Prof. Krzysztof Patan, Ph.D, D.Sc.
Institute of Control and Computation Engineering
University of Zielona Góra, Poland

Zielona Góra, May 10, 2021

Review

of the doctoral thesis

Statistical reasoning of fault occurrences in industrial applications

by Tian Cong

1 Legal ground

The review was prepared at the request of the Discipline Council of Automation, Electronics and Electrical Engineering of 4 February 2021. The doctoral dissertation is conducted in automatic control and robotics. However, according to the classification provided in Regulation of Ministry of Science and Higher Education in Poland of 20 September 2018 (Dz. U. 2018 poz. 1818) the doctoral dissertation is conducted in the engineering and technology field in the discipline automation, electronic and electrical engineering.

2 Scientific area of the dissertation

Nowadays, we are witnessing a rapid development of industrial plants and production lines. Such systems should satisfy very high reliability as well as safety standards. Unfortunately, as industrial installations become more and more complex they are vulnerable to faults, malfunctions and unexpected working modes. Therefore, fault diagnosis plays a crucial role when dealing with industrial processes. An early detection of a fault renders it possible to undertake the proper preventive actions so as to avoid installation damage, danger to people life and high repair costs.

The dissertation is devoted to the application of the statistical diagnostic methods to process condition monitoring (PCM). The author established a goal of developing monitoring algorithms which can handle a different range of system complexity and improve the fault detection performance including short detection time and as small false alarm rate as possible. To realize thesis objectives the author investigated the following problems: (i) Binary Classifier for Fault Detection (BaFFle) for monitoring a single-mode processes, (ii) Dirichlet Process-Gaussian Mixture Models (DP-GMMs) to automate data clustering without the knowledge about the number of classes, (iii) Field Kalman Filter (FKF) to cope with various operating modes of the process.

Therefore, the subject discussed in the dissertation is actual and very important. Any approach aimed at shortening the detection time and at the same time reducing the false and missed alarms is worthy of investigation. Moreover, such solutions make it possible to increase system reliability and safety.

The subject discussed in the dissertation is strictly connected to the **automation, electronics and electrical engineering** field and **automatic control and robotics** area.

3 Dissertation arrangement

The dissertation was prepared in English and contains also a summary in Polish. The dissertation consists of 113 pages and contains 7 main chapters, the bibliography with 238 positions and 2 appendices.

In Introduction the author very shortly describes the problem of process condition monitoring and then gives a list of new monitoring approaches investigated in the thesis. Introduction presents also a structure of the dissertation.

Chapter 2 introduces process condition monitoring defining its main tasks and characteristics. The chapter constitutes a review of different methodologies of process modelling providing their advantages and shortcomings. Next, the author describes multimode processes and defines problems emerging from analysis of data representing such processes. Moreover, this chapter discusses also the problem of decision making in multimode processes.

The third chapter deals with a heuristic algorithm which is called Binary Classifier for Fault Detection. The proposed approach is intended for use in cases of single-mode industrial processes. As the BaFFle algorithm is strongly dependent on the probability distribution function some density estimation methods are also discussed.

The next chapter presents methods for data clustering. The problem discussed in this chapter follows from the fact that in the monitoring plant production demands as well as loading conditions can vary in time. In result we can observe various operating modes of the plant. To proper data partition the author proposes to use a fusion of Dirichlet process and gaussian mixture models (GP-GMM). Implementation of GP-GMM to cluster data recorded in multi-mode simulation model is also illustrated.

Chapter 5 deals with field Kalman filter and its application to process monitoring. Clearly, the chapter proposes training of FKF by means of multivariate autoregressive state-space (MARSS) models. Then the proposed approach is tested on the example of multi-mode processes.

The following chapter shows the application of BaFFle to fault detection and FKF to mode identification and anomaly detection. All simulation studies are carried out using PRONTO benchmark.

The last chapter constitutes the summary and provides the contribution of the thesis to the area of automation, electronics and electrical engineering. The author defines also the future research directions.

In general, the overall structure of the dissertation seems to be proper. However, in my opinion the the first chapter should be thoroughly rearranged. The thesis should contain the so-called state-of-the-art in the fild of PCM. Based on the rigorous analysis of the existing approaches the main statement (or the main goal) of the thesis should be provided. This is of a crucial importance as through the material presented in the thesis the author should exhibit that the raised problem has been demonstrated and proven. Unfortunately, the thesis includes subobjectives only. Some kind of state-of-the-art is proposed in Section 2.2 but in the context of process modelling only. Moreover, the sections included in Chapter 2 are loosely connected. Finally, Chapter 6 should be much more comprehensive. There is a lack of results showing the work of the clusterization algorithm. Moreover, some aspects such as update of control limits of the BaFFle algorithm as well as FKF models developing should be provided in detail as these elements constitute the main research achievements of the thesis. Finally, it should be pointed out that the summary is not written in a proper Polish.

4 Dissertation contribution

The main objective of the thesis was to develop monitoring algorithms that can handle different range of system complexity as well as improve the detection performance. To realize this objective the author decided to consider statistic-based methods. The crucial problem investigated in the thesis is to cope with both the so-called single- and multi-mode processes. Developed approached were tested using PRONTO benchmark. The dataset was acquired from sensors distributed across a pilot-scale multiphase flow facility. Contribution of the dissertation to the field of **automation, electronics and electrical engineering** (formerly **automatic control and robotics**) can be listed as follows:

- to develop BaFFle algorithm with adaptive control limits to process monitoring and fault detection,
- to apply PCA to extract uncorrelated features from multivariate data in the context of BaFFle,
- to employ kernel density estimation to cope with nongaussian distributions in the framework of BaFFle,
- to analyse the impact of the initial hyperparameter setting of DP-GMM models on clusterization results,
- to propose a method for selecting κ_0 in the framework of DP-GMM,
- to develop FKF monitoring model for multimode processes by means of efficient MARSS approach,
- to propose a novel monitoring indicator for anomaly detection via FKF,
- to verify the proposed approaches using PRONTO benchmark data.

Undoubtedly, the dissertation contains both the theoretical contribution as well as application studies. Tian Cong demonstrated the ability to properly formulate research problems and solve them using suitable methods and accessible tools.

Tian Cong co-authored 6 research publications including 3 journal papers and 3 conference proceedings. All papers are indexed in the outstanding Web of Science (WoS) database. What is even more noteworthy, two papers has been published in the renowned and outstanding research journals, i.e. *Journal of Process Control* (IF=3.624) and *IEEE Transactions on Control Systems Technology* (IF=5.312). Declared contribution of the author varies from 10% to 90%. The average contribution is equal to 55% which can be rated as significant contribution.

5 Remarks and comments

In spite of the number of results reported in the dissertation some questions arise, among many the main drawbacks can be listed as follows:

1. Chapter 2 does not include the state-of-the-art regarding PCM. Then, we do not know what are limitations of the existing approaches. It is also unclear why did the author decide to consider statistical methods for process monitoring. What is the novelty of the dissertation contrary to the existing solutions used in real-engineering practice?

2. In Abstract the author presents main achievements of the thesis in the form of the flowchart. Analysing this flowchart I wonder what a mechanism is used to decide if we cope with a single-mode or multi-mode process. Could the author propose some mechanism, maybe of a heuristic nature, to distinguish the mode of operation?

3. BaFFle is strongly dependent on the form of a probability distribution function. How did the author perform the test for normality?

4. If we look at the update rules of monitoring models (the page 25) we observe that the confidence level for the system alert $\alpha_{\{1,j\}}$ is not updated at all, why? Moreover, please comment how one can optimally select the parameters λ, $s_{\uparrow}$ and $s_{\downarrow}$ for a specific process considered. What conditioning should we take into account?

5. To improve clustering performance the author proposes to set $\kappa_0 = \frac{1}{N}$ where N is the number of samples in the dataset. However, in case of a large dataset the value of κ_0 will be very small and consequently clusters of a very small spread will appear. How to prevent the situation that a cluster covers a very small number of samples or even one sample?

6. The author proposes the monitoring indicator L_t in the form (5.23). This indicator is compared with the level L_{LML} in order to carry out anomaly/fault detection. The author proposes to set L_{LML} on the value equal to 5th percentile of L_t, but why? How is this choice motivated? How to select this level in an optimal manner as to guarantee the best performance of fault detection.

Detailed remarks:

1. Both, a list of abbreviation as well as a nomenclature used are highly desirable.

2. Subsection 2.1.1, the author considers two commonly used approaches to PCM namely corrective maintenance and preventive maintenance. What about the remaining useful life approach in this context?

3. Section 2.2. There is a lack of rigorous analysis of the approaches presented therein. Advantages and drawbacks of the listed solutions should be provided.

4. The equation (2.1) does not represent a single mode process but gives a condition which should be satisfied by a process variable. Moreover t_0 as well as Δt are not explained. Finally, the equation is wrong as for $t = t_0$ is a division by zero!

5. In Subsection 2.3.1 there is no definition of a multi-mode process.

6. Subsection 2.3.3, it is not clear how presented methods can be used in the context of multi-mode processes. I guess that a method should point out a number of working modes of a process. Is it not?

7. In fact the model-based monitoring indices are not provided. The author should consider indices such as the detection time, the time of fault isolation, the false alarm rate, the fault size, etc.

8. Table 3.1 is simply a confusion matrix.

9. Equation (3.2) represents the index known as the specificity (not specification).

29. Please comment performance of the proposed clusterization method in cases when data is compatible with super- or sub-Gaussian distribution.

30. Subsection 4.6.3. In the experimental study clusters are rather easy to discriminate. The author should present a study with overlapping or surrounding clusters.

31. Figure captions and table headers are too wast. For clarity of the presentation it is recommended to discuss the results in the text.

32. Page 57, the second line, it should be "w.r.t. θ of appropriate dimensions...".

33. Subsection 5.2.2. From this section it does not follow what challenges has been considered in the dissertation.

34. In Section 5.3 the author considers the linear model only, why? What about nonlinear processes and nonlinear identification methods?

35. Subsection 5.3.1, I do not understand the wording "local dynamics existing in the healthy data". As regards of changing dynamics of the plant in time, multiple system models can be used both linear and nonlinear, however linear models are not at all better choice than nonlinear ones.

36. Autoregressive (AR) models are models without the external excitation. Then if such a model is converted to the state-space form we receive the so-called autonomous system described by:

$$x[t] = Ax[t-1]. \tag{1}$$

How is the model (1) related to FKF presented in Section 5.1?

37. I completely do not understand the meaning of (5.8). Intuitively

$$x(\theta_j)[t] = x_j[t], \quad i = 1, \ldots, k.$$

Is that right?

38. The FKF (5.5) is already discretised. Then, I do not catch the discretisation introduced in (5.12). The FKF (5.12) is rather a multi-mode version of FKF (5.5).

39. In simulation case studies it should be clearly pointed out what do Variable 1 and Variable 2 mean (Fig. 5.2). Moreover, some quality index is welcome here showing the modelling quality.

40. Subsection 5.6.2. Specification of faults would be more detailed. For example, what is the fault type (abrupt/incipient)?, What is the fault size? What is the fault type (multiplicative/additive)? What is the time a fault was introduced?

41. There is no fault detection in case of multivariate system (page 71).

42. It is not clear what anomaly is considered in the subsection 5.6.3.

43. Subsection 6.1.2, it should be "... developing faults...".

44. In subsection 6.1.2 the author introduces the so-called developing faults. However, from Fig. 6.4 it is obvious that the author considers a typical abrupt fault with intensity changing in time. The same behaviour of the detection system can be achieved introducing abrupt change equal to 60% of valve closing.

10. Interesting measure widely used in pattern recognition and classification tasks is F_1 score defined as follows:

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}.$$

11. Section 3.2. A correlation between process variables can be nonlinear in nature. Maybe a better solution is to use a nonlinear version of PCA, e.g. some kind of autoencoder. The autoencoder in the form of a neural network can easily adapt to the changing operating condition of the plant.

12. In (3.4) the symbol $\hat{\sigma}$ is not defined.

13. Line just after the equation (3.4), instead of "deviation" it should be "standard deviation".

14. Mathematical symbols should be unified to facilitate reading of the material concerning ED and SVD, e.g. there is no matrix V in the SVD description, however later on in this subsection the author uses the matrix V defining the projection.

15. Line just after (3.12) $\tilde{X}^{\top}$ is projected to v-dimensional space (not in v-dimensional space).

16. Equation (3.13), a symbol n in the denominator is not explained.

17. Equation (3.17), σ_j is not defined.

18. Equation (3.25) is not correct as it does not consider the case when $n_{D(j,i)=1} = n_{D(j,i)=0}$. What does it mean that $\boldsymbol{y}_t$ is undetermined. Please explain this from the computer implementation point of view. Simply, in a computer program if $n_{D(j,i)=1} = n_{D(j,i)=0}$, $D_{\boldsymbol{y}_t}$ is not set to a new value.

19. Line after (3.25), it should be "... 0 and 1 respectively represent normal and abnormal behaviour".

20. Mechanism represented by (3.25) is not robust to external disturbances or transient behaviour of the system. Decision making should be determined as a trade-off between the fault detection sensitivity and the false alarm rate. In order to avoid false alarms the decision making can be extended, e.g. with moving average or moving window mechanism. Then we are able to check if the conditions provided in (3.25) persist for some time.

21. Introduction to Section 4 should provide existing clusterization methods, e.g. k-means, self-organizing maps, fuzzy c-means, etc.

22. Paragraph **Binomial distribution**. The proper wording would be "Let x_i be".

23. In literature one can find a plenty of methods to estimate infinite GMMs namely a) multivariate distribution, b) inference methods, c) predictive distributions (see Neal (2000), Rasmussen (2000)). Why did the author decide to use Gaussian Process with Chinese Restaurant Process?

24. The title of Section 4.2 should be "Preliminaries".

25. Equations (4.29), (4.34)–(4.38) go beyond the text area.

26. Pages 34 and 39, missing brackets around the equation number.

27. Beginning of Section 4.6. What four parameters are we talking about?

28. Page 45, the first line of the paragraph **Default parameter setting**, superfluous dot after "matrix".

45. Illustration of LCL and UCL presented in Figs 6.2 and 6.3 are useless as both control limits are adaptive and depend on α.

46. To definitely evaluate the performance of BaFFle method it should be compared with other statistical methods, e.g. cumulative sum algorithm (CUSUM) and its variations, generalized likelihood ratio (GLR), etc. For reference see Baseville, M, Nikiforov I.V. Detection of Abrupt Changes: Theory and Applications, Prentice-Hall, 1993.

47. Fig. 6.4 clearly shows that BaFFle generates relatively large number of false alarms, see my previous remark 20 on this problem.

48. Table 6.3. To facilitate analysis of the achieved results the best values should be emphasized, e.g. using a bold face or frameboxes. The same problem is observed in Table 6.4.

49. Section 6.3, at the beginning of the second line it should be "modes".

50. Figures 6.6 and 6.7 are not very readable. The variable ranges are not portrayed.

51. Determining the number of working modes is not illustrated at all (Subsection 6.3.1). What is the reason of introducing the material in Chapter 5?

52. The method named FGMM-BIB is not presented.

53. How was L_t calculated?

6 Conclusions

In my opinion the subject presented in the dissertation is actual and achieved results interesting. Taking into account original research achievements as well as the fact that all critical remarks have arguable character, I declare the following:

1. The dissertation *Statistical reasoning of fault occurrences in industrial applications* by Tian Cong meets the requirements for doctoral dissertations imposed by legal act of scientific degrees and professor degree of 14 March 2003 (Dz.U. nr 65, poz. 595) with subsequent changes.

2. I request to accept the doctoral dissertation *Statistical reasoning of fault occurrences in industrial applications* by Tian Cong and to allow it to the public defence against the Discipline Council of Automation, Electronics and Electrical Engineering of University of Science and Technology in Kraków.

Professor Krzysztof Patan