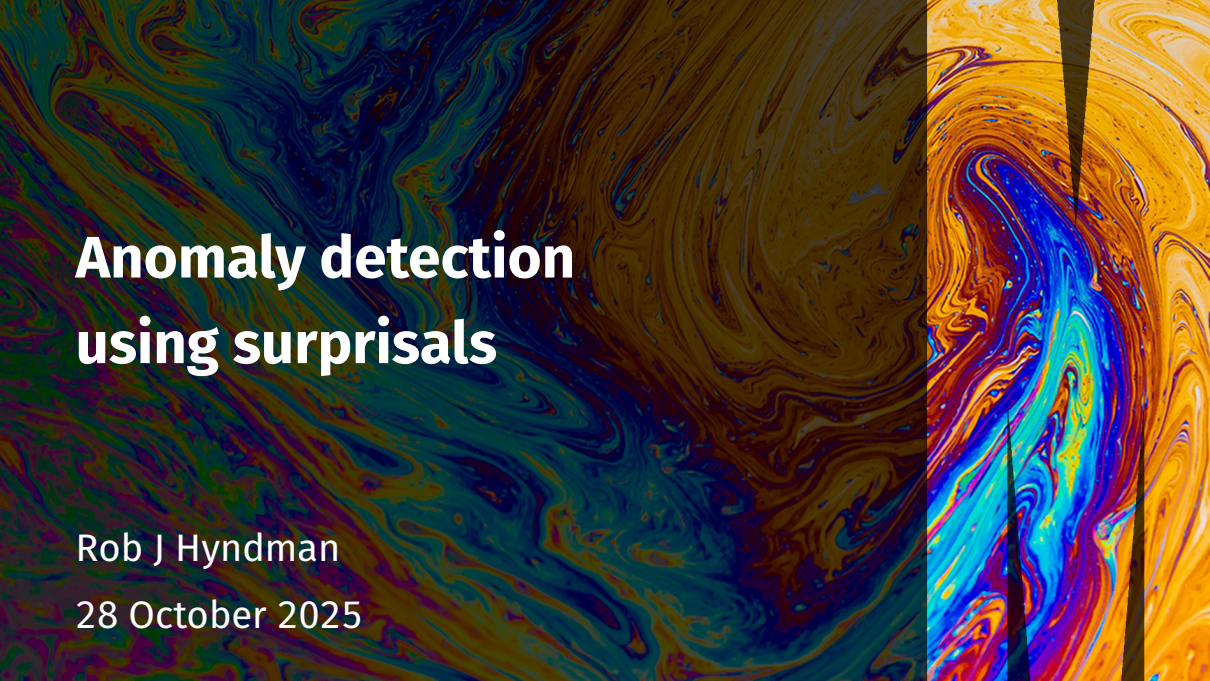


Anomaly detection using surprisals

Rob J Hyndman

28 October 2025

Coauthors

Sevvandi Kandanaarachchi
CSIRO

Kate Turner
ANU

David Frazier
Monash U

Outline

Outline

Definitions of anomalies

an observation (or a subset of observations) which appears to be inconsistent with the remainder of that set of data.

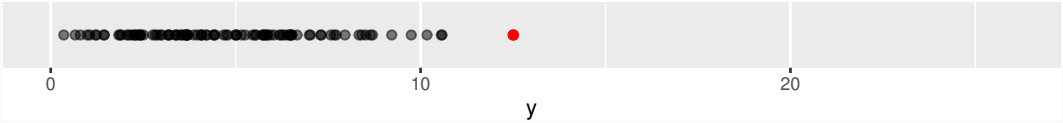
(Barnett & Lewis, 1978)

Definitions of anomalies

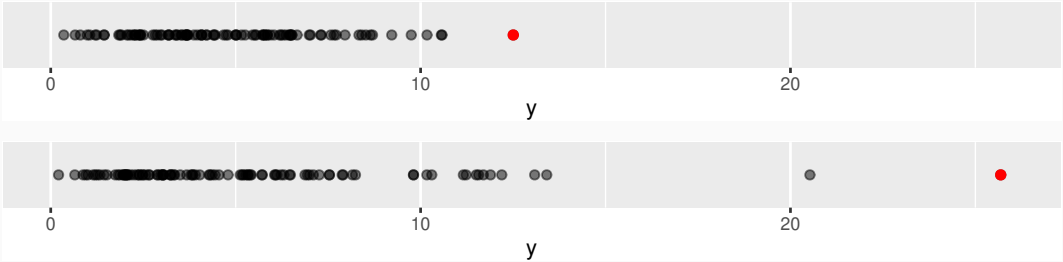
an observation (or a subset of observations) which appears to be inconsistent with the remainder of that set of data.
(Barnett & Lewis, 1978)

an observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism. (Hawkins, 1980)

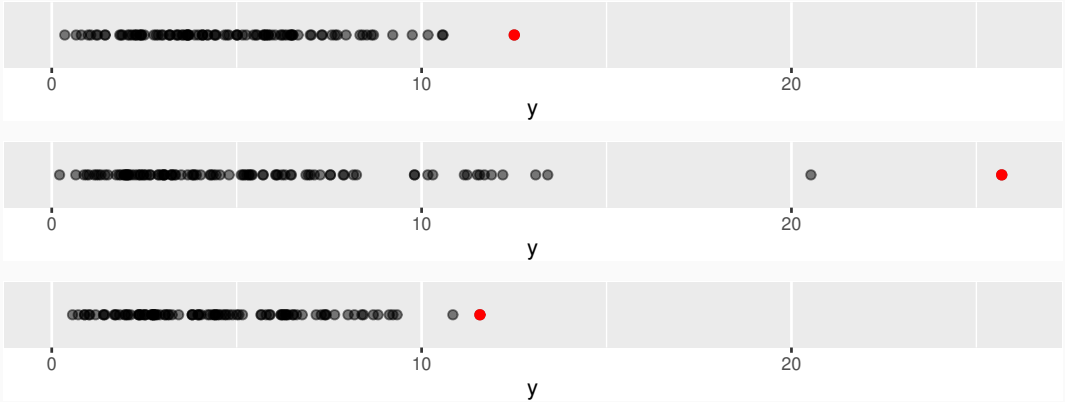
Is this an anomaly?



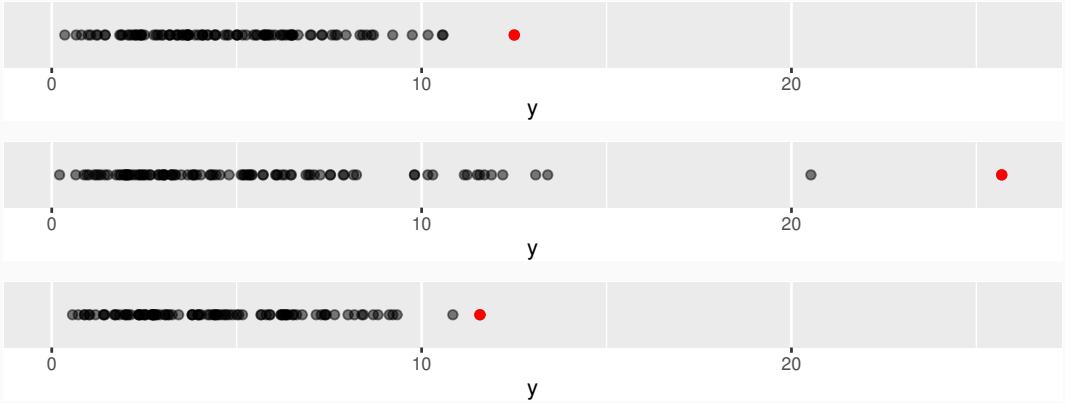
Is this an anomaly?



Is this an anomaly?

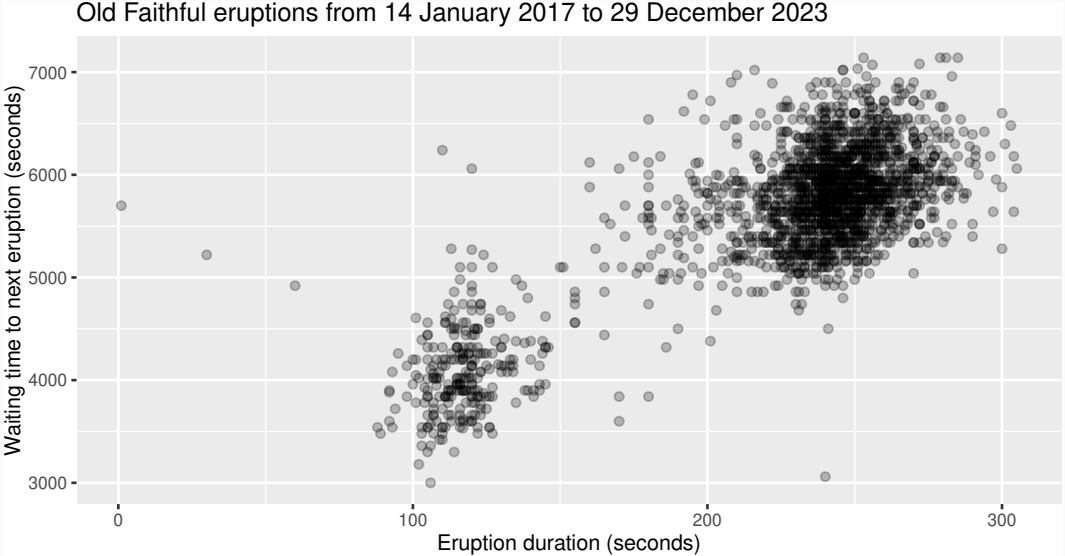


Is this an anomaly?



All points randomly generated from a χ^2_5 distribution.

Are there any anomalies?



Definitions of anomalies

Definition: Anomaly

Given a set of observations $\{y_1, \ldots, y_n\}$ and a generalized probability density f, the **anomaly score** of y_i wrt f is

$$p_i = \mathbb{P}(f(Y) \leq f(y_i))$$

where Y has density f. An observation is an **anomaly** wrt f if $p_i < \alpha$ for some threshold $\alpha > 0$.

Definitions of anomalies

Definition: Anomaly

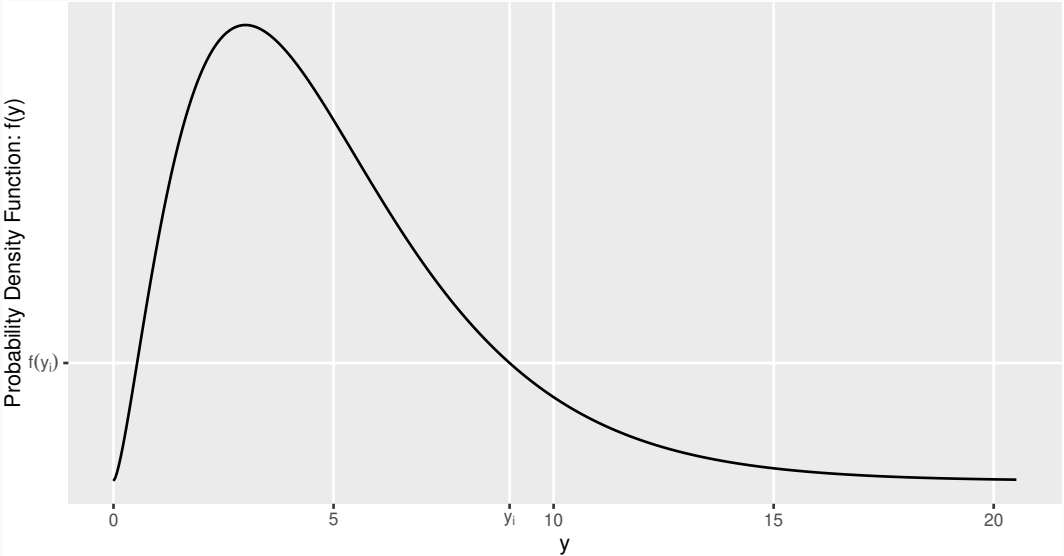
Given a set of observations $\{y_1, \ldots, y_n\}$ and a generalized probability density f, the **anomaly score** of y_i wrt f is

$$p_i = \mathbb{P}(f(Y) \leq f(y_i))$$

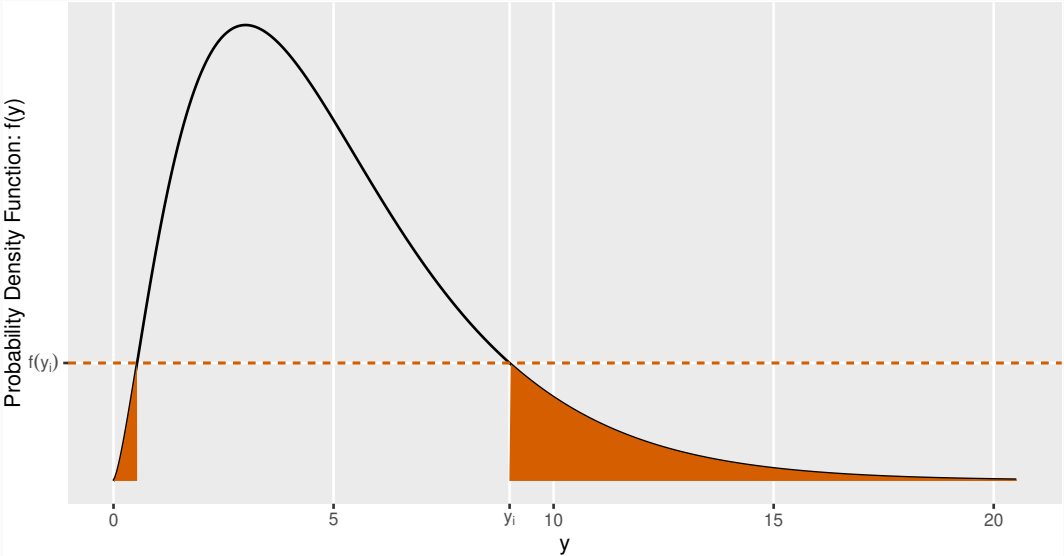
where Y has density f. An observation is an **anomaly** wrt f if $p_i < \alpha$ for some threshold $\alpha > 0$.

- y_i can be a scalar, vector or a more complex object
- f can be a conditional density, and can be known, assumed or estimated

Definitions of anomalies



Definitions of anomalies



Anomaly detection: Normal distribution

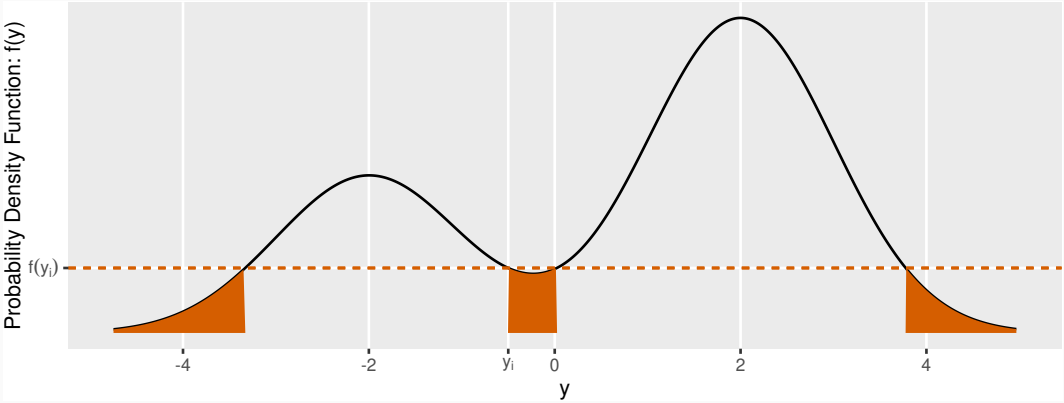
If $f \sim N(\mu, \sigma^2)$, then $p_i = 2\left[1 - \Phi(|y_i - \mu|/\sigma)\right]$

Equivalent to a two-sided p-value from a z-score test.

Anomaly detection: Highest density regions

HDR with probability $1-\alpha$ is $R_\alpha = \{y : f(y) \geq c_\alpha\}$ where c_α is largest constant s.t. $\mathbb{P}(Y \in R_\alpha) \geq 1-\alpha$.

An observation is an anomaly if $y_i \notin R_\alpha$.



Outline

Surprisals

Definition: Surprisal

The **surprisal** of an observation y_i drawn from a probability distribution with density f is defined as

$$s_i = -\log f(y_i)$$

- Better known as “log scores” in statistics.
- “Surprisal” coined by Tribus (1961).
- Expected surprisal = entropy of random variable
- Sum of surprisals = negative log likelihood

Anomaly detection using surprisals

Let $G(s) = \mathbb{P}(S \leq s)$ be the **surprisal distribution** where $S = -\log f(Y)$ and Y has density f.

$$G(s) = \mathbb{P}(-\log f(Y) \leq s) = \mathbb{P}(f(Y) \geq e^{-s})$$

Then $p_i = 1 - G(s_i)$.

Anomaly detection using surprisals

Let $G(s) = \mathbb{P}(S \leq s)$ be the **surprisal distribution** where $S = -\log f(Y)$ and Y has density f.

$$G(s) = \mathbb{P}(-\log f(Y) \leq s) = \mathbb{P}(f(Y) \geq e^{-s})$$

Then $p_i = 1 - G(s_i)$.

- An anomaly is an extreme value of the surprisal distribution.
- It is not necessarily extreme in the sample space of f.

Outline

Fisher-Tippett-Gnedenko theorem

Consider n iid rvs $S_1, \ldots, S_n$ with cdf G and $M_n = \max\{S_1, \ldots, S_n\}$. If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\mathbb{P}\left\{(M_n - b_n)/a_n \leq z\right\} \to H(z) \quad \text{as} \quad n \to \infty,$$

for a non-degenerate cdf H, then

$$H(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

- $\xi < 0$: Weibull distribution (G bounded upper tail)
- $\xi \to 0$: Gumbel distribution (G light-tailed)
- $\xi > 0$: Fréchet distribution (G heavy-tailed)

Pickands-Balkema-De Haan theorem

If G satisfies the FTG theorem, then the upper tail of G can be approximated by the Generalized Pareto Distribution (GPD):

$$\mathbb{P}(S \leq u + s \mid S > u) = 1 - \left(1 + \frac{\xi s}{\sigma_u}\right)^{-1/\xi}$$

for large enough u, where $\sigma_u = \sigma + \xi(u - \mu)$.

Pickands-Balkema-De Haan theorem

If G satisfies the FTG theorem, then the upper tail of G can be approximated by the Generalized Pareto Distribution (GPD):

$$\mathbb{P}(S \leq u + s \mid S > u) = 1 - \left(1 + \frac{\xi s}{\sigma_u}\right)^{-1/\xi}$$

for large enough u, where $\sigma_u = \sigma + \xi(u - \mu)$.

So in practice, we can approximate the upper tail of many distributions by a GPD.

Surprisals and EVT

- Suppose we have n iid observations $Y_1, \ldots, Y_n$ and a density f.
- Let $S_i = -\log f(Y_i)$ be the surprisal of Y_i wrt f
- Then $S_1, \ldots, S_n$ are iid from the surprisal distribution $G(s) = \mathbb{P}(S \leq s)$.
- If G satisfies the FTG theorem, then we can approximate the upper tail of G by a GPD fitted to the top $1 - \beta$ of the surprisal values.
- In practice, we typically use $\beta = 0.9$.

Three-type theorem for surprisals

A1: Sub-Gaussian: $S = -\log f(Y)$ satisfies, for all $\lambda \in \mathbb{R}$, and some $\nu > 0$, $\mathbb{E}\exp\{\lambda(S - \mathbb{E}[S])\} \leq \exp\{\lambda^2\nu^2/2\}$.
A2: Sub-exponential: S is sub-exponential with parameters ν and b, i.e., $\mathbb{E}\exp\{\lambda(S - \mathbb{E}[S])\} \leq \exp\{\lambda^2\nu^2/2\}$ for all $|\lambda| < 1/b$.
A3: Polynomial: $|S|$ has polynomial moments of order $p \geq 1$; i.e., $\mathbb{E}[|S|^p] \leq C^p$ for some $C > 0$ such that $C^p - 1 > 0$.

- A1 satisfied when f has bounded support
- A2 satisfied when $\log f$ unbounded below, and light tails (e.g., Gaussian)
- A3 satisfied when f has heavy tails (e.g., t with df ≥ 3)

Three-type theorem for surprisals

Let $y_1, \ldots, y_n$ be an iid sequence from density f, $s_i = -\log f(y_i)$, $M_n = \max\{s_1, \ldots, s_n\}$, and $S = -\log f(Y)$ where $Y \sim f$.

1. Under A1:

$$\sup_{s:s>0} \left| \mathbb{P}\left\{ \left| M_n - \mathbb{E}[S] \right| \geq \sqrt{2\nu^2 s} + \sqrt{2\nu^2 \log(2n)} \right\} - e^{-s} \right| = o(1).$$

2. Under A2:

$$\sup_{s:s>1/b} \left| \mathbb{P}\left\{ \left| M_n - \mathbb{E}[S] \right| \geq (2b)s + (2b)\log(2n) \right\} - e^{-e^{-s}} \right| = o(1).$$

3. Under A3:

$$\sup_{s:s>C} \left| \mathbb{P}\left\{ \left| M_n - \mathbb{E}[S] \right| \geq (Csn^{1/p}) \right\} - e^{-s^{-p}} \right| = o(1).$$

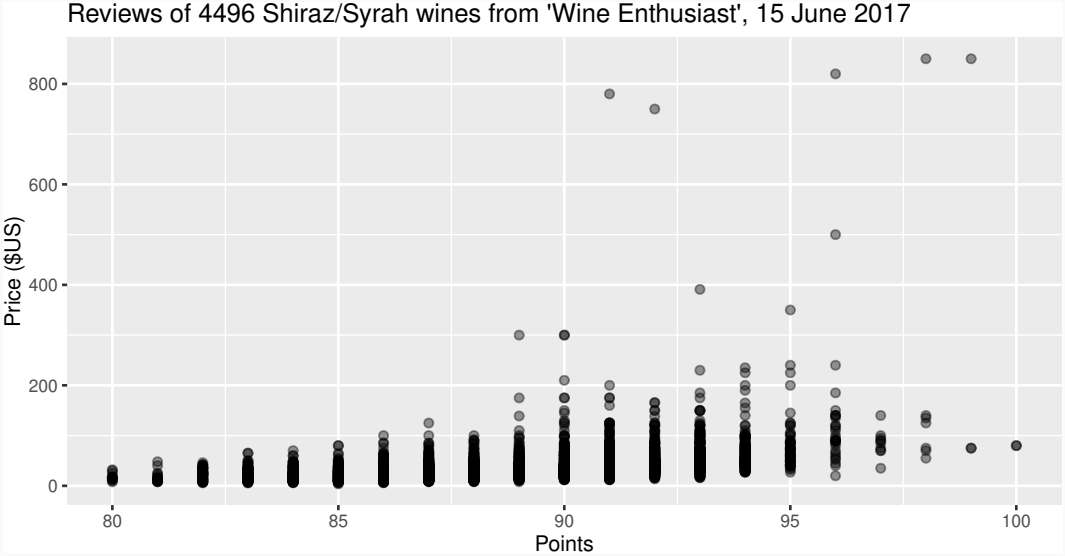
Three-type theorem for surprisals

- If surprisal has Gaussian like tails, then maximum surprisal is a reversed Weibull;
- If surprisal only has an exponential tail, then maximum surprisal is Gumbel;
- If surprisal only has a polynomial moment, then maximum surprisal is Fréchet.

Corollary

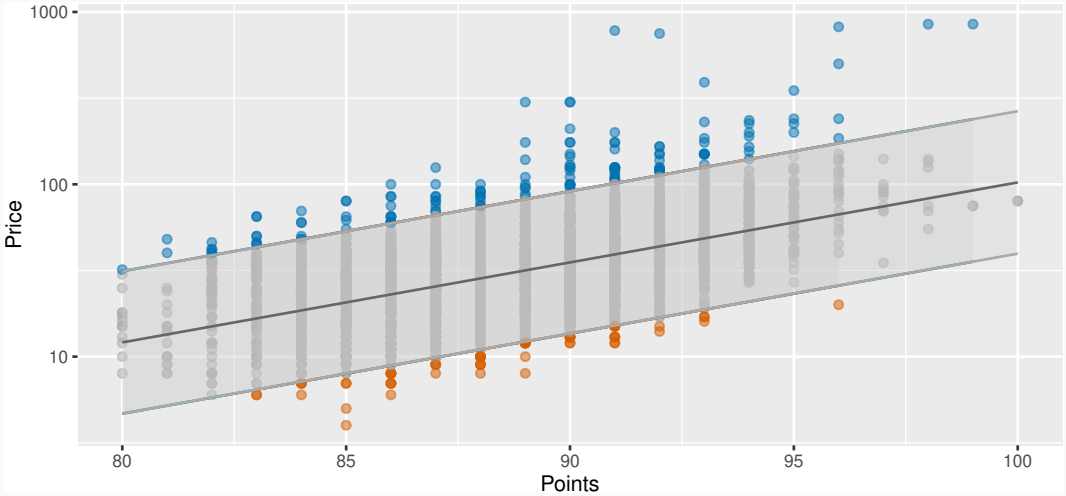
upper tail of the surprisal distribution can be approximated by a GPD, even if the assumed density f is incorrect, provided one of A1–A3 is satisfied.

Application to wine quality and prices



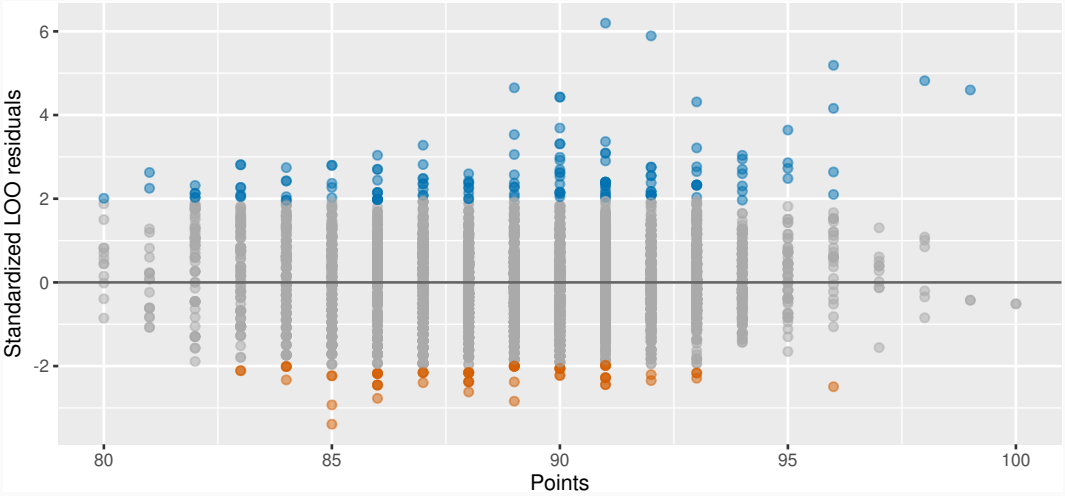
Application to wine quality and prices

Proposed model: $\log \text{Price} \mid \text{Points} \sim N(a + b\text{Points}, \sigma^2)$.



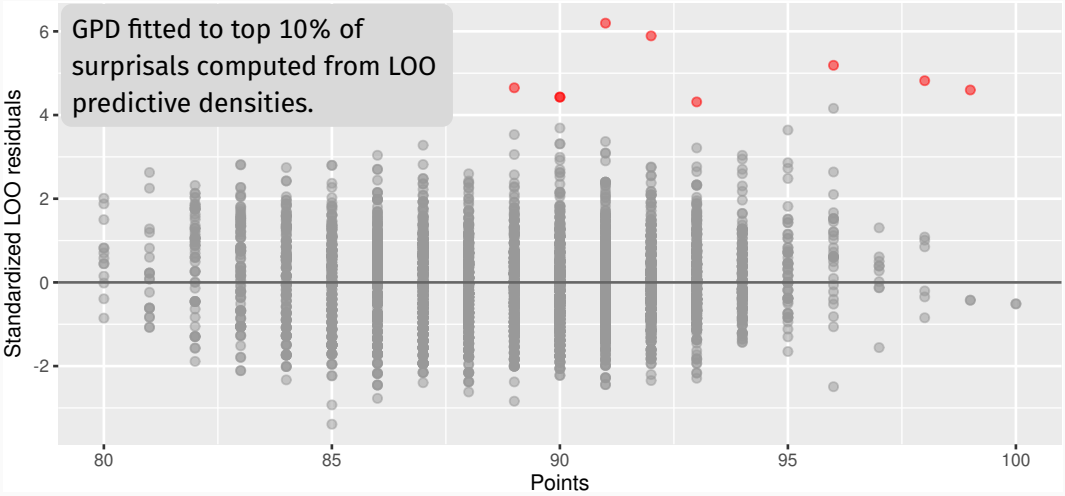
Application to wine quality and prices

Proposed model: $\log \text{Price} \mid \text{Points} \sim N(a + b\text{Points}, \sigma^2)$.

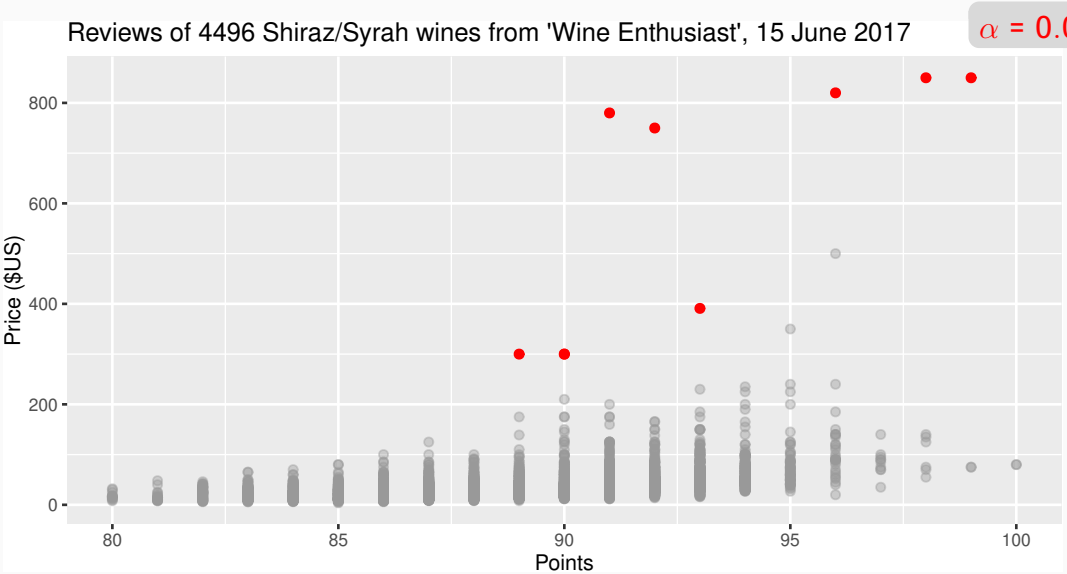


Application to wine quality and prices

Proposed model: $\log \text{Price} \mid \text{Points} \sim N(a + b\text{Points}, \sigma^2)$. $\alpha = 0.001$



Application to wine quality and prices

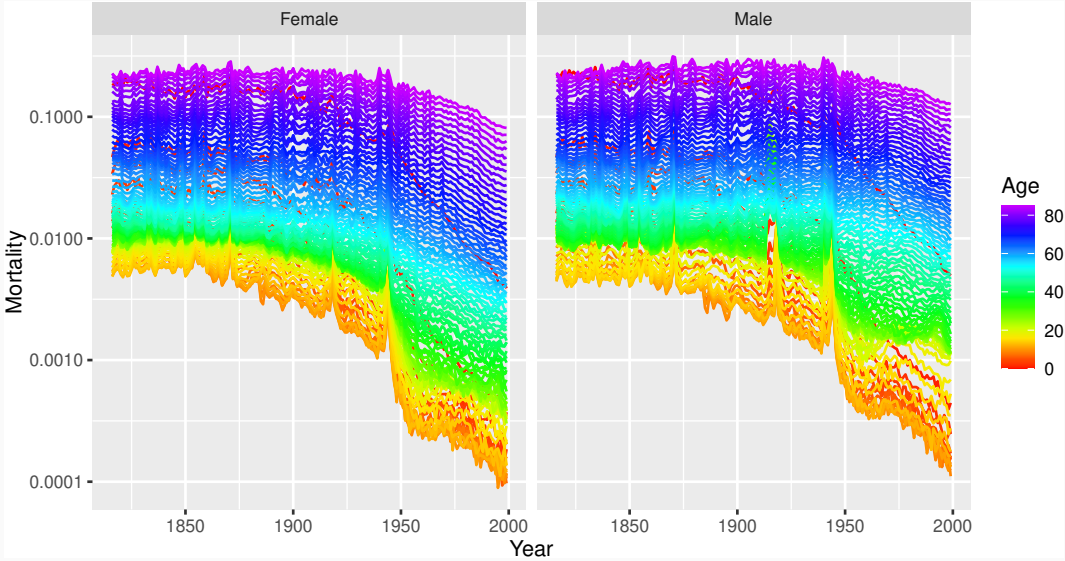


Application to wine quality and prices

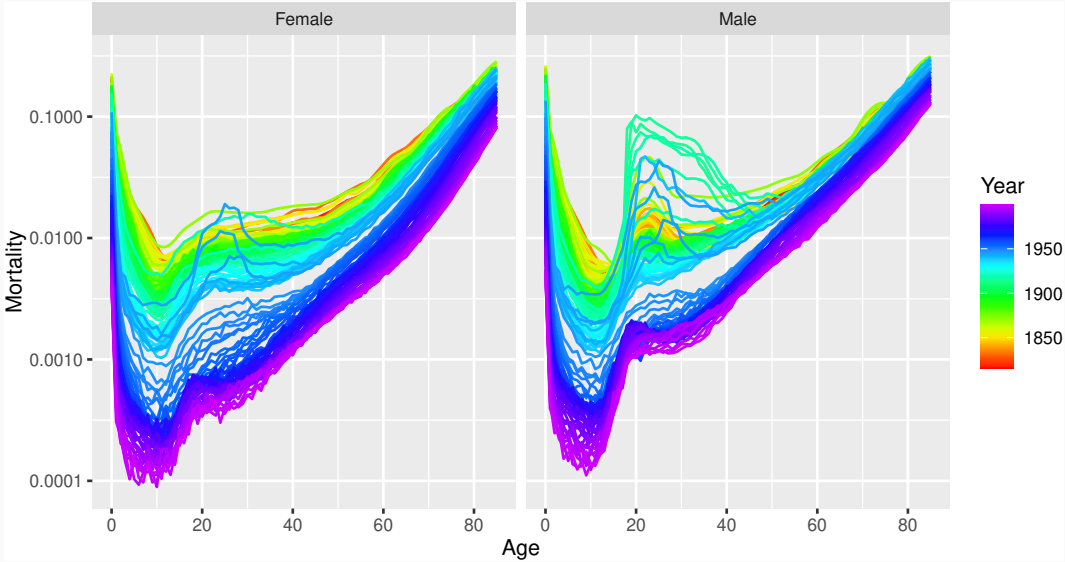
Anomalies detected (α = 0.001):

Area	Winery	Year	Points	Price
South Australia	Henschke	2009	91	780
California	Law	2013	92	750
South Australia	Henschke	2010	96	820
South Australia	Penfolds	2008	98	850
Tuscany	Tua Rita	2011	89	300
South Australia	Penfolds	2010	99	850
Tuscany	Tua Rita	2013	90	300
Tuscany	Tua Rita	2012	90	300
Rhône Valley	Domaine Jean-Michel Gerin	2013	93	391

French mortality



French mortality



Application to French mortality

Model: $\log y_t \sim N(m_t, a_t^2)$ where m_t and a_t are locally and robustly estimated in a window of size $2h + 1$ around time t:

$$\hat{m}_t = \text{median}(\log y_{t-h}, \ldots, \log y_{t+h})$$

$$\hat{a}_t = 1.4826 \times \text{median}(|\log y_{t-h} - \hat{m}_t|, \ldots, |\log y_{t+h} - \hat{m}_t|)$$

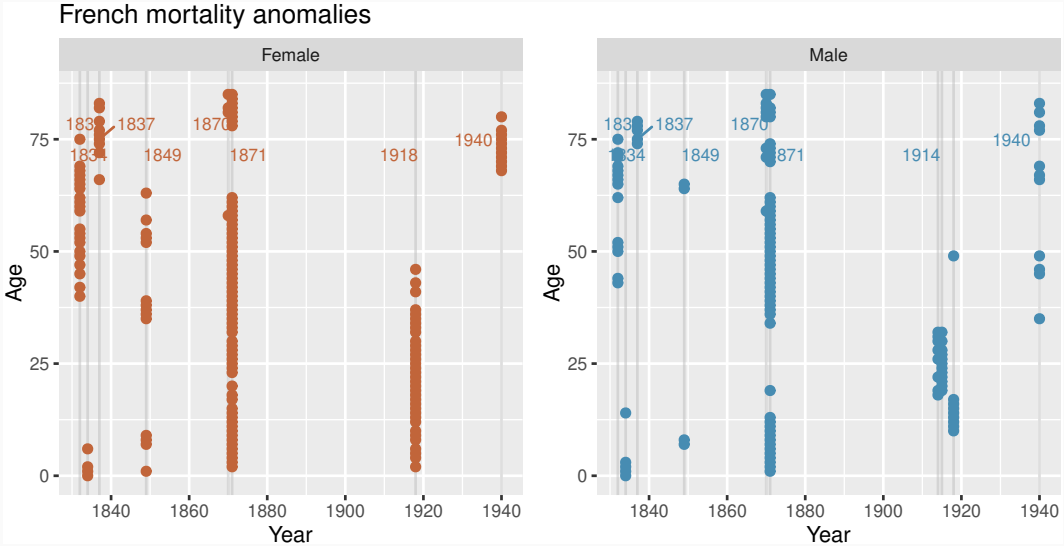
Application to French mortality

Model: $\log y_t \sim N(m_t, a_t^2)$ where m_t and a_t are locally and robustly estimated in a window of size $2h + 1$ around time t:

$$\hat{m}_t = \text{median}(\log y_{t-h}, \ldots, \log y_{t+h})$$
$$\hat{a}_t = 1.4826 \times \text{median}(|\log y_{t-h} - \hat{m}_t|, \ldots, |\log y_{t+h} - \hat{m}_t|)$$

- Male and female data from 1816–1999, over ages 0–85: 31648 observations.
- Compute surprisals under model, and surprisal probabilities under a GPD with $\alpha = 0.01$
- Identify when at least three age groups are anomalous in same year/sex.

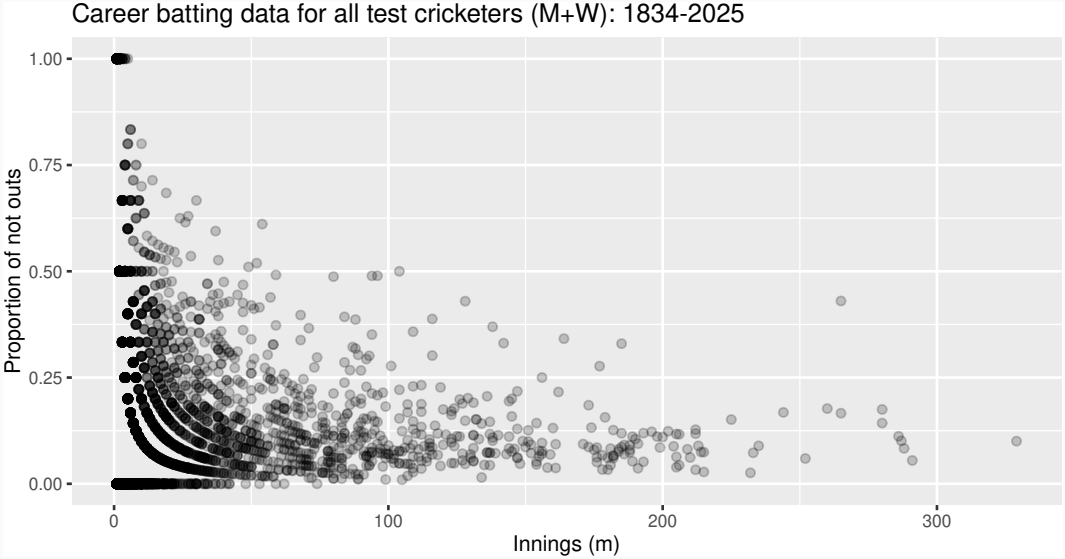
Application to French mortality



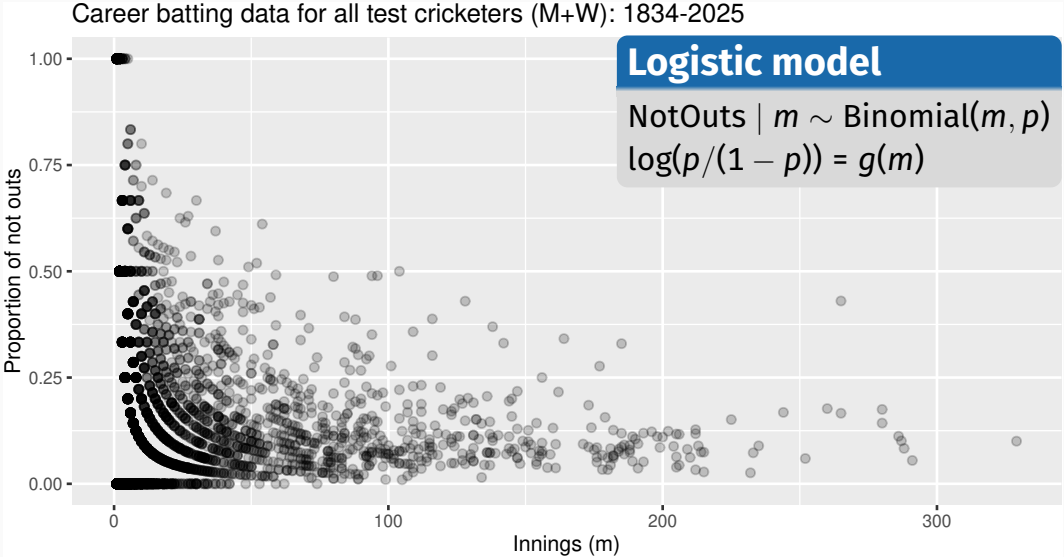
Application to French mortality

- 1832, 1849: Cholera outbreaks
- 1870: Franco–Prussian war
- 1871: Repression of the ‘Commune de Paris’
- 1914–1918: World War I
- 1918: Spanish flu outbreak
- 1940: World War II

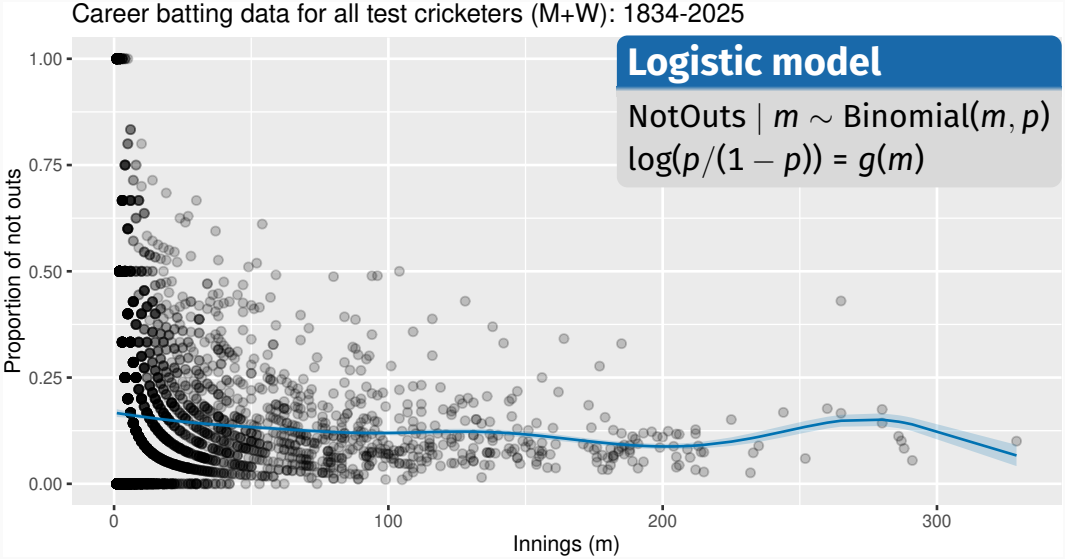
Application to test cricket not-outs



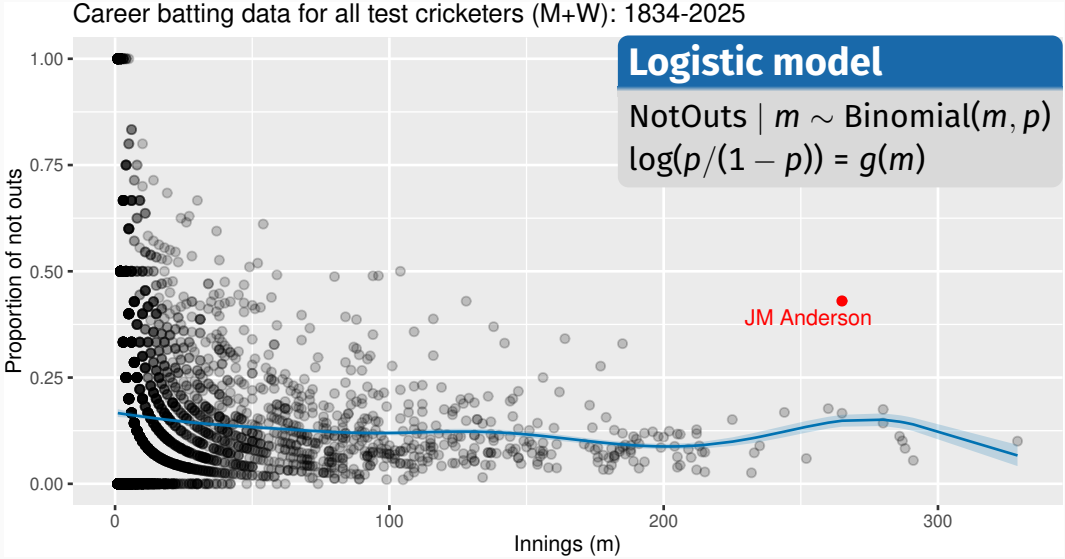
Application to test cricket not-outs



Application to test cricket not-outs



Application to test cricket not-outs



Univariate experiment

What happens when the distribution used to compute surprisals is mis-specified?

Data N(0,1)

- 1000 observations from a N(0,1) distribution
- Surprisals computed using a t(4) distribution.
- Estimate surprisal probabilities using N(0,1), t(4), surprisal ranks, GPD

Univariate experiment

What happens when the distribution used to compute surprisals is mis-specified?

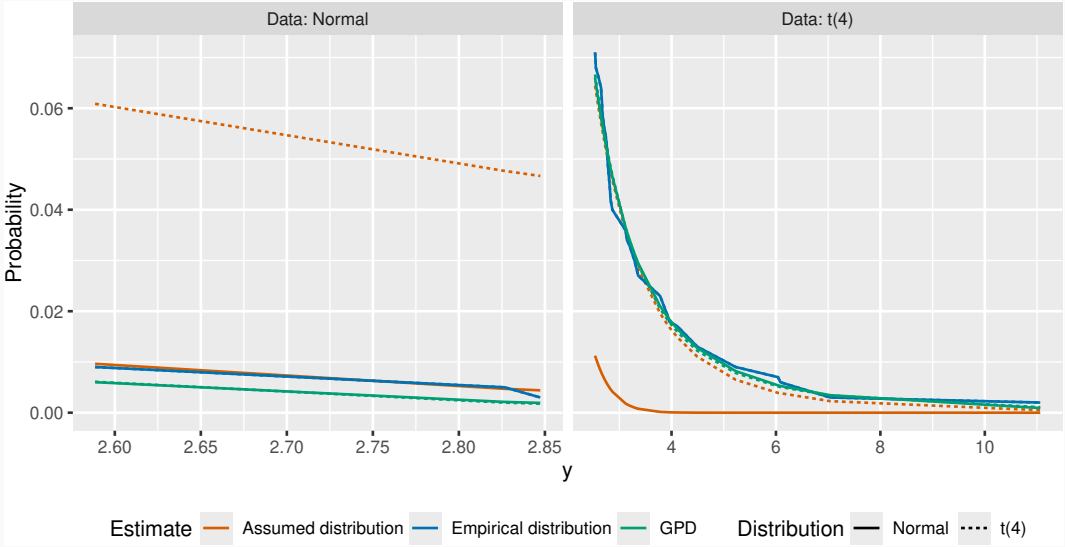
Data N(0,1)

- 1000 observations from a N(0,1) distribution
- Surprisals computed using a t(4) distribution.
- Estimate surprisal probabilities using N(0,1), t(4), surprisal ranks, GPD

Data t(4)

- 1000 observations from a t(4) distribution
- Surprisals computed using a N(0,1) distribution.
- Estimate surprisal probabilities using N(0,1), t(4), surprisal ranks, GPD

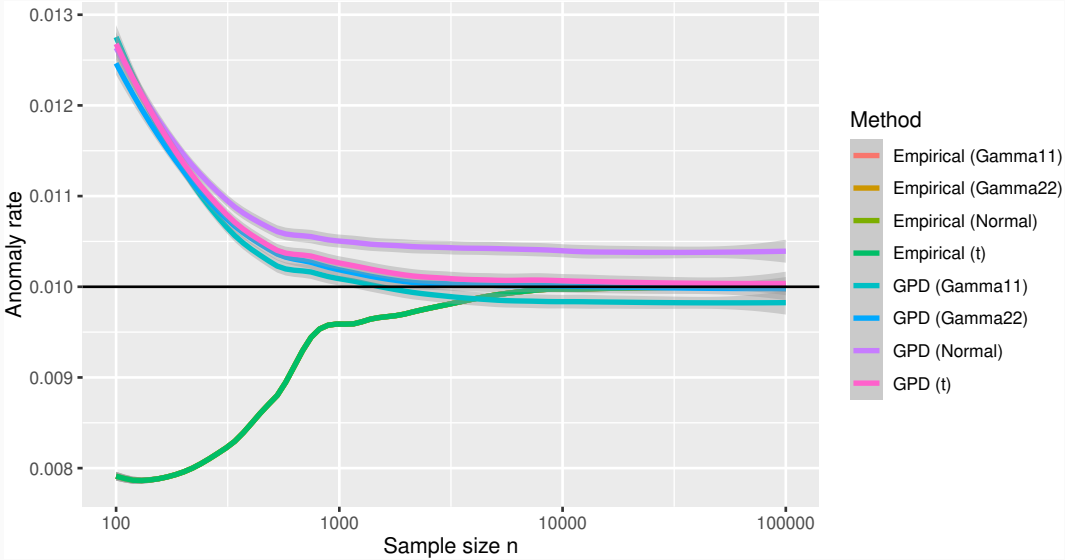
Univariate experiment



Bivariate experiment

- Data: 2 independent Gamma(2,2) variables
- Sample size: $n = 100, \ldots, 100000$
- Surprisals computed using:
  - Gamma(2,2) x 2
  - Gamma(1,1) x 2
  - Bivariate normal with correct mean and variance
  - Bivariate non-central t(4) with correct mean
- Surprisal probabilities computed using:
  - Surprisal ranks
  - GPD fit to top 10% surprisals

Bivariate experiment



Outline

Lookout algorithm

- Compute Kernel Density Estimate (KDE) of the observations, with a bandwidth matrix chosen using persistent homology.
- Compute surprisals from KDE values
- Fit a Generalized Pareto Distribution (GPD) to the largest $1 - \beta$ surprisals
- Compute surprisal probabilities p_i from the fitted GPD.
- An observation is an anomaly if $p_i < \alpha$.

Kernel density estimation

Observations: $\boldsymbol{y}_i \in \mathbb{R}^m$ for $i \in \{1, \ldots, n\}$.

KDE

$$\hat{f}(\boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} |\boldsymbol{H}|^{-1/2} K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{y}_i)\right),$$

- K is a square-integrable spherically-symmetric function, bounded below by 0, with a finite second-order moment and unit integral.
- $\boldsymbol{H}$ is a symmetric $m \times m$ positive-definite matrix.

Kernel density estimation

Observations: $\boldsymbol{y}_i \in \mathbb{R}^m$ for $i \in \{1, \ldots, n\}$.

KDE

$$\hat{f}(\boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} |\boldsymbol{H}|^{-1/2} K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{y}_i)\right),$$

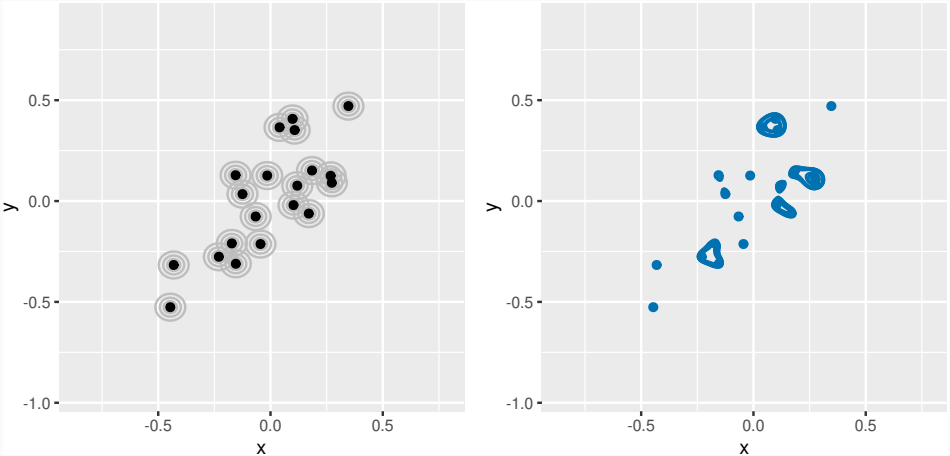
- K is a square-integrable spherically-symmetric function, bounded below by 0, with a finite second-order moment and unit integral.
- $\boldsymbol{H}$ is a symmetric $m \times m$ positive-definite matrix.

LOO KDE values

$$f_{-i} = \frac{1}{n-1}\left(n\hat{f}(\boldsymbol{y}_i) - \boldsymbol{H}^{-1/2} K(\boldsymbol{0})\right)$$

Kernel density estimation

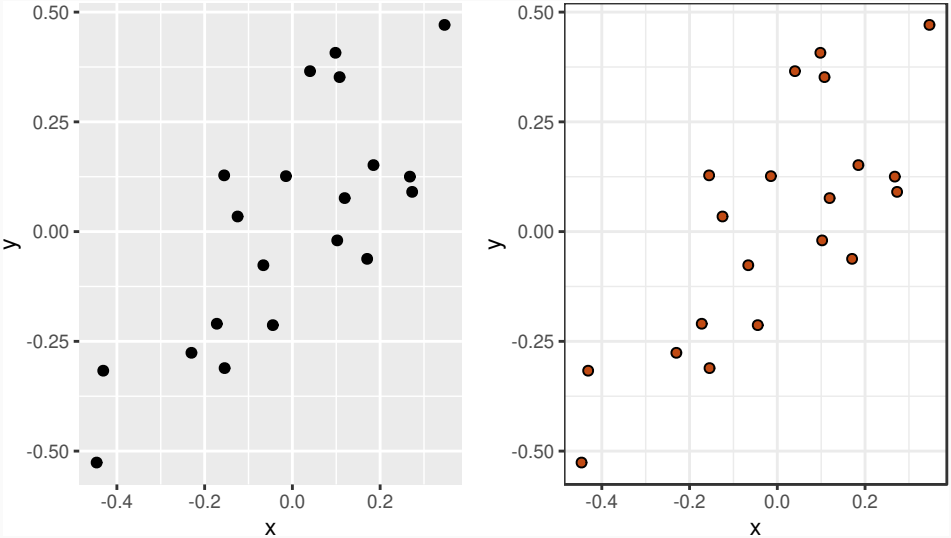
$$\hat{f}(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{H}|^{-1/2} K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{y}_i)\right)$$



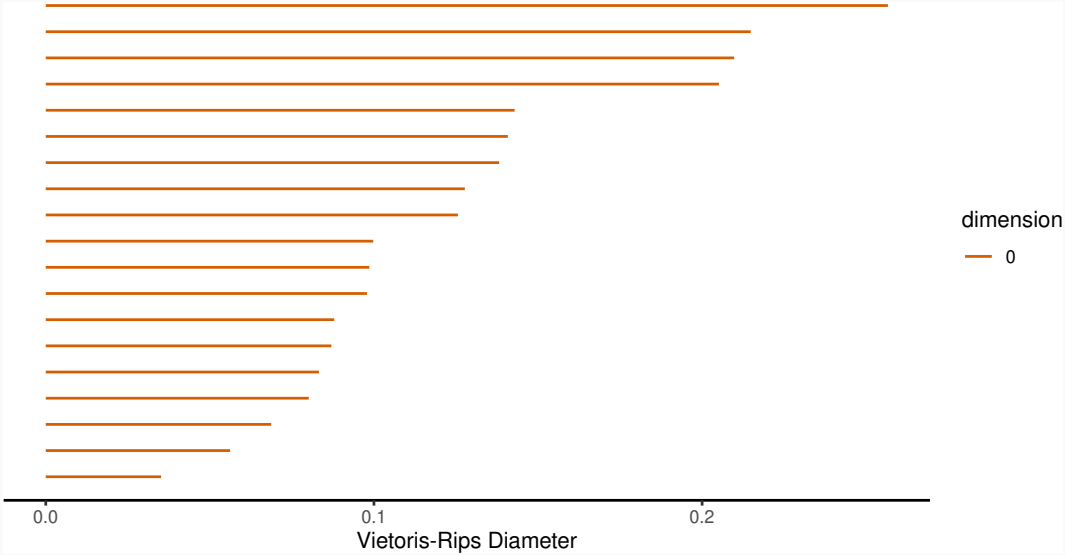
GPD fitted to KDE surprisals

- Compute surprisals from KDE values: $s_i = -\log f_i$, $i = 1, \ldots, n$.
- f_i are bounded above and below, so the surprisals are sub-Gaussian.
- Fit GPD to largest $1 - \beta$ of $\{s_i\}_{i=1}^n$, constraining shape parameter to be non-positive.

Persistent homology



Persistent homology

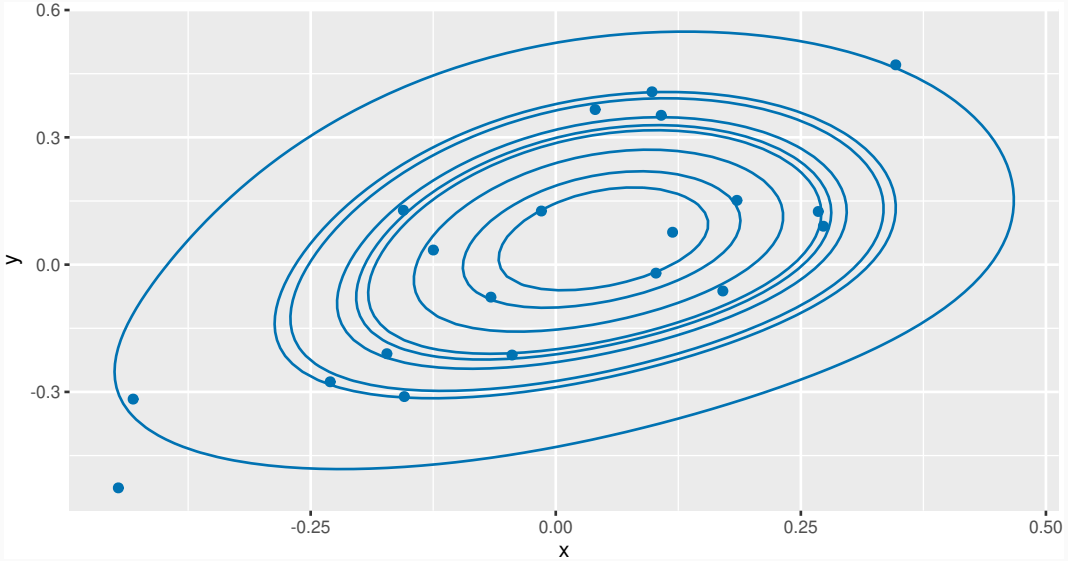


KDE bandwidth selection via VR diameters

1. Compute persistence homology barcode of data for dimension 0 using Vietoris-Rips diameter.
2. Obtain ordered death diameters $\{d_i\}_{i=1}^n$.
3. Let d_γ^* = γ sample quantile computed from $\{d_i\}_{i=1}^n$, for some large $\gamma < 1$
4. Set bandwidth matrix $\boldsymbol{H} = (d_\gamma^*)^{2/m}\boldsymbol{I}_m$.

- Assumes each dimension has the same scale.
- Ensures that persistently connected components have large weights in the KDE at each observed point.
- In practice, we use $\gamma = 0.97$.

KDE bandwidth selection via VR diameters



Kernel density estimation (consistency)

Let $Y_1, \ldots, Y_n$ be iid from square integrable density function f. Then KDE is a *consistent* estimator of f if

$$\lim_{n \to 0} \mathrm{E}\left[\left(\hat{f}(\boldsymbol{u}) - f(\boldsymbol{u})\right)^2 du\right] = 0$$

For $\hat{f}$ to be a consistent estimator, $\boldsymbol{H}$ must be a symmetric positive-definite matrix that satisfies the following conditions:

- All elements of $\boldsymbol{H}$ approach zero as $n \to \infty$
- $n^{-1}|\boldsymbol{H}|^{-1/2} \to 0$ as $n \to \infty$

KDE bandwidth selection via VR diameters

Theorem

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a Lipschitz density function. For any $\gamma \in (0, 1)$, $d^*_\gamma \to 0$ a.s., where $d^*_\gamma = \gamma$ quantile of finite death times.

KDE bandwidth selection via VR diameters

Theorem

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a Lipschitz density function. For any $\gamma \in (0, 1)$, $d^*_\gamma \to 0$ a.s., where $d^*_\gamma = \gamma$ quantile of finite death times.

Theorem

Let $d_k = k$th smallest death time of 0-homology for Rips filtrations of n points from distribution over $\mathbb{R}^2$ s.t. density bounded above. Choose any sequence of integers $\{\omega(n)\}$ s.t. $0 < \omega(n) < n \; \forall n$ and $\lim_{n\to\infty} \omega_n = \infty$. Then $nd_{\omega(n)} \to \infty$ as $n \to \infty$.

KDE bandwidth selection via VR diameters

Theorem

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a Lipschitz density function. For any $\gamma \in (0, 1)$, $d^*_\gamma \to 0$ a.s., where $d^*_\gamma = \gamma$ quantile of finite death times.

Theorem

Let $d_k = k$th smallest death time of 0-homology for Rips filtrations of n points from distribution over $\mathbb{R}^2$ s.t. density bounded above. Choose any sequence of integers $\{\omega(n)\}$ s.t. $0 < \omega(n) < n \ \forall n$ and $\lim_{n\to\infty} \omega_n = \infty$. Then $nd_{\omega(n)} \to \infty$ as $n \to \infty$.

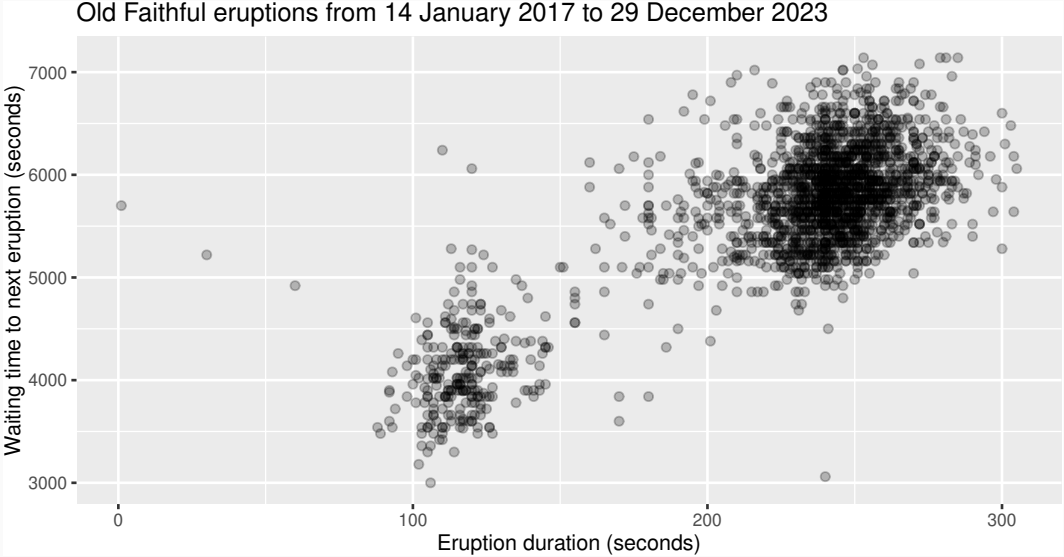
Theorem

Let $d_1 =$ smallest death time of 0-homology for Rips filtrations of n points from distribution over $\mathbb{R}^m$, $m \geq 3$, s.t. density bounded above. Then $nd_1 \to \infty$ as $n \to \infty$.

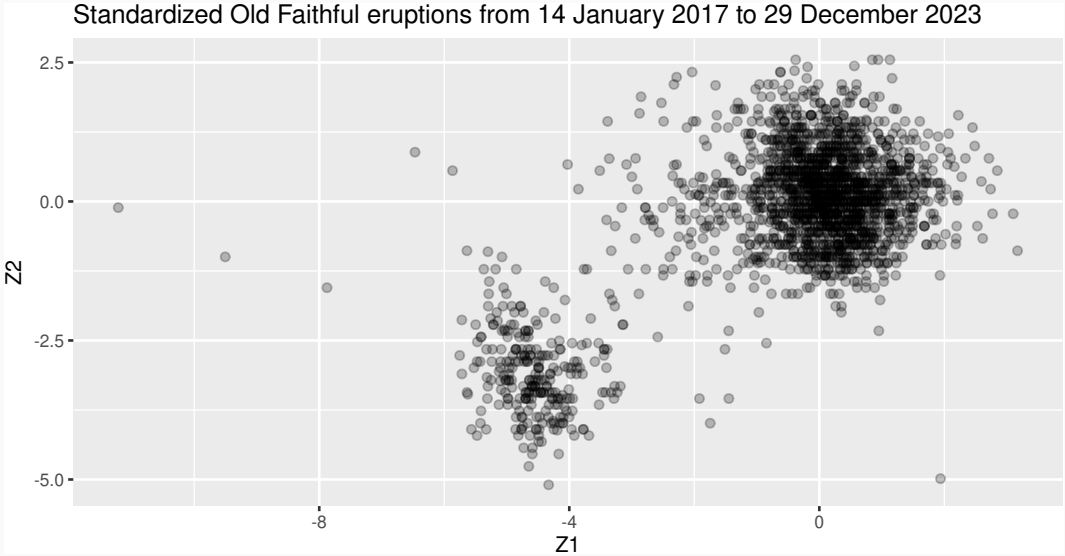
Lookout algorithm

1. $\boldsymbol{Y}$ = data matrix with rows $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$.
2. $\hat{\Sigma}$ = orthogonalized Gnanadesikan-Kettenring estimate of $\mathrm{Cov}(\boldsymbol{Y})$, with eigendecomposition $\hat{\Sigma} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$.
3. Rotate and scale the data: $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{Y}$.
4. Compute persistence homology barcode of $\boldsymbol{Z}$ for dim zero using Vietoris-Rips diameter; obtain ordered death diameters $\{d_i\}_{i=1}^n$.
5. d_γ^* = γ sample quantile computed from $\{d_i\}_{i=1}^n$.
6. Compute kde: $f_i = \hat{f}(\boldsymbol{z}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{H} = (d_\gamma^*)^{2/m}\boldsymbol{I}_m$.
7. Compute LOO kde values $f_{-i} = \frac{1}{n-1}\left(nf_i - \boldsymbol{H}^{-1/2}K(\boldsymbol{0})\right)$, $i = 1, \ldots, n$.
8. Fit GPD to largest $1-\beta$ of surprisals $\{-\log f_i\}_{i=1}^n$, constraining shape parameter to be non-positive.
9. $p_i = (1-\beta)P(-\log f_{-i} \mid \hat{\mu}, \hat{\sigma}, \hat{\xi})$, P = GPD cdf.

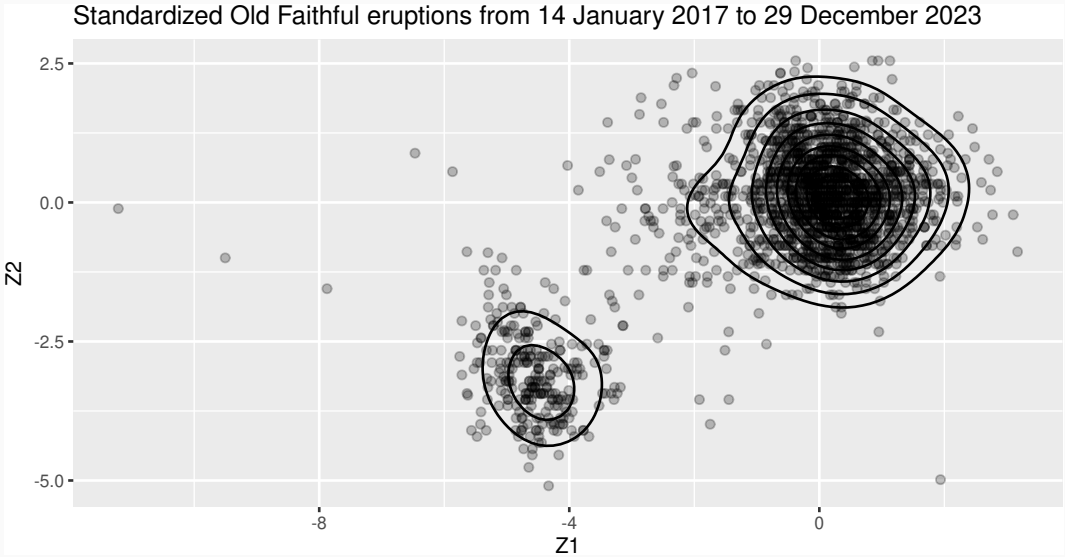
Old Faithful eruptions



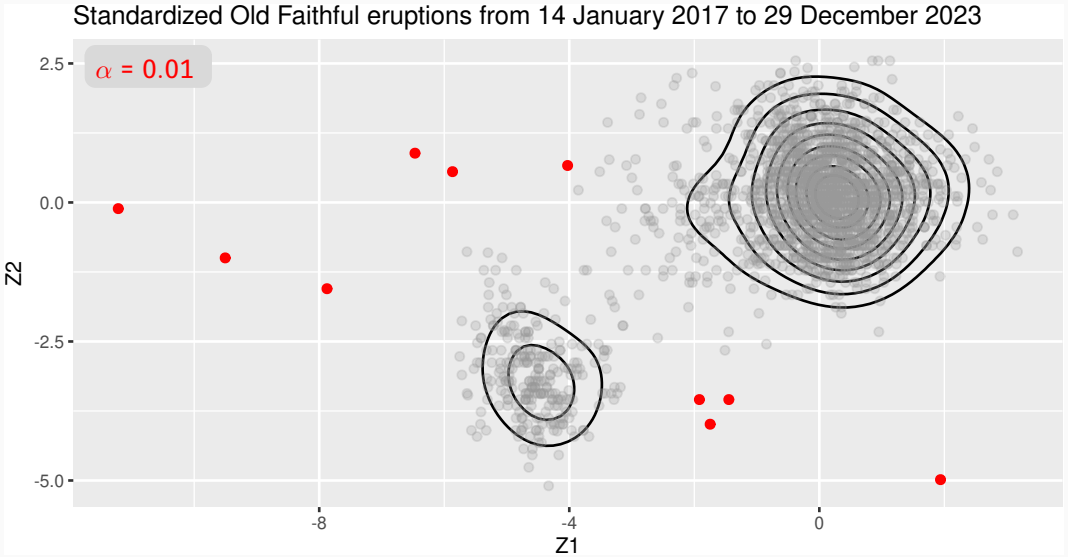
Old Faithful eruptions



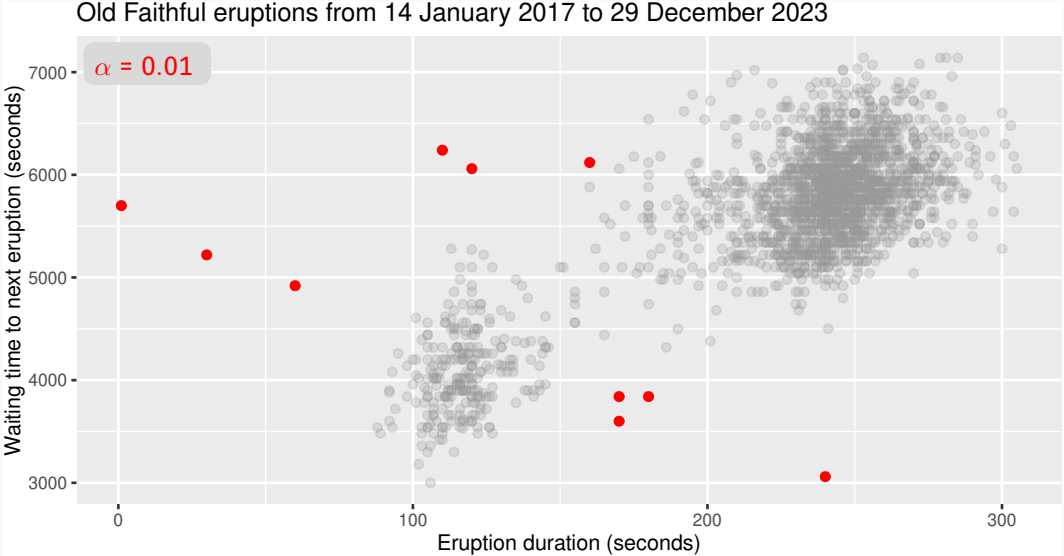
Old Faithful eruptions



Old Faithful eruptions



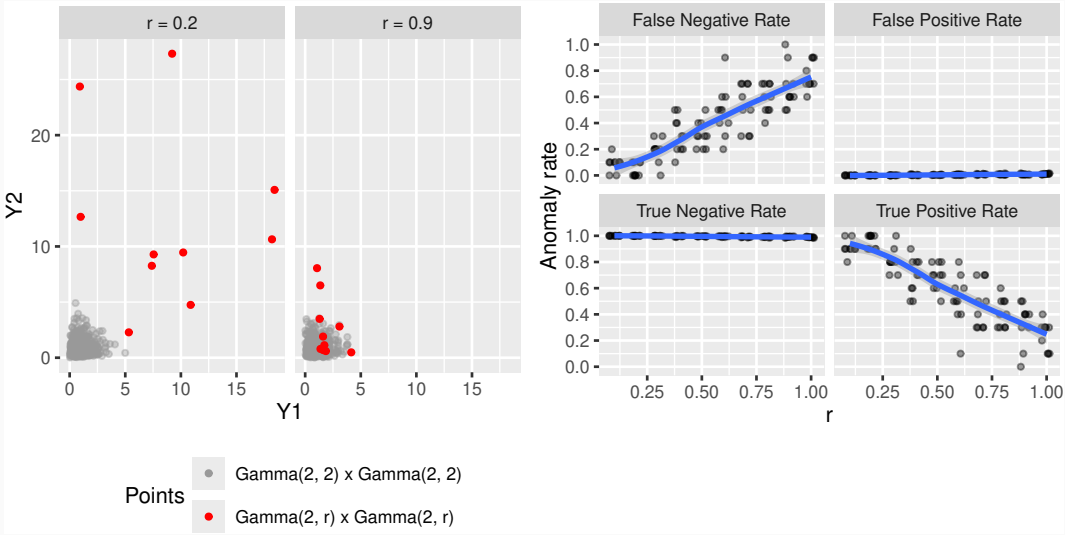
Old Faithful eruptions



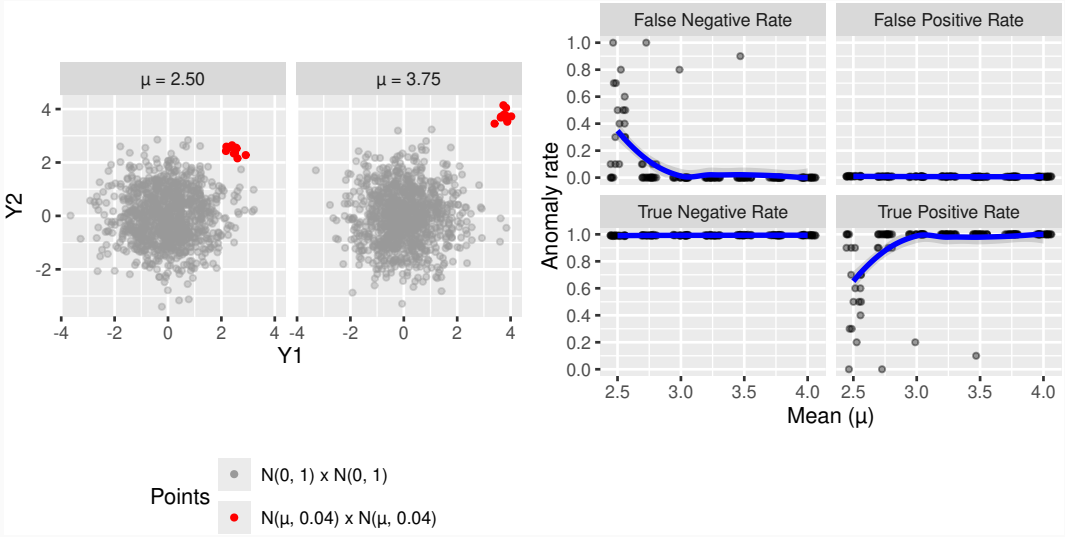
Old Faithful eruptions

time	recorded_duration	duration	waiting	prob
2018-04-25 19:08:00	1s	1	5700	0.0000000
2022-12-07 17:19:00	~4 30s	30	5220	0.0000000
2023-07-04 12:03:00	~1 minute 55ish seconds	60	4920	0.0000612
2020-09-04 01:38:00	>1m 50s	110	6240	0.0000012
2020-06-01 21:04:00	2 minutes	120	6060	0.0001390
2020-09-16 14:44:00	>2m40s	160	6120	0.0078057
2020-08-31 09:56:00	~2m50s	170	3840	0.0076247
2021-01-22 18:35:00	2m50s	170	3600	0.0029265
2022-11-29 14:51:00	~3m	180	3840	0.0033112
2022-12-03 16:20:00	~4m	240	3060	0.0000000

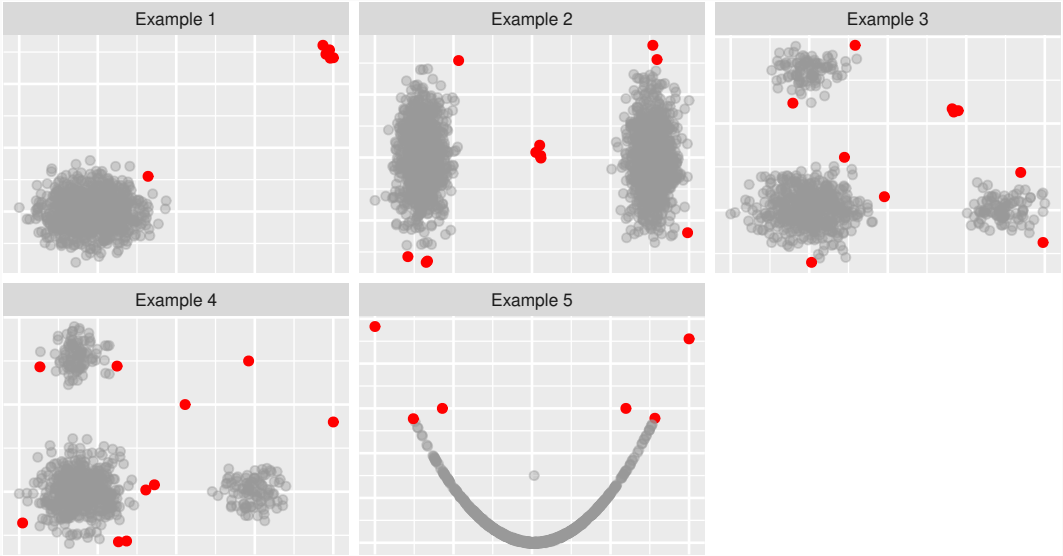
Experiment 1



Experiment 2



Experiment 3



Outline

Conclusions

- Surprisal-based anomaly detection is a flexible, probabilistic approach that can be used in any context where a probability distribution can be defined on the space of observations
- EVT theory on surprisals due to Hyndman & Frazier (in preparation)
- Original lookout algorithm due to Kandanaarachichi & Hyndman (*JCGS*, 2022). `https://robjhyndman.com/publications/lookout`
- Modified lookout algorithm due to Hyndman, Kandanaarachichi & Turner (in preparation)
- R packages `lookout` and `weird` available on CRAN
- Book in preparation at `https://OTexts.com/weird`
- Slides and links: `https://robjhyndman.com/toptime2025`