School of Mathematical Sciences | Lancaster University

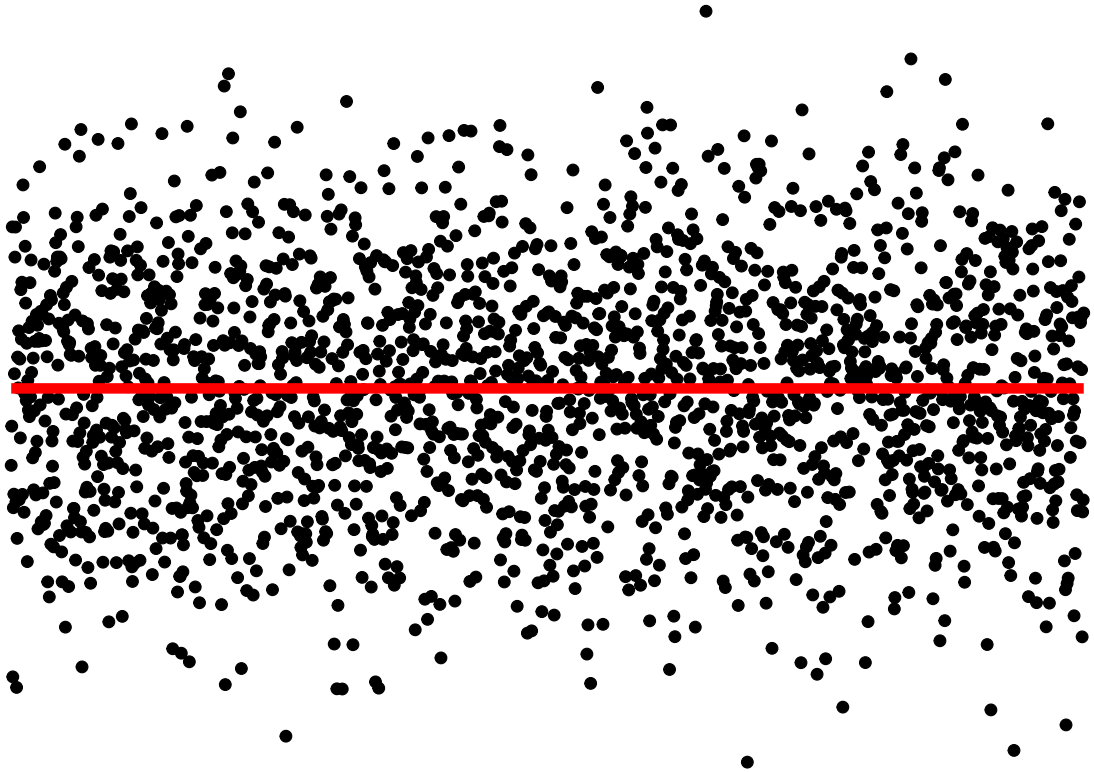
The Art of the Pivot: Detecting Shifts in your Data Streams

Rebecca Killick(r.killick@lancs.ac.uk)
DSAA Oct 2025

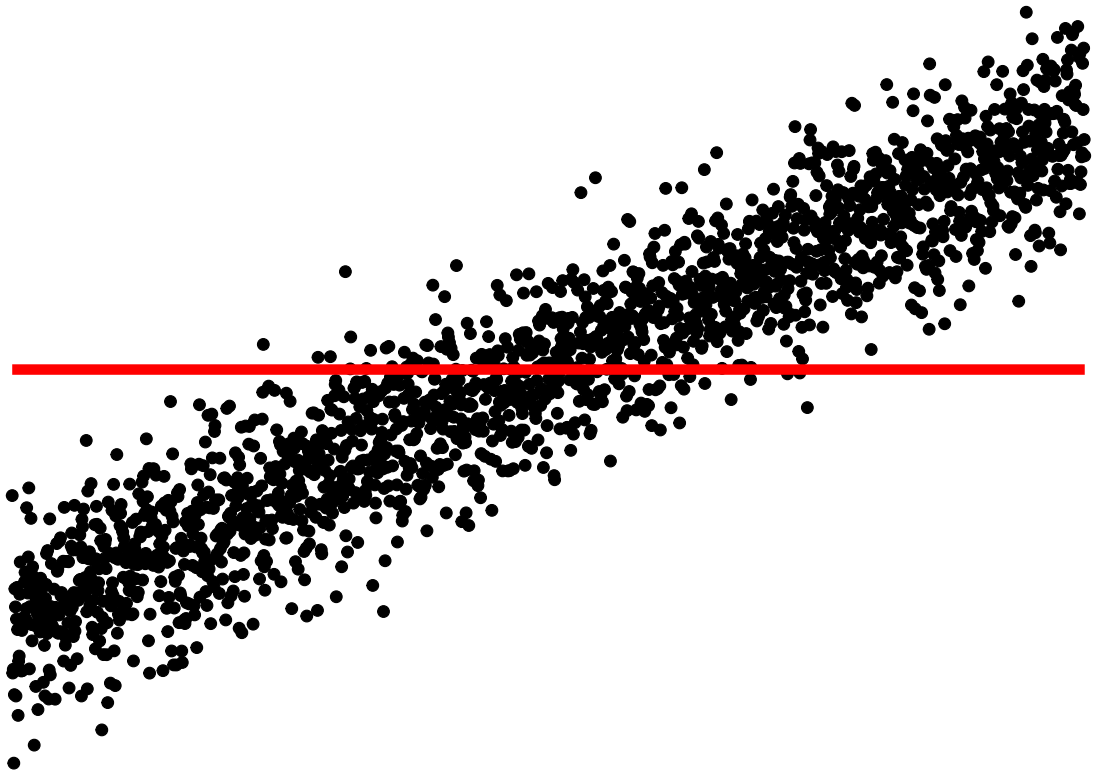
In collaboration

Over 140 co-authors over 13 years.

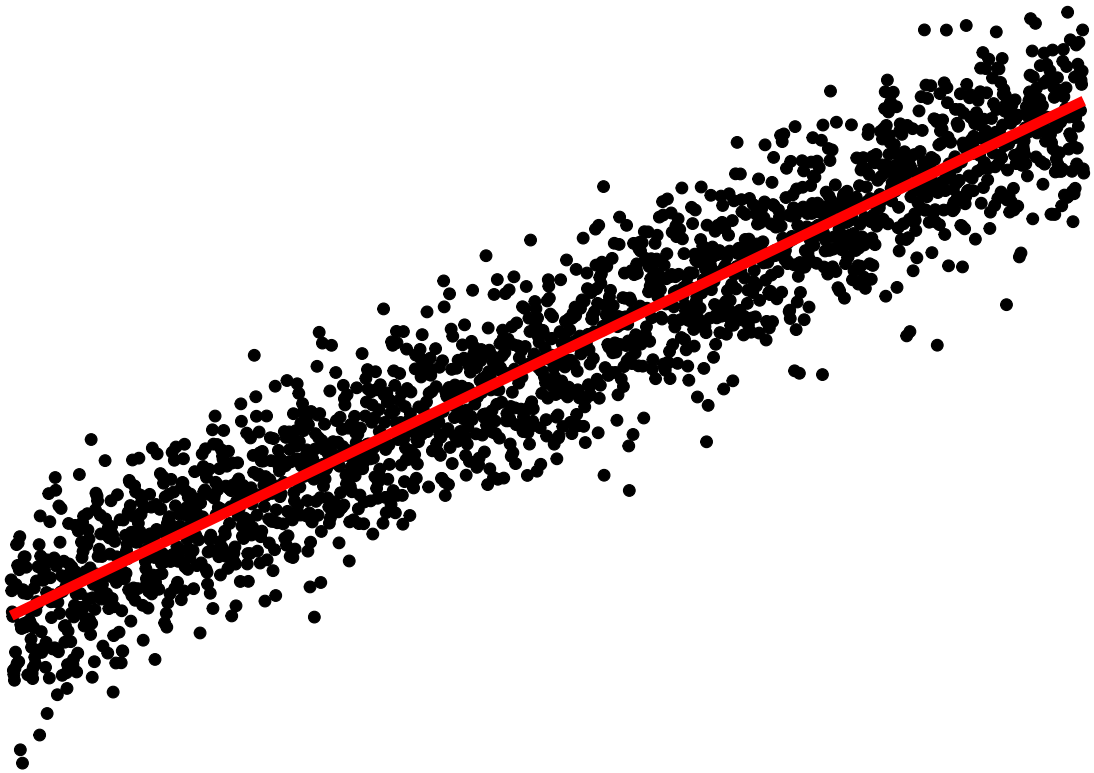
Model development



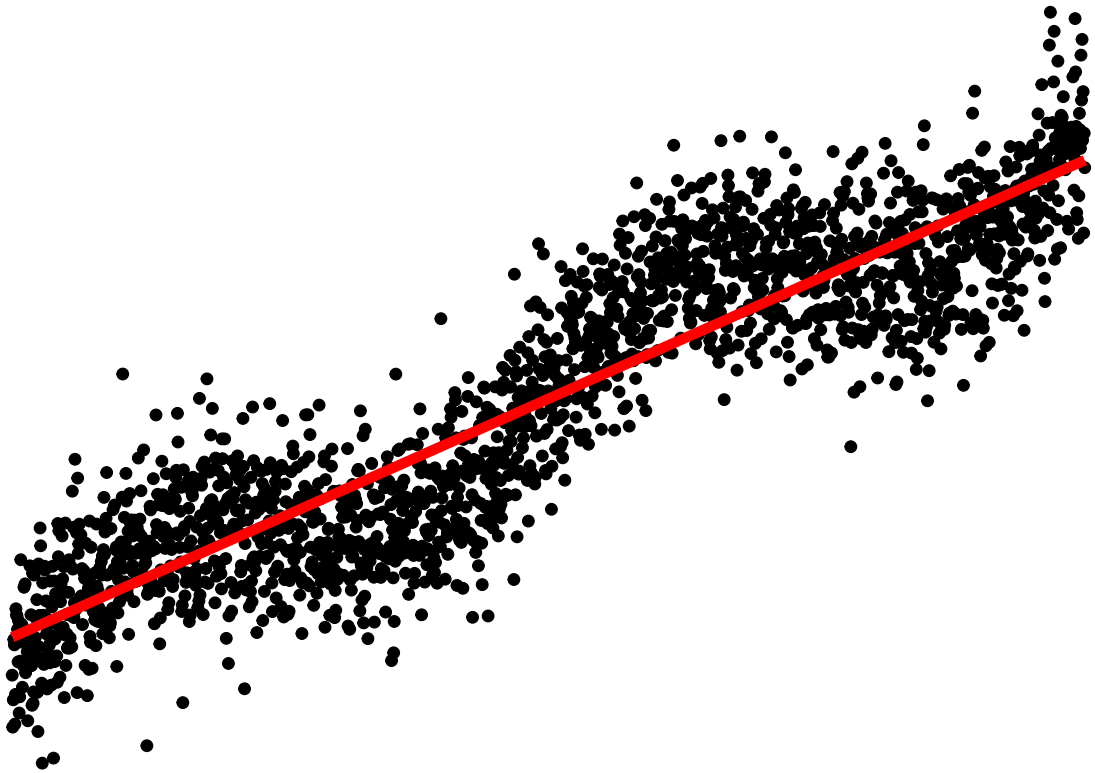
Model development



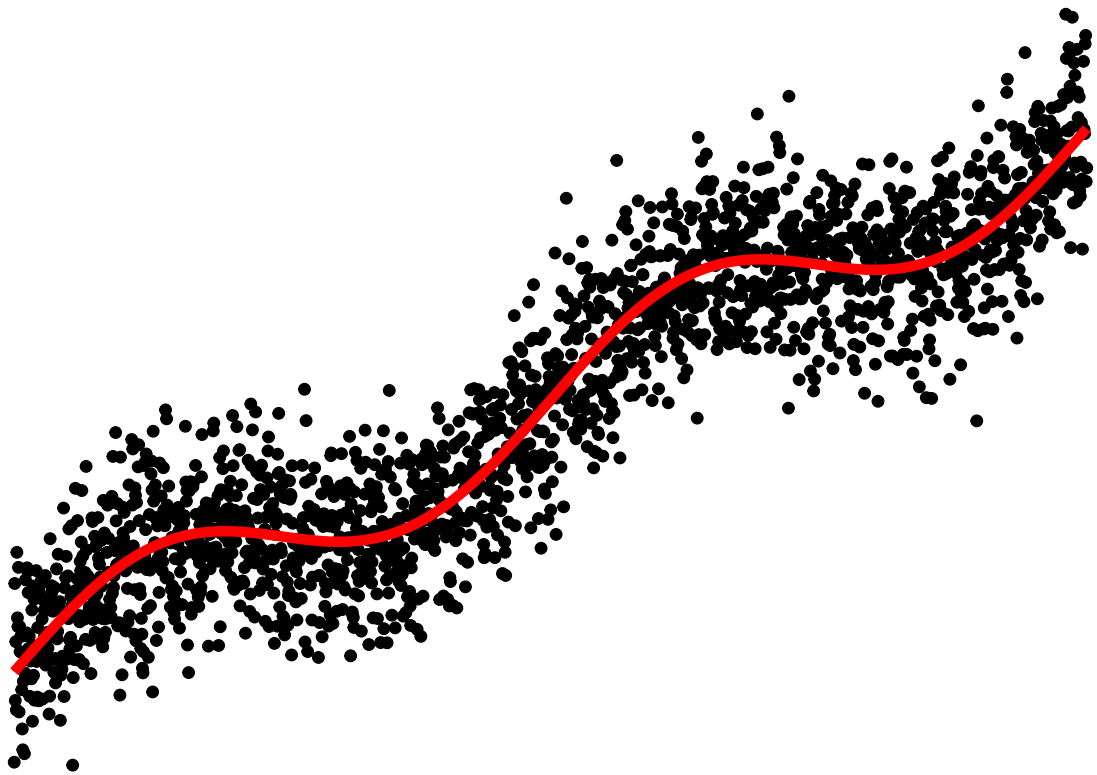
Model development



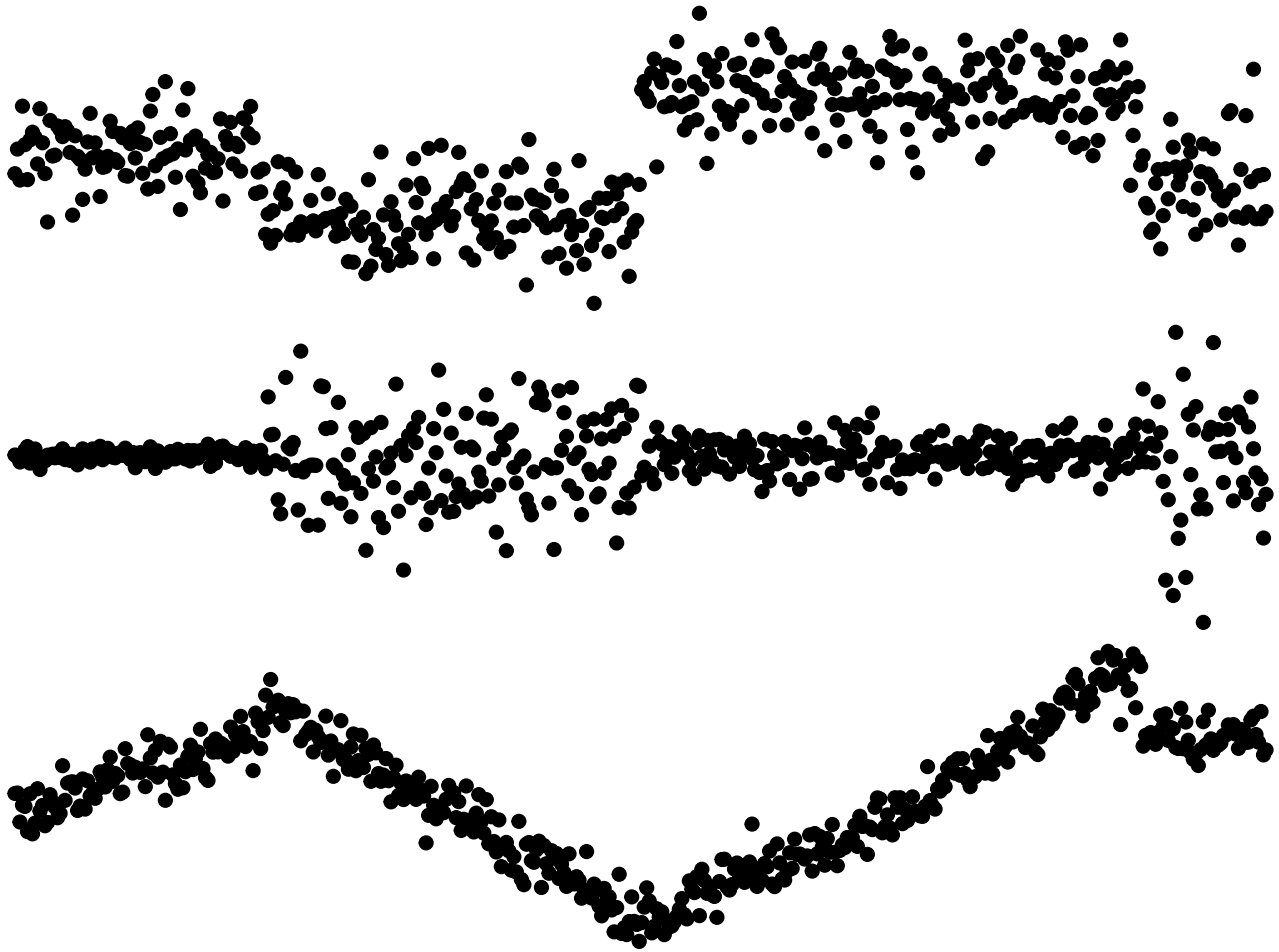
Model development



Model development



Model development



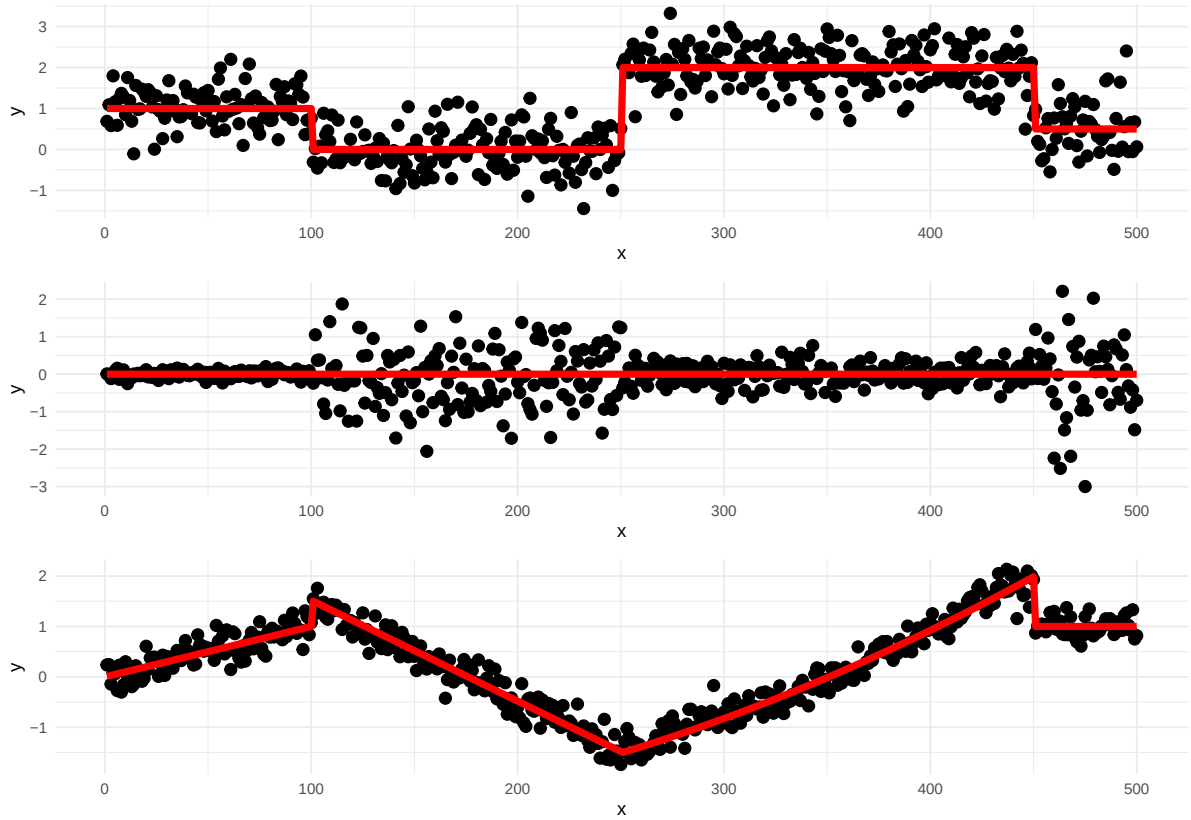
Outline

- What are changepoints?
- How do we fit changepoint models?
- Some key considerations
- Extensions

What are changepoints?

For data $y_1, \ldots, y_n$, if a changepoint exists at τ, then $y_1, \ldots, y_\tau$ differ from $y_{\tau+1}, \ldots, y_n$ in some way.

There are many different types of change.

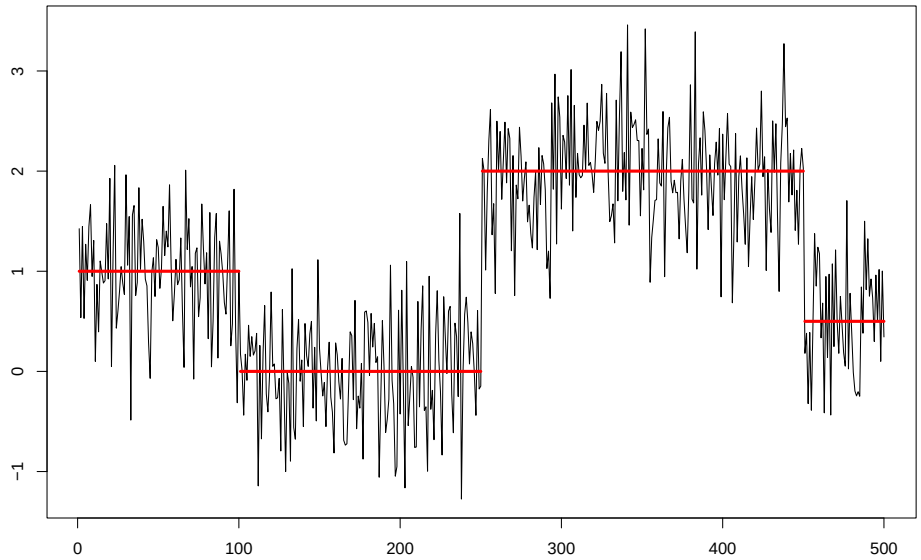


Change in mean

Assume we have time-series data where

$$Y_t|\theta_t \sim \mathrm{N}(\theta_t, 1),$$

but where the means, θ_t, are piecewise constant through time.



Inferring Changepoints

We want to infer the number and position of the points at which the mean changes. One approach:

Likelihood Ratio Test

To detect a single changepoint we can use the likelihood ratio test statistic:

$$LR = \max_{\tau}\{\ell(y_{1:\tau}) + \ell(y_{\tau+1:n}) - \ell(y_{1:n})\}.$$

We infer a changepoint if $LR > \beta$ for some (suitably chosen) β. If we infer a changepoint its position is estimated as

$$\tau = \arg\max\{\ell(y_{1:\tau}) + \ell(y_{\tau+1:n}) - \ell(y_{1:n})\}.$$

This can test can be repeatedly applied to new segments to find multiple changepoints.

Likelihood Ratio Tests

Define m to be the number of changepoints, with positions $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_{m+1})$ where $\tau_0 = 0$ and $\tau_{m+1} = n$.

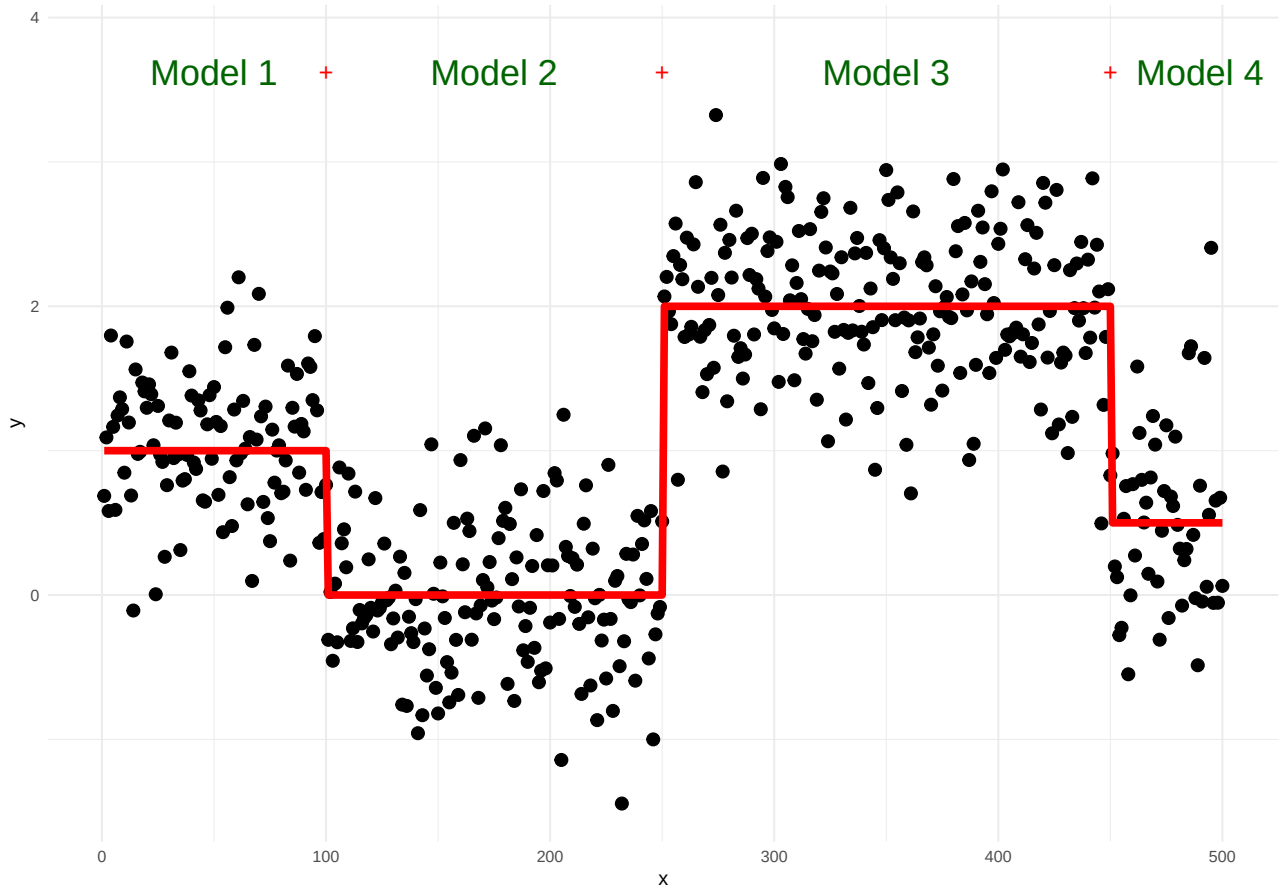
Then one application of the Likelihood ratio test can be viewed as

$$\min_{m \in \{0,1\}, \boldsymbol{\tau}} \left\{ \sum_{i=1}^{m+1} \left[-\ell(y_{\tau_{i-1}:\tau_i}) \right] + \beta m \right\}$$

Repeated application is thus aiming to minimise

$$\min_{m, \boldsymbol{\tau}} \left\{ \sum_{i=1}^{m+1} \left[-\ell(y_{\tau_{i-1}:\tau_i}) \right] + \beta m \right\}$$

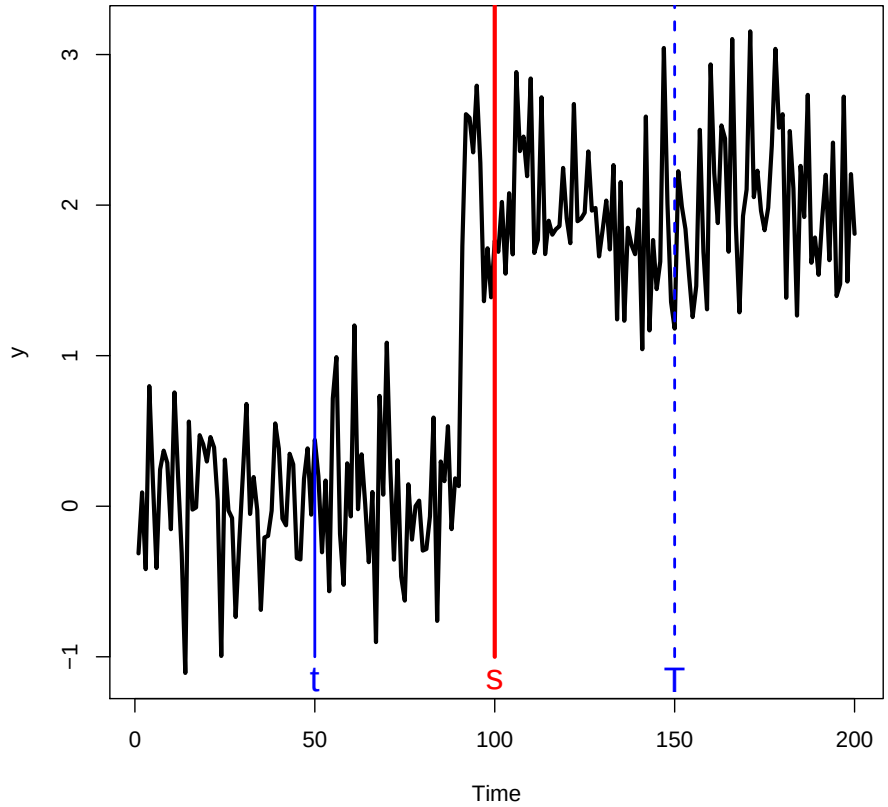
Problem



- How many changes?
- Where are the changes? 2^{n-1} possible solutions!

PELT in a nutshell

- Dynamic programming allows us to only worry about the location of the *last* change.
- Pruning means that as we go through the data we are smart about which locations are potential last change locations.

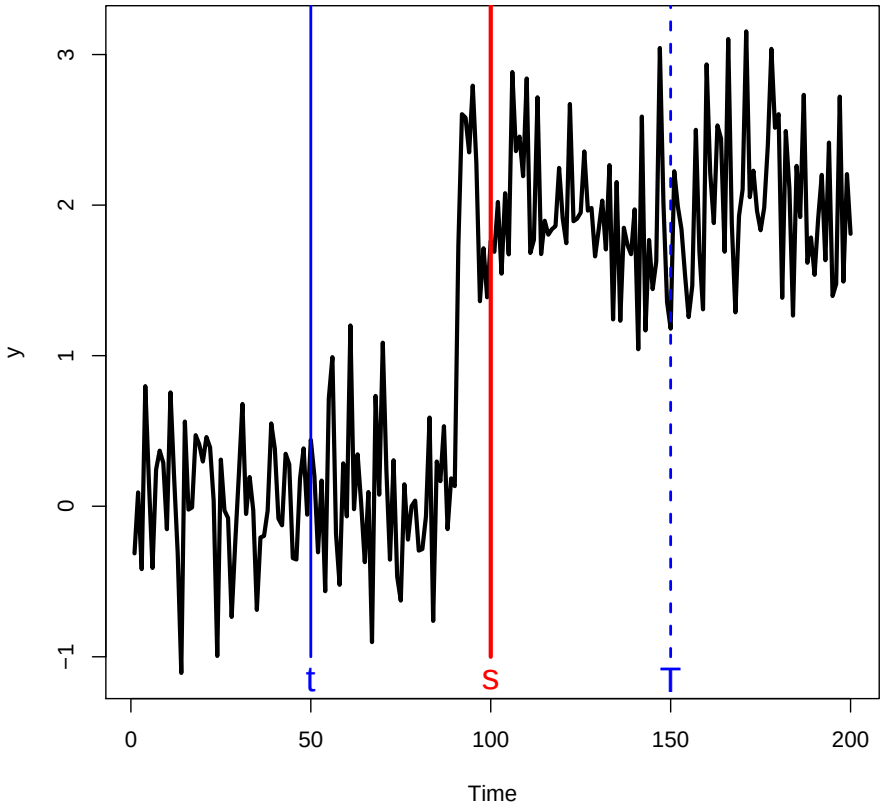


PELT: Pruning

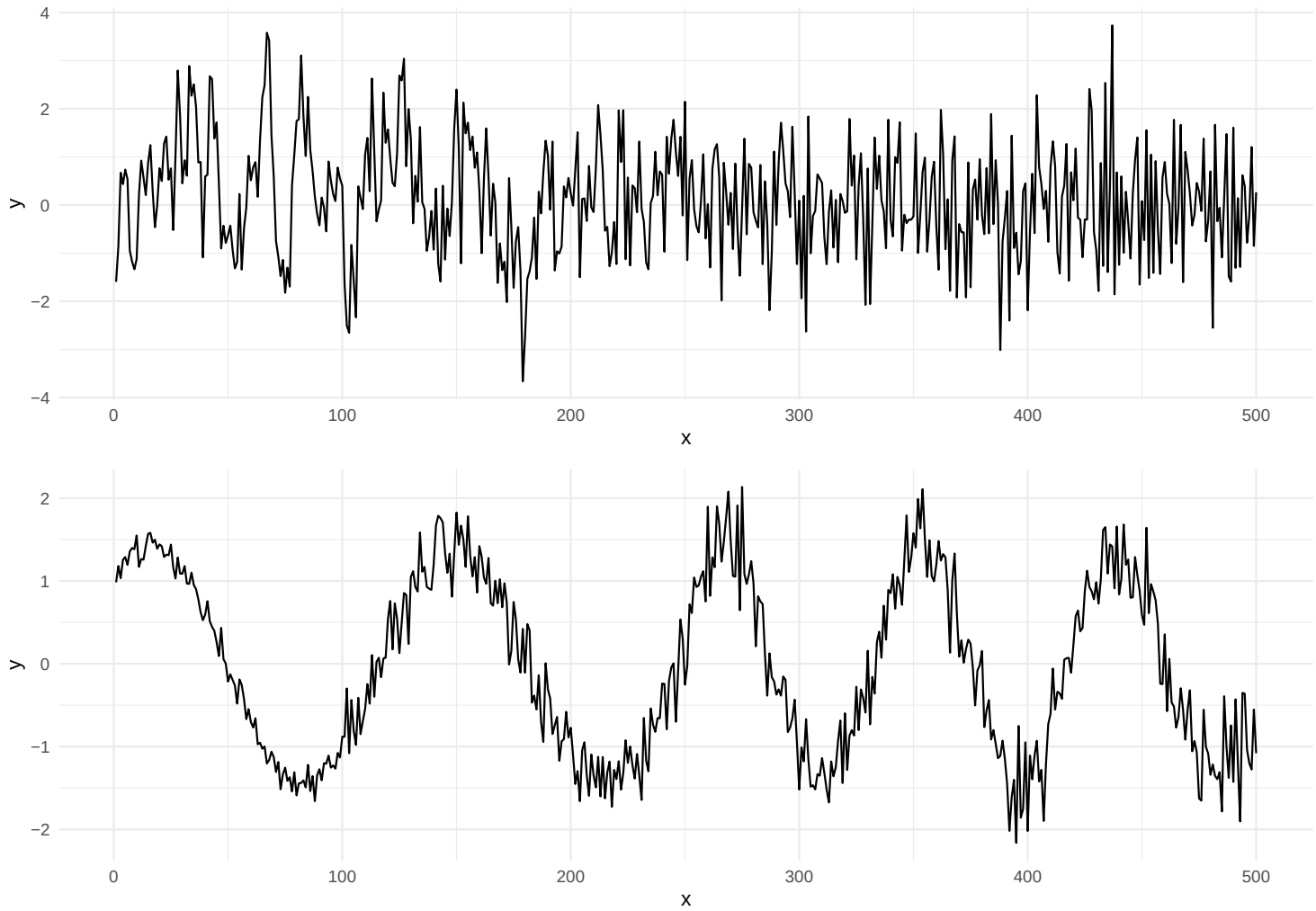
Let $0 < t < s < T$, if

$$F(t) + \mathcal{C}(y_{(t+1):s}) < F(s)$$

then at any future time $T > s$, t can never be the optimal last changepoint prior to T.
We can prove that, under certain regularity conditions, the expected computational complexity will be $\mathcal{O}(n)$.



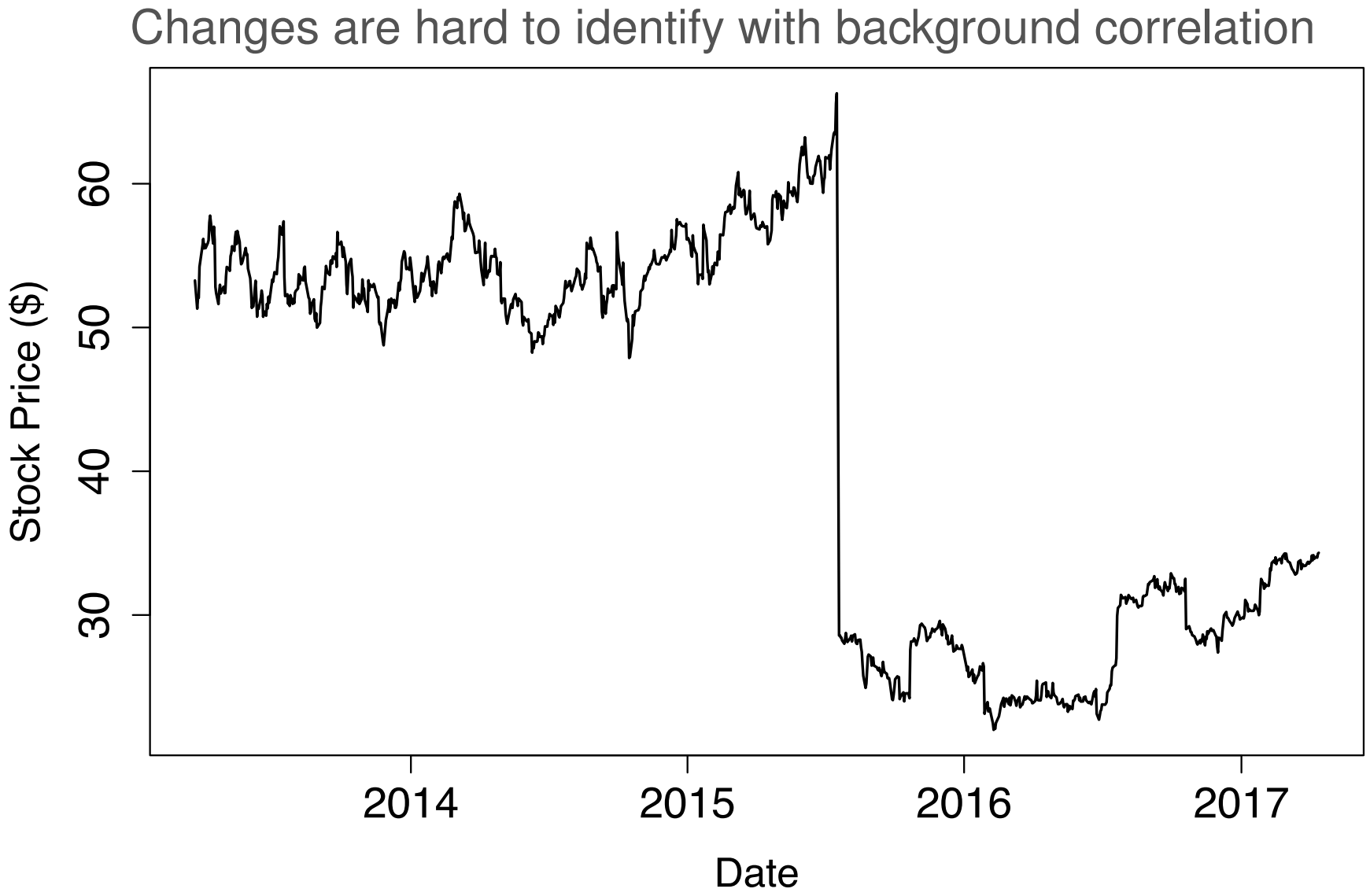
More complicated changes



Interesting Challenges

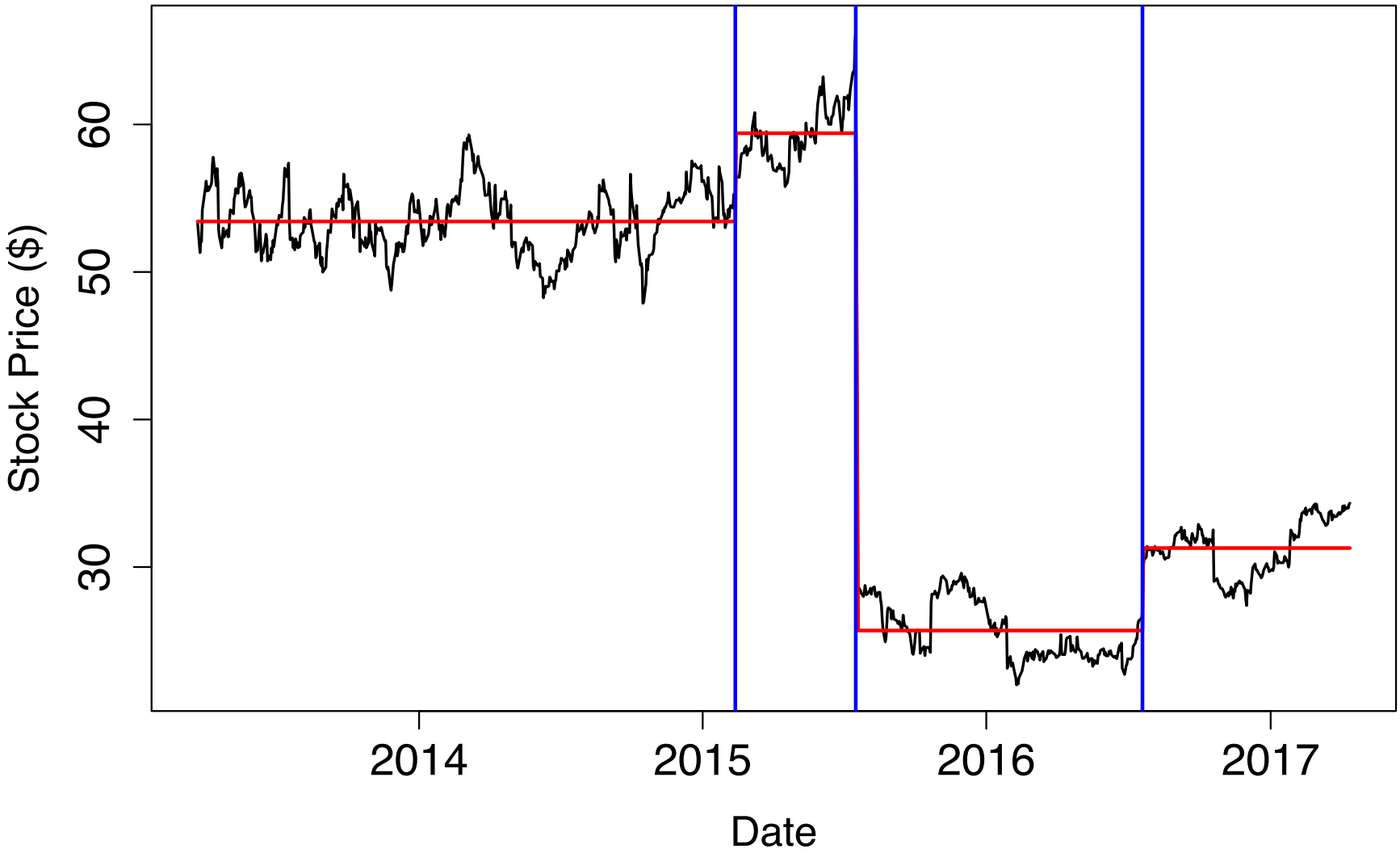
- Optimization
- Model development
- Theoretical properties
- Limits of detection
- Small sample and large sample in the same data

Stock Data


Changes are hard to identify with background correlation
Stock Price ($)
30
40
50
60
2014
2015
2016
2017
Date

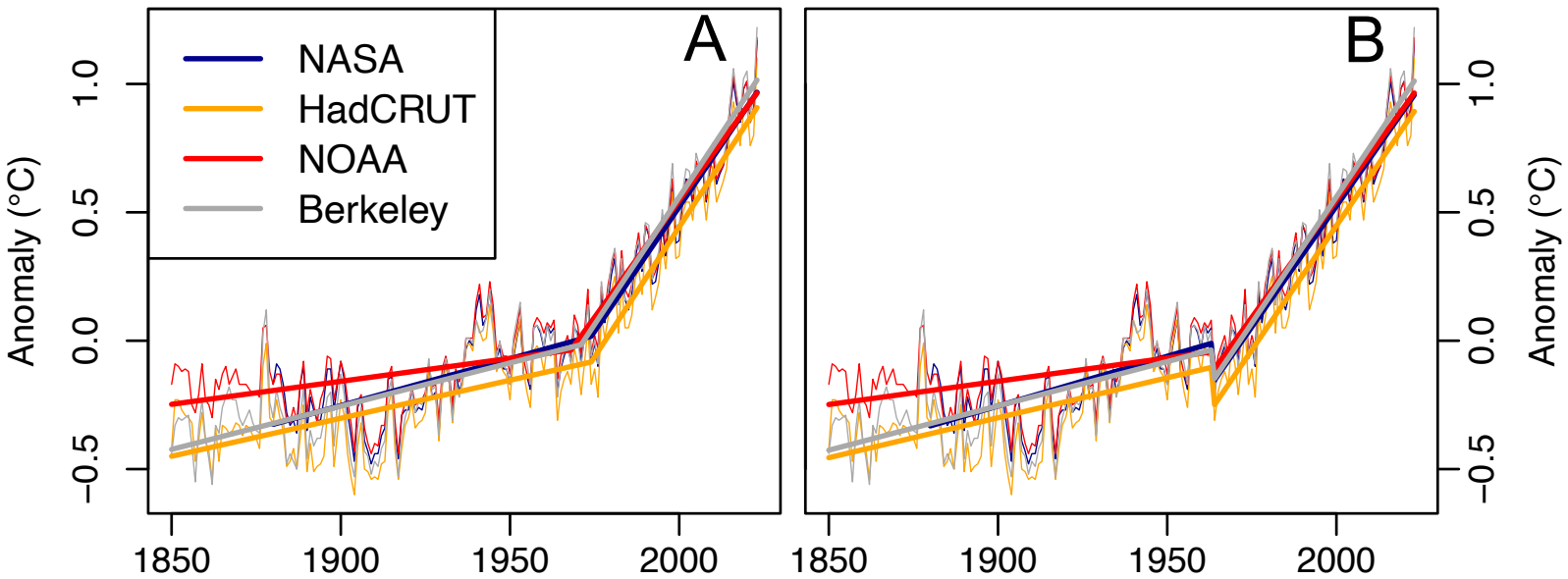
Stock Data

Change in mean with non-stationary errors



Climate Change

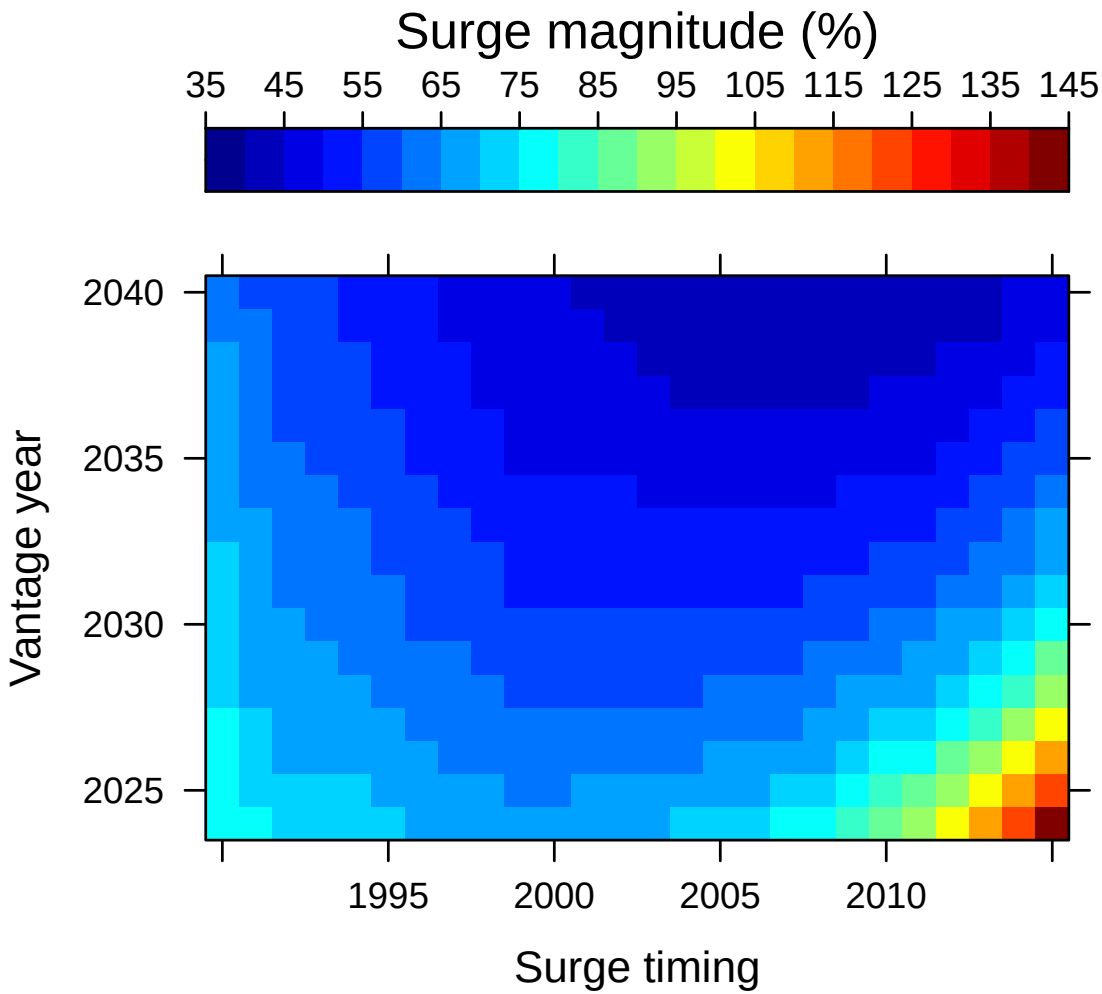
- Has there been a recent surge in global warming?
- Analysed several surface temperature time series estimates.



- Considered all possible models prevalent in the literature
- None show changes in the recent data

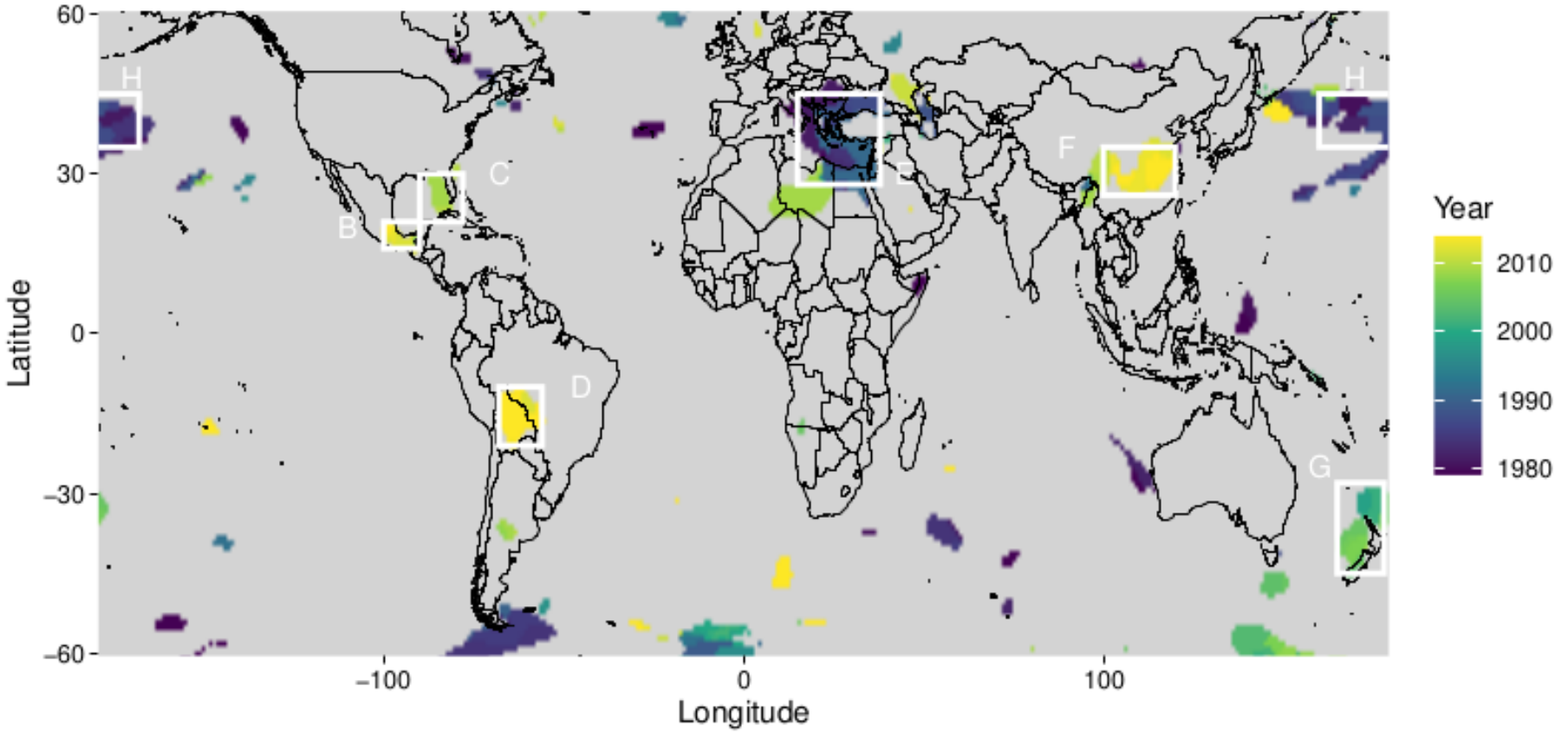
So what?

What would we need to see (in the statistically preferred model).



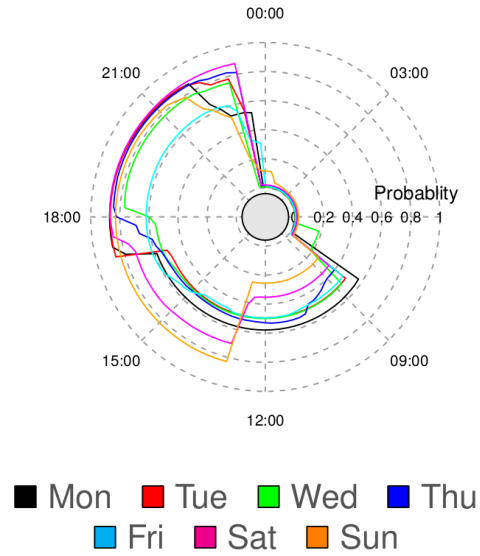
Climate Dynamics

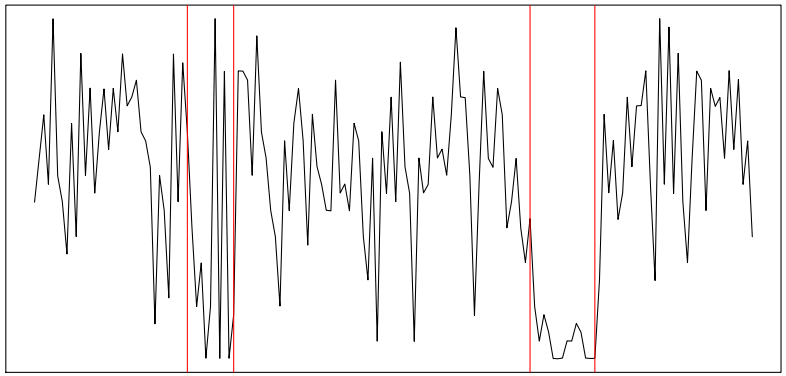
School of Mathematical Sciences | Lancaster University



Extending . . .

- Passively monitor patients at home
- Desire to have early intervention before crisis point is reached





- View data as circular, e.g., each day.
- Developed circular changepoint detection to detect changes in activity levels within a day
- Changes in the detected routine signal early intervention

Extension to multivariate

What do we mean by a multivariate changepoint?

- Change occurs in all series at the same time?
- Change occurs in groups of series, are they known or unknown?
- Change is allowed to be in any combination at each change time?

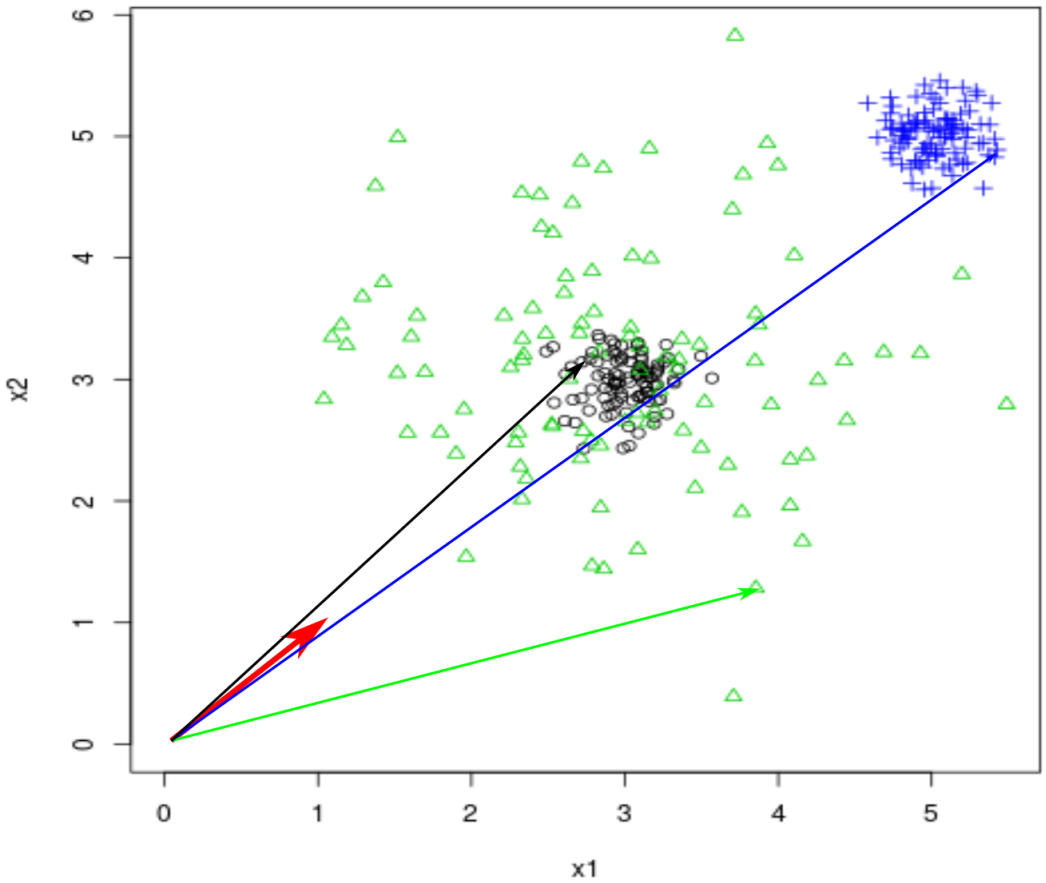
What about model structures?

- Are the series correlated?
- If yes, does the correlation change too?
- Is the model form the same for each series?
- Are we assuming the same sampling for each series?

...

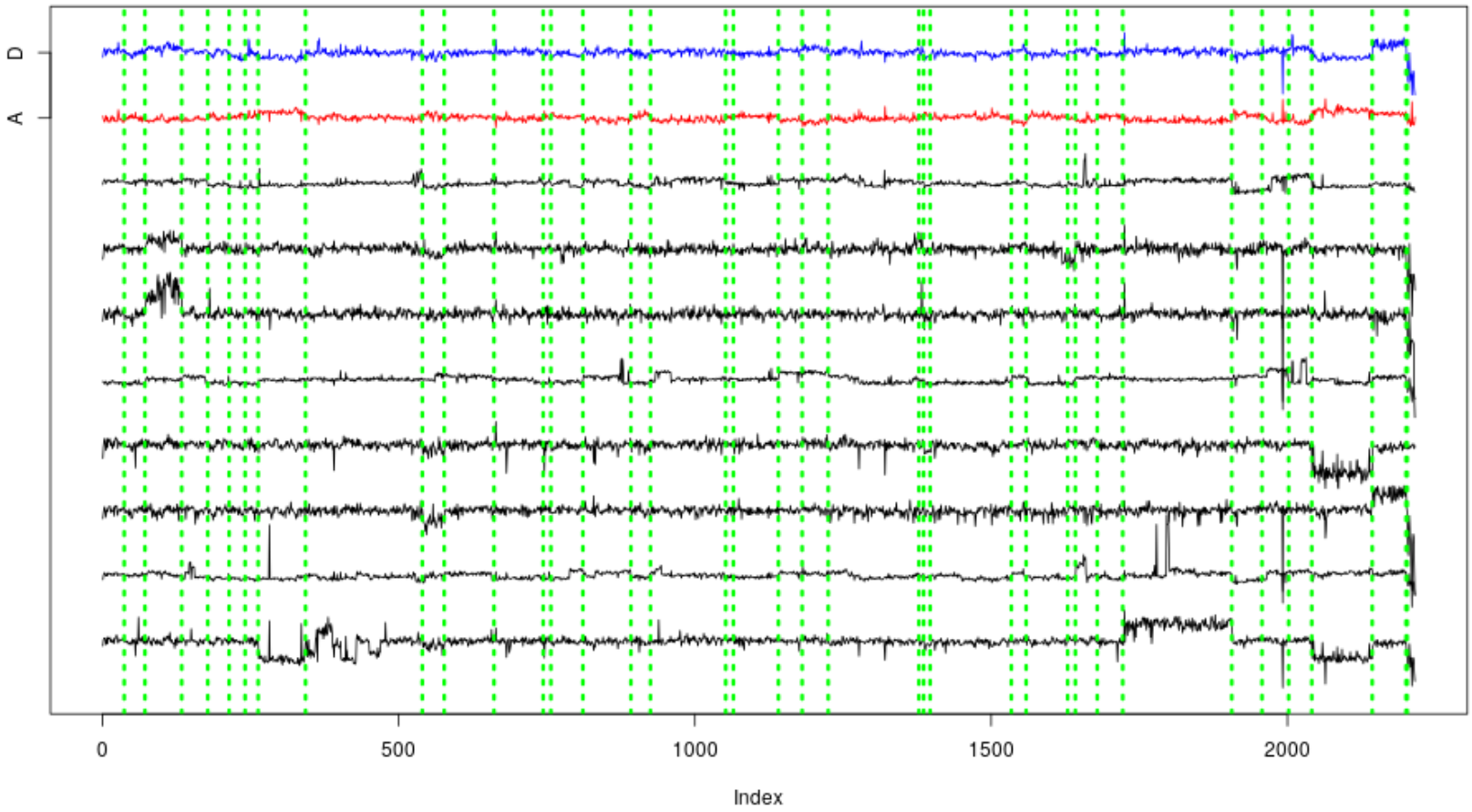
Health

- Which genomic markers could be indications for disease?
- Mean and variance changes



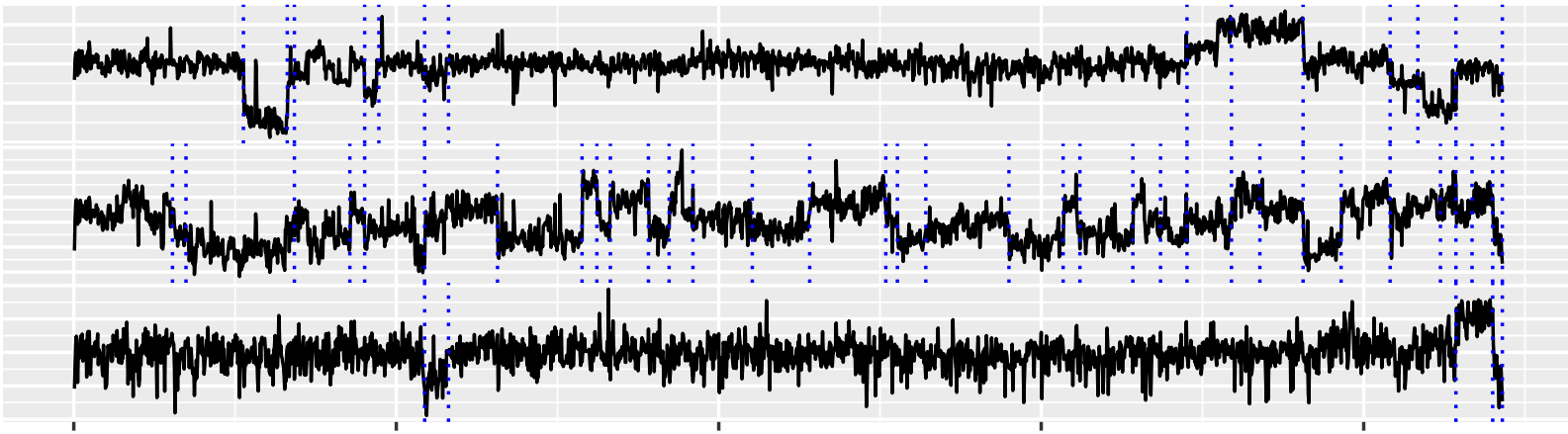
Health

- Which genomic markers could be indications for disease?
- Use two transformations instead of one



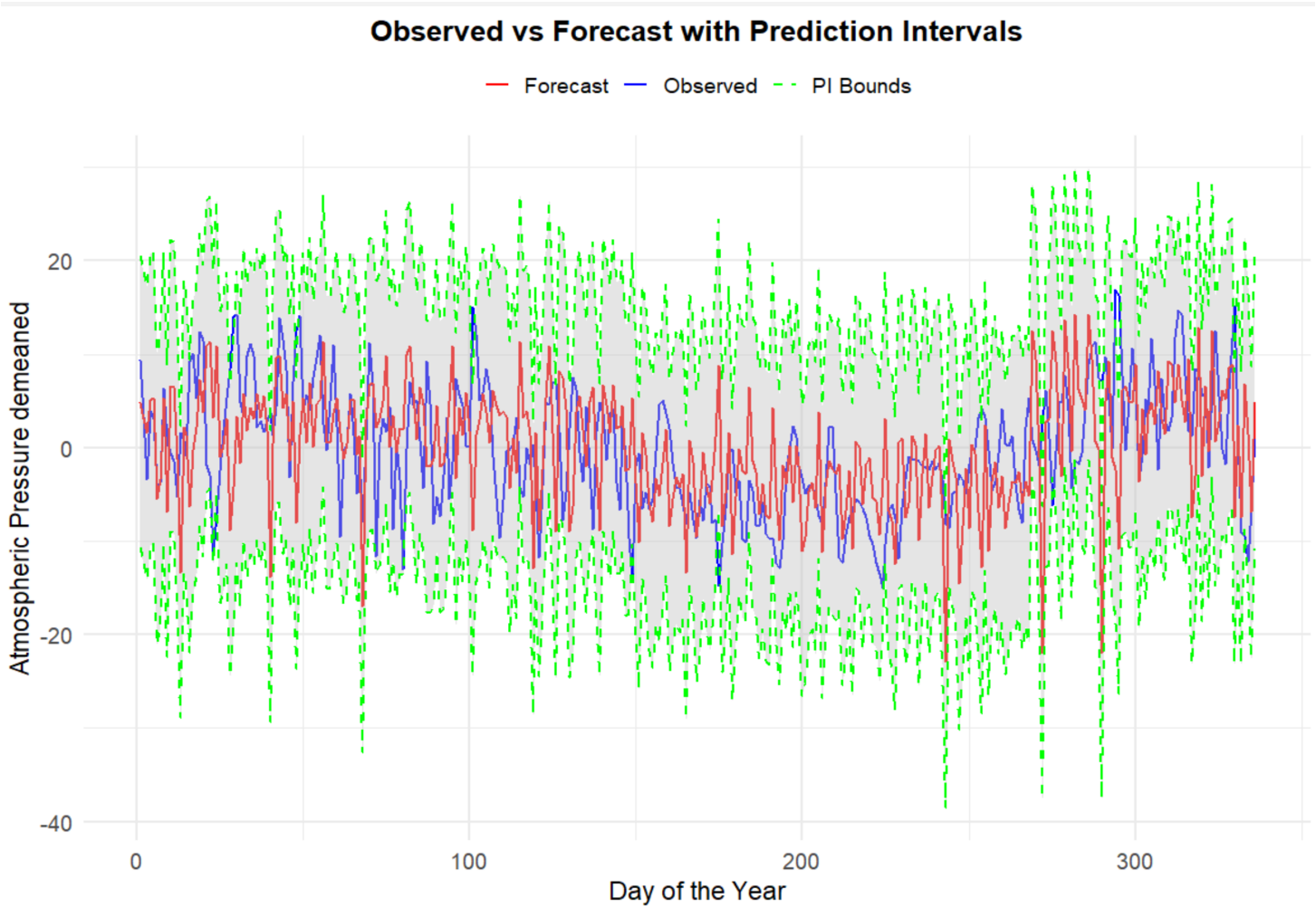
Health

- Which genomic markers could be indications for disease?
- Only markers present in lots of patients are worth investigating.



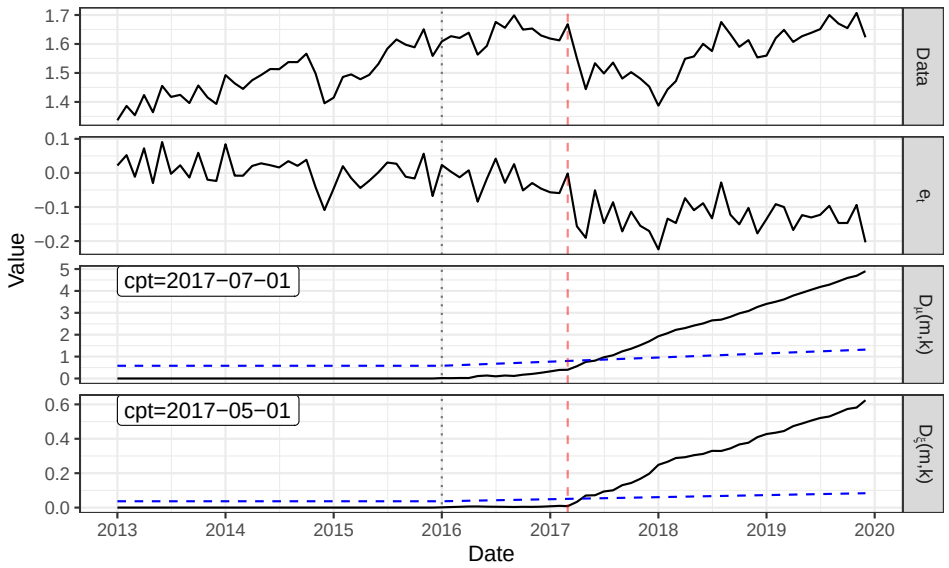
- Developed computationally efficient subset multivariate changepoint methodology.
- Only 10 changes are present in more than 60% of the patients.
- Over 50 detected by competitors.

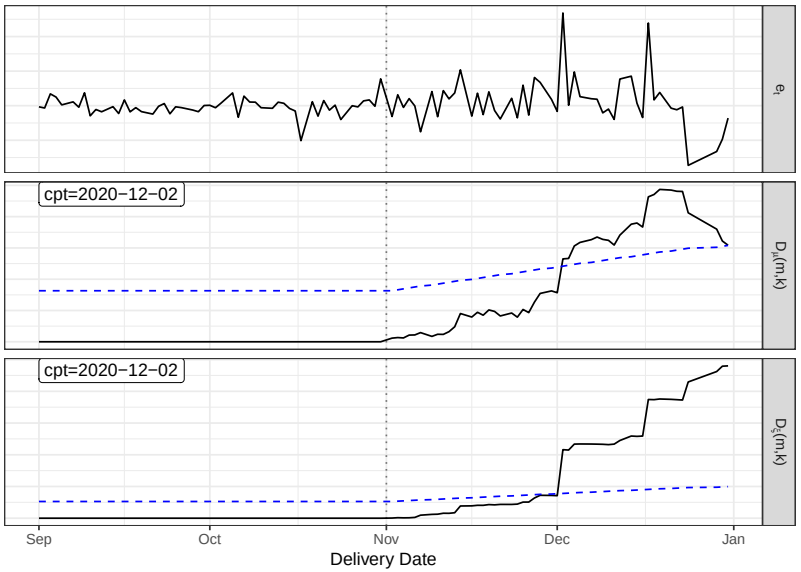
Forecasting change



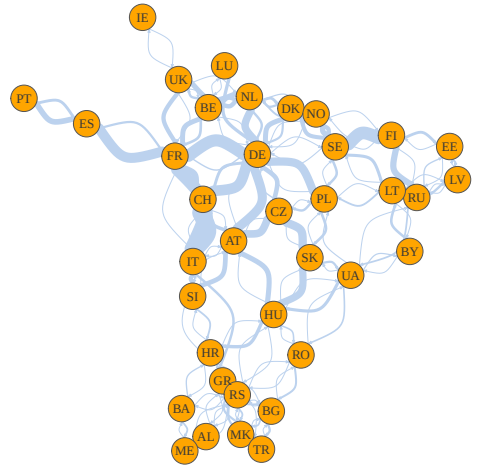
Concept drift

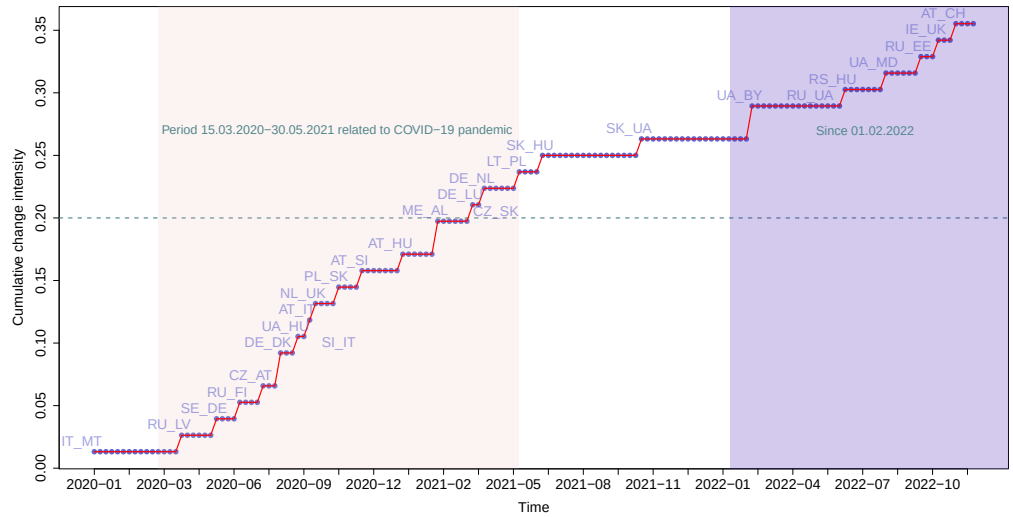
Monitor models for concept drift





Concept drift in time series on networks





Open Problems

- Optimization
- Autocorrelation
- Multivariate data
- Model complexity
- Online changes

Summary

- Changepoints occur in most data
- Detecting and documenting them improves inference
- Multivariate, online, different data types mostly isn't a problem
- But there are still plenty of open problems in the field
- I enjoy working with messy real life data . . .
- . . . as often it sparks my next research challenge.

Papers discussed

Preprints of all available at: www.lancs.ac.uk/~killick/pub.html

PELT: https://doi.org/10.1080/01621459.2012.737745

Model Choice: https://doi.org/10.1175/JCLI-D-17-0863.1 & https://doi.org/10.1002/qre.2712

LMvsCpts: https://doi.org/10.1007/s11222-017-9731-0 & https://doi.org/10.1002/env.2568

Multivariate: https://doi.org/10.1038/s41598-019-46836-y & https://doi.org/10.1007/s11222-020-09940-y

Autocorrelation: https://doi.org/10.1002/sta4.351

Papers discussed

Circular time: http://doi.org/10.1111/rssc.12472

Sensitivity: https://doi.org/10.1080/10618600.2021.2000873

Multivariate: https://doi.org/10.1080/00401706.2023.2183261

Autocorrelation: https://doi.org/10.1007/s42952-022-00173-5

Warming Surge: https://doi.org/10.1038/s43247-024-01711-1

Concept drift: https://doi.org/10.1111/jtsa.12843

Networks: https://arxiv.org/abs/2312.16357

Autocorrelation Change

Change in variance with nonstationary errors:
https://doi.org/10.1002/env.2576

Changes in second order structure within nonstationary series:
https://doi.org/10.1214/13-EJS799

Uncertainty of autocovariance changes:
https://doi.org/10.1080/00401706.2014.902776

Code

Links or stars from: https://www.github.com/rkillick

Online: https://github.com/grundy95/changepoint.forecast

Multivariate:

Covariance:
https://github.com/s-ryan1/Covariance_RMT_simulations

Mean/Var: https://github.com/grundy95/changepoint.geo