Supplementary Material for the Paper

Simultaneous Credible Regions for Multiple Changepoint Locations

Tobias Siems, Marc Hellmuth, Volkmar Liebscher
Department of Mathematics and Computer Science
University of Greifswald

December 11, 2017

1 Introduction

In Section 2 we provide a thorough treatment of the NP-completeness of the SBP and we give a proof of correctness of the ILP. Furthermore you can find two additional examples: In Section 3 we use credible regions to examine the robust Laplacian change in mean model on the Well-Log dataset from Fearnhead (2006). This dataset consists of outliers and small undesirable mean changes. We also compare different priors for the segment length: a geometric and a negative binomial distribution. Finally, we apply our credible regions to the changepoint model proposed in Fearnhead (2006). Section 4 examines a coal mining disasters dataset.

Attached to this supplement you can find:

- an R Package (called "SimCredRegR") to compute credible regions according to `Greedy`. This package also provides drawing routines and all the datasets and sampling algorithms that where used throughout the paper.

- an R Package (called "SimCredRegILPR") to compute credible regions according to the ILP. This package can only be installed if IBM's ILP solver CPLEX (IBM, 2016) is in place.

- the file “collection_of_different_data_simulations.pdf” which illustrates about 100 different credible regions, and joined confidence intervals from `stepR`. In each case the data was generated as in the Gaussian change in mean example, using the same mean values and changepoints.

- three video files “gauss_known_mean.mp4”, “gauss_known_var.mp4” and “laplace_known_var_r=5.mp4”. They demonstrate how credible regions evolve at different choices of the success rate in the prior for the segment length.

The supplement is available at: https://github.com/siemst/simcredreg .

2 Algorithmic view of the sample based problem

In this section we introduce the k minimum edge union problem (k-MINEU), which is shown to be equivalent to the SBP. We show the NP-completeness of the k-MINEU. In order to establish the theory, we introduce hypergraphs and some of its basic notions. Indeed, there is a close connection between hypergraphs and the statistical theory established in this research. As we shall see, a hypergraph can be defined in terms of a family of sequences of binary random variables with equal lengths and vice versa. We further provide an Integer Linear Program (ILP) formulation to k-MINEU that allows to compute exact solutions and we proof its correctness.

2.1 Hypergraphs

Here, we briefly discuss (multi-)hypergraphs and their structure and refer to Berge (1984); Voloshin (2009) for the interested reader. Before we start with the formal definitions, we recall that multisets are a natural generalization of usual sets (Syropoulos, 2001). Whereas a usual set contains each element only once, a multiset can contain each element arbitrary often. Therefore, a multiset over a set A is defined in terms of a mapping from A to $\mathbb{N}$, that assigns to each $a \in A$ the number of occurrences of a in the multiset.

Definition 1. *A* hypergraph $\mathscr{H}$ *is a pair* (V,h) *where* V *is a finite nonempty set and* h *is a multiset over* 2^V*, i.e.* $h : 2^V \to \mathbb{N}$.

Definition 2. *We say that $e \in 2^V$ is an* edge *of the hypergraph $\mathscr{H} = (V,h)$ iff $h(e) > 0$ and write $e \in \mathscr{H}$. Moreover, the* cardinality $\#\mathscr{H}$ *of $\mathscr{H}$ is given as the number of edges $\sum_{e\in\mathscr{H}} h(e)$. The* edge union *of a hypergraph $\mathscr{H}$ is defined as the set $\mathsf{U}(\mathscr{H}) := \bigcup_{e\in\mathscr{H}} e$.* Removing *an edge e of $\mathscr{H}$ is achieved by setting $h(e)$ to* $\max\{0, h(e)-1\}$*, whereas* completely removing *an edge e from $\mathscr{H}$ means to set $h(e) = 0$.*

Without loss of generality let $V = \{1,...,n\}$. Every edge $e \in \mathscr{H}$ can be represented by the unique sequence $(c_1,\dots,c_n) \in \{0,1\}^n$ with $\mathfrak{f}(c_1,\dots,c_n) = e$ and vice versa. With this in mind, we can treat random hypergraphs and families of sequences of binary random variables with identical lengths as equal. Therefore, a set of CP samples $s_1,\dots,s_m \in \{0,1\}^n$ represents a practical example of a hypergraph. There, the hypergraph is constructed through $V = \{1,\dots,n\}$ and $h(e) := \#\{i \mid e = \mathfrak{f}(s_i)\}$.

Definition 3. *A hypergraph $\mathscr{G} = (V,g)$ is a* sub-hypergraph *of a hypergraph $\mathscr{H} = (V,h)$ iff $g(e) \leq h(e)$ for all $e \in 2^V$ and we write $\mathscr{G}\subseteq\mathscr{H}$. Moreover, for a given vertex $x \in V$ and a hypergraph $\mathscr{H} = (V,h)$ we define the sub-hypergraph $\mathscr{D}(\mathscr{H},x) = (V,g)$ through*

$$g(e) = \begin{cases} h(e) & ,\textit{if } x \in e \\ 0 & ,\textit{otherwise} \end{cases}$$

$\mathscr{D}(\mathscr{H},x)$ represents the sub-hypergraph of $\mathscr{H}$ that consists of all edges that contain vertex x.

2.2 Computational complexity of k-MINEU and SBP

For a given hypergraph $\mathscr{H}$ and an integer k with $0 \leq k \leq \#\mathscr{H}$, we want to solve the problem of finding a sub-hypergraph $\mathscr{G} \subseteq \mathscr{H}$ that has at least k edges but an edge union of minimum cardinality. Equivalently, we want to find at least one element of the set $\operatorname*{argmin}_{\mathscr{G}\subseteq\mathscr{H}} \left\{ \#\mathsf{U}(\mathscr{G}) \,\middle|\, \#\mathscr{G} \geq k \right\}$ We refer to this task as the *k minimum edge union (optimization) problem* (k-MINEU).

Remark 1. *Provided that a solution $\mathscr{G} = (V,g)$ to a k-*MINEU *instance $(\mathscr{H},k)$ with $\mathscr{H} = (V,h)$ is known, one can easily determine further solutions $\mathscr{G}' = (V,g')$ by choosing for all edges $e \in \mathscr{H}$ an arbitrary integer ℓ_e with $g(e) \leq \ell_e \leq h(e)$ and setting*

$$g'(e) := \begin{cases} \ell_e & , \textit{ if } g(e) > 0 \\ 0 & , \textit{ otherwise} \end{cases}$$

Clearly, $\#\mathsf{U}(\mathscr{G}) = \#\mathsf{U}(\mathscr{G}')$ *and* $\#\mathscr{G}' \geq \#\mathscr{G} \geq k$.

Theorem 1. *The problems* k-MINEU *and SBP are equivalent.*

Before proving this theorem, we need to consider the following lemma.

Lemma 1. *Given a hypergraph* $\mathscr{H} = (\{1,\dots,n\},h)$ *and a sequence* $s_1,\dots,s_m \in \{0,1\}^n$ *with* $m = \#\mathscr{H}$ *and* $h(e) = \#\{i \mid e = \mathfrak{f}(s_i)\}$ *for all* $e \subseteq \{1,\dots,n\}$*, the following applies*

$$\min_{A\subseteq\{1,\dots,n\}} \left\{\#A \;\middle|\; \sum_{i=1}^{m} \mathbb{1}\langle \mathfrak{f}(s_i) \subseteq A\rangle \geq k\right\} = \min_{\mathscr{G}\subseteq\mathscr{H}} \left\{\#\mathsf{U}(\mathscr{G}) \;\middle|\; \#\mathscr{G} \geq k\right\} \tag{1}$$

Proof of Lemma 1. For $A \subseteq \{1,\dots,n\}$ with $\sum_{i=1}^{m} \mathbb{1}\langle \mathfrak{f}(s_i) \subseteq A\rangle \geq k$, we construct the hypergraph $\mathscr{G} = (\{1,\dots,n\},g)$ with $g(e) = \#\{i \mid e = \mathfrak{f}(s_i), \mathfrak{f}(s_i) \subseteq A\}$. Since $\#\mathscr{G} \geq k, \mathscr{G} \subseteq \mathscr{H}$ and $\#\mathsf{U}(\mathscr{G}) \leq \#A$ we conclude that $\geq$ holds in Equation (1).

Given a hypergraph $\mathscr{G} \subseteq \mathscr{H}$ with $\#\mathscr{G} \geq k$, we can chose an $I \subseteq \{1,\dots,m\}$ with $\#I \geq k$ and $\bigcup_{i\in I} \mathfrak{f}(s_i) \subseteq \mathsf{U}(\mathscr{G})$. Since $\sum_{i\in I} \mathbb{1}\langle \mathfrak{f}(s_i) \subseteq \mathsf{U}(\mathscr{G})\rangle \geq k$ we can also conclude that $\leq$ holds in Equation (1). □

Proof of Theorem 1. Let $(\mathscr{H},k)$ be a k-MINEU instance with $\mathscr{H} = (\{1,\dots,n\},h)$. We construct an equivalent SBP instance with $s_1,\dots,s_m \in \{0,1\}^n$ and $\alpha \in [0,1]$ such that a solution to this SBP instance provides a solution to the k-MINEU instance. Therefore, choose $m = \#\mathscr{H}$, $\alpha = 1 - \frac{k}{m}$ and $s_1,\dots,s_m$ so that $h(e) = \#\{i \mid e = \mathfrak{f}(s_i)\}$. If A is a solution to this SBP instance, then the hypergraph $\mathscr{G} = (\{1,\dots,n\},g)$ with $g(e) = \#\{i \mid e = \mathfrak{f}(s_i), \mathfrak{f}(s_i) \subseteq A\}$ is a solution to the k-MINEU instance $(\mathscr{H},k)$. Indeed, since $\#\mathsf{U}(\mathscr{G}) = \#A$ and $\#\mathscr{G} \geq k$, Lemma 1 implies that $\mathscr{G}$ is a solution to the k-MINEU instance.

Conversely, given an SBP instance with $\alpha \in [0,1]$ and $s_1,\dots,s_m \in \{0,1\}^n$, we construct an equivalent k-MINEU instance $(\mathscr{H},k)$ such that a solution to this k-MINEU instance provides a solution to the SBP instance. Therefore, let $k = \lceil m\cdot(1-\alpha)\rceil$ and $\mathscr{H} = (\{1,\dots,n\},h)$ with $h(e) = \#\{i \mid e = \mathfrak{f}(s_i)\}$. If $\mathscr{G}$ is a solution to this k-MINEU instance, then $\mathsf{U}(\mathscr{G})$ is a solution to the SBP instance. Indeed, since $\sum_{i=1}^{m} \mathbb{1}\langle \mathfrak{f}(s_i) \subseteq \mathsf{U}(\mathscr{G})\rangle \geq m\cdot(1-\alpha)$, Lemma 1 implies that $\mathsf{U}(\mathscr{G})$ is a solution to the SBP instance. □

We now show that (the decision version of) k-MINEU (and hence of SBP) is NP-complete (Garey and Johnson, 1979). Thus, there is no polynomial time algorithm to solve this problem, unless $P = NP$.

The decision version of k-MINEU is as follows:

Problem (Decision Version of k-MINEU).

Input: *A hypergraph $\mathscr{H} = (V,h)$ and integers k,l with $0 \leq k \leq \#\mathscr{H}$ and $0 < l \leq \#\mathsf{U}(\mathscr{H})$.*

Question: *Is there a sub-hypergraph $\mathscr{G} = (V,g) \subseteq \mathscr{H}$ such that $\#\mathscr{G} \geq k$ and $\#\mathsf{U}(\mathscr{G}) \leq l$?*

In order to prove the NP-completeness of k-MINEU, we use the well-known NP-complete KNAPSACK-problem (Karp, 1972; Garey and Johnson, 1979).

Problem (KNAPSACK).

Input: *A finite set U, for each $u \in U$ a weight $w(u) \in \mathbb{N}$ and a value $v(u) \in \mathbb{N}$, and positive integers a and b.*

Question: *Is there a subset $U' \subseteq U$ such that $\sum_{u \in U'} w(u) \leq b$ and $\sum_{u \in U'} v(u) \geq a$?*

Theorem 2. k-MINEU *is NP-complete.*

Proof. We begin with showing that k-MINEU $\in \mathsf{NP}$. To this end, it suffices to demonstrate that a candidate solution to k-MINEU can be verified in polynomial time. However, this is easy to see, since we only need to check whether for a possible solution $\mathscr{G} \subseteq \mathscr{H}$ it holds that $\#\mathsf{U}(\mathscr{G}) = \sum_{e \in \mathscr{G}} g(e) \leq l$ and $\#\mathscr{G} = \sum_{e \in \mathscr{G}} 1 \geq k$. Both tasks can be done in linear time in the number of edges of $\mathscr{G}$.

We proceed to show by reduction from KNAPSACK that k-MINEU is NP-hard. Thus, let us assume we are given an arbitrary instance of KNAPSACK, that is, a finite set U, for each $u \in U$ the weight $w(u) \in \mathbb{N}$ and the value $v(u) \in \mathbb{N}$, as well as positive integers b and a. Now, we construct an instance of k-MINEU as follows. For each $u \in U$ we set an edge $e_u := \{(u,1), \ldots, (u,v(u))\}$ and $h(e_u) = w(u)$. The vertex set of the hypergraph $\mathscr{H} = (V,h)$ is then $V = \cup_{u \in U} e_u$. Note, the edges in $\mathscr{H}$ are pairwise disjoint. Clearly, this reduction can be done in polynomial time in the number of elements in U and the values $v(u)$.

In what follows, we show that KNAPSACK has a solution for given integers b,a if and only if k-MINEU has a solution with $k = \#\mathscr{H} - b$ and $l = \#\mathsf{U}(\mathscr{H}) - a$.

Let $U' = \{u_1, \ldots, u_n\} \subseteq U$ such that $\sum_{i=1}^{n} w(u_i) \leq b$ and $\sum_{i=1}^{n} v(u_i) \geq a$. Completely remove all corresponding edges e_{u_i}, $1 \leq i \leq n$ from $\mathscr{H}$ to obtain the sub-hypergraph $\mathscr{G}$. Hence, $\#\mathscr{G} = \#\mathscr{H} - \sum_{i=1}^{n} w(u_i) \geq \#\mathscr{H} - b = k$ and $\#\mathsf{U}(\mathscr{G}) = \#\mathsf{U}(\mathscr{H}) - \sum_{i=1}^{n} v(u_i) \leq \#\mathsf{U}(\mathscr{H}) - a = l$.

Conversely, assume that $\mathscr{G} = (V,g)$ is a valid solution for the hypergraph $\mathscr{H} = (V,h)$ (as constructed above) and given integers $k \geq \#\mathscr{H}$ and $l \leq \#\mathsf{U}(\mathscr{H})$. Thus, we can write $k = \#\mathscr{H} - b$

and $l = \#\mathsf{U}(\mathscr{H}) - a$. Hence, $\#\mathscr{G} \geq \#\mathscr{H} - b$ and $\#\mathsf{U}(\mathscr{G}) \leq \#\mathsf{U}(\mathscr{H}) - a$. Therefore, at least b edges must have been removed from $\mathscr{H}$ resulting in $\mathscr{G}$ where the edge union of $\mathscr{G}$ has at least a fewer vertices than $\mathsf{U}(\mathscr{H})$. Note, to obtain fewer vertices in $\mathsf{U}(\mathscr{H})$ one needs to completely remove edges from $\mathscr{H}$. Let $E_0 = \{e \in \mathscr{H} \mid g(e) = 0\}$ be the set of all edges that have been completely removed from $\mathscr{H}$. By construction, each edge e_u is uniquely identified with an element $u \in U$. We show that $U' = \{u \in U \mid e_u \in E_0\}$ provides a solution for SBP. To this end, observe that $\#\mathscr{H} - \sum_{e \in E_0} h(e) \geq \#\mathscr{G} \geq \#\mathscr{H} - b$ which implies that $\sum_{e \in E_0} h(e) = \sum_{u \in U'} w(u) \leq b$, as desired. Moreover, by construction we have $\#\mathsf{U}(\mathscr{G}) = \#\mathsf{U}(\mathscr{H}) - \sum_{e \in E_0} \#e = \#\mathsf{U}(\mathscr{H}) - \sum_{u \in U'} v(u) \leq \#\mathsf{U}(\mathscr{H}) - a$. Thus, $\sum_{u \in U'} v(u) \geq a$, which completes the proof. □

Combining Theorem 1 and 2 we obtain the following

Corollary 1. *The decision version of SBP is an NP-complete problem.*

2.3 ILP formulation

Since k-MINEU is NP-complete, we cannot hope for polynomial-time algorithms that solve the corresponding optimization problem. Nevertheless, we show here that k-MINEU is in some cases tractable in practice by formulating it as an Integer Linear Program (ILP).

We introduce for a hypergraph $\mathscr{H} = (V, h)$ and an integer k the following binary variables $U_x, F_e \in \{0, 1\}$:

$U_x = 1$ if and only if vertex x of $\mathscr{H}$ is contained in the edge union $\mathsf{U}(\mathscr{G})$ of $\mathscr{G} \subseteq \mathscr{H}$.

$F_e = 1$ if and only if the edge $e \in \mathscr{H}$ is contained in $\mathscr{G}$.

Moreover, the number of edges that contain a vertex x is given by the constants $D_x := \#\mathscr{D}(\mathscr{H}, x)$ for all $x \in V$.

To find a solution for the k-MINEU problem, we need to minimize the number of vertices in the edge union of $\mathscr{G} = (V, g) \subseteq \mathscr{H}$, which is achieved by minimizing the objective function

$$\sum_{x \in \mathsf{U}(\mathscr{H})} U_x \tag{2}$$

By Remark 1 it is always possible to find a sub-hypergraph $\mathscr{G}$ with minimum edge union such that $g(e) = h(e)$ for all $e \in \mathscr{G}$. We will construct such a sub-hypergraph.

To ensure that $\#\mathscr{G} \geq k$ we add the constraint

$$\sum_{e \in \mathscr{H}} h(e) \cdot F_e \geq k \tag{3}$$

Note, e is contained in $\mathscr{G}$ if and only if $F_e = 1$ and by construction, $\#\mathscr{G} = \sum_{e \in \mathscr{H}} (h(e) \cdot F_e)$. Thus, constraint (3) is satisfied if and only if $\#\mathscr{G} \geq k$.

Finally, we have to ensure that $U_x = 1$ if and only if there is an edge in $\mathscr{G}$ that contains x. To this end, we add for all $x \in \mathsf{U}(\mathscr{H})$ the constraint

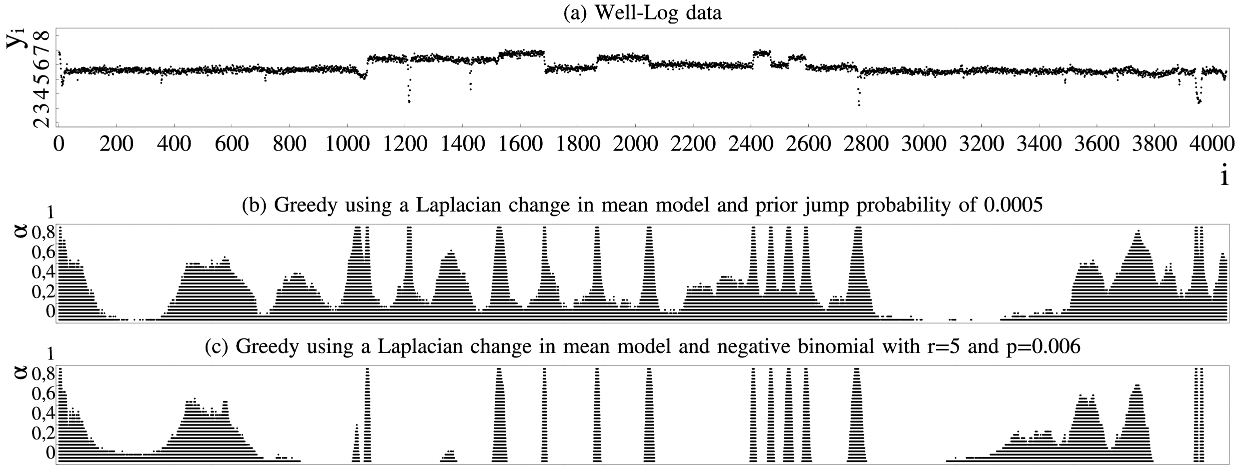
$$\sum_{e \in \mathscr{D}(\mathscr{H},x)} h(e) \cdot (1 - F_e) \geq D_x \cdot (1 - U_x) \tag{4}$$

Now, if there is no edge containing x in $\mathscr{G}$ and thus, $F_e = 0$ for all $e \in \mathscr{D}(\mathscr{H},x)$, then Constraint (4) implies that $\sum_{e \in \mathscr{D}(\mathscr{H},x)} h(e) \geq D_x(1 - U_x)$. Since $\sum_{e \in \mathscr{D}(\mathscr{H},x)} h(e) = D_x$, we have two choices for $U_x \in \{0,1\}$. However, the optimization function ensures that U_x is set to 0. Conversely, assume that there is an edge e that contains x and hence, $F_e = 1$. Thus, $\sum_{e \in \mathscr{D}(\mathscr{H},x)} h(e) \cdot (1 - F_e) < D_x$. The only way to satisfy Constraint (4) is achieved by setting $U_x = 1$.

Taken together the latter arguments we can infer the following result.

Theorem 3. *The ILP formulation in Eqs.* (2) - (4) *correctly solves the k-*MINEU *problem.*

3 Well-Log Data



(a) Well-Log data

(b) Greedy using a Laplacian change in mean model and prior jump probability of 0.0005

(c) Greedy using a Laplacian change in mean model and negative binomial with r=5 and p=0.006

Figure 1: (a) Well-Log data. (b) and (c) Credible regions inferred by `Greedy`.

Now we consider a well-log dataset (see Figure 1(a)) that stems from nuclear-magnetic response of underground rocks (Fearnhead, 2006). We model the data as Laplacian distributed with changing means. The segment lengths are geometrically distributed with parameter $p = 0.0005$ and the distribution of a jump is a Laplacian with mean 0 and a variance of 8. Since the data in (a) includes a lot of outliers, we decided to use the Laplace distribution because it is more robust than the Gaussian. Furthermore, inference in this model is still tractable, albeit computationally more demanding. Another difficulty with this data is the fact, that there are a lot small changes in mean, which should not be recognized as CP's.

The regions inferred by `Greedy` in (b) show that this model expresses a good sensitivity towards the desired CP's. At the same time it tends to ignore the outliers very well. Furthermore, this model is able to distinguish small mean changes from bigger ones, even on longer segments. However, the model still infers CP's at around 1200 which are the result of a small cluster of outliers. In order to avoid these kind of CP's we use a negative binomial distribution with $r = 5$ and $p = 0.006$ as the distribution for the time between two successive CP's. The result of this model choice is shown in (c). We see that the negative binomial helps the model to recognize small clusters of outliers.

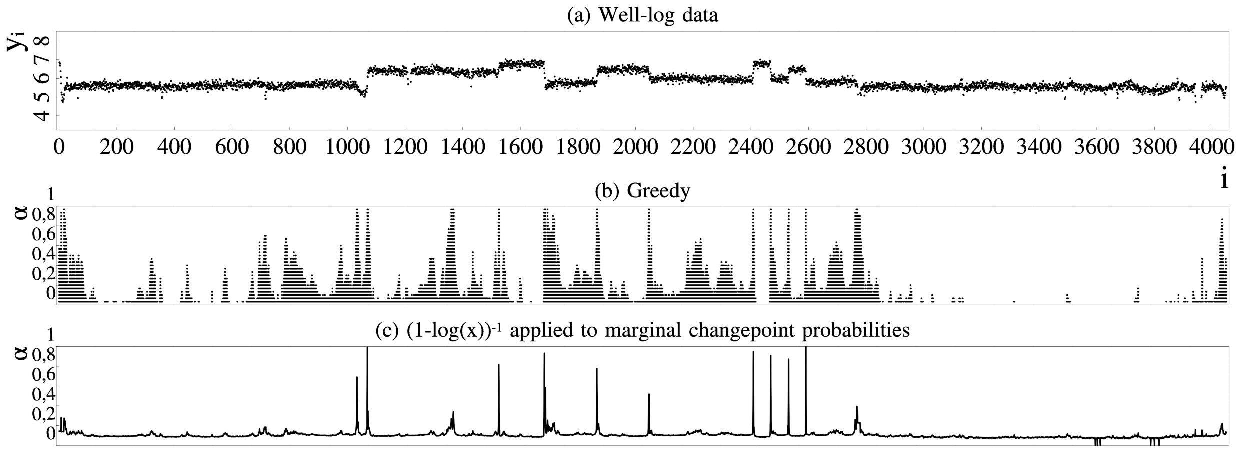


Figure 2: (a) Well-Log Data with outliers removed (b) Credible regions inferred by `Greedy`. (c) Posterior marginal CP probabilities transformed through $\left(1 - log(x)\right)^{-1}$.

Fearnhead (2006) considers the same dataset. He removes outliers in a preprocessing step (see Figure 2(a)) and fits a state space model for the segments in order to allow small changes in mean between successive CP's. (b) illustrates regions inferred by `Greedy`. We see that this model infers

CP's that stem from larger changes in mean, but ignores smaller ones. In contrast to the Laplace distribution it is able to disregard the dip at around 3800.

Marginal CP probabilities are used in order to evaluate the model choice and to adjust model parameters. By looking at the importance read from the credible regions at around 1300 and 1400, we see that there is a strong indication for a CP. As a matter of fact, 100% of the samples contain a CP there. In contrast, the marginal CP probabilities in (c) are not able to express the importance of this CP. This might be an undesirable CP that could have been avoided if credible regions had been used in order to evaluate the parameter choice.

4 Geometric change in success rate model

Now we examine another real world dataset consisting of 191 successive timepoints of coal mine explosions, that killed ten or more men between March 15, 1851 and March 22, 1962 (Trenkler, 1995). Figure 3(a) displays the data. The number of explosion can be read from the abscissa and the cumulated days up to the corresponding explosion can be read from the axis of ordinate. The black dot marks the Coal Mines Regulations Act in 1887. The picture shows that by means of the regulations act the time from one explosion to the next slightly increases.

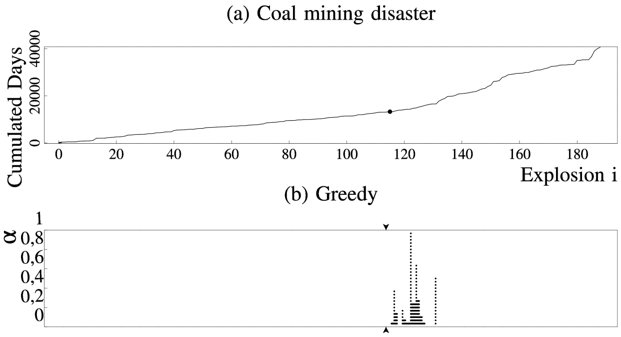


Figure 3: (a) 191 successive coal mine explosions. (b) Regions inferred by `Greedy`.

Unlike but not entirely different from Adams and MacKay (2007), we model the days from one accident to another as geometrically distributed in order to infer changes in the success rate. We pin the model down down to have 1 CP. In case of a jump, the success rates change their value according to a uniform distribution over $[0,1]$. The credible regions in (b) already show evidence for a CP directly after the regulations act took place.

References

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

Berge, C. (1984). *Hypergraphs: Combinatorics of Finite Sets*. North-Holland Mathematical Library. Elsevier Science.

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

Garey, M. R. and D. S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co.

IBM (2004–2016). IBM ILOG CPLEX C++ Optimizer. `http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/`.

Karp, R. M. (1972). Reducibility among combinatorial problems. In R. E. Miller, J. W. Thatcher, and J. D. Bohlinger (Eds.), *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations, held March 20–22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, and sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department*, Boston, MA, pp. 85–103. Springer US.

Syropoulos, A. (2001). *Mathematics of Multisets*, Chapter 4, pp. 347–358. Springer Berlin Heidelberg.

Trenkler, D. (1995). A handbook of small data sets : Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. & Ostrowski, E. (1994): Chapman & Hall, London. xvi + 458 pages, including one diskette with data files (MS-DOS), 40 Br. *Computational Statistics & Data Analysis*.

Voloshin, V. (2009). *Introduction to Graph and Hypergraph Theory*. Nova Science Publishers.