École nationale
de la statistique
et de l'analyse
de l'information

École Nationale Supérieure de Statistique et d'Analyse de l'Information - Rennes

Signature Theory in Oncology

Élèves :
Borrelli Claire
Debeve Merlin
Huang Ting
Montzamir Antoine
Lemoine Paulin

Encadrant :
Chrétien Stéphane,
Laboratoire ERIC,
Université Lyon 2, France

Année Universitaire 2024-2025

Signature Theory applied to the identification of malign tumor in Oncology

March 31, 2025

Abstract. The emergence of the Signature theory associated to Terry Lyons based on the formalism introduced by Kuo-Tsai Chen provides a new way to compute signals. It has been successfully applied in finance, cognitive sciences and spectrography to reconstruct or predict. This paper will tackle a classification problematic in oncology with a method to identify abnormal level of Circulating Tumor DNA (ctDNA) associated to the early stage of development of malign tumors. The nature of the problematic requires a classifier with the lowest False Discovery Rate (FDR) and an adequate model to train on. The paper shows that a classifier applied to signature has great results which can be further refined by controlling the FDR. This paper was written while keeping in sight possible practical applications.

Contents

Acknowledgments. We wanted to deeply thanks our tutor, Stéphane Chrétien, for his commitment to this project and his numerous explanations on the statistical concepts and ideas. He had a great impact on the way we understood the signature theory and provided us with constant articles and thesis on the matter. This article is based on analysis and models made by his PhD student, Rémi Vaucher, who we also wanted to thanks for his amazing work.

1 Introduction

"*In 2022, there were an estimated 20 million new cancer cases and 9.7 million deaths. The estimated number of people who were alive within 5 years following a cancer diagnosis was 53.5 million. About 1 in 5 people develop cancer in their lifetime, approximately 1 in 9 men and 1 in 12 women die from the disease.*" according to WHO in the alarming article "*Global cancer burden growing, amidst mounting need for services*". With the population aging and growth, the projected cancer burden is set to increase by 77% (over 35 millions new cancer cases) in 2050. Cancer is the main cause of death worldwide accouting for 1 in 6 deaths. However, let's not lose hope as some cancers can be cured and treatment options have expanded significantly in the recent years. The main factor is the ability to detect cancers at early stage of development. Cancer should be considered as an abnormal cell growth in the human body which can spread to other parts of the said body in opposition to fixated benign tumors.

Early detection allows for a more efficient treatment and patients are more likely to be cured than patients with late-stage cancer. For example, the 5-year survival rate of patients with lung cancer who are diagnosed at a localized stage is 57%, while for those diagnosed with distant metastases, it is only 5% [0]. Despite the high mortality and benefits of early detection, only 16% of lung cancers are diagnosed at a localized stage. Testing and screening has became a public problem and is now part of the agenda. Many healthcare systems cover at least partially the costs associated to the cure.

Early detection is also a way to reduce government spending while achieving a higher life expectancy. Nonetheless testing and screening are infamous for there invasive methods, thus discouraging population who already had difficulties accessing those tests. Many equalities arose when considering cancer. A simple, costless, non-invasive and efficient method of early detection would address a lot of this issues.

Recent progresses have casted a new light on circulating tumor DNA (ctDNA) as a biomarker for cancerous tumors. The collection method of ctDNA does also comply with the said issues. Despite a limited understanding of the generation process underlying ctDNA, many correlations between tumor sizes and rates of ctDNA in the blood led to studies looking for groundbreaking discoveries. This paper explores the application of ctDNA to the detection of malign tumors and extends the thesis work of Rémi Vaucher. This paper does not contain nor is based on any patient data to preserve anonymity.

At first, the reader will be provided insight on ctDNA and the biological generation process associated. It will lead to a model used to simulate patient data and methods to handle missing and incomplete data. Synthetic data will be analyzed to determine specific criteria to distinguish between benign and malignant patients. This analysis is based on the Signature theory developed by Terry Lyons and his team. A classifier based on the Levy Area will be implemented with controlled false discovery rate using the Benjamini-Hochberg method. Finally, alternative classifiers will be applied to the extracted signatures.

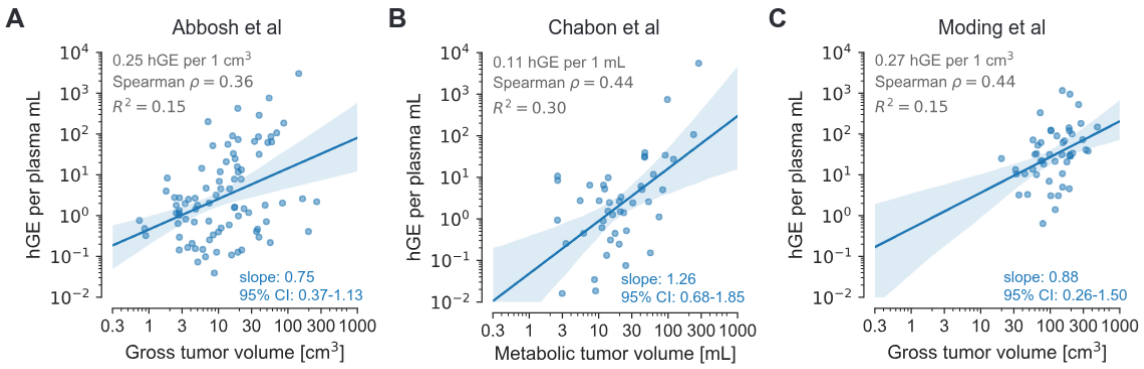
2 A model for ctDNA in patient's blood

2.1 Using ctDNA as a biomarker

2.1.1 What is ctDNA?

The bloodstream contains a great diversity of biological objects dispatched in the plasma and the blood cells. Among the plasma are proteins, hormones, DNA ... This DNA is called cell-free DNA (cfDNA) which is also a generic name to regroup different types of DNA. CtDNA stands for Circulating tumor DNA, a specific type of fragmented DNA found in the plasma. It is part of the cfDNA circulating in the bloodstream. cfDNA can be seen as a biomarker for various diseases like cancer [2], sepsis [3] but also myocardial infarction [4] and transplant graft rejection [5]... CtDNA is cfDNA that shares a part of the genome of tumors (not the entire genome). It allows for sequencing methods and addresses cancer classification problems. The interested reader can read about EPIC-seq and its application for lung cancer detection [6]. An increase in ctDNA concentration in blood samples is telltale of increasing tumor sizes in cancer patients [7].

Figure 1: **ctDNA increase in blood linked to tumor size**



This figure (borrowed from [10]) depicts the correlation between ctDNA and tumor size using linear regression in stage I to III in lung cancer. The original description can be found in annex [a-0].

However the precise process associated to the production of ctDNA remains unknown. CtDNA comes from the tumor or circulating tumor cells (CTC - tumor cells in the blood responsible for the propagation of the tumor). Multiples studies (see [**7**] for example) make the realistic hypothesis that apoptosis and necrosis of cells are responsible for ctDNA in the bloodstream. It causes the cytoplasmic DNA of, let's say CTC, to cross the cell's membrane. This cytoplasmic DNA is a biomarker in aging and disease [8].

2.1.2 Difficulties

Many difficulties arise when considering ctDNA as a cancer biomarker. One of them is distinguishing ctDNA can be distinguished cfDNA. Non-malignant cells during normal cellular turnover, but also during medical procedures releases cfDNA.

Another difficulty concerns the qualities and usefulness of ctDNA to determine the presence, type and stage of cancer. It is still actively studied and questioned. As of 2017, a study demonstrates that ctDNA burden does not always correlate with traditional cancer staging[9].

Finally ctDNA is a non-lasting quantity. It depletes naturally due to the body's clearance mechanism. One must think of the ability of the Liver to break down the ctDNA while the Kidneys will take out small chunks of this DNA. Specific enzymes are also able to break down ctDNA. Phagocytosis also acts by digesting cell debris including ctDNA. The half-life of ctDNA is typically 30 minutes to a few hours.

2.1.3 How is ctDNA monitored?

The intensive research around cfDNA has led to an innovative method of biopsy called liquid biopsy. Instead of extracting patient's tissues or cells to detect a disease, a liquid biopsy demands a simple, non-invasive blood sample.

The ctDNA can be monitored using the plasma contained in this blood sample. Moreover it has been shown than plasma is a better source of ctDNA than serum in lung tumor patients [1]. Such a method of blood collection minimizes the patient's discomfort and the probability of patient's defect in the program. The blood is processed twice by centrifugation to extract the plasma and to clean it from debris. The method is easy and cheap to implement. It doesn't require extensive formation (a biopsy is typically done by a surgeon) and can be performed everywhere as long as the blood is processed within 2 to 4 hours.

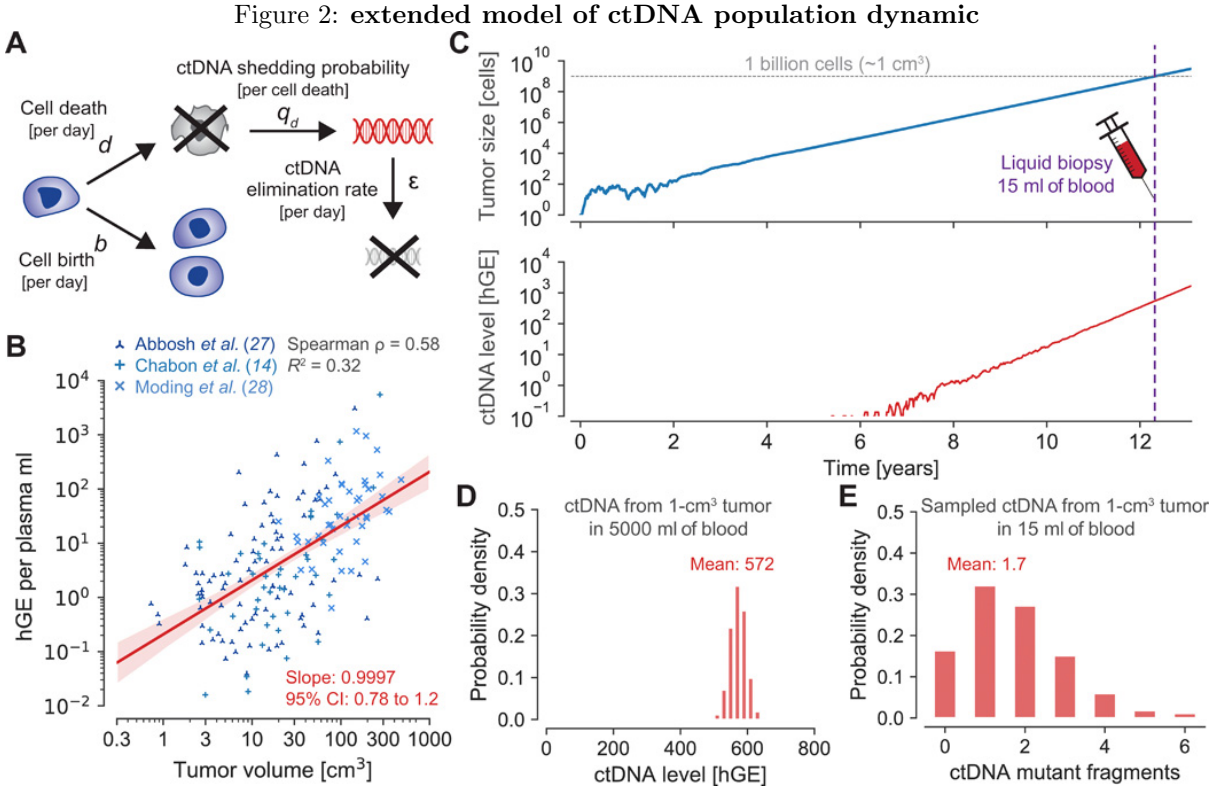
This study will not explore the sequencing methods at the heart of the analysis of ctDNA. The indicator at stake here is the amount of ctDNA in the patient blood and its modelization.

2.1.4 Toward a model?

A good understanding of the mechanism underlying the rates and trends of ctDNA in a patient blood allows for a pertinent model. CtDNA evolution is governed by a birth rate and a death rate. Both mechanisms have been theoretically explained but also empirically observed. According to a model [10], colorectal cancer cells divide with a birth rate of b = 0.25 (e.g., colorectal cancer cells). Lung cancer cells approximately divide with a birth rate of b = 0.14 per day (and a death rate of d = 0.136) [11]. Breast cancer cells have b = 0.1 .

A shedding rate and a ctDNA elimination rate could also be added such that "approximately 0.014% of a cancer cell's genome is shed into the bloodstream after it undergoes apoptosis. For a ctDNA half-life time of t1/2 = 30 min (16), we calculated a ctDNA elimination rate of $\varepsilon \approx 33$ per day.". The present paper chooses not to include this parameters.

Figure 2: **extended model of ctDNA population dynamic**



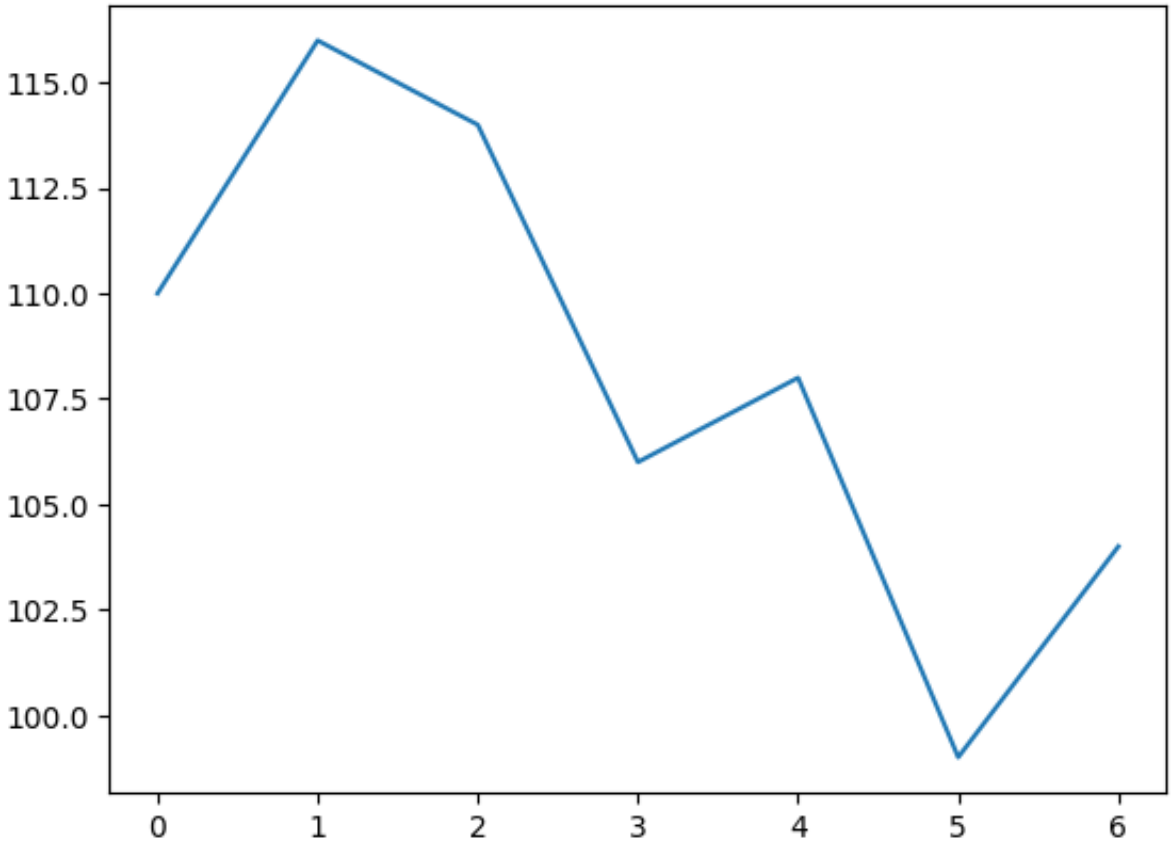
*This figure (borrowed from [**10**]) depicts the biological process at hands (A), the positive correlation between tumor size and ctDNA in plasma samples (B and C) and the estimated levels of ctDNA for a 1 cm3 tumor in 15 mL and 5L of blood (D and E). The original description can be found in annex [a-1].*

In conclusion, ctDNA seems to be a pertinent biomarker for cancer detection prone to be modelized using birth and death rates. Moreover, ctDNA in human Genome Equivalent (hGE) can be seen as a population that will grow when a malign tumor is present. While the exact birth and death mechanism remain unknown, a birth rate and a death rate can still be observed. This insight is the justification for the following model.

2.2 The generation process of the dataset

The nature of the data at hand is specific to a patient, thus raising the issue of anonymity. The Institut Pasteur cannot give public access to the levels of ctDNA it has monitored due to regulations and ethics [12]. To overcome the issue of anonymity, the creation of a model is needed to generate artificial data that closely resemble clinical data. At the core of this model is a birth-and-death process based on empirical parameters found in the literature. The following model is inspired by Rémi Vaucher thesis. The main idea is to generate ctDNA values in a fictitious patient's blood.

Figure 3: **A ctDNA trajectory by hGE through time**



This figure depicts a fictitious trajectory of ctDNA quantity in hGE per plasma in a patient's blood.

As stated previously, the amount of ctDNA in the blood depends on a birth rate and a death rate as ctDNA appears and depletes in the patient's blood. A continuous birth-and-death process (BDP) can be fitted to match this ctDNA levels. Moreover, the process to generate malign data can be decomposed into a benign dynamic, an unknown transition time and a malign dynamic. It means that at transition time, the first occurrence of a tumor is recorded through the level of the biomarker. Finally, the BDP will be discretized in 7 values corresponding to 7 blood samples at regular time intervals.

2.2.1 What is a Birth and Death Process?

A Birth and Death Process (BDP) is a type of Continuous Markov Chain. The transition probability of a Continuous Markov Chain is given by:

$$P_{ij}(t) = P(X(t+u) = j \mid X(u) = i)$$

where $t > 0$ and $u > 0$ are independent.

In a BDP, the transitions from state i are restricted to $\{i+1, i-1\}$, representing a birth or a death. The birth rate λ_i and the death rate μ_i vary according to state i. The transition probabilities are as follows: the process goes from i to $i+1$ with probability $\frac{\lambda_i}{\lambda_i+\mu_i}$, and from i to $i-1$ with probability $\frac{\mu_i}{\lambda_i+\mu_i}$.

Adding or removing from the population is decided by considering two exponentially distributed random variables: $B(i)$, the time until the next birth, and $D(i)$, the time until the next death, at state i. The distribution of $\min(B(i), D(i))$ is exponential with parameter $\lambda_i + \mu_i$, assuming independence. The process goes from i to $i+1$ if $B(i) < D(i)$, with the following probability:

$$P[B(i) < D(i)] = \frac{\lambda_i}{\lambda_i + \mu_i}.$$

Now, let's rewrite the transition probability of the Continuous Markov Chain for the case of a birth in a BDP:

$$P_{i,i+1}(h) = P(X(t+h) = i+1 \mid X(t) = i) = P(X(t+h) - X(t) = 1 \mid X(t) = i) \quad \text{with} \quad h > 0.$$

$X(t+h) - X(t)$ represents what happened between t and $t+h$, or simply during the time period h. This follows a Poisson distribution, such that:

$$P(X(t+h) - X(t) = k) = \frac{(\lambda h)^k \exp(-\lambda h)}{k!}$$

where λh is the birth rate intensity per time period. The transition probability for a birth can be rewritten as:

$$P(X(t+h) - X(t) = 1) = \frac{(\lambda h)\exp(-\lambda h)}{1!} = (\lambda h)\sum_{n=0}^{\infty}\frac{(-\lambda h)^n}{n!} = (\lambda h)\left(1 - \lambda h + \frac{1}{2}(\lambda h)^2 - \cdots\right) = \lambda h + o(h)$$

where $o(h)$ represents the remainder term.

This expression also applies to the death probability over a time period h. Finally, the transition probability from i to $i+1$ can be written as:

$$P_{i,i+1}(h) = P(X(t+h) - X(t) = 1 \mid X(t) = i) = \frac{(\lambda_i h)^1 \exp(-\lambda_i h)}{1!} \cdot \frac{(\mu_i h)^0 \exp(-\mu_i h)}{0!} + o(h)$$

$$= (\lambda_i h)\sum_{n=0}^{\infty}\frac{(-h(\lambda_i + \mu_i))^n}{n!} = (\lambda_i h)\left(1 - h(\lambda_i + \mu_i) + \frac{1}{2}h^2(\lambda_i + \mu_i)^2 - \cdots\right) + o(h)$$

$$= \lambda_i h + o(h).$$

$$\text{and} \quad P_{i,i-1}(h) = P(X(t+h) - X(t) = -1 \mid X(t) = i) = \mu_i h + o(h).$$

Finally, the transition probabilities of a continuous BDP are defined as:

$$p_{ij}(h) = \begin{cases} \lambda_i h + o(h), & \text{if } j = i+1, \\ \mu_i h + o(h), & \text{if } j = i-1, \\ 1 - h(\lambda_i + \mu_i) + o(h), & \text{if } j = i, \\ o(h), & \text{otherwise.} \end{cases}$$

The Birth and Death Process can be used as a powerful model to represent population dynamics. The interested reader will find more details in this lecture [13].

2.2.2 Using a tailored BDP to generate benign and malign data

The model can be refined by using a specific set of parameters. Let's take $\lambda_i(\theta)$ for the birth rate and $\mu_i(\theta)$ for the death rate such that,

$$\lambda_i(\theta) = \gamma i + \alpha \quad \text{and} \quad \mu_i(\theta) = \nu i.$$

where γ and ν are fixed rates and α is a migration constant symbolizing an additional influx in the population. Notice how both rates are multiplied by i, the state of the process or the number of individuals in the population at a given time. When the population grows rapidly, both the birth and death rates will also increase. This can cause the population to grow or deplete at linear speeds. Due to the migration constant α, this model is referred to as a linear-migration BDP. The migration rate can be used to add a constant influx of population.

One may wonder why the migration constant is relevant here. It was decided to fix $\gamma = 0$ and $\alpha = z_0\nu$, where z_0 is the initial size of the population, and ν is the death rate. This implies that:

$$\lambda_i(\theta) = \nu z_0 \quad \text{and} \quad \mu_i(\theta) = \nu i.$$

This ensures a constant influx of population based on the initial population while maintaining a dynamic death rate that oscillates around z_0. The model is also called an "M/M/inf" BDP, where the birth and death rates are equal. This oscillation is specific to benign data, as no mechanism (such as Circulating Tumor Cells (CTC)) produces more ctDNA during the process. In practice, the death rate is a value in the set $\{0.1, 0.12, 0.14, 0.17\}$. This set corresponds to the observed data mentioned earlier. The population starting point, z_0, is a random value generated by a Poisson distribution centered around 100. For each death rate, n trajectories were generated using the Galton-Watson approximation method and sampled 7 times at regular intervals, yielding a total of $28n$ observations in the benign dataset.

Algorithm 1: Pseudo-code for Generating Benign Data

Input: $n = 1000$ (number of trajectories), $b = 0$ (birth rate), d (death rate), jump_times (times at which jumps occur),
Output: Object containing all generated trajectories

Function `generate_benign_data`$(n, b, jump\_times, d)$**:**
  Initialize an empty matrix *data* of size $[8, n]$
  foreach *d in set of* $[np.log(2)/i$ *for i in range*$(4, 8)]$ **do**
    for $i = 0$ **to** $n - 1$ **do**
      Sample $B0 \sim Poisson(\mu = 100)$
      Compute $\alpha = B0 \cdot d$
      `/* Simulate with BDP */`
      biom_traj ← `birdepy.simulate.discrete`([b, d, α], 'linear-migration', $z_0 = B0$, times = jump_times, method = 'gwa')
      Store the trajectory data[:, i] ← biom_traj
  return *Object containing all the trajectories in data*

A slightly different model is used to generate the malign dataset by combining a benign and a malign dynamic into a Cox process. The core idea is to add a transition time when the generation process switches to a malign dynamic. This reflects the apparition of a tumor in the body and the rising ctDNA levels associated with it. Let's assume that the transition time follows a uniform distribution in a fixed time interval in the middle of the generation process timeline. Before the transition time, the data will be generated according to a linear BDP with equal birth and death rates. After the transition time, the data will be generated according to a linear BDP with the birth rate from the empirical set $\{0.07, 0.1, 0.15, 0.2, 0.25\}$ and a lower death rate, corresponding to the death rate subtracted by values from the set $\{0.001, 0.002, 0.004, 0.007, 0.01, 0.015\}$. Cases of negative or null death rates are avoided.

Finally, each trajectory of the dataset is generated using a Poisson distribution with a mean rescaled from the initial trajectory produced by the linear BDP. The choice not to use a linear migration BDP for the same dynamic is justified. In fact, a linear BDP is a Poisson process. Therefore, the malign dataset is generated using a doubly stochastic Poisson process (also known as a Cox process [14]), meaning that the data is generated by a Poisson distribution with an intensity parameter that evolves according to the linear BDP (see annex [a-2] for the pseudo code). This model makes the ctDNA level a random variable tied to the intensity parameter of the Poisson distribution, which is the evolving population of the model. The intensity parameter allows for a more accurate representation by considering the specific time dynamics of the apparition and multiplication of ctDNA at transition time, while reducing the linear aspect of the BDP. This approach is deeply grounded in today's reality, as there is evidence of the positive correlation between ctDNA and growing tumors, but no clear understanding of the exact generation process.

Algorithm 2: Pseudo-code for Generating Malign Data

Input: b (set of birth rates), r (set of death rate adjustments), 1000 (number of trajectories), 7 (jump times)

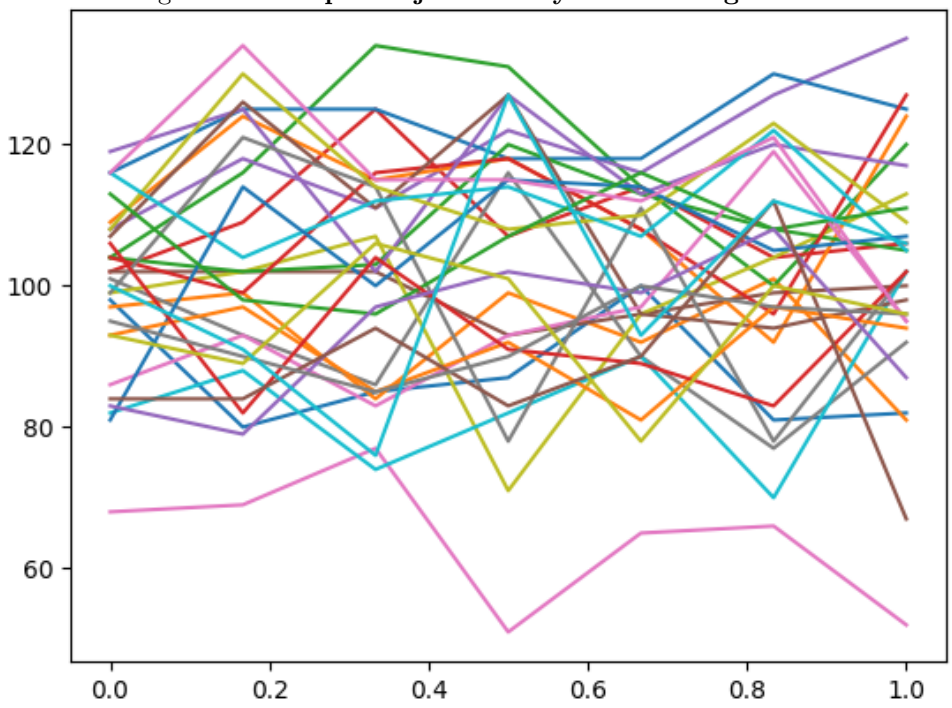
Output: Object containing all generated trajectories for malign data

Function `generate_malign_data`(b, r)**:**
  data $\leftarrow$ empty list ;
  foreach $b \in [0.07, 0.1, 0.15, 0.2, 0.25]$ **do**
    foreach $r \in [0.001, 0.002, 0.004, 0.007, 0.01, 0.015]$ **do**
      `/* Calculate adjusted death rate */`
      $d \leftarrow b - r$;
      if $d > 0$ **then**
        data_matrix $\leftarrow$ empty list ;
        for $i = 0$ **to** 999 **do**
          `/* Generate the Cox process trajectory for given` b `and` d `*/`
          biom_traj $\leftarrow$ `generate_cox_process`(b, d) ;
          data_matrix.`append`(biom_traj) ;
        data.`append`(data_matrix) ;
  return *data*

2.2.3 The dataset

The dataset is a combination of benign and malign data generated by the described processes. Both processes are based on empirical birth and death rate and should closely resemble real life data as it is observed by the Institut Pasteur.

Figure 4: **Multiple trajectories by hGE through time**



This figure depicts the various trajectories generated. Both malignant and benign trajectories are present.

Both dynamics appear to be undistinguishable at first sight. The next step is to provide a pertinent statistical analysis of the trajectories. However the resulting model hides a crucial difficulty : missing and censored data.

2.2.4 Censored Data

Our project is based on simulated data. However, since our goal is real-world application, we had to anticipate the challenges we might encounter with real data. One of the main issues we identified is that we could have to deal with censored data. Our data comes from routine blood tests performed on patients. However, a key issue is that patients may discontinue their blood tests before the end of the study period. This could happen for various reasons, such as the patient's passing or their decision to withdraw from the medical study. In this case, our data would be subject to right-censoring, meaning that while each data sample has a defined start, it may not always include observations until the end of the study period. We then had to find a way to face this issue. For that, we decided to find an imputation method that could impute the censored data based on the available non-censored data. In this way, our algorithms could work on an incomplete data sample and would still be able to determine whether a patient has cancer, even if he does not respond until the end of the study. But for that, we first needed to simulate the censoring process on the available data.

To simulate the censoring process we might encounter with real data, we developed an algorithm that systematically applies it to each available data sample. Here is how this algorithm works :

1. First, we set a proportion of censorship to apply to the data. In our case chose 30 %, as oncological studies typically encounter censorship rates within this range (generally between 20% and 40 %).
2. Then, the algorithm selects 30% of the patients in each data file (with each patient corresponding to a column).
3. Finally, for each column selected to be censored, the algorithm censors all data starting from a specific period. Censorship cannot begin at the first period, as a patient cannot be considered censored if they left the study before the follow-up started or if they never participated in the study.

Thanks to this algorithm, we obtain a realistic simulation of right-censorship for patients who leave the study and provide data only up to a certain period. Then, once our data is censored, we have to find a way to impute it.

We tested several methods to impute the missing data, including matrix completion, KNN imputation, and mean imputation. However, the method that yielded the best results was MICE imputation, which stands for Multiple Imputation by Chained Equations. It's an algorithm that works this way :

1. First, the algorithm imputes all the missing data, in our case, by using the mean of the dataset.
2. Then, the algorithm applies an imputation model to the imputed data to refine their values. This process is repeated multiple times, generating several imputed datasets. In our case, the process is repeated 10 times.
3. Finally, once the multiple imputations are created, the algorithm combine the different imputations using Rubin's Rule. These rules combine the estimates from each imputed dataset by calculating the average of the imputations for each missing value and adjusting for the variance between the imputed datasets.

In our case, the imputation model used by the MICE algorithm is linear regression. The values in the datasets appear to have a linear relationship, which was confirmed by the effective imputation results when comparing with the real datasets.

To assess the quality of the imputation, we compared the imputed data with the non-censored datasets. For this, we used the Root Mean Squared Error (RMSE) to measure the difference between the real and imputed values. The imputation for the benign datasets performed very well, with an RMSE of 3.5 across all benign imputed datasets, while the benign data values range from less than 70 to more than 1100.

This is illustrated in the graph, which shows the mean value for each period in both the imputed dataset and the real dataset it is based on.

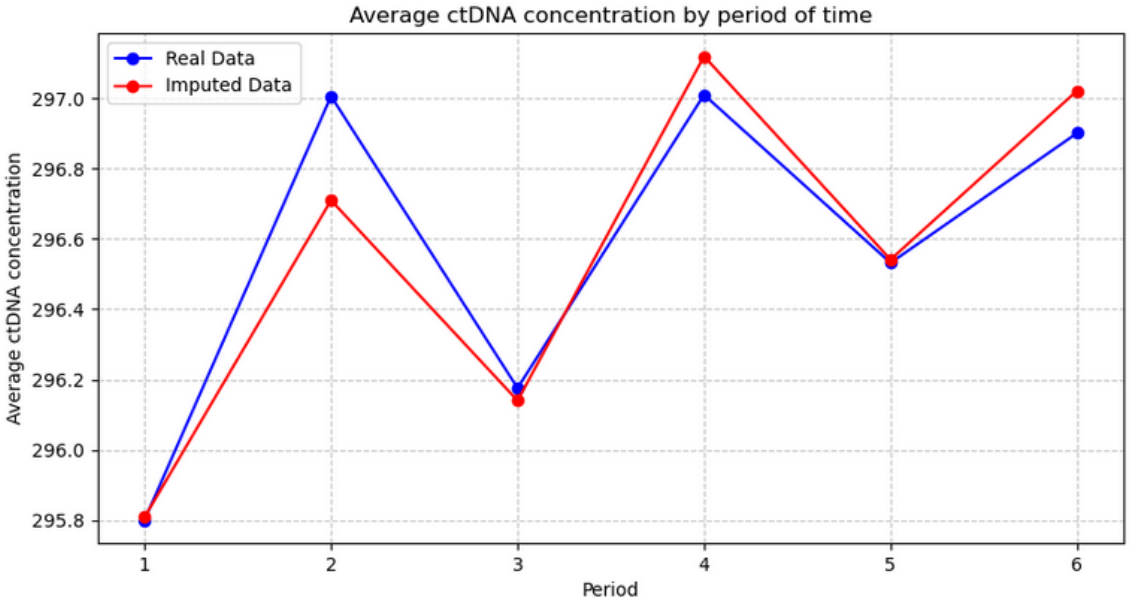


Figure 5: Difference between imputed and real benign ctDNA values by hGE

For the malign dataset, we were initially surprised to find that the RMSE across all the malign imputed datasets was 3.25e9, which seemed excessively large at first. However, upon closer inspection of the malign data, we discovered that some datasets contain extreme values up to 1e10. This indicates that the RMSE is heavily influenced by these extreme values. Despite this, the imputation is actually performing very well on the malign data, as shown in the following graph:

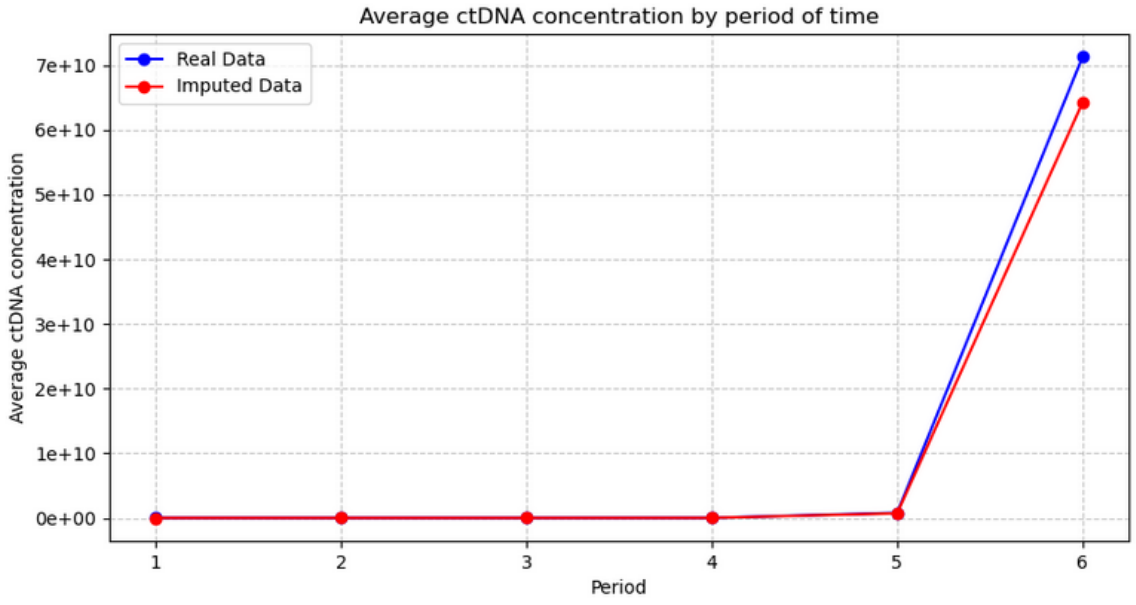


Figure 6: Difference between imputed and real malign ctDNA values by hGE

Eventually, the imputation using the MICE method yielded excellent results, enabling us to work with data that closely reflects reality.

2.3 Basic analysis isn't enough

Since the goal of our project is to distinguish between benign and malign data, we began with a basic analysis to evaluate the effectiveness of this approach. Initially, the results were very promising, as we observed a clear distinction in ctDNA levels: the mean ctDNA concentration value for all the benign data we had at our disposal was 355, while for malign data, it reached 1e10.

This appeared to be a significant difference, suggesting that distinguishing between malign and benign data could be straightforward. However, further analysis was needed to confirm this assumption.

We then examined the medians of the data, and the results were far less striking. The median ctDNA concentration for benign data was 99, while for malign data, it was 652. These much closer values suggested that extreme values in the malign data might be inflating the mean, raising the possibility that most malign data points were not drastically different from the benign ones.

We then decided to take a closer look at the data to understand the gap between the means and medians, particularly for the malign cases, and to determine whether distinguishing between benign and malign data was indeed straightforward. The following graph displays all the datasets together, with each path representing an individual dataset.

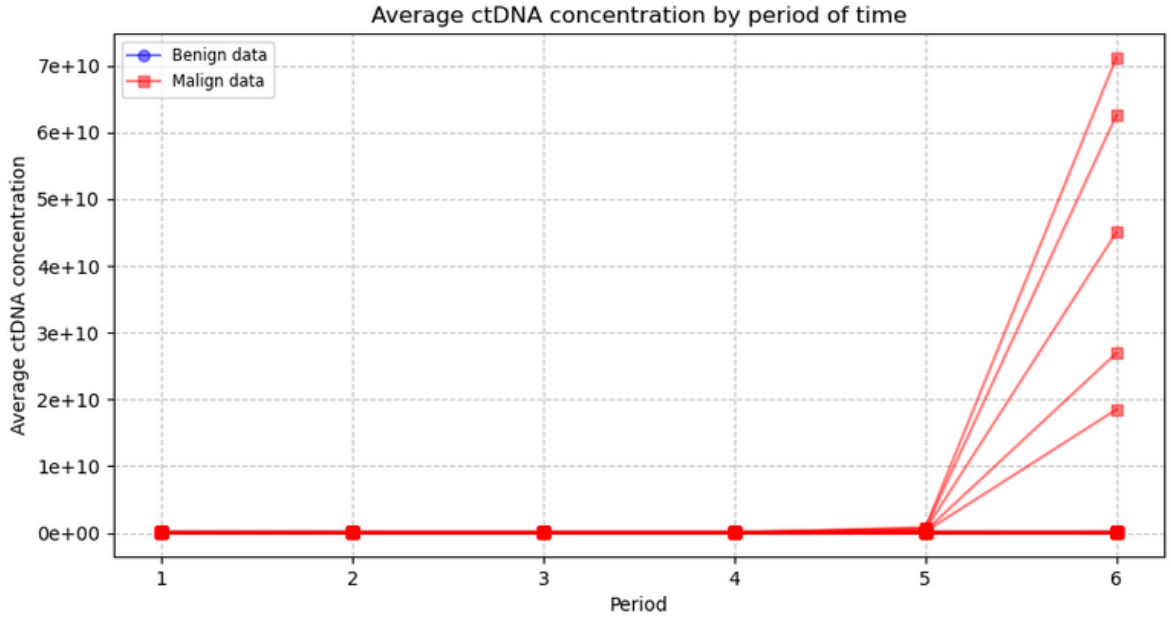


Figure 7: Means of all the ctDNA datasets by hGE through time

This graph reveals that some malign data points are indeed significantly higher than benign data, particularly in the last period, where benign data is not even visible. However, since the median is much lower than the mean, this suggests that the malign data points seen at the lower end of the graph in the last period might not be drastically different from the benign data. While certain ctDNA concentrations can clearly be classified as malign, this indicates that classifying the data in all cases may not be as straightforward.

We then decided to examine more closely whether some benign and malign trajectories could be similar. This led us to generate the following graph, which is similar to the previous one but includes only a smaller selection of dataset trajectories.

This graph shows that malign and benign data can have similar values, making it difficult to distinguish between the two, especially when ctDNA concentration values fall between 200 and 1500.

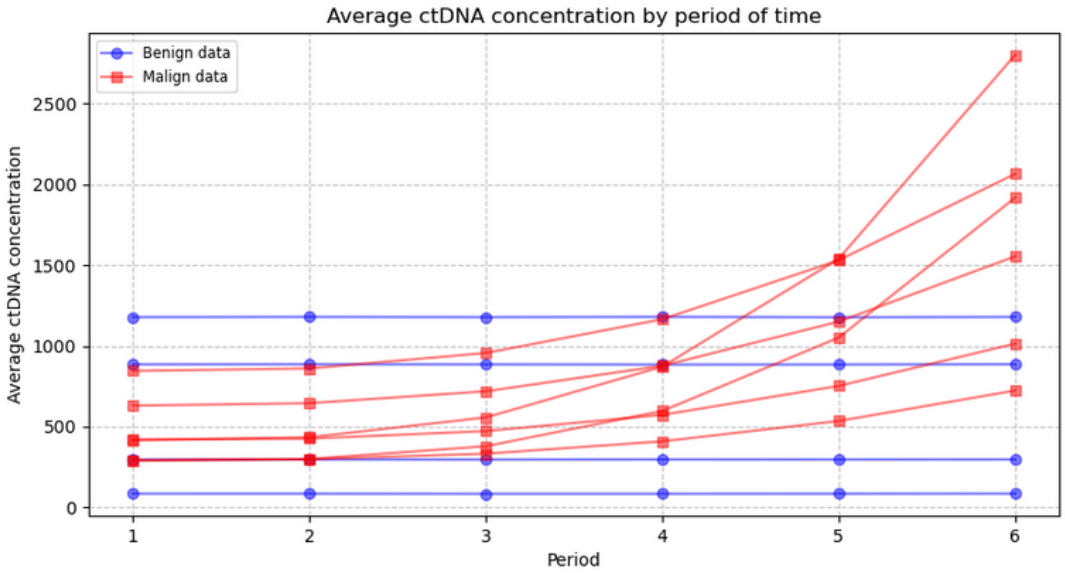


Figure 8: Means of some ctDNA datasets by hGE through time

However, the trajectory appears to be the key differentiator, as benign values seem to remain stable, while malign values exhibit exponential growth. Despite this, we were unsure if basing our distinction only on whether the values are growing was sufficient. Indeed, the values shown are averages of datasets and do not capture the variability within each trajectory, and we were then uncertain whether all benign trajectories were completely stable or if all malign ones exhibited enough growth to be significant. But as we saw that benign and malign data could have similar values, it became clear that the most reliable way to distinguish between the two was to focus on the trajectory path itself.

We then needed to find a tool that would enable us to establish a strong classification between benign and malign data based on their trajectories. And so, we decided to use signatures, as they provide valuable insights into the data by analyzing their paths.

3 What is signature theory and why do we use it?

In order to treat the dataset that we were given, we needed to understand signature theory and how it applies to our problem. Signature theory takes a geometrical approach for encoding the information contained within time series data.

3.1 What is a signature?

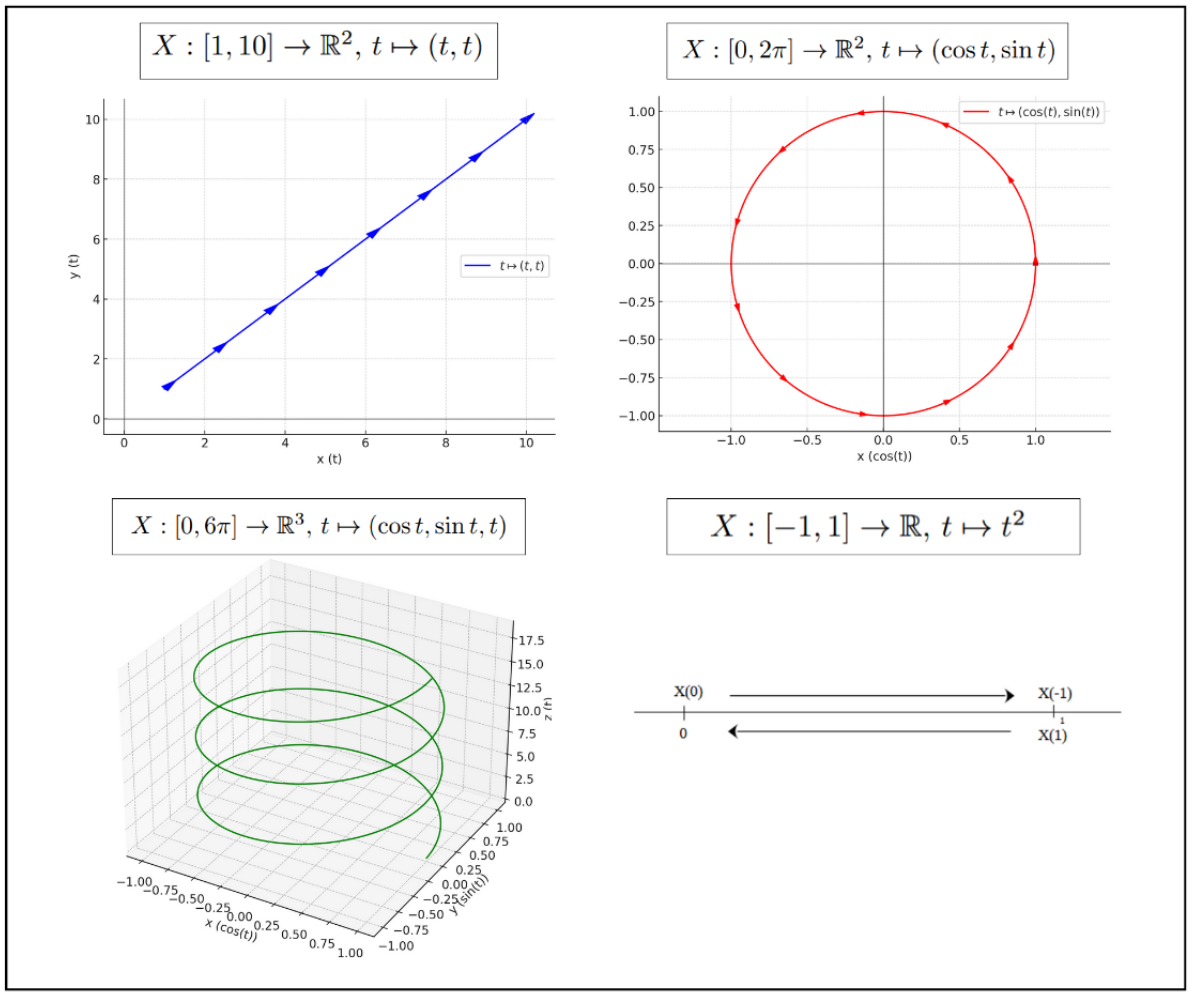
3.1.1 Parameterized Curve

A parameterized curve is a mapping such as :

$$X : [a, b] \to \mathbb{R}^d$$

The idea is to see the stochastic data as a path in an R^d space.

Examples of parameterized curves :



Goal :

Encode a path $X : [a, b] \to \mathbb{R}^d$ by a countable infinite series of numbers.

3.1.2 Signature of $X : [a, b] \to \mathbb{R}$

The signature of a path is a series of coefficients that capture the geometric characteristics of a trajectory. If d=1, that sequence is given by:

$$S_{a,b}(X) = \left(1, S^1_{a,b}(X), S^{11}_{a,b}(X), S^{111}_{a,b}(X), \dots\right)$$

It is defined recursively as follows:

$$S^1_{a,b}(X) = \int_a^b dX(t) = X(b) - X(a) \qquad S^{11}_{a,b}(X) = \int_a^b S^1_{a,b}(X)\, dX(t) \qquad S^{111}_{a,b}(X) = \int_a^b S^{11}_{a,b}(X)\, dX(t)$$

Evaluation of Integrals

$$S^{11}_{a,b}(X) = \int_a^b S^1_{a,t}\, dX(t) = \int_a^b [X(t) - X(a)]X'(t)\, dt = \left[\frac{(X(t) - X(a))^2}{2}\right]_a^b = \frac{(X(b) - X(a))^2}{2}$$

$$S^{111}_{a,b}(X) = \int_a^b S^{11}_{a,b}\, dX(t) = \int_a^b \frac{(X(t) - X(a))^2}{2} X'(t)\, dt = \left[\frac{(X(t) - X(a))^3}{6}\right]_a^b = \frac{(X(b) - X(a))^3}{6}$$

General Formula

For $n \geq 1$, the general formula is given by :

$$\boxed{S^{\overbrace{1\dots1}^{n \text{ times}}}_{a,b}(X) = \frac{(X(b) - X(a))^n}{n!}}$$

Note: Invariance of the signature under time reparameterization

Consider a simple linear path $X(t) = 2t$ on the interval $[0, 1]$. We sample the path at regular intervals:

$$X(0) = 0, \quad X(0.5) = 1, \quad X(1) = 2$$

Signature of a regularly sampled path:

$$\boxed{S_{0,1}(X) = (2, 2)}$$

$$S^1_{0,1}(X) = \int_0^1 dX(t) = X(1) - X(0) = 2 - 0 = 2 \qquad S^{11}_{0,1}(X) = \frac{(X(1) - X(0))^2}{2} = \frac{2^2}{2} = \frac{4}{2} = 2;$$

Now let's sample at irregular time intervals:

$$X(0) = 0, \quad X(0.3) = 0.6, \quad X(0.8) = 1.6, \quad X(1) = 2$$

Signature of an irregularly sampled path:

$$\boxed{S_{0,1}(X) = (2, 2)}$$

$$S^1_{0,1}(X) = \int_0^1 dX(t) = X(1) - X(0) = 2 - 0 = 2 \qquad S^{11}_{0,1}(X) = \frac{(X(1) - X(0))^2}{2} = \frac{2^2}{2} = \frac{4}{2} = 2$$

Despite the difference in sampling times, the signature remains the same. This example shows that the signature of the path is **invariant under time reparameterization** because it only depends on the relative position of the points and not on the specific time intervals at which they are sampled. The crucial point is that the order of the points is preserved.

3.1.3 Signature for $X : [a, b] \to \mathbb{R}^2$

The signature of $X : [a, b] \to \mathbb{R}^2$ is a sequence of real numbers :

$$S_{a,b}(X) = \left(1, S^1_{a,b}(X), S^2_{a,b}(X), S^{11}_{a,b}(X), S^{12}_{a,b}(X), S^{21}_{a,b}(X), \dots\right)$$

where $X = (X_1, X_2)$.

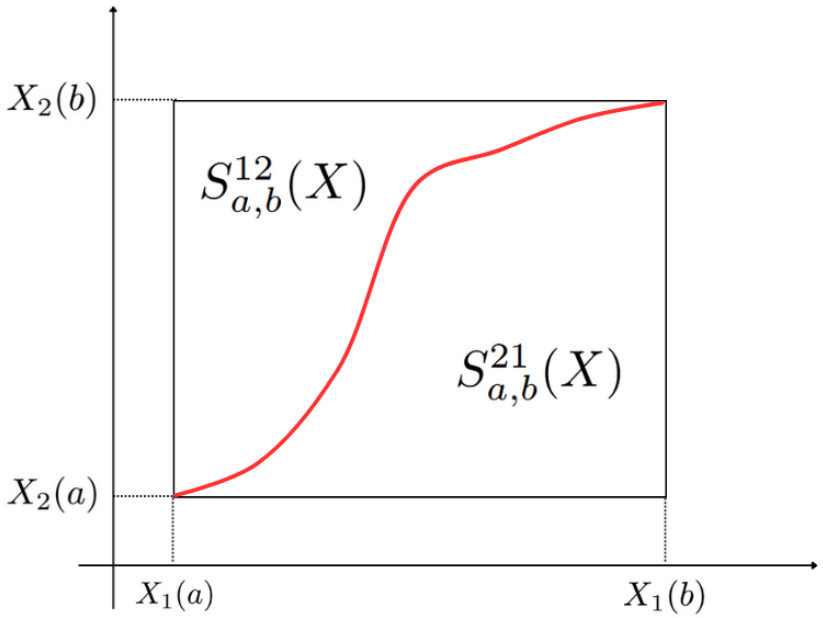
Components of the Signature

- $S^1_{a,b}(X) = \int_a^b dX_1(t) = X_1(b) - X_1(a)$
- $S^2_{a,b}(X) = \int_a^b dX_2(t) = X_2(b) - X_2(a)$
- $S^{11}_{a,b}(X) = \int_a^b S^1_{a,t}(X)\, dX_1(t) = \frac{(X_1(b)-X_1(a))^2}{2}$
- $S^{12}_{a,b}(X) = \int_a^b S^1_{a,t}(X)\, dX_2(t)$
- $S^{22}_{a,b}(X) = \int_a^b S^2_{a,t}(X)\, dX_2(t) = \frac{(X_2(b)-X_2(a))^2}{2}$
- $S^{i,j,k}_{a,b}(X) = \int_a^b S^{i,j}_{a,t}(X)\, dX^k(t)$

There is no general formula here, unlike the case when d=1, due to the rectangular terms that do not simplify easily.

Observation

The terms $S^{12}_{a,b}(X)$ et $S^{21}_{a,b}(X)$ represent areas.



Shuffle Product

From this graph, we derive the formula for the Shuffle Product. We observe a property of the signature that resembles the exponential function, but we will explore this connection later.

$$\boxed{S^{1,2}_{a,b}(X) + S^{2,1}_{a,b}(X) = S^1_{a,b}(X) \cdot S^2_{a,b}(X)}$$

3.2 The log-signature

The log-signature is interesting for two reasons. Firstly, the log-signature allows us to work in a linear space and so we can apply methods like PCA (with signatures, we deal with a curved space, which distorts the usual metrics). Secondly, because it brings out an important term for conducting tests, the Levy area.

3.2.1 General Idea

The signature $S(X)$ is a generalization of the exponential function. It then becomes natural to take its logarithm.

An interesting example : Let $X_s : [0, s] \to \mathbb{R}$ be a family of path, defined by :

$$t \mapsto t.$$

According to the calculations made in Annexe (3), we have:

$$S(X_s) = \big(1, s, \frac{s^2}{2!}, \frac{s^3}{3!}, \dots\big).$$

This sequence of numbers defines the exponential function by simple summation:

$$\boxed{e^s = \sum_{k=0}^{\infty} \frac{s^k}{k!}.}$$

Thus, the function $s \mapsto S(X_s)$ resembles $s \mapsto e^s$.

Link with the Exponential: If $X : [0, s] \to \mathbb{R}$ is a path whose signature is given by

$$S_{[0,1]}(X) = \big(1, S^1(X), S^{11}(X), S^{111}(X), \dots\big),$$

We associate to this signature the function:

$$s \mapsto 1 + S^1(X)s + S^{11}(X)s^2 + S^{111}(X)s^3 + \dots.$$

This function satisfies:

$$\boxed{S(X) = e^s.}$$

3.2.2 Let's Be Abstract

Let's associate a path $X : [0, a] \to \mathbb{R}^2$ with a non-commutative polynomial. The signature is given by:

$$S_{0,a}(X) = \big(1, S^1(X), S^2(X), S^{11}(X), S^{12}(X), S^{21}(X), S^{22}(X), \dots\big).$$

We associate a non-commutative polynomial to this signature:

$$S(X) = 1 + S^1(X)e_1 + S^2(X)e_2 + S^{11}(X)e_1^2 + S^{12}(X)e_1e_2 + S^{21}(X)e_2e_1 + \dots,$$

with the rule $e_1e_2 \neq e_2e_1$. These polynomes can multiply. For example :

$$S(X) = 1 + 2e_1 - e_2^2, \quad P(X) = 1 - e_2,$$

$$S(X) \otimes P(X) = (1 + 2e_1 - e_2^2) \otimes (1 - e_2) = 1 + 2e_1 - e_2^2 - e_2 - 2e_1e_2 + e_2^3.$$

3.2.3 Chen's identity

Definition - Concatenation of Two Paths: Let $X : [a, b] \to \mathbb{R}^d$ and $Y : [b, c] \to \mathbb{R}^d$, we define :

$$(X * Y)(t) = \begin{cases} X(t) & \text{si } t \in [a, b], \\ X(t) + Y(t) - Y(b) & \text{si } t \in [b, c]. \end{cases}$$

Chen's identity :

$$\boxed{S_{a,c}(X * Y) = S_{a,b}(X) \otimes S_{b,c}(Y)}$$

Morality : S generalizes the exponential function in the following sense:

- X, Y are replaced by real numbers,
- $*$ (concatenation) is replaced by addition,
- the tensor product $\otimes$ is replaced by multiplication,
- S is replaced by exp.

Thus, we recover the classic identity:

$$\boxed{e^{X+Y} = e^X e^Y}.$$

3.2.4 Where there is an exponential, there is a logarithm

Recall that:

$$\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$$

Since $S(X) = 1 + \big(S^1(X)e_1 + S^2(X)e_2 + S^{11}(X)e_1^2 + \dots\big)$, then $x = S(X) - 1$.

The log-signature is defined by :

$$\log S = (S(X) - 1) - \frac{(S(X) - 1)^{\otimes 2}}{2} + \frac{(S(X) - 1)^{\otimes 3}}{3} - \dots$$

Examples : 1. In a one-dimensional set up, if $S(X) = \sum_{k=0}^{\infty} \frac{1}{k!} e_1^k$, then :

$$\log S(X) = e_1.$$

2. In a two-dimensional set up:

$$\boxed{\log S(X) = S^1(X)e_1 + S^2(X)e_2 + \frac{1}{2}\big(S^{12}(X) - S^{21}(X)\big)[e_1, e_2] + \dots}$$

where $[e_1, e_2] = e_1e_2 - e_2e_1$.

Now that we have good grasp of why we find the logarithm function into signature theory, we can compute the log-signatures with a program similar to the one in the 3.2 section and then apply a PCA to our data in order to distinct between maline and benine trajectories.

3.3 PCA with log-signature

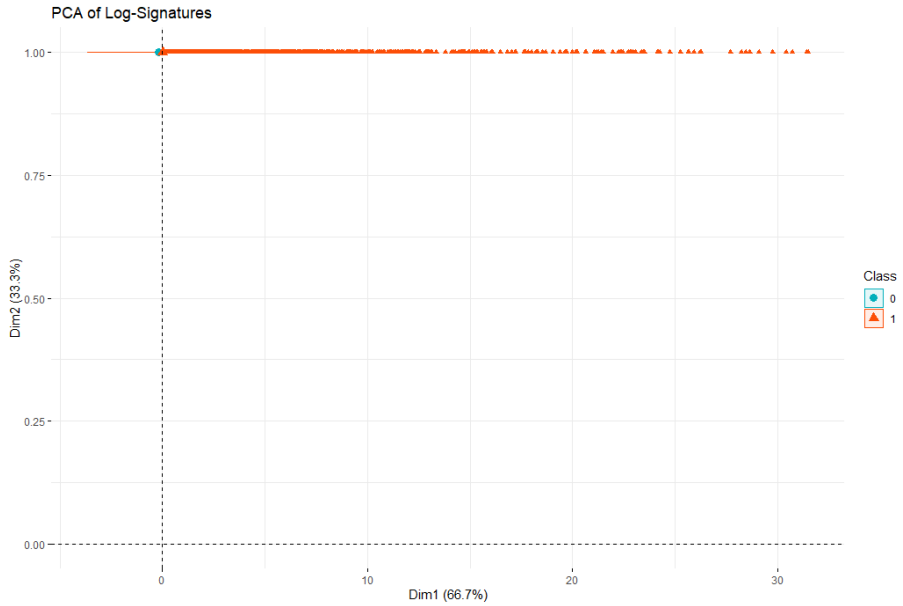


Figure 9: PCA of the log-signatures (Dim 1: 66.7%, Dim2: 33.3%)

Although we expected these dimensions to bring out a clear separation between benign (blue) and malignant (red) classes, the points corresponding to the log-signature of malign trajectories all align and the points corresponding to a benign trajectory all overlap on the same blue point (0,1). We see that they spread remain on the same ligne parallel to the first axis. By plotting the contribution we see that this axe is determined at 95% by the variance Levy Area. It seems to be the key term to distinguish between malign and benign.

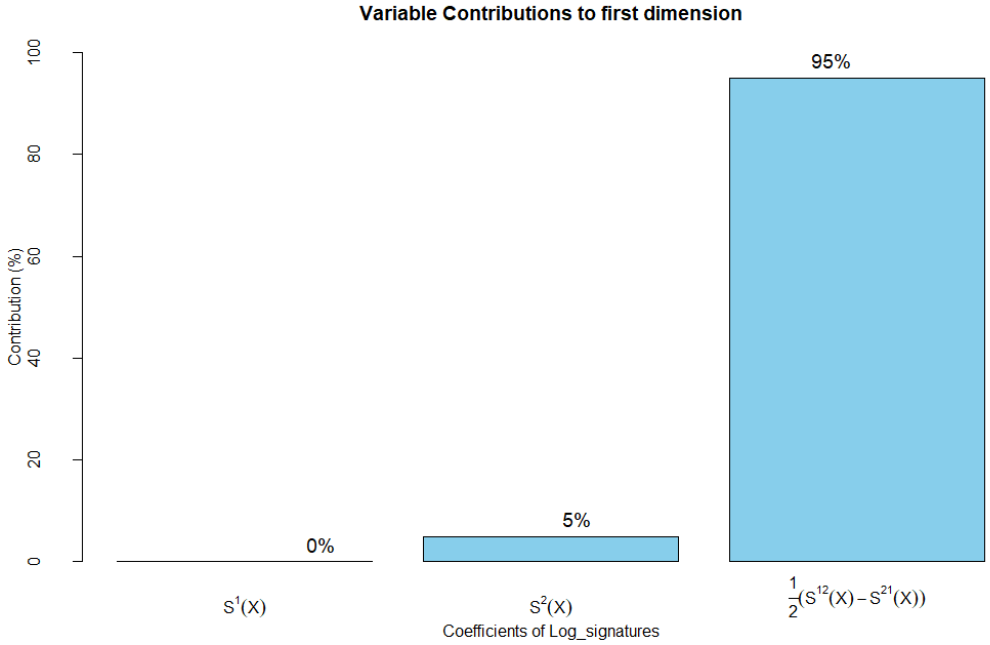
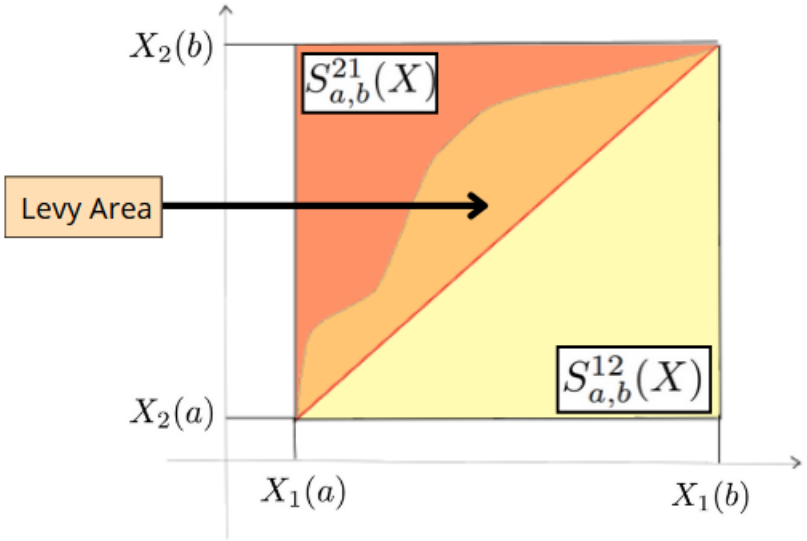


Figure 10: Contribution of each signature coefficient to the axes.

3.4 Levy Area

Definition: Let $X : [a, b] \to \mathbb{R}^2$. The Levy area is the geometric area enclosed by the closed curve formed by X, followed by the segment connecting $X(b)$ to $X(a)$.



The Levy area is a key term in the log-signature:

$$\log S(X) = S^1(X)e_1 + S^2(X)e_2 + \frac{1}{2}\big(S^{12}(X) - S^{21}(X)\big)[e_1, e_2] + \dots$$

We observe that the information here is reduced to three terms, in contrast to the signature which contains six terms at order 2. The third term represents the Levy area.

Rectangle area:

$$AR = (X^1(1) - X^1(0)) \cdot (X^2(1) - X^2(0))$$

Levy Area (LA) computation:

$$S^{2,1} = \frac{AR}{2} - LA$$

$$S^{1,2} = \frac{AR}{2} + LA$$

$$S^{1,2} - S^{2,1} = 2LA \quad \Rightarrow \quad \boxed{LA = \frac{1}{2}(S^{1,2} - S^{2,1})}$$

As highlighted by the Principal Component Analysis (PCA), the Levy Area is a key factor in distinguishing between benign and malignant paths. Therefore, we will test the Levy Area to differentiate between benign and malignant trajectories. Subsequently, we will plot the distribution of the signature coefficient corresponding to the benign dynamics.

We there see the advantage of using the signature theory and the log-signature. We can now sum up the information of one path in a few numbers and test the Levy area. Plus, the signature theory allows to follow path on which we have information at non constant interval. The reports one ct-DNA were not made at regular intervals by doctors. This would make analysis with time series very difficult since it is base on regular observations of the same variable.

4 Testing under signature

It is necessary to increase the number of benign data points from 4,000 to 40,000 to balance the dataset. This helps the classifier to recognize benign patterns more accurately and reduces the bias towards malignant predictions, which improves the accuracy and reliability of detecting benign cases. However, when plotting the distribution of signature coefficients of the benign dataset at different orders, familiar patterns are emerging.

4.1 Levy Area

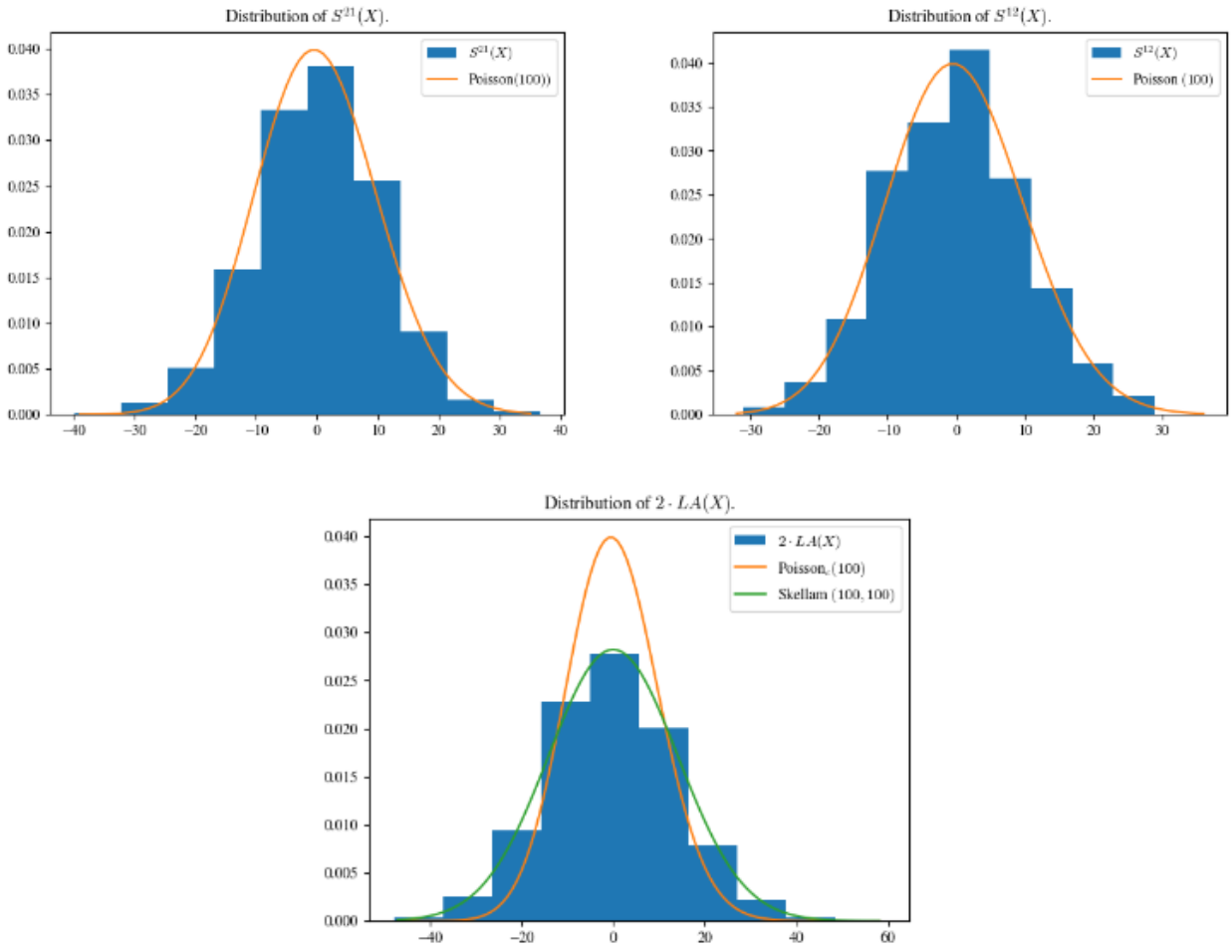


Figure 11: Histogram of the distribution of log-signature coefficients for benign data

Both signature distributions are matching Poisson distribution and we observe that $2 * LA$ seems to best fit a Skellam distribution with parameters $(100, 100)$. The Skellam distribution Skellam(μ_1, μ_2) is the discrete probability distribution of the difference between two independent Poisson-distributed random variables with respective parameters μ_1 and μ_2. Its probability mass function is given by:

$$f(k; \mu_1, \mu_2) = e^{-(\mu_1+\mu_2)} \left(\frac{\mu_1}{\mu_2}\right)^{k/2} I_{|k|}(2\sqrt{\mu_1\mu_2}),$$

Knowing that:

$$2LA = S_{12} - S_{21}$$

and $S_{12} \sim$ Poisson(100), $S_{2,1} \sim$ Poisson(100), the result we obtain is the result we expected which is that $2 * LA$ follows a Skellam distribution with parameters $(100, 100)$. Moreover, the Levy Area itself follows a Poisson distribution.

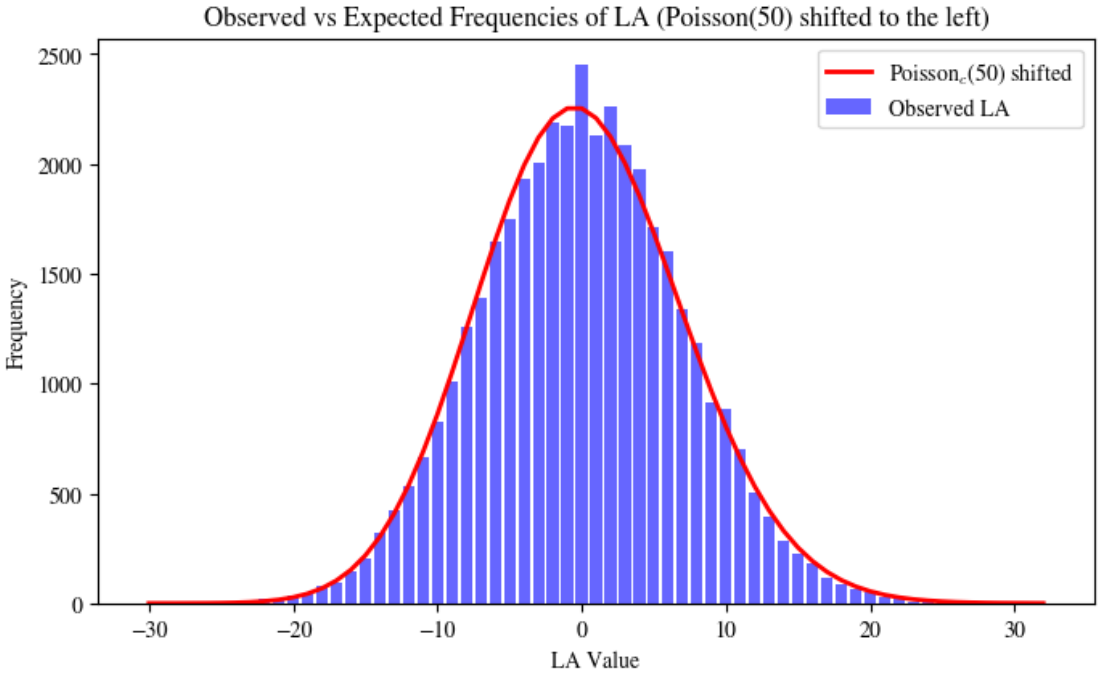


Figure 12: Histogram of the LA distribution of log-signature coefficients for benign data

Finally, we now know that benign data have coefficients following Skellam and Poisson distributions. When testing on malignant dynamics, we do not observe this pattern (see appendix). The malign data seems not to follow the same distribution. Thus, we have an effective tool to distinguish between different dynamics: we only need to test whether the signature coefficients follow a Poisson law.

Chi-Square Test Results		
Test Statistic	p-value	Result
199.147	2.29523e-17	*Significantly corresponding*

The Chi-Square Test serves as proof to consider that the LA follows in fact a Poisson distribution when considering benign trajectories.

4.2 Evaluation of Classification Performance

To quantify the effectiveness of our method, we perform a statistical test on the Levy Area coefficients. The LA should follow a Poisson distribution in presence of benign trajectories.

4.2.1 Test Statistic Definition

With X the trajectories data and $M = mean(X)/S$ the mean intensity of the process (which is half the intensity of the data). Let's consider P following a Poisson distribution of intensity M and compute the Levy Area associated to the signature coefficient of the data such that :

$$LA = \frac{S_{12} - S_{21}}{2}$$

The probability at stake is $P(P > LA)$ which is also the survival function of P evaluated at LA. The proximity of P and the data can be assessed by comparing this evaluation to a a fixed error like 5%. When X is benign, it is also true positive when the probability is lower than 5% (ie. $P(P > LA) < 0.05$) and when X is malign it is also true negative when it's higher than 5%. This test must be evaluated to see if it succeed to distinguish between benign and malign data.

Algorithm 3: Pseudo-code to process the survival function

Input: X (data matrix)
Output: Poisson survival function values for each signature coefficients

Function `test_LA(`X`):`
    `/* Calculate M, the mean of the last column of` X `divided by 2 */`
    $M \leftarrow$ mean of $X[:, :, -1]/2$
    $sig \leftarrow$ Signature(depth $= 2$)
    $X_{\text{centered}} \leftarrow X -$ mean of each row of X
    $signatures \leftarrow \text{sig}(X_{\text{centered}})$
    `/* Calculate the Levy Area */`
    $s \leftarrow 0.5 \times (\text{signatures}[:, 3] - \text{signatures}[:, 4])$
    `/* Return the Poisson survival function */`
    return *Poisson.sf(s, M, loc = -M)*

4.2.2 Results

Applying this test to benign and malignant datasets, we obtain:

- Proportion of benign data passing the test: 0.9798 (97.98%).
- Proportion of malignant data failing the test: 0.9997 (99.97%).

We define the following classification metrics:

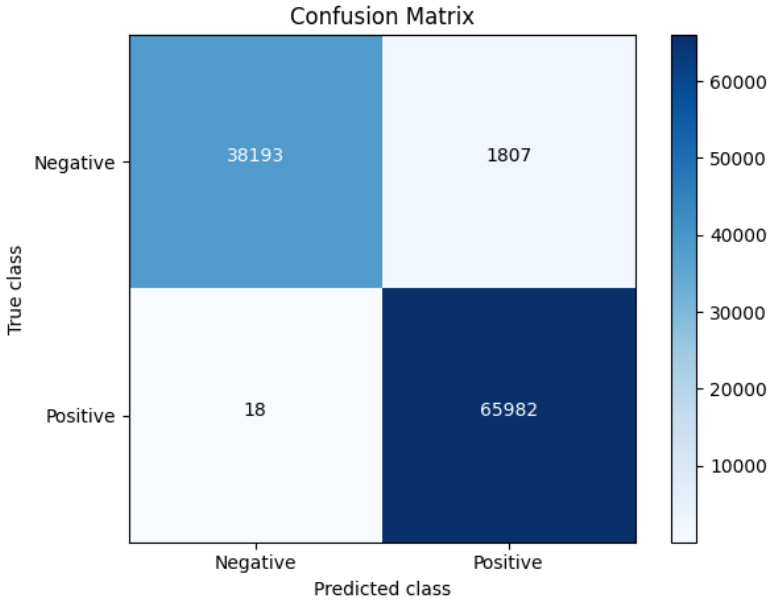


Figure 13: Confusion Matrix obtained from the simple classification results.

From these, we compute:

- **Accuracy**: $\frac{TP+TN}{TP+TN+FP+FN} = 0.9828$ (98.28%)
- **Precision**: $\frac{TP}{TP+FP} = 0.9733$ (97.33%)
- **Recall**: $\frac{TP}{TP+FN} = 0.9997$ (99.97%)
- **F1-score**: $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} = 0.9864$ (98.64%)

4.2.3 Results concerning censored data

In this section, we compare the confusion matrices obtained from different imputation methods following a censoring of the data to identify which method most closely resembles the complete data. The objective is to determine the method that maintains the highest consistency with the complete data in terms of classification performance.

The confusion matrix derived from the complete data serves as a reference. The following five methods are compared: column mean imputation, KNN imputation, matrix completion imputation, mice smoothed imputation, and row mean imputation. We examine each matrix to evaluate the similarity with the complete data in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

An important thing to note here is that we only work on the half on malign data because we don't need full data to compare the different methods to complete data and it is faster to compute.

Upon comparing the confusion matrices, it is evident that the row mean imputation method exhibits the most similar pattern to the complete data in terms of TP, TN, FP, and FN. This indicates that that method preserves the data structure more effectively compared to other imputation methods.

4.3 Multiple testing

These results demonstrate that the proposed method is highly effective in distinguishing benign and malignant dynamics. The classification performance is excellent, with an almost perfect recall and precision, making it a valuable tool for identifying different dynamic behaviors. However, they can be further refined with multiple testing. It refers to the statistical scenario where multiple hypotheses are tested simultaneously. This is common in modern data analysis, especially in fields dealing with high-dimensional data. When performing multiple hypothesis tests, the likelihood of committing at least one Type I error (false positive) increases. If each individual test is conducted at a significance level α, the probability of obtaining at least one false positive among m independent tests is given by:

$$P(\text{at least one false positive}) = 1 - (1 - \alpha)^m.$$

As m increases, this probability approaches 1, making unadjusted significance thresholds unreliable. This issue is known as the *multiple testing problem*, which leads to an inflated number of false discoveries.

This problem is particularly relevant in cancer detection using gene expression signatures. Cancer classification often relies on identifying gene expression patterns that distinguish between healthy and diseased states. Typically, thousands of genes are analyzed simultaneously to find those significantly associated with cancer. Without multiple testing corrections, many genes could appear falsely significant due to random chance, leading to unreliable biomarkers.

4.4 Benjamini-Hochberg correction method

The Benjamini-Hochberg (BH) method is designed to control that False Discovery Rate (FDR). A false positive occurs when a null hypothesis is incorrectly rejected, suggesting an effect that does not actually exist. Unlike more conservative methods, such as Bonferroni correction, which strictly control the probability of any false positives, the Benjamini-Hochberg procedure allows for some false discoveries while maintaining an overall balance between sensitivity and specificity. This makes it particularly useful in fields like genomics and neuroscience, where detecting true effects among many hypotheses is crucial.

4.4.1 False Discovery Rate (FDR)

The False Discovery Rate (FDR) is a statistical measure used to control the expected proportion of false positives when performing multiple hypothesis tests. It quantifies the expected ratio of Type I errors (false discoveries) among all rejected null hypotheses. More formally, if we denote:

- V: the number of false positives (incorrectly rejected null hypotheses),
- R: the total number of rejected null hypotheses,

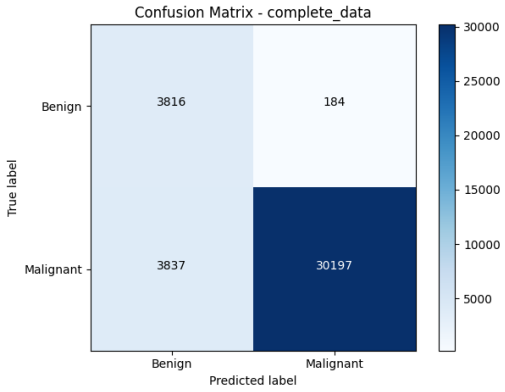


Figure 14: Complete Data

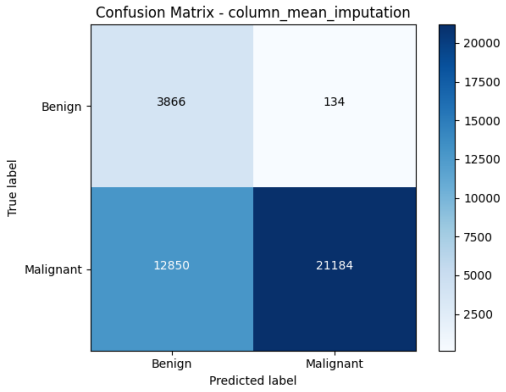


Figure 15: Column Mean Imputation

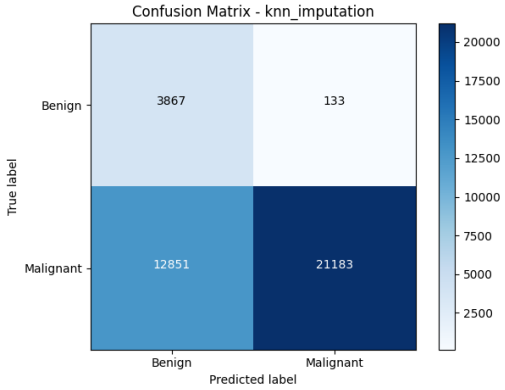


Figure 16: KNN Imputation

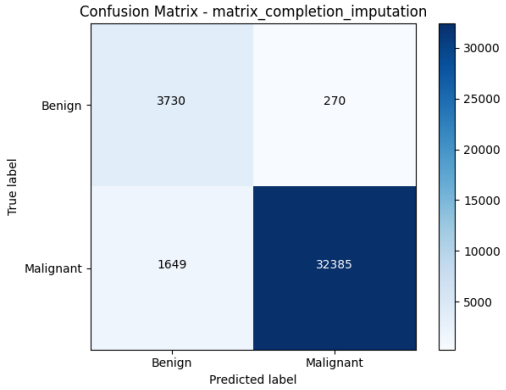


Figure 17: Matrix Completion Imputation

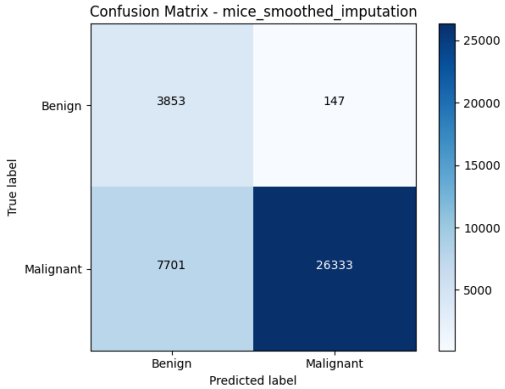


Figure 18: MICE Smoothed Imputation

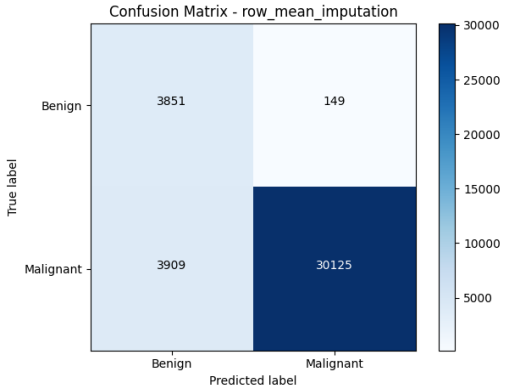


Figure 19: Row Mean Imputation

Figure 20: Comparison of Confusion Matrices for Different Imputation Methods

then the FDR is defined as:

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R}\right],$$

with the convention that $\frac{V}{R} = 0$ when $R = 0$. FDR control is widely used in fields involving large-scale testing, such as genomics or neuroimaging, as it provides a less conservative and more powerful alternative to the Family-Wise Error Rate (FWER).

4.4.2 How it works ?

The Benjamini-Hochberg procedure involves the following steps:

1. **Specify FDR Threshold**: Choose a desired FDR level q (e.g., 0.1 or 0.2)
2. **Compute P-values**: Calculate p-values for all m null hypotheses
3. **Order P-values**: Sort p-values from smallest to largest
4. **Calculate Threshold**: For each p-value, compute its critical threshold using the formula:

$$\text{Threshold} = \frac{\text{rank}}{\text{total number of tests}} \cdot q$$

5. **Identify Rejection Cutoff**: Find the largest index L where the p-value is less than its corresponding threshold
6. **Reject Hypotheses**: Reject all null hypotheses with p-values less than or equal to the p-value at index L

4.4.3 Results on simple data

First of all, we applied the Benjamini-Hochberg correction on the Levy area classification, Once new p-values were calculated, we then obtained the following new confusion matrix :

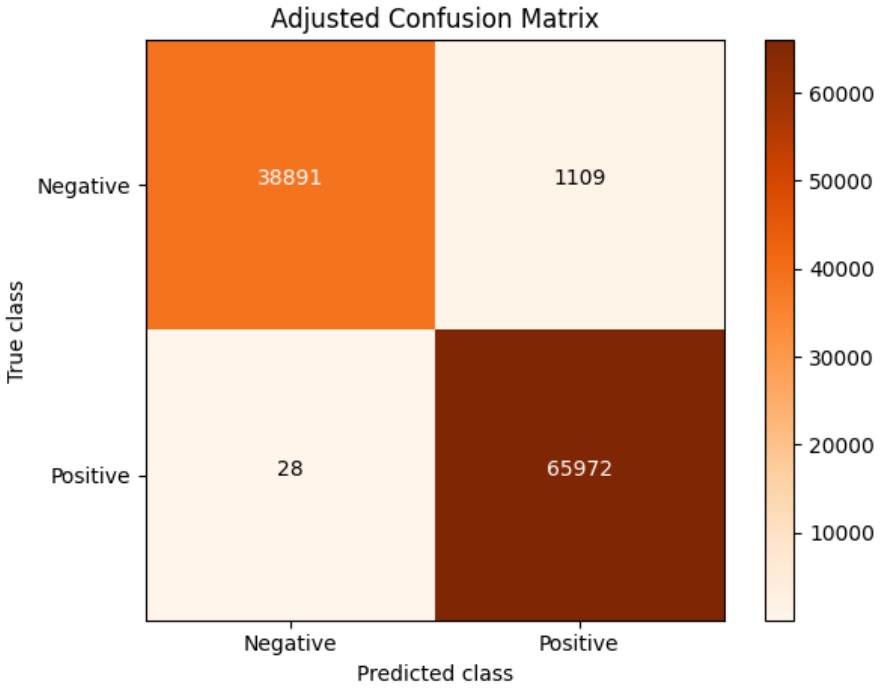


Figure 21: Confusion Matrix obtained from the simple classification results after Benjamini Hochberg correction.

In that graph, two things may keep our attention. To begin with, the false discovery rate is reduced by 698, so a diminution of (39%) which finally represents 2.8% of all benign data, which is the original objective of using that method. However, we can also observe that the false negative has increased by 10 (55%), 0.04% of all malign data, which can be a big problem in our oncology field but to balance considering the weak proportion it represents on benign data Regarding the various classification metrics, we obtained the following results:

Effect of Benjamini-Hochberg Adjustment on Performance Metrics			
Metric	Before BH	After BH	Impact
Accuracy	0.99711	0.9972	+
Precision	0.99722	0.9974	+
Recall	0.99973	0.9997	-
F1 Score	0.99847	0.9985	+

Figure 22: Difference of classification metrics

As for the different classification metrics, the results are as follows: While most metrics remain almost unchanged, the slight variations observed—such as the minor drop in recall—highlight the subtle impact of the Benjamini-Hochberg adjustment on model sensitivity, without significantly affecting overall performance.

4.4.4 Results on censored data

5 Different methods of classification applied to the signature

5.1 Support Vector Machines (SVM)

5.1.1 Why Use SVM in This Context

SVM are supervised learning algorithms particularly well-suited for optimally separating data into classes and here it's perfectly what you expects to do with signatures. Some of the advantages that they offer may interess us like:

- **Robustness**: SVMs are less sensitive to outliers thanks to their principle of maximizing the margin between the data points and the separating hyperplane.
- **Effective Generalization**: By focusing primarily on the points closest to the decision boundary (the support vectors), they tend to generalize well to new data.

When the goal is to classify or predict a phenomenon based on complex data such as signatures, SVMs can be an good choice for both classification and regression tasks.

5.1.2 SVM Model Construction and Hyperparameter Tuning

We used Support Vector Machine (SVM) in our approach to classify benign and malignant signals. We experimented with different kernels and chose the Radial Basis Function (RBF) kernel, as it turned out to be more appropriate for our data than linear kernels. We chose the RBF kernel because it allows for non-linear decision boundaries, which better captures the complexity present in our data.

For the best performance and generalization, we have chosen the following method. We have used the Support Vector Machine (SVM) to classify the benign and malignant signals. We have experimented with different kernels and we have chosen the Radial Basis Function (RBF) kernel as it is much more suitable for our data than the linear kernels. The RBF kernel was chosen because it allows for the non-linear decision boundaries that better capture the complexity of our data.

then, we performed hyperparameter tuning using cross-validation. We used a grid search strategy to explore combinations of the penalty parameter C and the kernel coefficient γ. To evaluate the model's robustness, we performed a 5-fold cross-validation, where the data was split into five subsets, and the model was trained and tested 5 times with different training and validation sets. The performance metrics, including accuracy, precision, recall, and F1-score, were averaged across the folds to provide a reliable estimate of the model's performance.

The main goal was to find the optimal combination of C and that maximized the F1-score while keeping the accuracy high. The RBF kernel was chosen because linear kernels gave bad classification results. They often classified all cases as benign. The cross-validation process helped to avoid overfitting and gave more reliable evaluation of the models. In this case, each data point was used for both training and testing.

5.1.3 SVM Results with Normal Data

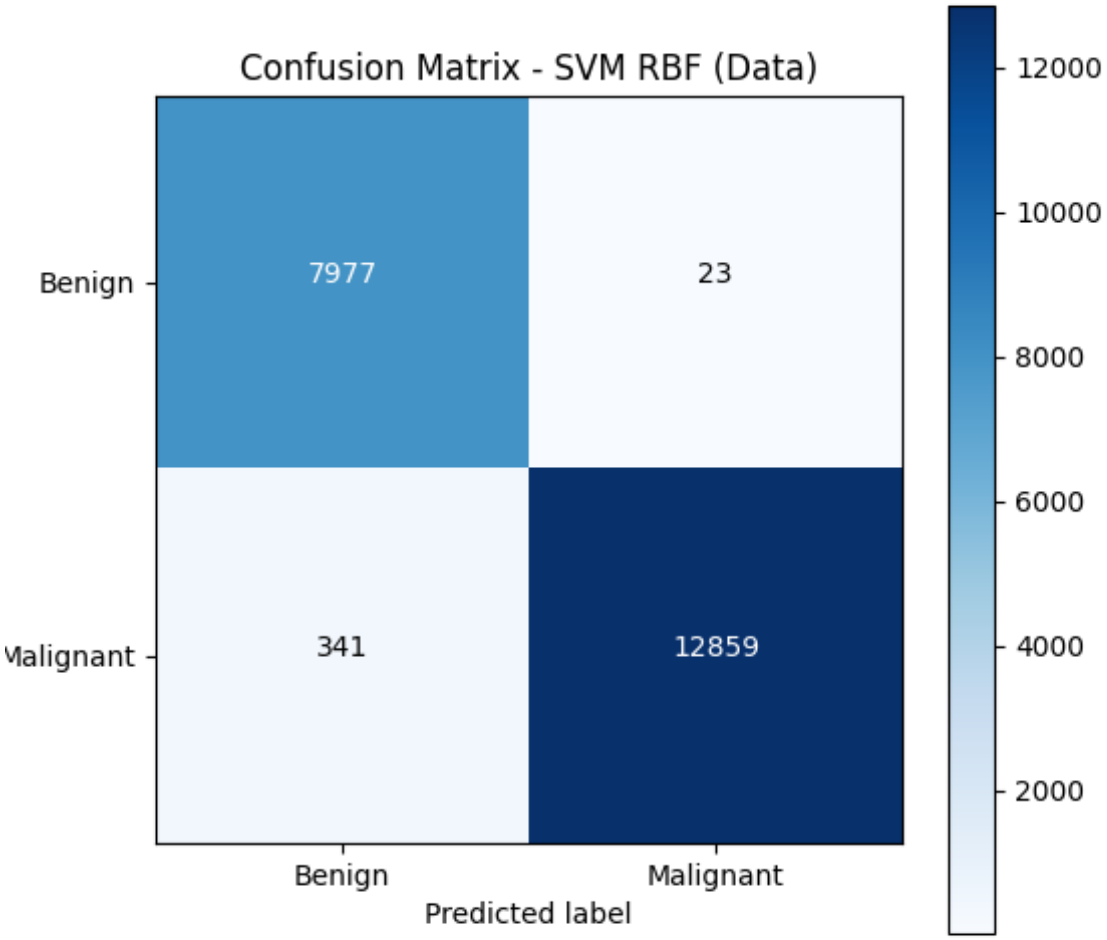


Figure 23: Confusion Matrix - SVM RBF with Normal Data

The results show that the model correctly classifies the majority of both benign and malignant cases, we obtained excellent results for the false positive rate, we have here 23 which represents 0.29% of benign data. However, results for false negative are not pretty good, 341 which represents 2.6% of malign data. Even if the false positive are better than for the levy area classification, the false negative in that case isn't acceptable. In fact it's around 13 times the number for LA testing, and it is just the test sample, which is 20% of the real malign data.

5.1.4 SVM Results with Signature Data

In contrast, the second SVM model was trained using data transformed with signatures. We again chose the RBF kernel to accommodate potential non-linear relationships. The corresponding confusion matrix is shown in Figure 24.

The model's performance when using signatures shows a significant difference compared to the raw data. In fact, the results are worse than those obtained with the raw data. More than 50% of benign and malignant data are incorrectly predicted.

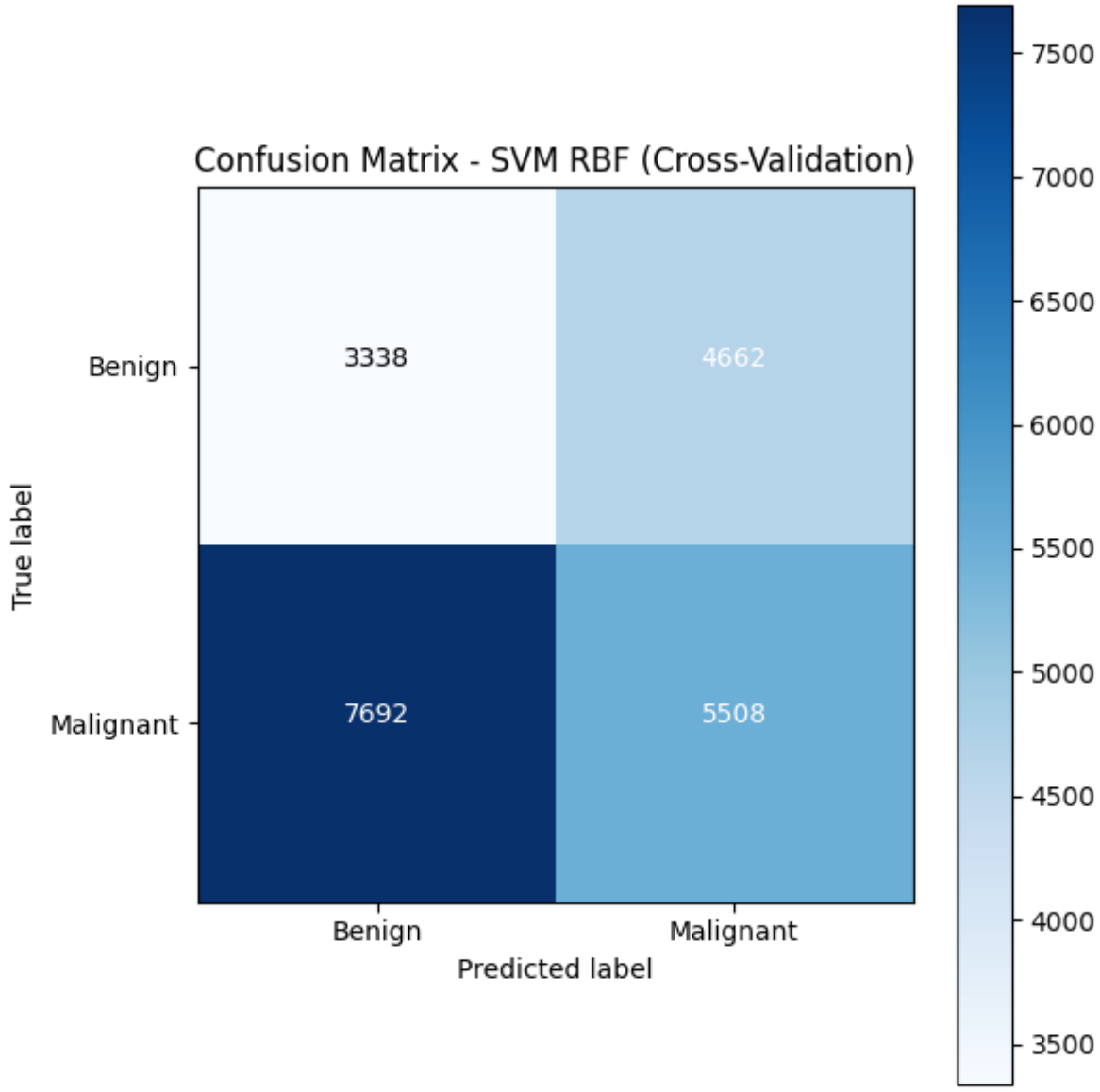


Figure 24: Confusion Matrix - SVM RBF with Signature Data (Cross-Validation)

5.2 Generalized Additive Models

5.2.1 Why GAM are pertinent in this context

Generalized Additive Models (GAMs) extend generalized linear models (GLMs) to handle non-linear relationship between predictors and the response variable, making them suitable for classification tasks. Here we use Logistic GAM, which is a GAM with a Binomial error distribution and a logit link. Some advantages in this time series data can be:

- **Nonlinear Pattern Capture**: Splines model temporal trends (e.g., sudden spikes or gradual changes).
- **Interpretability**: Unlike "black-box" models (e.g., neural networks), GAMs allow visualization of feature effects (e.g., partial dependence plots).
- **Additivity**: The contribution of each time step to the prediction is additive and interpretable.

5.2.2 GAM Construction

We employed a Generalized Additive Model (GAM) with logistic regression to classify benign and malignant time series signals. GAMs were selected for their ability to model non-linear relationships through additive smooth functions while retaining interpretability, a critical advantage given the complex temporal patterns observed in our data. Features were standardized to mitigate scale sensitivity during spline fitting. To address class imbalance, a weighted loss function was applied, assigning higher penalties to misclassifications of the minority class (benign) to ensure balanced performance.

The GAM architecture incorporated cubic splines with 8 basis functions per feature, enabling flexible nonlinear modeling of individual time steps. Regularization was applied via a penalty term on the splines' second derivatives to prevent overfitting. The model was trained using penalized iteratively reweighted least squares (PIRLS), optimized to maximize likelihood with class-weighted adjustments. Hyperparameters, including the number of splines and regularization strength, were selected empirically through preliminary experiments, balancing flexibility and computational efficiency. The dataset was split into stratified training (80%) and testing (20%) sets, and performance was evaluated using accuracy, precision, recall, and F1-score.

5.2.3 GAM Results with Normal Data

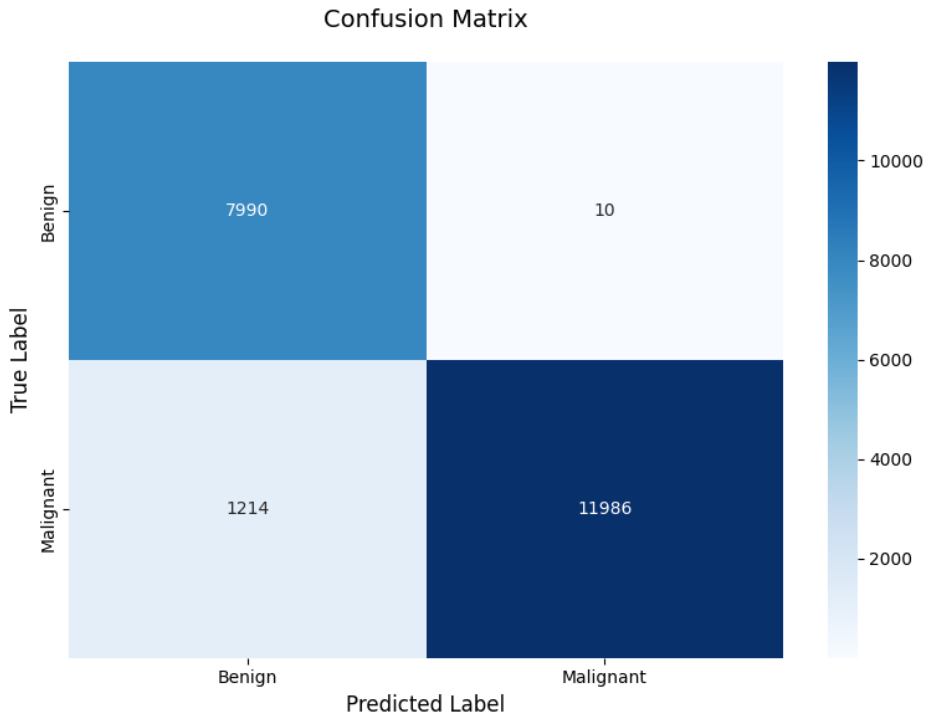


Figure 25: Confusion Matrix - GAM with Normal Data

The Generalized Additive Model (GAM) achieved 94% overall accuracy on the test set, demonstrating strong performance in distinguishing benign and malignant data. Only 10 benign cases were misclassified as malignant, while 9% of malignant samples (1,214) were misclassified as benign, which could be critical in clinical settings.

5.2.4 GAM Results with Signature Data

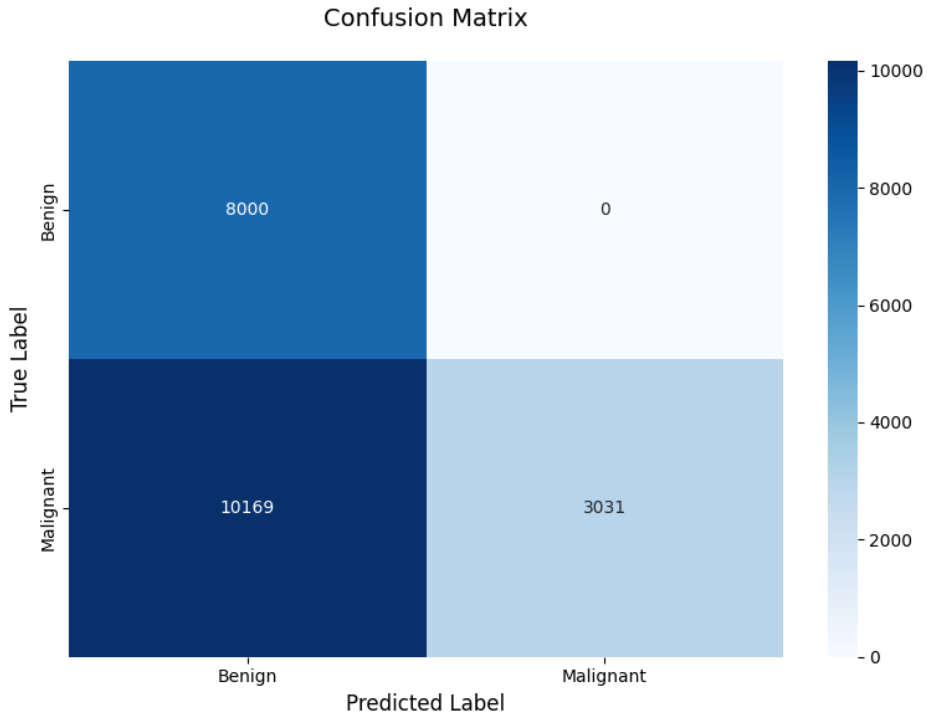


Figure 26: Confusion Matrix - GAM with Signature Data

When we use signature components (e.g., S_2, S_{12}, S_{21}, Levy Area) as features in GAM, the model is ineffective. The result refers to a github issue in pygam package, which it's still needed to be solved. Overall, classification in GAM can be seen as a trying method in this context, and it's not so robust and reliable.

6 Discussion

The signature theory has provided us with interesting results. It reveals distinction between signals where common statistical tools failed. It allowed for the engineering of a test statistic to identify malign trajectories within a ctDNA trajectories dataset. The false positive rate associated to this statistic was refined. It unveils promising results in the field of oncology and could lead to a reliable method for early stage detection thus saving lives. Moreover, the following sections addresses the issues encountered along the study, the crucial assumptions and provides food for thoughts on the matter.

6.1 Neural network

6.2 The data

The nature of the data imposed a generation model which could be inaccurate regardless of the effort made and the associated literature. The biological process behind ctDNA generation is only partially known. Due to this, Birth and Death Process may need new migration rates or more accurate birth and death rates. An interesting approach would be to fit the values of the BDP on clinical data on a metric such as the Least Square Estimator. In this theoretical approach, any noise has been silenced, meaning that the monitoring process should be perfect. This assumption is unrealistic but adding noise could suppress the ability to distinguish between benign and malign trajectories.

Regarding censored data and their imputation, we can assume that with real data and real censorship, the imputation results might not be as accurate. This is because the data were generated using an algorithm that follows certain regularities, which may differ from real-world patterns. Similarly, the censorship process we applied is quite simple and may not fully reflect actual censoring mechanisms. These factors combined could lead to less effective imputation and, consequently, less reliable results.

6.3 Interpolation Methods

In the current study, One have chosen to link the generated points using linear interpolation. This choice was motivated by the simplicity and computational efficiency of calculating signatures from paths. The linear interpolation method is straightforward and well-suited for the calculation of signatures, allowing us to rapidly compute and analyze large datasets.

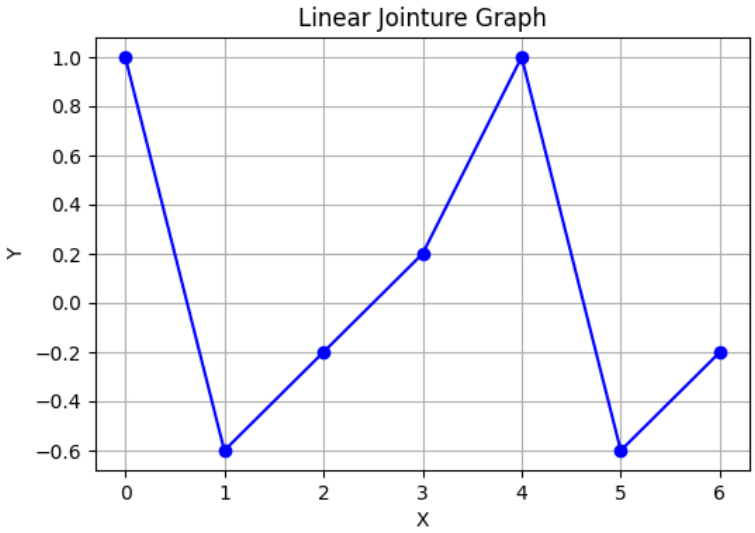


Figure 27: Linear Interpolation of Cell Life Data

However, an alternative approach that could be considered is the use of splines or other smooth interpolation methods. In particular, splines would provide a more continuous and smooth representation of the generated data points. This approach would be more consistent with the biological context, as the decay or evolution of molecular signals often follows smooth, non-linear patterns rather than abrupt linear transitions.

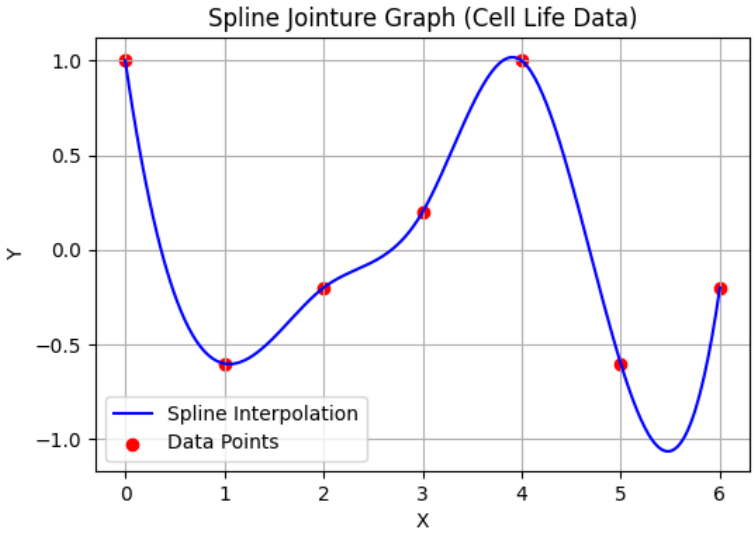


Figure 28: Smooth Interpolation of Cell Life Data using Splines

The main advantage of using splines is that they can capture more realistic transitions between points, which could lead to a more accurate modeling processes. Moreover, the smoothness of the spline paths would potentially enhance the interpretability of the signature-based models, as they would better mimic the actual physiological changes over time.

However, adopting a spline-based approach would also come with significant challenges. First, the computational complexity of calculating signatures from spline-interpolated data would be much higher compared to linear interpolation. Additionally, the entire data preprocessing and signature calculation workflow would need to be redesigned from scratch, as the current methods are optimized for linear paths. The additional complexity might outweigh the potential benefits, especially if the resulting improvements in classification accuracy are not substantial.

Thus, while the idea of spline-based interpolation presents an interesting direction, its practical implementation would require careful consideration of the computational cost and the potential gain in accuracy. Further research would be necessary to evaluate whether the benefits of a smoother, more biologically consistent model would justify the additional computational effort and complexity.

7 Bibliography

References

[1] Siegel, R. L., Miller, K. D., Jemal, A. (2020). Cancer statistics, 2020.

[2] Vallée, A., Marcq, M., Bizieux, A., El Kouri, C., Lacroix, H., Bennouna, J., Douillard, J.-Y., Denis, M. G. Plasma is a better source of tumor-derived circulating cell-free DNA than serum for the detection of EGFR alterations in lung tumor patients.

[3] Schwarzenbach, H., Hoon, D. S. B., Pantel, K. (2011). Cell-free nucleic acids as biomarkers in cancer patients.

[4] Rhodes, A., Wort, S. J., Thomas, H., Collinson, P., Bennett, E. D. Plasma DNA concentration as a predictor of mortality and sepsis in critically ill patients.

[5] Chang, C. P., Chia, R. H., Wu, T. L., Tsao, K. C., Sun, C. F., Wu, J. T. (2003). Elevated cell-free serum DNA detected in patients with myocardial infarction.

[6] Beck, J., Oellerich, M., Schulz, U., Schauerte, V., Reinhard, L., Fuchs, U., Knabbe, C., Zittermann, A., Olbricht, C., Gummert, J. F., Shipkova, M., Birschmann, I., Wieland, E., Schütz, E. (2015). Donor-Derived Cell-Free DNA Is a Novel Universal Biomarker for Allograft Rejection in Solid Organ Transplantation.

[7] Esfahani, M. S., Hamilton, E. G., Mehrmohamadi, M. (2022). Inferring gene expression from cell-free DNA fragmentation profiles.

[8] Miller, K. N., Victorelli, S. G., Salmonowicz, H., Dasgupta, N., Liu, T., Passos, J. F., Adams, P. D. Cytoplasmic DNA: sources, sensing, and role in aging and disease.

[9] Butler, T. M., Spellman, P. T., Gray, J. (2017). Circulating-tumor DNA as an early detection and diagnostic tool.

[10] Avanzini, S., Kurtz, D. M., Chabon, J. J., Moding, E. J., Hori, S. S., Gambhir, S. S., Alizadeh, A. A., Diehn, M., Reiter, J. G. A mathematical model of ctDNA shedding predicts tumor detection size.

[11] Rew, D. A., Wilson, G. D. (2000). Cell production rates in human tissues and tumors and their significance. Part II: clinical data. Eur. J. Surg. Oncol., 26, 405–417.

[12] Loi Informatique et Libertés art.78-17, Code de déontologie des professions de santé, Code de la Santé Publique.

[13] Mitrofanova, A. (2007). Lecture 3: Continuous times Markov chains. Poisson Process. Birth and Death process. NYU, Department of Computer Science.

[14] Cox, D. R. (1955). Some Statistical Methods Connected with Series of Events. Journal of the Royal Statistical Society.

8 Appendix

[a-0]

Amount of ctDNA correlates with tumor volume in stage I-III lung cancer. Haploid genome equivalents (hGE) correlate with tumor volume with a linear regression slope of close to 1 in logarithmic space. Gross tumor volumes were inferred from CT-scans. The blue shaded region depicts the 95% confidence interval of a linear regression in log-log space.

A — Reanalyzed data of 87 early stage non-small-cell lung cancer patients of the TRACERx cohort, where accurate volumetric estimates were available. Haploid genome equivalents were calculated from the mean variant allele frequency (VAF) of all clonal mutations per subject (see Materials and Methods for details). For subject CRUK0053, clonality estimates were not available, and hence all mutations were assumed to be clonal. If tumors with lower volumes (close to the limit of detection, with higher variance) are excluded, the slope approaches 1. Excluding tumors with volumes below 2 and 3 cm^3 leads to slopes

of 0.85 and 0.91, respectively. Linear regression with a fixed slope of 1 predicts 0.25 hGE per plasma mL for 1 cm^3 of tumor volume (95% CI: 0.15–0.40 hGE per plasma mL).

B — Reanalyzed data of 46 early stage non-small-cell lung cancer patients. Haploid genome equivalents were reported in Supplementary Table 3 of the original study. Linear regression with a fixed slope of 1 predicts 0.11 hGE per plasma mL for 1 mL of tumor volume (95% CI: 0.055–0.21 hGE per plasma mL).

C — Reanalyzed data of 49 locally advanced non-small-cell lung cancer patients. Haploid genome equivalents were reported in Table S2 of the original study. Note that only for six subjects were mutations known a priori. For the remaining 43 subjects, mutation calls were based on the blood samples, and hence required stricter calling thresholds, decreasing the sensitivity. Linear regression with a fixed slope of 1 predicts 0.27 hGE per plasma mL for 1 mL of tumor volume (95% CI: 0.18–0.41 hGE per plasma mL).

[a-1]

(A) Mathematical model of cancer evolution and ctDNA shedding. Tumor cells divide with birth rate b and die with death rate d per day. During cell apoptosis, cells shed ctDNA into the bloodstream with probability q_d. ctDNA is eliminated from the bloodstream with rate ε per day according to the half-life time of ctDNA, $t_{1/2} = 30$ min. (B) hGE per plasma ml correlate with tumor volume with a slope of 0.9997 in 176 patients with lung cancer ($R^2 = 0.32$; $P = 2.6 \times 10^{-16}$). Red shaded region depicts 95% confidence interval (CI). Linear regression predicts 0.21 hGE per plasma ml for a tumor volume of 1 cm^3, leading to a shedding probability of $q_d = 1.4 \times 10^{-4}$ hGE per cell death (Materials and Methods). (C) A tumor starts to grow at time zero with a growth rate of $r = b - d = 0.4\%$ ($b = 0.14$, $d = 0.136$), typical for early-stage lung cancers (tumor doubling time of 181 days). Tumor sheds ctDNA into the bloodstream according to the product of the cell death rate d and the shedding probability per cell death q_d. (D) Distribution of ctDNA hGE in the entire bloodstream at the time [purple dashed line in (C)] when the tumor reaches 1 billion cells [$\approx$ 1 cm^3; gray dashed line in (C)], leading to a mean of 0.21 hGE per plasma ml and a tumor fraction of 0.022% at a mean DNA concentration of 6.3 ng per plasma ml. (E) Probability distribution of ctDNA mutant fragments present in a liquid biopsy of 15 ml of blood when the tumor reaches 1 billion cells (assuming that the covered somatic heterozygous mutation is present in all cancer cells).

[a-2]

Algorithm 4: Pseudo-code for Generating Cox Process Data

Input: b (birth rate), d (death rate)
Output: Biomarker trajectory for the Cox process

Function `generate_cox_process(`b, d`):`
  `/* jump times for malign and benign data */`
  transition_time ← `np.random.uniform(400, 1500)` ;
  jump_times_b ← jump_times before transition time ;
  jump_times_m ← jump_times after transition time ;
  `/* Generate the cell population using birth-death process before and after transition time */`
  cell_pop_b ← `bd.simulate.discrete([d, d], 'linear', z0 = 1e9, times = jump_times_b, method = 'gwa')` ;
  cell_pop_m ← `bd.simulate.discrete([b, d], 'linear', z0 = cell_pop_b[-1], times = jump_times_m, method = 'gwa')` ;
  cell_pop ← `cell_pop_b + cell_pop_m` ;
  traj_cell ← cell trajectory and time;
  `/* Generate the intensity for the Cox process */`
  intensity ← $\frac{traj\_cell[:,:] \cdot d \cdot 1.4 \cdot 10^{-4}}{33+d-b}$;
  `/* Generate the biomarker trajectory using a Poisson distribution */`
  biom_traj ← `np.random.poisson(lam = intensity)` ;
  traj_biom ← biomarker trajectory and time;
  return *traj_biom*