Changepoint Detection in the Presence of Outliers

Paul Fearnhead[1,†] and Guillem Rigaill[2,3]

[1]Department of Mathematics and Statistics, Lancaster University

[2]Institute of Plant Sciences Paris-Saclay, UMR 9213/UMR1403, CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris-Diderot, Sorbonne Paris-Cité

[3]Laboratoire de Mathématiques at Modélisation d'Evry (LaMME), Université d'Evry Val d'Essonne, UMR CNRS 8071, ENSIIE, USC INRA

[†]Correspondence: p.fearnhead@lancaster.ac.uk

Abstract

Many traditional methods for identifying changepoints can struggle in the presence of outliers, or when the noise is heavy-tailed. Often they will infer additional changepoints in order to fit the outliers. To overcome this problem, data often needs to be pre-processed to remove outliers, though this is difficult for applications where the data needs to be analysed online. We present an approach to changepoint detection that is robust to the presence of outliers. The idea is to adapt existing penalised cost approaches for detecting changes so that they use loss functions that are less sensitive to outliers. We argue that loss functions that are bounded, such as the

classical biweight loss, are particularly suitable – as we show that only bounded loss functions are robust to arbitrarily extreme outliers. We present an efficient dynamic programming algorithm that can find the optimal segmentation under our penalised cost criteria. Importantly, this algorithm can be used in settings where the data needs to be analysed online. We show that we can consistently estimate the number of changepoints, and accurately estimate their locations, using the biweight loss function. We demonstrate the usefulness of our approach for applications such as analysing well-log data, detecting copy number variation, and detecting tampering of wireless devices.

Keywords: Binary Segmentation, Biweight loss, Cusum, M-estimation, Penalised likelihood, Robust Statistics

1 Introduction

Changepoint detection has been identified as one of the major challenges for modern, big data applications (National Research Council, 2013). The problem arises when analysing data that can be ordered, for example time-series or genomics data where observations are ordered by time or position on a chromosome respectively. Changepoint detection refers to locating points in time or position where some aspect of the data of interest, such as location, scale or distribution, changes. There has been a recent explosion in methods for detecting changes (e.g. Frick et al., 2014; Fryzlewicz, 2014; Cao and Wu, 2015; Haynes et al., 2017b; Ma and Yau, 2016, and references therein) in recent years, in part motivated by the range of applications for which changepoint detection is important. Exemplar areas of application include bioinformatics (Olshen et al., 2004; Futschik et al., 2014), ion channels (Hotz et al., 2013), climate records (Reeves et al., 2007), oceonagraphic data (Killick et al., 2010, 2012) and finance (Kim et al., 2005).

What has received less attention is the problem of distinguishing between changepoints and outliers. To give an example of the issue outliers can cause when attempting to detect changepoint, consider the problem of detecting changes in well-log data. An example of such data, taken originally from Ó Ruanaidh and Fitzgerald (1996), is shown in Figure 1. This data was collected from a probe being lowered into a borehole. As it is lowered the probe takes measurements of the rock that it is passing through. As the probe moves from one type of rock strata to another, there is an abrupt change in the measurements. It is these changes in rock strata that we wish to detect. The real motivation for collecting this data was to detect these changes in real-time. This would enable changes in rock strata that are being drilled through to be quickly detected, so that appropriate changes to the settings of the drill can be made.

The data in the top-left plot in Figure 1 has been analysed by many different change detection methods (e.g. Ó Ruanaidh and Fitzgerald, 1996; Fearnhead, 2006; Adams and MacKay, 2007; Wyse et al., 2011; Ruggieri and Antonellis, 2016). However, this plot actually shows data that has been pre-processed to remove outliers. The real data that was collected by the probe is shown in the top-right plot of Figure 1. There are a number of short periods of time where the probe mis-functions, and very low measurements are recorded. These are examples of what we are calling outliers. The real challenge with detecting the changes is to distinguish between actual changes and these outliers. Most existing methods for changepoint detection are unable to do so; hence the reason that most analysis of this data has used the "cleaned" data set in the top-left plot. For example in the bottom row of Figure 1 we show the results of estimating the changepoints based on minimising a square-error-loss criteria with a penalty for each detected changepoint. Whilst this method performs well when analysing the cleaned dataset, it is unable to distinguish between changes and outliers when analysing the real data.

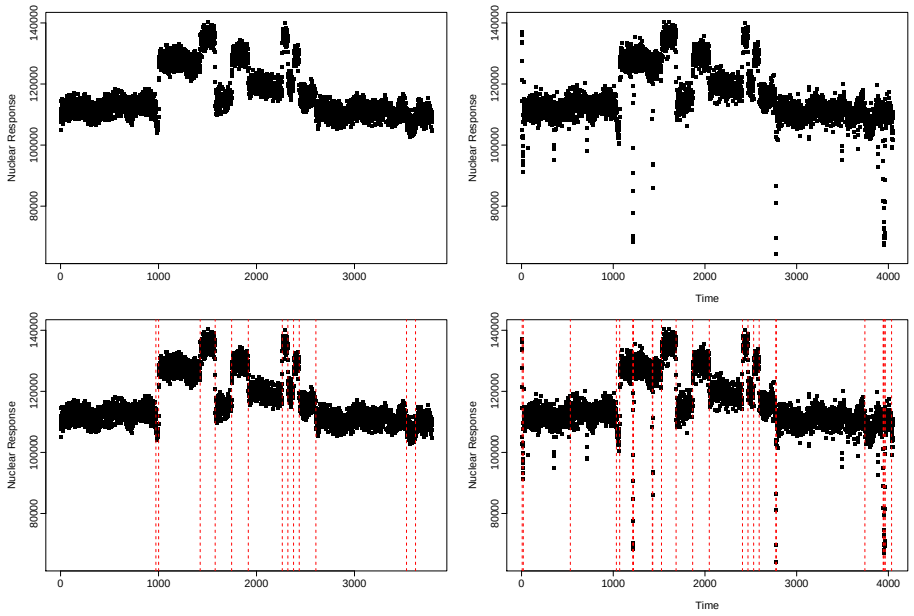

Figure 1: Well-log data: data with outliers removed (left column) and with outliers included (right column). Bottom row shows segmentations of the data under a least squares loss.

This lack of robustness for detecting a change in mean in the presence of outliers for many changepoint methods stems from explicit or implicit modelling assumptions of Gaussian noise. For example, methods based on a likelihood-ratio test for detecting a change (Worsley, 1979), or that use a penalised likelihood approach to detect multiple changes (Killick et al., 2012), or that do a Bayesian analysis (Yao, 1984), may be based on a Gaussian likelihood and thus explicitly assume Gaussian noise. Alternative methods, such as using an L_2 (square error) loss or a cusum based approach (Page, 1954), may not make such an assumption explicitly. However the resulting method are closely related to those based on a Gaussian likelihood (see e.g. Hinkley, 1971), and thus are implicitly making similar assumptions. Whilst these methods show some robustness to heavier-tailed noise (Lavielle and Moulines, 2000), in practice they can seriously over-estimate the number of changes in the presence of outliers.

Our approach is based on ideas from robust statistics, namely replacing an L_2 loss

with an alternative loss function that is less sensitive to outliers. We then use such a loss function within a penalised cost approach to estimating multiple changepoints. The use of alternative loss functions as a way to make changepoint detection robust to ouliers has been considered before (e.g. Hušková and Sen, 1989; Hušková, 1991; Hušková and Picek, 2005; Hušková, 2013). That work derives cusum-like tests for a single changepoint. Such a test for a single changepoint can then be used with binary segmentation to find multiple changes. As we discuss more fully in Section 2.3, this approach suffers from the draw-back that the test statistic is based upon how well the data can be modelled as not having a change, and does not directly compare this with how well we can fit the data with one or more changepoints. Thus it could spuriously infer a change if we have a cluster of outliers at consecutive time-points, even if the value of those outliers are not consistent with them coming from the same distribution.

One challenge with the penalised cost approach that we suggest is minimising this cost, which we need to do to infer the changepoints. We show how recent efficient dynamic programming algorithms (Maidstone et al., 2017; Rigaill, 2015) can be adapted to solve this minimisation problem. Our algorithm can use any loss function provided we are interested in the change of univariate parameter, such as the location parameter for univariate data, and the loss function is piecewise quadratic. Importantly these algorithms are sequential in nature, and thus can be directly applied in situations which need an online analysis of the data.

Whilst our approach can be used with a range of loss functions, we particularly recommend using a loss function that is bounded. We present a theoretical result that shows that we need a bounded loss function if we wish our method to be robust to any single outlier. The simplest such loss function is the biweight loss (Huber, 2011) which is the pointwise minimum of an L_2 loss and a constant. We show that, under mild conditions, we can consistently estimate the number of changepoints,

and accurately estimate their locations, if we use a penalised cost approach with the biweight loss.

To illustrate the usefulness of our approach, with the biweight loss, in practice, we present its use for three distinct applications. The first is for the online analysis of the well-log data of Figure 1. Secondly we show that it out-performs existing methods for detecting copy number variation. This includes performing better than methods that pre-process the data in an attempt to remove outliers. By comparison our approach is easier to implement as it does not require any pre-processing steps. Finally we consider the problem of detecting tampering of wireless security devices. Results here show our method can reliably distinguish between actual tampering events and changes in the data caused by short-term environmental factors.

Proofs of results are given in the Appendices in the supplementary material. Code implementing the new methods in this paper is available from `https://github.com/guillemr/robust-fpop`.

2 Model Definition

Assume we have data ordered by some covariate, such as time or position along a chromosome. Denote the data by $\mathbf{y} = (y_1, \ldots, y_n)$. We will use the notation that, for $s \leq t$, the set of observations from time s to time t is $\mathbf{y}_{s:t} = (y_s, ..., y_t)$. If we assume that there are k changepoints in the data, this will correspond to the data being split into $k+1$ distinct segments. We let the location of the jth changepoint be τ_j for $j = 1, \ldots, k$, and set $\tau_0 = 0$ and $\tau_{k+1} = n$. The jth segment will consist of data points $y_{\tau_{j-1}+1}, \ldots, y_{\tau_j}$. We let $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_{k+1})$ be the set of changepoints.

The statistical problem we are considering is how to infer both the number of changepoints and their locations. We assume the changepoints correspond to abrupt changes

in the location, that is mean, median or other quantile, of the data. We will focus on a minimum penalised cost approach to the problem. This approach encompasses penalised likelihood approaches to changepoint detection amongst others.

To define our penalised cost, we first introduce a loss function for a single observation, y, and a segment-specific location parameter θ. We denote this as $\gamma(y;\theta)$. For a penalised likelihood approach this loss would be equal to minus the log-likelihood. The class of losses we will consider are discussed below.

We can now define the cost associated with a segment of data, $y_{s:t}$. This is

$$\mathcal{C}(y_{s:t}) = \min_{\theta} \sum_{i=s}^{t} \gamma(y_i;\theta),$$

the minimum, over the segment-specific parameter θ, of the sum of the losses associated with each observation in the segment. The penalised cost for a segmentation is then

$$Q(y_{1:n};\tau_{1:k}) = \sum_{i=0}^{k} \left\{\mathcal{C}(y_{\tau_i+1:\tau_{i+1}}) + \beta\right\}, \tag{1}$$

where $\beta > 0$ is a chosen constant that penalises the introduction of changepoints. We estimate the number and position of the changepoints by the value of k and $\tau_{1:k}$ that minimise this penalised cost. The value of β has a substantial impact on the number of changepoints that are estimated (see Haynes et al., 2017a, for examples of this), with larger values of β leading to fewer estimated changepoints.

For inferring changes in the mean of the data, it is common to use the squared-error loss function (e.g. Yao and Au, 1989; Lavielle and Moulines, 2000)

$$\gamma(y;\theta) = (y-\theta)^2.$$

In this case, the penalised cost approach corresponds to a penalised likelihood approach where the data within a segment are IID Gaussian with common variance. Minimising a penalised cost of this form is closely related to binary segmentation procedures based on cusum statistics (e.g. Vostrikova, 1981; Bai, 1997; Fryzlewicz,

2014), as discussed in Killick et al. (2012). Use of the square-error loss function results in an approach that is very sensitive to outliers. For example, this loss function was the one used in the analysis of the well-log data in Figure 1, where we saw that it struggles to distinguish outliers from actual changes of interest.

2.1 Penalised Costs based on M-estimation

To develop a changepoint approach that can reliably detect changepoints in the presence of outliers we need a loss function that increases at a slower rate in $|y - \theta|$. Standard examples (Huber, 2011) are absolute error, $\gamma(y;\theta) = |y - \theta|$, Huber loss

$$\gamma(y;\theta) = \begin{cases} (y-\theta)^2 & \text{if } |y-\theta| < K, \\ 2K|y-\theta| - K^2 & \text{otherwise,} \end{cases}$$

and the biweight loss,

$$\gamma(y;\theta) = \begin{cases} (y-\theta)^2 & \text{if } |y-\theta| < K, \\ K^2 & \text{otherwise,} \end{cases} \tag{2}$$

or if interest lies in changes in the uth quantile for $0 < u < 1$,

$$\gamma(y;\theta) = \begin{cases} 2u(y-\theta) & \text{if } y > \theta, \\ 2(1-u)(\theta-y) & \text{otherwise.} \end{cases}$$

These are summarised in Figure 2.

We will develop an algorithm for finding the best segmentation under a penalised cost criteria that can deal with any of these choices for the loss. In practice we particularly advocate the use of the biweight loss. For a penalised cost approach to detecting changepoints to be robust to extreme outliers we will need the loss function to be bounded. For unbounded loss functions, such as the absolute error loss or Huber loss, a penalised cost approach will place an outlier in a segment on its own if that outlier is sufficiently extreme. This is shown by the following result.

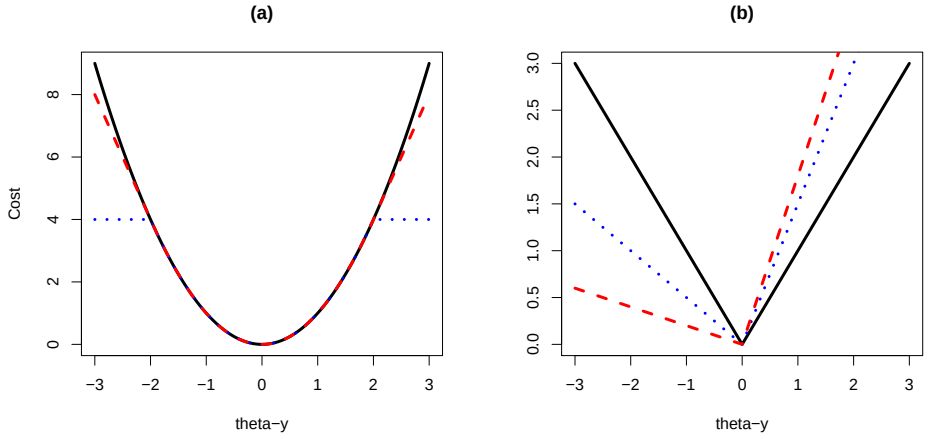

Figure 2: Example of different losses. (a) The square error loss (full-line), and the related Huber loss (red dashed) and biweight loss (blue dotted). (b) The absolute error loss (full-line), and its generalisation for detecting change in quantiles for $u = 0.1$ (red dashed) and $u = 0.25$ (blue dotted).

Theorem 2.1 *Assume that the loss function satisfies* $\gamma(y;\theta) = g(|y-\theta|)$ *where* $g(0) = 0$ *and* $g(\cdot)$ *is an unbounded, increasing function. Choose any* $t \in \{1,\ldots,n\}$ *and fix the set of other observations,* y_s *for* $s \neq t$*. Then there exists values of* y_t *such that the segmentation that minimises the penalised cost (1) will have changepoints at* $t-1$ *and* t*.*

If we choose a loss function, such as the biweight loss, that is bounded, then this will impose a minimum segment length on the segmentations that we infer using the penalised cost function. Providing this minimum segment length is greater than 1, our inference procedure will be robust to the presence of extreme outliers – unless these outliers cluster at similar values, and for a number of consecutive time-points greater than our minimum segment length.

Theorem 2.2 *If the loss function satisfies* $0 \leq \gamma(y;\theta) \leq K$*, and we infer changepoints by minimising the penalised cost function (1) with penalty* β *for adding a changepoint, then all inferred segments will be of length greater than* β/K*.*

The other conclusion to draw from this result is that, for any choice of K and β, we would want the minimum segment length to be smaller than any segment we expect, or that we wish to detect, in the data. Any real segments shorter than the minimum segment length are unlikely to be detected, with the observations in such short segments being identified as outliers instead. Furthermore our procedure can lose power to detect real segments that are only slightly longer than the minimum segment length (see empirical results for scenario 4 in Section 5.2).

2.2 Consistency under Biweight Loss

As mentioned above, and as suggested by Theorem 2.1, a particular focus will be on the use of the biweight loss (2). Here we give conditions under which we can consistently estimate the number and location of changepoints when using this loss.

We will consider the standard in-fill asymptotics, as we let the number of data points, n, increase. To be able to consistently estimate the number of changepoints we will need the penalty for adding a changepoint, β in (1), to increase with n. We will thus denote the choice of penalty for a given number of data points to be β_n.

Data Generating Model: We assume a fixed number of changepoints, k_0, and fixed constants $0 < u_1 < \cdots < u_{k_0} < 1$ so that for a data set of size n we have the ith changepoint at $\tau_i = \lfloor nu_i \rfloor$, for $i = 1, \ldots, k_0$. As above we let $\tau_0 = 0$ and $\tau_{k_0+1} = n$. We further assume fixed segment-specific location parameters, $\mu_0, \ldots, \mu_{k_0}$, with the obvious constraint that $\mu_i \neq \mu_{i-1}$ for $i = 1, \ldots, k_0$. Finally we let $Z_1, Z_2, \ldots$ be IID noise random variables, so that for $t = 1, \ldots, n$ the observations are realisations of

$$Y_t = \mu_i + Z_t,$$

where i is such that $\tau_i < t \leq \tau_{i+1}$.

Our results require two mild conditions on the distribution of the noise random vari-

ables. Firstly introduce the mean of the loss function, $M(\theta) = \mathrm{E}\{\gamma(Z_i;\theta)\}$. We assume $M(\theta)$ takes it minimum value at $\theta = 0$. We can make this assumption without loss of generality, as if $M(\theta)$ has its minimum at θ^* we can just re-parameterise our model with new noise random variables set to $Z_i - \theta^*$ and with location parameters re-defined to be $\mu_i + \theta^*$.

Condition 1: Our first condition is that there exists constants $c_1 > 0$ and $c_2 > 0$ such that

$$M(\theta) = \mathrm{E}\left[\min\left\{(Z_i - \theta)^2, K^2\right\}\right] \geq M(0) + \min\left\{c_1\theta^2, c_2\right\}. \tag{3}$$

This is a weak assumption, and will hold if $M(\theta)$ has a positive second derivative for all θ in a neighbourhood around 0 and that $M(\theta) - M(0) \geq c_2 > 0$ for all θ outside this region. The latter requirement is a common assumption made to ensure identifiability of estimates of a location parameter when using a given loss function.

Condition 2: Our second condition is slightly stronger. Let $p = \Pr(|Z_i| > K)$ and $\sigma^2 = \mathrm{E}(Z_i^2 \mid |Z_i| \leq K)$, then we need

$$K^2(1 - 2p) - (1 - p)\sigma^2 > 0. \tag{4}$$

This condition can be achieved by taking K large enough. If the noise has finite variance then, using Chebyshev's inequality, it is easy to show that any choice with $K > \sqrt{3}\mathrm{E}(Z_i^2)$ will ensure this condition holds. However we do not need the noise to have a variance. For example it is sufficient to choose $K > \sqrt{3}\mathrm{E}(\min\{Z_i^2, K^2\})$, or, if Z_i has a unimodal density function with mode at 0, then $\sigma^2 \leq K^2/3$ and it suffices to choose K so that $p = \Pr(|Z_i| > K) < 2/5$. By comparison, we would recommend taking K sufficiently large that $|Z_i| > K$ is relatively rare, and thus $p \approx 0$. In line with Theorem 2.1, this condition does not depend on the distribution of the noise conditional on $|Z_i| > K$.

Theorem 2.3 *Consider the data generating model described above, and suppose conditions 1 and 2 hold. For a given n let $\hat{k}_n$ be the estimate of the number of change-*

points, and $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{k}_n}$ *their estimated locations, obtained by minimising the penalised cost (1) using the biweight loss function and a penalty* β_n*. Then there exists constants* $C_1 > 0$ *and* $C_2 > 0$ *such that*

$$\Pr\left[\hat{k}_n = k_0 \text{ and } \max_{i=1,\ldots,k_0}\left\{\min_{j=1,\ldots,\hat{k}_n} |\tau_i - \hat{\tau}_j|\right\} \leq C_2 \log(n)\right] \to 1, \text{ as } n \to \infty,$$

provided that $C_1 \log(n) < \beta_n = o(n)$*.*

The theorem shows that for an appropriate choice of β_n we can obtain a consistent estimate of the true number of parameters, and that the error in estimating any of the changepoint locations will be less than $C_2 \log(n)$ with probability 1. The latter order of error is in line with asymptotic results for the accuracy of changepoint estimates using wild binary segmentation with the cusum test (Fryzlewicz, 2014). We require much weaker conditions on the distribution of the noise, but our result assumes stronger conditions on the number of changes, the segment lengths and the size of change of mean at each changepoint than, for example, results in Fryzlewicz (2014) and Baranowski et al. (2016). The result supports the use of a penalty, β_n, that is proportional to $\log(n)$, a choice that is common for other penalised cost procedures, but it does not specify the constant of proportionality.

2.3 Alternative Robust Changepoint Methods

There have been other proposed M-procedures for robust detection of changepoints (Hušková and Sen, 1989; Hušková, 1991; Hušková and Picek, 2005; Hušková, 2013). These differ from our approach in that they are based on sequentially applying tests for single changepoints. One approach is to use a Wald-type test. For a convex loss function $\gamma(y;\theta)$ which depends only on $y - \theta$, define $\gamma(y;\theta) = \rho(y - \theta)$ and define ϕ to be the first derivative of ρ. Then we can estimate a common θ for data $y_{1:n}$ by

minimising

$$\sum_{i=1}^{n} \rho(y_i - \theta),$$

with respect to θ. In many cases this is equivalent to solving

$$\sum_{i=1}^{n} \phi(y_i - \theta) = 0.$$

If $\hat{\theta}$ denotes the estimate we obtain, we can define residuals as $\phi(y_i - \hat{\theta})$, and their partial sums, or cusums, by

$$S_m = \sum_{i=1}^{m} \phi(y_i - \hat{\theta}).$$

A Wald-type test is then based on a test-statistic of the form

$$T_n = \max_{1 \leq m \leq n-1} \frac{n}{m(n-m)} S_m^2,$$

where the term $n/(m(n-m))$ is introduced so that the variability of the term on the right-hand side will be similar for each value of m. Large values of T_n are taken as evidence for a change. The position of a changepoint is then inferred at the position m which maximises the right-hand side. To detect multiple changepoints, this Wald-type test needs is currently used within a binary segmentation procedure; though it can also be used with improved versions of binary segmentation, such as wild binary segmentation (Fryzlewicz, 2014).

There are two main differences between the Wald-type test approach and our penalised cost approach. The first is that the Wald-type test statistic is appropriate only for convex loss functions. So, for example, the biweight loss is not appropriate for use with this approach. To see this note that the derivative of the biweight loss satisfies $\phi(x) = 0$ for $|x| > K$. Thus large abrupt changes in the data will lead to M-residuals which are 0, and hence provide no evidence for a change in the test statistic.

Secondly any loss function which increases linearly in $|y-\theta|$ for sufficiently large $|y-\theta|$ will result in $\phi(y_i-\theta)$ being constant for large $|y_i-\theta|$. Thus, large residuals will have

a bounded contribution to the test statistic. To see this consider the Wald-type test with the Huber loss. To calculate this test statistic we first calculate our estimate of the location parameter for the data, $\hat{\theta}$, assuming the data is from a single segment. The ith residual is then K if $y_i > \hat{\theta} + K$, $-K$ if $y_i < \hat{\theta} - K$, and $y_i - \hat{\theta}$ otherwise. The cusum statistic is just the sum of these residuals. This is equivalent to winsorizing the data, where we shrink extreme positive or negative values to be K above or below our estimate of the location parameter, and then using a cusum test for detecting a changepoint. The actual value of the data points that are above $\hat{\theta} + K$ or below $\hat{\theta} - K$ will not affect the cusum values, and hence not affect the value of the Wald-type test statistic.

The use of Huber loss within a Wald-type test will thus have a similar robustness to extreme outliers that bounded loss functions have for the penalised cost approach. The main difference is that the Wald-type test statistic does not consider whether the data after a putative changepoint is consistent with data from a single segment. Thus a cluster of outliers of the same sign that occur concurrently but which are very different in value, such as we observe for the well-log data, will produce a similar value for the test-statistic as a set of concurrent observations that are very different to the other data points but are also very similar to one another. By comparison, the penalised cost based approach would, correctly, say the latter provided substantially more evidence for the presence of a change.

3 Minimising the Penalised Cost

An issue with detecting changepoints using any of these loss functions, is how can we efficiently minimise the resulting penalised cost over all segmentations? We present an efficient dynamic programming algorithm for performing this minimisation exactly. This algorithm is an extension of the pruned DP algorithm of Rigaill (2015) and the

FPOP algorithm of Maidstone et al. (2017) (see also Johnson, 2013) to the robust loss functions. We will call the resulting algorithm R-FPOP.

3.1 A Dynamic Programming Recursion

We develop a recursion for finding the minimum cost (1) of segmenting data $y_{1:t}$ for $t = 1, \ldots, n$. In the following we let $\boldsymbol{\tau}$ denote a vector of changepoints. Furthermore we let $\mathcal{S}_t$ denote the set of possible changepoints for the $y_{1:t}$, so

$$\mathcal{S}_t = \left\{\boldsymbol{\tau} = \tau_{1:k} : 0 < \tau_1 < \cdots < \tau_k < t\right\}.$$

Note that $\mathcal{S}_t$ has 2^{t-1} elements. Define

$$Q_t = \min_{\boldsymbol{\tau} \in \mathcal{S}_t} Q(y_{1:t}; \tau_{1:k}) = \min_{\boldsymbol{\tau} \in \mathcal{S}_t} \sum_{i=0}^{k} \left\{\mathcal{C}(y_{\tau_i+1:\tau_{i+1}}) + \beta\right\},$$

where here and later we use the convention that k is the number of changepoints in $\boldsymbol{\tau}$, and that $\tau_0 = 0$ and $\tau_{k+1} = t$. First we introduce the minimum penalised cost of segmenting $y_{1:t}$ conditional on the most recent segment having parameter θ,

$$Q_t(\theta) = \min_{\boldsymbol{\tau} \in \mathcal{S}_t} \left[\sum_{i=0}^{k-1} \left\{\mathcal{C}(y_{\tau_i+1:\tau_{i+1}}) + \beta\right\} + \sum_{j=\tau_k+1}^{t} \gamma(y_j; \theta) + \beta\right],$$

where we take the first summation on the right-hand side to be 0 if $k = 0$. Trivially we have $Q_t = \min_\theta Q_t(\theta)$ and $Q_1(\theta) = \gamma(y_1; \theta) + \beta$.

The idea is to recursively calculate $Q_t(\theta)$ for increasing values of t. To do this, we note that each element in $\mathcal{S}_t$ is either an element in $\mathcal{S}_{t-1}$ or an element in $\mathcal{S}_{t-1}$ with

the addition of a changepoint at $t-1$. So

$$\begin{aligned} Q_t(\theta) &= \min_{\boldsymbol{\tau}\in\mathcal{S}_{t-1}}\left[\min\left\{\sum_{i=0}^{k-1}\left(\mathcal{C}(y_{\tau_i+1:\tau_{i+1}})+\beta\right)+\sum_{j=\tau_k+1}^{t}\gamma(y_j;\theta)+\beta,\right.\right. \\ &\qquad \left.\left.\sum_{i=0}^{k}\left(\mathcal{C}(y_{\tau_i+1:\tau_{i+1}})+\beta\right)+\gamma(y_t;\theta)+\beta\right\}\right] \\ &= \min\left\{\min_{\boldsymbol{\tau}\in\mathcal{S}_{t-1}}\left[\sum_{i=0}^{k-1}\left(\mathcal{C}(y_{\tau_i+1:\tau_{i+1}})+\beta\right)+\sum_{j=\tau_k+1}^{t-1}\gamma(y_j;\theta)+\beta\right],\right. \\ &\qquad \left.\min_{\boldsymbol{\tau}\in\mathcal{S}_{t-1}}\left[\sum_{i=0}^{k}\left(\mathcal{C}(y_{\tau_i+1:\tau_{i+1}})+\beta\right)+\beta\right]\right\}+\gamma(y_t;\theta) \\ &= \min\left\{Q_{t-1}(\theta), Q_{t-1}+\beta\right\}+\gamma(y_t;\theta). \end{aligned} \tag{5}$$

The first equality comes from splitting the minimisation into the minimisation over the changepoints for $y_{1:t-1}$ and then whether there is or is not a changepoint at $t-1$. The second equality comes from interchanging the order of the minimisations, and taking out the common $\gamma(y_t;\theta)$ term. The final equality comes from the definitions of $Q_{t-1}(\theta)$ and Q_{t-1}. The right-hand side just depends on $Q_{t-1}(\theta)$, as $Q_{t-1} = \min_\theta Q_{t-1}(\theta)$.

3.2 Solving the Recursion

We now show how we can efficiently solve the dynamic programming recursion from the previous section for loss functions like those introduced in Section 2. We make the assumption that the loss for any observation, $\gamma(y_t;\theta)$, viewed as function of θ, can be written as a piecewise quadratic in θ. Note that by quadratic we include the special cases of linear or constant functions of θ, and this definition covers all the loss functions introduced in Section 2.

As the set of piecewise quadratics is closed under both addition and minimsation, it follows that $C_t(\theta)$ can be written as a piecewise quadratic for all t. We summarise $C_t(\theta)$ by N_t intervals $(a_i^{(t)}, b_i^{(t)}]$, and associated quadratics $q_i^{(t)}(\theta)$. We assume that the intervals are ordered, so $a_1^{(t)} = -\infty$, $a_i^{(t)} = b_{i-1}^{(t)}$ for $i = 2, \ldots, N_t$ and $b_{N_t}^{(t)} = \infty$.

To make this summary of $C_t(\theta)$ unique we further assume that $q_i^{(t)}(\theta) \neq q_{i-1}^{(t)}(\theta)$ for $i = 2, \ldots, N_t$. If this were not the case we could merge the neighbouring intervals.

We can split (5) into two steps. The first is

$$Q_t^*(\theta) = \min \left\{ Q_{t-1}(\theta), Q_{t-1} + \beta \right\}, \tag{6}$$

and the second is

$$Q_t(\theta) = Q_t^*(\theta) + \gamma(y_t; \theta).$$

For the first step we first calculate Q_{t-1} by first minimising the N_{t-1} quadratics defining $Q_{t-1}(\theta)$ on their respective intervals, and then calculating the minimum of these minima. We then solve the minimisation problem (6) on each of the N_{t-1} intervals. For interval i, the solution will either be $q_i^{(t)}(\theta)$, $Q_{t-1} + \beta$ or we will need to split the interval into two or three smaller intervals, on which the solution will change between $q_i^{(t)}(\theta)$ and $Q_{t-1} + \beta$. Thus we will end with a set of N_{t-1}, or more, ordered intervals and corresponding quadratics that define $Q_t^*(\theta)$. We then prune these intervals by checking whether any neighbouring intervals both take the value $Q_{t-1} + \beta$, and merging these if they do. This will lead to a new set of N_t^*, say, ordered intervals, and associated quadratics, $q_{t,i}^*(\theta)$ say.

For each of the N_t^* intervals from the output of the minimisation problem we then add $\gamma(y_t; \theta)$ to the corresponding $q_{t,i}^*(\theta)$. This may involve splitting the ith interval into two or more smaller intervals if one or more of the points of change of the function $\gamma(y_t; \theta)$ are contained in it. This will lead to the N_t intervals and corresponding quadratics that define $Q_t(\theta)$.

The above describes how we recursively calculate $Q_t(\theta)$. In practice we also want to then extract the optimal segmentation under our criteria. This is straightforward to do. For each of the intervals corresponding to different pieces of $Q_t(\theta)$ we can associate a value of the most recent changepoint prior to t. When we evaluate Q_t,

we need to find which interval contains this value, and then the optimal value for the most recent changepoint prior to t is the value associated with that interval. We can store these optimal values for all t, and after processing all data we can recursively track back through these values to extract the optimal segmentation. So we would first find the value of the most recent changepoint prior to n, τ say, then find the value of the most recent changepoint prior to τ. We repeat this until the most recent changepoint is at 0, corresponding to no earlier changepoints.

Pseudo-code for R-FPOP is given in Appendix D. An example of the steps involved in one iteration is given in Figure 3.

4 Computational Cost of R-FPOP

We now present results which bound the computational cost and storage requirements of R-FPOP. As above we will assume that $\gamma(y;\theta)$ can be written as a piecewise quadratic with L pieces. The bounds that we get differ depending on whether, for a given y, $\gamma(y;\theta)$ is convex in θ. We first consider the convex case, which includes all the examples in Section 2 except the biweight loss.

Theorem 4.1 *If γ is convex in θ and defined in L pieces R-FPOP stores at most $2t - 1 + t(L-1)$ quadratics and intervals at step t.*

Corollary 4.2 *If γ is convex in θ and defined in L pieces, the space complexity of R-FPOP is $\mathcal{O}(n)$, and the time complexity of R-FPOP is $\mathcal{O}(n^2)$.*

For the biweight loss, which is not convex, we get worse bounds on the complexity of R-FPOP.

Theorem 4.3 *For the biweight loss R-FPOP stores $\mathcal{O}(t^2)$ intervals at step t.*

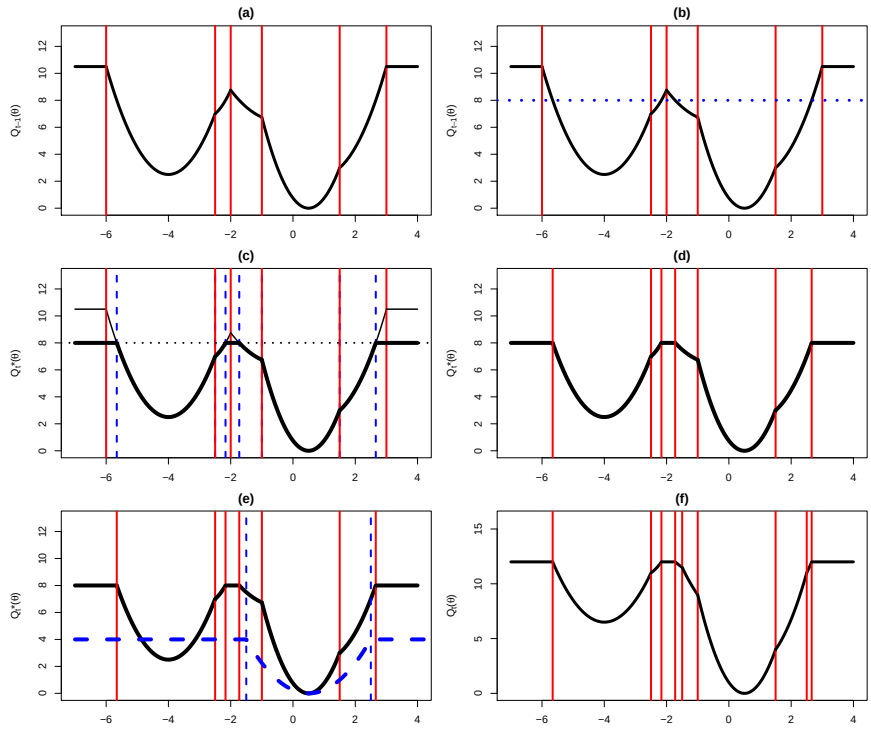

Figure 3: Example of one iteration of R-FPOP: (a) $Q_{t-1}(\theta)$ (black solid line), and set of intervals stored (split by vertical red lines) at start of iteration. (b) Find the pointwise minimum of $Q_{t-1}(\theta)$ and $Q_{t-1} + \beta$ (blue dashed line). (c) This is done by solving the minimisation on each interval, which splits some intervals into two or three. New splits are shown by blue dashed vertical lines. (d) Merge neighbouring intervals if they both take the value $Q_{t-1} + \beta$. (e) Now add the loss for the new observation (blue dashed curve). (f) This further splits intervals at the points where the form of $\gamma(y_t; \theta)$ changes, the blue vertical lines in plot (e). Shown is the final representation of $Q_t(\theta)$. At all stages only piecewise quadratic functions need to be stored.

Corollary 4.4 *For the biweight loss R-FPOP has worst-case space complexity that is* $\mathcal{O}(n^2)$*, and time complexity that is* $\mathcal{O}(n^3)$*.*

These results give worst-case bounds on the time and storage complexity of R-FPOP. Below we investigate empirically the time and storage cost and observe an average computational cost that is linear in n when the number of changepoints is large and less than quadratic when there is no changepoint.

5 Results

5.1 Simulation Study: Computational Cost

This paper is mostly concerned with the statistical performances of our robust estimators. Thus an in-depth analysis of the runtime of our approach is outside the scope of this paper. In this section we just aim at showing that our approach is easily applicable to large profiles ($n = 10^3$ to $n = 10^6$) in the sense that its runtime is comparable to other commonly used approach like FPOP (Maidstone et al., 2017), PELT (Killick et al., 2012), WBS (Fryzlewicz, 2014) or smuceR (Frick et al., 2014).

We used a standard laptop with an Intel Core i7-3687U CPU with 2.10GHz x 4 Core and 7.7 Gb of Ram. For the biweight loss, for a profile of length $n = 10^6$ and in the absence of any true change the runtime is around 4 seconds (slightly larger than FPOP see Figure 4 left L_2). As a matter of comparison on the same computer the runtime of competitor methods WBS, PELT and smuceR for a profile of length $n = 10^5$ are respectively around 7 seconds, 40 seconds and 175 seconds. For an increasing number of changes runtimes are smaller (see Figure 4 right). Runtimes for the L_1 and Huber loss are quite a bit larger: in the absence of changes and for $n = 10^6$ the L_1 runtime is around 500 seconds and the Huber runtime is around 200 seconds (see Figure 4

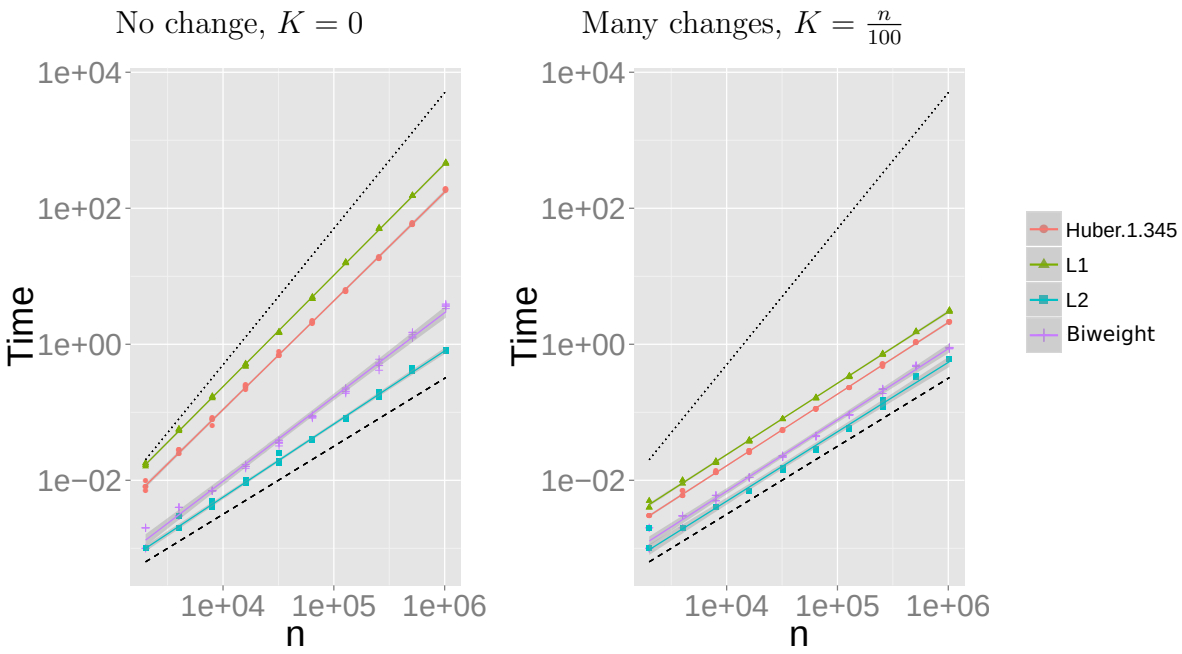

Figure 4: Runtime in seconds of R-FPOP for different loss functions. We simulated profiles with n going from 2000 to 1024000, with or without changes and using IID Gaussian noise. The axes use a log-scale, and we have added lines of slope 1 (dashed) and 2 (dotted).

left).

Most importantly, we see that with many changepoints, the average CPU cost of all penalised cost approaches increases only linearly with the number of data points (parallel to the dashed black line in Figure 4 right). With no changepoints, the average CPU cost increases faster in particular for the L_1 and Huber losses however it is less than quadratic (slopes smaller than the dotted black line in Figure 4 left). The CPU cost of the biweight loss is very close to the CPU cost of the L_2 loss.

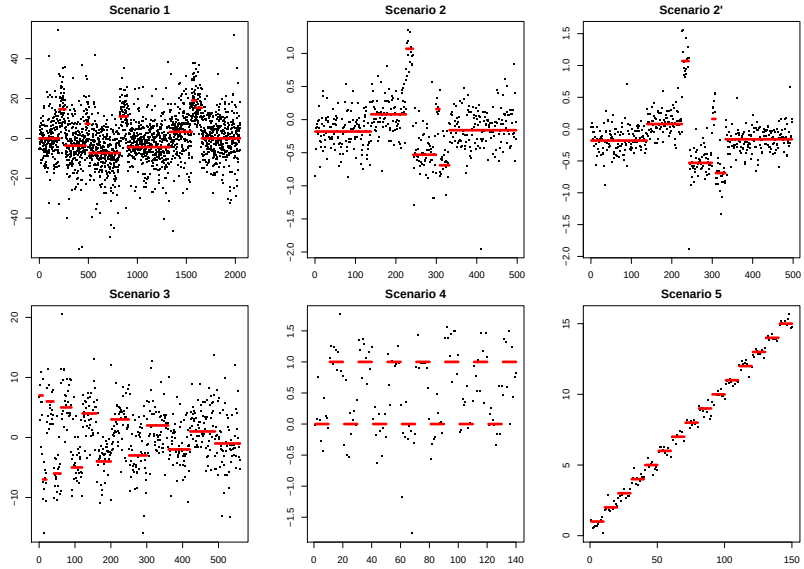

Figure 5: The signal, and example data, for each of the scenarios considered for the simulation study. Data was generated with the noise having a t-distribution with 5 degrees of freedom.

5.2 Simulation Study: Accuracy

We assessed the performance of our robust estimators using the simulation benchmark proposed in the WBS paper (Fryzlewicz, 2014). In that paper 5 scenarios are considered. These vary in length from $n = 150$ to $n = 2048$ and contain a variety of short and long segments, and a variety of sizes of the change in location from one segment to the next. We considered an additional scenario from Frick et al. (2014) corresponding to scenario 2 of WBS with a standard deviation of 0.2 rather than 0.3. In our simulation study we are interested to see how the presence of outliers or heavy tailed noise affect different changepoint methods, and so we will test each method assuming t-distributed noise. The underlying signals and example data for the three scenarios are shown in Figure 5.

For all approaches we need to choose the value K in the loss function and the penalty/threshold for adding a changepoint. These will depend on the standard

deviation of the noise. Our approach is to estimate this standard deviation using the median absolute deviation of the differenced time-series, as in Fryzlewicz (2014), which we denote as $\hat{\sigma}$. We compared our various robust estimators (Huber and biweight loss) to binary segmentation using the robust cusum test (Hušková and Sen, 1989), described in Section 2.3 (Cusum). For the biweight loss we chose $K = 3\hat{\sigma}$, so that extreme residuals according to a Gaussian model are treated as outliers. For the Huber loss we chose $K = 1.345\hat{\sigma}$, a standard choice for trading statistical efficiency of estimation with robustness. We further set the penalty/threshold to be $\beta = 2\hat{\sigma}^2 \log(n)\mathrm{E}(\phi(Z)^2)$, where ϕ is the gradient of the loss function and Z is a standard Gaussian random variable. This is based on the Schwarz information criteria, adapted to account for the variability of loss function that is used (see, e.g., theoretical results in Hušková and Marušiaková, 2012, for further justification of this), and for the biweight loss this is inline with Theorem 2.3, which suggested the use of a penalty that is proportional to $\log(n)$. We also compared to just using the standard square-error loss: implemented using FPOP (Maidstone et al., 2017); and to the WBS (Fryzlewicz, 2014) approach that uses a standard cusum test statistic for detecting changepoints. Again we used $\beta = 2\hat{\sigma}^2 \log(n)\mathrm{E}(\phi(Z)^2)$, which in this case simplifies to the standard BIC penalty $\beta = 2\hat{\sigma}^2 \log(n)$, and is the value that gave the best results for these methods across the 6 scenarios when there is normal noise (see Maidstone et al., 2017).

We consider analysing data where the noise was from a t-distribution. We vary the degrees of freedom from 3 to 100 to see of how varying how heavy-tailed the noise is affects the performance of different methods.

In Figures 6 and 7 we show the results of all approaches as a function of the degrees of freedom. We compare methods based on how they estimate the underlying piecewise constant mean function, measured in terms of mean square error; and how well they estimate the segmentation, measured using the normalized rand-index. The normal-

ized rand-index measures the overlap between the true segmentation and the inferred segmentation, with larger values indicating a better estimation of the segmentation.

In terms of mean square error, for almost all scenarios we consider the biweight loss performs best when the degrees of freedom is small. It also appears to lose little in terms of accuracy when the degrees of freedom is large, and the noise is close to Gaussian. The robust cusum approach also performs well when the degrees of freedom are small, but in most cases it shows a marked drop in accuracy relative to the alternative methods when the noise is close to Gaussian. The one scenario where the biweight loss performs poorly when the noise is close to Gaussian is Scenario 4. In this case we have short segments, only slightly larger than the minimum segment length for the biweight loss, with the segment mean being the same for all odd segments. We can get a reasonable fit under the biweight loss by, for example, ignoring the changepoints and treating all observations in the even segments as outliers. The problem of distinguishing between this case and the presence of actual changepoints causes the poor performance.

The results in terms of the quality of the segmentation, as measured using the rand-index, are more mixed. The biweight loss is clearly best in scenarios 1 and 2, but performs poorly for scenario 3. Here the use of the Huber loss appears to give best results across the different scenarios. Again we see that the use of the L2 loss, using either FPOP or WBS, performs poorly when the degrees of freedom are small.

5.3 Online analysis of well-log data

We return to the well-log data of Figure 1. For this data, due to the presence of substantial outliers, we choose to use the biweight loss function. We set the threshold, K in (2), to be twice an estimate of the standard deviation of the observation noise. We set β to be 70 times the estimated variance of the noise. This is larger than that

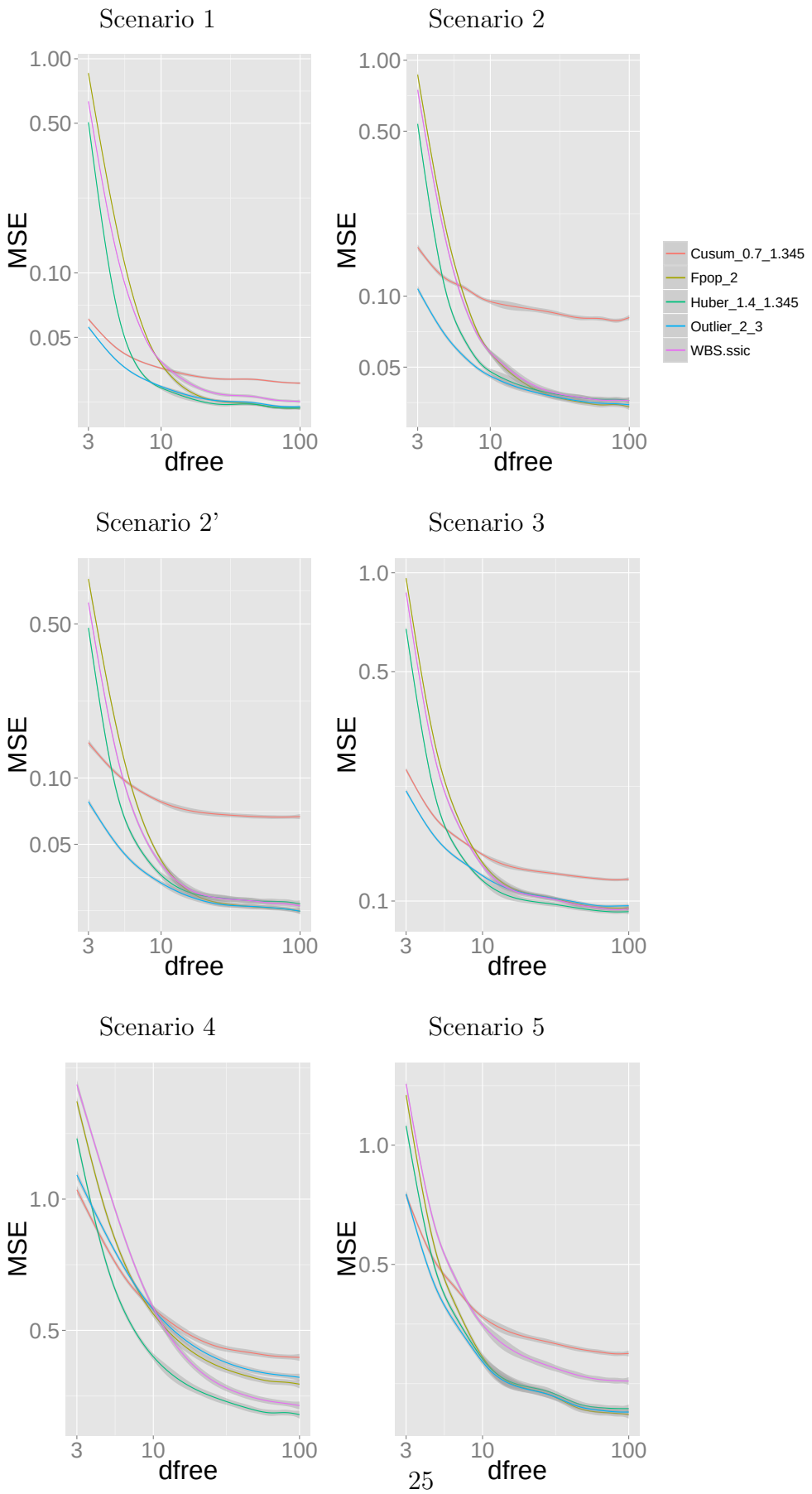

Figure 6: Smoothed log MSE of all tested approaches on the 6 scenarios using a student-noise with the degrees of freedom ranging from 3 to 100.

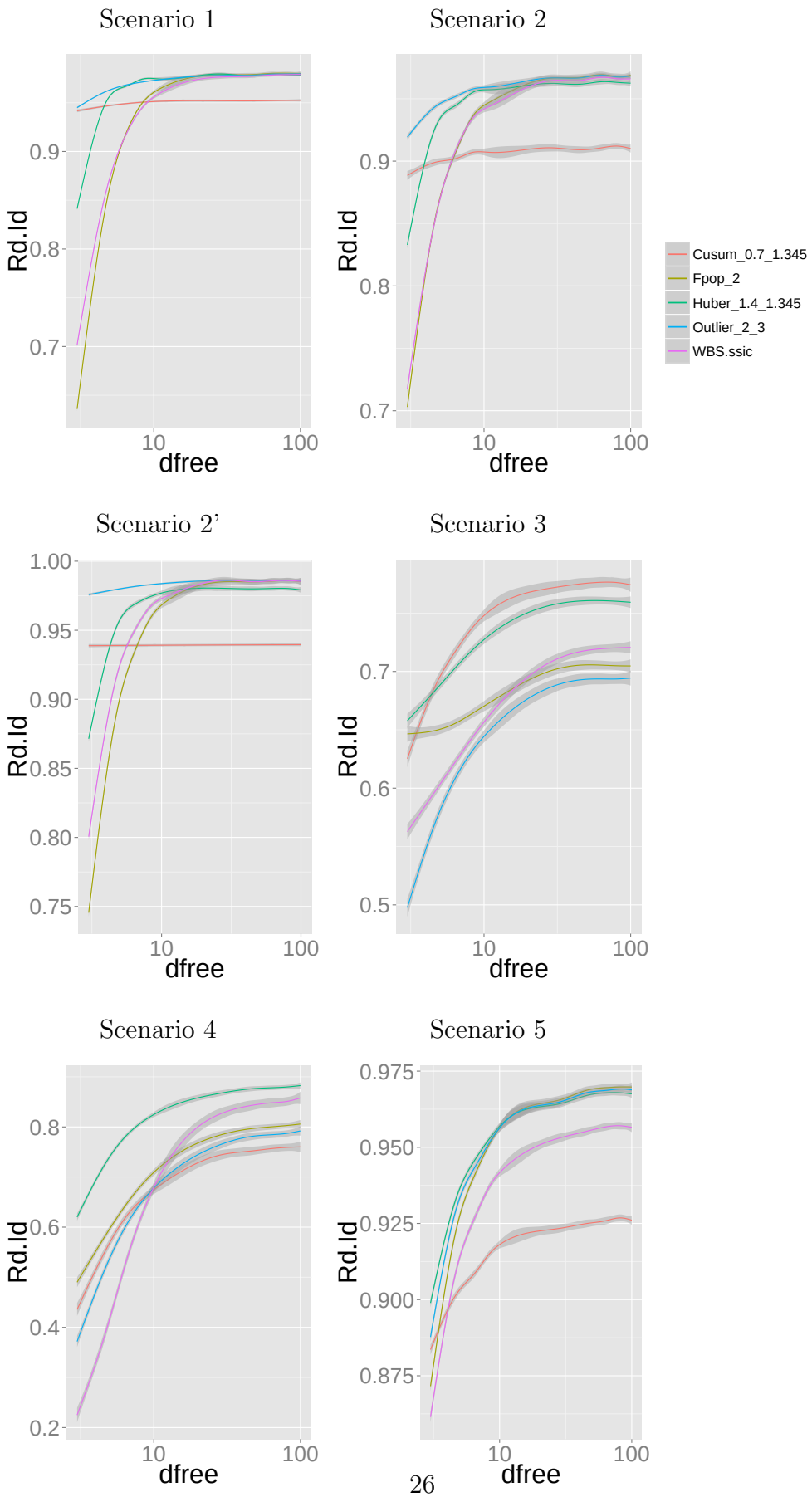

Figure 7: Smoothed normalized Rand-index of all tested approaches on the 6 scenarios using student-noise with the degrees of freedom varying from 3 to 100.

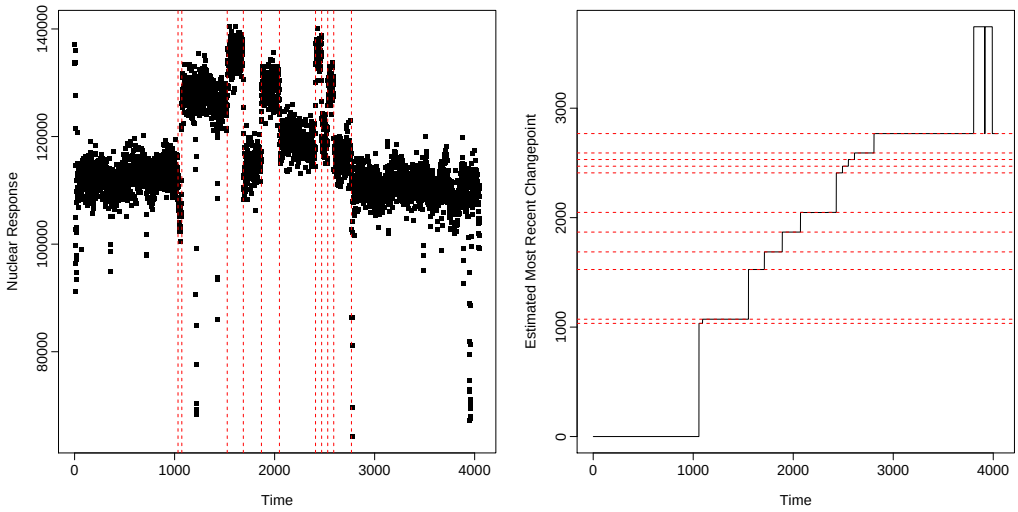

Figure 8: Estimated changepoints from batch analysis of the well-log data (left-hand plot) under biweight loss. Estimate of location of most-recent changepoint from online analysis (right-hand plot). The black line shows the estimate of the most-recent changepoint against the number of data points analysed. The red dashed horizontal lines show the locations of the changepoints detected from the batch analysis.

of the BIC penalty, but this is needed due to the presence of auto-correlation in the observation noise (Lavielle and Moulines, 2000), and is the same penalty used for the analysis presented in Figure 1.

Figure 8 shows the estimated changepoints we obtain from a batch analysis of the data. As we can see, using the biweight penalty makes the changepoint detection robust to the presence of the outliers. All obvious changes are detected, and we do not detect a change at any point where the outliers cluster.

As mentioned in the introduction, the motivation for analysing this data requires an online analysis. We present output from such an online analysis in the right-hand plot of Figure 8. Here we plot the estimate of the most recent changepoint prior to t, given data $y_{1:t}$, as a function of t. To help interpret the result we also show the locations of the changepoints inferred from the batch analysis. We see that we are able to quickly

detect changes when they happen, and we have only one region where there is some fluctuation in where we estimate the most recent changepoint. Whilst by eye the plot may suggest we immediately detect the changes, there is actually some lag. This is inevitable when using the biweight loss, due to the presence of a minimum segment length that can be inferred (see Theorem 2.2). The lag in detecting the changepoint is between 21 and 27 observations for all except the final changepoint. The final inferred changepoint is less pronounced, and is not detected until after a lag of 40 observations. This lag can be reduced by increasing K, but at the expense of less robustness to outliers. The region of fluctuation over the estimate of the most recent changepoint corresponds to uncertainty about whether there are changepoints in the last inferred segment (corresponding to the final two changepoints inferred in the bottom-left plot of Figure 1). One disadvantage of detection methods that involve minimising a penalised cost, and of other methods that produce a single estimate of the changepoint locations, is that they do not quantify the uncertainty in the estimate.

5.4 Estimating Copy Number Variation

Healthy human cells have two copies of DNA. In tumor cells, parts of chromosomes of various sizes (from kilobases to a chromosome arm) may be deleted or amplified several times, and this can lead to the copy number of the DNA from such regions being different from 2. Copy numbers (CN) can be measured using microarray or sequencing experiments. They are piecewise constant along the genome, and interest lies in detecting whether, and where, the copy number changes. For many samples we would have a mixture of healthy and tumor cells, and the signal to noise ratio for changes in copy number will go down with the tumor fraction. The detection of changes in copy number is further complicated by the presence of outliers. We illustrate this in Figure 9 using output from the jointseg package (Pierre-Jean et al.,

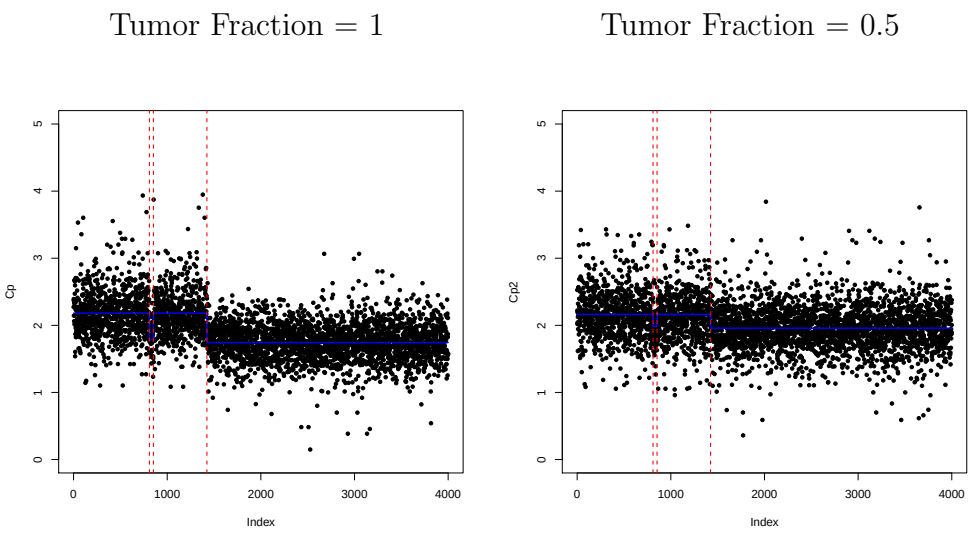

Figure 9: Two DNA copy number profiles obtained using the jointseg package with a tumor fraction of 1 (left) and 0.5 (right). The true change-points are represented with red dotted lines. It can be seen that a number of data-points are quite far from from the blue line. The size of each jump is larger when the tumor fraction is larger.

2015) which enables simulation of realistic CN profiles by resampling real datasets for which the truth is known.

A standard way to analyse such data is to use the smooth.CNA function of the well known DNAcopy package (Bengtsson et al., 2016). This function shrinks outliers towards the value of its neighbors. Once this is done one can run a preferred segmentation approach. As we will see below, this heuristic preprocessing procedure greatly improves changepoint detection. We want to compare such a two-stage approach to a simpler analysis where we analyse data using our penalised cost approach with the biweight loss.

To assess the performance of our approach on DNA copy number data we used the jointseg package. We simulated profiles of length $n = 4000$ with 10 change-points with segments of at least 40 data-points. The package propose two real datasets, GSE11976 and GSE29172, to resample from. For both we considered four levels

of difficulty corresponding to different tumor fractions: 0.34, 0.50, 0.79 and 1 for GSE11976; and 0.3, 0.5, 0.7 and 1 for GSE29172.

We consider four approaches: FPOP (L2), FPOP after using smooth.CNA to remove outliers (Rout L2), robust binary segmentation (Cusum) and our biweight loss with a threshold value of 3. All approaches are implemented for a range of penalty values. For every simulated profile and each run of a method we computed the number of true positive (TP) and false positive (FP) change-points. For all true change-point we counted one TP if there is at least one change-point identified within a window of 15 data-points. We then computed the number of FPs as the number of predicted changes minus the number of TPs. We then average, over 200 simulated profiles, the number of TPs and FPs per approach, penalty value and difficulty to recover ROC curves.

Overall our robust biweight loss outperforms the L2 loss following outlier removal and the Cusum approach. For low tumor fractions (0.3 and 0.5 GSE29172 and 0.34 GSE11976) the biweight loss is possibly slightly better than the Cusum approach. For a tumor fraction of 1 the biweight loss is slightly better than the L2 following outlier removal. In other cases it is clearly better. Results are shown for the two datasets and a tumor fraction of 0.7 and 0.79 in Figure 10. Results for other tumor fractions are provided in figures in Appendix E.

5.5 Wireless Tampering

We now consider an application which looks at security of the Internet of Things (IoT). Many IoT devices use WiFi to communicate. Often, for example with surveillance systems, these need a high level of security. Thus it is important to be able to detect if a device has been tampered with. WiFi signals include a “preamble” which is used by the receiver to determine channel state. One approach that can be used to detect

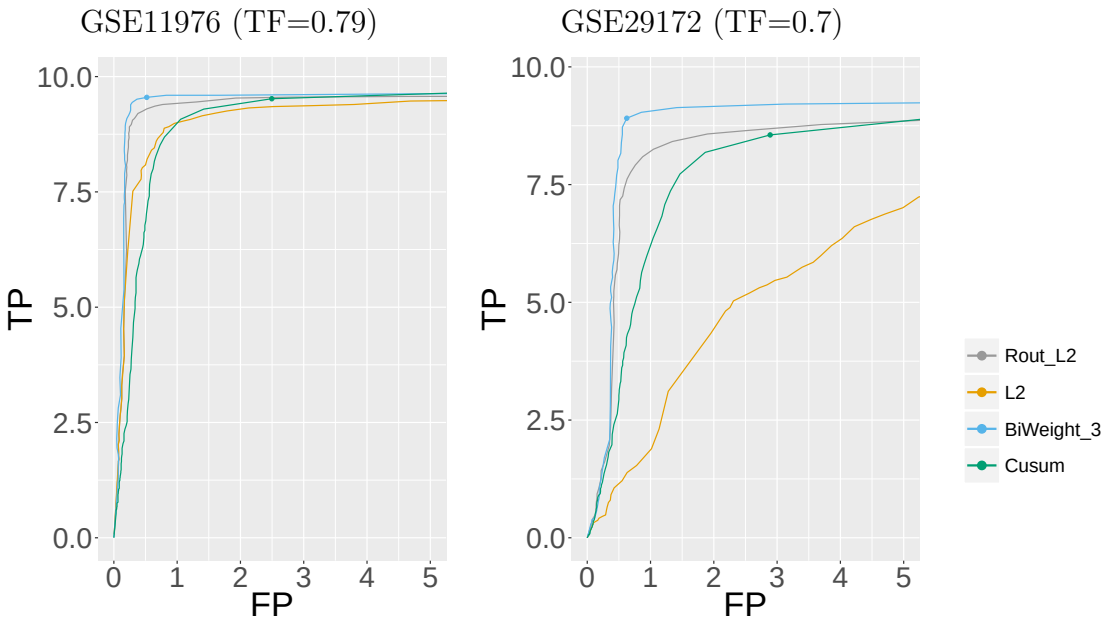

Figure 10: Average ROC curve on the GSE11976 and GSE29172 datasets for a tumor fraction of respectively 0.79 and 0.7, for the Cusum, L2, L2 with outlier removal (Rout L2) and our robust biweight loss (Biweight 3).

tampering is to monitor channel state variation (Bagci, 2016). Abrupt changes in it could indicate some tampering event. However changes can also be caused by less sinister events, such as movement of people within the communication environment. Thus the challenge is to detect a change caused by tampering as opposed to any "outliers" caused by such temporary environmental factors.

Figure 11 shows some time-series of channel state information (CSI) that has been extracted from the preamable from a signal sent by a single IoT device. This data is taken from Bagci et al. (2015), where a controlled experiment was performed, with an actual tampering event occurring after 22 minutes. Before this tampering event, there was movement of people around the device, which has a short-term effect on the time-series data.

In practice the channel state information from an IoT device is multi-dimensional, and we show time-series for 6 out of 90 dimensions. Whilst ideally we would jointly analyse the data from all 90 time-series that we get from the device, we will just consider analysing each time-series individually. Our interest is to see how viable it is to use our approach, with the biweight loss, to accurately distinguish between tampering event and any effects due to temporary environmental factors. The six time-series we show each show different patterns, both in terms of the change caused by tampering, and the effect of people walking near the device. As such they give a thorough testing of any approach. We implemented the biweight loss with the SIC penalty for a change, and with K chosen so that the minimum segment length (see Theorem 2.2) corresponds to a period of 20 seconds. Results are shown in Figure 11, where we see that we accurately only detect the change that corresponds to the tampering event in all cases.

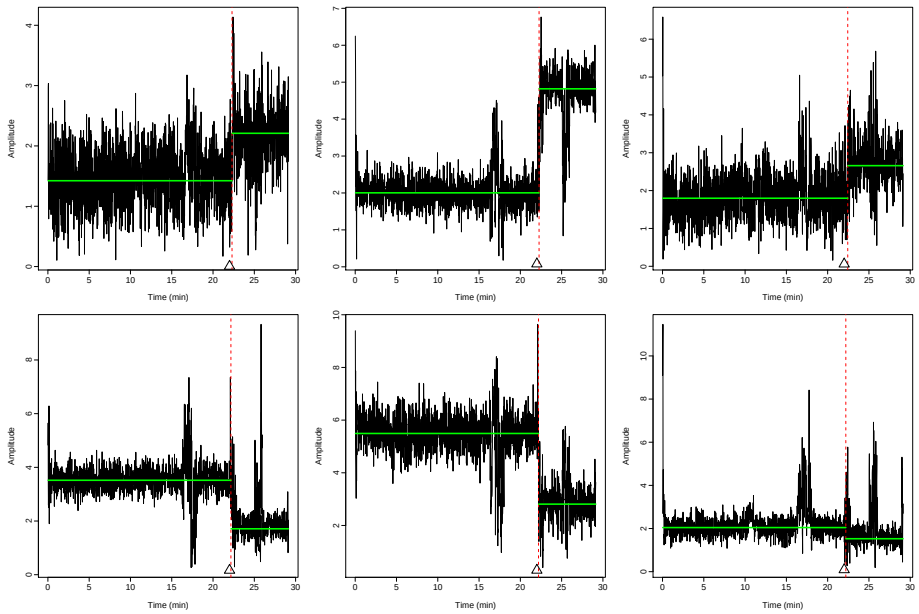

Figure 11: Examples from the analysis of the wireless tampering data. We show six examples of the data, with different structure before and after a change, and with different patterns of outliers caused by temporary environmental factors. In each case there is a single changepoint, after 22 minutes (denoted by the triangle). The inferred changepoint (vertical dashed line) and inferred mean function (green full horizonal line) from our method with the biweight loss function are shown in each case.

6 Discussion

We have presented an algorithm for detecting changepoints by minimising a penalised cost which measures fit to the data by a loss function that is piecewise quadratic. In particular we have shown that by using bounded loss functions we can develop algorithms that are robust to the presence of arbitrarily large outliers. We particular recommend the use of the biweight loss function, and have shown that using such a loss function can lead to consistent esimation of the number of changepoints and accurate estimation of their location under weak conditions on the noise distribution.

If we use the biweight loss we have to choose an appropriate value for K. To some extent this is a modelling decision, but a reasonable default is to choose this to be around 2 to 3 times an estimate of the standard deviation of the noise. This will mean the loss performs similarly to the square error loss, as most observations will be within K of the segment location parameter, but with the added benefit of robustness to extreme outliers.

We have shown that using the biweight loss with a penalty for adding a changepoint that is $C_1 \log(n)$ for some suitable constant C_1 can lead to consistent estimation of the number of changepoints. If K is chosen as suggested, it is natural to choose C_1 to be similar to choices that are known to work well with the square-error loss, as we did within Section 5.2. Such choices are not guaranteed to produce large enough constants to ensure consistency. If this is a concern, it is possible to use the idea behind that strengthened Schwarz information criteria of Fryzlewicz (2014), and choose a penalty $C_1(\log n)^{1+\epsilon}$ for some small $\epsilon > 0$.

Care must be taken if there are violations of the IID assumption for the noise. In such cases it is known that consistent estimation of the number of changepoints is still possible if we appropriately inflate the penalty (Lavielle and Moulines, 2000), and we would suggest using a similar inflation when using the biweight loss. Choosing

how much to inflate is difficult in practice, and thus it makes sense to try a range of penalties (which can be done efficiently, e.g. using the CROPS algorithm of Haynes et al., 2017a). For applications which involve analysing multiple similar data sets, we would recommend using a small set of training data to help choose an appropriate constant (see e.g. Rigaill et al., 2013).

Finally, the joint choice of K and β can be informed by the minimum segment length that can be inferred for such a choice; see Theorem 2.2. To have robustness to extreme outliers we need this minimum segment length to be greater than 1. Equally it should be chosen to be smaller than the shortest segment we wish to identify. This choice is linked to the question of how many similar observations would we require before we would classify them as coming from a new segment as opposed to being correlated outliers.

Acknowledgements We thank Utz Roedig and Ethem Bagci for supplying and discussions around the wireless tampering data, and to Lawrence Bardwell for help with analysing this data. This work was supported by EPSRC grant EP/N031938/1 (StatScale) and an ATIGE grant from Génopole.

References

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

Bagci, I. E. (2016). *Novel security mechanisms for wireless sensor networks*. PhD thesis, Lancaster University, Lancaster, UK.

Bagci, I. E., Roedig, U., Martinovic, I., Schulz, M., and Hollick, M. (2015). Using channel state information for tamper detection in the internet of things. In

Proceedings of the 31st Annual Computer Security Applications Conference, pages 131–140. ACM.

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric theory*, 13(3):315–352.

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2016). Narrowest-over-threshold detection of multiple change-points and change-point-like features. *arXiv:1609.00293*.

Bengtsson, H., Neuvial, P., Seshan, V. E., Olshen, A. B., Spellman, P. T., and Olshen, R. A. (2016). Package pscbs.

Cao, H. and Wu, W. B. (2015). Changepoint estimation: another look at multiple testing problems. *Biometrika*, 102(4):974–980.

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

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

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

Futschik, A., Hotz, T., Munk, A., and Sieling, H. (2014). Multiscale DNA partitioning: statistical evidence for segments. *Bioinformatics*, 30:2255–2262.

Haynes, K., Eckley, I. A., and Fearnhead, P. (2017a). Computationally efficient changepoint detection for a range of penalties. *Journal of Computational and Graphical Statistics*, 26:134–143.

Haynes, K., P, F., and Eckley, I. (2017b). A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, 27:1293–1305.

Hinkley, D. V. (1971). Inference about the change-point from cumulative sum tests. *Biometrika*, 58(3):509–523.

Hotz, T., Schütte, O. M., Sieling, H., Polupanow, T., Diederichsen, U., Steinem, C., and Munk, A. (2013). Idealizing ion channel recordings by a jump segmentation multiresolution filter. *IEEE Transactions on Nanobioscience*, 12(4):376–386.

Huber, P. J. (2011). *Robust statistics*. Springer.

Hušková, M. (1991). Recursive M-tests for the change-point problem. In *Economic Structural Change*, pages 13–33. Springer.

Hušková, M. (2013). Robust change point analysis. In *Robustness and Complex Data Structures*, pages 171–190. Springer.

Hušková, M. and Marušiaková, M. (2012). M-procedures for detection of changes for dependent observations. *Communications in Statistics-Simulation and Computation*, 41(7):1032–1050.

Hušková, M. and Picek, J. (2005). Bootstrap in detection of changes in linear regression. *Sankhyā: The Indian Journal of Statistics*, pages 200–226.

Hušková, M. and Sen, P. K. (1989). Nonparametric tests for shift and change in regression at an unknown time point. In *Statistical Analysis and Forecasting of Economic Structural Change*, pages 71–85. Springer.

Johnson, N. A. (2013). A dynamic programming algorithm for the fused lasso and l_0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260.

Killick, R., Eckley, I. A., Ewans, K., and Jonathan, P. (2010). Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, 37(13):1120–1126.

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.

Kim, C.-J., Morley, J. C., and Nelson, C. R. (2005). The structural break in the equity premium. *Journal of Business & Economic Statistics*, 23:181–191.

Lavielle, M. and Moulines, E. (2000). Least-squares estimation of an unknown number of shifts in a time series. *Journal of time series analysis*, 21(1):33–59.

Ma, T. F. and Yau, C. Y. (2016). A pairwise likelihood-based approach for change-point detection in multivariate time series models. *Biometrika*, 103(2):409–421.

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

National Research Council (2013). Frontiers in massive data analysis.

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):557–72.

Ó Ruanaidh, J. J. K. and Fitzgerald, W. J. (1996). *Numerical Bayesion Methods Applied to Signal Processing*. New York: Springer.

Page, E. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

Pierre-Jean, M., Rigaill, G., and Neuvial, P. (2015). Performance evaluation of DNA copy number segmentation methods. *Briefings in Bioinformatics*, 16:600–615.

Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915.

Rigaill, G. (2015). A pruned dynamic programming algorithm to recover the best segmentations with 1 to $K_{\max}$ change-points. *Journal de la Société Française de Statistique*, 156(4):180–205.

Rigaill, G., Hocking, T. D., Bach, F., and Vert, J.-P. (2013). Learning sparse penalties for change-point detection using max margin interval regression. In *Proceedings of the 30th International Conference on Machine Learning, JMLR W&CP*, volume 28, pages 172–180.

Ruggieri, E. and Antonellis, M. (2016). An exact approach to Bayesian sequential change point detection. *Computational Statistics & Data Analysis*, 97:71–86.

Vostrikova, L. (1981). Detection of the disorder in multidimensional random-processes. *Doklady Akademii Nauk SSSR*, 259(2):270–274.

Worsley, K. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, 74(366a):365–367.

Wyse, J., Friel, N., 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.

Yao, Y.-C. and Au, S. (1989). Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 370–381.

Supplementary Material for Changepoint Detection in the Presence of Outliers

Appendix A Proof of Theorem 2.3

We first introduce notation, and present some basic properties that will be used within the proof. Let $\gamma(y;\theta)$ be the biweight loss, so $\gamma(y;\theta) = \min\{(y-\theta)^2, K^2\}$, and we will keep K fixed throughout. To simplify notation we will write $\gamma(y;0) = \gamma(y)$, and further note that $\gamma(y;\theta) = \gamma(y-\theta;0) = \gamma(y-\theta)$. We also note that $\gamma(y)$ is bounded, $0 \leq \gamma(y) \leq K^2$, and satisfies a Lipschitz property

$$|\gamma(y_1) - \gamma(y_2)| \leq 2K|y_1 - y_2|.$$

Let $Z_1, Z_2, \ldots$ be independent, identically distributed (IID) random variables whose distribution is that of the residual process within our model. The following random functions will play an important role

$$X_i(\theta) = \gamma(Z_i - \theta) - \gamma(Z_i).$$

By assumption (3), we have

$$\mathrm{E}\{X_i(\theta)\} = M(\theta) - M(0) =\geq \min\{c_1\theta^2, c_2\},$$

for constants $c_1 > 0$ and $c_2 > 0$. Define $v(\theta) = \mathrm{E}\{X_i(\theta)^2\}$. By the bounded and Lipschitz properties of $\gamma(y)$ we have

$$v(\theta) = \mathrm{E}\left\{X_i(\theta)^2\right\} \leq \min\{4K^2\theta^2, K^4\}.$$

We will be interested in sums of $X_i(\theta)$, and thus define

$$S_l(\theta) = \sum_{i=1}^{l} X_i(\theta).$$

To simplify notation in the following we will use C_1, C_2, α, α' etc. to denote constants, and allow these constants to differ for different results and for different parts of the proofs.

Our consistency results states that asymptotically we estimate the correct number of changepoints and that each changepoint location is estimated within some degree of accuracy. The proof of consistency requires two preliminary results – the proofs of which appear at the end of this section. The first is used to bound how much the un-penalised cost can be reduced by adding extra changepoints, and is used to show that asymptotically we do not over-estimate the number of changepoints.

Lemma A.1 *Under assumptions (3) and (4), there exists strictly positive constants α, C_1 and δ such for any $l \leq n$,*

$$\Pr\left\{\min_{\theta} S_l(\theta) < -\alpha \log(n)\right\} \leq C_1 n^{-2-\delta}.$$

The second lemma will give us a probabilistic bound on the increase in the cost we get if we miss a true changepoint. This will be used to show both that asymptotically we cannot under-estimate the number of changepoints, and that any estimated segmentation must have an estimated changepoint "close to" a real changepoint.

Lemma A.2 *Let $S'_l(\theta)$ be an independent copy of $S_l(\theta)$. Then under assumptions (3) and (4), for any $\Delta > 0$, there exists positive constants, C_1, C_2 and α such that*

$$\Pr\left(\min_{\theta}\left\{S_l(\theta) + S'_l(\theta - \Delta)\right\} \leq l\alpha\right) \leq C_1 \exp\{-C_2 l\}.$$

Remember that our asymptotic results are for esimators of the changepoints obtained by minimising a penalised cost

$$Q(y_{1:n}; \hat{\tau}_{1:k}) = \sum_{i=0}^{k} \left\{\mathcal{C}(y_{\hat{\tau}_i+1:\hat{\tau}_{i+1}}) + \beta_n\right\},$$

where we allow the penalty for adding a changepoint, β_n, to depend on n. To simplify the notation we will write this cost function as $Q(\hat{\tau}_{1:k})$ from now on. We will further introduce the notation

$$Q_0(\hat{\tau}_{1:k}) = \sum_{i=0}^{k} \mathcal{C}(y_{\hat{\tau}_i+1:\hat{\tau}_{i+1}}),$$

to denote the un-penalised cost of a segmentation. We will extend both these functions to be defined when the argument is a set of un-ordered changepoints. Thus for an unordered vector $\hat{\tau}_{1:k}$, if we let $\hat{\tau}_{(1)} < \hat{\tau}_{(2)} < \cdots < \hat{\tau}(k)$ be the ordered changepoint locations then, for example,

$$Q_0(\hat{\tau}_{1:k}) = \sum_{i=0}^{k} \mathcal{C}(y_{\hat{\tau}_{(i)}+1:\hat{\tau}_{(i+1)}}),$$

where as before $\tau_{(0)} = 0$ and $\tau_{(k+1)} = n$. Finally we will use, for example, $Q_0(\hat{\tau}_{1:k}, \hat{\tau}'_{1:k'})$ to be the unpenalised cost for a segmentation with changepoints given by the union of $\hat{\tau}_{1:k}$ and $\hat{\tau}'_{1:k'}$.

Proof of Theorem 2.3

The proof has three parts. Each showing that the probability of "bad" segmentations, for a different definition of "bad", goes to 0 as n increases. In each case the idea is to show that the penalised cost for such "bad" segmentations must be greater than some other segmentation. The segmentation we compare with will either be the true segmentation of the data, or a slight adaptation of the "bad" segmentation which adds one or more true changepoints. Furthermore we can upper bound the un-penalised cost of, say, the correct segmentation as

$$Q_0(\tau_{1:k_0}) \leq \sum_{i=1}^{n} \gamma(Z_i), \tag{7}$$

where $Z_1, \ldots, Z_n$ are the noise random variables used to generate the data. This upper bound comes from the fact it is the cost if we fix the segment location parameter for each segment to be the true value for that segment, as opposed to minimising the cost over all possible location parameter values. A similar bound will also be used to

bound the contribution to the cost function from individual segments that are subsets of a true segment. We also repeatedly used the fact that if we add changepoints then the un-penalised cost can never increase.

Part 1

The first part is to show that the estimated number of changepoints, $\hat{k}_n$, satisfies

$$\Pr(\hat{k} > k_0) \to 0$$

as $n \to \infty$, provided $\beta_n > C_1 \log(n)$ for a suitable C_1. To do this we use Lemma A.1.

Let E_n be the event that

$$\min_{1 \le s < t \le n} \min_{\theta} \sum_{i=s}^{t} \{\gamma(Z_i - \theta) - \gamma(Z_i)\} > -\alpha \log(n),$$

where α is defined as in Lemma A.1. Then by Lemma A.1 it follows that $\Pr(E_n) \to 1$ as $n \to \infty$.

Now assume that E_n occurs and consider a segmentation $\hat{\tau}_{1:\hat{k}}$ with $\hat{k} > k$. We have

$$\begin{aligned} Q_0(\hat{\tau}_{1:\hat{k}}) &\geq Q_0(\hat{\tau}_{1:\hat{k}}, \tau_{1:k_0}) \\ &= \sum_{j=0}^{k^*} \mathcal{C}(Y_{\tau^*_{(j)}+1:\tau^*_{(j+1)}}) \\ &= \sum_{j=0}^{k^*} \min_{\theta} \sum_{i=\tau^*_{(j)}+1}^{\tau^*_{(j+1)}} \gamma(Z_i - \theta), \end{aligned}$$

where $k^* \leq k_0 + \hat{k}$ is the number of distinct changepoints in the union of $\hat{\tau}_{1:\hat{k}}$ and $\tau_{1:k_0}$, and $\tau^*_{(1)}, \ldots, \tau^*_{(k^*)}$ are the ordered changepoint locations.

As E_n occurs, and using (7),

$$\begin{aligned} Q_0(\hat{\tau}_{1:\hat{k}}) - Q_0(\tau_{1:k_0}) &\geq Q_0(\hat{\tau}_{1:\hat{k}}) - \sum_{i=1}^{n} \gamma(Z_i) \\ &= \sum_{j=0}^{k^*} \min_{\theta} \sum_{i=\tau^*_{(j)}+1}^{\tau^*_{(j+1)}} \{\gamma(Z_i - \theta) - \gamma(Z_i)\} \\ &> -(k^* + 1)\alpha \log(n). \end{aligned}$$

Thus for the penalised costs,

$$Q(\hat{\tau}_{1:\hat{k}}) - Q(\tau_{1:k_0}) > -(k^* + 1)\alpha \log(n) + (\hat{k} - k_0)\beta_n.$$

Now as $k^* \leq k_0 + \hat{k}$ we have that the right-hand side will be positive if $\beta_n > 2(k_0 + 1)\alpha \log(n)$. For such a β_n we have that the true segmentation will have a lower penalised cost than any segmentation with more than k_0 changepoints if E_n occurs. As $\Pr(E_n) \to 1$ we have that $\Pr(\hat{k}_n > k_0) \to 0$ as required.

Part 2

The second part of the proof is to show that asymptotically we cannot underestimate the number of changepoints if $\beta_n = o(n)$.

Let l be the largest integer less than half the length of the smallest segment. Let $\tilde{E}_n$ be the event that

$$\min_{j \in \{1,\ldots,k_0\}} \min_{\theta} \sum_{i=\tau_j - l + 1}^{\tau_j + l} \{\gamma(Y_i - \theta) - \gamma(Z_i)\} > \alpha' l,$$

for some suitable α' which is specified below. Now

$$\begin{aligned} \min_{\theta'} \sum_{i=\tau_j - l + 1}^{\tau_j + l} \gamma(Y_i - \theta') &= \min_{\theta'} \left\{ \sum_{i=1}^{l} \gamma(\mu_{j-1} + Z_{\tau_j - l + i} - \theta') + \gamma(\mu_j + Z_{\tau_j + i} - \theta') \right\} \\ &= \min_{\theta} \left\{ \sum_{i=1}^{l} \gamma(Z_{\tau_j - l + i} - \theta) + \gamma(Z_{\tau_j + i} - \theta + \Delta_j) \right\} \end{aligned}$$

where μ_{j-1} and μ_j are the segment location parameters before and after the jth changepoint and $\Delta_j = \mu_j - \mu_{j-1}$ is the change at the jth changepoint. We have $|\Delta_j| > 0$ for each j as by assumption $\mu_j \neq \mu_{j-1}$.

If we choose $\alpha' > 0$ to be a value such that the statement of Lemma A.2 holds for all $\Delta = |\Delta_j|$, then, as $l \to \infty$ as $n \to \infty$, we have that $\Pr(\tilde{E}_n) \to 1$ as $n \to \infty$.

Now assume event $\tilde{E}_n$ occurs. Consider a segmentation $\hat{\tau}_{1:\hat{k}}$ with $\hat{k} < k_0$. As $\hat{k} < k_0$ there exists a changepoint τ_j such that no estimated changepoint is within l of τ_j.

We can bound the un-penalised cost of the segmentation $\hat{\tau}_{1:\hat{k}}$ by comparing with a segmentation that also includes changes at $\tau_j - l$, τ_j and $\tau_j + l$:

$$\begin{aligned} Q_0(\hat{\tau}_{1:\hat{k}}) &\geq Q_0(\hat{\tau}_{1:\hat{k}}, \tau_j - l, \tau_j + l) \\ &\geq Q_0(\hat{\tau}_{1:\hat{k}}, \tau_j - l, \tau_j, \tau_j + l) + \min_{\theta} \sum_{i=\tau_j - l+1}^{\tau_j + l} \{\gamma(Y_i - \theta) - \gamma(Z_i)\} \\ &> Q_0(\hat{\tau}_{1:\hat{k}}, \tau_j - l, \tau_j, \tau_j + l) + \alpha' l. \end{aligned}$$

The second inequality comes from a bound on the reduction in the un-penalised cost from adding a changepoint at τ_j. It is obtained using (7) for data $Y_{\tau_j - l+1:\tau_j + l}$ with a change at τ_j.

Thus the penalised cost satisfies

$$Q(\hat{\tau}_{1:\hat{k}}) - Q(\hat{\tau}_{1:\hat{k}}, \tau_j - l, \tau_j + l) > \alpha l - 3\beta_n,$$

where the $3\beta_n$ term comes from adding 3 changepoints.

Now l increases linearly in n, thus the right-hand side is positive for large enough n providing $\beta_n = o(n)$. Thus for large enough n, the segmentation $\hat{\tau}_{1:\hat{k}}$ cannot minimise the penalised cost if $\tilde{E}_n$ occurs. This argument applies for all such segmentations with fewer than k_0 changes. Thus as $\Pr(\tilde{E}_n) \to 1$ we have that $\Pr(\hat{k}_n < k_0) \to 0$ as required.

Part 3

Taken together, the first two parts of the proof show that $\Pr(\hat{k}_n = k_0) \to 1$. The third part of the proof relates to the accuracy of the estimated changepoint locations. As $\Pr(\hat{k}_n = k_0) \to 1$ we need consider only segmentations of the data with k_0 changepoints.

We introduce an event $\bar{E}_n$ similar to $\tilde{E}_n$ used in the second part of the proof, but with $l = \lfloor C_2 \log(n) \rfloor$. Consider this event only for n sufficiently large that l is less

than the smallest segment length. That is $\bar{E}_n$ is the event

$$\min_{j\in\{1,\ldots,k_0\}} \min_{\theta} \sum_{i=\tau_j-l+1}^{\tau_j+l} \{\gamma(Y_i-\theta)-\gamma(Z_i)\} > \alpha' l,$$

where α' is a constant. By the same argument as in Part 2 of the proof, we can choose $\alpha' > 0$ such that, by Lemma A.2, $\Pr(\bar{E}_n) \to 1$.

Now assume both $\bar{E}_n$ and E_n, which was defined in the first part of the proof, occur. Consider any segmentation with k_0 changepoints, $\hat{\tau}_{1:k_0}$, for which

$$\max_{i=1,\ldots,k_0} \left\{ \min_{j=1,\ldots,\hat{k}_n} |\tau_i - \hat{\tau}_j| \right\} > C_2 \log(n).$$

Let i be the index of a changepoint for which

$$\min_{j=1,\ldots,\hat{k}_n} |\tau_i - \hat{\tau}_j| > C_2 \log(n),$$

and let $\tau'_{1:k_0+1} = (\tau_{1:i-1}, \tau_i - l, \tau_i + l, \tau_{i+1:k_0})$. This is the set of all actual changepoints except the ith one together with the two changes at a distance l from τ_i. We will compare the penalised cost of segmentation $\hat{\tau}_{1:k_0}$ with the cost of the true segmentation:

$$\begin{aligned} Q(\hat{\tau}_{1:k_0}) - Q(\tau_{1:k_0}) &= Q_0(\hat{\tau}_{1:k_0}) - Q_0(\tau_{1:k_0}) \\ &\geq Q_0(\hat{\tau}_{1:k_0}, \tau'_{1:k_0+1}) - \sum_{j=1}^{n} \gamma(Z_j), \end{aligned}$$

where the last inequality comes from using (7). We can write the final term as a sum over the $k^* + 1$ segments of the segmentation with change-points given by the union of $\hat{\tau}_{1:k_0}$ and $\tau'_{1:k_0+1}$. This union contains all true changepoints except for τ_i. Thus the k^* segments that do not contain τ_i will contribute a term

$$\min_{\theta} \left\{ \sum_{j=s}^{t} \gamma(Z_j - \theta) - \gamma(Z_j) \right\},$$

for some suitable $t > s$, to this sum. As event E_n holds, this term is bounded below by $-\alpha \log(n)$. The remaining term is of the form

$$\min_{\theta} \sum_{j=\tau_i-l+1}^{\tau_i+l} \{\gamma(Y_j-\theta)-\gamma(Z_j)\}$$

and this is bounded below by $\alpha' l$. Thus we have

$$Q(\hat{\tau}_{1:k_0}) - Q(\tau_{1:k_0}) > \alpha' \lfloor C_2 \log(n) \rfloor - (2k_0 + 1)\alpha \log(n),$$

where the first term on the right-hand side comes from the bound on the term for the segment that include τ_i, the other term is from the $k^* \leq 2k_0 + 1$ other terms. Thus for $C_2 > (2k_0+1)\alpha/\alpha'$ this will be positive for large enough n, and the penalised cost would prefer the true segmentation to $\hat{\tau}_{1:k_0}$. This will hold for all $\hat{\tau}_{1:k_0}$ for which the error in estimating at least one changepoint is greater than $C_2 \log(n)$. Thus if E_n and $\bar{E}_n$ hold then for large enough n all segmentations with k_0 changepoints will have

$$\max_{i=1,\ldots,k_0} \left\{ \min_{j=1,\ldots,\hat{k}_n} |\tau_i - \hat{\tau}_j| \right\} \leq C_2 \log(n),$$

as required. □

We finish by giving the proofs of our two lemmas.

Proof of Lemma A.1.

The proof proceeds in two parts. The first involves considering $\min_\theta S_l(\theta)$ for $-2K \leq \theta \leq 2K$, and the second considers $|\theta| > 2K$. In each case we need to show that there is a sufficiently large α such that the probability of the minimum of $S_l(\theta)$, for the respective ranges of θ, being less than $-\alpha \log(n)$ is bounded by a constant times $n^{-2-\delta}$.

For the first part, we initially give a bound on the probability of small values for $S_l(\theta)$ for a fixed value of θ. To do this we use Bennett's inequality (see Theorem 2.9 in Boucheron et al., 2013) for

$$\tilde{S} = \sum_{i=1}^{l} \{X_i(\theta) - M(\theta)\} = S_l(\theta) - lM(\theta).$$

As $X_i(\theta) \geq -K^2$ and $\mathrm{E}\{X_i(\theta)^2\} = v(\theta)$, Bennett's inequality gives us

$$\Pr(\tilde{S} \leq -t) \leq \exp\left(-\frac{t^2}{2(lv(\theta) + K^2 t/3)} \right).$$

Now the event $S_l(\theta) \leq -\alpha' \log(n)$ is equivalent to the event $\tilde{S} \leq -\alpha' \log(n) - lM(\theta)$, so

$$\begin{aligned}\Pr\left\{S_l(\theta) \leq -\alpha' \log(n)\right\} &\leq \exp\left(-\frac{\{\alpha' \log(n) + lM(\theta)\}^2}{2[lv(\theta) + K^2\{\alpha' n \log(n) + lM(\theta)\}/3]}\right) \\ &= \exp\left\{-\alpha' \log(n)\left[\frac{\alpha' \log(n) + 2lM(\theta) + l^2 M(\theta)^2/\{\alpha' \log(n)\}}{2lv(\theta) + 2K^2\{\alpha' \log(n) + lM(\theta)\}/3}\right]\right\}.\end{aligned}$$

Now there exists a $D > 0$ such that $v(\theta) < DM(\theta)$ for all θ. If we further write $\psi = lM(\theta)/\{\alpha' \log(n)\}$, and note that $\psi \geq 0$, we get

$$\Pr\left\{S_l(\theta) \leq -\alpha' \log(n)\right\} \leq \exp\left\{-\alpha' \log(n)\left(\frac{1 + 2\psi + \psi^2}{2K^2/3 + (2D + 2K^2/3)\psi}\right)\right\}$$

For $\psi \geq 0$, we can lower bound the bracketed term in the exponent by some strictly positive constant. Thus for sufficiently large α' we will have that this probability is less than $n^{-3-\delta}$.

Now the above argument is for $S_l(\theta)$ at a specific value of θ, but we are interested in $\min_{|\theta| \leq 2K} S_l(\theta)$. To deal with this minimisation we use the Lipschitz property of $\gamma(y)$, which gives that

$$|S_l(\theta) - S_l(\theta')| \leq 2Kl|\theta - \theta'|.$$

Thus if we choose $\epsilon > 0$ we can partition the interval $[-2K, 2K]$ into $\lceil 4K^2 n/\epsilon \rceil$ intervals of width at most $\epsilon/(Kn)$. For one such interval, as $l \leq n$,

$$\Pr\left\{\min_{\theta:|\theta-\theta'|<\epsilon/(2Kn)} S_l(\theta) \leq -\alpha \log(n)\right\} \leq \Pr\left\{S_l(\theta') \leq -\alpha \log(n) + \epsilon\right\}.$$

Thus if we choose $\alpha > \alpha'$ such that, for sufficiently large n, $-\alpha \log(n) + \epsilon < -\alpha' \log(n)$, we have that, for sufficiently large n, this probability is less than $n^{-3-\delta}$. We require the bound to hold for all $O(n)$ intervals, and thus we get that

$$\Pr\left(\min_{\theta:|\theta|\leq 2K} S_l(\theta) \leq -\alpha \log(n)\right) \leq C_2 \frac{1}{n^{2+\delta}},$$

for some constant C_2, as required.

We now consider the case where $|\theta| > 2K$. We will use the following bound

$$X_i(\theta) = \gamma(Z_i - \theta) - \gamma(Z_i) \geq \begin{cases} K^2 - Z_i^2 & \text{if } |Z_i| < K, \\ -K^2 & \text{otherwise,} \end{cases}$$

for $|\theta| > 2K$. Let $\tilde{X}_i$ be the random variable defined by the right-hand side of this equation, and $\tilde{S}_l = \sum_{i=1}^{l} \tilde{X}_i$. Then we have

$$\Pr\left\{\min_{\theta:|\theta|>2K} S_l(\theta) < -\alpha\log(n)\right\} \leq \Pr\left\{\tilde{S}_l < -\alpha\log(n)\right\}.$$

Using the notation in assumption (4), where $p = \Pr(|Z_i| > K)$ and, if $\tilde{Z}$ is a random variable whose distribution is that of Z_i conditional on $|Z_i| \leq K$, $\sigma^2 = \mathrm{E}(\tilde{Z}^2)$, we have

$$\mathrm{E}(\tilde{X}_i) \geq K^2(1-2p) - (1-p)\sigma^2,$$

which we will denote by $\tilde{M}$. By assumption (4), $\tilde{M} > 0$. As $|\tilde{X}_i| \leq K^2$ we have $\mathrm{E}(\tilde{X}_i^2) < K^4$.

Thus using Bennett's inequality, and a similar argument to above,

$$\begin{aligned}\Pr\left\{\tilde{S}_l < -\alpha\log(n)\right\} &\leq \exp\left\{\frac{-\{\alpha\log(n) + l\tilde{M}\}^2}{2[lK^4 + K^2\{\alpha\log(n) + l\tilde{M}\}/3]}\right\} \\ &= \exp\left\{-\alpha\log(n)\left[\frac{(1+\psi)^2}{2(D\psi + K^2(1+\psi)/3)}\right]\right\},\end{aligned}$$

where $\psi = l\tilde{M}/\{\alpha\log(n)\}$, and $D = K^4/\tilde{M}$. We can bound from below the bracketed term in the exponent for all $\psi > 0$. Thus we can choose α large enough so that the right-hand side is less than $n^{-2-\delta}$ for all l as required. □

Proof of Lemma A.2.

Fix $\Delta > 0$. Let $Z_1', Z_2', \ldots$ be IID copies of Z_1. Let $X_i'(\theta) = \gamma(Z_i' - \theta) - \gamma(Z_i')$. Then

$$S_l(\theta) + S_l'(\theta - \Delta) = \sum_{i=1}^{l} \{\gamma(Z_i - \theta) - \gamma(Z_i) + \gamma(Z_i' - \theta + \Delta) - \gamma(Z_i')\} = \sum_{i=1}^{l} \{X_i(\theta) + X_i'(\theta - \Delta)\}.$$

As we have fixed Δ, we will write $\bar{X}_i(\theta) = X_i(\theta) + X_i'(\theta - \Delta)$, and $\bar{S}_l(\theta) = \sum_{i=1}^{l} \bar{X}_i(\theta)$. We are thus interested in bounding

$$\Pr\left(\min_{\theta} \bar{S}_l < \alpha l\right).$$

To get the required bound, our argument will closely follow that of Lemma A.1. In particular we will show that we get the required exponential bound on this probability first for the case where we minimise θ over $|\theta| \leq \Delta + 2K$ and second for the case where we minimise θ over $|\theta| > \Delta + 2K$.

By condition (3) we have that, for any θ,

$$\mathrm{E}\{\bar{X}_i(\theta)\} = M(\theta) + M(\theta - \Delta) \geq \min\left\{c_1\frac{\Delta^2}{2}, c_2\right\},$$

and we denote the right-hand side, which is a constant as we have fixed Δ, by $\bar{M}$. As $|\bar{X}_i(\theta)| \leq 2K$ we have $\mathrm{E}\{\bar{X}_i(\theta)^2\} \leq 4K^2$.

Thus, for any chosen θ, we can use Bennett's inequality to bound the lower tail for $\bar{S}_l(\theta)$. We get

$$\begin{aligned}
\Pr\left\{\bar{S}_l(\theta) \leq l\bar{M}/2\right\} &= \Pr\left\{\bar{S}_l(\theta) - l\bar{M} \leq -l\bar{M}/2\right\} \\
&\leq \exp\left\{-\frac{l^2\bar{M}^2}{8(4lK^4 + Kl\bar{M}/3)}\right\} \\
&\leq \exp\left\{-lC_2'\right\},
\end{aligned}$$

where $C_2' = \bar{M}^2/(32K^4 + 8K\bar{M}/3)$.

The above inequality is for a single, fixed, value of θ. To deal with the minimisation over θ such that $|\theta| \leq \Delta + 2K$ we partition this interval into intervals of length $\bar{M}/(8K)$, or smaller. As $\bar{X}_i(\theta)$ is Lipschitz in θ, with constant $4K$, we have that for θ in an interval of length $\bar{M}/(8K)$ the value of $\bar{S}_l(\theta)$ can vary by at most $l\bar{M}/4$ from the value in the centre of the interval. Thus if $\bar{S}_l(\theta) \leq l\bar{M}/2$ then we must have $\bar{S}_l(\theta') \leq l\bar{M}/4$ for all $\theta' \in [\theta - \bar{M}/(8K), \theta + \bar{M}/(8K)]$.

Our partition of the region $|\theta| \leq \Delta + 2K$ requires a fixed number of intervals of length at most $\bar{M}/(8K)$. Thus we have

$$\Pr\left\{\min_{\theta:|\theta|<\Delta+2K} \bar{S}_l(\theta) \leq l\bar{M}/4\right\} \leq C_1' \exp\left\{-lC_2'\right\},$$

where C_1 is the number of intervals.

We now consider the case where we minimise $\bar{S}_l(\theta)$ over $|\theta| > \Delta + 2K$. As in the proof to Lemma A.1, for such θ we can bound from below both $X_i(\theta)$ and $X_i(\theta - \Delta)$ by a random variable that takes the value $K^2 - Z_i^2$ if $|Z_i| < K$ and that takes the value $-K^2$ otherwise. Let $\tilde{X}_i$ denote this random variable, and $\tilde{X}_i'$ the corresponding random variable defined from Z_i'. We then have that for any α

$$\Pr\left\{\min_{\theta:|\theta|>\Delta+2K} \bar{S}_l(\theta) \leq \alpha l\right\} \leq \Pr\left\{\sum_{i=1}^{l}(\tilde{X}_i + \tilde{X}_i') \leq \alpha l\right\}.$$

Defining $\tilde{M}$ to be the lower bound on $\mathrm{E}(\tilde{X}_i)$ that was given in the proof of Lemma A.1 we have, using Bennett's inequality again,

$$\begin{aligned}\Pr\left\{\sum_{i=1}^{l}(\tilde{X}_i + \tilde{X}_i') \leq l\tilde{M}\right\} &= \Pr\left\{\sum_{i=1}^{l}(\tilde{X}_i + \tilde{X}_i' - 2\tilde{M}) \leq -l\tilde{M}\right\} \\ &\leq \exp\left\{-\frac{l^2\tilde{M}}{2(4K^4 l + 2K\tilde{M}l/3)}\right\}.\end{aligned}$$

The final expression can be written as $\exp\{-lC_2''\}$ for some suitable C_2''.

Putting this together with the result when we minimise over $|\theta| \leq \Delta + 2K$ we get the required inequality with $\alpha = \min\{\bar{M}/4, \tilde{M}\}$, $C_1 = C_1' + 1$ and $C_2 = \min\{C_2', C_2''\}$. □

Appendix B Proofs from Section 2

Proof of Theorem 2.1. Consider any segmentation of the data that does not include changepoints at both $t-1$ and t. We will show that for sufficiently large y_t, that adding changepoints at both $t-1$ and t (or at just one of these if the segmentation has a change at the other time-point) will reduce the penalised cost. Thus the optimal segmentation must have changes at both $t-1$ and t.

Let the segment in the original segmentation that contains y_t contain $y_{s:u}$ for $s < t$ and $u > t$. The change in cost between this segmentation and one with changepoints

added at $t-1$ and t will be

$$\min_{\theta} \sum_{i=s}^{t-1} \gamma(y_i;\theta) + \min_{\theta} \sum_{i=t+1}^{u} \gamma(y_i;\theta) + 2\beta - \min_{\theta} \sum_{i=s}^{u} \gamma(y_i;\theta). \tag{8}$$

Here we have used the fact that the only change in cost will be for fitting $y_{u:t}$ as the other segmentations are unchanged. The new segmentation has segments which include $y_{s:t-1}$, y_t and $y_{t+1:u}$ and introduces two extra changepoints. The cost of the segmentation which just includes y_t is 0. We need to show that for large enough y_t this will always be negative. To see this we use the fact that

$$\min_{\theta} \sum_{i=s}^{u} \gamma(y_i;\theta) \geq \min\{\gamma(y_t;(y_t+y_{t+1})/2), \gamma(y_{t+1};(y_t+y_{t+1})/2)\},$$

and this tends to infinity as $y_t \to \infty$. The other terms in (8) do not depend on y_t, and hence (8) will be negative for sufficiently large y_t.

The proof for $s \leq t$ or $u \geq t$ follows similarly. ■

Proof of Theorem 2.2. We will prove this by showing that for any segmentation with a segment which is shorter than β/K, we can reduce the penalised cost by removing the changepoint at either the start or end of the segment.

Consider a segmentation of the data with neighbouring segments $y_{s:t}$ and $y_{t+1:u}$. If either $t-s < \beta/K$ or $u-t-1 < \beta/K$ then removing the changepoint at t will reduce the penalised cost. Without loss of generality assume $t-s < \beta/K$. The change in cost of removing the changepoint at t is

$$\begin{aligned}
&\min_{\theta} \sum_{i=s}^{u} \gamma(y_i;\theta) - \min_{\theta} \sum_{i=s}^{t} \gamma(y_i;\theta) - \min_{\theta} \sum_{i=t+1}^{u} \gamma(y_i;\theta) - \beta \\
&< \quad (t-s)K + \min_{\theta} \sum_{i=t+1}^{u} \gamma(y_i;\theta) - \min_{\theta} \sum_{i=s}^{t} \gamma(y_i;\theta) - \min_{\theta} \sum_{i=t+1}^{u} \gamma(y_i;\theta) - \beta \\
&= \quad (t-s)K - \beta
\end{aligned}$$

The first inequality uses the fact that the cost for segmenting $y_{s:u}$ is less than the cost if we fix θ to the optimal value for segmenting $y_{t+1:u}$. We then bound the

contribution of the cost for each of $y_{s:t}$ by K .The second inequality the fact that the costs are positive. We have that this change in cost is negative, as required, because $(t-s) < \beta/K$. ■

Appendix C Proofs from Section 4

Proof of Theorem 4.1. Given that γ is convex then, following the proof of the worst case complexity of the pDPA (Rigaill (2015) appendix A), we get that the function $Q_t(\theta)$ can be described in at most $2t-1$ intervals such that for each interval there is a single value for the best time of the most recent changepoint. That is we can define intervals I_k for $k = 1, \ldots, K$, with $K < 2t$, such that for a given k there exists s such that

$$\forall \theta \ \in I_k = [s_k, e_k] \qquad Q_s + \sum_{i=s+1}^{n} \gamma(y_i; \theta) + \beta = Q_t(\theta).$$

All $\tilde{c}_{t+1:n}(\theta)$ are themselves defined in pieces, as sums of $\gamma(y_i, \theta)$. For each $\gamma(y_i, \theta)$ we need to consider L intervals and thus $L-1$ points between intervals. We call these points $T_{i,j}$ for j in 1 to $L-1$. For a given $I_k = [s_k, e_k]$ define N_k to be the number of the $T_{i,j}$ points that are in the open interval (s_k, e_k). As all I_k are disjoint then each $T_{i,j}$ can only appear in a single interval. So $\sum_{k=1}^{K} N_k \leq t(L-1)$.

On I_k R-FPOP will thus define $Q_t(\theta)$ using $N_k + 1$ intervals. Thus R-FPOP uses

$$\sum_{k=1}^{K} (N_k + 1) \leq t(L-1) + 2t - 1$$

intervals to define $Q_t(\theta)$ ■.

Proof of Corollary 4.2. The space complexity is obtained using the fact that the number of intervals is bounded by $2n-1+n(L-1)$ and the fact that on each interval we need to store a quadratic.

As for the time complexity, at step t R-FPOP needs to consider at most $2t-1+t(L-1)$ ordered intervals. The key to bounding the time-complexity is that we can split up the operations at step t of R-FPOP into a series of operations on each interval. The cost for each operations on an interval is $\mathcal{O}(1)$, and as there are $\mathcal{O}(t)$ intervals the overall cost of the t iteration is $\mathcal{O}(t)$. The details are as follows.

On each of these intervals we will compute the roots of $Q_t(\theta)$ minus a constant function. According to the number of roots the interval will be split in at most three (because on each interval $Q_t(\theta)$ is convex). Calculating these roots can be done in $\mathcal{O}(1)$ for each interval. Thus we get a new ordered list of intervals in $\mathcal{O}(t)$ time. Iterating on this list, successive intervals having the same analytical decomposition can be fused in $\mathcal{O}(t)$ time. Once this is done, it is possible to add $\gamma(y_t, \mu)$ to $Q_t(\mu)$ on all intervals in $\mathcal{O}(t)$ time. Finally the minimum of $Q_{t+1}(\theta)$ on each interval is recovered in $\mathcal{O}(t)$ time. Overall, step t is performed in $\mathcal{O}(t)$. Summing over all t we get a quadratic complexity ■

Proof of Theorem 4.3. From observations up to and including time t, the biweight loss function will define (at most) $2t + 1$ intervals which are separated by points of non-diffentiability of the loss function, $y_i - K$ and $y_i + K$ for $i = 1, \ldots, t$. Denote these intervals as I_k, $k = 1, \ldots, 2t + 1$. Following the proof of the worst case complexity of the pDPA (Rigaill (2015) appendix A) we see that R-FPOP will need at most $2t - 1$ intervals to describe $Q_t(\theta)$ on any given I_k. Summing over k we recover a quadratic complexity. ■

Proof of Corollary 4.4 The proof followa the proof of corollary 4.2 replacing the $\mathcal{O}(n)$ bound on the number of intervals by the $\mathcal{O}(n^2)$ bound on the number of intervals. ■

Appendix D Pseudo Code for R-FPOP

Here we provide some pseudo-code of the algorithm. In FPOP the algorithm is working on candidate change-points τ. Each of those change can be associated to one or more intervals on which it is optimal. In R-FPOP we directly work on these intervals. R-FPOP essentially relies on three sub-routines or sub-algorithms that manipulate the functions $Q_t^*(\theta)$ and $Q_t(\theta)$:

- Sub-routine 2 to compute the function $Q_t(\theta) = Q_t^*(\theta) + \gamma(y_t, \theta)$;
- Sub-routine 3 to recover the minimum and best change of the function $Q_t(\theta)$;
- Sub-routine 4 to compare the function $Q_t(\theta)$ to a constant and recover the function $Q_{t+1}^*(\theta)$.

Using these sub-routines the pseudo-code of R-FPOP is fairly simple and provided below in Algorithm 1. We then provide the pseudo-code of each sub-routine in Algorithms 2, 3 and 4.

Algorithm 1: Robust FPOP algorithm

Input : Set of data of the form $\mathbf{y}_{1:n} = (y_1, \ldots, y_n)$,

A measure of fit $\gamma(\cdot, \cdot)$ dependent on the data and the mean,

A penalty β which does not depend on the number or location of the changepoints.

Set $Q_1^*(\theta) = 0$ on the interval $(\min_i\{y_i\}, \max_i\{y_i\}]$;

for $t = 1, \ldots, n$ **do**

Compute $Q_t(\theta) = Q_t^*(\theta) + \gamma(y_t, \theta)$ using sub-routine 2 ;

Compute Q_t and τ_t the min and $\arg\min$ of $Q_t(\theta)$ using sub-routine 3 ;

Set $cp(t) = (Q_t, \tau_t)$;

Compute $Q_{t+1}^*(\theta)$ by comparing the function $Q_t(\theta)$ to $Q_t + \beta$ using sub-routine 4 ;

Output: The changepoints recorded in $cp(n)$.

Appendix E ROC curve for all tumor fractions

References

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Rigaill, G. (2015). A pruned dynamic programming algorithm to recover the best segmentations with 1 to $K_{\max}$ change-points. *Journal de la Société Française de Statistique*, 156(4):180–205.

Algorithm 2: Sub-routine to compute $Q_t(\theta) = Q_t^*(\theta) + \gamma(y_t, \theta)$

Input : $Q_t^*(\theta)$ a function defined on N_t^* intervals: $(a_i^{(t)}, b_i^{(t)}]$

For each interval we also have a change: $\tau_i^{*(t)}$

$\gamma(y_t, \theta)$ a function defined on l intervals: $(c_j^{(t)}, d_j^{(t)}]$

We assume $a_1^{(t)} = c_1^{(t)}$ and $b_{N_t^*}^{(t)} = d_l^{(t)}$

Set current number of intervals for $Q_t(\theta)$ to $N_t = 0$;

Set current Q_t^* interval to $i = 1$;

Set current $\gamma(y_t, \theta)$ interval to $j = 1$;

while $i \leq N_t^*$ *and* $j \leq l$ **do**

$N_t = N_t + 1$;

Create the new interval $(A_{N_t}^{(t)}, B_{N_t}^{(t)}] = \left(\max\{a_i^{(t)}, c_j^{(t)}\}, \min\{b_i^{(t)}, d_j^{(t)}\}\right]$;

For θ in interval $(A_{N_t}^{(t)}, B_{N_t}^{(t)}]$ set $Q_t(\theta) = Q_t^*(\theta) + \gamma(y_t, \theta)$;

Set $\tau_{N_t}^{(t)} = \tau_i^{*(t)}$;

if $B_{N_t}^{(t)} = b_i^{(t)}$ **then**

$i = i + 1$;

if $B_{N_t}^{(t)} = d_j^{(t)}$ **then**

$j = j + 1$;

Output: The function $Q_t(\theta) = Q_t^*(\theta) + \gamma(y_t, \theta)$ defined on N_t intervals:

$(A_i^{(t)}, B_i^{(t)}]$

Algorithm 3: Sub-routine to recover the minimum Q_t and best change τ_t of $Q_t(\theta)$

Input : $Q_t(\theta)$ a function defined on N_t intervals: $(A_i^{(t)}, B_i^{(t)}]$

For each $(A_i^{(t)}, B_i^{(t)}]$ we also have an associated change: $\tau_i^{(t)}$

Set $Q_t = \infty$;

Set $\tau_t = 0$;

while $i \leq N_t$ **do**

On the intervals $(A_i^{(t)}, B_i^{(t)}]$ recover $m = \min\{Q_t(\theta)\}$;

if $m < Q_t$ **then**

Set $Q_t = m$;

Set $\tau_t = \tau_i^{(t)}$

$i = i + 1$;

Output: Q_t and τ_t

Algorithm 4: Sub-routine to compare $Q_t(\theta)$ to a constant function and recover $Q^*_{t+1}(\theta)$

Input : $Q_t(\theta)$ a function defined on N_t intervals: $(A_i^{(t)}, B_i^{(t)}]$

For each $(A_i^{(t)}, B_i^{(t)}]$ we have an associated change: $\tau_i^{(t)}$

C a constant function $(= Q_t + \beta)$

Set current number of intervals for $Q^*_{t+1}(\theta)$ to $N^*_{t+1} = 0$;

while $i \leq N_t$ **do**

Find the n_t roots of $(Q_t(\theta) - C)$ in the intervals $(A_i^{(t)}, B_i^{(t)})$;

Sort the n_t roots and store them in a vector R_{tmp} ;

Create the vector $R = (A_i^{(t)}, R_{tmp}, B_i^{(t)})$;

for $j = 1$ *to* $j = n_t + 1$ **do**

$N^*_{t+1} = N^*_{t+1} + 1$;

Create a new interval $\left(a^{(t+1)}_{N^*_{t+1}}, b^{(t+1)}_{N^*_{t+1}}\right] = (R_j, R_{j+1}]$;

if $Q_t(\theta) \geq C$ *on* $(R_j, R_{j+1}]$ **then**

For θ in $(R_j, R_{j+1}]$ set $Q^*_t(\theta) = C$;

Set $\tau^{*(N^*_{t+1})}_{t+1} = t$;

else

For θ in $(R_j, R_{j+1}]$ set $Q^*_t(\theta) = Q_t(\theta)$;

Set $\tau^{*(N^*_{t+1})}_{t+1} = \tau^{(i)}_t$;

$i = i + 1$;

Output: The function $Q^*_{t+1}(\theta)$ defined on N^*_{t+1} intervals: $(a_i^{(t+1)}, b_i^{(t+1)}]$

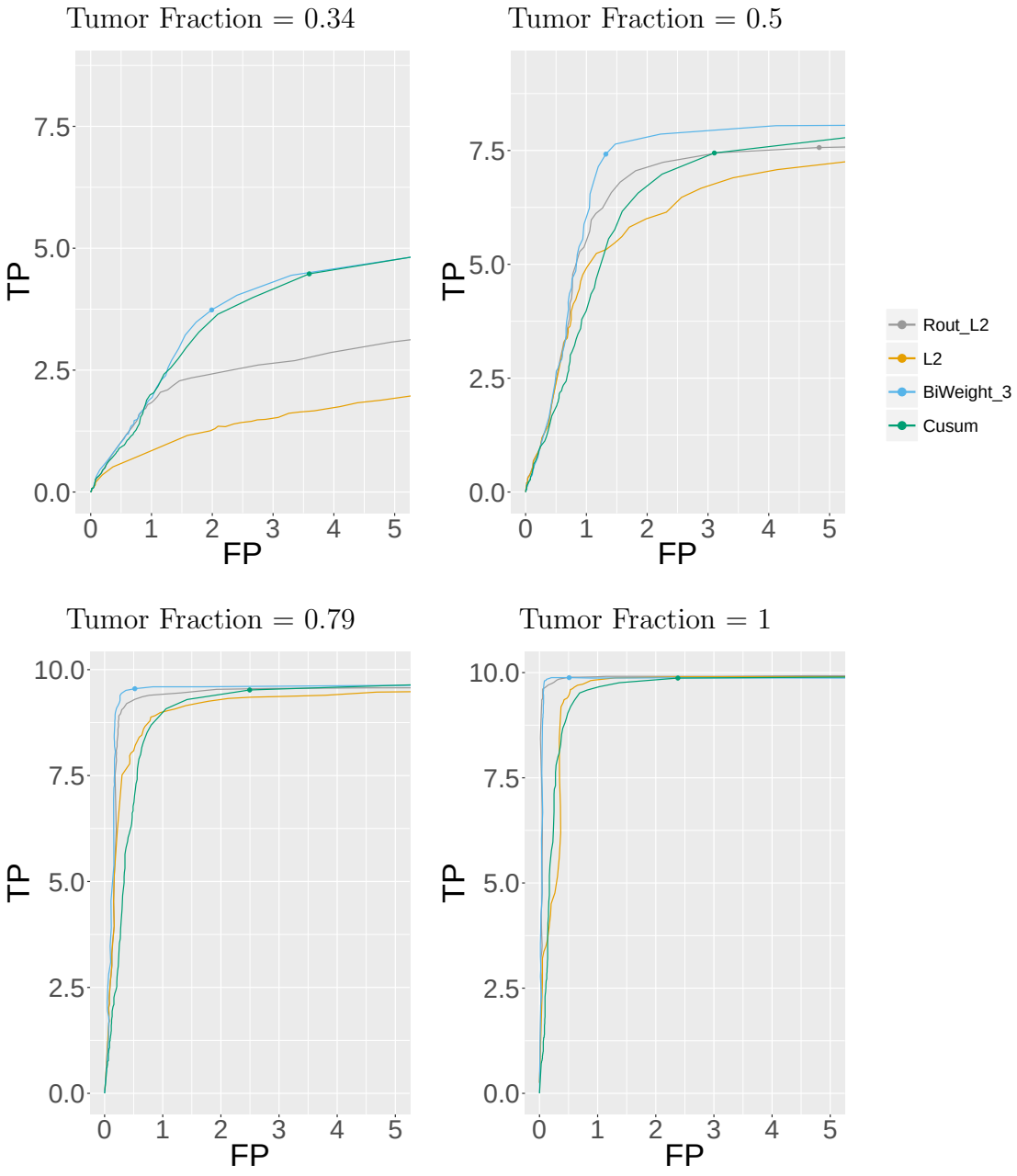

Figure 12: Average ROC on the GSE11976 datasets for the Cusum, L2, L2 with outlier removal (Rout L2) and our robust biweight loss (Biweight 3) for four tumor fraction.

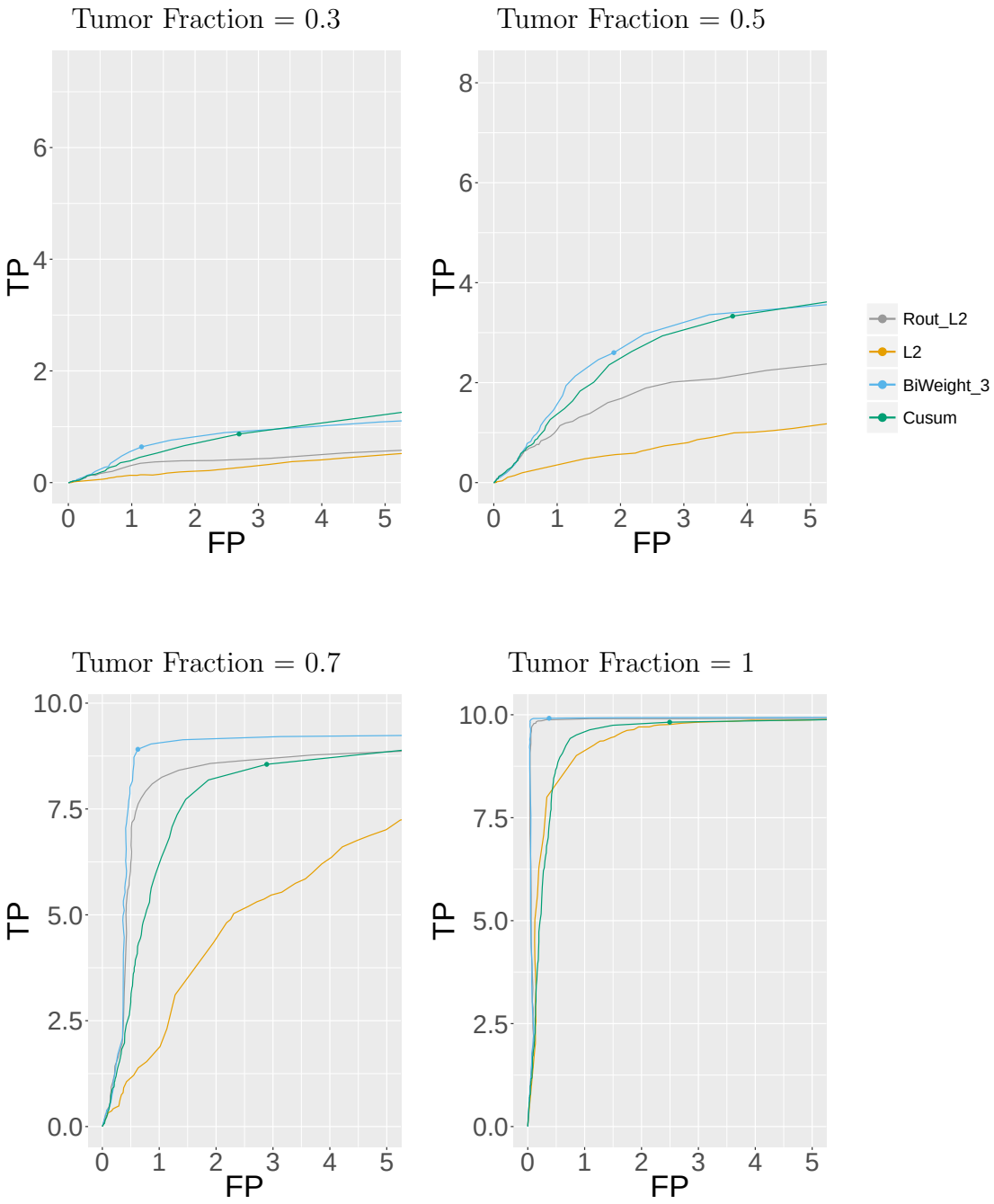

Figure 13: Average ROC on the GSE29172 datasets for the Cusum, L2, L2 with outlier removal (Rout L2) and our robust biweight loss (Biweight 3) for four tumor fraction.