Stochastic Conformance Checking based on Variable-length Markov Chains: on metrics for probabilistic languages in Process Mining

Emilio Incerto 🖂
IMT School for Advanced Studies Lucca, Italy

Andrea Vandin 🖂
Sant'Anna School of Advanced Studies Pisa, Italy

Sima Sarv Ahrabi 🖂
Sant'Anna School of Advanced Studies Pisa, Italy

Abstract

We present a work recently appeared in the journal Information Systems on the use of techniques from Markov chain learning to the problem of Stochastic Conformance Checking in Process Mining. The connection with the workshop is that we see stochastic conformance checking as a possible application domain for behavioural metrics and quantitative logics.

2012 ACM Subject Classification Theory of computation → Timed and hybrid models

Keywords and phrases Stochastic conformance checking, Earth Movers' Stochastic Conformance, Higher-order Markov chains

Funding *Andrea Vandin*: The work has been partially supported by project SMaRT COnSTRUCT (CUP J53C24001460006), in the context of FAIR (PE0000013, CUP B53C22003630006) under the National Recovery and Resilience Plan (Mission 4, Component 2, Line of Investment 1.3) funded by the European Union - NextGenerationEU.

Acknowledgements We thank Andrea Burattin, DTU Denmark, for discussions and precious suggestions in several phases of the preparation of this work.

Extended Abstract

Process Mining (PM) is an interdisciplinary research area with the goal of extracting insights and knowledge from execution traces of a process, bridging the gap between data science and process science [27]. PM consists of a wide range of techniques structured in three macro domains: process discovery, process enhancement, and conformance checking. Process discovery is about learning a graphical representation of a process starting from logs of its executions. Process enhancement regards enriching a model with additional information, such as the frequency of executed activities or paths. Lastly, conformance checking is a key problem of PM, enabling the identification, quantification, and analysis of deviations among reference and mined processes [3]. Several proposals in PM focus on the stochastic nature of the studied process (see, e.g., [12, 13, 25], just to cite a few). However, usually PM, and conformance checking techniques in particular, do not focus on the stochastic aspects of the studied process, and consider qualitative models. In other words, most of the conformance checking techniques ignore the stochastic perspective of the process model (see, e.g., the discussion in [21]). Recently, there has been a growing interest towards *stochastic conformance checking* (SCC, see, e.g., [16, 22, 14]). These are approaches to conformance checking that emphasize stochastic aspects of the underlying process, like the frequency and probability of traces, and consider quantitative models. The most recent among these approaches are based of stochastic distances like the famous Earth Mover's Distance (EMD, also known as Wasserstein distance) [24]. There exist SCC measures based on it named Earth

Movers' Stochastic Conformance (EMSC), or unit EMSC (uEMSC) [16, 14]. All approaches to SCC start from a reference model and from a group of traces. Both are transformed into stochastic languages (i.e., traces and their probabilities). To *compare* the two obtained stochastic languages, stochastic variants of the EMD distance are computed among them. This distance is then ultimately used to establish the conformance of the group of traces to the model.

Over the years, the *software performance engineering* (PE) community developed techniques for synthesizing Markovian models that accurately describe the stochastic process underlying programs (see., e.g., [9, 5, 1, 20, 8, 18, 19, 7, 10]). However, surprisingly, stochastic conformance checking is not central in PE.

In a recent paper [11], we proposed to the PM community a novel approach to SCC based on PE techniques developed for synthetizing Markovian models. Given a log, we mine, or learn, a Variable-length Markov Chain (VLMC, higher-order Markovian models equipped with memory) [9]. Intuitively, VLMCs are Markov chains which use memory, partially departing from the *memory-less* property. For each trace, they build a trace-specific dependency on the previous events (the *memory*). This is used to compute a precise probability distribution for the next event. For this reason, VLMCs are well-suited for compactly expressing complex path dependencies in the process. In [11], we further equip VLMCs with a method for stochastic conformance checking. We do this using the VLMC notion of *likelihood* of a trace in a discovered stochastic process, that is, the probability for the model to generate that trace. Taking inspiration from the SCC literature (e.g., [17, 15]), we use likelihood to do SCC of a log against a model. In particular, we use the conformance measure uEMSC, standard in the SCC literature. Our method is therefore an innovative approach to stochastic conformance checking. Our method is accurate: it gives high uEMSC values, often close to 1, for logs conformant to models.

Our claims are supported by a rich experimental evaluation in [11]. We considered 11 benchmark datasets from the PM literature, and 18 competitor SCC techniques. In particular, we used all datasets considered in [17], a paper presenting a previous approach to SCC. Furthermore, we benchmark with all the 15 SCC techniques considered in [17], and with the 3 additional ones considered in [15]. The results in [11] clearly show that our approach outperforms all 18 competitor SCC techniques in terms of uEMSC values on 10 out of 11 datasets. That is, we get uEMSC values closer to 1. Such good performances may be due to the fact that all the considered competitor techniques are actually combinations of a qualitative discovery step, to mine the structure of a (qualitative) model, followed by a stochastic step where weights are assigned to the qualitative model to make it stochastic. Instead, our approach is natively stochastic: we directly learn a stochastic model (a higher-order Markov Chain). Another reason may be connected to the use of memory, which is central in the analysis of stochastic models in several domains (see, e.g. [23, 6, 2]). In fact, it allows to handle issues connected to the so-called phenomenon of path dependency [4, 26, 28]. Nevertheless, none of the considered competitor approaches is based explicitly on memory. In the paper we also perform a preliminary study of the impact of noise in traces, showing a decrease in performance linear in the amount of noise in the dataset, and sketch an extension mitigating this effect.

Relation to BMQL (Behavioural Metrics and Quantitative Logics).

We see stochastic conformance checking as a possible application domain for behavioural metrics and quantitative logics. Indeed, SCC aims at measuring the behavioural dissimilarities between systems (a log and a model, two logs, two models). In particular, SCC aims at estimating such dissimilarities from observations of the system. However, the metrics used for SCC are still somehow *simple*. In particular, the mentioned conformance meausre uEMSC simply compares the likelihood of traces. In particular, uEMSC is computed by overlappings of probability mass between the stochastic language of an event log and that of a model. The uEMSC sums up the (positive) difference between the probability of each trace in the log (the frequency of the trace in the log), and its probability in the stochastic process model M [14] (the likelihood of the trace in the model):

$$\text{uEMSC}(L, M) = 1 - \sum_{\sigma \in L} \max(L(\sigma) - M(\sigma), 0)$$

where $L(\sigma)$ is the probability of the trace σ in the log L, while $M(\sigma)$ is the probability of the trace as computed by the model.

Studies of SCC involving actual behavioral metrics are still missing. We believe that this is due to the fact that stochastic extensions to conformance checking are still quite recent. The aim of this talk is to introduce the SCC problem to the community of behavioural metrics and quantitative logics, so to foster SCC discussions and research in that direction.

References

1 Simonetta Balsamo, Antinisca Di Marco, Paola Inverardi, and Marta Simeoni. Model-based performance prediction in software development: A survey. 30(5):295–310, 2004.

2 Anne-Laure Bianne-Bernard, Fares Menasri, Laurence Likforman-Sulem, Chafic Mokbel, and Christopher Kermorvant. Variable length and context-dependent hmm letter form models for arabic handwritten word recognition. In *Document Recognition and Retrieval XIX*, volume 8297, pages 52–59. SPIE, 2012.

3 Josep Carmona, Boudewijn van Dongen, and Matthias Weidlich. *Conformance Checking: Foundations, Milestones and Challenges*, pages 155–190. Springer International Publishing, Cham, 2022. `doi:10.1007/978-3-031-08848-3_5`.

4 Alexandre Checoli Choueiri and Eduardo Alves Portela Santos. Discovery of path-attribute dependency in manufacturing environments: A process mining approach. *Journal of Manufacturing Systems*, 61:54–65, 2021.

5 Vittorio Cortellessa, Antinisca Di Marco, and Paola Inverardi. *Model-Based Software Performance Analysis*. Springer, Berlin, Heidelberg, 2011.

6 Antonio Galves and Eva Löcherbach. Infinite systems of interacting chains with memory of variable length—a stochastic model for biological neural nets. *Journal of Statistical Physics*, 151:896–921, 2013.

7 Giulio Garbi, Emilio Incerto, and Mirco Tribastone. Learning queuing networks by recurrent neural networks. In *Proceedings of the ACM/SPEC International Conference on Performance Engineering*, pages 56–66, 2020.

8 Carlo Ghezzi, Mauro Pezzè, Michele Sama, and Giordano Tamburrelli. Mining behavior models from user-intensive web applications. In *Proceedings of the 36th International Conference on Software Engineering*, pages 277–287, 2014.

9 Emilio Incerto, Annalisa Napolitano, and Mirco Tribastone. Statistical learning of markov chains of programs. In *28th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, MASCOTS 2020, Nice, France, November 17-19, 2020*, pages 1–8, USA, 2020. IEEE. `doi:10.1109/MASCOTS50786.2020.9285947`.

10 Emilio Incerto, Annalisa Napolitano, and Mirco Tribastone. Learning queuing networks via linear optimization. In *Proceedings of the ACM/SPEC International Conference on Performance Engineering*, pages 51–60, 2021.

11 Emilio Incerto, Andrea Vandin, and Sima Sarv Ahrabi. Stochastic conformance checking based on variable-length markov chains. *Inf. Syst.*, 133:102561, 2025. URL: https://doi.org/10.1016/j.is.2025.102561, doi:10.1016/J.IS.2025.102561.

12 Sander J. J. Leemans, Tian Li, Marco Montali, and Artem Polyvyanyy. Stochastic process discovery: Can it be done optimally? In *Advanced Information Systems Engineering: 36th International Conference, CAiSE 2024, Limassol, Cyprus, June 3–7, 2024, Proceedings*, page 36–52, Berlin, Heidelberg, 2024. Springer-Verlag. doi:10.1007/978-3-031-61057-8_3.

13 Sander J. J. Leemans, Fabrizio Maria Maggi, and Marco Montali. Reasoning on labelled petri nets and their dynamics in a stochastic setting. In *Business Process Management: 20th International Conference, BPM 2022, Münster, Germany, September 11–16, 2022, Proceedings*, page 324–342, Berlin, Heidelberg, 2022. Springer-Verlag. doi:10.1007/978-3-031-16103-2_22.

14 Sander J. J. Leemans, Anja F. Syring, and Wil M. P. van der Aalst. Earth movers' stochastic conformance checking. In Thomas Hildebrandt, Boudewijn F. van Dongen, Maximilian Röglinger, and Jan Mendling, editors, *Business Process Management Forum*, pages 127–143, Cham, 2019. Springer International Publishing.

15 Sander J.J. Leemans, Fabrizio Maria Maggi, and Marco Montali. Enjoy the silence: Analysis of stochastic petri nets with silent transitions. *Information Systems*, 124:102383, 2024. URL: https://www.sciencedirect.com/science/article/pii/S0306437924000413, doi:10.1016/j.is.2024.102383.

16 Sander J.J. Leemans, Wil M.P. van der Aalst, Tobias Brockhoff, and Artem Polyvyanyy. Stochastic process mining: Earth movers' stochastic conformance. *Information Systems*, 102:101724, 2021. URL: https://www.sciencedirect.com/science/article/pii/S0306437921000041, doi:10.1016/j.is.2021.101724.

17 Felix Mannhardt, Sander J. J. Leemans, Christopher T. Schwanen, and Massimiliano de Leoni. Modelling data-aware stochastic processes - discovery and conformance checking. In Luis Gomes and Robert Lorenz, editors, *Application and Theory of Petri Nets and Concurrency*, pages 77–98, Cham, 2023. Springer Nature Switzerland.

18 Geoff Mazeroff, Victor De, Cerqueira Jens, Gregor Michael, and G Thomason. Probabilistic trees and automata for application behavior modeling. In *41st ACM Southeast Regional Conference Proceedings*, 2003.

19 Geoffrey Mazeroff, Jens Gregor, Michael Thomason, and Richard Ford. Probabilistic suffix models for api sequence analysis of windows xp applications. *Pattern Recognition*, 41(1):90–101, 2008.

20 Tony Ohmann, Michael Herzberg, Sebastian Fiss, Armand Halbert, Marc Palyart, Ivan Beschastnikh, and Yuriy Brun. Behavioral resource-aware model inference. pages 19–30, 2014.

21 Eduardo Goulart Rocha, Sander J. J. Leemans, and Wil M. P. van der Aalst. Stochastic conformance checking based on expected subtrace frequency. In *2024 6th International Conference on Process Mining (ICPM)*, pages 73–80, 2024. doi:10.1109/ICPM63005.2024.10680673.

22 Eduardo Goulart Rocha, Sander J. J Leemans, and Wil M.P. van der Aalst. Stochastic conformance checking based on expected subtrace frequency. In *2024 6th International Conference on Process Mining (ICPM)*, 2024.

23 Dana Ron, Yoram Singer, and Naftali Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine learning*, 25(2-3):117–149, 1996.

24 Ludger Rüschendorf. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1):117–129, 1985.

25 Arik Senderovich, Matthias Weidlich, Liron Yedidsion, Avigdor Gal, Avishai Mandelbaum, Sarah Kadish, and Craig A. Bunnell. Conformance checking and performance improvement in scheduled processes: A queueing-network perspective. *Information Systems*, 62:185–206, 2016. URL: https://www.sciencedirect.com/science/article/pii/S0306437915301095, doi:10.1016/j.is.2016.01.002.

26 Hongwei Sun, Wei Liu, Liang Qi, Xiaojun Ren, and Yuyue Du. An algorithm for mining indirect dependencies from loop-choice-driven loop structure via petri nets. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(9):5411–5423, 2021.

27 Wil M. P. van der Aalst. *Process Mining: Data Science in Action*. Springer, Germany, 2 edition, 2016. `doi:10.1007/978-3-662-49851-4`.

28 Lijie Wen, Jianmin Wang, and Jiaguang Sun. Detecting implicit dependencies between tasks from event logs. In *Frontiers of WWW Research and Development-APWeb 2006: 8th Asia-Pacific Web Conference, Harbin, China, January 16-18, 2006. Proceedings 8*, pages 591–603. Springer, 2006.