Degree project specification: Estimating the probability of event occurrence

ALEXANDRE GUINAUDEAU

Master in Computer Science
Date: June 5, 2018
Supervisor: Pawel Herman
Examiner: Hedvig Kjellström
School of Electrical Engineering and Computer Science

Degree project specification: Estimating the probability of event occurrence

Alexandre GUINAUDEAU

June 5, 2018

1 Background & objective

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.

2 Interest and assignment

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.

3 Objective

The desired outcome of the degree project is to detect change points in the probability of occurrence of events. 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.

4 Research question & method

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 & expected scientific results First, we will generate sequences of probabilities of occurrence in a *realistic* way (See 6), and derive events based on those probabilities. Then, we will try to estimate the underlying probabilities based on the events. Once this is set up, define a metric to measure the error in the estimation of the underlying sequence of probabilities, and compare different candidate models to find the one that performs best.

Depending on the complexity of the data, we will either use the mathematical definition or approximate it using Monte Carlo methods to evaluate the performance of models.

As we will generate the data, it is important to clearly define the assumptions. Otherwise, the results will be biased by the way we generated the data and will not correctly evaluate the models. We will generate probabilities of occurrences that either change abruptly, for instance when a part breaks, or that slowly increase, to model the natural deterioration of a part (as explained in 6).

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 99%). 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 this, 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

5 Evaluation

To measure the accuracy of our estimations, we can use the $\mathcal{L}_2$ score to measure the error with the true underlying distribution. We will also measure the error in the detection of a change point, to determine whether we are able to pinpoint the factor that triggered the change.

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.

6 Pilot 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.

7 Conditions & schedule

Limitations We will only work on generated data and therefore we will not confirm the validity of the results on real world data.

Collaboration with principal supervisor My supervisor will mostly contribute in general thesis and problem definition.

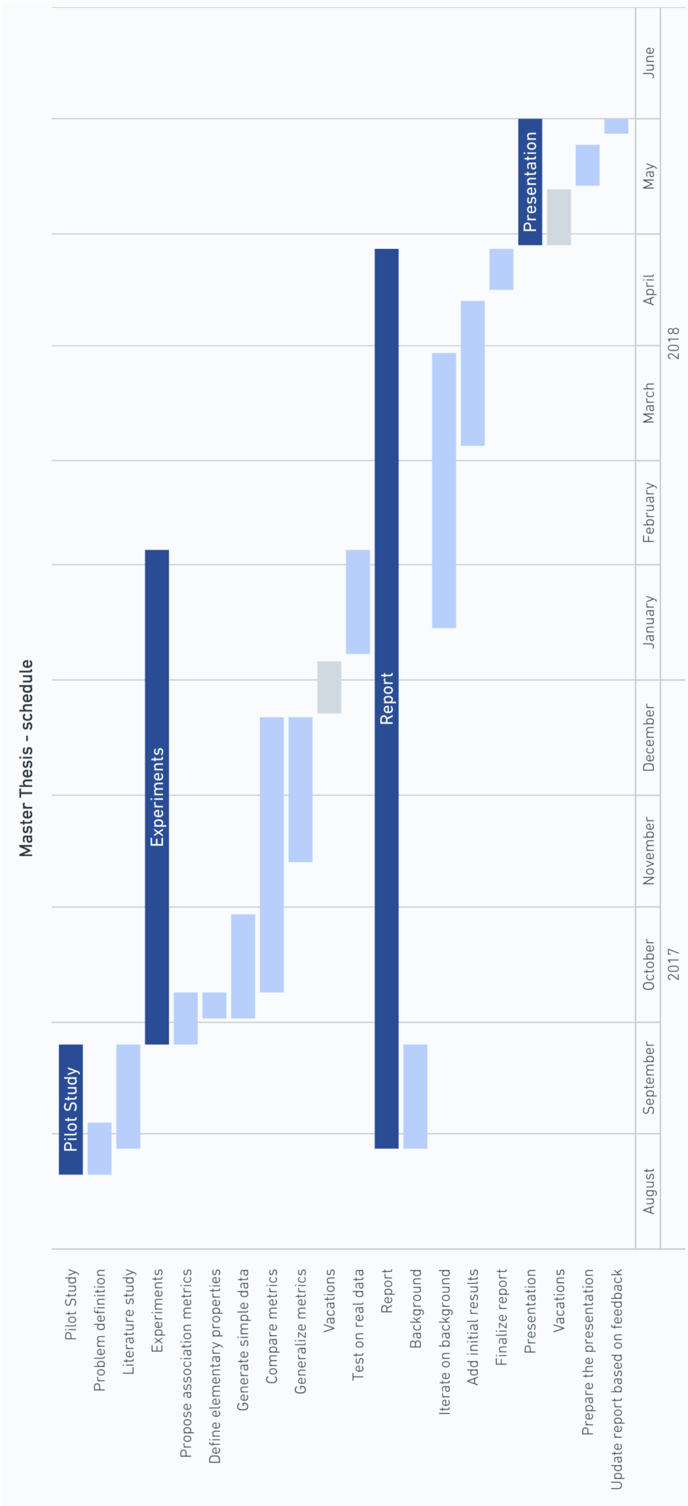

Figure 1: Schedule

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.