Estimating the probability of event occurrence

ALEXANDRE GUINAUDEAU

Master in Computer Science
Date: June 24, 2018
Supervisor: Pawel Herman
Examiner: Hedvig Kjellström
Swedish title: Uppskattning av sannolikheten för händelse
School of Electrical Engineering and Computer Science

Abstract

English abstract goes here.

Sammanfattning

I will translate my abstract once the English version has been approved.

Contents

Introduction

In complex systems, errors can occur intermittently and in a non-deterministic way, which makes it harder to diagnose real errors among spurious ones. In manufacturing for instance, intermittent errors could be due to physical properties, either internal, like bad contacts, or external, e.g. extreme temperatures. In any case, these errors are often hard to troubleshoot and require close attention. By analogy with *flaky tests* in computer science, we will also refer to these as *flaky errors*.

Non-deterministic errors are often considered as unreliable and therefore discarded, which creates an important risk of ignoring a real error. On the other hand, troubleshooting each occurrence of a flaky error is very time-consuming and is not always an option. Therefore, it is critical to detect when flaky errors occur at an unexpected rate and to pinpoint when the rate of failure is likely to have evolved. This enables engineers to understand which elements have an impact on the error. In computer science, the usual workaround for flaky tests is to re-run tests that failed a certain number of times until they pass. In other fields, this is not always possible as errors are triggered in a production environment.

In this thesis, we intend to estimate the underlying probability of occurrence of an error. We assume its distribution to be piecewise stationary. This corresponds to the fact that the probability of occurrence changes when a part breaks, wears out, or is repaired. Given this assumption, estimating the underlying probability of occurrence is equivalent to finding its change points.

Chapter 1

Pilot study

1.1 Objective

Given a sequence of events, we intend to detect changes in the probability of occurrence. This has two main interests:

- Alert when events start occurring more frequently.
- Troubleshoot errors by pinpointing when the probability of occurrence changed.

For instance, in the case of intermittent errors due to bad contact, the detection of change in frequency could enable users to understand that high vibrations triggered the bad contact, and that some specific maintenance work fixed it.

According to [1], two binomial distributions $Bin(n, p)$ and $Bin(n, p+\varepsilon)$ (where n is the number of observations, and $p \gg \varepsilon$ the probability of occurrence of the event) can be distinguished after

$$n \sim \frac{K}{\varepsilon^2}$$

observations, for some K independent of ε. So a change in probability of 0.1 should be detected within the order of $n \sim 100$ observations.

News value In complex systems, it may be too time-consuming to monitor all the errors that occur. in that case, unusual errors are carefully troubleshooted, but intermittent and non-deterministic errors can end up being ignored. In that case, being able to flag when an error occurs

more frequently than usual can be vital. Indicating precisely when the frequency increased significantly is also very useful to quickly understand the reason for this increase in frequency and troubleshoot the error.

1.2 Research question

What is the sensitivity of detection of changes in the frequency of event occurrence? In other words, given a target level of confidence, what is the minimum change in the frequency that can be identified, and what delay is necessary to reliably identify this change?

Examination method First, we will find the tuning of the parameters both candidate models that provide the target confidence level. Therefore, we will use the mathematical definition to estimate the how the parameters should be tuned, and we will approximate the actual parameters using Monte Carlo methods.

Then, we will compare the sensitivity and delay of the models. We will generate data with a single change point, and measure the number of observations required to detect the change, depending on the amplitude of the change.

Finally, we will discuss how well both models could be generalized to real data. To emulate real data do so, we will generate sequences of probabilities of occurrence in a *realistic* way (See 1.3), and derive events based on those probabilities. Then, we will estimate the underlying probabilities based on the events, and compare the candidate models to find the one that performs best. We will also compare efficiency of the algorithms (in terms of memory and computation), and see how easily they could be adapted to online data.

Expected scientific results The hypothesis being tested is that we are able to detect *significant* changes in probability of occurrence, *shortly* after the change. The only parameter of our model should be the target level of confidence (for instance 95%). Performing maintenance is usually very expensive as it often requires to replace a part. Therefore, we only want to trigger an alert if we have sufficiently high confidence that a change in the frequency actually occurred.

Given these, we want to measure:

- the *sensitivity* of the detection: the minimal increase or decrease in the frequency of occurrence that we are able to detect
- the *delay* of the detection: the minimal number of observations after a change in frequency required to detect a change

To measure the accuracy of our estimations, we will compare their sensitivity and delay on generated data with a single change point.

For data with multiple change points, we would use the $\mathcal{L}_2$ score to measure the error with the true underlying distribution.

1.3 Literature study

Generating events To generate our data, we will simulate part failures that lead to an increased probability of triggered events. Interestingly, the lifetime of organisms, devices, structures, materials in both biological and engineering sciences have very similar behaviors. For example, business mortality [2], failures in the air-conditioning equipment of aircrafts or in semiconductors [3] and integrated circuit modules [4] all have similar behaviors. These can be modeled with a mixture of exponential or Weibull-Lomax distributions. In particular, the Weibull distribution [5] is the most widely used to model the lifetime of parts [6], as it has a limited number of parameters which can easily be interpreted, and captures both the *infant mortality* of defective parts and the exponential distribution of events that occur independently at a constant average rate, for normal parts. The two parameters are used to reflect these two elements, the defects in the material and the average rate of failure.

Detection of probability change-points The detection of a change point can be formulated as a null hypothesis that determines whether all events were drawn from the same binomial distribution [7]. To generalize this to multiple change points, we could use a sliding window. However, this is often unstable because the change point detection is performed on small regions, and therefore isn't always statistically significant [8], [9]. Therefore, we will also explore binary segmentation or dynamic programming to generalize our algorithm to sequences of events with multiple change points [10], while still running in linear time [11].

Another lead could be to use Hidden Markov Models [12], which seems well-adapted to our use case, as we are trying to guess the hidden probability state that was used to generate the events. Chis and Harrison [13] suggest a solution to adapt the model to online problems by estimating its updated parameters. This can be used to avoid recomputing the hidden state with new values.

1.4 Specific problem definition

Definitions As we will generate our data, it is important to clearly specify how the data is generated and what assumptions are made.

We consider a complex system composed of many different physical parts, and equipped with a logging system that triggers an error when an unusual behavior is detected. This logging system stores a sequence of observations t, for which events either occurred or did not. These observations could be of various time scales, either periodic (e.g. 1 per second) or based on iterations (e.g. one per cycle in manufacturing). We consider a single event e, independently of the other ones that were recorded by the logging system.

- We call *observation* t each iteration where a value is stored by the logger.
- We call *event* or *error* e a specific kind of event that was recorded by a logging system. For each observation t, we note $y_t = 1$ if the event occurred and $y_t = 0$ otherwise. We note $p(t) = \mathbb{P}[y_t = 1]$ the probability that e occurs at instant t.
- We call *part* a physical part of our system; it can be anything from a valve to a sensor. The failure or wear of such a part could increase the probability of occurrence of e, and maintenance or replacement of the part could decrease it.
- We call *change point* any of these changes affecting a part which has an impact on e.

Assumptions We assume that the event occurs with a fixed probability p, independent of previous events, between two consecutive change points. Note that this is not a lost in generality intrinsically, given we

could consider change points at every observation. However, we will in general consider a low frequency of change points.

To tune the parameters of our models, we will consider events generated with a constant probability p. To evaluate their sensitivity and delay, we will consider events generated with to probabilities p_1 and p_2, and a single change point. To evaluate their capacity to generate to real-life data, we will use a Weibull distribution to model the lifetime of parts, and derive series from successive underlying probabilities.

1.5 Using a null hypothesis to tune our models

As mentioned earlier, we want to define models with a single parameter, the target level of confidence α. We will tune the parameters of our models such that the probability that we erroneously detect a change point is at most α. Given a sequence of events $y_0, \ldots, y_{T-1}$, we consider the null hypothesis:

H_0: All events $y_0, \ldots, y_{T-1}$ were drawn from the same Bernoulli distribution $B(1, \lvert y \rvert)$

If the model is correctly tuned, we should reject this null hypothesis with a probability of α for distributions drawn from a constant distribution. Therefore, when our model detects a change point, we can assert that the underlying probability of occurrence has changed with a confidence of $1 - \alpha$.

Chapter 2

Hidden Markov Model

2.1 Mathematical definition

The first model we consider to estimate the underlying probabilities is a Hidden Markov Model, where the hidden states are the underlying probabilities.

We note x_t the hidden state at t and y_t the corresponding observation. The parameters of a HMM are:

- The number of states N. In our case states correspond to probabilities $p_0, \dots, p_{N-1}$
- The number of observations T, observations are noted $y_0, \dots, y_{T-1}$
- The initial hidden state $\varphi_i = \mathbb{P}[x_0 = p_i]$
- The transition probabilities $\phi_{i,j} = \mathbb{P}[x_{t+1} = p_j \mid x_t = p_i]$
- The distribution from which observations are drawn. Here, observations are drawn from a Bernoulli distribution: $y_t \sim B(1, x_t)$

Given N, there are $\mathcal{O}(N^2)$ independent parameters φ_i and $\phi_{i,j}$. We will make a few additional assumptions to reduce the complexity.

Uniformly distributed states We assume the states are uniformly distributed on the interval $[0, 1]$:

$$\forall i \in [|0, N-1|], p_i = \frac{i}{N-1}$$

This is not a loss in generality, as we can increase the number of states to cover any possible probability $p \in [0, 1]$

Uniform prior distribution No particular prior distribution is assumed, so the initial states are sampled from a discrete uniform distribution over all possible states:

$$\forall i \in [|0, N-1|], \varphi_i = \frac{1}{N}$$

Uniform change amplitude We also assume that when a change occurs, all new states are equiprobable, in other words that, given i, $\phi_{i,j}$ is constant for any $j \neq i$.

Uniform change likelihood We also assume that probability that a change occurs is independent of the current state, in other words that $\phi_{i,i}$ is constant. This assumption is more debatable, as when errors occur frequently, operators are likely to do maintenance operations that will decrease the frequency of occurrence. For the sake of simplicity, we will make this assumption for now and discuss this choice in the later (See 4).

With these two assumptions on ϕ, the matrix of transition probabilities now has two values, one for diagonal coefficients $\phi_{i,i}$ and one for extra-diagonal coefficients $\phi_{i,j}$. We note ρ the ratio between these two coefficients:

$$\rho = \frac{\phi_{0,1}}{\phi_{0,0}}$$

As for each row, the sum of its coefficients is 1, we can derive the value of all the transition probabilities from ρ:

$$\phi_{i,j} = \begin{cases} \dfrac{1}{1+(N-1)\rho} & \text{if } i = j \\ \dfrac{\rho}{1+(N-1)\rho} & \text{if } i \neq j \end{cases}$$

With these assumptions, we have reduced the complexity of our model to a single parameter ρ. In the next section, we will find the best value of ρ given the number of states N, the number of observations T, the probability of occurrence p and the target level of confidence α.

2.2 Tuning the Hidden Markov Model

The Baum-Weich algorithm [12] provides the most likely set of underlying probabilities, given the parameters of the model and a sequence of observations. We will use this algorithm to detect change points, when the most probable state contains at least two different states.

2.3 Mathematical interpretation

Influence of ρ on the detection of change points Our Hidden Markov Model has been reduced to a single parameter ρ. It important to first understand how this parameter will impact the behavior of the HMM. When ρ increases, the relative probability of changing state increases, so the HMM is more likely to detect a change point. Our goal is to find the maximum value of ρ that detects a change point with a probability at most α.

Influence of N on ρ To determine the most likely sequence of underlying states, the Baum-Weich algorithm does a forward pass during which it computes the most likely state up to the current observation, and then a backward pass where it updates the likelihoods that could have lead to the final state.

This can be seen as a trade-off between two opposite "energies": changing the underlying state from i to j costs $\mathcal{E}_{i\to j} \sim \frac{\phi_{i,j}}{\phi_{i,i}}$, while staying in the current state i costs $\mathcal{E}_{i\not\to j} \sim \frac{\overline{p_i}}{\overline{p_j}}$ where $\overline{p_k}$ represents the probability of predicting the wrong state ($1 - p_k$ if the even occurred, p_k otherwise). The Baum-Weich algorithms finds the sequence of states that minimizes this energy: it is only worth changing state if many future observations are very unlikely the current state i.

Using our assumptions in 2.1, these energies become $\mathcal{E}_{i\to j} \sim \rho$ and $\mathcal{E}_{i\not\to j} \sim \frac{\overline{p_i}}{\overline{p_j}}$. Note that these energies are now independent of N: the only impact is that, for small values of N, all probabilities p_i and p_j will not be represented. In other words, there might not be the state corresponding to the actual underlying probability. However, for sufficient large value of N, ρ is independent of N.

The choice of ρ as the single parameter to tune the model was not arbitrary: it is independent on the number of states N. Figure 2.1 shows an example using Monte Carlo method with $n_{tries} = 1000$ tries, where

we approximate the optimal ρ with a precision of 0.001. In the rest of this report, we will use $N = 101$ states, so that the states represent $0\%, 1\%, 2\%, \ldots, 100\%$.

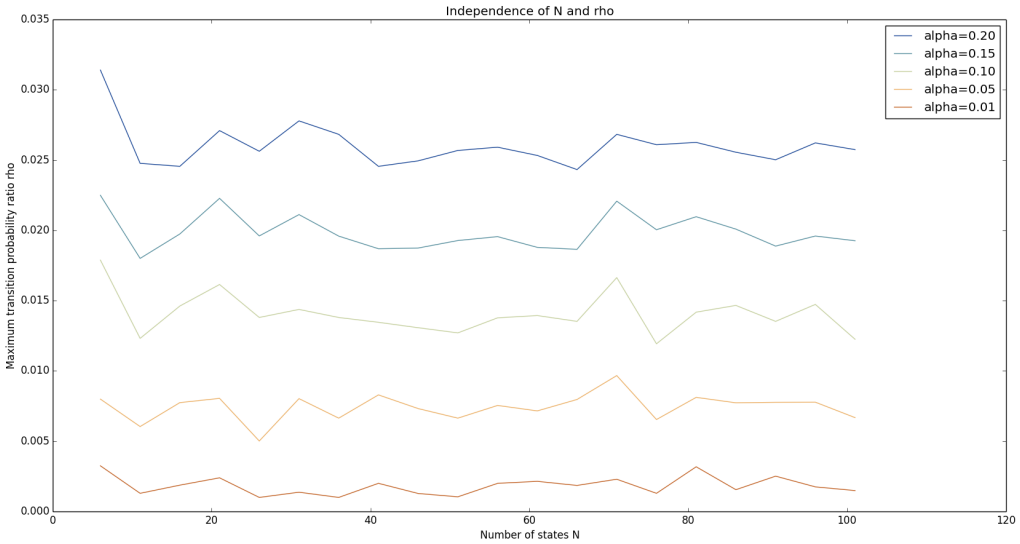

Figure 2.1: Maximum transition probability ratio ρ that does not detect a change point, for different levels of confidence α
For $p = 20\%$, $T = 100$, $n_{tries} = 1000$, $precision = 0.001$

Influence of α on ρ We expect ρ to be an increasing function of the level of confidence α: if we loosen our restrictions and accept a higher risk of erroneously detecting a change, then we can increase the transition probability ratio. When plotting some examples **??**, it seems clear that for small values of α, there is a linear correlation between α and ρ.

2.3.1 Application to events derived from a constant probability

2.3.2 Detection of a single change point

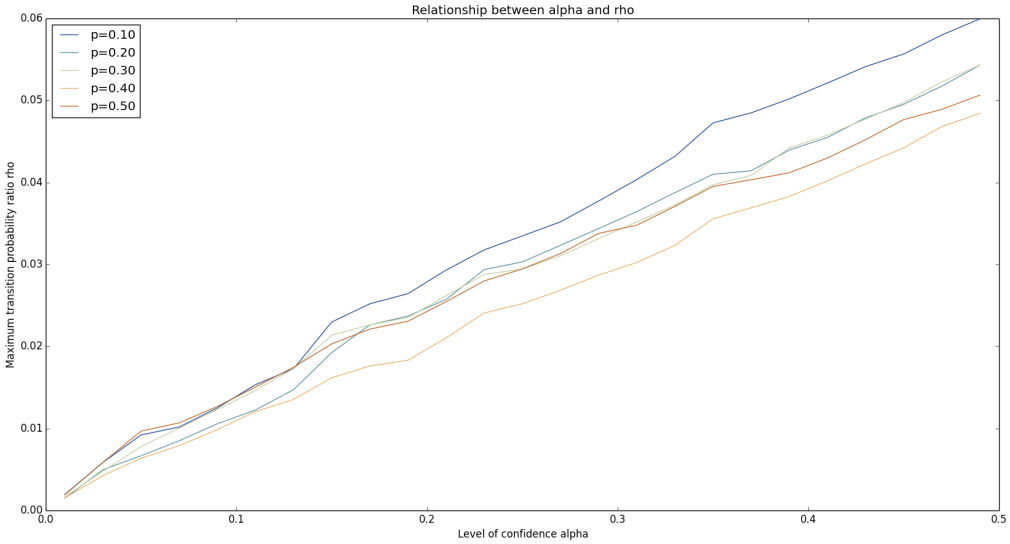

Figure 2.2: Maximum transition probability ratio ρ that does not detect a change point, for different probabilities of occurrence p
For $N = 101$, $T = 100$, $n_{tries} = 1000$, $precision = 0.001$

Chapter 3

Binary Segmentation Model

Chapter 4

Model comparison and Discussion

4.1 Detection of single change points

4.2 Generalization to multiple change points

4.3 Efficiency and generalization to an online algorithm

Chapter 5

Conclusion

5.1 Discussion

5.2 Future Work

Bibliography

[1] T. Baignères, P. Junod, and S. Vaudenay, "How far can we go beyond linear cryptanalysis?", *Advances in Cryptology - ASIACRYPT 2004, 10th International Conference on the Theory and Application of Cryptology and Information Security*, Lecture Notes in Computer Science, vol. 3329, pp. 432–450, 2004.

[2] K. S. Lomax, "Business failures: Another example of the analysis of failure data", *Journal of the American Statistical Association*, vol. 49, no. 268, pp. 847–852, 1954. DOI: 10.1080/01621459.1954.10501239.

[3] F. Proschan, "Theoretical explanation of observed decreasing failure rate", *Technometrics*, vol. 5, no. 3, pp. 375–383, 1963. DOI: 10.1080/00401706.1963.10490105.

[4] S. C. Saunders and J. M. Myhre, "Maximum likelihood estimation for two-parameter decreasing hazard rate distributions using censored data", *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 664–673, 1983. DOI: 10.1080/01621459.1983.10478027.

[5] W. Weibull, "A statistical distribution function of wide applicability", *Journal of Applied Mechanics*, vol. 18, pp. 293–297, 1951.

[6] T. L. Anderson, *Fracture Mechanics: Fundamentals and Applications, 3rd Edition*. Taylor & Francis Ed., 2005.

[7] L. Wasserman, *All of Statistics : A Concise Course in Statistical Inference*. Springer-Verlag, 2004, pp. 179–203.

[8] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Litt, "A comparison of waveform fractal dimension algorithms", *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 2, pp. 177–183, 2001, ISSN: 1057-7122. DOI: 10.1109/81.904882.

[9] Z. Harchaoui and C. Lévy-Leduc, "Multiple change-point estimation with a total variation penalty", *Journal of the American Statistical Association*, vol. 105, no. 492, pp. 1480–1493, 2010. DOI: 10.1198/jasa.2010.tm09181.

[10] B. Jackson, J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. San, L. Tan, and T. T. Tsai, "An algorithm for optimal partitioning of data on an interval", *IEEE Signal Processing Letters*, vol. 12, pp. 105–108, Feb. 2005. DOI: 10.1109/LSP.2001.838216. eprint: math/0309285.

[11] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost", *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012. DOI: 10.1080/01621459.2012.737745.

[12] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains", *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1554–1563, Dec. 1966. DOI: 10.1214/aoms/1177699147.

[13] T. Chis and P. G. Harrison, "Adapting hidden markov models for online learning", *Electronic Notes in Theoretical Computer Science*, vol. 318, pp. 109 –127, 2015, Twenty-ninth and thirtieth Annual UK Performance Engineering Workshops (UKPEW), ISSN: 1571-0661. DOI: https://doi.org/10.1016/j.entcs.2015.10.022.