Variance change point detection with credible sets

Lorenzo Cappello

Joint work with Oscar Madrid Padilla (UCLA)

arXiv:2211.14097

StatScale ECR Meeting, Brighton,
Dec 14 to 16 2022

upf. Universitat Pompeu Fabra Barcelona

BSE Barcelona School of Economics

An invite not to leave Bayes Stat behind

- Tremendous expansion of the literature in change-point detection
- This has not interested much Bayesian statistics. Comparably, very little work [Fearnhead ’06, Liu et al. ’20, Wang et al. ’20, C. et al. ’21]

✗ Computationally expensive

✗ (Almost) No theoretical guarantees
  - On the properties of the estimator (*e.g.*, localisation rate)
  - On the finite sample convergence of the algorithms used for inference (*e.g.* MCMC)

- Should we care?

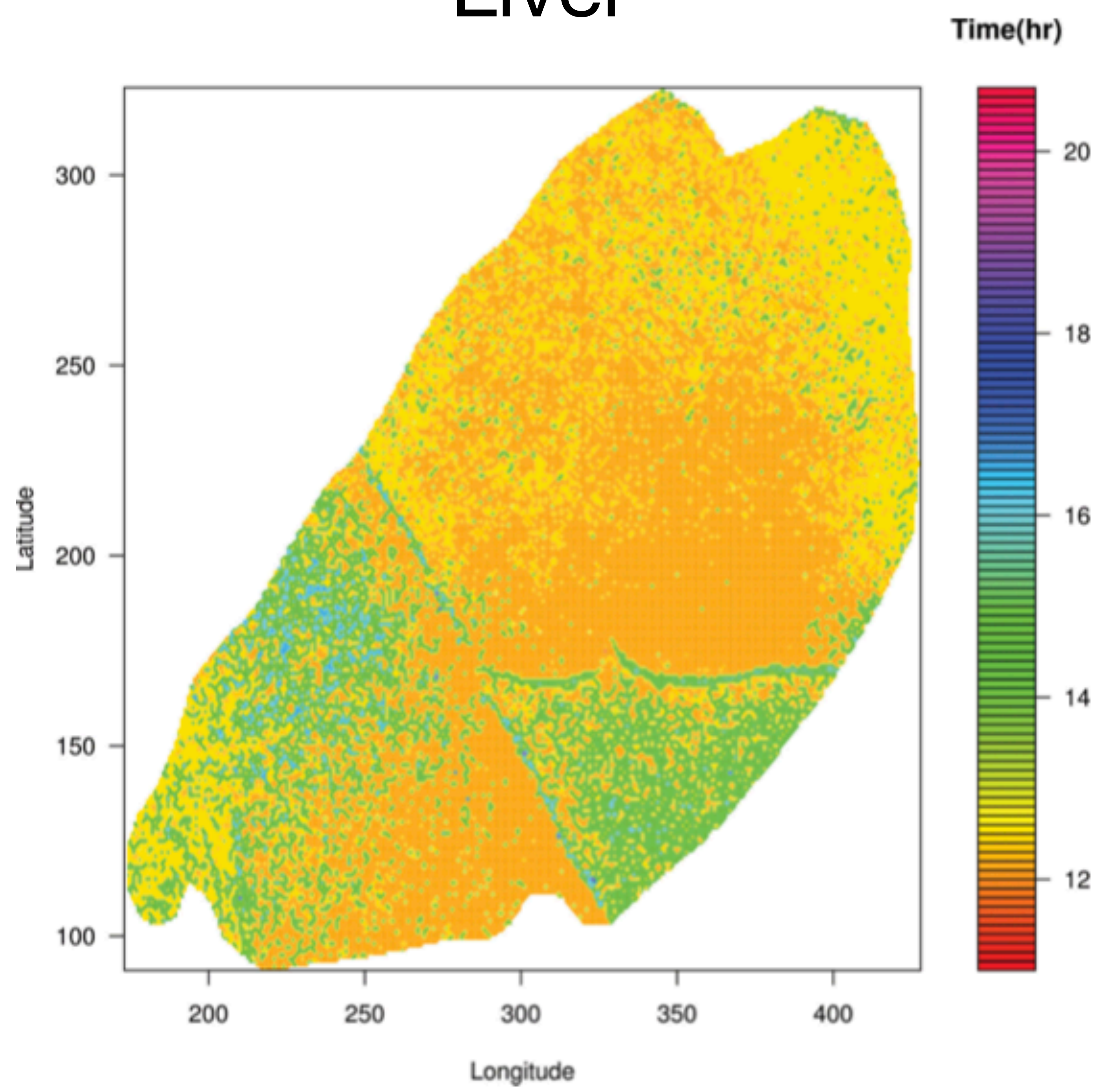
+ Natural Uncertainty quantification

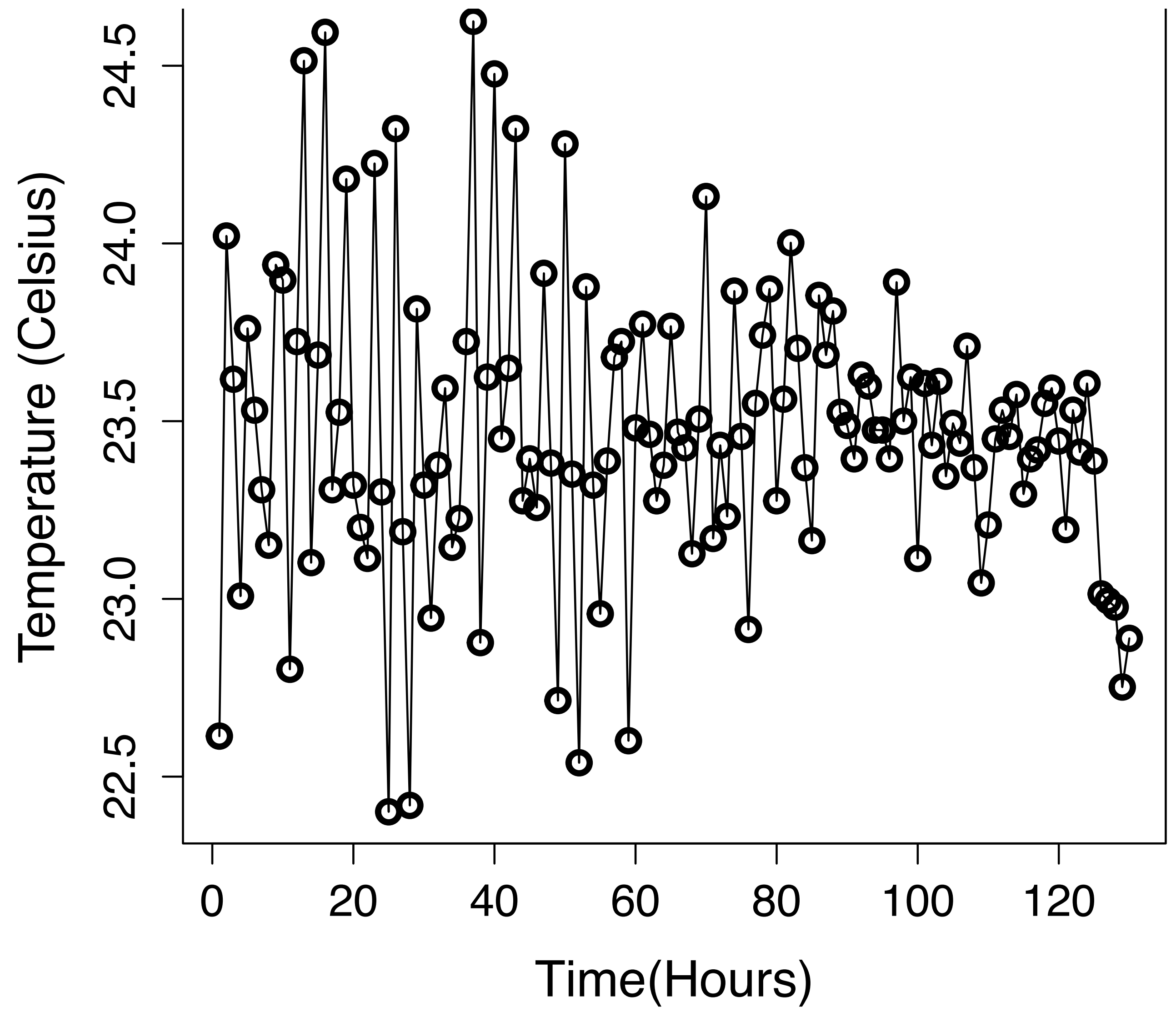
+ Modular/Generalizible

Today

- New Bayesian procedure to estimate changes in the variance of a Gaussian sequence
  - ✶ Point estimates and credible sets
  - ✶ Offline setting
  - ✶ Fast algorithm for inference
  - ✶ Theory
- **Motivation:** Liver procurement [Gao et al, 2019]
  - ✶ Surface temperature avoids invasive biopsy
  - ✶ Less risk to ruin the organs
- Many of the ideas generalise to other settings (hopefully we will discuss at least one)

Liver





Surface temperature at a randomly location

Detection of single change in variance with UQ

We collect T observations from a Gaussian sequence with a change in variance at t^*

A working (Bayesian) model is

Change point (cp) location

$$\begin{cases} Y_i \mid \gamma_t = 1, \sigma^2 \sim N(0, \sigma^2) & \text{if } \; 1 \le i < t, \\ Y_i \mid \gamma_t = 1, \sigma^2, \tau \sim N(0, \tau^{-2}\sigma^2) & \text{if } \; t \le i \le T. \end{cases}$$

Likelihood $\mathscr{L}(Y_{1:T} \mid \gamma, \sigma, \tau)$

Unknown scaling parameter

$\boldsymbol{\gamma} \sim$ Categorical$(T^{-1}, \ldots, T^{-1})$ Prior on cp location

$\tau^2 \mid a_0 \sim$ Gamma(a_0, a_0) Prior on scale parameter

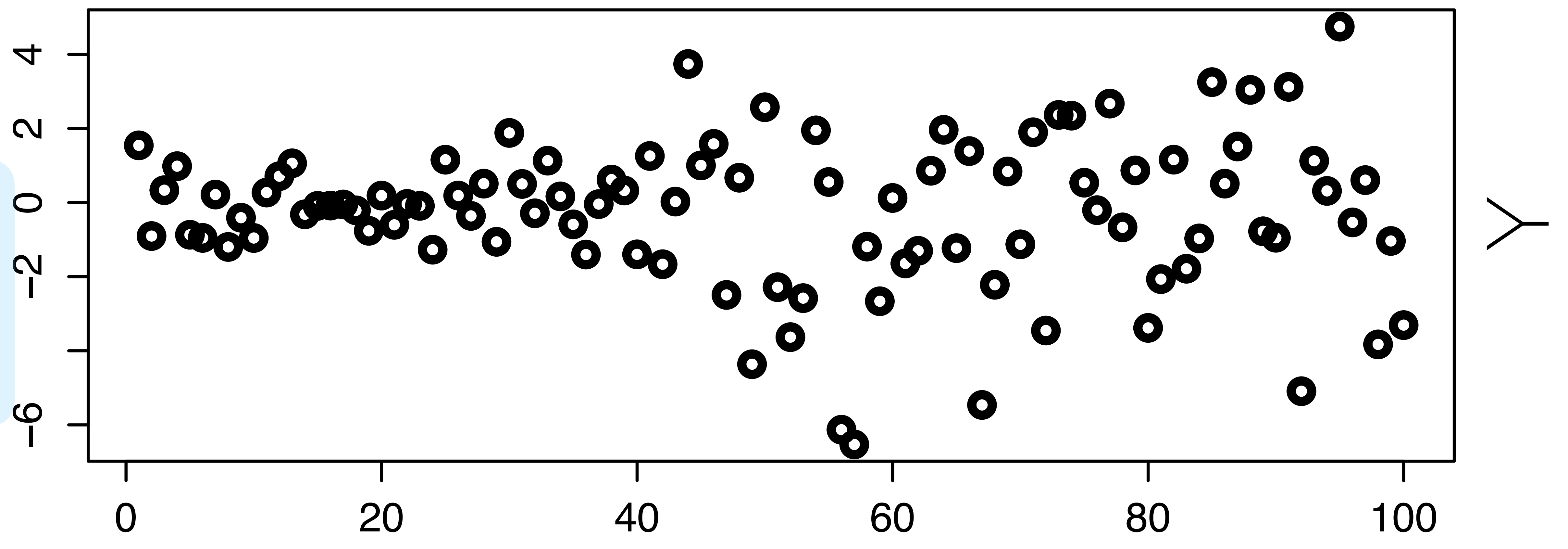
Prior $\pi(\sigma, \tau)$

Bayesian change point detection

In the spirit of [Chernoff Zacks,'64; Smith, '75; Raftery Akman, '86; Wang et al. '20]

- Posterior available in closed form

$$P(\gamma_t = 1 \,|\, y_{1:T}, \sigma^2) = \frac{P(y_{1:T} \,|\, \gamma_t = 1, \sigma^2)}{\sum_i P(y_{1:T} \,|\, \gamma_i = 1, \sigma^2)}$$



- I.e. minimal computations
- $\boldsymbol{\gamma}$ naturally describes uncertainty on change point location

Point estimate and credible sets

- Change-point detection here is an estimation problem, not a testing problem
- A point estimates could be $\max_t P(\gamma_t = 1 \mid y_{1:T}, \sigma^2)$
- Let $\alpha_t = P(\gamma_t = 1 \mid y_{1:T}, \sigma^2)$, we can a ***credible sets*** *of level p*, describing the uncertainty. $\mathcal{CS}(\alpha, p) := \arg \min_{S \subset \{1,\ldots,T\} : \sum_{t \in S} \alpha_t > p} |S|$.
- Variance estimates are also available

$$\bar{\tau}^2 = E[\tau^2] = \left(\alpha_1 \hat{s}_1^2 + 1 - \alpha_1, \alpha_1 \hat{s}_1^2 + \alpha_2 \hat{s}_2^2 + 1 - \alpha_1 - \alpha_2, \ldots, \sum_{i=1}^{t} \alpha_i \hat{s}_i^2 + 1 - \sum_{i=1}^{t} \alpha_i, \ldots, \sum_{i=1}^{T} \alpha_i \hat{s}_i^2 \right),$$

Theoretical guarantees

Thm. [C. & Padilla, '22] Under mild conditions, the Bayesian point estimator described attains a minimax localization rate* up for a logarithm factor for a single change in variance of a Gaussian sequence

- Minimax localization rate for variance is $\sqrt{T \log T}$ [Wang et al. 2021]
- *: the rate is in the multiple change-point case

Towards multiple change point

Single change-point model

$$\begin{cases} Y_i \mid \gamma_t = 1, \sigma^2 \sim N(0, \sigma^2) & \text{if} \quad 1 \le i < t, \\ Y_i \mid \gamma_t = 1, \sigma^2, \tau \sim N(0, \tau^{-2}\sigma^2) & \text{if} \quad t \le i \le T. \end{cases}$$

$$\boldsymbol{\gamma} \sim \text{Categorical}(T^{-1}, \ldots, T^{-1})$$

$$\tau^2 \mid a_0 \sim \text{Gamma}(a_0, a_0)$$

Two change-points model

2nd change point location

$$\begin{cases} Y_i \mid \gamma_{t,1} = 1, \gamma_{s,2} = 1, \sigma^2 \sim N(0, \sigma^2) & \text{if} \quad 1 \le i < t, \\ Y_i \mid \gamma_{t,1} = 1, \gamma_{s,2} = 1, \sigma^2, \tau_1 \sim N(0, \tau_1^{-2}\sigma^2) & \text{if} \quad t \le i < s, \\ Y_i \mid \gamma_{t,1} = 1, \gamma_{s,2} = 1, \sigma^2, \tau_1, \tau_2 \sim N(0, \tau_1^{-2}\tau_2^{-2}\sigma^2) & \text{if} \quad s \le i < T, \end{cases}$$

$$\boldsymbol{\gamma}_1 \sim \text{Categorical}(T^{-1}, \ldots, T^{-1})$$

$$\boldsymbol{\gamma}_2 \sim \text{Categorical}(T^{-1}, \ldots, T^{-1})$$

$$\tau_1^2 \mid a_0 \sim \text{Gamma}(a_0, a_0)$$

$$\tau_2^2 \mid a_0 \sim \text{Gamma}(a_0, a_0)$$

2nd Unknown scaling parameter

PRISCA: PRoduct Single SCAle effect

Arbitrary number L of change points

$$
\begin{cases}
Y_i \mid \gamma_{t_1,1} = 1, \ldots, \gamma_{t_L,L} = 1, \sigma^2, \tau_1, \ldots, \tau_L \sim N(0, \sigma^2) & \text{if} \quad 1 \leq i < t_1, \\
\ldots \\
Y_i \mid \gamma_{t_1,1} = 1, \ldots, \gamma_{t_L,L} = 1, \sigma^2, \tau_1, \ldots, \tau_L \sim N(0, \prod_l \tau_l^{-2} \sigma^2) & \text{if} \quad t_L \leq i < T,
\end{cases}
$$

$$\boldsymbol{\gamma}_l \sim \text{Categorical}(T^{-1}, \ldots, T^{-1})$$

$$\tau_l^2 \mid a_0 \sim \text{Gamma}(a_0, a_0)$$

- L is like an upper bound on the number of change-points
- a_0 shared and center the mean at 1

Fitting PRISCA

- Ideally, we would like $p(\tau^2_{1:L}, \gamma_{1:L} \,|\, y_{1:T}, \sigma^2)$ or marginals $p(\tau^2_l, \gamma_l \,|\, y_{1:T}, \sigma^2)$ for all l
- The model we wrote is *conditionally conjugate*, *i.e.*, given $(\tau^2_i, \gamma_i)_{i \neq l}$, we can do an update.
- This suggests an easy Gibbs sampler.
- We don't want to do that
- If I had $(\bar{\tau}^2_i)_{i \neq l}$, we could compute **residuals** $\boldsymbol{r}^2_l = \boldsymbol{y}^2 \circ \prod_{i \neq l} E[\boldsymbol{\tau}^2_i]$ and the posterior distribution $p(\tau^2_l, \gamma_l | \boldsymbol{r}^2_l, \sigma^2)$ [similar to Wang et al, 2020]
- The idea/hope is that $p(\tau^2_l, \gamma_l | \boldsymbol{r}^2_l, \sigma^2)$ is a good approximation for $p(\tau^2_l, \gamma_l \,|\, y_{1:T}, \sigma^2)$

Algorithm 1

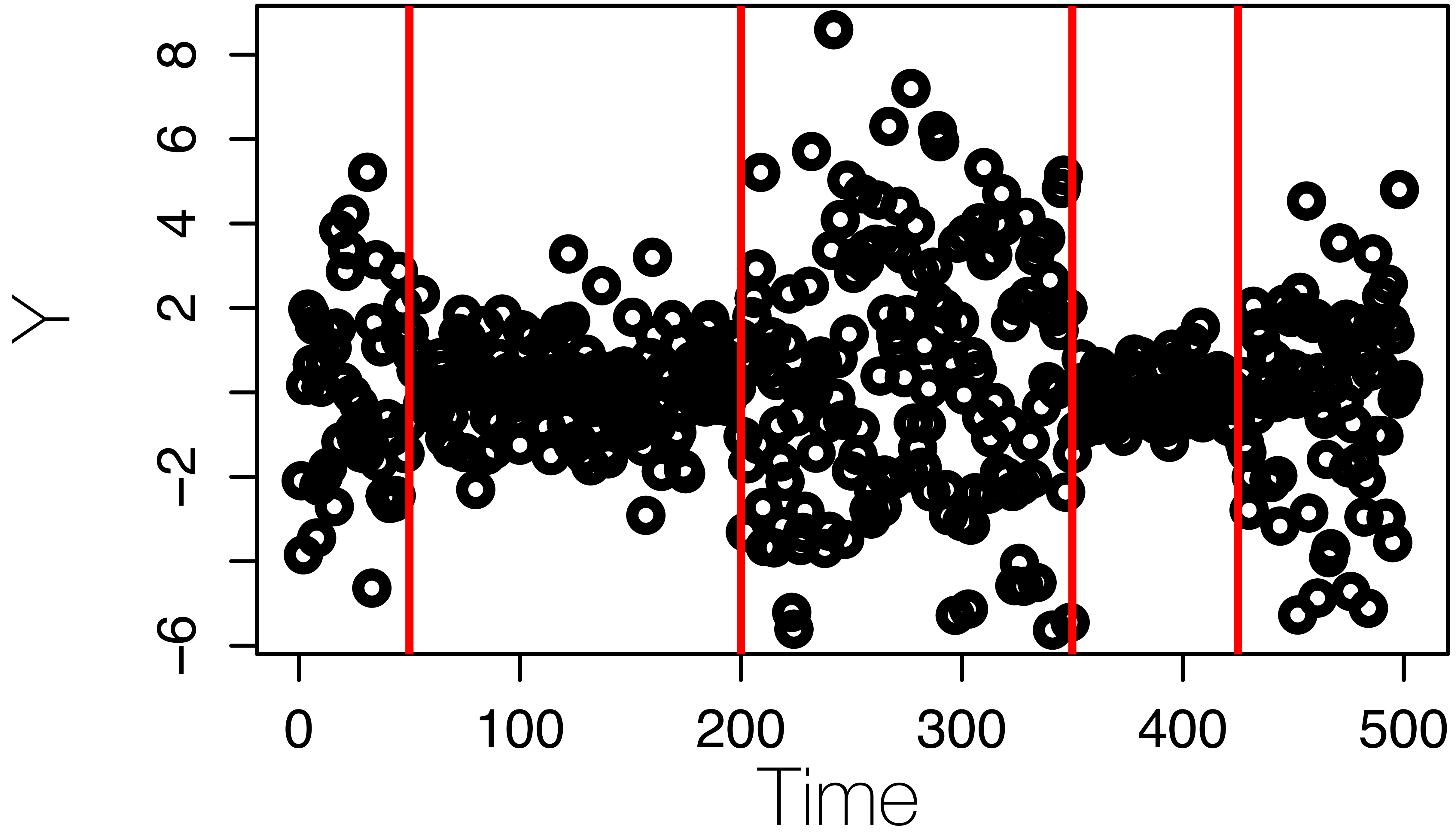
Input: L, a_0

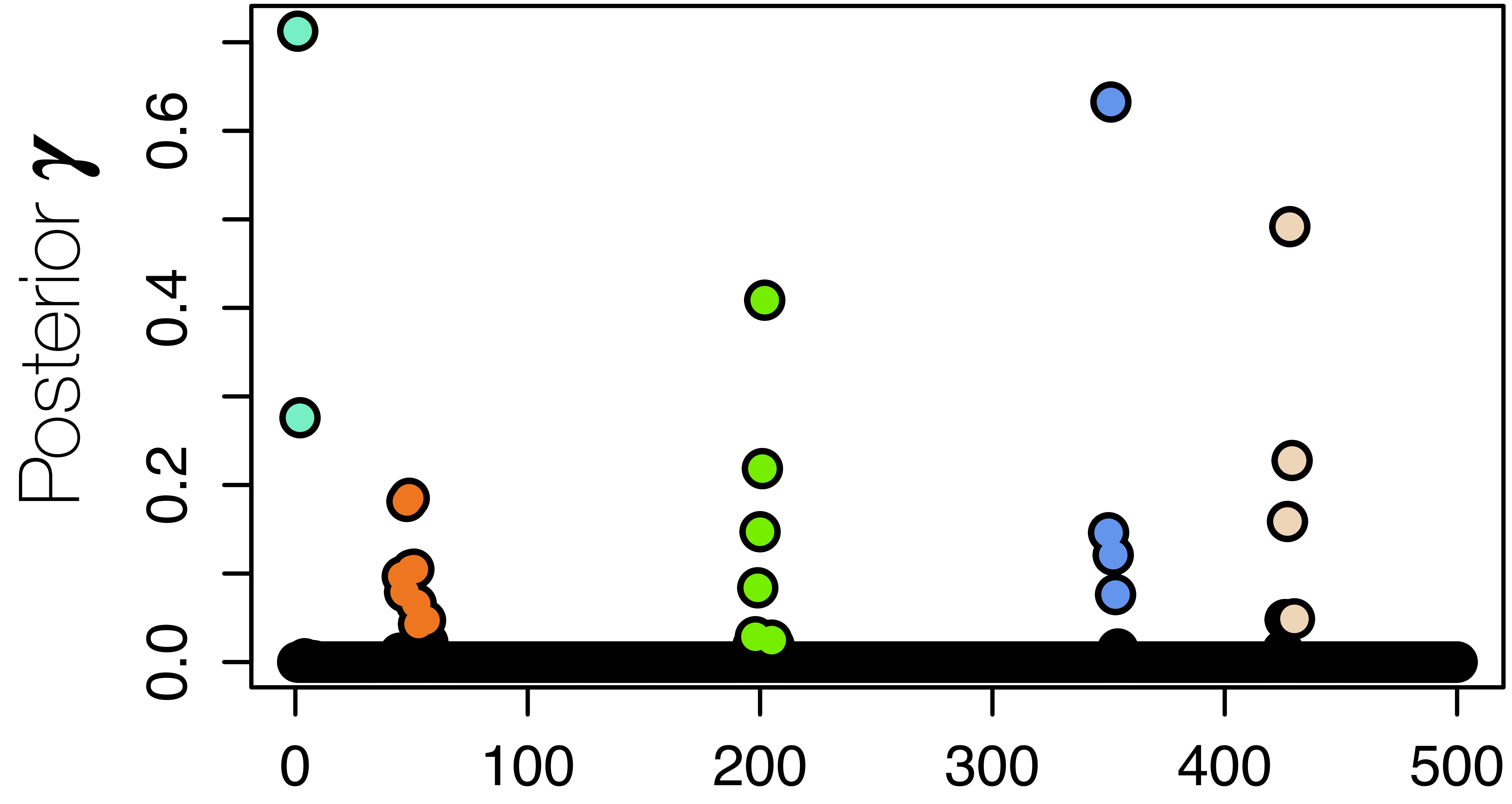
0. Initialize $\bar{\boldsymbol{\tau}}_l^2 = E[\boldsymbol{\tau}_l^2] = 1$ for all l
1. For *l* in *1:L* repeat
   a. *Compute residuals:* $\boldsymbol{r}_l^2 = \boldsymbol{y}^2 \circ \prod_{l' \neq l} \bar{\boldsymbol{\tau}}_{l'}^2$
   b. *Fit the single change point to the residuals*: compute posterior $\boldsymbol{\tau}_l^2$ and $\boldsymbol{\gamma}_l$
   c. Update $\bar{\boldsymbol{\tau}}_l^2$
2. Repeat 1 until convergence (*backfitting*)

Notation
$\mathbf{x}^k = (x_i^k)_{1:T}$

<u>Output is posterior distribution of</u> $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L$ and $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_L$

Output





- T=500
- L set to 9

Simulations

- Essential taken from the PELT paper [Killick et al. ‘12]
  - ✦ Random change point locations and variances
  - ✦ Various sample sizes, #changepoints, and multiple datasets per combination

- Compared to
  - ✦ PELT (dynamic programming)
  - ✦ Binary Segmentation [Scott Knott ’74]
  - ✦ Segment Neighbourhood (dynamic programming) [Auger Lawrence ’89]

- We report bias ($K - \widehat{K}$), Hausdorff-like quantity ($d(\widehat{\mathscr{C}} \mid \mathscr{C}^*) = \max_{\eta \in \mathscr{C}^*} \min_{x \in \widehat{\mathscr{C}}} |x - \eta|$)

Simulations: results

- Averages across T
- PRISCA has $L = \lfloor \sqrt{T}/30 \rfloor$
- ora-PRISCA $L = K$
- auto-PRISCA automatic L
- $a = 10^{-3}$

Method	$K - \widehat{K}$	$d(\widehat{\mathcal{C}}, \mathcal{C}^*)$	Time
auto-PRISCA	2.11	124.49	0.91
BINSEG	2.96	212.28	0
PELT	2.6	176.42	0
PRISCA	1.98	131.44	0.66
ora-PRISCA	1.9	118.17	0.07
SEGNEI	2.42	193.2	0.36

Why it works?

Prop1. [C. & Padilla, '22] Algorithm 1 is a specific Variational approximation to the model presented

- Intuition from [Wang et al, 2020]
- This means that the algorithm is a coordinate ascent
- For PRISCA we can compute the ELBO to have convergence criterion

Prop2. [C. & Padilla, '22] Algorithm 1 converges to a limit point that is a stationary point of the objective function.

Modular algorithm easy to generalise

Input: $\boldsymbol{L}, \boldsymbol{a_0}$

0. Initialisation
1. A. For *l* in *1:L*, repeat
    a. *Compute residuals:* $\boldsymbol{r}_l^2 = \boldsymbol{r}_B^2 \circ \prod_{l' \neq l} E[\boldsymbol{\tau}_{l'}^2]$
    b. *Fit the single change point to the residuals*: compute posterior $\boldsymbol{\tau}_l^2$ and $\boldsymbol{\gamma}_l$

   B. a. *Fit any arbitrary procedure that "may benefit" from variance estimates*
    b. *Compute Compute residuals:* $\boldsymbol{r}_B$
2. And 3. Same as before

- e.g. "may benefit" is something to solve with weighted least square
- e.g. autoregression or smooth trend

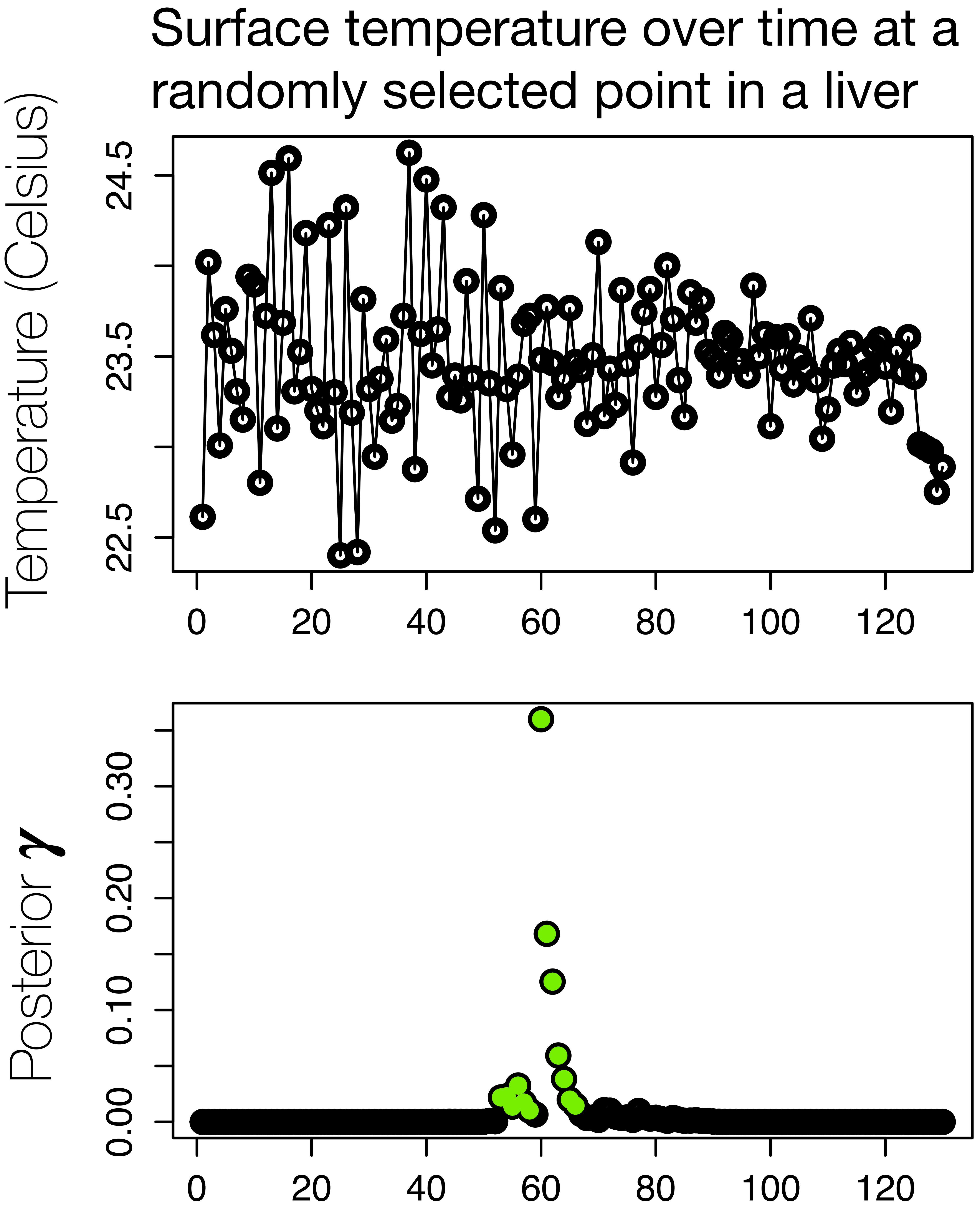
E.g. Trendfiltering with heteroskedasticty

- $Y_t = f_t + \epsilon_t$ with f_t "smooth" and $(\epsilon_t)_{1:T}$ piecewise constant taking K values
- B. Can be [Tibshirani,'14] trend filtering solved with weighted least squares
- Set up of [Gao et al. '19] generalised to multiple K
- Simulation
  - $\mathscr{C} = \{.15\,T, .4\,T, .75\,T, .85\,T\}$
  - $f_t = 20 + 12t/T(1 - t/T)$

T	Method	$K - \widehat{K}$	$d(\widehat{\mathcal{C}}, \mathcal{C}^*)$	Time
200	PRISCA	3	169	0.11
	ora-PRISCA	0.4	32.42	0.04
	pre-PRISCA	3	169	0.01
	TF-PRISCA	0.16	30.44	4.09
500	PRISCA	3	424	0.11
	ora-PRISCA	-0.29	12.14	0.11
	pre-PRISCA	3	424	0.01
	TF-PRISCA	-0.35	15.61	2.45
1000	PRISCA	3	849	0.13
	ora-PRISCA	-0.34	9.74	0.22
	pre-PRISCA	3	849	0.03
	TF-PRISCA	-0.38	9.5	5.02

Liver procurement

- L=4
- Detrended with Tibshirani's (2014) trend filtering
- It is possible to include a trend filtering in the loop of Algorithm 1
- Highest posterior probability at 60. PELT gives 59



Conclusion

- Procedure to estimate multiple changes in the variance of a Gaussian sequence
  - ✓ It is computationally efficient
  - ✓ We still get credible sets
- (Some) Theory available:
  - ✓ Single change point estimators attains minimax rate up to a logarithm factor
  - ✓ Convergence of the algorithm in the multiple change point case

Future directions

Methodological

- Generalizations: count data, multivariate (e.g. covariance), times-series
- Fix “known issues” in Variational Bayes
- Related to the credible sets: FDR control using our posterior estimates [*Bayesian linear programming,* Spector Janson, ’22]

Theory

- Consistency/Localization rate in the multiple change-points case.
- Credible sets frequentist coverage (*Bernstein-von Mises* type of theorems)

Thanks

A draft recently online (arXiv:2211.14097)

lorenzo.cappello@upf.edu comments welcome!

There is a R package as well to try it out

Theoretical guarantees (zoom in)

Assumption 1. *Let t_0 be the time instance such that $Y_t \overset{iid}{\sim} N(0, \sigma_r^2)$ for $t \geq t_0$ and $Y_t \overset{iid}{\sim} N(0, \sigma_l^2)$ for $t < t_0$, and let $\tau^2 = \sigma_l^2/\sigma_r^2$.*

a. *There exists a constant $c > 0$ such that $\min\{t_0, T - t_0\} > cT$.*

b. *For some fixed intervals $I_1 \subset (1, \infty)$ and $I_2 \subset (0, 1)$ we have that $\tau^2 \in I_1 \cup I_2$.*

c. *The hyperparameters are chosen such that $a_0 > 0$ and $\boldsymbol{\pi}$ satisfies that $\pi_t > 0$ for all t and $\sum_t \pi_t = 1$.*

Theorem 1. *Supposed that Assumption 1 holds. Then, for $\epsilon > 0$ there exists a constant $c_1 > 0$ such that, with probability approaching one, we have that*

$$\max_{t:\min\{t, T-t\} > cT, |t-t_0| > c_1\sqrt{T \log^{1+\epsilon} T}} \alpha_t < \alpha_{t_0}.$$

Why it works? Preliminaries

- Based on an intuition in [Wang et al. 2020]
- Let $\boldsymbol{p}$ be the target posterior distribution, for $\boldsymbol{q} \in \mathcal{Q}$, we can see Bayesian posterior computation as an optimization problem:

$$\arg\min_q KL(q \,||\, p) = \arg\min_q [\log p(\mathrm{y} \,|\, \sigma^2) - ELBO(q, \sigma^2, \mathrm{y})] = \arg\max_q ELBO(q, \sigma^2, \mathrm{y})$$

- With no restrictions on $\mathcal{Q}$, the posterior computation is exact: KL=0
- Variational Bayes is an approximation only if $\mathcal{Q}$ is restricted
- Assuming $\boldsymbol{q(\tau)} = \prod_l q_l(\boldsymbol{\gamma}_l, \boldsymbol{\tau}_l)$, we can maximize the ELBO component-wise

$$\arg\max_{q_1} ELBO(q, \sigma^2, \mathrm{y}), \ldots, \arg\max_{q_L} ELBO(q, \sigma^2, \mathrm{y})$$

PRISCA convergence

Prop1. [C. & Padilla, '22] The solution of $\operatorname*{arg\,max}_{q_l} ELBO(q, \sigma^2, \mathrm{y})$ is equal to the solution of the single change point model applied to the residuals $\boldsymbol{r}_l^2 = \boldsymbol{y}^2 \circ \prod_{l' \neq l} E[\boldsymbol{\tau}_l^2]$

- This is exactly what we are doing at each iteration when we fit PRISCA
- This means that the algorithm is a coordinate ascent
- For PRISCA we can compute the ELBO to have convergence criterion

Prop2. [C. & Padilla, '22] Algorithm 1 converges to a limit point that is a stationary point of the objective function.