Changepoints and Outliers

Paul Fearnhead

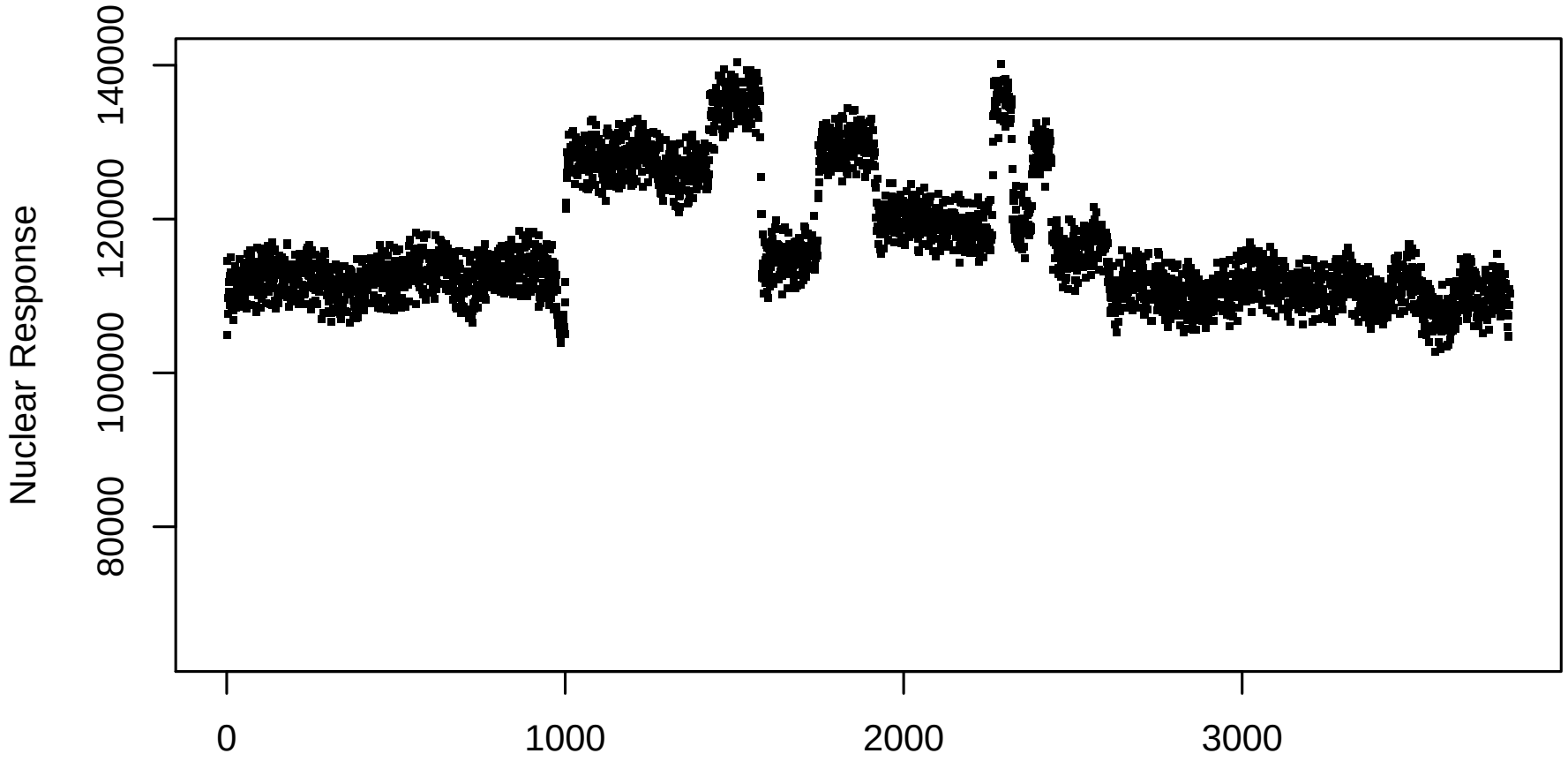
Department of Mathematics & Statistics, Lancaster University
Joint work with Guillem Rigaill

15th November

StatScale | Lancaster University

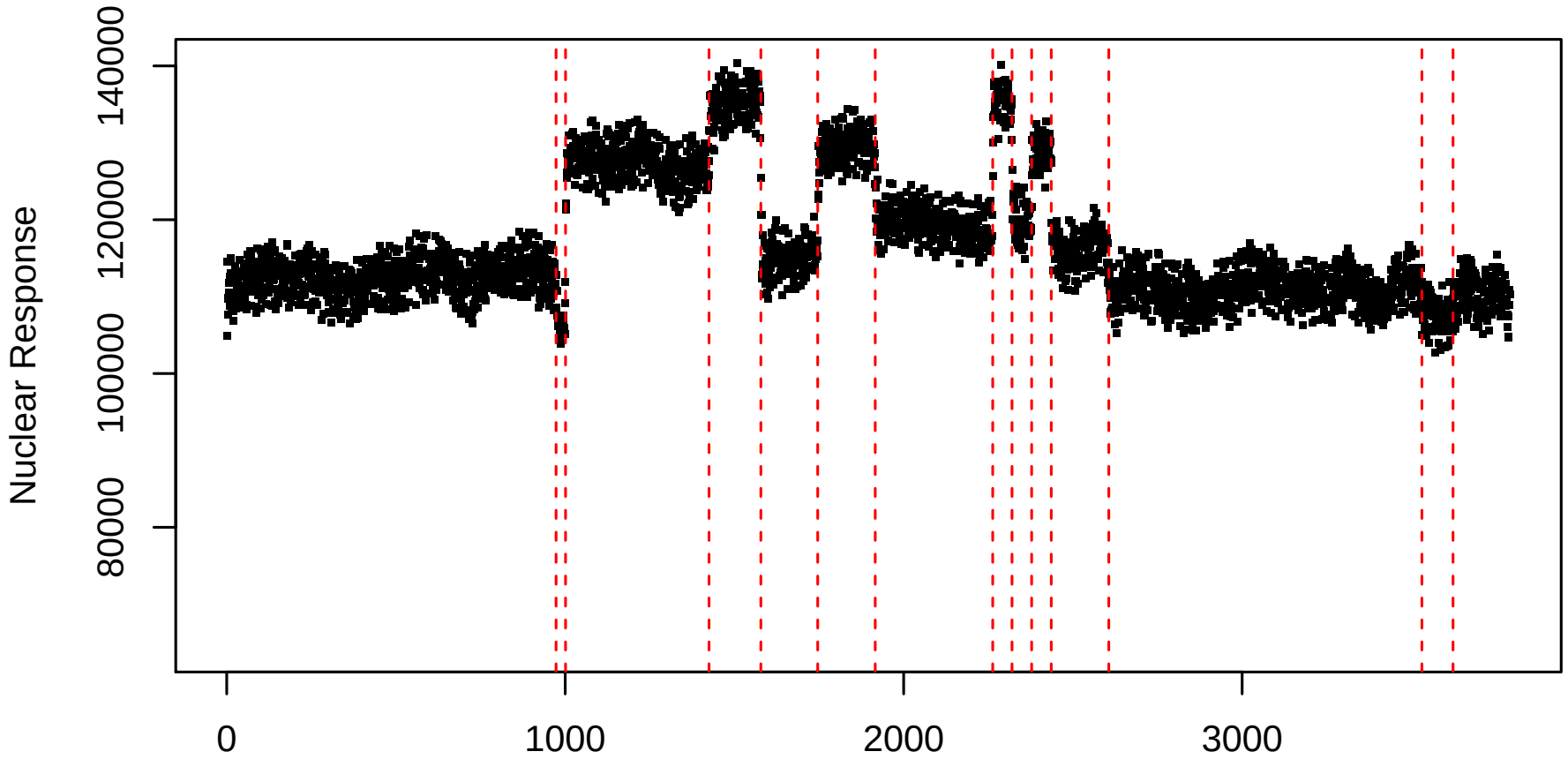
MOTIVATION

Motivation: Well-log data

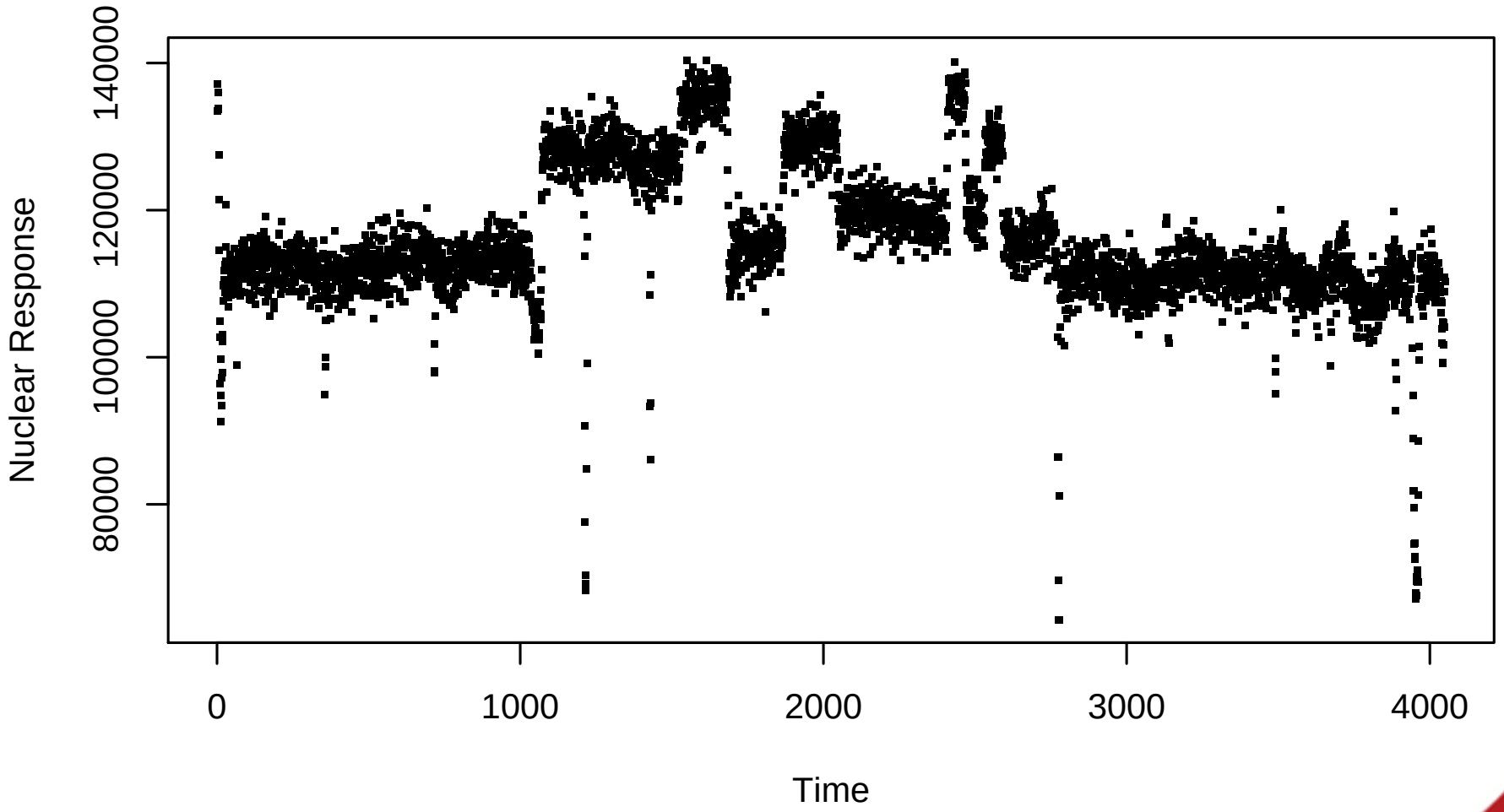


Measurements as probe lowered through bore hole.

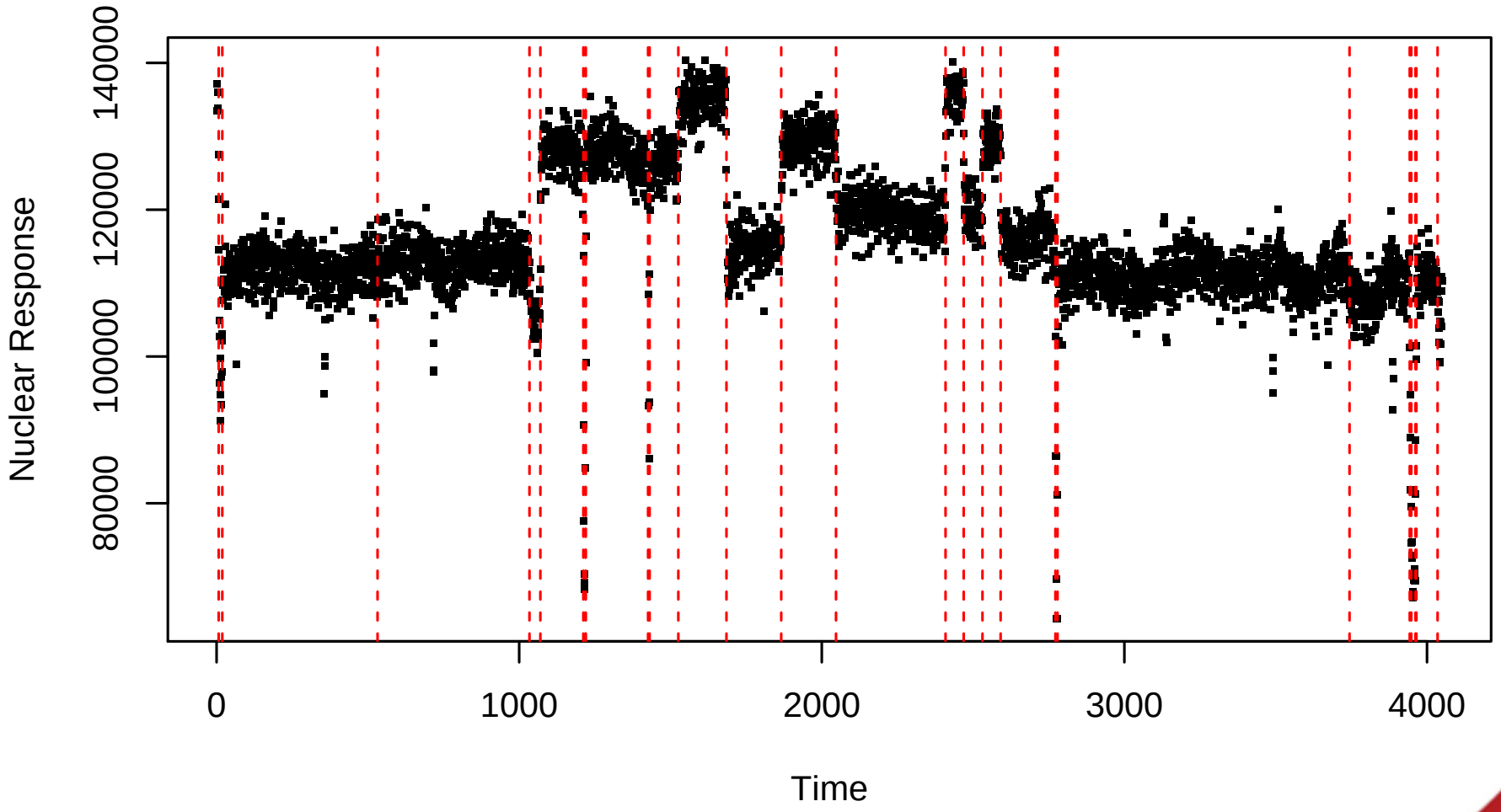
Motivation: Well-log data



Motivation: Well-log data



Motivation: Well-log data



Motivation: Well-log data

StatScale | Lancaster University

Can we develop statistical methods that can reliably distinguish between changes we are interested in and outliers?

And also detect these changes for streaming data?

So we need online algorithms that scale to well to large data (long/high-frequency time-series).

DISTINGUISHING CHANGES AND OUTLIERS

Detecting Changepoints

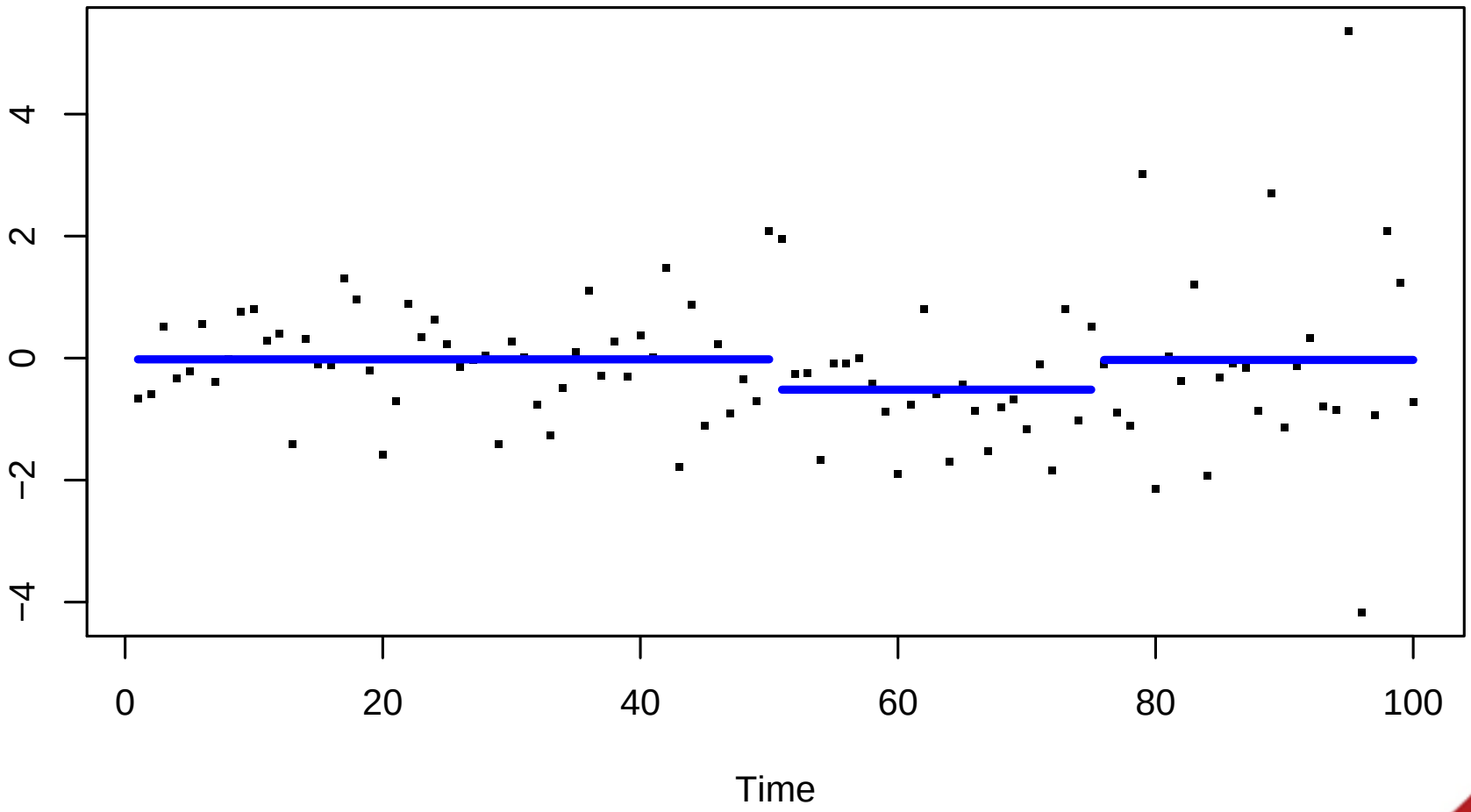
StatScale | Lancaster University

For our "change-in-mean" problem, detecting changepoints is equivalent to fitting a piecewise-constant function to the data.

The changepoints correspond to where the piecewise-constant function changes.

We can define (parameterise) a piecewise-constant function by the number of changes, m, the changepoints, $\tau_{1:m} = (\tau_1, \ldots, \tau_m)$, and the value of the function for each segment $\theta_{0:m} = (\theta_0, \ldots, \theta_m)$.

Example



Detecting Changepoints

A common approach to detecting changepoints is to find the "best" piecewise-constant function that fits the data.

"Best" is defined in terms minimising a cost:

"Fit to Data" + "Penalty for Complexity"

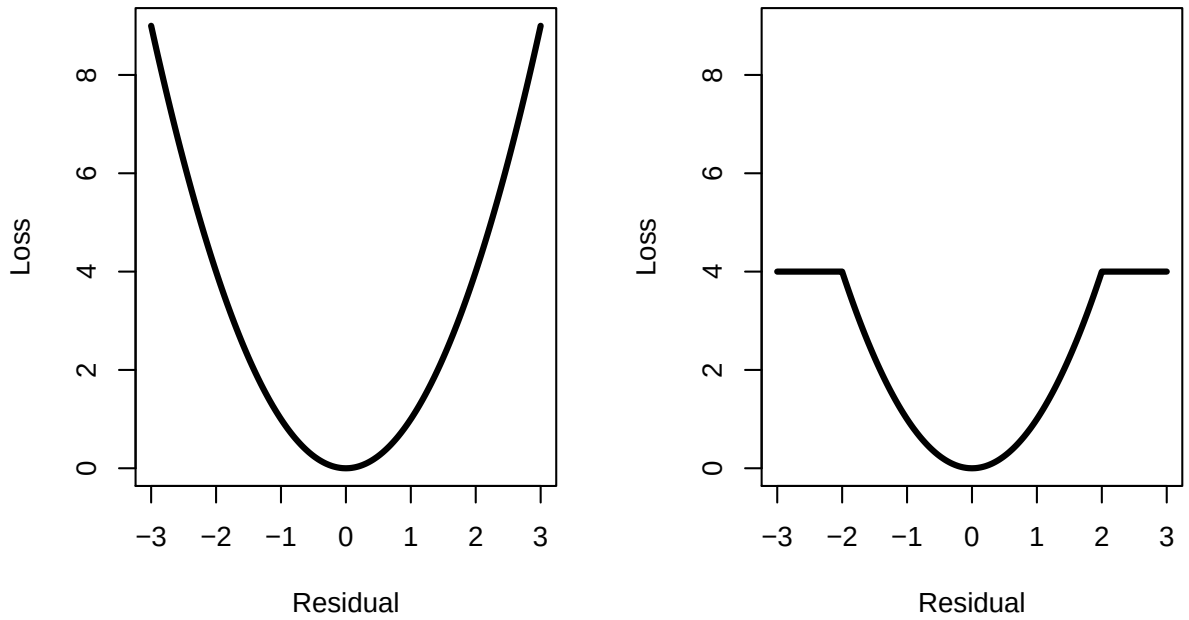
The most common approach minimises the following cost:

$$\min_{m,\tau_{1:m},\theta_{0:m}} \left\{ \sum_{i=0}^{m} \left[\sum_{t=\tau_i+1}^{\tau_{i+1}} (y_t - \theta_i)^2 \right] + \beta m \right\},$$

for some chosen constant β.

Robustness to outliers

The lack of robustness to outliers is due to measuring fit through a quadratic loss (left).



By changing the loss function (right), we can produce a more robust approach.

Detecting Changepoints

StatScale | Lancaster University

So define an outlier loss-function for residual z:

$$\ell(z) = \min\{K, z^2\},$$

for some constant K.

Then detect changepoints by solving the following optimisation problem:

$$\min_{m,\tau_{1:m},\theta_{0:m}} \left\{ \sum_{i=0}^{m} \left[\sum_{t=\tau_i+1}^{\tau_{i+1}} \ell(y_t - \theta_i) \right] + \beta m \right\}.$$

StatScale | Lancaster University

ONLINE ALGORITHM

Solving the Optimisation Problem

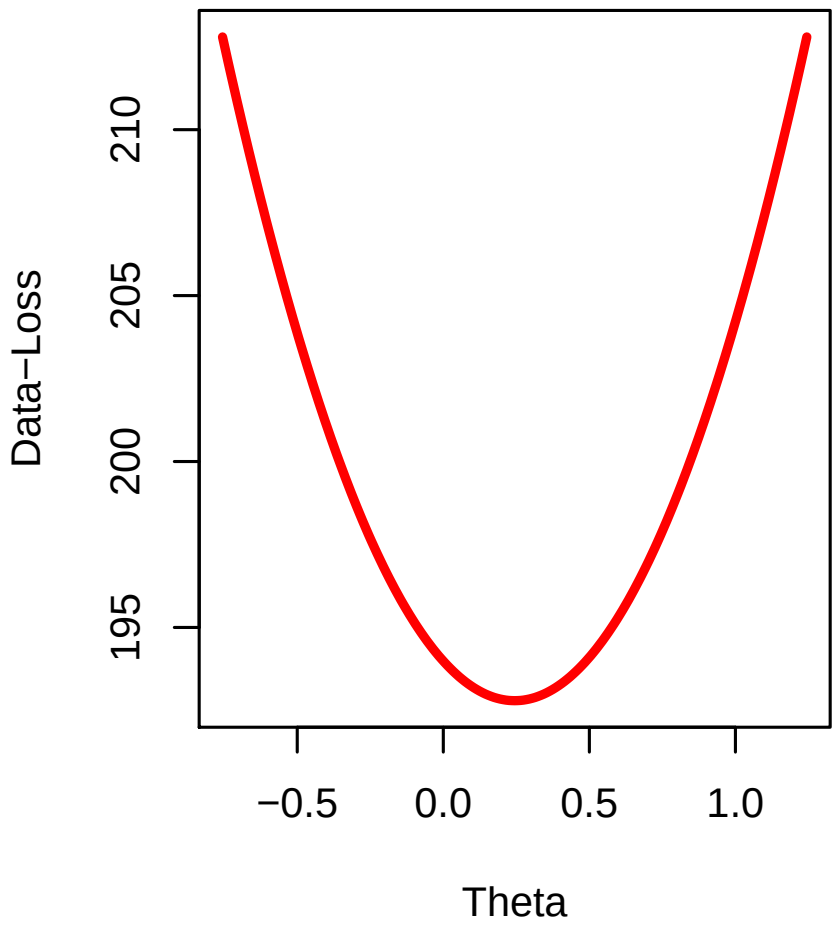
StatScale | Lancaster University

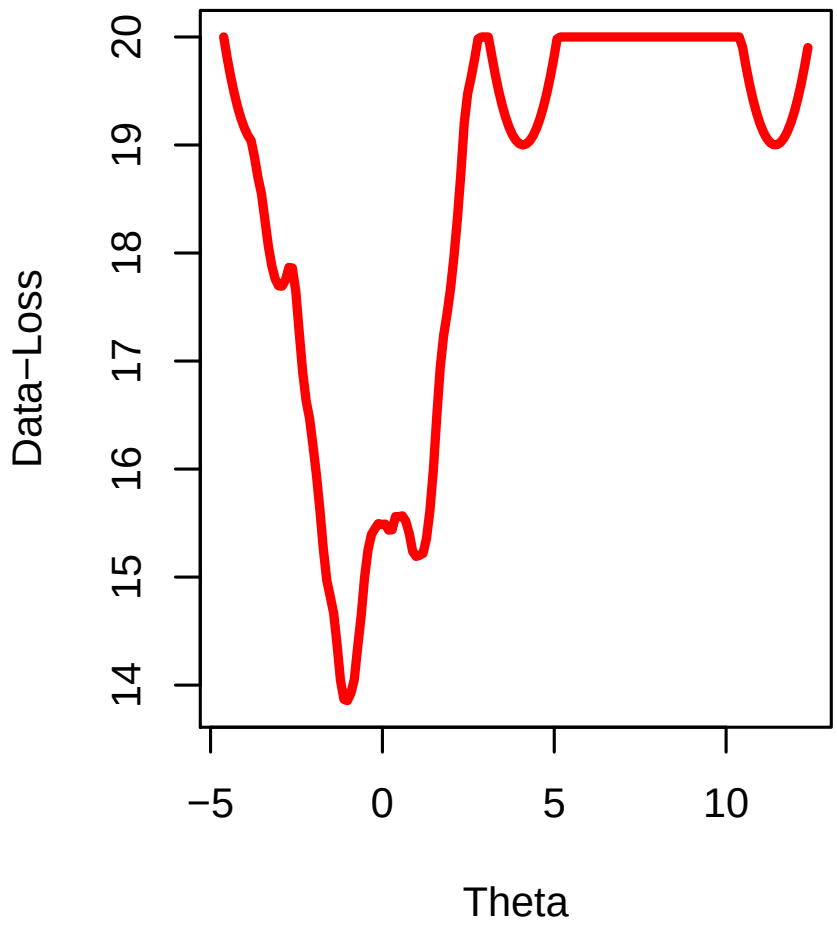
This outlier-loss function is not convex.

That makes the optimisation problem more challenging.

To see the effect, consider the simplest case of no-changepoints. We then need to just estimate the sample mean, θ.

Minimising Outlier Loss





Example of the overall loss when using the quadratic loss (left) and outlier loss (right).

Solving the Optimisation Problem

StatScale | Lancaster University

Our approach is to develop a dynamic programming algorithm for solving:

$$\min_{m,\tau_{1:m},\theta_{0:m}} \left\{ \sum_{i=0}^{m} \left[\sum_{t=\tau_i+1}^{\tau_{i+1}} \ell(y_t - \theta_i) \right] + \beta m \right\}.$$

The key property of this problem that we will use is that if we know the value of the piecewise-constant function at some time t, then we can separately solve the minimisation problem for the data before and after t.

Given this information we can segment data after t without needing to know anything about the data before t.

Dynamic Programming Recursion

StatScale | Lancaster University

So we let $F_t(\phi)$ be the minimum value of our cost for segmenting $y_{1:t}$ given that the piecewise-constant function at time t takes the value ϕ.

Then the value of the minimisation problem for segmenting $y_{1:t}$ is

$$F_t = \min_{\phi} F_t(\phi).$$

Using the separability property we get the following recursion

$$F_{t+1}(\phi) = \min\{F_t(\phi), F_t + \beta\} + \ell(y_{t+1} - \phi)$$

[If we solve this recursion, we get "for-free" the value of the most-recent changepoint and most-recent segment-mean in the optimal segmentation.]

Dynamic Programming Recursion

StatScale | Lancaster University

Solving the dynamic programming recursion for $F_t(\phi)$ is non-trivial.

The key insight is that (assuming $\ell(\cdot)$ is piecewise quadratic) $F_t(\phi)$ is piecewise quadratic.

Hence it is low-dimensional: we can define it by storing a set of intervals and the corresponding quadratic function that $F_t(\phi)$ is equal to on that interval.

We can then derive recursions for these intervals and quadratics.

Algorithm Properties

The dynamic programming algorithm is sequential: hence suitable for online analysis.

The computational cost depends on the dimension of $F_t(\phi)$, and how it changes as t increases.

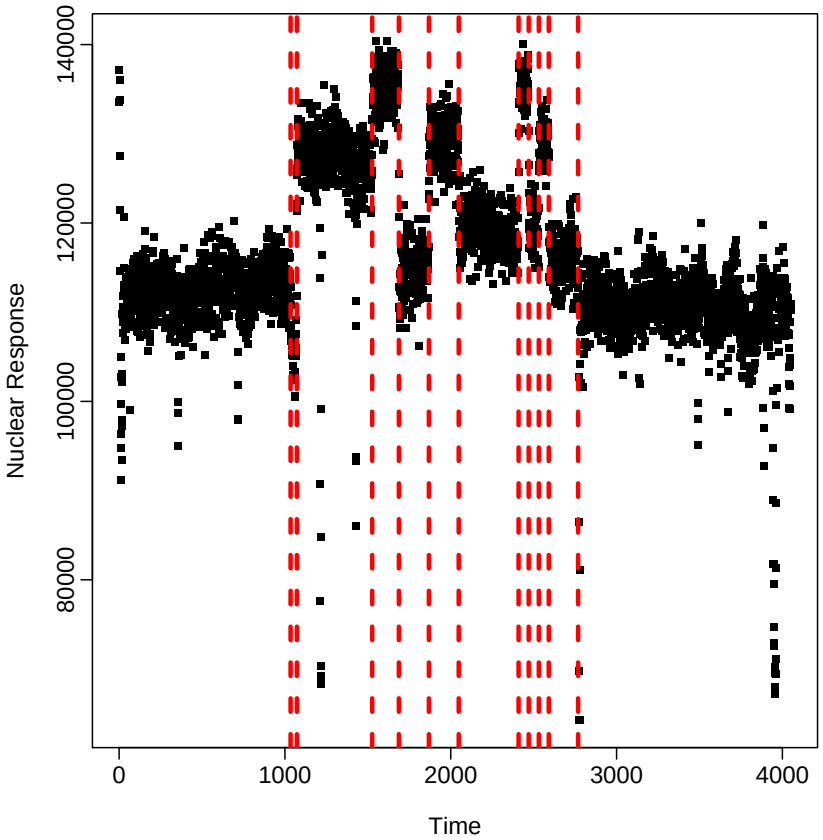
In practice we observe that $F_t(\phi)$ has relatively few intervals (≈ 10) and, on average, this does not increase with t.

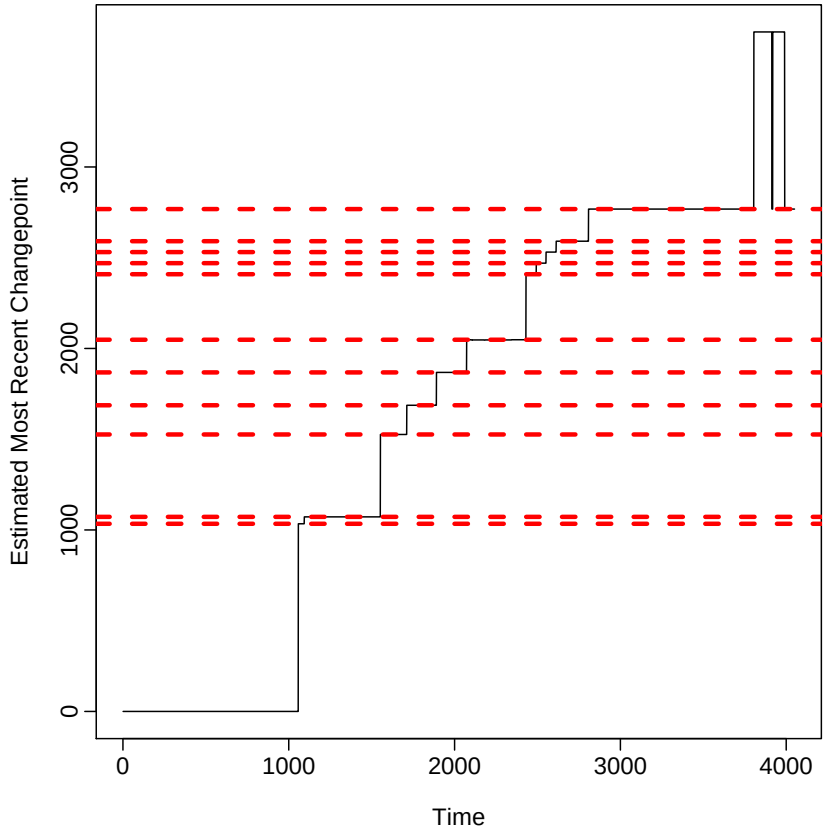
Hence, empirically, we observe the algorithm has a (low) computational cost that is linear in the length of time-series, n.

StatScale | Lancaster University

APPLICATIONS

Well-log Data





Copy Number Variation

Another common application of change-in-mean methods is to detect copy number variation.

Healthy human cells have two copies of DNA (e.g. a gene); tumor cells have deletions/amplifications that mean they have fewer/more copies.

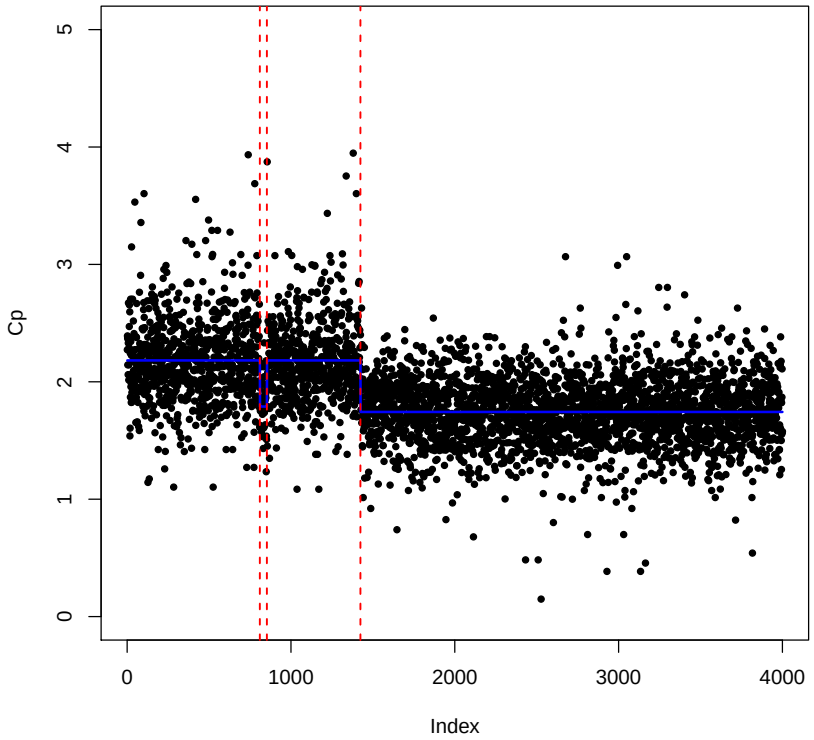
Often samples will consist of a mix of healthy and tumor cells; and data (measurement of copy number along a chromosome) will have outliers.

We tested our robust changepoint approach on some benchmark data.

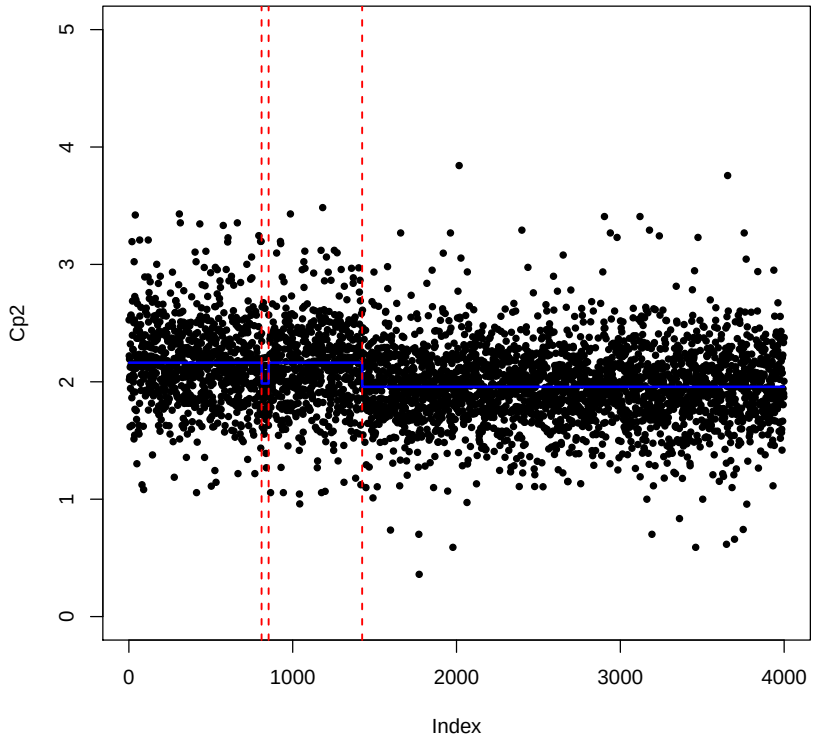
Copy Number Variation

StatScale | Lancaster University

Tumor Fraction = 1



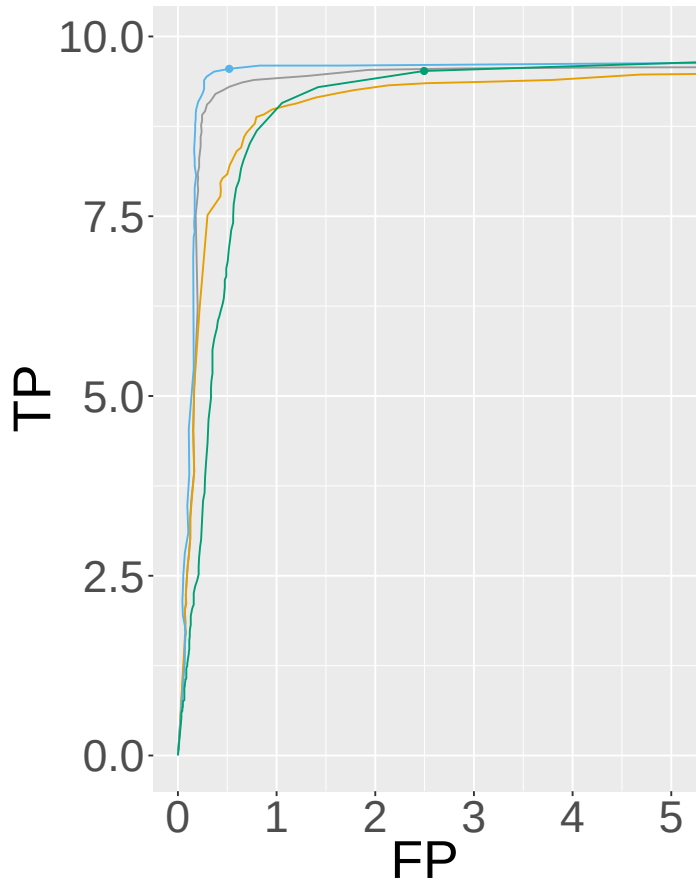
Tumor Fraction = 0.5



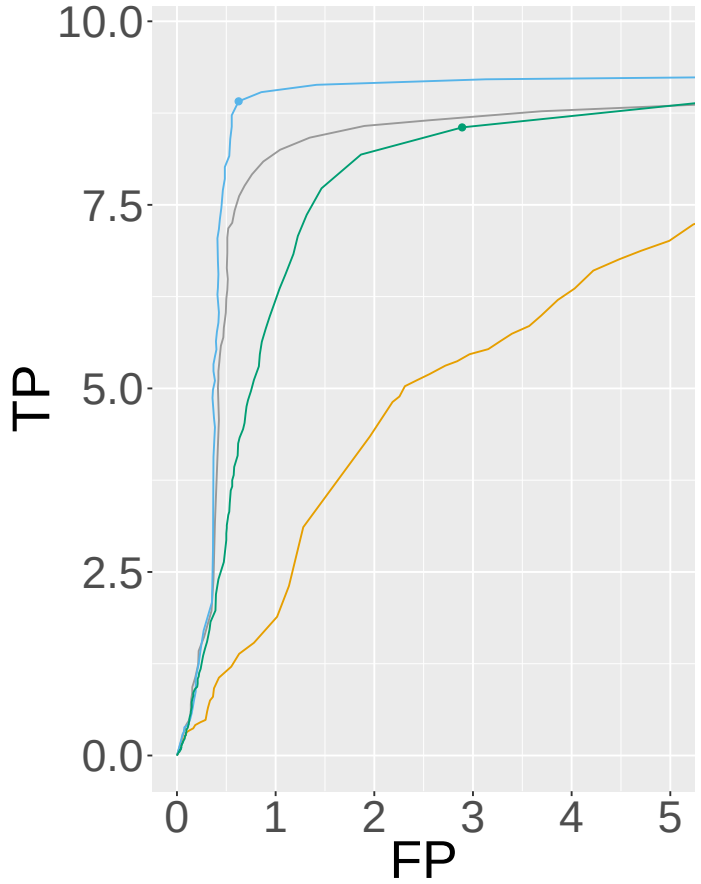
Example of parts of data sets with different tumor fractions.

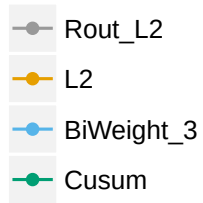
Copy Number Variation

GSE11976 (TF=0.79)



GSE29172 (TF=0.7)





ROC Curve for different methods.

WiFi Tampering

StatScale | Lancaster University

Many secuity devices (e.g. surveillance systems) use WiFi to communicate. 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. Tampering will result in changes in this channel state information.

However other factors (e.g. people moving near the device) may also lead to temporary changes in the channel state information (outliers!).

WiFi Tampering

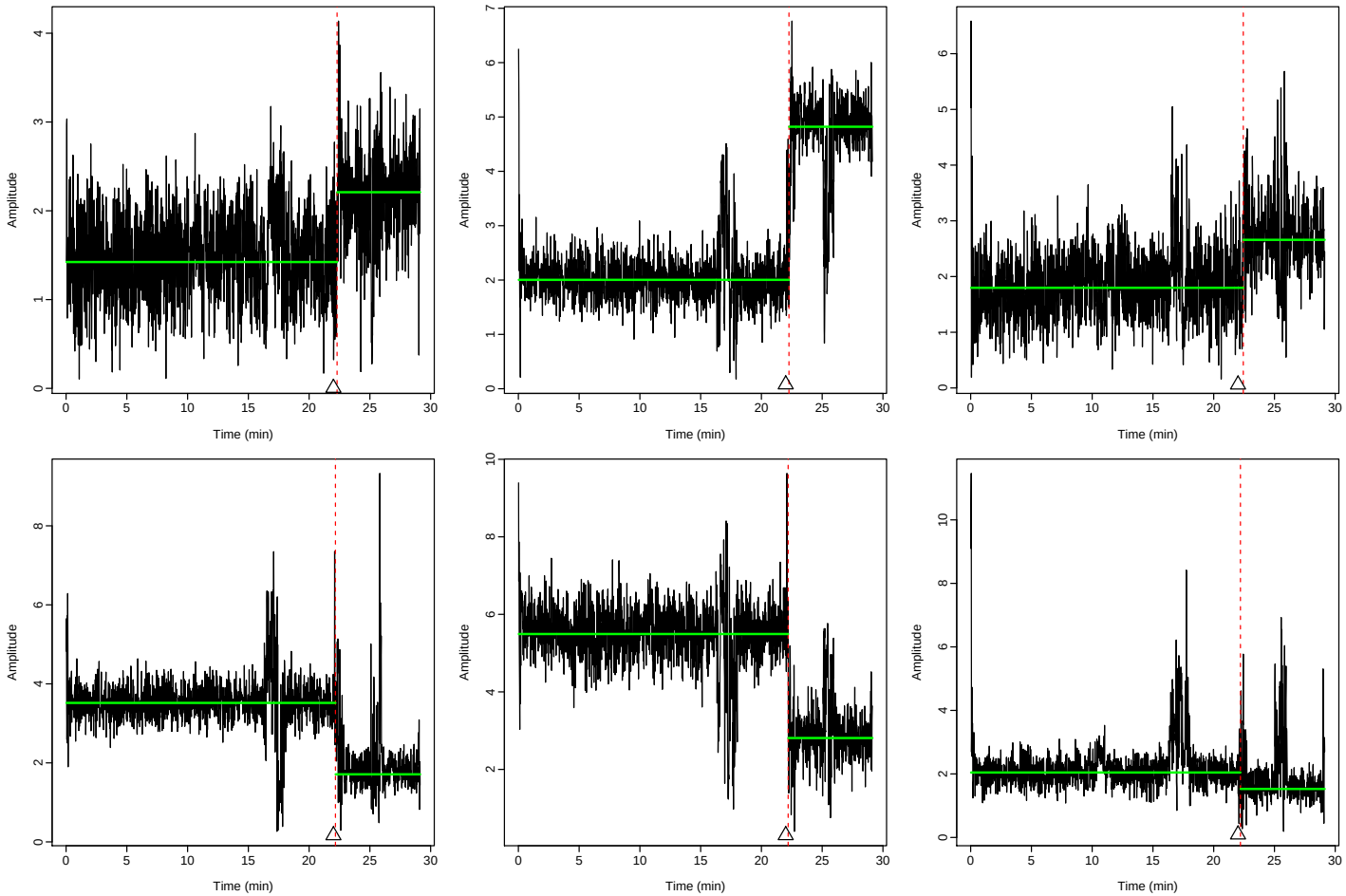
Colleagues at Lancaster have conducted a controlled experiment.

They set up a wireless device in an office.

At known time-points they tampered with the device.

Can we detect these tampering events and distinguish them from “outliers”.

WiFi Tampering



CONCLUSION

Looking Forward

Applications have pointed to open challenges:

- How do we efficiently detect changes in highly multivariate data?
- How do we assess uncertainty in the changepoint locations?
- How do we make methods robust to local-dependence within a segment?
- We need to do all these with scalable algorithms.

Full details are availble: arXiv:1609.07363.
Code is available `https://github.com/guillemr/robust-fpop`