Stockholms universitet

Multidimensional change point detection using likelihood ratio statistics

Max Brehmer

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2026:2
Matematisk statistik
Februari 2026

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2026:2**
http://www.math.su.se

Multidimensional change point detection using likelihood ratio statistics

Max Brehmer*

February 2026

Abstract

This thesis tackles binary splitting of regression trees through the lens of change-point detection. Consider a dataset with multidimensional features and a one-dimensional response variable. A binary split attempts to form partitions of observations with similar response values. A typical Classification and Regression Tree (CART) lacks an inherent stopping mechanism to avoid over-partitioning which leads to overfitting. CARTs tend to rely on cross-validation to reduce overfitting, but then one loses out on valuable training data. We propose a method that succeeds at generalizing without removing any data from the training set. We model this setup as a change point problem, where the change point is the index of an ordered dataset where the partitions are optimal. A likelihood ratio test is used to determine the significance of each recurring optimal change point. We first study the one-dimensional asymptotic distribution of the split location under the null hypothesis (that there is no change point).

Using a likelihood ratio statistic we recover the argmax of a Brownian bridge, which has an arcsine distribution, when the noise has finite variance. In the case where the noise has infinite variance, a stable-bridge limit results in an approximate Beta distribution approaching to uniformity as tails thicken. The limiting distribution of the statistic is approximated by a Gumbel distribution that changes by an affine scaling as dimensionality grows. Across Gaussian and t-distributed response variables, this method provides a solid method for partitioning datasets, while avoiding overfitting, and could be useful when regularizing regression trees.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: mabr3613@student.su.se. Supervisor: Johannes Heiny and Taariq Fahran Nazar.

Contents

1. Introduction

Change point detection is a fundamental problem in statistics, with applications across various fields such as time series analysis, finance, and supervised and unsupervised machine learning. The core idea is to identify points in a sequence of data where the statistical parameters, such as the mean, undergo a significant shift. In the univariate case, this problem has been studied extensively, where methods based on likelihood ratios have been demonstrated to be useful. The asymptotic properties of these methods have been rigorously analyzed, e.g. in [9, 15] for independent data points. With the increasing prevalence and importance of multidimensional datasets in contemporary data science applications, there is high demand for effective methods to detect change points in multidimensional data. Unlike univariate data, where each observation is a scalar value, multidimensional data represents features as vectors, adding complexity to the analysis. Recent progress has been made e.g. in [4, 8, 10].

This thesis aims to extend the framework of change point detection to the multidimensional setting. More precisely, the project focuses on adapting likelihood-based methods to handle the challenges introduced by high-dimensional datasets. A key objective is to analyze the asymptotic behaviour of likelihood ratio test statistics in settings where the sample size n and/or the dimensionality p are large. Understanding the asymptotic behaviour of likelihood ratio test statistics will allow us to apply the framework of change point detection to regularize regression trees. Engler et. al. [7] provides some inspiration for using change point detection to regularize CART trees.

1.1. **Notation.**

Convergence in distribution (resp. probability) is denoted by $\xrightarrow{\mathrm{d}}$ (resp. $\xrightarrow{\mathbb{P}}$), equality in distribution by $\overset{\mathrm{d}}{=}$, and unless explicitly stated otherwise all limits are for $n \to \infty$. We write random variables as X, Y, vectors as $\mathbf{x}, \mathbf{y}$, and $\mathbf{X}, \mathbf{Y}$ to represent matrices. Subscript X_k refers to the k-th observation of X, while superscript $X^{(i)}$ is used to describe the i-th feature dimension. T describes a single partition, while $\mathcal{T}$ describes an optimal partition. Assume $Y \sim \mathcal{N}$ and all $X \sim i.i.d.$ unless specified. Λ refers to a standard Gumbel distribution which has C.D.F. $\exp(-e^x)$.

2. Decision trees with binary splitting

In machine learning we are often interested in partitioning data into groups, where the simplest case is a binary partition. However, we are usually interested in partitioning the data in a data driven manner. Consider a dataset $\mathcal{D} = \{(X_k, Y_k)\}_{k=1}^n$ where $X_k \in \mathcal{X}$ for some p-dimensional feature space $\mathcal{X}$ and $Y_k \in \mathcal{Y}$ for some 1-dimensional response space $\mathcal{Y}$. We are interested in splitting the data in terms of X, into partitions which have similar values of Y. We achieve all this by minimizing some loss L and choose the split that results in the smallest loss. Trying to compare partitions of n data points is computationally infeasible. Instead, we can recursively split the data into two parts to obtain a satisfactory partition. This procedure is implemented for instance on Classification and Regression Trees (CART) [2].

A simple algorithm for obtaining binary partitions is to order the data according to the features $\left\{X^{(i)}\right\}_{i=1}^p$. That is, for each feature i, order the data $\mathcal{D}$ such that $X_1^{(i)} \leq X_2^{(i)} \leq ... \leq X_n^{(i)}$. Then construct a list $\boldsymbol{\pi}_i = \{\pi_i(1), \pi_i(2), ..., \pi_i(n)\}$ of time-indexes $\pi_i(k)$ which maps the k-th smallest observation to the index at which the observation exists in the original data. The ordered sequence is thus

$$X_{\pi_i(1)} \leq X_{\pi_i(2)} \leq ... \leq X_{\pi_i(n)}\,.$$

Which means that

$$\left\{\left(X_{\pi_i(k)}, Y_{\pi_i(k)}\right)\right\}_{k=1}^n$$

can be modeled as a time series with respect to feature i. The loss function with a binary split at time index r is

$$\begin{aligned} L_r &:= \sum_{k=1}^{r}(Y_k - \overline{Y}_{1:r})^2 + \sum_{k=r+1}^{n}(Y_k - \overline{Y}_{r+1:n})^2 \\ &= S_{1:r} - S_{r+1:n}\,, \end{aligned}$$

where $\overline{Y}_{l:r} = \frac{1}{r-l+1}\sum_{k=l}^{r} Y_k$ is the mean of the observations Y_k on the interval $l \leq k \leq r$ and the sum of squares is defined

$$S_{l:r} := \sum_{k=l}^{r}(Y_k - \overline{Y}_{l:r})^2 = \sum_{k=l}^{r} Y_k^2 - \frac{1}{r-l+1}\overline{Y}_{l:r}^2\,.$$

An optimal splitting index with respect to feature i is then chosen as $r^* = \arg\min_r L_r$. Comparing the optimal split across all features yields a global optimal split in the form of the pair (i^*, r^*),

where i^* is the dimension in which the optimal split occurs. This pair is then used to form the following binary partitions:

$$D_1 = \left\{ \left(X_{\pi_i(k)}, Y_{\pi_i(k)} \right) \in \mathcal{D} \mid X_{\pi_{i^*}(k)} \leq X_{\pi_{i^*}(r^*)} \right\}, \qquad k \in \{1, ..., n\}$$

and

$$D_2 = \left\{ \left(X_{\pi_i(k)}, Y_{\pi_i(k)} \right) \in \mathcal{D} \mid X_{\pi_{i^*}(k)} > X_{\pi_{i^*}(r^*)} \right\}, \qquad k \in \{1, ..., n\}.$$

This means only the dimension in which the change occurs is partitioned. Recursively applying the same procedure to each subsequent partition we can obtain finer partitions across various features and call the resulting final partitions a binary tree. Continually partitioning the data leads to overfitting, since in many cases the data cannot reasonably be partitioned further. The algorithm needs some stopping criterion to determine if the data can be partitioned or not. Without knowing the data generating process, we can never be sure if there is a change point in the data, however, we can use statistical tests to determine a confidence level for the partition and stop the procedure when we are confident, to some level, that the data can't be partitioned further.

3. Decision trees as a change point detection problem

We notice that our problem is similar to the problem of finding a change point, as explored in [7]. That is, we are interested in finding the index k where a change in the data occurs, we can directly use results from change point detection literature. The results from [15] provide an explicit construction for a statistical test for binary partitioning. In the case, where the feature space $\mathcal{X}$ is one-dimensional, we can use these results. However, the premise of this thesis is that we are interested in the case where the features are multidimensional. Therefore, we need to extend the model to the multidimensional setting. More precisely, we want to compute the asymptotic distribution for the loss L when the feature space $\mathcal{X}$ has dimension $p > 1$.

Consider the feature space $\mathcal{X} = \mathbb{R}^p$ and response space $\mathcal{Y} = \mathbb{R}$ where each feature is a sequence of $n \in \mathbb{N}$ random variables

$$\left\{ X_1^{(i)}, ..., X_n^{(i)} \sim F^{(i)} \right\}_{i=1}^{p}.$$

where F is some continuous distribution with nonzero variance. Response variables are constructed as

$$Y = \mu(X) + \varepsilon,$$

where $\mu(X)$ is a regression function of $X^{(1)}, ..., X^{(p)}$ and noise ε has i.i.d. Gaussian components with mean zero and variance $\sigma^2 > 0$. This means that our dataset is

$$\mathcal{D} = \left\{ \left(X_k^{(1)}, ..., X_k^{(p)} \right)^\top ; (Y_k, ..., Y_k)^\top \right\}_{k=1}^{n}.$$

The goal is to find the optimal change point (i^*, r^*) to partition at, i.e. the index r^* and feature i^* which minimizes the loss, then determine if it is statistically significant. For this we use a likelihood ratio test as in [15].

The null hypothesis is that there is no change point, i.e.

$$H_0 : r^* = n,$$

while the alternative hypothesis is that a change point does exist

$$H_1 : r^* < n\,.$$

If H_0 is rejected, meaning (i^*, r^*) is in fact a change point, then the process is repeated with the optimal partition. We continue to repeat this algorithm until after K iterations the K:th optimal change point $(i^*, r^*)_K$ is not statistically significant. The results from [15] show that the generalized negative log likelihood ratio of H_0 in the univariate Gaussian case is proportional to:

$$\mathcal{T}_n := \max_{1\le r\le n-1} \left[S_{1:n} - S_{1:r} - S_{r+1:n}\right]\,,$$

where

$$S_{l:m} = \sum_{k=l}^{m} \left(Y_k - \overline{Y}_{l:m}\right)^2$$

Under H_0, [15, Thm. 2.1] asserts that $\mathcal{T}_n$ has the asymptotic distribution of a non-standard Gumbel,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\mathcal{T}_n - \tilde{b}_n}{\tilde{a}_n} < x\right) = \exp\left(-2\pi^{-\frac{1}{2}} e^{-x}\right), \quad -\infty < x < \infty\,,$$

where the constants are $\tilde{a}_n = (2\log\log n)^{-\frac{1}{2}}$ and $\tilde{b}_n = \tilde{a}_n^{-1} + 2^{-1}\tilde{a}_n \log^{(3)} n$. In the case where there are $p > 1$ features, the test statistic takes the form

$$\mathcal{T}_n := \max_{1\le i\le p} \max_{1\le r\le n} T_r^{(i)}$$

where the loss for a particular partition (i, r) is

$$T_r^{(i)} := S^{(i)}_{\pi_i(1:n)} - S^{(i)}_{\pi_i(1:r)} - S^{(i)}_{\pi_i(r+1:n)}\,.$$

The optimal partition is found by maximizing T over elements r and features i

$$(i^*, r_i^*) := \operatorname*{arg\,max}_{\substack{1\le i\le p\\ 1\le r\le n-1}} T_r^{(i)}\,.$$

Throughout this paper we would prefer to work with statistics that converge on a standard Gumbel rather than the non-standard Gumbel shown above. To do this we must find a_n and b_n such that, under H_0:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\mathcal{T}_n - b_n}{a_n} < x\right) = \exp\left(-e^{-x}\right), \quad -\infty < x < \infty.$$

We call this standardized statistic $\tilde{\mathcal{T}}_n$, where

$$\tilde{\mathcal{T}}_n = \frac{\mathcal{T}_n - b_n}{a_n}\,.$$

We can derive the solution to a_n and b_n by using the Convergence to Types theorem from [6].

Theorem 3.1. *(Convergence to types theorem). Let A, B, A_1, A_2 be r.v.s. and $b_n > 0$, $\beta_n > 0$ and an a_n, $\alpha_n \in \mathbb{R}$ be constants. Suppose that*

$$b_n^{-1}(A_n - a_n) \xrightarrow{d} A\,.$$

Then the relation

$$\beta_n^{-1}(A_n - \alpha_n) \xrightarrow{d} B \qquad (*)$$

holds if and only if

$$\lim_{n\to\infty} b_n/\beta_n = b \in [0,\infty), \quad \lim_{n\to\infty} (a_n - \alpha_n)/\beta_n = a \in \mathbb{R}\,. \qquad (**)$$

If $(*)$ *holds then* $B \overset{d}{=} bA + a$ *and* a, b *are the unique constants for which this holds.*

When $(*)$ *holds* A *is non-degenerate if and only if* $b > 0$ *and then* A *and* B *belong to the same type.*

It is immediate from $(**)$ *that the constants* a_n *and* b_n *are uniquely determined only up to the asymptotic relation* $(**)$*.*

The constants from [15, Thm. 2.1] are as mentioned,

$$\tilde{a}_n = [2\log^{(2)} n]^{-\frac{1}{2}}$$
$$\tilde{b}_n = \tilde{a}^{-1} + \frac{1}{2}\tilde{a}_n \log^{(3)} n$$

where $\log^{(k)} n$ is the k-th iterated logarithm of n. Using Theorem 3.1, one finds that

$$a_n = [2\log^{(2)} n]^{-\frac{1}{2}}$$
$$b_n = \frac{1}{a_n} + \frac{1}{2}a_n \log^{(3)} n + a_n \log(2\pi^{-\frac{1}{2}})\,.$$

are the constants for which the statistic $\tilde{\mathcal{T}}_n$ converges on a standard Gumbel. See Appendix B for more detail.

3.1. **Quadratic forms of the test statistics.**

Moving forward it will be beneficial to work with the data in quadratic form. Our motivation for doing so are primarily that matrix operations are easier to generalize and perform on a large scale, and that the notation becomes clearer. In this section we present some useful lemmas and definitions in regard to the quadratic forms. Consider a response vector $\mathbf{y}$ of observations $Y_{k:l}$

$$\mathbf{y}_{k:l} := (Y_k, \quad Y_{k+1}, \quad ..., \quad Y_l)^\top\,.$$

Define a vector of ones by

$$\mathbf{1}_k := (1,\ 1,\ ...,\ 1)^\top \in \mathbb{R}^k$$

and set

$$\mathbf{J}_k := \mathbf{1}_k \mathbf{1}_k^\top\,.$$

Furthermore, for $1 \le k \le l \le n$ we set $\mathbf{1}_{k:l} = \mathbf{1}_{\{k,k+1,...,l\}} \in \mathbb{R}^n$, where $\mathbf{1}_{\{k,k+1,...,l\}}$ is a vector of length n that has ones at indices $\{k, k+1, ...l\}$ and zeroes elsewhere.

Lemma 3.2. *For* $1 \le k \le l \le n$*, one has*

$$\mathbf{J}_{k:l}^2 = \mathbf{1}_{k:l}\mathbf{1}_{k:l}^T\mathbf{1}_{k:l}\mathbf{1}_{k:l}^T = (l-k+1)\mathbf{1}_{k:l}\mathbf{1}_{k:l}^T = (l-k+1)\mathbf{J}_{k:l}\,,$$

where $\mathbf{J}_{k:l} \in \mathbb{R}^{n\times n}$ *is a block matrix with ones at places where the row and column are on the range* $\{k, ..., l\}$ *and zero elsewhere.*

Lemma 3.3. *It holds that the sum of squares corresponding to* $\mathbf{y}_{k:l}$ *is*

$$\begin{aligned} S_{k:l} &= \sum_{i=k}^{l} Y_i^2 - (l-k+1)\,\overline{Y}_{k:l}^2 \\ &= (\mathbf{y}_{k:l})^\top \mathbf{I}_{l-k+1}\,\mathbf{y}_{k:l} - \frac{1}{l-k+1}\,(\mathbf{y}_{k:l})^\top \mathbf{J}_{l-k+1}\,\mathbf{y}_{k:l}\,. \end{aligned}$$

Lemma 3.4. *We can write* T_r *in quadratic form as*

$$T_r = \mathbf{y}_{1:n}^\top \mathbf{Q}_r\,\mathbf{y}_{1:n}$$

where $\mathbf{Q}_r$ *is the following block matrix*

$$\mathbf{Q}_r = \left(\begin{array}{c|c} (-\frac{1}{n}+\frac{1}{r})\,\mathbf{J}_r & -\frac{1}{n}\mathbf{1}_r(\mathbf{1}_{n-r})^\top \\ \hline -\frac{1}{n}\mathbf{1}_{n-r}(\mathbf{1}_r)^\top & (-\frac{1}{n}+\frac{1}{n-r})\,\mathbf{J}_{n-r} \end{array}\right).$$

Proof. An application of Lemma 3.3 yields

$$\begin{aligned} T_r &= S_{1:n} - S_{1:r} - S_{r+1:n} \\ &= (\mathbf{y}_{1:n})^\top \mathbf{I}_n\,\mathbf{y}_{1:n} - \frac{1}{n}(\mathbf{y}_{1:n})^\top \mathbf{J}_n\,\mathbf{y}_{1:n} \\ &\quad - (\mathbf{y}_{1:r})^\top \mathbf{I}_r\,\mathbf{y}_{1:r} + \frac{1}{r}(\mathbf{y}_{1:r})^\top \mathbf{J}_r\,\mathbf{y}_{1:r} \\ &\quad - (\mathbf{y}_{r+1:n})^\top \mathbf{I}_{n-r}\,\mathbf{y}_{r+1:n} + \frac{1}{n-r}(\mathbf{y}_{r+1:n})^\top \mathbf{J}_{n-r}\,\mathbf{y}_{r+1:n} \\ &= (\mathbf{y}_{1:n})^\top \left(\begin{array}{c|c} (-\frac{1}{n}+\frac{1}{r})\,\mathbf{J}_r & -\frac{1}{n}\mathbf{1}_r(\mathbf{1}_{n-r})^\top \\ \hline -\frac{1}{n}\mathbf{1}_{n-r}(\mathbf{1}_r)^\top & (-\frac{1}{n}+\frac{1}{n-r})\,\mathbf{J}_{n-r} \end{array}\right)\mathbf{y}_{1:n} \\ &= (\mathbf{y}_{1:n})^\top \mathbf{Q}_r\,\mathbf{y}_{1:n}\,. \end{aligned}$$

□

4. Asymptotic distribution of optimal change points

Consider the one-dimensional case. For $k = 1, ..., n-1$ we want to understand the probability $P_k := \mathbb{P}(k \text{ maximizes } \mathcal{T}_n)$. We know that it must be symmetric around $\frac{n-1}{2}$, since under H_0 the time series is exchangeable. We model the probability by setting up the system of equations

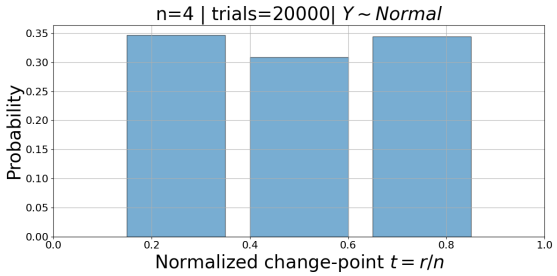
$$\sum_{k=1}^{n-1} P_k = 1\,, \qquad P_k = P_{n-k}\,.$$

In the case that n is even, there is a unique mid-point $r = \frac{n}{2}$ that partitions the data into equally sized groups where the system of equations is

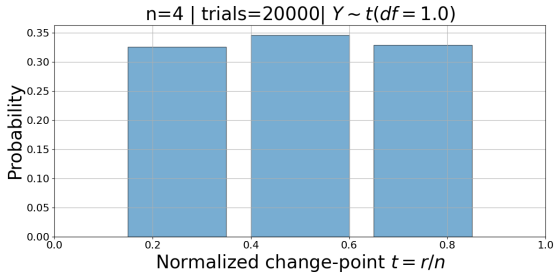
$$\sum_{k=1}^{n-1} P_k = 1\,, \qquad P_{n/2} = 1 - 2\sum_{k=1}^{\frac{n}{2}-1} P_k\,.$$

In the case that n is odd, the equation becomes

$$P_{\lfloor \frac{n}{2} \rfloor} = \left(1 - 2\sum_{k=1}^{\lfloor \frac{n}{2} \rfloor - 1} P_k\right) \cdot \frac{1}{2} = P_{\lfloor \frac{n}{2}+1 \rfloor}\,.$$



(A) Distribution of P_{r^*} in the case $Y \sim \mathcal{N}$



(B) Distribution of P_{r^*} in the case $Y \sim t_{\nu=1}$

Figure 1. Probability estimates for the location of optimal change points r^* using t-distributed outputs Y, based on 20 000 simulation trials. A Gaussian distribution is equivalent to the t-distribution when the degrees of freedom are $\nu \to \infty$, while a Cauchy distribution is when $\nu = 1$ and thus has heavier tails. The heavier tailed Cauchy distribution gives a higher probability to the partition occurring at the mid-point $r = \lfloor n/2 \rfloor$.

We conclude that a key component of understanding the general probability of optimal change points P_r, is to find the probability of the change point occurring at the middle-index $P_{\lfloor n/2 \rfloor}$. Before we attempt to find an analytical formula to determine this probability we want to get an idea of the distribution by simulating the time series $Y_1, ..., Y_n \sim D_Y$ and finding the arg max of T_n. In this section we normalize the change points by setting $t = \frac{r}{n}$, representing a time on $t \in [0, 1]$. This makes it easier to compare the various distributions. In figure 1 we compare the outcome of optimal change points in the smallest non-trivial case, that being $n = 4$, from a simulation of 20 000 trials, with the only difference being the distribution of Y. The figure clearly shows that the heavier tailed Cauchy distribution results in an optimal partition at the center more often than

when output variable has a thinner tailed Normal distribution. This finding establishes that P_r depends on the distribution D_Y when n is small. Next we explore what patterns emerge when the number of observations is larger.

In figure 2, we analyze the optimal change point distribution when the number of observations n is higher. Indeed; the difference in concavity between Normal and Cauchy distributed Y's persists for moderately sized n. In the figure we have also included a discrete approximation of the arcsine distribution for reference. This is because we observe P_r to converge on what appears to be an arcsine distribution. Note that the arcsine distribution in turn is a special case of the Beta distribution, $D_{\text{arcsin}} = \text{Beta}(\frac{1}{2}, \frac{1}{2})$. As we discuss further along this section, P_r is not expected to precisely follow the discrete approximation, rather the discrete approximation itself converges to the true arcsine distribution in the limit $n \to \infty$, as seen in figure 3. Further, we observe that regardless of the kurtosis (i.e. tailedness) of the distribution D_Y, the distribution P_r converges on a concave distribution, meaning change points near or at the extremes are more likely than points closer to the center. By understanding that the heavier the tail of a t-distribution, the more likely extreme outliers are, we can deduce that when there is an extreme outlier in the dataset, i.e.

$$Y_{k^*} = \max_k Y_k >> \overline{Y}_{1:n}\,,$$

the optimal partition is typically that Y_{k^*} belongs to the smaller set. This can be interpreted as the outlier skewing the mean away from most other observations. Leaving a heavy outlier outside the larger partition, decreases the loss, and thus ensures T_r gets maximized. Y_{k^*} is to be as isolated as possible. By observation, it appears that the distribution of $\frac{r}{n}$ converges to an arcsine distribution, which has density

$$f_{\text{arcsin}}(u) = \frac{1}{\pi\sqrt{u(1-u)}}, \quad 0 < u < 1\,.$$

Since the change points are integers, it is useful to use a discrete probability function to compare with. We therefore introduce a discrete approximation of the arcsine distribution. This distribution has the probability mass function

$$\mathbb{P}(r = k) = \frac{p_n(k)}{Z_n}\,,$$

where

$$p_n(k) = \frac{1}{\sqrt{k(n-k)}}, \quad k = 1, ..., n-1$$

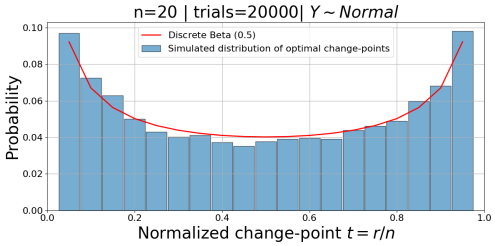
and

$$Z_n = \sum_{j=1}^{n-1} p_n(j)\,.$$

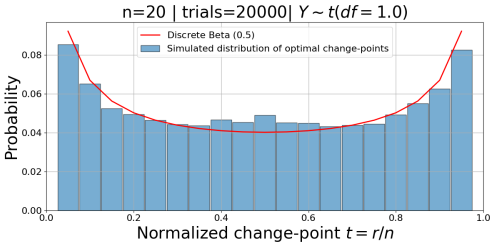
In the limit $n \to \infty$, it follows that

$$Z_n = \int_1^{n-1} \frac{1}{\sqrt{j(n-j)}}\, dj = \pi$$

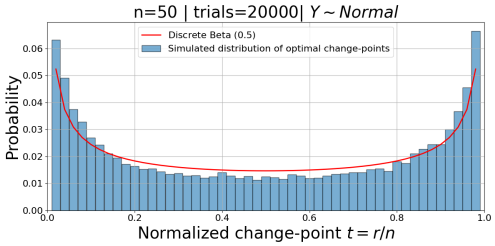
therefore, $Z_n \sim \pi$ and one recovers the probability $\mathbb{P}(r = k) = \frac{1}{\pi\sqrt{k(n-k)}}$.



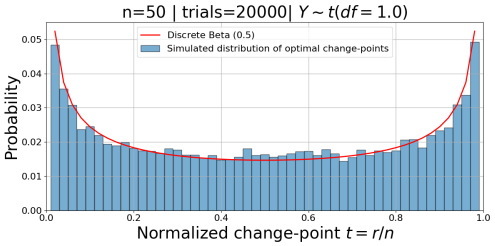
(A) $Y \sim \mathcal{N}$, $n = 20$



(B) $Y \sim t_1$, $n = 20$

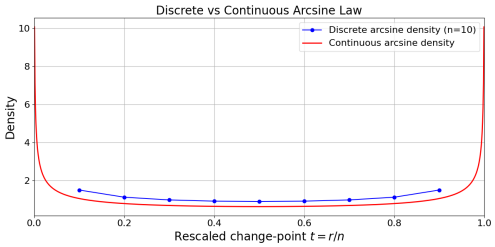


(C) $Y \sim \mathcal{N}$, $n = 50$

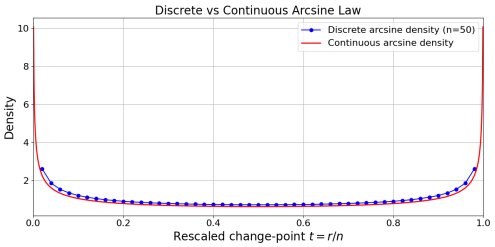


(D) $Y \sim t_1$, $n = 50$

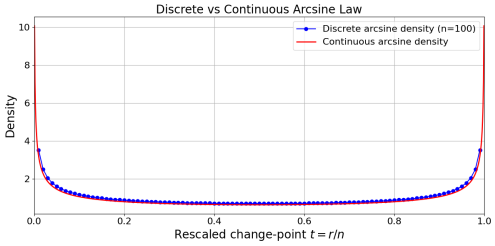
FIGURE 2. Empirical distribution of P_r at various levels of n and distributions of Y. Asymptotically, P_r can be approximated by a discrete variant of the arcsin /Beta$(\frac{1}{2}, \frac{1}{2})$ function. Even for larger n, P_r is less concave when Y has a heavy tailed distribution.



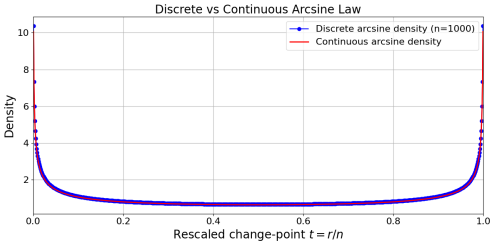
(A) $n = 10$



(B) $n = 50$

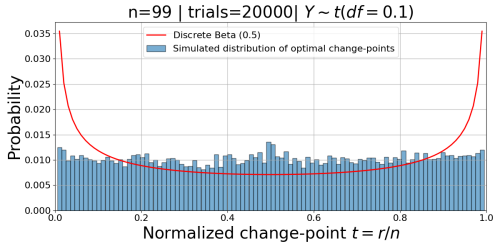


(C) $n = 100$

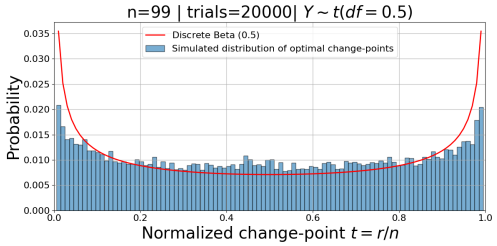


(D) $n = 1000$

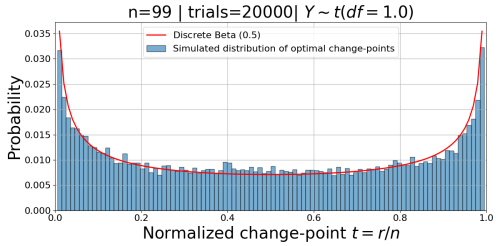
FIGURE 3. Comparison of the continuous arcsine density f_{arcsin} with the discrete arcsine PMF. The values of the PMF are scaled by a factor of one unit length $\frac{1}{\Delta_k} = n$ so that the density at point k is $d_k = \frac{p_k}{\Delta_k} = n \cdot p$. We observe that the discrete arcsine PMF converges to the density function as $n \to \infty$, but is also a good approximation for relatively small n.



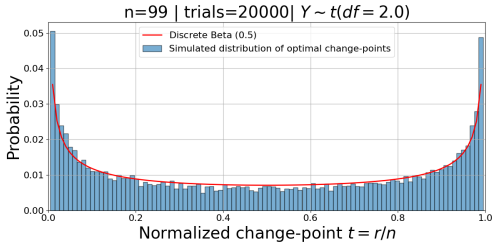
(A) $Y \sim t_{0.1}$, $n = 99$



(B) $Y \sim t_{0.5}$, $n = 99$



(C) $Y \sim t_1$, $n = 99$

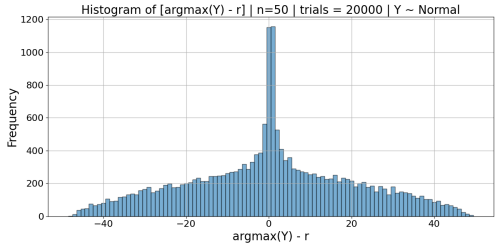


(D) $Y \sim t_2$, $n = 99$

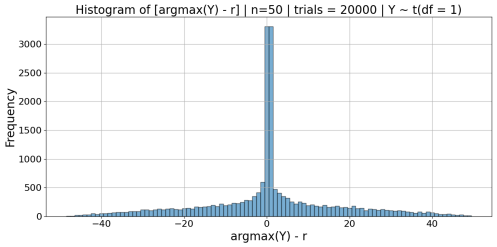
FIGURE 4. When tailedness of Y is high, the magnitude of extreme outliers are so great that only the location of the greatest outlier, which is uniform by independence of Y's, determines the change point. So it appears $P_r \approx U = \text{Beta}(1, 1)$ as $\nu \to 0$. In the case that Y has thinner tails, the magnitude of the greatest outlier is not strong enough to dominate the partition, instead we see that P_r is closer in distribution to $f_{\text{arcsin}} = \text{Beta}(\frac{1}{2}, \frac{1}{2})$.

In figure 4 we explore the normalized distribution P_t for various cases of the t-distribution, using $n = 99$. The conclusion we make is that thinner tailed distributions, such as Gaussians, converge on a limiting distribution that are close to $f_{\text{arcsin}} = \text{Beta}(\frac{1}{2}, \frac{1}{2})$, while the thicker tailed t-distributions ($\nu \to 0$) converge on distributions that approach a Uniform distribution $U = \text{Beta}(1, 1)$.

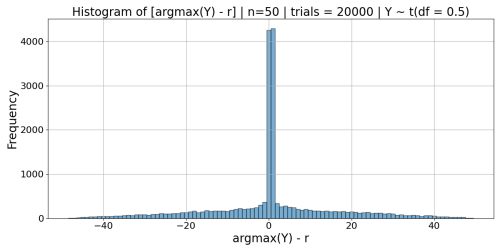
We have alluded to the importance of extreme outliers. To get a hint of how it affects the location of the optimal change point, we explore the distance between the largest outlier and the optimal change point, that is $[\hat{r} - r]$, where $\hat{r} := \arg\max_k(Y_k)$ in figure 5. We observe that the average distance is smaller when the tails are thick, meaning that the higher the kurtosis, and subsequently the larger the expected outliers $\max(Y)$, the impact of the greatest outlier grows. For very thick tailed distributions, the optimal change point is almost always the same as $\arg\max(Y)$. In figure 6 we see how the probability $\mathbb{P}\left(r = \arg\max_k(Y_k)\right)$ decreases with higher degrees of freedom (thinner tails). However if the greatest outlier is truly dominant, one would expect $\mathbb{P}(\hat{r} - r \in \{0, 1\}) \approx 1$. We observe no greater than ~ 0.5 for Y-distributions as thick-tailed as $t_{0.05}$, which is almost uniform.



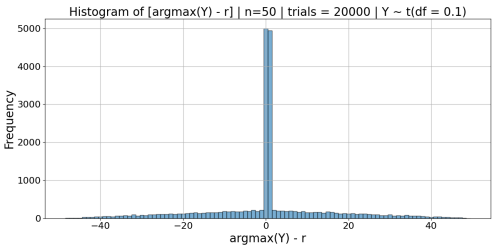
(A) $Y \sim \mathcal{N}$



(B) $Y \sim t(\nu = 1)$



(C) $Y \sim t(\nu = \frac{1}{2})$



(D) $Y \sim t(\nu = \frac{1}{10})$

FIGURE 5. Distribution of $\hat{r} - r$ where $\hat{r} = \arg\max_k Y_k$ and r is the optimal change point.

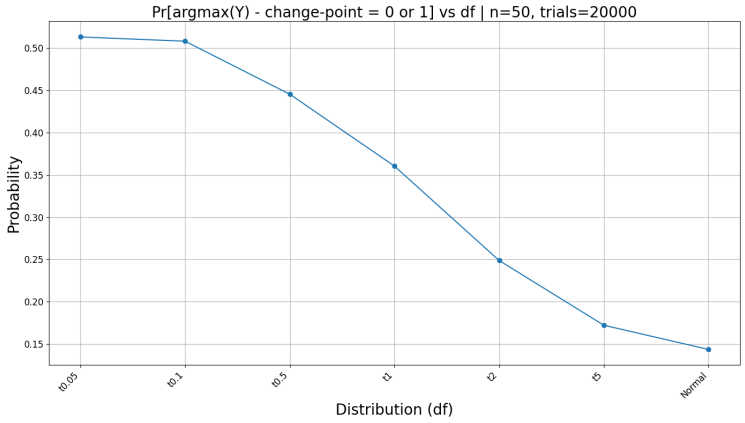


FIGURE 6. The probability $\mathbb{P}(\hat{r} - r \in \{0, 1\})$ (location of change point exactly same as maximum Y_k) across various degrees of freedom for t-distributed Y (Discrete steps, Not to scale). Note that change point index is right shifted, so $r = k$ means that the partition occurs after observation Y_k, which is why $\hat{r} - r = 1$ is a partition with $\arg\max(Y)$ as its final observation.

4.1. **Arcsine and beta laws.**

In order to find a proof for why the asymptotic distribution of the location of change points is an arcsine distribution, we begin by rewriting the expression for T_r as

$$T_r = S_{1:n} - S_{1:r} - S_{r+1:n} = \frac{(\tilde{S}_{1:r} - \frac{r}{n}\tilde{S}_{1:n})^2}{r(1-\frac{r}{n})}$$

where $\tilde{S}_{k:l} = \sum_{i=k}^{l}(Y_i)$, see [15, Eq. 2.1]. For simplicity we assume $\mathbb{E}[Y_k] = 0$. Theorem 4.1 provides insight into the asymptotic distribution.

Theorem 4.1 (Asymptotic distribution of the change-point location under H_0)**.** *Let* $(Y_k)_{k\geq 1}$ *be i.i.d. with* $\mathbb{E}[Y_1] = 0$ *and finite* $\mathrm{Var}(Y_1) = \sigma^2 > 0$*. Let*

$$\hat{r}_n = \operatorname*{arg\,max}_{1\leq r\leq n-1} T_r\,.$$

Then

$$\frac{\hat{r}_n}{n} \xrightarrow{d} D_{\arcsin}(0,1)$$

which has density

$$f(t) = \frac{1}{\pi\sqrt{t(1-t)}},\ 0 < t < 1\,.$$

Proof. **Step 1 (CUSUM process and scaling).** Since we are working with the limit $n \to \infty$, and $\tilde{S}_m := \sum_{i=1}^{m}\tilde{Y}_i$. Define the process

$$B_n(t) := \frac{\tilde{S}_{\lfloor nt\rfloor} - t\tilde{S}_n}{\sigma\sqrt{n}}, \quad t \in [0,1],$$

so that $T_{\lfloor nt\rfloor} = \frac{\left[\left(\tilde{S}_{\lfloor nt\rfloor} - t\,\tilde{S}_n\right)\frac{1}{\sqrt{n}}\right]^2}{t(1-t)} = \sigma^2\,\frac{B_n(t)^2}{t(1-t)}$.

Step 2 (Donsker's invariance principle). Define the partial-sum process $W_n(t) := \frac{\tilde{S}_{\lfloor nt\rfloor}}{\sigma\sqrt{n}}$. By Donsker's invariance principle [1, Thm. 8.2],

$$W_n(t) \xrightarrow{d} W(t) \quad \text{as } n \to \infty,$$

where W is standard Brownian motion. Noticing that $\frac{\tilde{S}_n}{\sigma\sqrt{n}} = W_n(1)$, we can apply the continuous mapping theorem to $(W(t), W(1))$, which gives

$$B_n(t) \xrightarrow{d} B(t) := W(t) - t\,W(1)\,,$$

where B is a standard Brownian bridge.

Step 3 (Argmax mapping). Define the estimated change-point fraction

$$\hat{t}_n := \operatorname*{arg\,max}_{t\in\{1/n,\ldots,(n-1)/n\}} \frac{B_n(t)^2}{t(1-t)}.$$

From Step 2, we have $B_n \xrightarrow{d} B$, where B is a standard Brownian bridge. We now view $\hat{t}_n$ as the image of B_n under the *argmax functional*

$$\mathcal{A}(f) := \operatorname*{arg\,max}_{t\in(0,1)} \frac{f(t)^2}{t(1-t)}\,.$$

For the continuous mapping theorem to apply, $\mathcal{A}$ must be continuous at the limit process B with probability one. This holds because:

- B has continuous sample paths almost surely,
- $\frac{B(t)^2}{t(1-t)}$ almost surely attains its global maximum at a *unique* $t \in (0,1)$.

Under these conditions, $\mathcal{A}$ is a.s. continuous at B (see [13, Thm. 3.2.2]).

Therefore, by the continuous mapping theorem,

$$\hat{t}_n = \mathcal{A}(B_n) \xrightarrow{d} \mathcal{A}(B) =: \hat{t}\,.$$

That is

$$\hat{t} := \operatorname*{arg\,max}_{t\in(0,1)} \frac{B(t)^2}{t(1-t)}.$$

The restriction of $\hat{t}_n$ to the n^{-1}-grid is asymptotically negligible, since the grid becomes dense in $(0,1)$ as $n \to \infty$ and $\frac{B(t)^2}{t(1-t)}$ is continuous a.s.

Step 4 (Change of coordinates). Consider the Brownian bridge

$$B(t) \overset{d}{=} (1-t)\, W\left(\frac{t}{1-t}\right), \quad t \in (0,1),$$

with W a standard Brownian motion. By the transformation $u := \frac{t}{1-t}$, We find that

$$\frac{B(t)^2}{t(1-t)} \overset{d}{=} \frac{W(u)^2}{u}\,.$$

The map $t \mapsto u = \frac{t}{1-t}$ is strictly increasing for $t \in (0,1)$. As $\hat{t}$ maximizes the bridge B, the Brownian Motion maximizer for u is

$$\hat{u} := \operatorname*{arg\,max}_{u>0} \frac{|W(u)|}{\sqrt{u}}, \quad u \in (0,\infty)\,.$$

The maximizer in t-space is

$$\hat{t} = \frac{\hat{u}}{1+\hat{u}}\,.$$

Thus, it suffices to identify the distribution of $\frac{\hat{u}}{1+\hat{u}}$.

Step 5 (Lévy's arcsine law). By the bijection $u \mapsto t = \frac{u}{1+u}$, we have

$$\hat{t} \;=\; \frac{\hat{u}}{1+\hat{u}}, \qquad \hat{u} \;:=\; \operatorname*{arg\,max}_{u>0} \frac{|W(u)|}{\sqrt{u}}.$$

Lévy's third arcsine law implies that the location of the maximum of $|W|$ on a finite interval is arcsine distributed [12, Ch. VI, Thm. 2.7.]. Using Brownian scaling and the monotone compactification $u \mapsto t = \frac{u}{1+u}$, it follows that

$$\hat{t} \;\sim\; \text{arcsine}(0,1), \qquad f_{\hat{t}}(t) \;=\; \frac{1}{\pi\sqrt{t(1-t)}}, \quad 0<t<1.$$

This completes the proof. □

Remark 4.2. *For increments with infinite variance, such as for t-distributions where $\nu < 2$, the Donsker invariance principle does not hold [3] and is replaced by the stable invariance principle of [11, Thm. 2], which shows that the CUSUM process, with scaling $n^{-\frac{1}{\alpha}}$, converges to an α-stable bridge $B_\alpha(t)$, where α is said to be the index of the stable law, related to tailedness. Doney [5] proves that the distribution of the time of the maximum of such stable processes is concave and symmetric about $t = \frac{1}{2}$. However it depends on α and is no longer arcsine. Thus, the arcsine law holds only in the finite-variance case ($\alpha = 2$), while for ($0 < \alpha < 2$) one must instead work with stable bridges, whose maximizer law lacks a closed form.*

Despite the limiting distribution case being unknown in the infinite variance, we suspect that it lies in the family of Beta distributions. More specifically, we know that for t_ν distributions

$$T_r \xrightarrow{d} \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right), \qquad \text{as } n,\ \nu \to \infty,$$

and we expect

$$T_r \xrightarrow{d} \text{Beta}\,(1,1), \qquad \text{as } n \to \infty,\ \nu \to 0,$$

however this remains unproven.

(A) $n = 1000$, $\nu \in (0, 2]$ (B) $n = 1000$, $\nu \in [1, 5]$

FIGURE 7. Maximum Likelihood Estimated parameters of the asymptotic distribution Beta(γ, γ) vs degrees of freedom of the underlying t-distributed observations. The parameter γ represents both standard parameters (α, β), since the limiting distribution must be symmetric. 50000 trials, $n = 1000$.

We model the cases $0 < \nu < 2$ as symmetric Beta-distributions with parameter $\frac{1}{2} < \gamma < 1$. Observing the simulated change points, we compute the MLE $\hat{\gamma}$ for such Beta-distribution. The results can be seen in figure 7*a*, which shows a somewhat linear or weak exponential relationship between γ and ν. Figure 7*b* looks at cases up to $\nu = 5$, which has infinite variance and in theory should yield parameters $\gamma = \frac{1}{2}$. We observe a result closer to $\lim_{\nu \to \infty} \gamma = 0.37$. A limitation of this estimation is that n is finite.

5. Multidimensional settings

5.1. **Set notation.**

In a multidimensional setting, it becomes difficult to follow which variables belong to which vectors, and in which order the observations occur. Fortunately there is another approach that does not require vectors to be re-ordered and that simplifies the notation while making many calculations a lot more forgiving. Thus, in this section we will describe the model in the form of sets rather than ordered sequences. Let $N = \{1, 2, ..., n\}$ where n is the number of observations in the dataset. Consider the the non-empty set $R_i \subsetneq N$ and its complement $R_i^c := N \setminus R_i$. When a partition of $\mathcal{D}$ is made, R_i contains the elements which belong to the partition before the change occurs, while the elements on the other side of the change point belong to the set R_i^c. The pre-change point set R_i may only contain elements that are ordered in terms of $\boldsymbol{\pi}_i$. This means that the order

$$\pi_i(1),\ \pi_i(2),\ ..., \pi_i(n)$$

sets the restriction

$$\pi_i(k) \in R_i \implies \pi_i(k-1) \in R_i \qquad \forall k \in N\,.$$

Each feature i allows R_i to select a new collection of indexes based on the order π_i. The null hypothesis that there is no change point is described as

$$H_0 : R^* = N$$

and the alternative hypothesis that a change point does exist

$$H_1 : R^* \subsetneq N\,.$$

In the case that we want to compare two T_i's, where the feature i is different, we use the notation

$$\begin{aligned} M &= R_i \\ B &= R_j \qquad \text{where } j \neq i\,. \end{aligned}$$

Many expressions in this section use the cardinality of the sets N, M, B and $M \cap B$, denoted as

$$n := |N|, \quad m := |M|, \quad b := |B|, \quad s := |M \cap B|\,.$$

Further, this means for the complementary sets that

$$|M^c| = n-m, \quad |R^c| = n-b, \quad |M \cap B^c| = m-s, \quad |M^c \cap B| = b-s, \quad |M^c \cap B^c| = n-m-b+s\,.$$

The following definitions and lemmas are useful to describe and solve the problem using a set approach.

Definition 5.1. *The matrix $\mathbf{J}_M$ is defined as*

$$\mathbf{J}_M := \mathbf{1}_M \mathbf{1}_M^\top$$

with vectors $\mathbf{1}_M$ constructed by the following

$$\mathbf{1}_M := \sum_{k \in M} \mathbf{e}_k \in \mathbb{R}^n \quad \textit{for } M \subset N.$$

where $\mathbf{e}_k$ is a zero-vector with 1 at position k and M is a set that represents indexes at which observations belong to said class.

Lemma 5.2. *(a) Multiplying matrices $\mathbf{J}_M$ and $\mathbf{J}_B$ forms the expression*

$$\mathbf{J}_M\mathbf{J}_B = \mathbf{1}_M \underbrace{\mathbf{1}_M^\top \mathbf{1}_B}_{s} \mathbf{1}_B^\top = s\left(\mathbf{1}_M\mathbf{1}_B^\top\right) \quad \text{for } M, B \subset N.$$

(b) In the case that $M = B$, the square $\mathbf{J}_M^2$ can be expressed as

$$\mathbf{J}_M^2 = \mathbf{1}_M\mathbf{1}_M^\top\mathbf{1}_M\mathbf{1}_M^\top = |M|\left(\mathbf{1}_M\mathbf{1}_M^\top\right) = m\,\mathbf{J}_M\,.$$

(c) The trace of the matrix $\mathbf{1}_M\mathbf{1}_B^\top$ that forms by multiplying two $\mathbf{J}$-matrices is

$$\operatorname{tr}\left(\mathbf{1}_M\mathbf{1}_B^\top\right) \overset{\text{cyclic property}}{=} \operatorname{tr}\left(\mathbf{1}_B^\top\mathbf{1}_M\right) \overset{\text{scalar}}{=} \mathbf{1}_R^\top\mathbf{1}_M = s\,.$$

(d) The Hadamard product between two $\mathbf{J}$-matrices with observations belonging to sets M and B takes the form

$$\mathbf{J}_M \circ \mathbf{J}_B = \left(\mathbf{1}_M\mathbf{1}_M^\top\right) \circ \left(\mathbf{1}_B\mathbf{1}_B^\top\right) = \sum_{i,j\in M}\sum_{k,l\in B}\left(\mathbf{e}_i\mathbf{e}_j^\top\right) \circ \left(\mathbf{e}_k\mathbf{e}_l^\top\right)\,.$$

Lemma 5.3. *(a) Trace of the matrix product of $\mathbf{J}_M$, $\mathbf{J}_B$*

$$\operatorname{tr}\left(\mathbf{J}_M\mathbf{J}_B\right) = |M \cap B|^2 = s^2 \quad \text{for } M, B \subset N.$$

(b) Trace of the Hadamard product of $\mathbf{J}_M$, $\mathbf{J}_B$

$$\operatorname{tr}\left(\mathbf{J}_M \circ \mathbf{J}_B\right) = |M \cap B| = s \quad \text{for } M, B \subset N.$$

Proof. (a) By multiplying two one-vector matrices

$$\mathbf{J}_M\mathbf{J}_B = \mathbf{1}_M\mathbf{1}_M^\top\mathbf{1}_B\mathbf{1}_B^\top\,.$$

Using properties of scalar multiplication we obtain

$$\mathbf{1}_M\mathbf{1}_M^\top\mathbf{1}_B\mathbf{1}_B^\top = s\cdot\mathbf{1}_M\mathbf{1}_B^\top\,.$$

By properties of the trace and the result from 5.2 (d), it follows that

$$\operatorname{tr}\left(s\cdot\mathbf{1}_M\mathbf{1}_B^\top\right) = s\cdot\operatorname{tr}\left(\mathbf{1}_M\mathbf{1}_B^\top\right) = s\cdot s = s^2\,.$$

(b) It is established in lemma 5.2 (d) that

$$\mathbf{J}_M \circ \mathbf{J}_B = \sum_{i,j\in M}\sum_{k,l\in B}\left(\mathbf{e}_i\mathbf{e}_j^\top\right) \circ \left(\mathbf{e}_k\mathbf{e}_l^\top\right)\,.$$

By linearity of the trace, it follows that

$$\begin{aligned}
\operatorname{tr}\left(\sum_{i,j\in M}\sum_{k,l\in B}\left(\mathbf{e}_i\mathbf{e}_j^\top\right) \circ \left(\mathbf{e}_k\mathbf{e}_l^\top\right)\right) &= \sum_{i,j\in M}\sum_{k,l\in B}\operatorname{tr}\left(\left(\mathbf{e}_i\mathbf{e}_j^\top\right) \circ \left(\mathbf{e}_k\mathbf{e}_l^\top\right)\right) \\
&= \sum_{i,j\in M}\sum_{k,l\in B}\delta_{ijkl} \\
&= |M \cap B| \\
&= s
\end{aligned}$$

where $\delta_{ijkl} = 1$ when $i = j = k = l$ and 0 otherwise are Kronecker deltas. Thus the trace only depends on the number of intersected elements. □

Using Lemma 3.4 we write the quadratic form of $T_{i,R}$ as

$$T_{i,R} = \left(\mathbf{y}_{1:n}^{(i)}\right)^\top \mathbf{Q}_R \left(\mathbf{y}_{1:n}^{(i)}\right),$$

where

$$\mathbf{Q}_R = -\frac{1}{n}\mathbf{J}_N + \frac{1}{r}\mathbf{J}_R + \frac{1}{(n-r)}\mathbf{J}_{R^c}\,.$$

The test statistic using a set approach is

$$\mathcal{T}_N = \max_{i\in\mathcal{P}} \max_{\substack{R\subsetneq N \\ R\neq\emptyset}} T_{i,R}\,.$$

where $\mathcal{P} = \{1, 2, ..., p\}$ is the set of input dimensions. Next, we collect some facts about Q_M and Q_B.

Corollary 5.4. *(a) Trace of the matrix* $\mathbf{Q}_M$

$$\mathrm{tr}\left(\mathbf{Q}_M\right) = 1\,.$$

(b) Trace of the matrix product of $\mathbf{Q}_M$, $\mathbf{Q}_B$

$$\mathrm{tr}\left(\mathbf{Q}_M\mathbf{Q}_B\right) = \frac{(mb-ns)^2}{mr(n-m)(n-b)}\,.$$

(c) Trace of the Hadamard product of $\mathbf{Q}_M$, $\mathbf{Q}_B$

$$\mathrm{tr}\left(\mathbf{Q}_M \circ \mathbf{Q}_B\right) = \frac{1}{n} + \frac{(ns-mb)(n-2m)(n-2b)}{nmb(n-m)(n-b)}\,.$$

Proof. (a) Write $Q_M = -\frac{1}{n}\mathbf{J}_N + \frac{1}{m}\mathbf{J}_M + \frac{1}{n-m}\mathbf{J}_{M^c}$. We have

$$\begin{aligned}
\mathrm{tr}(Q_M) &= \mathrm{tr}\left(-\frac{1}{n}\mathbf{J}_N + \frac{1}{m}\mathbf{J}_M + \frac{1}{n-m}\mathbf{J}_{M^c}\right) \\
&= -\frac{1}{n}\,\mathrm{tr}\left(\mathbf{J}_N\right) + \frac{1}{m}\,\mathrm{tr}\left(\mathbf{J}_M\right) + \frac{1}{n-m}\,\mathrm{tr}\left(\mathbf{J}_{M^c}\right) \\
&= -\frac{n}{n} + \frac{m}{m} + \frac{n-m}{n-m} = 1\,.
\end{aligned}$$

(b) Breaking the down the matrix we get

$$\begin{aligned}
\mathbf{Q}_M\mathbf{Q}_B &= \left(-\frac{1}{n}\mathbf{J}_N + \frac{1}{m}\mathbf{J}_M + \frac{1}{n-m}\mathbf{J}_{M^c}\right)\left(-\frac{1}{n}\mathbf{J}_N + \frac{1}{b}\mathbf{J}_B + \frac{1}{n-b}\mathbf{J}_{B^c}\right) \\
&= \frac{1}{n^2}\mathbf{J}_N^2 - \frac{1}{nb}\mathbf{J}_N\mathbf{J}_B - \frac{1}{n(n-b)}\mathbf{J}_N\mathbf{J}_{B^c} - \frac{1}{nm}\mathbf{J}_M\mathbf{J}_N + \frac{1}{mb}\mathbf{J}_M\mathbf{J}_B \\
&\quad + \frac{1}{m(n-b)}\mathbf{J}_M\mathbf{J}_{B^c} - \frac{1}{n(n-m)}\mathbf{J}_{M^c}\mathbf{J}_N + \frac{1}{b(n-m)}\mathbf{J}_{M^c}\mathbf{J}_B + \frac{1}{(n-m)(n-b)}\mathbf{J}_{M^c}\mathbf{J}_{B^c}\,.
\end{aligned}$$

Taking the trace we have

$$\begin{aligned}
\mathrm{tr}(\mathbf{Q}_M\mathbf{Q}_B) &= \frac{1}{n^2}\,\mathrm{tr}(\mathbf{J}_N^2) - \frac{1}{nb}\,\mathrm{tr}(\mathbf{J}_N\mathbf{J}_B) - \frac{1}{n(n-b)}\,\mathrm{tr}(\mathbf{J}_N\mathbf{J}_{B^c}) - \frac{1}{nm}\,\mathrm{tr}(\mathbf{J}_M\mathbf{J}_N) + \frac{1}{mb}\,\mathrm{tr}(\mathbf{J}_M\mathbf{J}_B) \\
&\quad + \frac{1}{m(n-b)}\,\mathrm{tr}(\mathbf{J}_M\mathbf{J}_{B^c}) - \frac{1}{n(n-m)}\,\mathrm{tr}(\mathbf{J}_{M^c}\mathbf{J}_N) + \frac{1}{b(n-m)}\,\mathrm{tr}(\mathbf{J}_{M^c}\mathbf{J}_B)
\end{aligned}$$

$$+\frac{1}{(n-m)(n-b)}\operatorname{tr}(\mathbf{J}_{M^c}\mathbf{J}_{B^c})$$
$$=\frac{n^2}{n^2}-\frac{b^2}{nb}-\frac{(n-b)^2}{n(n-b)}-\frac{m^2}{nm}+\frac{s^2}{mb}+\frac{(m-s)^2}{m(n-b)}-\frac{(n-m)^2}{n(n-m)}+\frac{(b-s)^2}{b(n-m)}+\frac{(n-m-b+s)^2}{(n-m)(n-b)},$$

which reduces to

$$\operatorname{tr}(\mathbf{Q}_M\mathbf{Q}_B)=\frac{(mb-ns)^2}{mr(n-m)(n-b)}.$$

(c) Using the same simplifications as in (b), we write the expression as

$$\operatorname{tr}(\mathbf{Q}_M\circ\mathbf{Q}_B)=\frac{1}{n^2}\operatorname{tr}(\mathbf{J}_N\circ\mathbf{J}_N)-\frac{1}{nb}\operatorname{tr}(\mathbf{J}_N\circ\mathbf{J}_B)-\frac{1}{n(n-b)}\operatorname{tr}(\mathbf{J}_N\circ\mathbf{J}_{B^c})-\frac{1}{nm}\operatorname{tr}(\mathbf{J}_M\circ\mathbf{J}_N)+\frac{1}{mb}\operatorname{tr}(\mathbf{J}_M\circ\mathbf{J}_B)$$
$$+\frac{1}{m(n-b)}\operatorname{tr}(\mathbf{J}_M\circ\mathbf{J}_{B^c})-\frac{1}{n(n-m)}\operatorname{tr}(\mathbf{J}_{M^c}\circ\mathbf{J}_N)+\frac{1}{b(n-m)}\operatorname{tr}(\mathbf{J}_{M^c}\circ\mathbf{J}_B)$$
$$+\frac{1}{(n-m)(n-b)}\operatorname{tr}(\mathbf{J}_{M^c}\circ\mathbf{J}_{B^c})$$
$$=\frac{n}{n^2}-\frac{b}{nb}+\frac{(n-b)}{n(n-b)}+\frac{m}{nm}+\frac{s}{mb}+\frac{m-s}{m(n-b)}+\frac{n-m}{n(n-m)}+\frac{b-s}{b(n-m)}+\frac{n-m-b+s}{(n-m)(n-b)},$$

which reduces to

$$\operatorname{tr}(\mathbf{Q}_M\circ\mathbf{Q}_B)=\frac{1}{n}+\frac{(ns-mb)(n-2m)(n-2b)}{nmb(n-m)(n-b)}.$$

□

Lemma 5.5. *(a) The covariance of the variables $T_M := T_{i,R}$ and $T_B := T_{j,R}$ takes the form*

$$\operatorname{Cov}(T_M,T_B)=f(n,m,b,s)$$

where $f(\cdot)$ is a function the intersection length s, the full lengths of observations n and the individual set lengths m, b, which models the covariance of the test statistics. This function is

$$f(n,m,b,s)=2\left[\frac{(mb-ns)^2}{mb(n-m)(n-b)}\right]$$
$$+(\nu_4-3)\left[\frac{1}{n}+\frac{(ns-mb)(n-2m)(n-2b)}{nmb(n-m)(n-b)}\right].$$

In the case the fourth moment of Y is $\nu_4 = 3$, such as when $Y \sim \mathcal{N}(\mu,\Sigma)$, the covariance is simply

$$\operatorname{Cov}(T_M,T_B)=2\left[\frac{(mb-ns)^2}{mb(n-m)(n-b)}\right].$$

(b) The variance of T_M can be expressed as

$$\operatorname{Var}(T_M)=2+(\nu_4-3)\left[\frac{1}{n}+\frac{(n-2m)^2}{nm(n-m)}\right].$$

(c) The correlation is given by

$$\operatorname{Corr}(T_M,T_B)=\frac{\operatorname{Cov}(T_M,T_B)}{\sigma_M\sigma_B}.$$

where $\sigma_M := \sqrt{\operatorname{Var}(T_M)}$ and $\sigma_B := \sqrt{\operatorname{Var}(T_B)}$ are the standard deviations of the respective test statistics.

Proof. (a) To find this function one must compute the second moment and the product of first moments

$$\mathrm{Cov}(T_M, T_B) = \mathbb{E}[T_M T_B] - \mathbb{E}[T_M]\mathbb{E}[T_B]\,.$$

By Lemma *A*.1 it follows that

$$\mathbb{E}[T_M T_B] = \mathrm{tr}(\mathbf{Q}_M)\,\mathrm{tr}(\mathbf{Q}_B) + 2\,\mathrm{tr}(\mathbf{Q}_M\mathbf{Q}_B) + (\nu_4 - 3)\,\mathrm{tr}(\mathbf{Q}_M \circ \mathbf{Q}_B).$$

We also have

$$\mathbb{E}[T_M]\mathbb{E}[T_B] = \mathrm{tr}(\mathbf{Q}_M)\,\mathrm{tr}(\mathbf{Q}_B) = 1.$$

Using Corollary 5.4, the covariance computes to

$$\begin{aligned}\mathrm{Cov}(T_M, T_B) = 2&\left[\frac{(mb-ns)^2}{mb(n-m)(n-b)}\right] \\ &+ (\nu_4 - 3)\left[\frac{1}{n} + \frac{(ns-mb)(n-2m)(n-2b)}{nmr(n-m)(n-b)}\right].\end{aligned}$$

(b) One can find $\mathrm{Var}(T_M)$ by setting $s, b = m$ in the expression for $\mathrm{Cov}(T_M, T_B)$

$$\begin{aligned}\mathrm{Var}(T_M) &= \mathbb{E}[(T_M)^2] - \mathbb{E}[T_M]^2 \\ &= \mathrm{tr}(\mathbf{Q}_M)^2 + 2\,\mathrm{tr}(\mathbf{Q}_M\mathbf{Q}_M) + (\nu_4 - 3)\,\mathrm{tr}(\mathbf{Q}_M \circ \mathbf{Q}_M) - \mathrm{tr}(\mathbf{Q}_M)^2 \\ &= 2\,\mathrm{tr}(\mathbf{Q}_M\mathbf{Q}_M) + (\nu_4 - 3)\,\mathrm{tr}(\mathbf{Q}_M \circ \mathbf{Q}_M) \\ &= 2 + (\nu_4 - 3)\left[\frac{1}{n} + \frac{(n-2m)^2}{nm(n-m)}\right].\end{aligned}$$

□

5.2. **General expression for covariance.** In order to understand the relationship between two test statistics, a key component is understanding the pairwise covariance, which may be beneficial to finding a joint distribution for change points in higher dimensional settings. The full probabilistic expression for the covariance between two test statistics is

$$\mathrm{Cov}(\mathcal{T}_n^{(i)}, \mathcal{T}_n^{(j)}) = \sum_{M,B \subsetneq N} w_{M,B} \cdot f(n, m, b, s)$$

where $w_{M,B} = \mathbb{P}(M, B \text{ maximizes } \mathcal{T}_n^{(i)} \text{ resp. } \mathcal{T}_n^{(j)})$ is a weight describing the probability of the set combination M and B occurring, while $f(n, m, b, s) = \mathrm{Cov}\,(T_M, T_B)$ is the covariance of two T's with chosen sets M, B. Consider the case where the feature correlation is $\rho_{ij} = 0$, this means that the order π_1 and π_2, and subsequently the choices of M and B are independent of one another. Define the probability that M, which has length $m = |M|$, maximizes $\mathcal{T}_n$ as

$$P_M := \mathbb{P}(M \text{ maximizes } \mathcal{T}_n)\,.$$

Now consider the set

$$\tilde{M} = \{1, 2, ..., m\} \qquad \text{where } m \leq n - 1\,.$$

It follows that

$$P_{\tilde{M}} = P_M$$

since all observations are i.i.d. and thus the probability that any set of length m maximizes the test statistic is the same. It also holds that

$$P_M = P_{M^c}$$

as both sets M and M^c imply that the change point occurs at the same place. Combining both properties means that $P_{\tilde{M}} = P_{\tilde{M}^c}$ and thus the probability for $m = k$ is equivalent to the probability for $m = n-k$, from which it follows that only the cases where $m \leq \lceil \frac{n-1}{2} \rceil$ form unique probabilities. The number of ways to pick B is

$$\left[\sum_{i=1}^{n-1} \binom{n}{i} \right] = 2^n - 2$$

since $1 \leq b \leq n-1$ and one can draw b elements from a total n.

Finding the weights $w_{M,B}$ seems complicated, however in Theorem 5.6 we manage to reduce the problem to a joint distribution of m and b.

Theorem 5.6. *Given i.i.d. features $X^{(1)}, ..., X^{(p)}$, and the set sizes m, b are known, the full covariance is*

$$\mathrm{Cov}(\mathcal{T}_n^{(1)}, \mathcal{T}_n^{(2)}) = \binom{n}{b}^{-1} \sum_{s=0}^{\min(m,b)} \binom{m}{s} \binom{n-m}{b-s} f(n, m, b, s)$$

where $f(n, m, b, s)$ is known from Lemma 5.5.

Proof. Consider π as a random permutation and write

$$T_M^{(1)} = T_M, \qquad \text{w.l.o.g. permutation is identity.}$$
$$T_B^{(2)} = T_{\pi(B)}, \qquad (2) \text{ indicates second component of } X.$$

$\mathrm{Cov}(\mathcal{T}_n^{(1)}, \mathcal{T}_n^{(2)})$ can be represented as $\mathbb{E}\left[T_M T_{\pi(B)}\right] - \mathbb{E}\left[T_M\right]\mathbb{E}\left[T_{\pi(B)}\right]$, as component (1) has optimal set M and component (2) has optimal set $\pi(B)$. We understand that

$$\mathbb{E}\left[T_M T_{\pi(B)}\right] - \mathbb{E}\left[T_M\right]\mathbb{E}\left[T_{\pi(B)}\right] = \sum_{\tilde{B}} \mathbb{E}\left[T_M T_{\tilde{B}} \mathbb{1}_{\{\pi(B)=\tilde{B}\}}\right] - \mathbb{E}\left[T_M\right]\mathbb{E}\left[T_{\tilde{B}} \mathbb{1}_{\{\pi(B)=\tilde{B}\}}\right],$$

where $\tilde{B}$ are possible permutations $\pi(B)$. Set $\delta_b := \{A \subseteq \{1, ..., n\} : |A| = b\}$. Note that for every $\tilde{B} \in \delta_b$ there are $b!(n-b)!$ permutations q such that $q(B) = \tilde{B}$. It follows that

$$\begin{aligned}
\mathbb{E}\left[T_M T_{\pi(B)}\right] - \mathbb{E}\left[T_M\right]\mathbb{E}\left[T_{\pi(B)}\right] &= \sum_{\tilde{B} \in \delta_b} \sum_{q: q(B)=\tilde{B}} \mathbb{P}(\pi = q) \left(\mathbb{E}\left[T_M T_{\tilde{B}}\right] - \mathbb{E}\left[T_M\right]\mathbb{E}\left[T_{\tilde{B}}\right]\right) \\
&= \sum_{\tilde{B} \in \delta_b} \frac{b!(n-b)!}{n!} \left(\mathbb{E}\left[T_M T_{\tilde{B}}\right] - \mathbb{E}\left[T_M\right]\mathbb{E}\left[T_{\tilde{B}}\right]\right) \\
&= \sum_{\tilde{B} \in \delta_b} \binom{n}{b}^{-1} \left(\mathbb{E}\left[T_M T_{\tilde{B}}\right] - \mathbb{E}\left[T_M\right]\mathbb{E}\left[T_{\tilde{B}}\right]\right) \\
&= \binom{n}{b}^{-1} \sum_{s=0}^{\min(m,b)} \sum_{\substack{\tilde{B} \in \delta_b \\ \tilde{s}=s}} f(n, m, b, s),
\end{aligned}$$

where $\tilde{s} = |M \cap \tilde{B}|$. Set $c_{s,m,b} := |\{\xi \in \delta_b : |\xi \cap \{1, ..., m\}| \} = s|$. If $s > \min(m, b)$, we have $c_{s,m,b} = 0$. For $s \leq \min(m, b)$

$$c_{s,m,b} = \binom{m}{s}\binom{n-m}{b-s}.$$

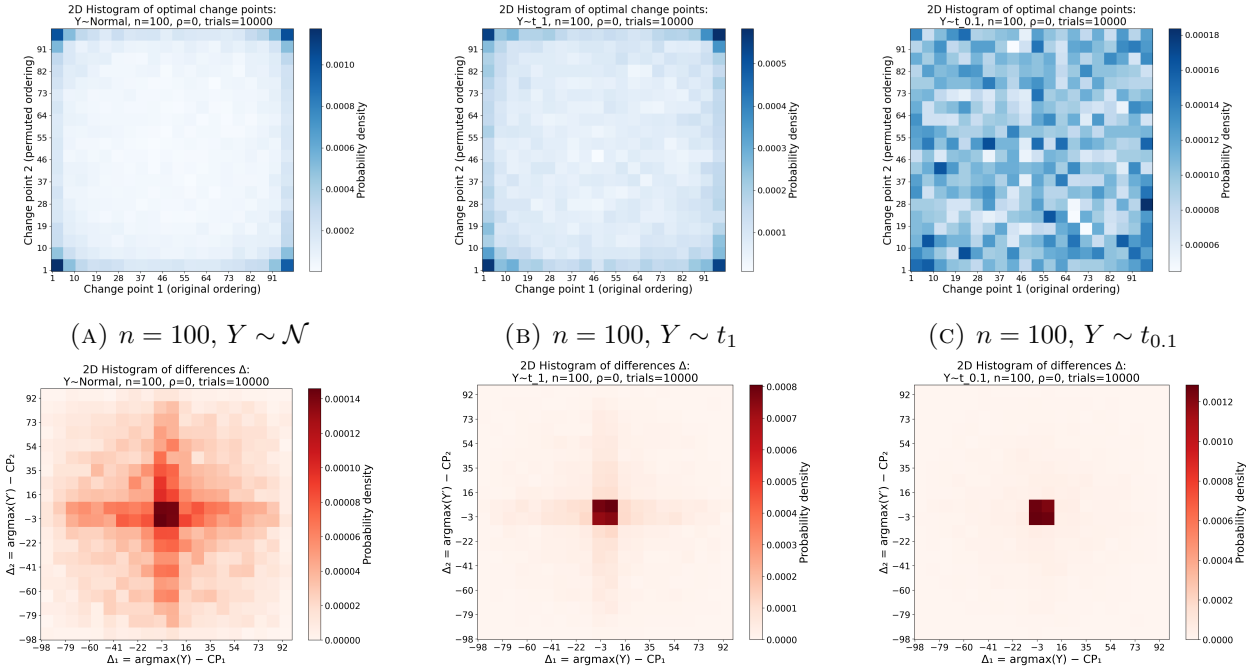
It follows

$$\mathbb{E}[T_M T_{\pi(B)}] - \mathbb{E}\left[T_M\right]\mathbb{E}\left[T_{\tilde{B}}\right] = \binom{n}{b}^{-1} \sum_{s=0}^{\min(m,b)} \binom{m}{s}\binom{n-m}{b-s} f(n, m, b, s).$$

□

5.3. **The case $p = 2$.**

We use the same methodology to find optimal change points as in the one-dimensional case, with the additional step of performing a permutation of the observations $Y_1, ..., Y_n$, which is akin to creating a new X-dimension. After finding the change points in the second dimension, a joint optimal change point $(r^{(1)}, r^{(2)})$ is established. In figure 8 we observe the distribution of two-dimensional change points using underlying distributions $(\mathcal{N}, t_1, t_{0.1})$ for Y. We observe that the density is the greatest at the corners as one would expect for two independent arcsin-distributed variables, with increasing uniformity as the distributions get more heavy tailed. We also compare the distance between the largest observation $\hat{r}$ and change point r in the 2-dimensional setting.



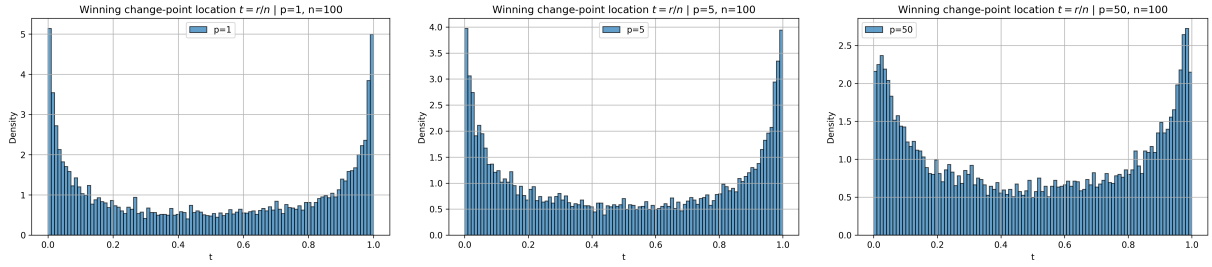
(A) $n = 100$, $Y \sim \mathcal{N}$ (B) $n = 100$, $Y \sim t_1$ (C) $n = 100$, $Y \sim t_{0.1}$

(D) $n = 100$, $Y \sim \mathcal{N}$ (E) $n = 100$, $Y \sim t_1$ (F) $n = 100$, $Y \sim t_{0.1}$

FIGURE 8. $2d$ histogram of optimal change point location (blue), and difference between the largest outlier vs change point $\hat{r} - r$ (red). Here we have simulated independent outputs $Y_1, ..., Y_n$ and found the optimal change point $r^{(1)}$, then randomly permuted the Y's. This permutation is the order of the second dimension and from there the optimal change point w.r.t dimension 2 is found. The graphs show only the location of the joint optimal change point $(r^{(1)}, r^{(2)})$. The number of bins are capped at 20.

5.4. **The case $p \to \infty$.**

We have seen that the joint distribution of two variables is very similar to the product of their marginals, next we investigate what the distributions look like when there are significantly more features. Figure 9 shows that the distribution of the optimal change points in each feature changes with increasing dimensionality. In figure 10 we use a 2-dimensional histogram to compare the difference between the asymptotic change point distribution and a 2-dimensional arcsine distribution. This allows us to understand how increased dimensionality changes the distribution in relation to arcsine probability. We observe that the corners give a higher probability to the empirical probability. This suggests that the change point distribution is more likely to generate an extreme outcome than the theoretical $2d$-arcsine distribution. The plots in figure 11 show that the distribution of $\tilde{\mathcal{T}}_{n,p}$ loosely approximates a standard Gumbel, but we find that the standardization constants a_n and b_n are not sufficient for stabilization to a standard Gumbel for $p > 1$. The QQ-plots in figure 12 corresponding to the distributions in figure 11 show that the mean shifts significantly with p, while the variance stays relatively stable, albeit the constant a_n for $p = 1$ may need refining as we observe a thinner tail than expected. Figure 13 explores taking a shift $d_p = \log p$ to compensate for the rightward shift observed.



(A) $p = 1, n = 100, Y \sim \mathcal{N}$ (B) $p = 5, n = 100, Y \sim \mathcal{N}$ (C) $p = 50, n = 100, Y \sim \mathcal{N}$

FIGURE 9. Location of optimal change point within the feature in which it lies. The figure shows how the distribution changes with dimensionality p.

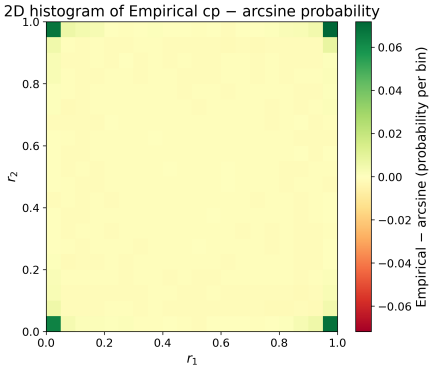
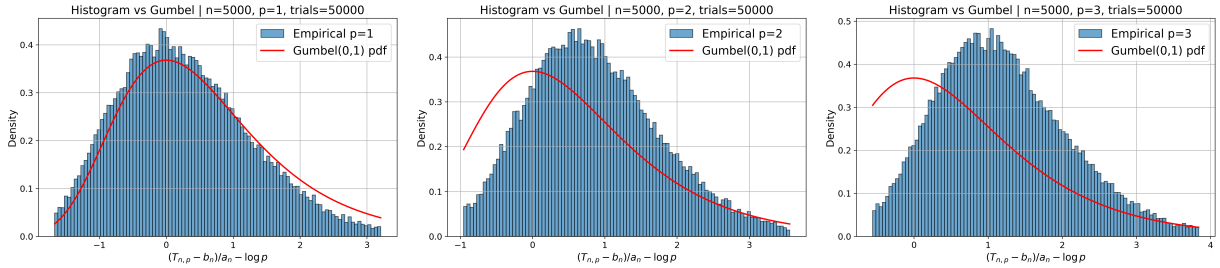
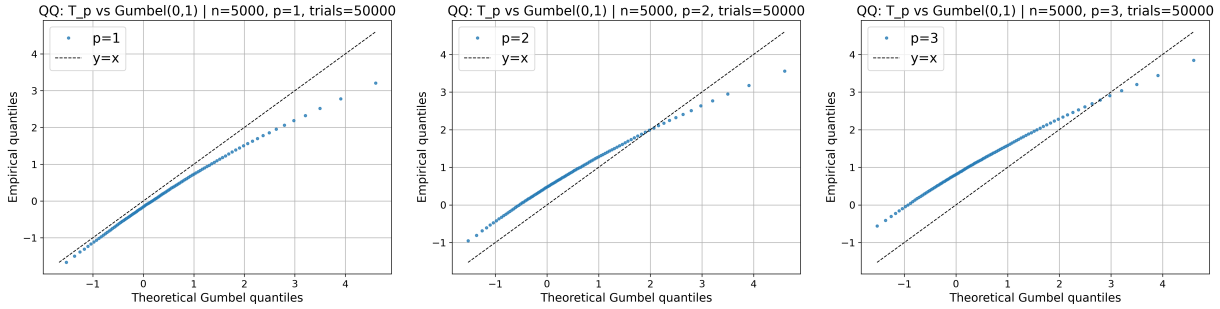


FIGURE 10. 2-dimensional histogram comparing the observed probability of change point location against an arcsine distribution.



(A) $p = 1, n = 5000$ (B) $p = 2, n = 5000$ (C) $p = 3, n = 5000$

FIGURE 11. Asymptotic distribution of $\tilde{\mathcal{T}}_{n,p}$ compared to a standard Gumbel.



(A) $p = 1, n = 5000$ (B) $p = 2, n = 5000$ (C) $p = 3, n = 5000$

FIGURE 12. QQ-plot comparing the distributions of simulated test statistics against a standard Gumbel. The dashed lines represent the theoretical quantiles of a standard Gumbel, while the blue data points are the simulated quantiles.

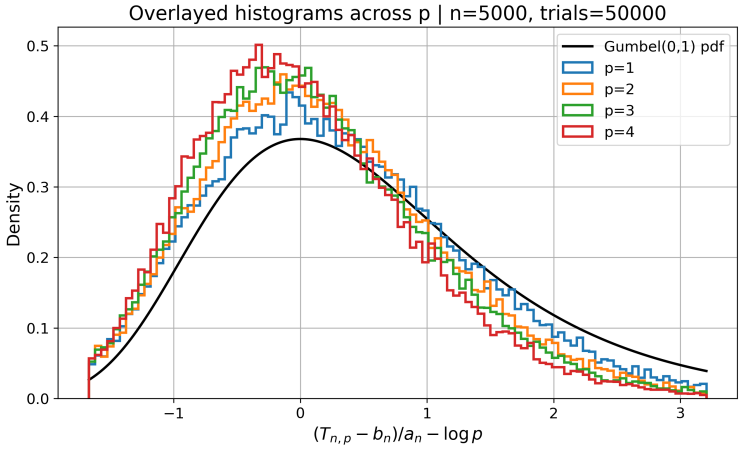


FIGURE 13. Distribution of $\tilde{\mathcal{T}}_{n,p} = \frac{\mathcal{T}_{n,p} - b_n}{a_n} - \log(p)$ under H_0. The constant $d_p = \log(p)$ appears to center the mode fairly well around 0, but some additional scaling constant c_p may be needed.

With each dimension we need to consider another Gumbel-distributed statistic $G_i = \tilde{\mathcal{T}}_n^{(i)}$. This means when the dimensionality of $\mathcal{D}$ is p one takes the maxima of p Gumbels. Fortunately, the Gumbel distribution is max-stable [6, Thm. 3.2.2]. Max-stability means that the maximum of p draws from the same distribution F has the same distribution as an affine scaling of F. In our case, a Gumbel distributed r.v. G satisfies

$$\max\{G_1, ..., G_p\} \overset{d}{=} c_p G + d_p$$

where $G_1, ..., G_p$ are i.i.d. and $c_p > 0,\ d_p \in \mathbb{R}$. Consider the variable $M_p = \max_{1 \leq i \leq p} G_i$, then for i.i.d. Gumbels $G_1, ..., G_p$, it holds that

$$\begin{aligned}
\max\{G_1, ..., G_p\} &= \mathbb{P}(M_p \leq t) \\
&= \prod_{i=1}^{p} \mathbb{P}(G_i \leq t) \\
&= \Lambda(t)^p \\
&= \exp\{-e^{-t}\}^p = \exp\{-p\, e^{-t}\} = \exp\{-e^{-t+\log p}\} \\
&= \Lambda(t + \log p)\,.
\end{aligned}$$

Thus the constants are found to be $c_p = 1$ and $d_p = \log p$ as shown in [6, Def. 3.2.6]. Heuristically this suggests that

$$c_p \left(\frac{\mathcal{T}_{n,p} - b_n}{a_n} \right) - d_p \sim \Lambda(x)$$

where a_n, b_n are the 1-dimensional constants needed for $\mathcal{T}_n$ to become standard Gumbel, while c_p, d_p are max-stable constants to compensate for growing dimensionality.

The constants a_n and b_n inspired by [15] are approximations and not proven to hold perfectly. Not all literature on this topic uses the same constants, so we began exploring with various similar constants a_n^*, b_n^*. By using the squared statistic $T_n^2 = \frac{(\tilde{S}_{1:r} - \frac{r}{n}\tilde{S}_{1:n})^2}{r(1-\frac{r}{n})}$ instead of the unsquared version used in [15], we set up the system of equations

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\mathcal{T}_n - b_n}{a_n} < x \right) = \lim_{n\to\infty} \mathbb{P}\left(\mathcal{T}_n < a_n x + b_n\right) = \exp\left(-e^{-x}\right) \qquad (\star)$$

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\mathcal{T}_n^2 - b_n^*}{a_n^*} < x \right) = \lim_{n\to\infty} \mathbb{P}\left(\mathcal{T}_n^2 < a_n^* x + b_n^*\right) = \exp\left(-e^{-x}\right) \qquad (\star\star)$$

In order to express the desired constants a_n^*, b_n^* in terms of a_n, b_n we can square the expression in $(\star)$ to get

$$\mathbb{P}\left(\mathcal{T}_n^2 < a_n^2 x^2 + a_n b_n x + b_n^2\right) = \mathbb{P}\left(\mathcal{T}_n^2 < a_n^* x + b_n^*\right)$$

and subsequentely

$$a_n^2 x^2 + 2 a_n b_n x + b_n^2 = a_n^* x + b_n^*\ .$$

Here the second order term $a_n^2 x^2 = 0 \cdot x^2$ vanishes and are left with

$$2 a_n b_n x + b_n^2 = a_n^* x + b_n^*\ .$$

Thus, we find that

$$\begin{aligned} a_n^* = 2a_n b_n &= 2a_n \left(\frac{1}{a_n} + \frac{1}{2} a_n \log^{(3)} n + a_n \log(2\pi^{-\frac{1}{2}}) \right) \\ &= 2 + a_n^2 \log^{(3)} n + a_n^2 \log \left(\frac{4}{\pi} \right) \\ &= 2 + \frac{1}{2 \log^{(2)} n} \left(\log^{(3)} n + \log \left(\frac{4}{\pi} \right) \right) \end{aligned}$$

where the $o(\log^{(2)} n)^{-1}$ term dominates the $o(\log^{(3)} n)$ and $o(1)$ terms as $n \to \infty$, meaning

$$\lim_{n \to \infty} \; a_n^* = 2 \, .$$

We also have

$$\begin{aligned} b_n^* = b_n^2 &= \left(\frac{1}{a_n} + \frac{1}{2} a_n \log^{(3)} n + a_n \log(2\pi^{-\frac{1}{2}}) \right)^2 \\ &= \frac{1}{a_n^2} + \log^{(3)} n + \log \left(\frac{4}{\pi} \right) \\ &\quad + a_n^2 \left(\frac{1}{4} (\log^{(3)} n)^2 + \log^{(3)} n \cdot \log(2\pi^{-\frac{1}{2}}) + (\log(2\pi^{-\frac{1}{2}}))^2 \right) \end{aligned}$$

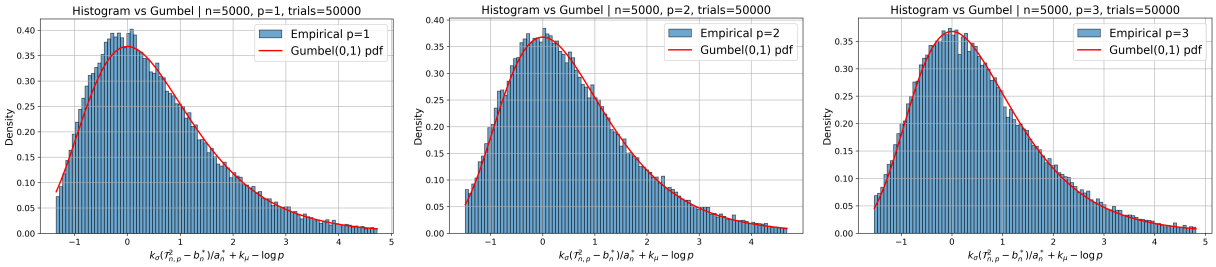
where $a_n^2 = o(\log^{(2)} n)^{-1}$ dominates the terms inside the brackets which are

$$o([\log^{(3)} n]^2) + o(1 \cdot \log^{(3)} n) + o(1) \; .$$

Asymptotically, dropping the terms that $\to 0$ as $n \to \infty$ we get the constant

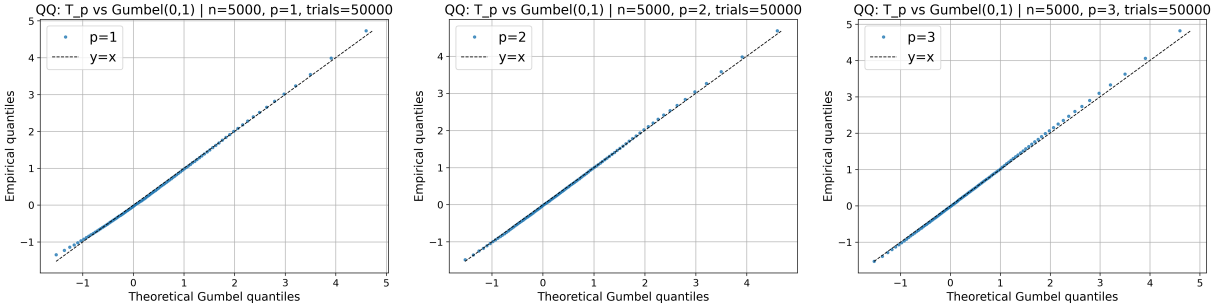
$$\lim_{n \to \infty} b_n^* = \lim_{n \to \infty} \left\{ 2 \log^{(2)} n + \log^{(3)} n + \log \left(\frac{4}{\pi} \right) \right\} .$$

Which order terms to keep is an arbitrary decision. Keeping the linear and greater terms appears to work quite well. Using the limit constants for a_n^* and b_n^* we find that, when $n = 5000$, the mean and standard deviation of the distribution of $\mathcal{T}_n^* := \frac{\mathcal{T}_n^2 - b_n^*}{a_n^*}$ are $\mu_{obs}(n = 5000) = 0.473$ and $\sigma_{obs}(n = 5000) = \; 1.335$. The theoretical values of a standard Gumbel are $\mu = \gamma \approx 0.577$, where γ is the Euler-Mascheroni constant, and $\sigma = \frac{\pi}{\sqrt{6}} \approx 1.283$. Correcting for this disparity we introduce a variance-correction factor $k_\sigma(n = 5000) = \frac{\sigma}{\sigma_{obs}} = \frac{1}{1.04}$ and mean-correction term $k_\mu(n = 5000) = \mu - \mu_{obs} = 0.104$. This means we get a good standard Gumbel approximation with $\tilde{\mathcal{T}}_n^* := k_\sigma(n) \, \mathcal{T}_n^* + k_\mu(n)$. Figures $14 - 16$ demonstrate the accuracy of these constants including the shift $d_p = \log p$. Why the constants we derived for the squared statistic appear to work significantly better than the constants we derived for the unsquared statistic from [15] is not clear. The final step is to find constants c_p and d_p such that $c_p \tilde{\mathcal{T}}_{n,p}^* - d_p \sim \Lambda$. Recall that we expect $c_p = 1$ and $d_p = \log p$ due to max-stability. Table 1 explores the effectiveness of $c_p = 1$ and $d_p = \log p$ and finds that these constants do not deviate μ_{obs} and σ_{obs} of the empirical distribution from μ and σ of Λ by more than 5% for $p \leq 6$.



(A) $p = 1, n = 5000$ (B) $p = 2, n = 5000$ (C) $p = 3, n = 5000$

FIGURE 14. Asymptotic distribution of the statistic $c_p\tilde{\mathcal{T}}^*_{n,p} - d_p$, where $\tilde{\mathcal{T}}^*_{n,p} := k_\sigma(n)\frac{\mathcal{T}^2_{n,p}-b^*_n}{a^*_n} + k_\sigma(n)$ and $c_p = 1$, $d_p = \log p$.



(A) $p = 1, n = 5000$ (B) $p = 2, n = 5000$ (C) $p = 3, n = 5000$

FIGURE 15. QQ-plot of $c_p\tilde{\mathcal{T}}^*_{n,p} - d_p$ against a standard Gumbel.

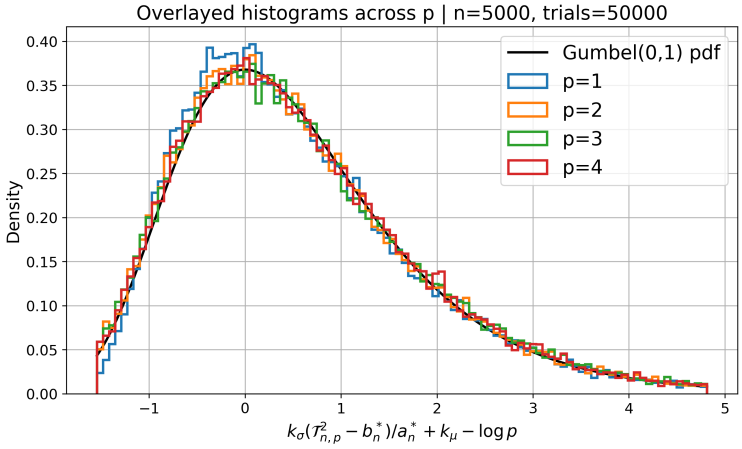


FIGURE 16. Distribution of $\tilde{\mathcal{T}}^*_{n,p} - \log(p)$ under H_0. For small p the constants $c_p = 1$ and $d_p = \log p$ produce a distribution very close to a standard Gumbel.

n	p	μ_{obs}	$\mu - \mu_{obs}$	$\frac{\mu_{obs}}{\mu}$	σ_{obs}	$\sigma - \sigma_{obs}$	$\frac{\sigma_{obs}}{\sigma}$
5000	1	0.577	~ 0	~ 1	1.283	~ 0	~ 1
5000	2	0.574	-0.003	0.995	1.302	$+0.019$	1.015
5000	3	0.592	$+0.015$	1.026	1.332	$+0.049$	1.038
5000	4	0.586	$+0.009$	1.016	1.316	$+0.033$	1.026
5000	5	0.599	$+0.022$	1.038	1.317	$+0.034$	1.027
5000	6	0.601	$+0.024$	1.042	1.319	$+0.035$	1.027

TABLE 1. Mean and standard deviations of $\tilde{\mathcal{T}}^*_{n,p} - \log(p)$ compared to a standard Gumbel. The table suggests that $c_p < 1,\ \forall p \geq 1$ and $d_p > \log p$ (for $p \geq 3$). The ratio of the observed and theoretical statistics is within $\pm 5\%$ for all observed p.

6. Conclusions

This thesis develops a change–point perspective on binary splitting of datasets and demonstrated that likelihood–ratio statistics can be used to regularize regression trees. Treating each split as a change–point problem, we use a likelihood ratio test to determine if the change is significant, under the null that there is no change. In the one–dimensional case we find that when Y is in the domain of a Gaussian distribution, Donsker's invariance principle maps the partial sum process, given by $\mathcal{T}_n$, to a Brownian bridge. The location of the maximizer r^* converges by the arcsine law to a Beta$(\frac{1}{2}, \frac{1}{2})$-distribution. For the asymptotic distribution of the statistic itself, we used a result from [15] to recover a standard Gumbel limit for $\widetilde{\mathcal{T}}_n = (\mathcal{T}_n - b_n)/a_n$ under H_0. Simulations suggest that $\widetilde{\mathcal{T}}^*_n = k_\sigma(n)(\mathcal{T}^2_n - b^*_n)/a^*_n + k_\mu(n)$ is a much better fit. The constants used are

$$\begin{aligned} \lim_{n\to\infty} a^*_n &= 2 \\ b^*_n &= 2\log^{(2)} n + \log^{(3)} n + \log\left(\frac{4}{\pi}\right) \\ k_\mu(5000) &= 0.104 \\ k_\sigma(5000) &= \frac{1}{1.04}\,. \end{aligned}$$

From there one can use critical values of a standard Gumbel to determine whether the candidate split is statistically significant. When the observations have infinite variance, a stable–bridge limit replaces the classic Donsker theorem. While a closed–form result for the maximizer has not been discovered, simulations show that the asymptotic distribution can be approximated by a Beta(γ, γ) where $\gamma \in [\frac{1}{2}, 1]$. The parameter γ grows continuously from $\frac{1}{2}$ (arcsine) in the finite variance case, to 1 (uniform) as the tails become thicker.

Beyond the univariate case, we investigate the pairwise covariances of single–feature statistics. It becomes clear that the amount of overlapping features $s = |M \cap B|$ determines the pairwise covariance. If features have dependence, one can plug in a weight factor for each set combination since larger overlaps are more likely. When dimensionality is high ($p \to \infty$) one finds that the distribution of the change points reduces the number of extreme splits, but overall has a smaller impact than n. Finding an analytical formula for how the asymptotic distribution behaves as $p \to \infty$

could be a natural extension to this thesis. The distribution of $\widetilde{\mathcal{T}}_{n,p}$ remains in the family of Gumbel distributions, but the mean grows $\propto \log p$ with $p \to \infty$. Thus, a standardization $c_p \approx 1,\ d_p \approx \log p$ is necessary. Simulations for $1 \leq p \leq 6$ show that $c_p < 1$ and $d_p > \log p$. Finding improved dimensionality constants $c_p,\ d_p$ could be further explored in another paper.
Our analysis works exclusively with i.i.d. variables within nodes and focuses on shifts in the mean. One could extend this thesis to handle multivariate responses $Y_{(1)}, ..., Y_{(d)}$, different underlying distributions, or dependent relationships between features. Theory for stable bridges in multiple dimensions remains an open topic. Finally, as the motivation for this thesis lies primarily with applying statistical tests and change point detection for regularizing regression trees, this would be a natural next step to take.

Appendix A. Auxiliary results

We need the following lemma; see for example parts *b*) and *d*) of Theorem in [14].

Lemma A.1 (Moments of quadratic forms)**.** *Let* $\mathbf{z} = (Z_1, \dots, Z_n)^\top$ *be a random vector with i.i.d. entries, with* $\mathbb{E}[Z_1] = 0, \mathbb{E}[Z_1^2] = \nu_2, \mathbb{E}[Z_1^4] = \nu_4 < \infty$*, and let* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *be real and symmetric* $n \times n$ *nonrandom matrices. Then*

$$\mathbb{E}[\mathbf{z}^\top \mathbf{A}\mathbf{z} \cdot \mathbf{z}^\top \mathbf{B}\mathbf{z}] = \operatorname{tr}(\mathbf{A}) \operatorname{tr}(\mathbf{B}) + 2 \operatorname{tr}(\mathbf{A}\mathbf{B}) + (\nu_4 - 3) \operatorname{tr}(\mathbf{A} \circ \mathbf{B}),$$

where $\circ$ *denotes the Hadamard product. As a special case we get the variance*

$$\operatorname{Var}(\mathbf{z}^\top \mathbf{A}\mathbf{z}) = 2 \operatorname{tr}(\mathbf{A}^2) + (\nu_4 - 3) \operatorname{tr}(\mathbf{A} \circ \mathbf{A}).$$

If additionally $\mathbb{E}[Z_1^6] = \nu_6 < \infty$*, one has*

$$\begin{aligned}
\mathbb{E}[\mathbf{z}^\top \mathbf{A}\mathbf{z} \cdot \mathbf{z}^\top \mathbf{B}\mathbf{z} \cdot \mathbf{z}^\top \mathbf{C}\mathbf{z}] &= \operatorname{tr}\mathbf{A} \operatorname{tr}\mathbf{B} \operatorname{tr}\mathbf{C} + 2\left(\operatorname{tr}\mathbf{A} \cdot \operatorname{tr}(\mathbf{B}\mathbf{C}) + \operatorname{tr}\mathbf{B} \cdot \operatorname{tr}(\mathbf{A}\mathbf{C}) + \operatorname{tr}\mathbf{C} \cdot \operatorname{tr}(\mathbf{A}\mathbf{B})\right) \\
&+ (\nu_4 - 3)\left(\operatorname{tr}\mathbf{A} \cdot \operatorname{tr}(\mathbf{B} \circ \mathbf{C}) + \operatorname{tr}\mathbf{B} \cdot \operatorname{tr}(\mathbf{A} \circ \mathbf{C}) + \operatorname{tr}\mathbf{C} \cdot \operatorname{tr}(\mathbf{A} \circ \mathbf{B})\right) \\
&+ 4(\nu_4 - 3)\left(\operatorname{tr}(\mathbf{A} \cdot (\mathbf{B} \circ \mathbf{C})) + \operatorname{tr}(\mathbf{B} \cdot (\mathbf{A} \circ \mathbf{C})) + \operatorname{tr}(\mathbf{C} \cdot (\mathbf{A} \circ \mathbf{B}))\right) \\
&+ (\nu_6 - 15\nu_4 + 30) \operatorname{tr}(\mathbf{A} \circ \mathbf{B} \circ \mathbf{C}) + 8 \operatorname{tr}(\mathbf{A}\mathbf{B}\mathbf{C}).
\end{aligned}$$

Particularly,

$$\mathbb{E}[(\mathbf{z}^\top \mathbf{A}\mathbf{z} - \mathbb{E}[\mathbf{z}^\top \mathbf{A}\mathbf{z}])^3] = 8 \operatorname{tr}(\mathbf{A}^3) + 12(\nu_4 - 3) \operatorname{tr}(\mathbf{A} \circ \mathbf{A}^2) + (\nu_6 - 15\nu_4 + 30) \operatorname{tr}(\mathbf{A} \circ \mathbf{A} \circ \mathbf{A}).$$

Appendix B. Solving constants for standard Gumbel distribution

The test statistic $\mathcal{T}_n$ converges to a standard Gumbel distribution when there are standardizing constants a_n and b_n. We call this standardized statistic $\tilde{\mathcal{T}}_n$, where

$$\begin{aligned}
\tilde{\mathcal{T}}_n &= \frac{\mathcal{T}_n - b_n}{a_n} \xrightarrow{n\to\infty} \Lambda(x) = \exp(-e^{-x}) \\
a_n &= [2 \log^{(2)} n]^{-\frac{1}{2}} \\
b_n &= \frac{1}{a_n} + \frac{1}{2} a_n \log^{(3)} n + a_n \log(2\pi^{-\frac{1}{2}}).
\end{aligned}$$

Here $\log^{(k)} n$ is the k-th iterative logarithm of n.

To derive these constants one sets

$$\frac{\mathcal{T}_n - \tilde{b}_n}{\tilde{a}_n} = x$$

and

$$\frac{\mathcal{T}_n - b_n}{a_n} = y\,,$$

which means

$$\begin{aligned}
&\exp(-2\pi^{-1/2}e^{-x}) = \exp(-e^{-y}) \\
\iff &2\pi^{-1/2}e^{-x} = e^{-y} \\
\iff &\log(2\pi^{-1/2}) - x = -y \\
\iff &y = x - \log(2\pi^{-1/2}) \\
\iff &\frac{\mathcal{T}_n - b_n}{a_n} = \frac{\mathcal{T}_n - \tilde{b}_n}{\tilde{a}_n} - \log(2\pi^{-1/2}) \\
\iff &\mathcal{T}_n - b_n = a_n\left[\frac{\mathcal{T}_n - \tilde{b}_n}{\tilde{a}_n} - \log(2\pi^{-1/2})\right].
\end{aligned}$$

In terms of x and constants it follows that

$$\begin{aligned}
T_n &= \tilde{a}_n x + \tilde{b}_n \\
T_n &= a_n y + b_n = a_n x - a_n \log(2\pi^{-1/2}) + b_n\,.
\end{aligned}$$

So we have

$$\tilde{a}_n x + \tilde{b}_n = a_n x - a_n \log(2\pi^{-1/2}) + b_n\,.$$

Since the x-terms satisfy $\tilde{a}_n x = a_n x$, it holds that $a = \tilde{a}$. We then solve for b_n, where the remaining terms are

$$b_n = \tilde{b}_n + a_n \log(2\pi^{-1/2})$$

and since we know

$$\begin{aligned}
\tilde{a}_n &= (2\log\log n)^{-1/2} \\
\tilde{b}_n &= \frac{1}{\tilde{a}_n} + \frac{1}{2}\tilde{a}_n \log^{(3)} n
\end{aligned}$$

we find that

$$b_n = \frac{1}{a_n} + \frac{1}{2}a_n \log^{(3)} n + a_n \log(2\pi^{-1/2})$$

Remark B.1. It is interesting to compare with the maximum of iid standard normal variables Z_i. In this case, the norming constants can be chosen as

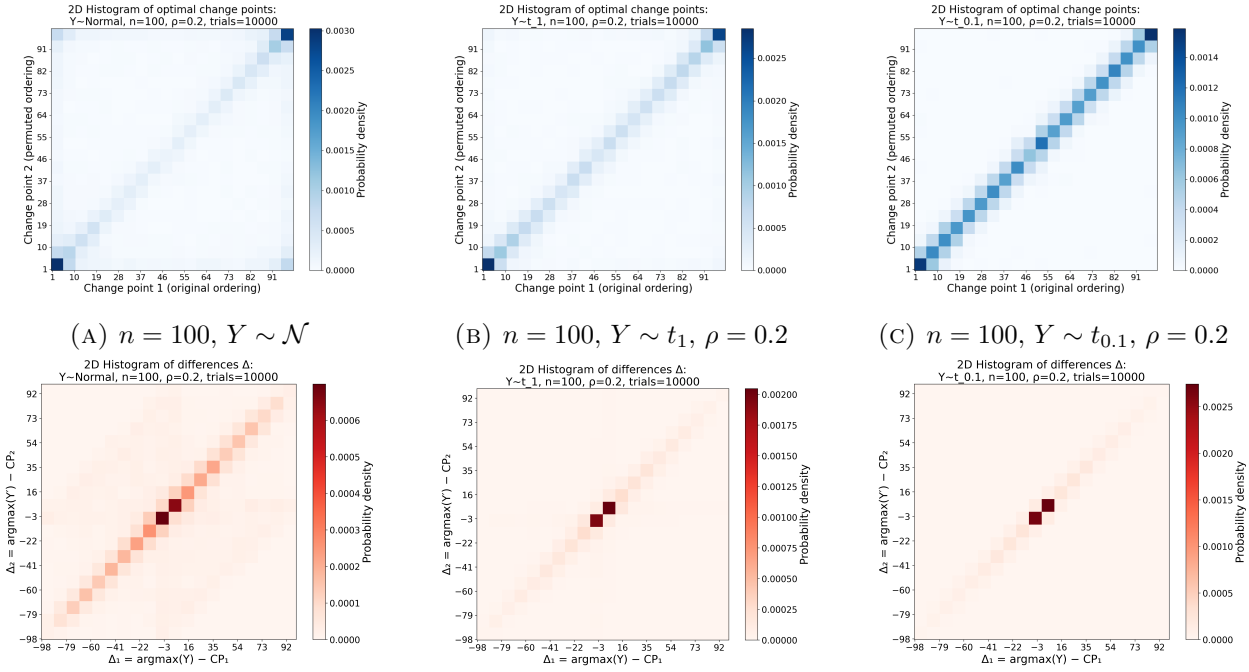
$$\tilde{a}_n = \frac{1}{\sqrt{2\log n}} \quad \text{and} \quad \tilde{b}_n = \sqrt{2\log n} - \frac{\log\log n + \log(4\pi)}{2\sqrt{2\log n}}\,.$$

and $\frac{\max(Z_1,\ldots,Z_n)-\tilde{b}_n}{\tilde{a}_n}$ converges in distribution to standard Gumbel Λ.

We notice that $\tilde{a}_{\log n} = a_n$ and $\tilde{b}_{\log n} \sim b_n$. Even the two leading order terms of $b_{\log n}$ and $\tilde{b}_n$ coincide. The third order terms are slightly different.

Appendix C. Multidimensional asymptotic distribution with dependent features

When the features $X^{(1)}, X^{(2)}$ are dependent, the joint distribution can no longer be modeled as the product of two independent arcsin-distributed r.v.'s, instead the joint distribution behaves increasingly similar to the one-dimensional case, as the location of outliers are more likely to occur at a similar position. Figure 17 visualizes the asymptotic joint distribution of the change points, and how much impact the largest outlier has for each underlying distribution. The setup for figure 17 is such that observations $Y_1, ..., Y_n \overset{i.i.d.}{\sim} \mathcal{N}$ from feature 1 are permuted such that, with probability ρ, the observation $Y_k,\ k \in \{1, ..., n\}$ is fixed to the same location as in $X^{(1)}$. So $\mathbb{P}\left(Y_k^{(1)} = Y_k^{(2)}\right) = \rho$. The remaining non-fixed observations are permuted randomly.



(A) $n = 100$, $Y \sim \mathcal{N}$ (B) $n = 100$, $Y \sim t_1$, $\rho = 0.2$ (C) $n = 100$, $Y \sim t_{0.1}$, $\rho = 0.2$

(D) $n = 100$, $Y \sim \mathcal{N}$, $\rho = 0.2$ (E) $n = 100$, $Y \sim t_1$, $\rho = 0.2$ (F) $n = 100$, $Y \sim t_{0.1}$, $\rho = 0.2$

Figure 17. (a-c) $2d$ histogram of optimal change point location. (d-f) $\hat{r} - r$ with dependent features. I.i.d outputs $Y_1, ..., Y_n$ were simulated and optimal output $r^{(1)}$ was found, then Y's were permuted with a correlation factor $\rho = 0.2$ such that each observation has probability ρ of being fixed at the same rank order as the original. The remaining $1 - \rho$ of observations are treated as independent.

References

[1] Billingsley, P. *Convergence of probability measures*, second ed. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1999. A Wiley-Interscience Publication.

[2] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984.

[3] Csörgő, M., Szyszkowicz, B., and Wang, Q. Donsker's theorem for self-normalized partial sums processes. *The Annals of Probability 31*, 3 (2003), 1228–1240.

[4] Dierickx, G., and Einmahl, U. A general darling–erdős theorem in euclidean space. *Journal of Theoretical Probability 31*, 2 (Dec. 2016), 1142–1165.

[5] Doney, R. A. On the maxima of random walks and stable processes and the arc-sine law. *Bulletin of the London Mathematical Society 19*, 2 (03 1987), 177–182.

[6] Embrechts, P., Klüppelberg, C., and Mikosch, T. *Modelling Extremal Events for Insurance and Finance*, vol. 33 of *Applications of Mathematics (New York)*. Springer, Berlin, 1997.

[7] Engler, N., Lindholm, M., Lindskog, F., and Nazar, T. Regularisation of cart trees by summation of p-values, 2025.

[8] Gösmann, J., Stoehr, C., Heiny, J., and Dette, H. Sequential change point detection in high dimensional time series. *Electron. J. Stat. 16*, 1 (2022), 3608–3671.

[9] Horvath, L. The maximum likelihood method for testing changes in the parameters of normal observations. *The Annals of Statistics 21*, 2 (1993), 671–680.

[10] Horváth, L., Kokoszka, P., and Steinebach, J. Testing for changes in multivariate dependent observations with an application to temperature changes. *Journal of Multivariate Analysis 68*, 1 (1999), 96–119.

[11] Philipp, W. Weak and L^p-Invariance Principles for Sums of B-Valued Random Variables. *The Annals of Probability 8*, 1 (1980), 68 – 82.

[12] Revuz, D., and Yor, M. *Continuous martingales and Brownian motion*, vol. 293. Springer Science & Business Media, 2013.

[13] van der Vaart, A., and Wellner, J. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer International Publishing, 2023.

[14] Wiens, D. P. On moments of quadratic forms in non-spherically distributed variables. *Statistics 23*, 3 (1992), 265–270.

[15] Yao, Y.-C., and Davis, R. A. The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates. *The Indian Journal of Statistics 48*, 3 (1986), 339–353.

Department of Mathematics, Stockholm University, Albano hus 1, 10691 Stockholm, Sweden
Email address: mabr3613@student.su.se