Machine Learning for Time Series
Computing Valid p-value for Optimal Changepoint by Selective Inference using Dynamic Programming

Toby Johnstone Waïss Azizian

March 25, 2021

1 Overview of our work

Our project was based on the work of [1]. We reimplemented the method described from scratch in Python (for a fixed number of changepoints K) and checked it was working by reproducing some simple performance tests from the article. We then went on to apply the method to some novel real world data: time series tracking bike usage in Paris. Since we found the reasoning behind the choice of conditioning events for the selective p-value to be rather opaque, we took the time to explain how the reasoning from [2] (which was only alluded to in [1]) can be applied in the setting of the project.

2 Description of the method

The method we implemented is described in [1]. The objective is to detect changepoints in the mean of a time series. This can be formalised in the following way. We consider a random sequence $X = (X_0, \dots, X_{N-1})^T \sim \mathcal{N}(\mu, \Sigma)$ where N is the length of the time-series, $\mu \in \mathbb{R}^N$ is the unknown mean vector, and $\Sigma \in \mathbb{R}^{N\times N}$ is the covariance matrix which is considered to be known (or at least already estimated). We are given a single observed sequence $x^{obs} = (x_0^{obs}, \dots, x_{N-1}^{obs})^T \in \mathbb{R}^N$.

2.1 Changepoint vectors

The vector of detected CP locations is denoted as $\tau_1^{det}, \dots, \tau_K^{det}$ where K is the number of CPs, and $\tau_1^{det} < \cdots < \tau_K^{det}$. We set $\tau_0^{det} = -1$ and $\tau_{K+1}^{det} = N - 1$. Note that our indexing convention is not that which is used by [1] but rather one that aligns more readily with indexing in Python. We denote $x_{s:e}$ the subsequence of x starting at s and ending at e. The mean of a given subsequence is written $\bar{x}_{s:e} = \frac{1}{e-s+1}\sum_i x_i$, and the cost function measuring its "homogeneity" is $C(x_{s:e}) = \sum_{i=s}^{e}(x_i - \bar{x}_{s:e})^2$.

For K changepoints is known, the CP detection problem is formulated as the following optimisation problem:

$$\tau^{det} = \mathcal{A}(x^{obs}) := \arg\min_{\tau} \sum_{k=1}^{K+1} C(x^{obs}_{\tau_{k-1}+1:\tau_k}) \tag{1}$$

Having solved this problem, it is desirable to be able to determine the statistical significance of a given detected change point. For the k^{th} changepoint, this can be formulated as testing the null hypothesis $H_{0,k}$ vs the alternative hypothesis $H_{1,k}$ where

$$H_{0,k} : \mu_{\tau_{k-1}^{det}+1} = \cdots = \mu_{\tau_k^{det}} = \mu_{\tau_k^{det}+1} = \cdots = \mu_{\tau_{k+1}^{det}}\,, \quad H_{1,k} : \mu_{\tau_{k-1}^{det}+1} = \cdots = \mu_{\tau_k^{det}} \neq \mu_{\tau_k^{det}+1} = \cdots = \mu_{\tau_{k+1}^{det}}$$

A natural choice of statistic is that defined by

$$\eta_k^T X = \bar{X}_{\tau_{k-1}^{det}+1:\tau_k^{det}} - \bar{X}_{\tau_k^{det}+1:\tau_{k+1}^{det}} \qquad \text{where } \eta_k = \frac{1}{\tau_{k-1}^{det} - \tau_k^{det}}\mathbf{1}^N_{\tau_{k-1}^{det}+1:\tau_k^{det}} - \frac{1}{\tau_k^{det} - \tau_{k+1}^{det}}\mathbf{1}^N_{\tau_k^{det}+1:\tau_{k+1}^{det}}$$

The p-value associated to this statistic is $p_k^{naive} = \mathbb{P}_{H_{0,k}}(|\eta_k^T X| \geq |\eta_k^T x^{obs}|)$. However, it is well known that this p-value is not valid in this situation because it suffers from *selection bias*. To counter this, the authors resort to the selective inference framework from [2] and *condition* the p-value on the selection event.

2.2 Selective inference and selective p-value

To explain the expression of the conditional p-value introduced by the authors [1], let us take a moment to demonstrate how the situation at hand fits into the framework of [2].

First, as in [2], let us build our *question* space. The family of models for the data under consideration are the,

$$M^\tau = \{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^N \text{ s.t. } \forall k = 0, \dots, K,\ \mu_{\tau_k+1} = \cdots = \mu_{\tau_{k+1}}\}\,,$$

for τ CP vector. Given a model M^τ, the null hypothesis for the k^{th} changepoint is the subset of models of M^τ defined by,

$$H^\tau_{0,k} = \{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^N \text{ s.t. } \mathcal{N}(\mu, \Sigma) \in M^\tau \text{ and } \mu_{\tau_k} = \mu_{\tau_k+1}\}\,.$$

The question space is now, $\mathcal{Q} = \{(M^\tau, H^\tau_{0,k}) : \tau \text{ CP vector}, k = 0, \dots, K\}$.

Selective inference now consists in two stages:

Selection: Based on the data x^{obs}, we select a set of question $\hat{\mathcal{Q}}(x^{obs}) \subset \mathcal{Q}$. In our case, this is,

$$\hat{\mathcal{Q}}(x^{obs}) = \{(M^{\tau^{det}}, H^{\tau^{det}}_{0,k}) : k = 0, \dots, K\}\,.$$

Inference: We try to answer a selected question $Q = (M^{\tau^{det}}, H^{\tau^{det}}_{0,k}) \in \hat{\mathcal{Q}}(x^{obs})$. In other words, we perform a statistical test $H^{\tau^{det}}_{0,k}$ against $H^{\tau^{det}}_{1,k} \coloneqq M^{\tau^{det}} \setminus H^{\tau^{det}}_{0,k}$ as described above.

To remove the selection bias, one must condition on the selection event, i.e. the event of asking a question Q when we look at the data,

$$E_Q \coloneqq \{Q \in \hat{\mathcal{Q}}(X)\}\,.$$

In our case, for $Q = (M^{\tau^{det}}, H^{\tau^{det}}_{0,k}) \in \hat{\mathcal{Q}}(x^{obs})$, the selection event becomes $E_Q = \{\mathcal{A}(X) = \tau^{det}\}$.

One could then want to consider the selective p-value $\mathbb{P}_{H^{\tau^{det}}_{0,k}}\left(|\eta_k^T X| \geq |\eta_k^T x^{obs}| \,\middle|\, \mathcal{A}(X) = \tau^{det}\right)$. However, another conditioning step is necessary to get rid of nuisance statistics, as explained in [2, §3.1].

For this, define q a linear mapping of $\mathbb{R}^N$ by,

$$q(x) = x - \frac{\eta_k^T x}{\eta_k^T \Sigma \eta_k} \Sigma \eta_k$$

Actually, $x \mapsto \frac{\eta_k^T x}{\eta_k^T \sigma \eta_k} \Sigma \eta_k$ is the orthogonal projection on the line $\operatorname{Span} \Sigma\eta_k$ w.r.t to the scalar product induced by Σ, $\langle x, y\rangle_\Sigma \coloneqq x^T \Sigma^{-1} y$. Therefore, q is the orthogonal projection on the orthogonal of this line.

Writing the density $p(x)$ of $X \sim \mathcal{N}(\mu, \Sigma)$ with $\langle \cdot, \cdot \rangle_\Sigma$ and decomposing them using the orthogonal projection q gives,

$$\begin{aligned}
p(x) &\propto \exp\left(-\frac{1}{2}\langle x, x\rangle_\Sigma + \langle x, \mu\rangle_\Sigma\right) \\
&\propto \exp\left(-\frac{1}{2}\langle x - q(x), x - q(x)\rangle_\Sigma - \frac{1}{2}\langle q(x), q(x)\rangle_\Sigma + \langle x - q(x), x - q(\mu)\rangle_\Sigma + \langle q(x), q(\mu)\rangle_\Sigma\right) \\
&\propto \exp\left(-\frac{1}{2}\frac{(\eta_k^T x)^2}{\eta_k^T \Sigma \eta_k} + \frac{(\eta_k^T x)(\eta_k^T \mu)}{\eta_k^T \Sigma \eta_k}\right) \exp\left(-\frac{1}{2}\langle q(x), q(x)\rangle_\Sigma + \langle q(x), q(\mu)\rangle_\Sigma\right)\,.
\end{aligned}$$

In particular, this shows that, when we do not condition on the selection event, $\eta_k^T X$ and $q(X)$ are independent. However, conditionally on the selection event $E_Q = \{\mathcal{A}(X) = \tau^{det}\}$, the density of the $Z = \eta_k^T X$ would be,

$$p(z|\mathcal{A}(X) = \tau^{det}) \propto \exp\left(-\frac{1}{2}\frac{z^2}{\eta_k^T \Sigma \eta_k} + \frac{z(\eta_k^T \mu)}{\eta_k^T \Sigma \eta_k}\right) \int_{\mathbb{R}^N} \exp\left(-\frac{1}{2}\langle q(x), q(x)\rangle_\Sigma + \langle q(x), q(\mu)\rangle_\Sigma\right) \mathbb{1}\left\{z = \eta_k^T x,\ \mathcal{A}(x) = \tau^{det}\right\} dx\,,$$

which is first, not tractable, and second, does depend on the distribution of $q(X)$, which we are not interested in for our test. In the literature, $q(X)$ is sometimes called a *nuisance statistic*.

Therefore, to get rid of it, we follow [2, §3.1], which recommends conditioning on the value of this nuisance statistics.

The final selective p-value is thus,

$$p_k^{selective} = \mathbb{P}_{H_{0,k}}\left(|\eta_k^T X| \geq |\eta_k^T x^{obs}| \;\middle|\; \mathcal{A}(X) = \mathcal{A}(x^{obs}), q(X) = q(x^{obs})\right)\,. \tag{2}$$

Now, the density of $Z = \eta_k^T X$ conditionally on these events is,

$$p(z) \propto \exp\left(-\frac{1}{2}\frac{z^2}{\eta_k^T \Sigma \eta_k} + \frac{z(\eta_k^T \mu)}{\eta_k^T \Sigma \eta_k}\right) \mathbb{1}\left\{\mathcal{A}\left(q(x^{obs}) + \frac{z}{\eta_k^T \Sigma \eta_k} \Sigma \eta_k\right) = \tau^{det}\right\}\,,$$

where we used that $x = q(x) + \frac{\eta_k^T x}{\eta_k^T \Sigma \eta_k} \Sigma \eta_k = q(x) + \frac{z}{\eta_k^T \Sigma \eta_k} \Sigma \eta_k$ with $z = \eta_k^T x$.

Hence, $\eta_k^T X$, conditionally on $\{\mathcal{A}(X) = \tau^{det}, q(X) = q(x^{obs})\}$, is a real Gaussian variable $Z \sim \mathcal{N}(0, \eta_k^T \Sigma \eta_k)$ conditioned on

$$Z \in \mathcal{Z} := \left\{ z \in \mathbb{R} : \mathcal{A}\left(q(x^{obs}) + \frac{z}{\eta_k^T \Sigma \eta_k} \Sigma \eta_k \right) = \tau^{det} \right\} .$$

Therefore, to compute the selective p-value, one must find an efficient way of computing the set $\mathcal{Z}$ above.

2.3 Computation of the truncation region

Let us introduce more convenient notations. Since, the observed signal x^{obs} and the index of the changepoint under consideration k are fixed, we write, for any $z \in \mathbb{R}$,

$$x(z) = q(x) + \frac{z}{\eta_k^T \Sigma \eta_k} \Sigma \eta_k .$$

Henceforth, in order to be consistent with [1], k will denote a number of changepoints. Denote, for $n \leq N$, $k \leq K$, the set of CP vectors for sequences of length n and with k CP by $\mathcal{T}_{k,n}$. The loss of segmenting $x(z)_{1:n}$ the first sub-sequence of length n of $x(z)$ with $\tau \in \mathcal{T}_{k,n}$ is,

$$L_{k,n}(z, \tau) = \sum_{l=1}^{k+1} C(x(z)_{\tau_{l-1}+1:\tau_l}) .$$

Finally, denote the optimal loss and the optimal CP vector by,

$$L_{k,n}^{opt}(z) = \min_{\tau \in \mathcal{T}_{k,n}} L_{k,n}(z, \tau) \qquad\qquad T_{k,n}^{opt}(z) = \arg\min_{\tau \in \mathcal{T}_{k,n}} L_{k,n}(z, \tau) .$$

The truncation region $\mathcal{Z}$ becomes $\mathcal{Z} = \{z \in \mathbb{R} : T_{K,N}^{opt}(z) = \tau^{det}\}$.

Remark that $L_{K,N}(z, \tau)$ is actually quadratic in z so that computing $T_{K,N}^{opt}(z)$ for any $z \in \mathbb{R}$ can be achieved by finding the minimal quadratic form among a finite number of them. That is why, in the next section, we briefly discuss how to solve this problem.

2.3.1 Finding the pointwise minimum of a family of quadratics

In this paragraph, we are concerned with the following problem: given P_i, for $i = 1, \ldots, I$ a family of polynomials of degree at most 2, compute the function $z \mapsto \min_{i=1,\ldots,I} P_i(z)$, which is a piece-wise polynomial, and the optimal polynomials on each pieces. This question is actually tractable and its solution is a fundamental subroutine of the algorithm for selective p-value computation.

Its algorithmic solution relies on the following elementary lemma,

Lemma 1. *Assume that, for some $z_0 \in \{-\infty\} \cup \mathbb{R}$, and some $\epsilon > 0$, $P_1 = \min_{i=1,\ldots,I} P_i$ pointwise on $[z_0, z_0 + \epsilon]$.*

For simplicity, also suppose that, for all $i = 2, \ldots, I$, $P_i - P_1$ is exactly of degree two and write $P_i - P = a_i X^2 + b_i X + c_i$, $\Delta_i := b_i^2 - 4 a_i c_i$. Then, if

$$z_1 := \min \left\{ \frac{-b_i - \sqrt{\Delta_i}}{2a_i} : i = 2, \ldots, I,\ \Delta_i > 0,\ \frac{-b_i - \sqrt{\Delta_i}}{2a_i} > z_0 \right\} ,$$

then, for all $z \in [z_0, z_1]$, $P_1(z) = \min_{i=1,\ldots,I} P_i(z)$.

Further, if the values in the minimum above are all distinct, and the minimum is reached for the root corresponding to index $j \geq 2$, then there is some $\epsilon > 0$ such that $P_j = \min_{i=1,\ldots,I} P_i$ pointwise on $[z_1, z_1 + \epsilon]$.

The iteration of this lemma now gives an algorithm to compute $z \mapsto \min_{i=1,\ldots,I} P_i(z)$: starting from $z_0 = -\infty$, this lemma indeed explains how compute the next *breakpoint* z_1 at which the minimum polynomial changes.

2.3.2 Parametric Dynamic Programming

Armed with the procedure above, one could be tempted to solve directly, $T_{K,N}^{opt}(z) = \arg\min_{\tau \in \mathcal{T}_{K,N}} L_{K,N}(z, \tau)$. However, this would require enumerating $\mathcal{T}_{K,N}$ which is huge, since it has size $\binom{N-1}{K}$.

To bypass this issue, the authors [1] propose to bring in ideas from Dynamic Programming.

The Bellman equation in this context indeed reads, for any $z \in \mathbb{R}$,

$$L_{k,n}^{opt}(z) = \min\{L_{k,n}\left(z, \texttt{concat}\left(T_{k-1,m}^{opt}, m-1\right)\right) : m = k, \ldots, n-1\} .$$

But, define $\mathcal{T}_{k,n}^{opt} = \{T_{k-1,m}^{opt}(z) : z \in \mathbb{R}\}$ and, since $T_{k,n}^{opt}(z) = \arg\min_{\mathcal{T}_{k,n}^{opt}} L_{k,n}(z, \cdot)$, we have

$$L_{k,n}^{opt}(z) = \min\{L_{k,n}\left(z, \texttt{concat}\left(\tau, m-1\right)\right) : \tau \in \mathcal{T}_{k-1,m}^{opt},\ m = k, \dots, n-1\}$$
$$= \min\left\{L_{k,n}\left(z, \tau\right)\right) : \tau \in \bigcup_{m=k}^{n-1} \texttt{concat}(\mathcal{T}_{k-1,m}^{opt}, m-1)\Bigg\}\ ,$$

and similarly,

$$T_{k,n}^{opt}(z) = \arg\min\left\{L_{k,n}\left(z, \tau\right)\right) : \tau \in \bigcup_{m=k}^{n-1} \texttt{concat}(\mathcal{T}_{k-1,m}^{opt}, m-1)\Bigg\}\ . \tag{3}$$

Therefore, given the piecewise quadratic functions $T_{k-1,m}^{opt}$ for $m = k, \dots, n-1$, one can compute the piecewise quadratic function $T_{k,n}^{opt}$ with the subroutine described in the previous section §2.3.1. Repeating this procedure in a dynamic programming fashion results in a tractable method to compute the full function $T_{K,N}^{opt}(z)$.

However, the authors of [1] do not stop here: they provide another result to further reduce the computations. Indeed, they show [1, Lemma 4] that (3) can be rewritten as,

$$T_{k,n}^{opt}(z) = \arg\min\left\{L_{k,n}\left(z, \tau\right)\right) : \tau \in \bar{\mathcal{T}}_{k,n}\right\}\ ,$$

where $\bar{\mathcal{T}}_{k,n}$ is defined inductively by $\bar{\mathcal{T}}_{k,k} = \varnothing$ and, for $n \geq k+1$,

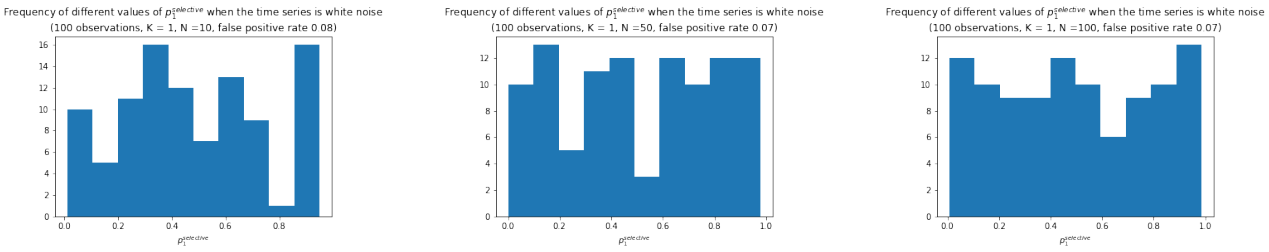
$$\bar{\mathcal{T}}_{k,n} = \texttt{concat}\left(\mathcal{T}_{k-1,n-1}^{opt}, n-2\right) \cup \left\{\tau \in \bar{\mathcal{T}}_{k,n-1} : \exists z \in \mathbb{R},\ L_{k,n-1}(z,\tau) \leq L_{k-1,n-1}^{opt}(z)\right\}\ .$$

Since only $\bar{\mathcal{T}}_{k,n-1}$ and $T_{k-1,n-1}^{opt}(z)$ are needed to build $\bar{\mathcal{T}}_{k,n}$, and that $\bar{\mathcal{T}}_{k,n}$ enables us to compute $T_{k,n}^{opt}(z)$, a similar dynamic programming scheme can be used to finally compute $T_{K,N}^{opt}(z)$.

3 Results and analysis

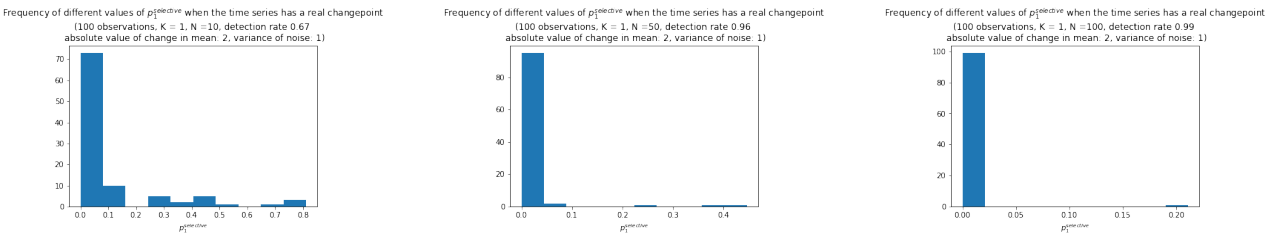
3.1 Synthetic data

A first simple test that what performed in the original article and that we reproduced, is to check that the distribution of $p^{selective}$ appears to be uniform when evaluated on white noise. The histograms of 100 observations of p values calculated on white noise can be found in Figure 1. This should be opposed to what happens in Figure 2 where there is a real changepoint and we expect the p-values to be small, allowing us to reject the null-hypothesis. The clear difference confirms the correctness of our implementation.



(a) $N = 10$ (b) $N = 50$ (c) $N = 100$

Figure 1: Distribution of $p^{selective}$ when applied to white noise



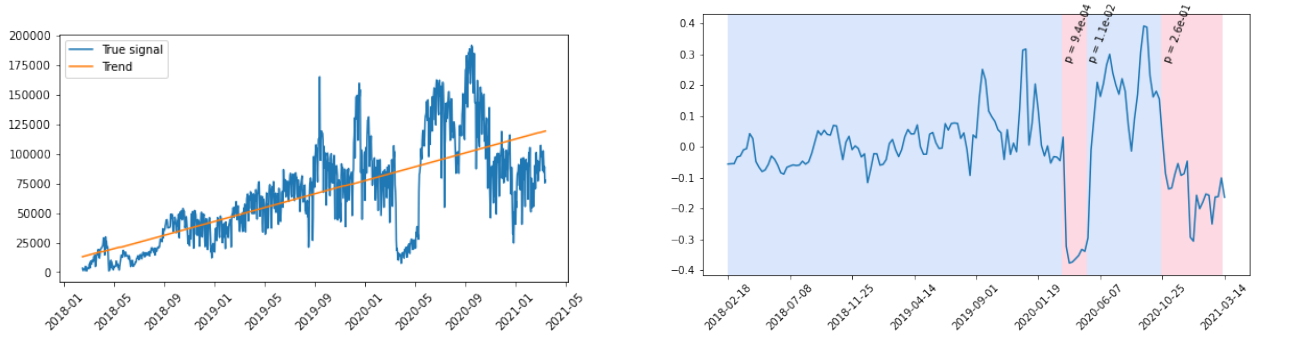
(a) $N = 10$ (b) $N = 50$ (c) $N = 100$

Figure 2: Distribution of $p^{selective}$ when applied to time series with changepoint

3.2 Bike data

We tested the algorithm on two realworld datasets based around bike usage in Paris.

The first is an estimation of the number of *Velib'* journeys made each hour. The data was collected and aggregated by `https://velib.philibert.info/` which relies on the official Velib' API. The number of journeys is estimated checking the number of velib at each station every 5 minutes and hence tracking which bikes moved. We noted that the data has a significant upwards trend as the service has grown more popular and more bikes were added over the years. We removed the trend before running any changepoint detection. We also decided to remove the the weekly periodic component by considering the number of journeys made each week. Our results are shown in Figure 3.

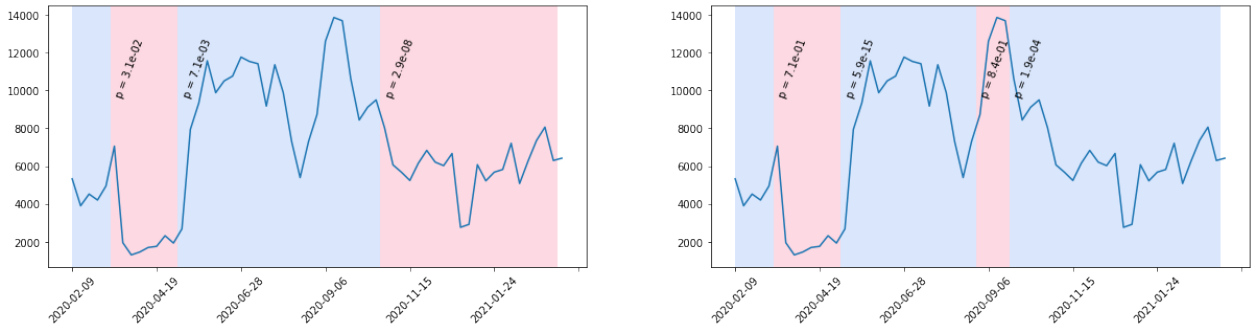


(a) Original Velib data

(b) Changepoint detection applied to detrended, renormalized, weekly data

Figure 3: Velib' data

The second dataset comes from Paris' Open Data initiative. The city has indeed deployed a network of sensors all around the city to measure bike traffic on its cycling equipment. This dataset consists of the hourly count of bikes passing by a given sensor for the last thirteen months. We arbitrarily selected the sensor from quai de Grenelle, and, as above, we removed the periodic component by summing over each week. We display the detected changepoints, as well as their associated p-values, for $K = 3$ and $K = 4$ in Figure 4 (where K denotes the total number of breakpoints).



(a) $K = 3$

(b) $K = 4$

Figure 4: Changepoints detected from bikelane sensor data

One of the big challenges when trying to calculate p-values associated with changepoints on real data is the estimation of the variance of the data. Indeed, we observed that the choice of Σ had a significant impact on the calculated p-values. In the end, we followed the recommendation of the paper: first determining τ^{det}, and then calculating an estimate of Σ by considering the maximal empirical variance of any of the subsections defined by τ^{det}. A possible extension of this work may be to mitigate this issue.

Another one of the shortcomings of our approach on these datasets is that the Gaussian model may not be adapted. Indeed, these series are counts journeys, a Poisson model might be a better fit. However, extending the work of [1] to losses other than the squared loss is far from straightforward.

References

[1] Vo Nguyen Le Duy et al. "Computing Valid p-value for Optimal Changepoint by Selective Inference using Dynamic Programming". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 11356–11367. URL: https://proceedings.neurips.cc/paper/2020/file/82b04cd5aa016d979fe048f3ddf0e8d3-Paper.pdf.

[2] William Fithian, Dennis Sun, and Jonathan Taylor. *Optimal Inference After Model Selection*. 2014. arXiv: 1410.2597 [math.ST].