Nonparametric anomaly detection: finding hidden radioactive sources with better statistics

James Scott
University of Texas at Austin

Describing work with:
Oscar Padilla (UT-Austin, now Berkeley)
Wesley Tansey (UT-Austin, now Columbia)
Alex Athey (UT-Austin)
Alex Reinhart (CMU)

September 5, 2018

Detecting a change in distribution from streaming data

Batches of data $y_t = \{y_{t,i}\}_{i=1}^{N_t}$ arrive in discrete time:

$$y_{t,i} \overset{\text{iid}}{\sim} f_t\,,\ i = 1, \ldots, N_t\,,\ t = 1, 2, \ldots$$

At some unknown time v, f_t changes:

$$f_t = \begin{cases} f_0 & \text{for } t \leq v \qquad \text{“pre-change” (known)} \\ f_c & \text{for } t > v \qquad \text{“post-change” (unknown)} \end{cases}$$

The statistical problem

- We want to detect the change as quickly as possible, while minimizing the number of false alarms.
- But no parametric forms for f_0 or f_c

This talk: two parts

A "windowed KS" test:

- Based on (but isn't quite) the Kolmogorov–Smirnov statistic
- It is simple, robust, efficient, and intuitive to calibrate.
- Both the false-alarm rate and the power can be rigorously analyzed.
- It outperforms existing sequential testing procedures in practice.

A objective Bayesian test based on "Pólya tree discounting":

- Harder and less intuitive to calibrate.
- Improves upon the KS test in simulation studies.
- Right now, just using a Bayes factor as a test statistic.
- Still a work in progress.

Our motivating example: detecting radiological anomalies

We work with physicists who build devices and software for radiological anomaly detection. These tools can be used to:

- Find and defuse a radiological dispersal device.
- Monitor a port for smuggled radiological material.
- Locate a lost source (e.g. at a hospital).

Three problems:

- Radiation is everywhere (NORM).
- NORM varies from place to place.
- Radiation is statistically noisy (due to quantum mechanics).

Current best solution = hire a Ph.D. in physics to stare at a computer monitor.

Our motivating example: detecting radiological anomalies

Our setup:

- Small cesium-iodide detector (on officer, in vehicle, etc.)
- Detector yields energies $y_{t,i}$ for photons arriving at time t
- Energies binned into discrete channels: 4,096 counts per second $\times$ 24/7/365 $\times$ D detectors
- Detector hooked up to Raspberry Pi + iPhone that continuously queries PostGIS database and compares y_t versus the known background spectrum f_0 at the officer's location

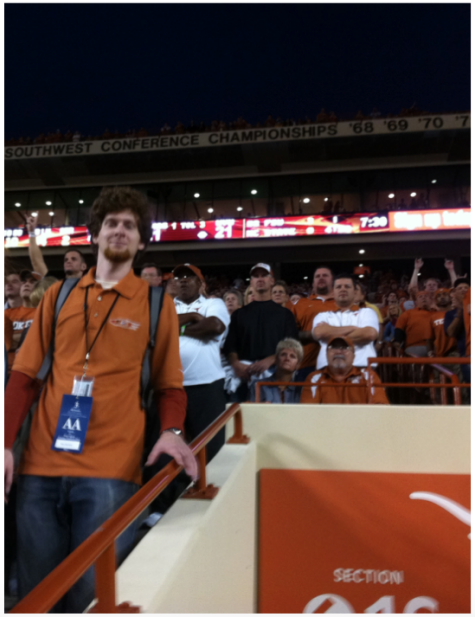
Radiological anomalies show up as changes in distribution:

- Are the recent $y_{t,i}$'s from f_0 (the background spectrum)?
- Or from f_c, a spectrum distorted by the presence of some unknown anomaly?

Our data


mophie
BRIDGEPORT


YoYo! LTE 12:13 PM
192.168.20.6:8080
41 cps
Histogram


SOUTHWEST CONFERENCE CHAMPIONSHIPS '68 '69 '70
SECTION


LONGHORNS

Our data


POLICE
2850
THE UNIVERSITY OF TEXAS AT AUSTIN


POLICE
714
THE UNIVERSITY OF TEXAS AT AUSTIN


SET YOUR CARTS 5 FEET APART
5 feet
RECYCLING

The whole pipeline

1) Instrument calibration (not discussed today)
   - Cheap ($\approx$ \$5000) detectors allow us wider coverage but are noisier (temperature, rain, instrument-level variability).

2) Background mapping (maybe a bit at the end)
   - Significant spatial variation due to NORM, mostly in buildings
   - "Sharp + smooth," both in spectral and spatial dimensions
   - Lots of data, unevenly distributed over monitoring area

3) Anomaly detection (most of this talk)

Toy example

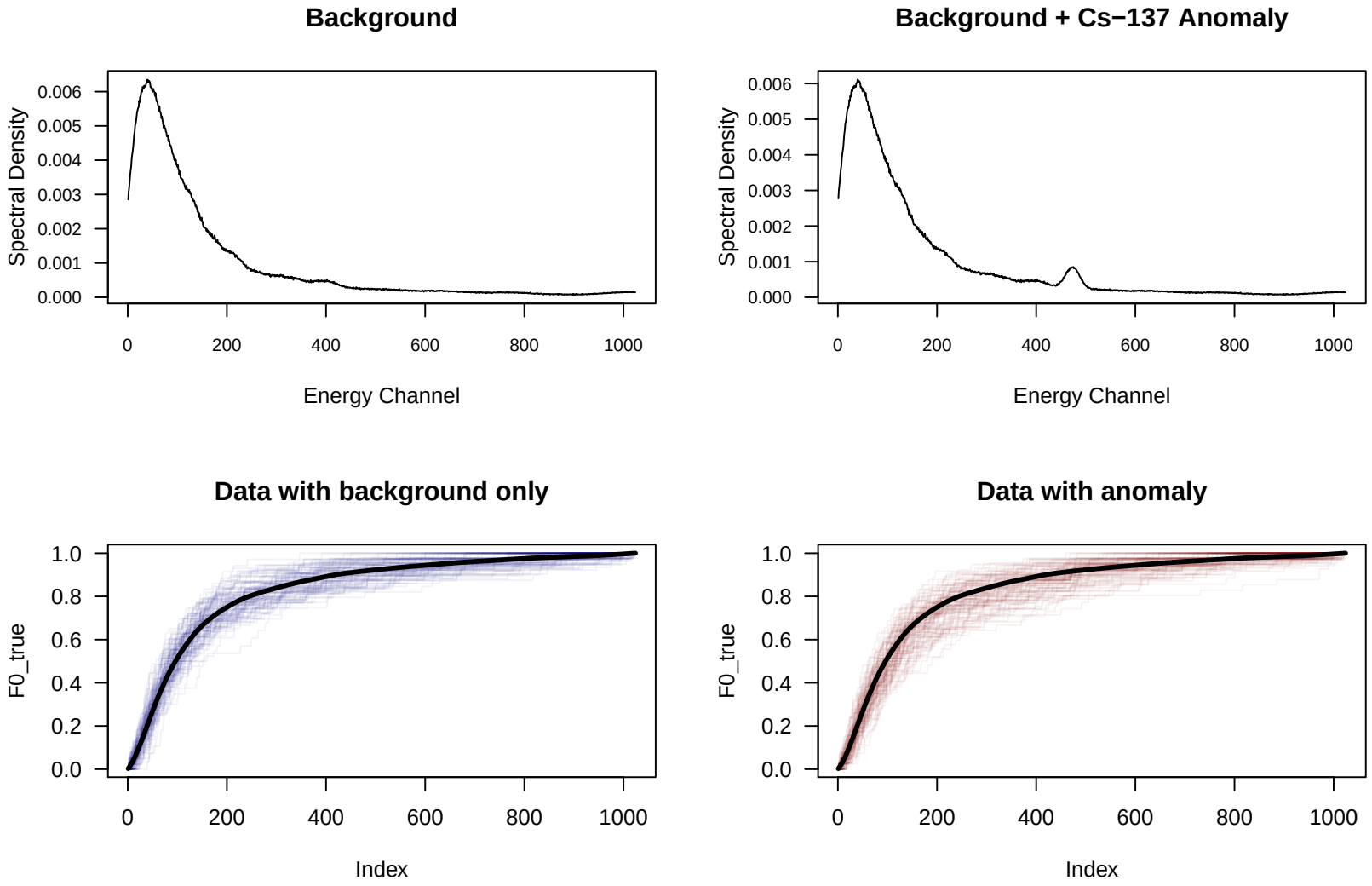


Figure: A synthetic injection of 100 milliCurie source of Cesium 137 located at a distance of 150m from the detector.

Toy example

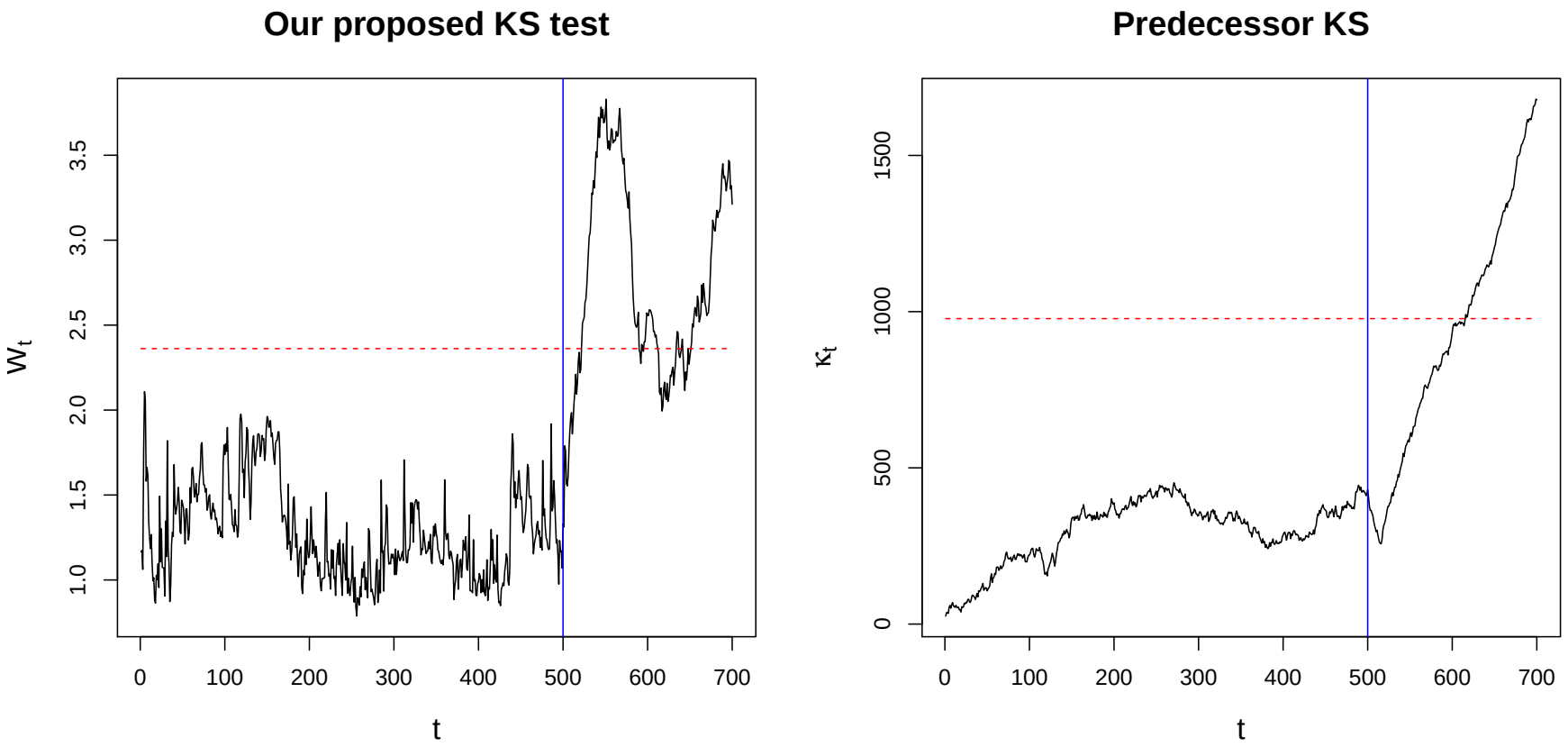


Figure: The proposed test versus the pre-existing state of the art. Both methods are calibrated to have a false-alarm rate of $\leq$1 in 1000. Left: threshold chosen from our theory. Right: threshold from simulations.

Basics

Let $\{y_{t,i}\}_{i=1}^{N_t}$ be the set of measured energies from the gamma rays arriving at time t. The laws of physics says that

$$y_{t,i} \overset{\text{iid}}{\sim} f_t\,,\ i = 1, \ldots, N_t\,, \quad N_t \sim \text{Poisson}(\mu)\,,$$

where f_t is the gamma-ray spectrum at time t, and $\mu > 0$.

Naïve approach: compare N_t with the background rate μ

- Different devices have different sensitivities to radiation: μ is a joint property of the world and the measurement device.
- We also find noticeable differences in N_t observed using the *same* detector from one day to the next
- Thus attempting to detect anomalies using N_t is too fraught.

Basics

A better approach is to look for a change in f_t:

$$y_{t,i} \overset{\text{iid}}{\sim} f_t\,,$$

where

$$f_t \;=\; \begin{cases} f_0 & \text{for } t \leq v \\ f_c & \text{for } t > v\,. \end{cases}$$

Key facts:

- Both v and f_c are unknown.
- f_c has no particular parametric form.
- f_0 is "known" (OK, estimated—another fun problem).
- After calibration, both f_0 and f_c are consistent across devices.

Basics

The goal: construct a stopping rule T:

- A procedure for detecting that a change-point has occurred, i.e. that $t \geq v$.
- When $T = t$, we stop the data-collection process and declare that a change-point has occurred at some time during the first t observations.

Performance of stopping rules typically evaluated using two criteria

- The expected "null" stopping time: $\mathbb{E}_0(T)$, or long ARL.
- The worst-case average detection delay, or short ARL:

$$\bar{E}_c(T) = \sup_{s \geq 1} \operatorname{ess\,sup} \mathbb{E}_s \left[(T - s + 1)^+ \,\middle|\, \{y_{t,i}\}_{i=1}^{N_t},\ t = 1, \ldots, s-1 \right].$$

- The ess sup takes the "worst-case" pre-change data.
- But not well understood when f_C is unknown.

Basics

Our approach:

- Define a stochastic process $\{\Delta_t : t \in \mathbb{N}\}$.
- Declare an alarm when $\Delta_t > c$.
- The number of false alarms up to a time horizon $T < v$ is:

$$A_T(\Delta, c) = |\{t \in \mathbb{N} : t \leq T \text{ and } \Delta_t \geq c\}|,$$

- The delay time is

$$D(\Delta, c) = \inf\{t \in \mathbb{N} : t > v \text{ and } \Delta_t \geq c\} - v - 1.$$

- Goal: constrain $\mathbb{E}[A_T(\Delta, c)] \leq \alpha$ and construct a procedure with small $\mathbb{E}(D(\Delta, c))$ under this constraint.

Existing work on anomaly detection

Retrospective detection:

- KS: Chan et al. (2014) and Reinhart et al. (2015)
- Spectral comparison ratio: Pfund et al. (2006), Du et al. (2010), Reinhart et al. (2014)
- Neither fit the design requirements of the streaming-data scenario.

Sequential SCR test:

- Pfund et al. (2010)
- Tuning parameter selection is opaque; seems underpowered in experiments.

Existing work on anomaly detection

Exponential-family methods:

- We could exploit the fact that the detector actually returns discrete bin/channel counts x_{tj} for bin j.
- Thus in principle, $x_{tj} \sim \text{Poisson}(\lambda_j)$.
- Many methods could then work: e.g. Pollak (1987), Basseville et al. (1993), Siegmund and Venkatraman (1995), Lai (1995).

Issues:

- This has the same (huge) problem as testing based on total count rate μ: per-bin rates λ_j are not comparable across devices.
- Moreover, theoretical guarantees apply to univariate, continuous distributions. Neither hold here.

Our approach: based on KS statistics

Notation:

- Let F_0 be the CDF associated with the background f_0.
- Define

$$\hat{F}_t(y) \;=\; \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{1}_{(-\infty, y_{t,i}]}(y),$$

A "single-window" KS test would use the statistic

$$D_t \;=\; \sqrt{N_t} \sup_y \left| F_0(y) - \hat{F}_t(y) \right| .$$

This does not yield a good protocol for sequential detection:

- Either we pool batches of data and test retrospectively. . .
- . . . or we use D_t one step at a time and give up power.

The proposed test

The idea is simple:

- Pool data across an series of backward-looking windows.
- Test using the maximal KS statistic over those windows.

Let $\hat{F}^{s:t}(y)$ be the empirical CDF constructed from all data collected from time $s < t$ to time t:

$$\hat{F}^{s:t}(y) = \frac{1}{\sum_{k=s}^{k=t} N_k} \sum_{k=s}^{t} \sum_{i=1}^{N_k} \mathbf{1}_{(-\infty, y_{k,i}]}(y).$$

Let $\Delta_{s:t}$ be the corresponding KS statistic:

$$\Delta_{s:t} = \sqrt{\sum_{k=s}^{t} N_k} \sup_y |F^0(y) - \hat{F}^{s:t}(y)|.$$

The proposed test

Define the window statistic W_t as

$$W_t = \max_{s:\max\{t-L,1\} \le s \le t} \Delta_{s:t} \, .$$

We propose to declare an anomaly at time

$$\tau_L = \min \{ t : W_t \ge c_L \} \, .$$

where $L \in \mathbb{N}$ and $c_L > 0$ are constants.

In words: we look back and see if there is evidence for a changepoint between times $\max\{t - L, 1\}$ and t. The window size L bounds the complexity for computing W_t as $O(L)$.

Example: 5 Steps After the Changepoint

Example: 19 Steps After the Changepoint

Example: 20 Steps After the Changepoint

And that's how we detected the anomaly

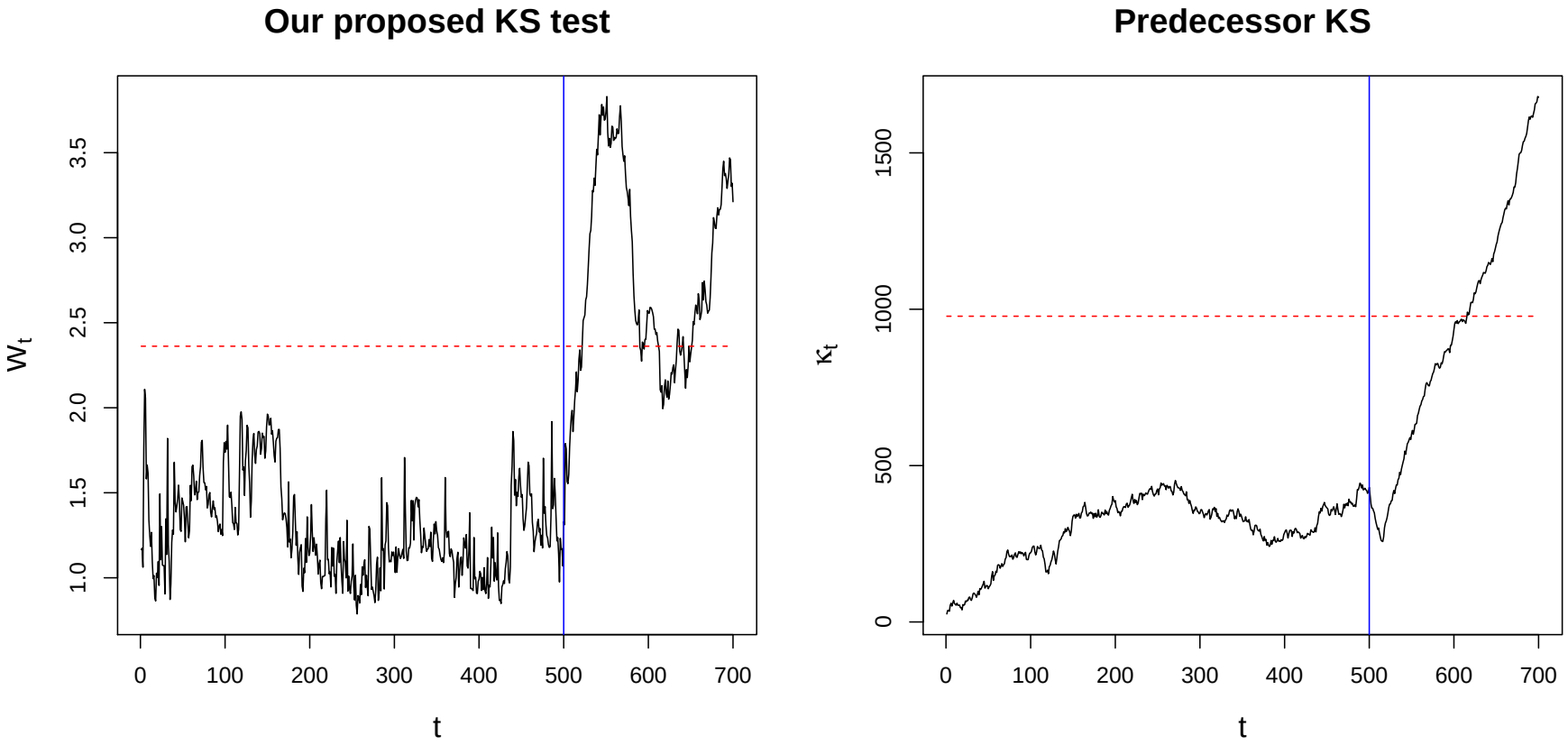


Figure: The proposed test versus the pre-existing state of the art. Both methods are calibrated to have a false-alarm rate of ≤ 1 in 1000. Left: threshold chosen from our theory. Right: threshold from simulations.

Key questions

What is the "lookback" penalty?

- Looking back across multiple lags = multiple testing.
- How should this affect the threshold for alarm?

Can we characterize the power of the procedure?

- Here power = time to detection.

How does it work in practice?

Main theorem

Let F_c be the post-change CDF (i.e. for $t \geq v$) and let

$$d(F_c, F_0) := \sup_y \left| F_0(y) - F_c(y) \right| ,$$

Theorem

Assume that $s > v$ is fixed. Then

$$\lim_{t \to \infty} \Delta_{s:t} = \infty \quad a.s.$$

provided that $d(F_0, F_c) > 0$. Moreover, for $c_L > 0$,

$$P\left(\Delta_{s:t} > -c_L + d(F_c, F_0) \sqrt{\sum_{k=s}^{t} N_k} \right) \geq 1 - 2\exp(-2\, c_L^2) .$$

A corollary

Define, for $T < v$,

$$A_T = \left| \left\{ t : t \leq T, \max_{\max\{1, t-L\} \leq s \leq t} \Delta_{s,t} \geq c_L \right\} \right|.$$

Here A_T can be thought as the number of times that the process exceeds the threshold within a window of length T when there is no change point in $\{1, \ldots, T\}$.

Corollary

If $T \leq v$, then

$$\frac{E(A_T)}{T} \leq 2L \exp\left(-2\, c_L^2\right).$$

This bounds the expected number of false alarms up to time T, provided that the change point happens after T.

A corollary

This corollary is immediately practical:

- For an acceptable false-alarm rate r, choose c_L so that the expect number of false alarms is less r

Example:

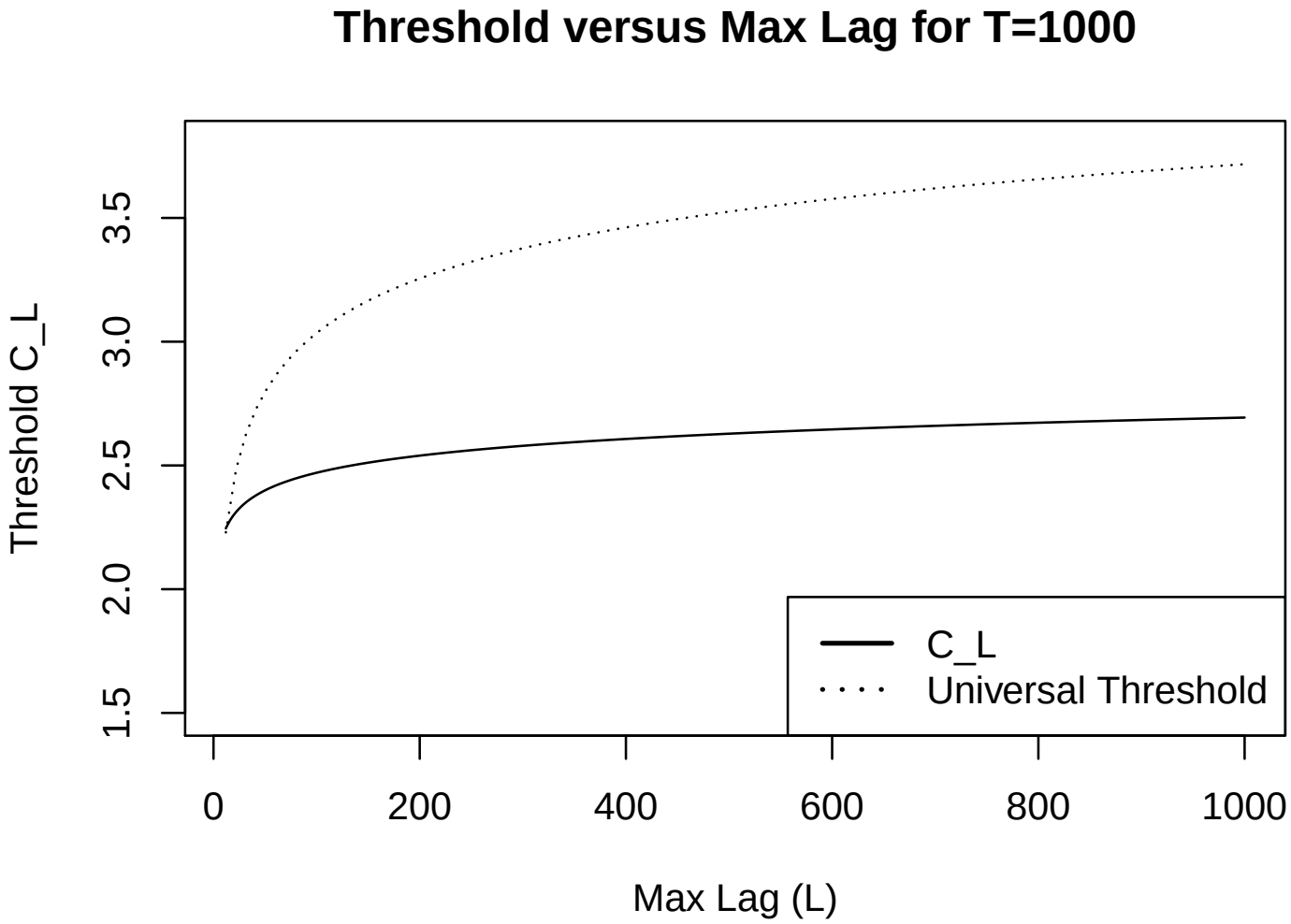
- Suppose that $L = 50$.
- Goal: expected rate of false alarms, $E(A_T)/T$, no more than $r = 1/1000$.
- Requirement:

$$\frac{E(A_T)}{T} \leq 2L \exp\left(-2\, c_L^2\right) \leq r$$

- If $r = 0.001$, this holds whenever

$$c_L \geq \sqrt{\frac{1}{2}\log(2L) - \frac{1}{2}\log r} \approx 2.4\,.$$

The penalty for multiple testing



Toy example

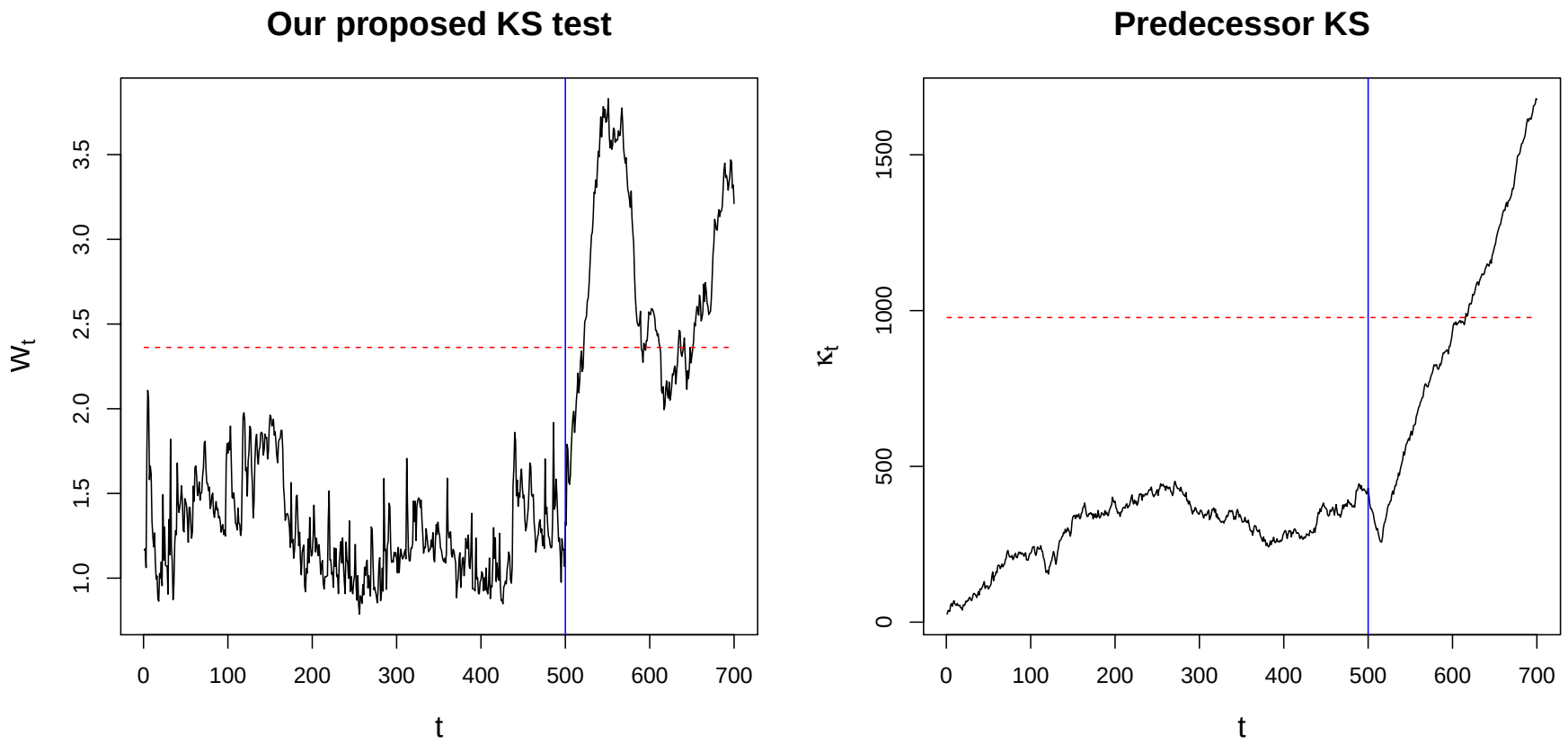
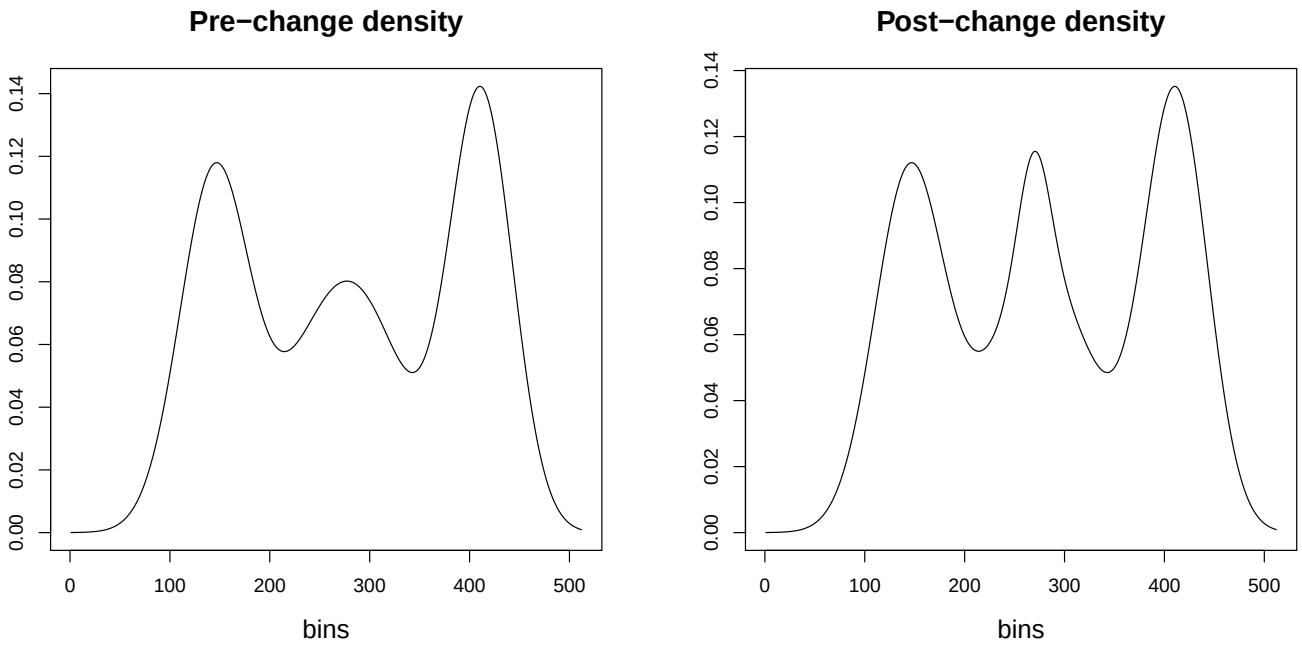


Figure: The proposed test versus that of Hawkins (1988).

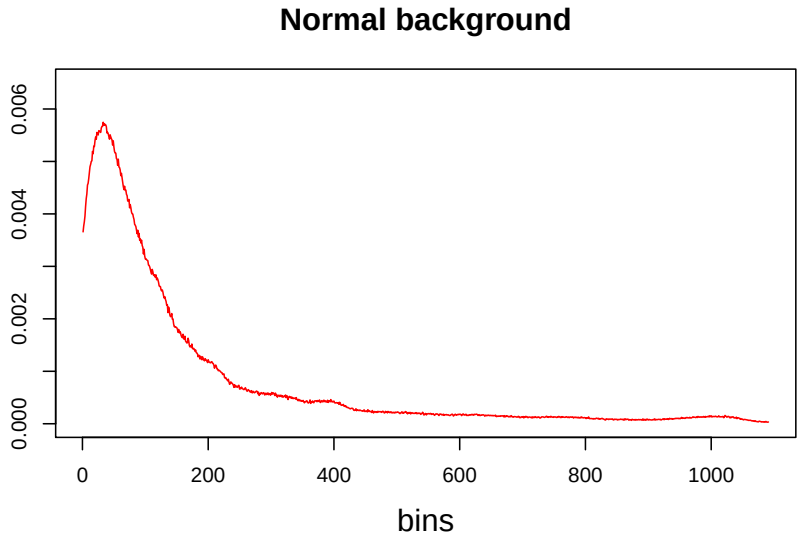
A small simulated-data example

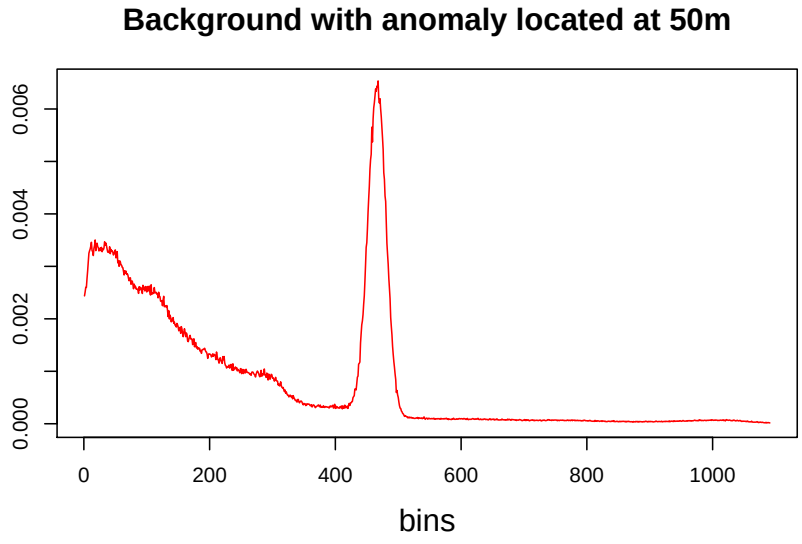


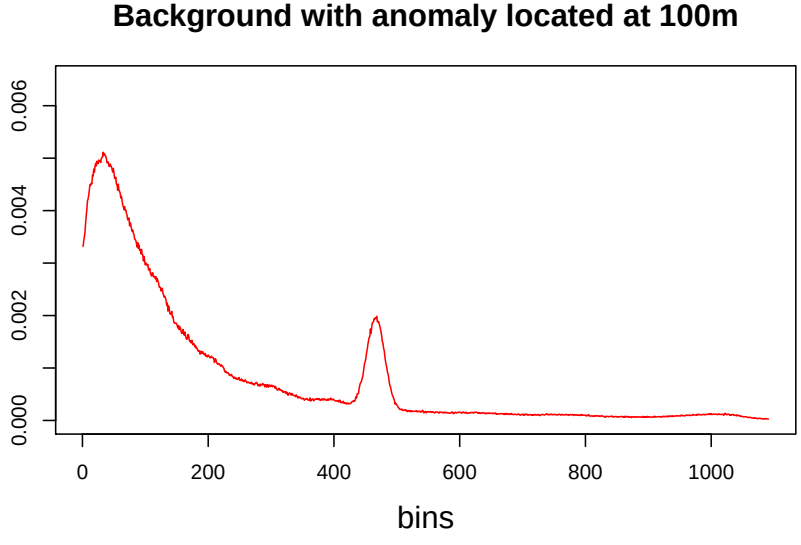
$E(N_t)$	KS	PKS	EF	GLR
100	**152.6**	197.0	317.1	257.2
500	**31.2**	94.7	45.2	41.0
1000	**12.3**	64.0	31.3	30.4

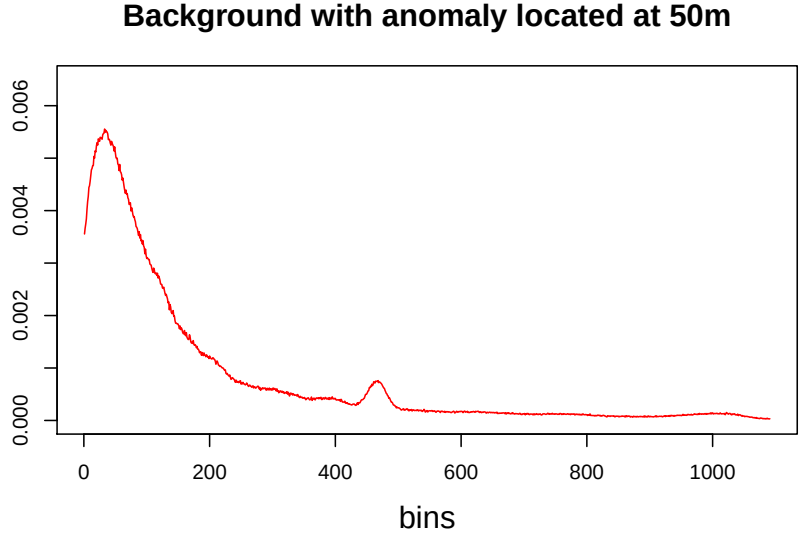
Table: Stopping times averaged over 100 data sets (smaller is better).

A cesium anomaly







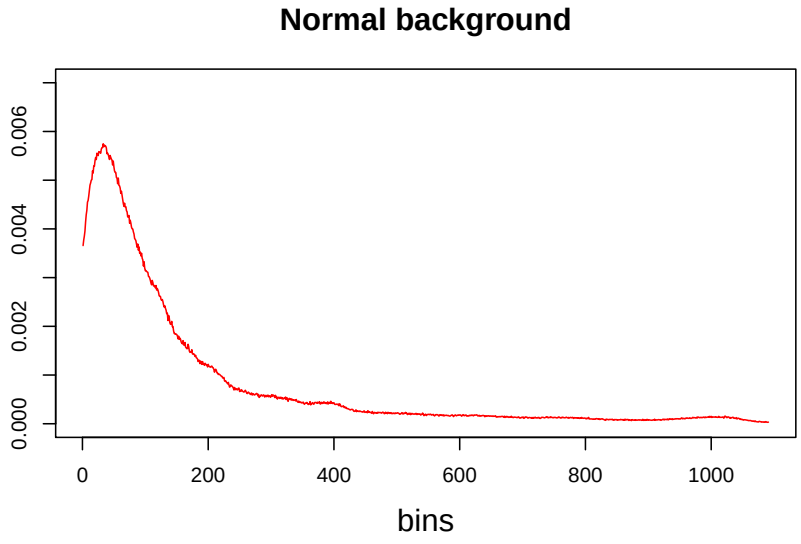


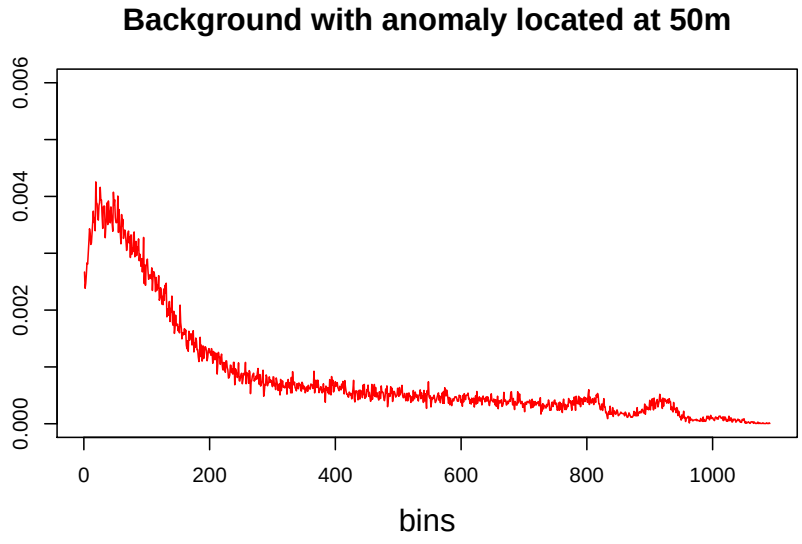
A cesium anomaly

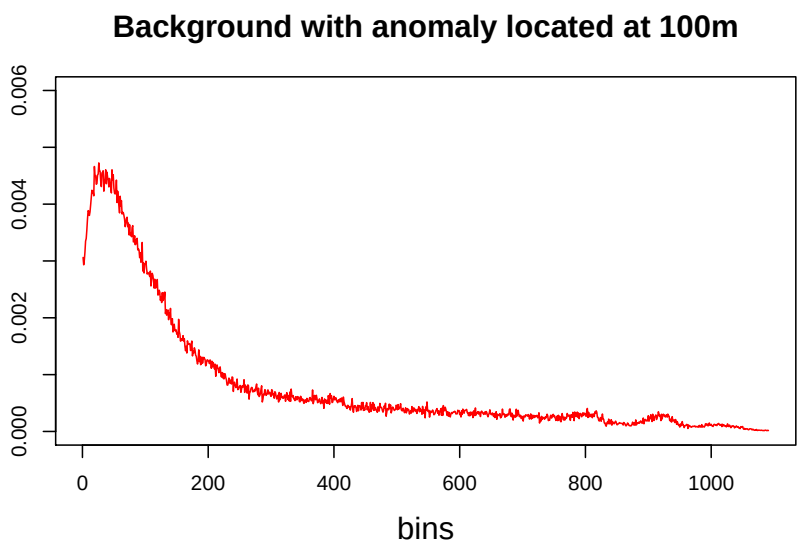
Dist.	$E(N_t)$	KS	KS*	SCR	PKS	EF	GLR
50m	100	1.3	1.6	200	19.0	7.1	5.9
50m	500	1.0	1.0	1.0	9.2	1.9	1.7
50m	1000	1.0	1.0	1.0	5.9	1.2	1.1
100m	100	9.8	12.0	200	66.8	24.1	19.7
100m	500	2.6	3.1	8.7	30.6	9.0	8.7
100m	1000	1.6	1.8	1.1	19.6	6.9	6.4
150m	100	111.2	161.3	146.5	208.0	143.4	117.9
150m	500	19.6	25.4	188.7	88.8	28.8	27.4
150m	1000	9.4	13.4	167.3	69.7	18.9	18.8

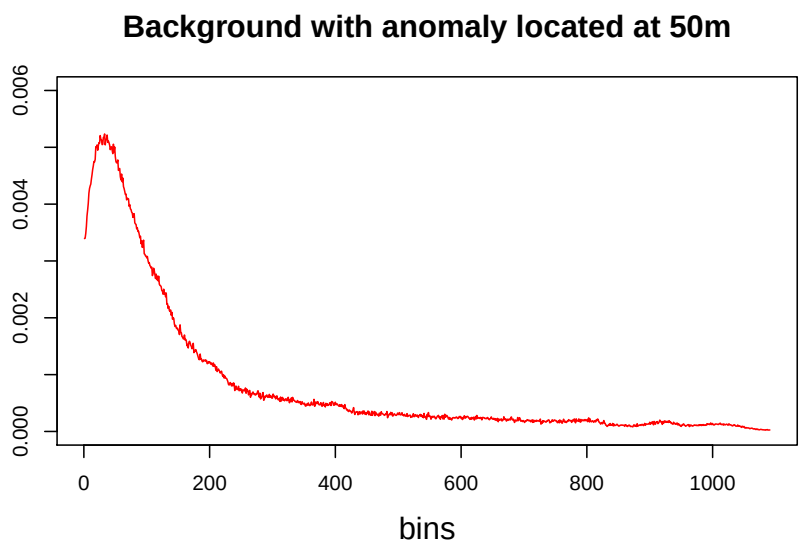
Table: Average time to detection for the cesium example.

A cobalt anomaly









A cobalt anomaly

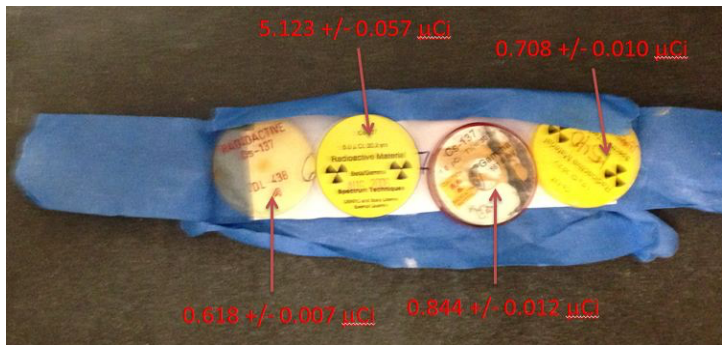
Dist.	$E(N_t)$	KS	KS*	SCR	PKS	EF	GLR
50m	100	2.3	2.7	200	26.7	13.5	17.7
50m	500	1.0	1.0	21.4	12.9	7.6	8.9
50m	1000	1.0	1.0	1.0	8.1	5.1	5.2
100m	100	5.0	5.5	200	44.1	21.6	28.4
100m	500	1.4	1.6	170.1	20.9	12.1	14.4
100m	1000	1.0	1.1	7.8	13.0	10.0	9.9
150m	100	21.1	23.9	200	98.6	57.0	111.5
150m	500	4.9	5.9	194.9	46.3	25.7	31.0
150m	1000	2.9	3.3	168.5	28.7	22.0	21.6

Table: Average time to detection for the cesium example.

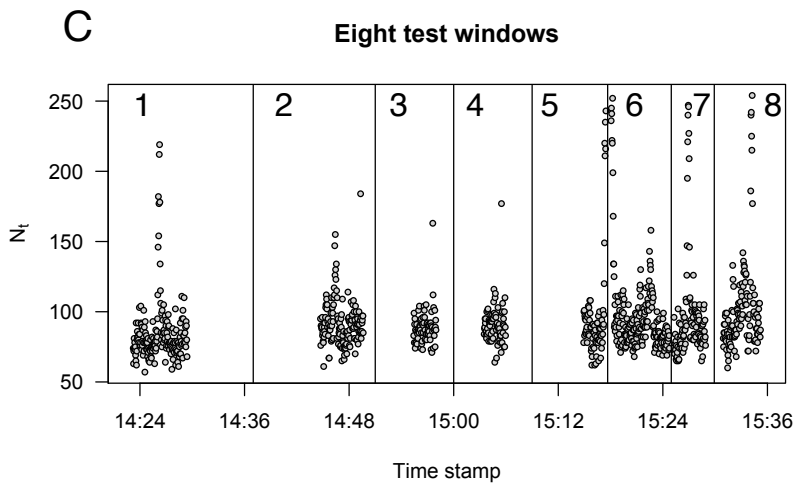
A real field experiment

A

B



C



A real field experiment

Test window	KS	KS*	PKS	EF	GLR	SCR
1	16	16	16	24	103	88
2	8	19	19	22	22	124
3	9	9	23	56	56	∞
4	17	17	25	∞	∞	∞
5	55	55	63	76	76	76
6	5	5	6	12	12	147
7	16	17	16	52	52	49
8	29	29	22	98	97	95

Table: Time to detection (measured by the number of discrete two-second time steps required to raise an alarm). A detection time of ∞ means that method was not able to detect the anomaly.

Summary

Our anomaly-detection method:

- Can be deployed in a streaming-data scenario.
- Has well-understood theoretical properties.
- Is easy to calibrate.
- Improves upon the state of the art.

Part 2: a Bayesian formulation

A full Bayes formulation would involve:

1. a prior for f_C, the space of possible post-change densities.
2. a set of prior probabilities over possible change-points (perhaps out to a certain lag)

We haven't tried to formulate model probabilities.

We've just focused on a prior for f_C, and investigated the behavior of the Bayes factor as a test statistic against the null.

Part 2: a Bayesian formulation

Our approach:

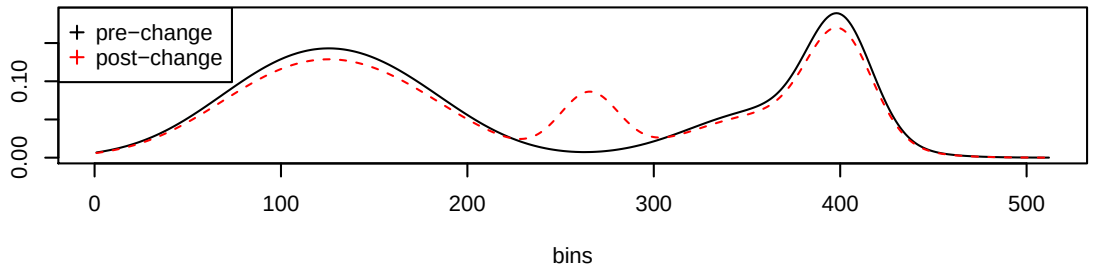
- Based on a Polya-tree prior.
- In our examples, the pre- and post-change densities are similar.
- It is therefore natural to "center" the alternative at the null, e.g. Berger and Guglielmi (2001).

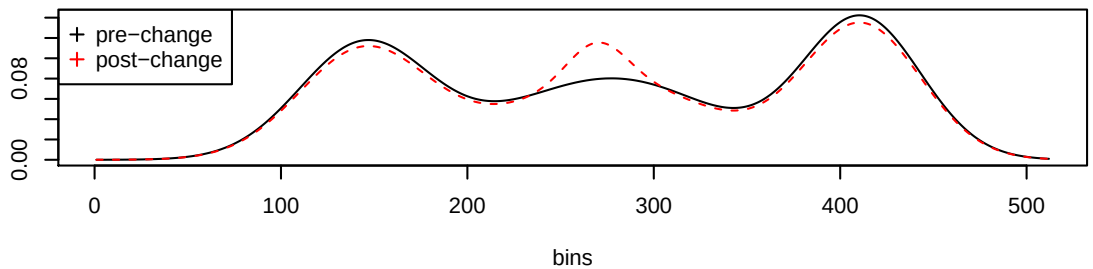
Suppose we want to ask: did the change-point just happen?

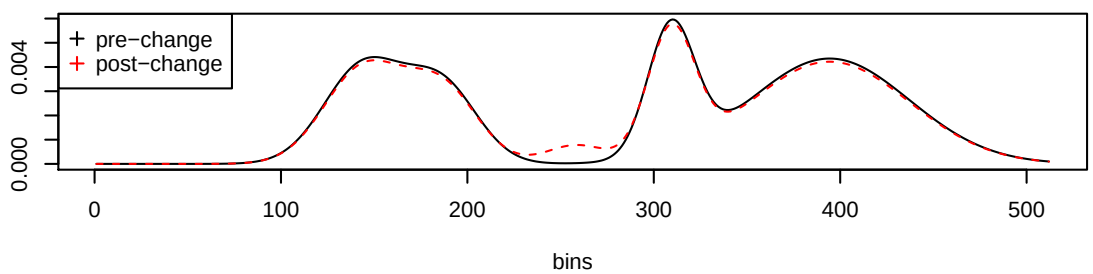
- $H_0 : y_i \sim f_0$ for $i = 1, \ldots N$
- $H_1 : (y_i \mid f_1) \sim f_1$ for $i = 1, \ldots N$, and $f_1 \sim PT(f_0, \alpha)$.
- Fix the partitioning subsets as the dyadic quantiles of f_0 (canonical centering)
- α optionally concentrates the beta-distribution splitting probabilities

And so on for lag 2, lag 3, etc...

Some simulated examples







Example 1

d	$\frac{E(N_t)}{D}$	KS	SCR	PT .1	PT .3	PT .5	PT .7
7	1.5	4.9	255.2	1.7	1.9	1.7	1.7
7	2.0	4.3	254.2	1.4	1.8	1.7	1.3
7	2.5	3.3	158.6	1.3	1.4	1.2	1.3
8	1.5	2.7	246.8	1.3	1.1	1.1	1.1
8	2.0	2.4	130.5	1.0	1.0	1.1	1.0
8	2.5	2.7	2.4	1.0	1.0	1.0	1.0
9	1.5	1.9	247.7	1.0	1.0	1.0	1.0
9	2.0	1.5	77.1	1.0	1.0	1.0	1.0
9	2.5	1.3	1.1	1.0	1.0	1.0	1.0

Example 2

d	$\frac{E(N_t)}{D}$	KS	SCR	PT .1	PT .3	PT .5	PT .7
7	1.5	24.4	247.6	32.3	35.9	19.4	27.6
7	2.0	20.5	254.8	22.0	16.8	14.0	19.5
7	2.5	12.6	254.5	12.1	13.5	11.1	10.9
8	1.5	13.9	258.1	29.0	7.6	10.7	17.1
8	2.0	10.1	258.3	21.8	6.0	6.6	11.5
8	2.5	8.7	256.1	8.4	7.4	8.7	7.5
9	1.5	7.7	244.3	34.5	21.3	7.9	8.7
9	2.0	6.0	238.5	7.2	5.7	4.3	4.7
9	2.5	5.4	240.9	3.8	3.1	3.2	3.2

Example 3

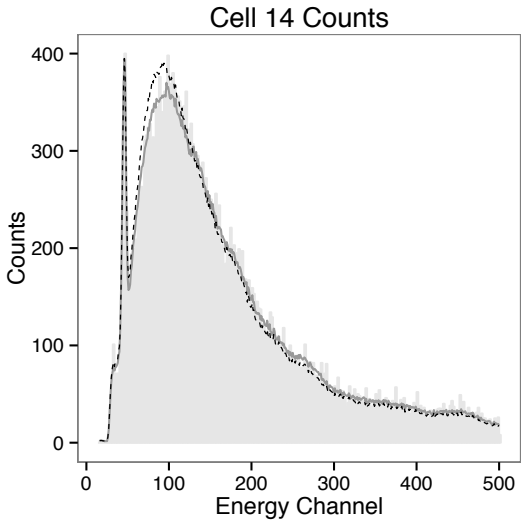
d	$\frac{E(N_t)}{D}$	KS	SCR	PT .1	PT .3	PT .5	PT .7
7	1.5	37.2	260.2	3.1	3.4	3.1	3.6
7	2.0	31.5	250.5	2.8	2.8	2.7	3.6
7	2.5	28.7	257.5	2.9	3.0	2.8	2.5
8	1.5	23.6	271.1	2.6	3.2	3.1	2.8
8	2.0	13.2	253.6	2.0	2.3	1.9	2.3
8	2.5	14.9	250.5	1.6	1.9	1.8	1.8
9	1.5	9.1	239.5	1.6	1.8	1.8	1.9
9	2.0	9.3	257.3	1.5	1.1	2.0	1.6
9	2.5	8.2	254.9	1.1	1.3	1.3	1.2

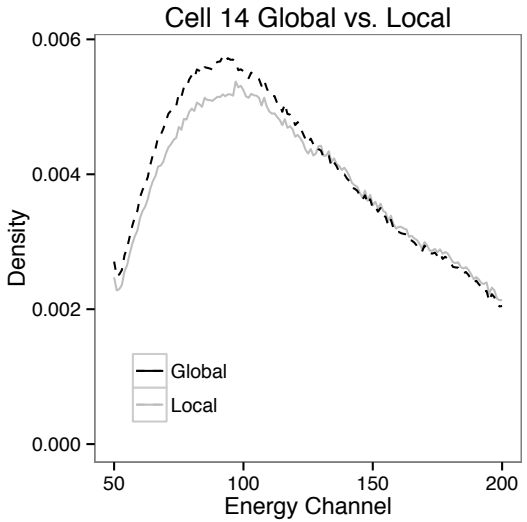
Thank you!

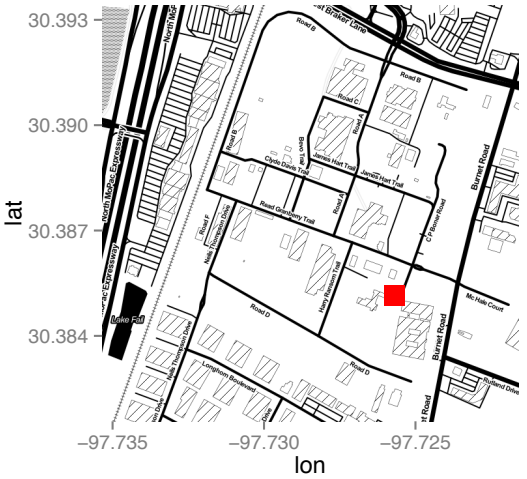
Multiscale spatial density smoothing: an application to large-scale radiological survey and anomaly detection. W. Tansey, A. Athey, A. Reinhart, and James G. Scott. *Journal of the American Statistical Association* 112(519): 1047–63 (2017).

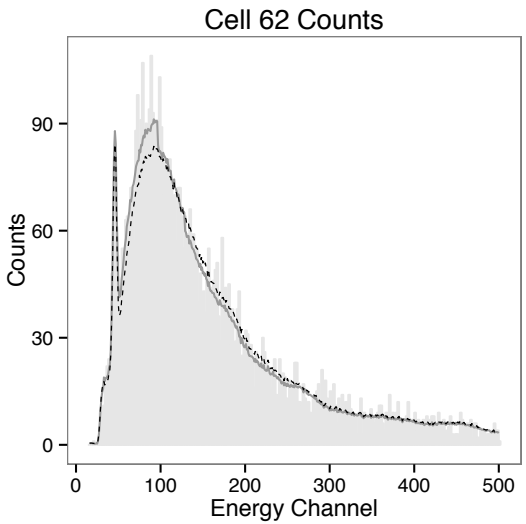
Sequential nonparametric tests for a change in distribution: an application to detecting radiological anomalies. O.H.M. Padilla, A. Athey, A. Reinhart, J.G. Scott. arXiv:1612.07867

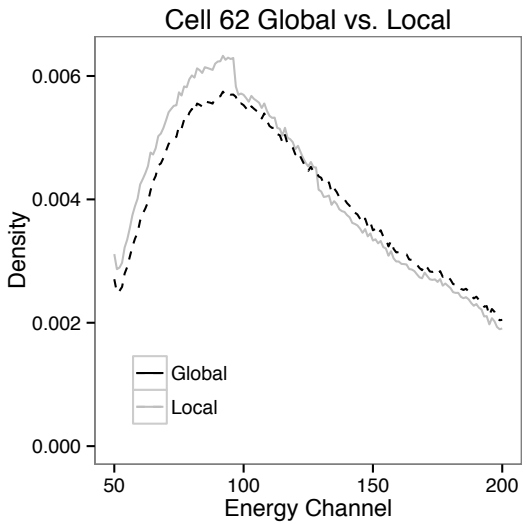
Spatial variation in spectrum











Two different locations.

Multiscale spatial density smoothing

The idea: motivated by Pólya trees (c.f. Hanson and Yang, 2007).

- **Split** into sub-problems via recursive partitioning.
- **Smooth** the half-space probabilities over the spatial lattice using binomial graph trend filtering.
- **Merge** the smoothed probabilities to yield $\hat{f}_0^{(s)}$, $s \in \mathcal{V}$.
- Reserve the power/hassle/expense of full Bayes analysis for the other parts of the pipeline.

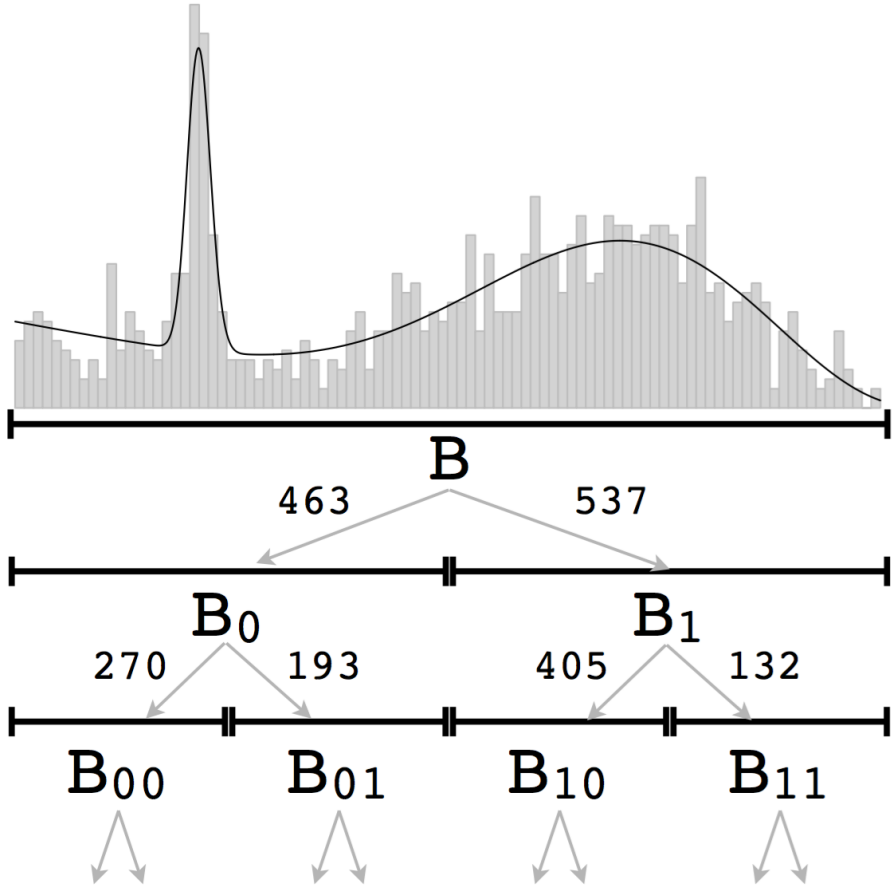
Notable points:

- Reduces the functional smoothing problem to a set of embarrassingly parallel scalar smoothing problems.
- Extremely fast and scalable to very large data sets (dominant cost = loading data into memory).

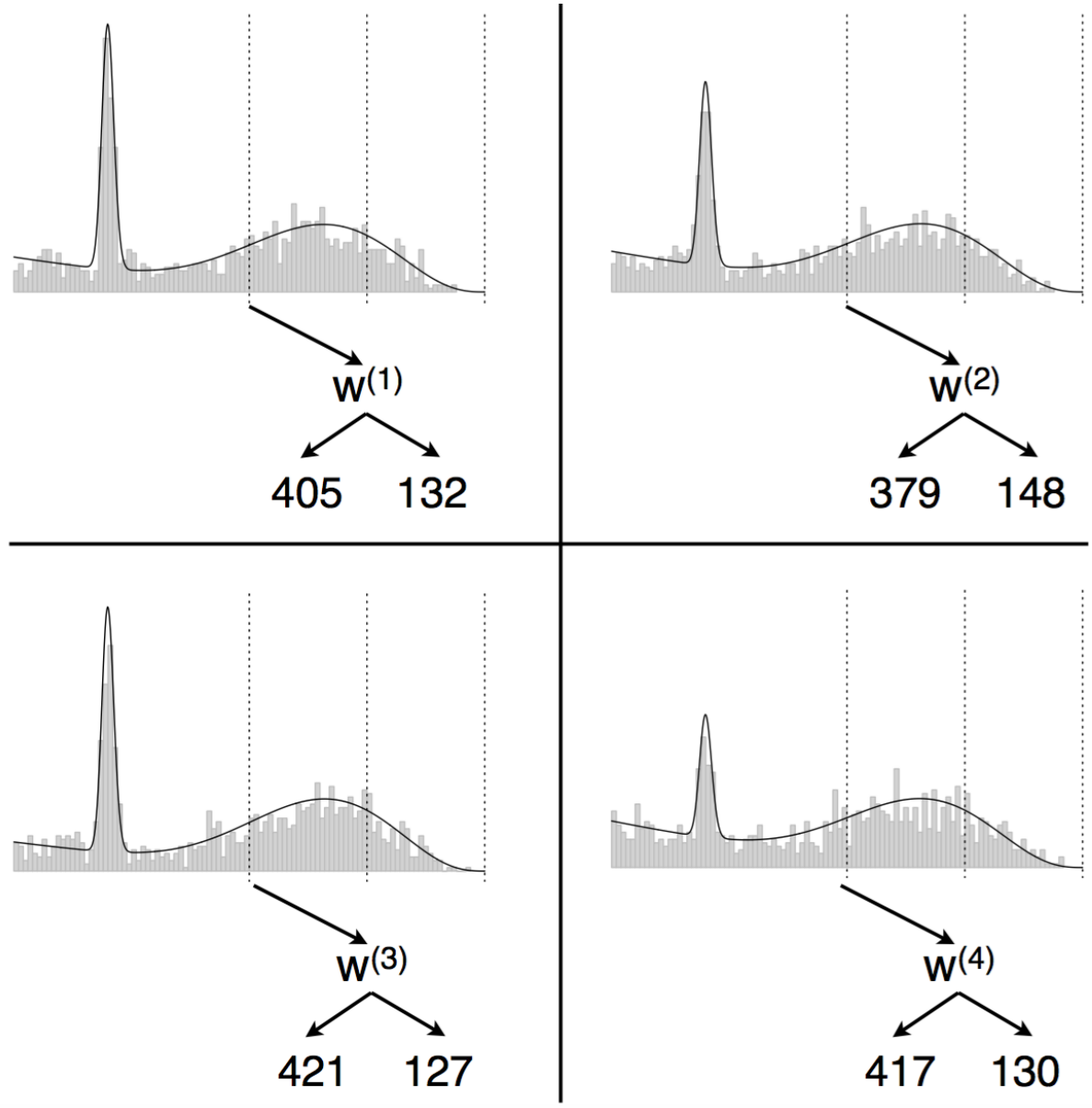
Recursive dyadic partitions

Let $(x_1, \ldots, x_n)$ be a sample from $f(x)$.

- n_γ: number of samples in the parent set B_γ.
- $y_{\gamma 0}$: number of samples in the left child set $B_{\gamma 0}$.



Spatial variation: a 2x2 example



Spatial variation

Now consider a specific split in the tree and drop the γ index. We want to estimate the "left-child" splitting probability across all spatial sites in our graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

$$y^{(s)} \sim \text{Binom}\left(n^{(s)}, \frac{e^{\beta^{(s)}}}{1 + e^{\beta^{(s)}}}\right), \quad s \in \mathcal{V}$$

We enforce spatial smoothness by solving the following optimization problem for all splitting nodes, in parallel:

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad \sum_{s \in \mathcal{V}} \left\{ n^{(s)} \log\left(1 + e^{\beta^{(s)}}\right) - y^{(s)} \beta^{(s)} \right\} + \lambda \|D\beta\|_1$$