection

Let $Y = (Y_1, \ldots, Y_n)$ be a sequence of independent random variables with the distribution of Y_i denoted as F_i. A change point is then a point in the set of indices of the sequence $\tau \in \{2, \ldots, n\}$ such that $F_{\tau-1} \neq F_\tau$. The central aim in change point detection lies in estimating the location of an instance τ. In multiple change point detection, the goal is to estimate both the locations of the change points and their total number, denoted as K.

Consider the multiple mean change point detection problem with K change points with their locations denoted by $\tau_1, \ldots, \tau_K$ where $\tau_k \in \{2, \ldots, n\}$ and $\tau_k < \tau_{k+1}$ for each $k \in \{1, \ldots, K-1\}$. Note that the case where $K = n-1$ (i.e., an instance of change with every next sequence member) has generally very little insights as the interest usually lies in detecting piece-wise independent and identically distributed (*iid*) random variables in between instances of change. On the other extreme, the case where $K \leq 1$ (i.e., whether a single change point is present or not) is of great interest in that it quantifies whether a full *iid* assumption on a sequence of data is justified. For the normal mean multiple change point model, which is central in this work, we assume $K > 1$ and take $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ for each $i \in \{1, \ldots, n\}$ with mean μ_i and common variance σ^2. A change point τ_k now refers to an unknown location in the sequence Y with $\mu_{\tau_k - 1} \neq \mu_{\tau_k}$. In addition, we make the usual assumption that the total number of change points K is unknown. Consequently, the normal mean multiple change point model can be understood as a sequence Y generated by $K+1$ piece-wise constant signals μ_i and zero-mean *iid* noise sequence $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$, so that

$$Y_i = \mu_i + \varepsilon_i, \quad i \in \{1, 2, \ldots, n\}.$$

The central challenge then translates to estimating the mean sequence

$$\mu_1 = \ldots = \mu_{\tau_1 - 1} \neq \mu_{\tau_1} = \mu_{\tau_1 + 1} = \ldots = \mu_{\tau_k - 1} \neq \mu_{\tau_k} = \ldots = \mu_{\tau_K - 1} \neq \mu_{\tau_K} = \ldots = \mu_n.$$

Besides, we introduce an artificial change point at $\tau_0 = 1$ whenever the initial mean $\mu_1 \neq 0$.

We now briefly discuss two major criteria in change point detection based on signal-to-noise ratio and sparsity; refer to, e.g., Niu et al. (2016) for more details. Both these criteria often provide the main assumptions and quantities of interest in investigation and analysis. In order for the instances of mean change to be insightful and identifiable, they need to have relevant degree of distinction from one another which is measured in shift between means relative to the noise. The corresponding parameter is the minimal difference in signal standardized by the noise standard deviation, formally defined as,

$$\delta = \min_{i \in \{2,3,\ldots,n\}} \frac{|\mu_i - \mu_{i-1}|}{\sigma},$$

which is hereafter simply referred to as signal-to-noise ratio. Additionally, two instances of change in too close proximity do not allow for any valuable inference as there is too little data in between. This motivates control of the minimal distance between instances of change, which is formally defined as

$$L = \min_{k \in \{1,2,\ldots,K\}} \{\tau_k - \tau_{k-1}\}.$$

The assumption on sparsity hereby also implicitly forms a boundary condition of distance of the first and last instance of change from the respective start and end of the data sequence. Further following Niu et al. (2016), both these parameters are combined into the central quantity of signal strength $S = \delta^2 L$. While establishing the weakest possible condition on S for full recovery of all change points is an open problem, results by Arias-Castro et al. (2005) show that $S \geq 2\log n$ could be a necessary condition for full recovery of multiple change points. Additionally, Rinaldo (2009) and Qian and Jia (2016) showed for their fused Lasso regression approach that the consistency $\mathbb{P}(\hat{\tau} = \tau) \to 1$ as $n \to \infty$ with $\tau = (\tau_1, \ldots, \tau_K) \in \mathbb{R}^K$ denoting the vector of change point locations, where $\hat{\tau}$ being the estimator of τ. However, this requires a fairly strong assumption of $\delta^2 \gg \log n$, which may not be applicable in practice as indicated in Niu et al. (2016).

The above mentioned linear regression approach to change point detection is best understood in terms of the statistical inverse problem of recovering the mean vector μ from the data sequence y, a realization of the random vector Y. While a L^1-distance for tackling the problem may be reasonable under considerations of robustness, we want to apply a L^2-norm instead for desirable properties on differentiability. This results in the ordinary least squares formulation $\|y - \mu\|_2^2$. Via introducing an artificial design matrix X and parameter vector β, we can further transform the problem to a typical statistical inverse problem. In particular, let $\beta_i = \mu_i - \mu_{i-1}$ for $i \in \{1,2,\ldots,n\}$ with $\mu_0 = 0$ and X be a $n \times n$ lower-triangular matrix with all the non-zero elements being 1. Then, we obtain $\mu = X\beta$ and $Y = X\beta + \varepsilon$. The resulting statistical inverse problem is then

$$\min_{\mu \in \mathbb{R}^n} \|y - \mu\|_2^2 = \min_{\beta \in \mathbb{R}^n} \|y - X\beta\|_2^2,$$

which is the ordinary least squares approach for estimating the parameter vector β in linear regression. By construction, the piece-wise constant structure of the mean vector μ translates to 0 entries in β as

$$\mu_i = \mu_{i-1} \quad \text{if and only if} \quad \beta_i = 0,$$

for all $i \in \{2,3,\ldots,n\}$. This motivates the use of modern high-dimensional sparse regression methods for recovery of μ, the central connection exploited in this work.

With the given definition of β, we observe that $\beta_i \neq 0$ for $i = \tau_k$ and $k \in \{0,1,\ldots K\}$, possibly including the artificial change point τ_0. That is, imposing sparsity in instances of change translates to controlling the total number of non-zero entries of β. Hence, we obtain

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{n} \|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_0 := \sum_{i=1}^{n} \mathbb{I}(\beta_i \neq 0) \leq K, \tag{1}$$

where $\mathbb{I}(\cdot)$ denotes the usual indicator function. This problem is a reformulation of the normal mean change point detection to an ordinary least squares regression problem with the sparsity induced by the L^0-constraint $\|\beta\|_0 \leq K$. This is a well studied highly non-convex best subset selection problem in linear regression; see Müller and Welsh (2010) and Hui et al. (2017) for details.

Among the existing methods for solving (1), the Lasso regression is a popular convex relaxation of the subset selection problem where a L^0-norm $\|\beta\|_0$ is replaced by a L^1-norm $\|\beta\|_1$. In particular, the convex relaxation of the Lasso regularization is given by

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1,$$

where the parameter λ in the penalty allows us to control the sparsity in the solution. Application of the Lasso regularization to the normal mean change point detection was studied in Huang et al. (2005) and (Harchaoui and Lévy-Leduc 2010) with additional constraints for a fused Lasso penalty explored in Tibshirani and Wang (2008) and further work in Rinaldo (2009) and Qian and Jia (2016). While this change yields desired properties in fast computation and estimation, it does not generally guarantee to address the best subset problem as above; refer, e.g., Zhu et al. (2020) and Hazimeh and Mazumder (2020). However, the main insight from all these studies is the necessity of tuning λ in penalty Friedman et al. (2007), e.g. via cross-validation as heuristic for parameter choice.

Two very recent works by Wang et al. (2020) and Verzelen et al. (2023) support the sketched approach and show minimax optimality of their respective estimators. Additionally, in emphasizing the value of post-processing of detected instances of change, they provide a valuable future line of work. Both these works focus on deriving lower bounds for the described multiple change point model, however, with some adaptations and extensions, i.e., a more general sub-Gaussian noise term. Each then introduces an estimator based on the best subset selection approach with additional focus on the estimation of the total number of changes. Yet in both cases, estimators constructed based on (1) are shown to reach minimax optimal rates with high computational efficiency. The gradient based estimator described in this work aims to fall in line with these results and provide a solution via continuous optimization similar to the Lasso approach.

2.2 Continuous Optimization for Subset Selection

We now introduce the key idea of Moka et al. (2024) that enables a continuous, gradient based optimization for best subset selection. Towards this, recall the optimization problem (1) and consider a vector $s \in \{0,1\}^n$ with

$$s_i = \begin{cases} 1, & \text{if } i = \tau_k \text{ for some } k \in \{0,1,\dots K\}, \\ 0, & \text{otherwise,} \end{cases}$$

that is, the indices of non-zero elements s correspond to the location of the change points. Hence, $|s| = \sum_{i=1}^n s_i = K$ is equivalent to $\|\beta\|_0 = K$. With this notation, for any $s \in \{0,1\}^n$, let $X_{[s]}$ be the matrix of size $n \times |s|$ obtained from X by keeping only the columns with indices j where $s_j = 1$. Then, for any $K \in \{1,2,\dots p\}$, the exact best subset problem (1) can be restated as

$$\min_{s \in \{0,1\}^n} \frac{1}{n} \|y - X_{[s]} \widehat{\beta}_{[s]}\|_2^2, \quad \text{subject to } |s| \leq K, \tag{2}$$

where $\hat{\beta}_{[s]}$ is the low-dimensional ordinary least squares solution utilizing $X_{[s]}$ instead of the full model design matrix X for estimation.

Even for a given K, the resulting problem in (2) is in general non-deterministic polynomial-time hard (NP-hard) as stated in Natarajan (1995). Exact methods are only feasible for small values of n. A more recent approach via mixed integer optimization as in Bertsimas et al. (2016), although faster than the exact method, remains fairly slow for practical application Hazimeh and Mazumder (2020). The Lasso

regularization, mentioned above, is a reasonable relaxation regarding feasibility concerns, however, it does not recover the best subset as in (1) or (2), as stated before recalling Hazimeh and Mazumder (2020) and Zhu et al. (2020). As a novel approach, continuous optimization method towards best subset selection, COMBSS, aims to combine computational feasibility and accuracy in enabling modern gradient based continuous optimization methods for the best subset selection problem; see COMBSS with SubsetMapV1 in Moka et al. (2024) for details.

The central idea of COMBSS is to relax the binary vector $s \in \{0,1\}^n$ in (2), which takes values at the corners of the hypercube $[0,1]^n$, to a vector $t \in [0,1]^n$ evolving on the entire hypercube. For each such t, let

$$M_t = X_t^\top X_t + n\,(I - \mathrm{Diag}\,(t \odot t)),$$

with $X_t = X\mathrm{Diag}(t)$, where I is the identity matrix, $\mathrm{Diag}\,(v)$ is the diagonal matrix with v as its diagonal, and $\odot$ denotes the Hadamard (or, element-wise) product. Further, let $\widetilde{\beta}_t$ be a solution of the linear equation (in terms of u),

$$M_t u = X_t^\top y. \tag{3}$$

Then, the continuous Boolean relaxation of (2) is

$$\min_{t\in[0,1]^n} \frac{1}{n}\|y - X_t\widetilde{\beta}_t\|_2^2, \quad \text{subject to} \sum_{i=1}^n t_i \le K. \tag{4}$$

The construction of $\tilde{\beta}_t$ hereby guarantees equality of (4) and (2) at the corner points of the hypercube $[0,1]^n$. Instead of solving this linear constrained problem, COMBSS aim to optimize its Lagrangian form as

$$\min_{t\in[0,1]^n} \frac{1}{n}\|y - X_t\widetilde{\beta}_t\|_2^2 + \lambda \sum_{i=1}^n t_i, \tag{5}$$

where the regularization penalty parameter λ, analogous to K in the exact problem, allows us to control the sparsity in the solution obtained. Further, the continuous relaxation inside the hypercube enables the desired smoothness in t of the objective function $\|y - X_t\widetilde{\beta}_t\|_2^2$ and the Lagrangian function in (5). COMBSS main insight lies in suggesting candidates of best subsets by utilizing this smoothness, without discrete constraints, while directly addressing the best subset selection unlike previous continuous optimization approaches. This way popular gradient based optimization methods in machine learning, such as Adam optimizer, can easily be applied to optimize (5). Further, via a transformation from $[0,1]^n$ to $\mathbb{R}^n$, COMBSS transforms the box-constrained optimization (5) to an equivalent unconstrained optimization so that the continuous optimizer does not face any boundary issues. The final point t for each λ is mapped to an approximate solution s of (2) by taking $s_i = 1$ if t_i is close to 1, otherwise $s_i = 0$, for all i.

The algorithm and exact mathematical framework with full details of the procedure as well as the respective guarantees for smoothness are given in Moka et al. (2024). The key enabling feature for computational feasibility lies in efficiently solving the linear equation (3). Further, there is software made available for both Python and R to implement the method. Our experimental results emphasize the ability of COMBSS to recover best subsets in both low and high dimensions. To summarize, by providing smooth solution paths in the continuous relaxation of the exact best subset selection problem, COMBSS has enabled a new and promising line of work. Among the benefactors may well be the best subset selection problem in simultaneous normal mean multiple change point detection.

3 METHODOLOGY

The central idea of this work is eminent: Normal mean multiple change point detection and estimation in recovering the piece-wise constant mean vector μ is equivalent to finding the best subset of covariates

via sparsity conditions imposed on the artificial design matrix X and parameter vector β with $\mu = X\beta$ in (1). This optimization problem is equivalent to the reduced lower-dimensional discrete optimization problem in (2). Solving the best subset selection using an exact method, such as *leaps-and-bounds*, is only computationally feasible when n is less than 31 Furnival and Wilson (2000). However, the novel results of Moka et al. (2024) enable approaching it via continuous optimization for solving (5). Furthermore, the regularisation penalty parameter λ enables us to control the total number of change points. Thus, we create a grid of λ values and execute COMBSS for each λ to produce a vector $t^*_\lambda \in [0,1]^n$, which is further mapped to a binary vector $s^*_\lambda \in \{0,1\}^n$. This binary vector s^*_λ has 1's at locations corresponding to the change points. So, using s^*_λ, we compute $\widehat{\beta}_{[s^*_\lambda]}$ with the non-zero elements corresponding to the shift in the mean values at the change points.

Taking a supervised perspective, where we assume that the number of change points K is known, the problem then translates into finding a suitable λ over a grid corresponding to K change points. This is straightforward, e.g., via a simple grid search or interval-halving algorithm. However, the more practical problem is to take an unsupervised perspective assuming that K is unknown. This turns the challenge at hand into a typical regularisation penalty parameter choice for statistical inverse problems. While we do have a better understanding of the respective penalty parameter λ compared to the previously introduced Lasso approaches, it still leaves us with a similar conclusion to Rinaldo (2009) in the unsupervised case – it is an important open problem.

To address this issue of penalty parameter choice, we intend to employ established methods for regularization choice in the context of inverse problems. In particular, we investigate the discrepancy principle and an adaptation of the discrepancy principle based on a lower confidence bound. Let β_λ denote the parameter vector β obtained for a given choice of λ. Following Richter (2021), the discrepancy principle for statistical inverse problems aims to choose λ such that the error of $\|y - X\beta_\lambda\|_2^2$ aligns with the expected model error, $\mathbb{E}[\|Y - X\beta\|_2^2]$. However, an analytical evaluation of this expectation is only feasible in specific cases for the data sequence y. Given the data sequence as introduced via the normal mean multiple change points and constant noise variance σ^2, we can standardize the inverse problem and find β_λ such that it recovers $\|\frac{1}{\sigma}(y - X\beta_\lambda)\|_2^2 = n$, as under the true model $\|\frac{1}{\sigma}(y - X\beta)\|_2^2 \sim \chi^2(n)$ with expectation equal to the degrees of freedom. Again, the basic idea hereby is that the regularization penalty is chosen so that the average or expected level of error given the estimated parameter vector should resemble that of a true parameter vector given the model. For a much more sophisticated and thorough analysis of the discrepancy principle for statistical inverse problems see Blanchard and Mathé (2012). Adaptation to a confidence bound extends this comparison via using a χ^2-quantile instead of the expectation, to control the effect of the model error in the inverse problem. This then leads to the choice of a regularization penalty that is at least as restrictive as with the discrepancy principle. The intuition here is that of a one-sided composite hypothesis of the regularisation penalty λ that recovers the parameter vector β in $H_0: \lambda \leq \lambda_0$ vs $H_1: \lambda > \lambda_0$ for λ_0 the penalty parameter recovering the true model. The χ^2-distribution of the standardized regularization again provides a decision threshold. The regularisation penalty is chosen such that the supposedly plausible error is maximal within the confidence bound $\max_{\lambda>0} \|\frac{1}{\sigma}(y - X\beta_\lambda)\|_2^2 \leq \chi^2_{1-\alpha}(n)$ for $\alpha \in (0,1)$. In essence this means increasing λ and thus decreasing K, right until H_0 needs to be rejected. The basic idea then is that under the true parameter vector the accumulated model error is plausible up to the threshold $\chi^2_{1-\alpha}(n)$. As noted, this accommodates at least as much error as the discrepancy principle and therefore may allow for more restrictive choices of regularisation penalty and smaller values of K. The main limitation of the discrepancy principle and its adaptation is that they require knowledge of the noise variance of the sequence. whereas they do not require a constant variance assumption.Solutions via estimating the variance from the sequence itself are difficult, but could be implemented via pooled variance estimators. Via an iterative scheme, repeated estimation of the pooled covariance and change point detection may provide ways of implementation of the discrepancy principle even for unknown covariance. However, this is likely to lead to problems in dependence of the estimated variance and the analysis of the expected error after

standardization. Accordingly, the main insight for the following experiments lies in a simple litmus test of whether and to what extent information about the variance helps in the choice of the regularisation penalty.

4 TIME COMPLEXITY

We now show that our implementation of COMBSS exhibits $O(n)$ running time complexity. In the general COMBSS approach of Moka et al. (2024), for optimizing (5), computational complexity of the gradients of the objective function is dictated by computations of solving linear equations of the form $M_t u = b$. The matrix M_t can be shown to be dense positive-definite for $t \in (0,1)^n$ and thus, exact gradient computation can be expensive with a time complexity of $O(n^3)$. When a conjugate gradient method is used for approximately solve such a linear equation, the general COMBSS exhibits $O(n^2)$ complexity.

It is important to note that in the general setting, the design matrix X is part of the data. However, in our case, X has an artificial design of being lower triangular matrix with the non-zero elements equal to 1. We now take advantage of this fact to modify the COMBSS algorithm to achieve a time complexity of $O(n)$. Towards this, Lemma 1 establishes that the inverse of $X^\top X$ is tridiagonal.

Lemma 1 Consider the lower triangular matrix X with the non-zero elements being 1. Then, $A = (X^\top X)^{-1}$ is a tridiagonal matrix with the principle diagonal being $(1,2,2,\ldots,2)$ (i.e., first element is 1 and all other elements are equal to 2) and both the sub-diagonals being $(-1,-1,\ldots,-1)$ (i.e., all elements equal to -1).

We can easily establish Lemma 1 using a proof by induction. Start with the basic case at $n = 2$ and assume that it is true for n. Then, we can show that this is true for $n+1$ via using the Banachiewicz inversion lemma Tian and Takane (2005).

Lemma 2 (Woodbury Matrix Identity) For any conformable matrices A, U, C and V,

$$(UAV + C)^{-1} = C^{-1} - C^{-1}U(A^{-1} + VC^{-1}U)^{-1}VC^{-1}.$$

Refer to Woodbury (1950) for a proof of Lemma 2. Note that with $T_t = \mathrm{Diag}\,(t)$ and $D_t = d(I - T_t^2)$, we can write $M_t = \left(T_t X^\top X T_t + D_t\right)/n$. Further, as a consequence of Lemma 2, we have the following result

Theorem 1 Let $\widetilde{M}_t = (X^\top X)^{-1} + T_t D_t^{-1} T_t$. Then, for the lower triangular matrix X, $\widetilde{M}_t$ is a tridiagonal matrix, and

$$M_t^{-1} = n D_t^{-1} - n D_t^{-1} T_t \widetilde{M}_t^{-1} T_t D_t^{-1}, \quad t \in [0,1)^n. \tag{6}$$

Proof of Theorem 1 is straightforward: Since $(X^\top X)^{-1}$ is tridiagonal given in Lemma 1 and $T_t D_t^{-1} T_t$ is diagonal, $\widetilde{M}_t$ must be tridiagonal. Selecting $A = X^\top X$, $C = D_t$, and $U = V = T_t$ in Lemma 2, we get (6).

Recall that we want to compute the expressions of the form $M_t^{-1} b$. Since D_t is diagonal, computing expressions of the form $D_t^{-1} v$ has $O(n)$ complexity. This means, the bottleneck is the complexity of computing expressions of the form $\widetilde{M}_t^{-1} v$. Since $\widetilde{M}_t$ is a tridiagonal matrix, this operation is easy to execute using the popular method of the *Thomas algorithm* or *tridiagonal matrix algorithm* which is known to have $O(n)$ complexity; see, e.g., Lee (2011) and Higham (2002). In conclusion, the computational complexity of solving a linear equation of the form $M_t u = b$ is $O(n)$ only. Moka et al. (2024) have shown that the COMBSS method approaches an ε-stationary point in $O(1/\varepsilon^2)$ iterations (independent of n). Thus, our implementation has $O(n)$ time complexity.

5 EMPIRICAL EVALUATION RESULTS

By construction and with the extensive simulation results in Moka et al. (2024), the proposed approach appears reliable in best subset selection in linear regression. The experimental simulation study in this paper aims to investigate the two levels of knowledge. First, there is the supervised level with knowledge about the number of change points K. How well does our continuous optimization method recover the

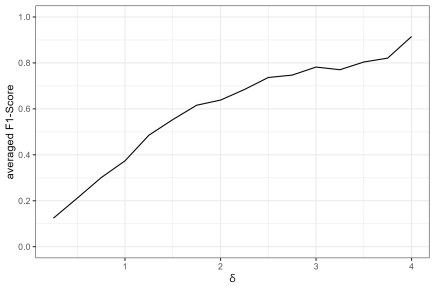
locations of change given K is known? Second is the unsupervised level with knowledge about the variance of the noise. How well do the discrepancy principle and the confidence bound perform for our method in recovering the true number of change points as well as their locations? All experiments follow a basic set up reduced to its essentials. We conduct two sets of experiments for known and unknown number of change points. In each set, the first respective experiment scales signal-to-noise ratio, δ^2, with a fixed distance between change points. The second experiment then scales minimal distance between change points, L, for fixed signal-to-noise ratio. We set $\sigma^2 = 1$ throughout all experiments only adjusting the difference in mean values to set signal-to-noise ratio. We conduct 100 Monte Carlo simulations of the noise for each experiment and value of the scaling quantities δ and L. Following Aminikhanghahi and Cook (2017) and Truong et al. (2020), the metrics for evaluation will be the F1-score and the Hausdorff distance metric. The F1-score hereby serves as a measure of accuracy combining sensitivity and precision with a tolerance level of $\tau_k \pm L/20$ to allow for negligible inaccuracies. The Hausdorff distance metric serves as evaluation of robustness of the detection procedure in measuring the largest difference between a change point and its closest estimated counterpart. For better comparison under scaling minimal distance, we also standardize the Hausdorff metric by L. The evaluation criteria are averaged over all Monte-Carlo simulations for a concrete value of δ or L. All experiments are conducted in R (version 4.2.2).

5.1 Experiment A - Known Number of Change Points

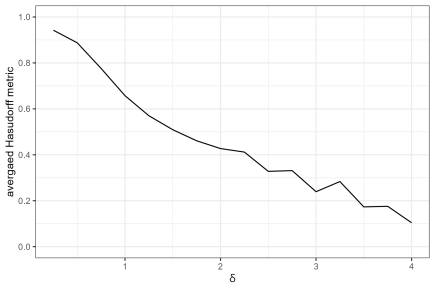
For each Monte-Carlo simulation, the regularization penalty λ for the proposed algorithm is chosen via a interval halving approach until it produces a result with the right number of change points K. In very rare instances, if the interval halving fails to produce a result with right number K in a certain number of steps, the corresponding Monte-Carlo sample will be skipped in the aggregation of the Hausdorff metric.

5.1.1 Experiment A1 - Scaling δ^2

Set $n = 150$ with a total of 4 change points at $\tau = (31, 61, 91, 121)$, so $L = 30$. The mean sequence has a simple staircase form with difference in mean at the change points $\mu_{\tau_k} - \mu_{\tau_k - 1}$ for $k \in \{1, 2, 3, 4\}$ ranging from 0.25 to 4 with step-size $\triangle\delta = 0.25$ and $\mu_{\tau_0} = 0$, so it should not be detected. The necessary condition in (Arias-Castro et al. 2005) is therefore $\delta_{\text{crit}} \geq \sqrt{\frac{2\log(150)}{30}} \approx 0.58$ and is surpassed after two steps of δ.



(a) Scaling signal-to-noise ratio via δ and resulting F1-score.



(b) Scaling signal-to-noise ratio via δ and resulting Hausdorff metric.

Figure 1

There are two main insight obtained from this experiment. First, the approach yields the desired result in that both measures, F1-score and Hausdorff metric, improve for increasing signal-to-noise ratio δ with reliable estimation after $\delta \approx 2.5$. The second, and much more valuable insight at this stage, is best portrayed by the histogram of change point locations over all Monte-Carlo simulations. The proposed algorithm seems to have a tendency in selecting change points in close proximity to each other for a given K and struggles to recover the last change point – even for larger values of δ. To give a concrete example, it tends

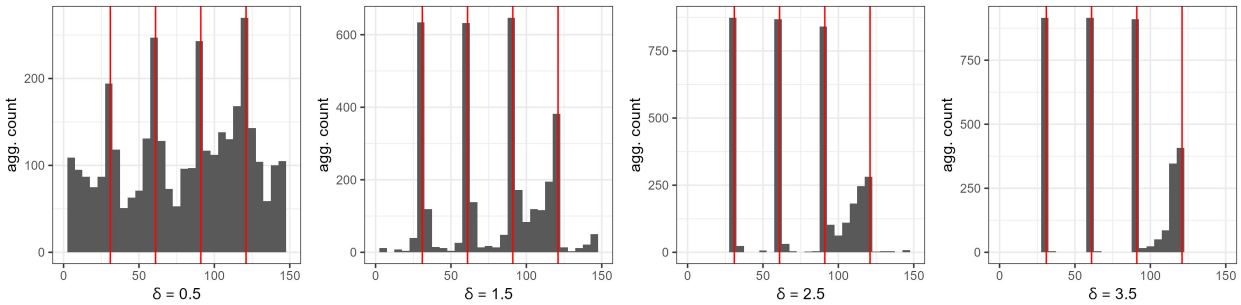
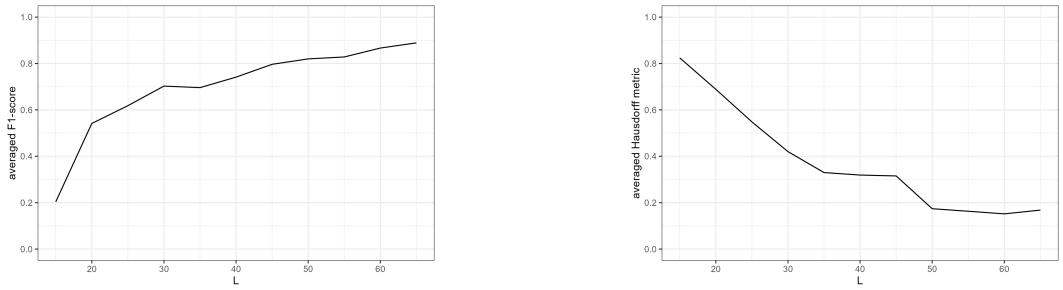


Figure 2: Histogram of estimated change points over all Monte Carlo runs for selected values of δ.

to choose $\hat{\tau} = (31, 61, 90, 91)$ or something similar for a given $K = 4$ instead of the actual vector τ and it seemingly does so consistently. This emphasizes the need of post-treatment of suggested candidates of change point locations, i.e., via clustering, before matching a given K. Furthermore, this behavior shows consistently throughout all experiments although we only include the histogram here.

5.1.2 Experiment A2 - Scaling L

Again, set $K = 4$ with the mean sequence of a simple staircase form with difference in mean $\delta = 2$ at the change points, so $\mu_{\tau_k} - \mu_{\tau_k - 1} = 2$ for $k \in \{1, 2, 3, 4\}$. We take $\mu_{\tau_0} = 0$ for the artificial change point $\tau_0 = 1$ so it should not be detected. The choice of δ is hereby taken from the previous experiment indicating good detection for a minimal distance $L = 30$ yet with potential for showing effect in either direction. Scaling now the distance between change points L taking values from 15 to 65 with step-size $\triangle L = 5$, we therefore have change points at $\tau = (L+1, 2L+1, 3L+1, 4L+1)$ and sequence length $n = 5L$.



(a) Scaling signal-to-noise ratio via L and resulting F1-score.

(b) Scaling signal-to-noise ratio via L and resulting Hausdorff metric.

Figure 3

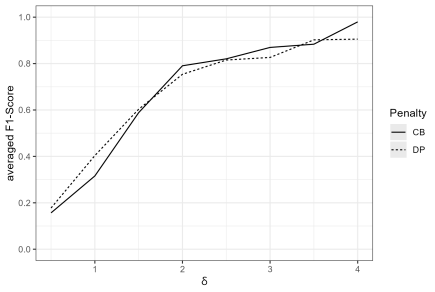
Again, w observe the desired behavior in improving performance measures for increasing L tempering off for larger values. Accordingly, we may conclude the generally desired behavior of the approach in both metrics of interest for increasing signal strength S even for this more pilot implementation.

5.2 Experiment B - Unknown Number of Change Points

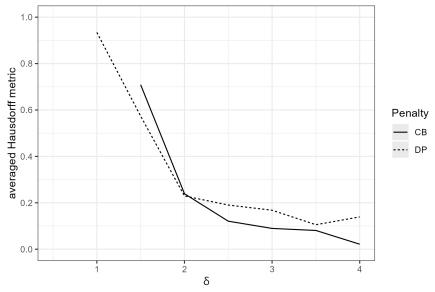
For each Monte-Carlo simulation, the regularization penalty λ for the COMBSS algorithm is chosen via both of the described methods. The discrepancy principle and the confidence bound hereby start at $\lambda = 0$ and then increase with a step size $\triangle\lambda = 0.005$ until the respective conditions are surpassed. The discrepancy principle then picks whichever λ is closer to the expected error and the confidence bound chooses the previous penalty value λ before the threshold is exceeded. Both estimate $\hat{\tau}$ for their respective choice of λ.

5.2.1 Experiment B1 - Scaling δ^2

Set $n = 100$ with a total of 3 change points at $\tau = (26, 51, 76)$, so minimum distance $L = 25$. The mean sequence has a simple staircase form with the difference of the mean at the change points $\mu_{\tau_k} - \mu_{\tau_k - 1}$ for $k \in \{1, 2, 3\}$, so δ, ranging from 1 to 4 with step-size $\triangle\delta = 0.5$. The artificial change point is set to $\mu_{\tau_0} = 0$, so it should not be detected. The necessary condition for detection derived in (Arias-Castro et al. 2005) is at$\delta_{\text{crit}} \geq \sqrt{\frac{2\log(100)}{25}} \approx 0.61$ and is exceeded after the first step.



(a) Scaling signal-to-noise ratio via δ and resulting F1-score for both penalty choice approaches.



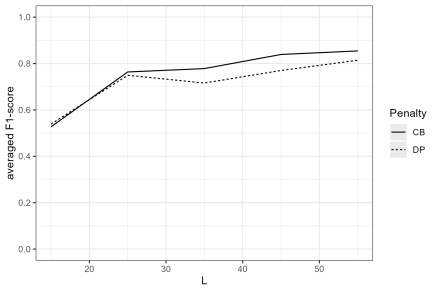
(b) Scaling signal-to-noise ratio via δ and resulting Hausdorff metric for both penalty choice approaches.

Figure 4

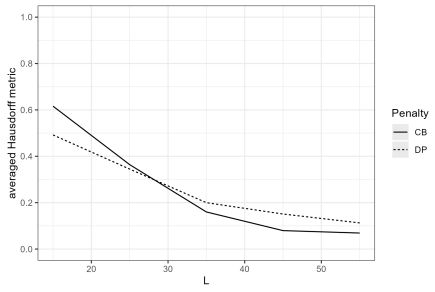
We observe the desired behavior for both curves, similar to the previous experiments. The confidence bound seems to provide a slightly better choice of λ for larger values of δ, but further investigation is needed.

5.2.2 Experiment B2 - Scaling L

Again, set $K = 3$ with the mean sequence of a simple staircase form with difference in mean $\delta = 2$ at the change points $\mu_{\tau_k} - \mu_{\tau_k - 1} = 1$ for $k \in \{1, 2, 3\}$ and $\mu_{\tau_0} = 0$ for the artificial change point $\tau_0 = 1$ so it should not be detected. δ is hereby chosen for the same reasons as in experiment A2 to provide range for chnage in performance in both directions. Scaling now the distance between change points L taking values from 10 to 50 with step-size $\triangle L = 5$, we therefore have change points at $\tau = (L+1, 2L+1, 3L+1)$ and sequence length $n = 4L$.



(a) Scaling signal-to-noise ratio via L and resulting F1-score for both penalty choice approaches.



(b) Scaling signal-to-noise ratio via L and resulting Hausdorff metric for both penalty choice approaches.

Figure 5

Similar to before, we observe improving performance in both metrics for increasing minimal distance L between change points. There does not seem to be a major difference in the two choices of penalty value λ. Both methods seem viable with good performance and should be included in follow up studies.

6 DISCUSSION, CONCLUSIONS AND OUTLOOK

The presented work is a proof of concept to show that there is great value to the approach utilizing COMBSS for candidates of change points in the normal mean change point model. This approach here still has plenty of room for improvement in implementation, yet the results provide a valuable basis for further investigation. While the experiments in the work at hand do not include a direct comparison to current Lasso approaches, results in Moka et al. (2024) suggest for COMBSS competitive results in the more general case. Additionally, a real world application is highly desirable but needs to be subject for future investigations.

The main insights currently are the computational cost and the spurious change points estimated. The term spurious change points as mentioned in Pilliat et al. (2023) hereby refers to the tendency to detect change points in close proximity to each other. In practice, they can simply be aggregated into a single change point in post treatment whenever they are in too close proximity to each other. This problem mainly influenced the first two supervised experiments while the unsupervised experiments simply gave too large of a number of change points without this treatment, however, neither of the metrics were influenced by that in a meaningful way. For the F1-score, the tolerance accounted for that and in the Hausdorff metric only the closes change point is of interest. A possible line of future work arising from this problem could be to investigate the vast toolbox of linear regression for testing significance of supposed spurious change points. Other future lines of work will lie in exploring advances in adjacent topics. One such, is in investigating the discrepancy principle for choice of regularisation penalty under the sophisticated results in Blanchard and Mathé (2012). The artificial design matrix is of trace type and can therefore be fit to their results for better statistical guarantees. Further, great interest lies in the gradient of the COMBSS algorithm given its specific choice of X. It promises high potential for improving computational efficiency and in combination with the approach of Niu et al. (2016) in utilizing symmetry by connecting the ends of a data sequence, it may help addressing the current issue of the neglected last change point.

To conclude, utilizing the key concept in COMBSS in continuous relaxation of the best subset selection problem for modern gradient based optimization yields a novel, promising and valuable line of work for simultaneous offline multiple change point detection in the normal mean multiple change point model. It offers a large variety of directions for detailed investigations, optimizations and adaptations. With ongoing research on COMBSS, both fields may very much profit from each other sharing their advances. With the addition of COMBSS to the vast topic of change point detection has a new, additional connection to popular research in machine learning, further opening it up to interested researchers.

REFERENCES

Aminikhanghahi, S. and D. J. Cook. 2017. "A Survey of Methods for Time Series Change Point Detection". *Knowledge and Information Systems* 51(2):339–367.

Arias-Castro, E., D. L. Donoho, and X. Huo. 2005. "Near-Optimal Detection of Geometric Objects by Fast Multiscale Methods". *IEEE Transactions on Information Theory* 51(7):2402–2425.

Bertsimas, D., A. King, and R. Mazumder. 2016. "Best Subset Selection via a Modern Optimization Lens". *The Annals of Statistics* 44(2):813–852.

Blanchard, G. and P. Mathé. 2012. "Discrepancy Principle for Statistical Inverse Problems with Application to Conjugate Gradient Iteration". *Inverse Problems* 28(11):115011.

Chernoff, H. and S. Zacks. 1964. "Estimating the Current Mean of a Normal Distribution Which is Subjected to Changes in Time". *The Annals of Mathematical Statistics* 35(3):999–1018.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani. 2007. "Pathwise Coordinate Optimization". *The Annals of Appplied Statistics* 1(2):302–332.

Furnival, G. M. and R. W. Wilson. 2000. "Regressions by Leaps and Bounds". *Technometrics* 42(1):69–79.

Harchaoui, Z. and C. Lévy-Leduc. 2010. "Multiple Change-Point Estimation with a Total Variation Penalty". *Journal of the American Statistical Association* 105(492):1480–1493.

Hazimeh, H. and R. Mazumder. 2020. "Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms". *Operations Research* 68(5):1517–1537.

Higham, N. J. 2002. *Accuracy and Stability of Numerical Algorithms*. SIAM.

Huang, T., B. Wu, P. Lizardi, and H. Zhao. 2005. "Detection of DNA Copy Number Alterations Using Penalized Least Squares Regression". *Bioinformatics* 21(20):3811–3817.

Hui, F. K., S. Müller, and A. Welsh. 2017. "Joint Selection in Mixed Models Using Regularized PQL". *Journal of the American Statistical Association* 112(519):1323–1333.

Lee, W. 2011. "Tridiagonal Matrices: Thomas Algorithm". *MS6021, Scientific Computation, University of Limerick*.

Lorden, G. 1971. "Procedures for Reacting to a Change in Distribution". *The Annals of Mathematical Statistics* 42(6):1897–1908.

Ma, L., A. Grant, and G. Sofronov. 2020. "Multiple Change Point Detection and Validation in Autoregressive Time Series Data". *Statistical Papers* 61(4):1507–1528.

Moka, S., B. Liquet, H. Zhu, and S. Muller. 2024. "COMBSS: Best Subset Selection via Continuous Optimization". *Statistics and Computing* 34(2):75.

Müller, S. and A. H. Welsh. 2010. "On Model Selection Curves". *International Statistical Review* 78(2):240–256.

Natarajan, B. K. 1995. "Sparse Approximate Solutions to Linear Systems". *SIAM Journal on Computing* 24(2):227–234.

Niu, Y. S., N. Hao, and H. Zhang. 2016. "Multiple Change-Point Detection: A Selective Overview". *Statistical Science* 31(4):611–623.

Page, E. S. 1954. "Continuous Inspection Schemes". *Biometrika* 41(1/2):100–115.

Pilliat, E., A. Carpentier, and N. Verzelen. 2023. "Optimal Multiple Change-Point Detection for High-Dimensional Data". *Electronic Journal of Statistics* 17(1):1240–1315.

Qian, J. and J. Jia. 2016. "On Stepwise Pattern Recovery of the Fused Lasso". *Computational Statistics & Data Analysis* 94:221–237.

Richter, M. 2021. *Inverse Problems: Basics, Theory and Applications in Geophysics*. Springer Nature.

Rinaldo, A. 2009. "Properties and Refinements of the Fused Lasso". *The Annals of Statistics* 37(5B):2922–2952.

Sen, A. and M. S. Srivastava. 1975. "On Tests for Detecting Change in Mean". *The Annals of Statistics* 3(1):98–108.

Tian, Y. and Y. Takane. 2005. "Schur Complements and Banachiewicz-Schur Forms". *The Electronic Journal of Linear Algebra* 13:405–418.

Tibshirani, R. and P. Wang. 2008. "Spatial Smoothing and Hot Spot Detection for CGH Data Using the Fused Lasso". *Biostatistics* 9(1):18–29.

Truong, C., L. Oudre, and N. Vayatis. 2020. "Selective Review of Offline Change Point Detection Methods". *Signal Processing* 167:107299.

Verzelen, N., M. Fromont, M. Lerasle, and P. Reynaud-Bouret. 2023. "Optimal Change-Point Detection and Localization". *The Annals of Statistics* 51(4):1586–1610.

Wang, D., Y. Yu, and A. Rinaldo. 2020. "Univariate Mean Change Point Detection: Penalization, CUSUM and Optimality". *Electronic Journal of Statistics* 14:1917–1961.

Woodbury, M. A. 1950. *Inverting Modified Matrices*. Department of Statistics, Princeton University.

Zhu, J., C. Wen, J. Zhu, H. Zhang and X. Wang. 2020. "A Polynomial Algorithm for Best-Subset Selection Problem". *Proceedings of the National Academy of Sciences* 117(52):33117–33123.

AUTHOR BIOGRAPHIES

HANS REIMANN is a Master student in Mathematics at the University of Potsdam. He is currently on track to finish his degree in 2024 with a profile in data assimilation. His research interests include statistical learning such as in filtering and change point detection as well as the statistics of complex systems. His email address is hans.reimann@uni-potsdam.de.

SARAT MOKA is Lecturer at the School of Mathematics and Statistics at The University of New South Wales. He received his PhD in Applied Probability from the School of Technology and Computer Science at Tata Institute of Fundamental Research, Mumbai. His research interests encompass applied probability, computational statistics, machine learning, and deep learning. His email address is s.moka@unsw.edu.au and his personal website is https://www.saratmoka.com.

GEORGY SOFRONOV is an Associate Professor in the School of Mathematical and Physical Sciences at Macquarie University. He received the PhD degree in Probability Theory and Mathematical Statistics from Moscow State University in 2002. His research interests include Markov chain Monte Carlo simulation, Cross-Entropy method, change-point problem and optimal stopping rules. He serves on the editorial boards of Statistical Papers and Methodology and Computing in Applied Probability. His email address is georgy.sofronov@mq.edu.au and his website is https://researchers.mq.edu.au/en/persons/georgy-sofronov.