Stochastic process discovery via model inference

Pierre Cry[1]

[1]MICS Laboratory, CentraleSupélec, Université Paris-Saclay at Gif-sur-Yvette, France

Abstract

Business processes often exhibit significant variability in both structure and execution probabilities. While existing process discovery techniques focus primarily on control-flow, they typically overlook the stochastic nature of processes, limiting their ability to accurately simulate and predict real-world behaviour. My research addresses this gap by developing computationally efficient and statistically grounded methods for indirect stochastic process discovery from event logs. I propose a dual approach: (i) optimization-based discovery, where exact stochastic languages are extracted from workflow nets via log-driven unfolding of the net space and aligned to observed behaviour using distance metrics such as Kullback–Leibler divergence and Earth Mover's Distance; and (ii) statistical inference-based discovery, where probabilistic parameters are inferred using Approximate Bayesian Computation Sequential Monte Carlo (ABC-SMC), providing a posterior distribution over model weights. A third contribution focuses on extending process tree semantics to the stochastic dimension, enabling both effective optimization and subsequent exploitation by humans and computational tools. The proposed methods are validated on real-life event logs, demonstrating improved accuracy, scalability, and interpretability compared to state-of-the-art stochastic process discovery techniques. This work aims to establish a unified, reproducible framework for discovering probabilistic behaviours in business processes. This Ph.D. project began in October 2022 under the supervision of Paolo Ballarini and Pascale Le Gall at the MICS Laboratory of the engineering school CentraleSupélec in Paris, France, with special thanks to András Horváth, from the university of Turin, Italy, for his significant contributions to the majority of the work presented in this thesis.

Keywords

Stochastic Process Discovery, Stochastic Workflow Nets, Stochastic Process Trees, Model Inference, Optimization, Approximate Bayesian Computation.

1. Research question and Motivation

Over the years, data has become a central asset for understanding, analyzing, and improving complex systems. Process mining is a research field that leverages such data, specifically, event logs, to model and analyze real-world processes. An event log records sequences of events, where each event captures a specific interaction between a system and its environment. Typical event attributes include the executed activity, timestamp, resource, and contextual information.

Within process mining, process discovery focuses on constructing process models directly from event logs. These models serve both as a means to communicate and understand observed behaviour and as a starting point for more advanced analyses, such as identifying strengths, inefficiencies, and potential improvements. Beyond capturing control-flow structure, an increasingly important challenge lies in exploiting the stochastic information in event logs: the probabilities associated with different execution paths. Properly modelling this probabilistic dimension enables richer analysis, more realistic simulations, and improved predictive capabilities. In domains such as healthcare, logistics, and manufacturing, understanding what is likely to happen and how likely it is is essential for decision-making, resource allocation, and risk analysis. While stochastic modelling in process mining is not new, recent interest has surged due to the growing demand for accurate, predictive, and resource-aware models of real-life processes. However, current stochastic process discovery approaches face limitations, including scalability issues when applied to real-life logs and limited interpretability regarding the influence of each parameter on the model's stochastic semantics. Moreover, the conformance results that evaluate the quality of a model to reproduce the studied log leave room for improvement in accurately capturing

ICPM Doctoral Consortium and Demo Track 2025, October 20-24, 2025, Montevideo, Uruguay

pierre.cry@centralesupelec.fr (P. Cry)

https://pierrecry98.github.io (P. Cry)

the observed process behaviour.

My thesis addresses key challenges in discovering stochastic process models that reproduce observed behaviours both accurately and efficiently. In particular, it seeks to answer the following research questions:

1. Given a stochastic Petri net model, can we compute its corresponding stochastic language efficiently?
2. Given a *parametric* stochastic Petri net model, can we design an optimisation procedure to identify the optimal parameters with respect to a chosen stochastic conformance measure?
3. Since the exact computation of the stochastic language issued by a stochastic Petri net can be computationally intensive and thus limited by scalability issues, can we devise techniques to alleviate these limitations?

To address these questions, we propose methods that combine formal model analysis, optimisation techniques, and statistical inference to uncover the probabilistic nature of processes from event data. Section 2 situates this work within the state of the art in stochastic process discovery. Section 3 presents the three main contributions of the thesis, and Section 4 discusses the results obtained so far, outlines current limitations, and highlights directions for future work.

2. Relation to state of the art

Only a limited number of methods have been proposed in the literature for discovering stochastic process models directly from event logs. One of the earliest contributions is by Rogge-Solti et al. [1], who introduced a framework for discovering generalized stochastic Petri nets (GSPNs) enriched with generally distributed timed transitions, enabling performance analysis of the mined process.

Burke et al. [2] addressed the problem of converting a workflow Petri net, discovered using a conventional algorithm, into a stochastic workflow Petri net through weight estimation. In contrast to [1], their focus was on a subclass of GSPNs containing only immediate transitions. This means that time delays are excluded, and the models capture only the probability of traces while ignoring timestamps. They proposed six lightweight weight estimators that combine summary statistics from the log (e.g., subsequence frequencies) with statistics from the model that consider structural relationships between Petri net nodes (e.g., transition causality). Although computationally efficient, since they do not require enumerating the model language, these methods often yield suboptimal conformance in practice. In a follow-up study [3], the authors introduced a novel framework for directly discovering untimed Generalized Stochastic Petri Nets (GSPNs) from event logs using trace frequencies. Unlike previous methods that apply weights to an existing control-flow model, this approach builds both the structure and the stochastic behaviour simultaneously. The key innovation lies in the application of reduction and abstraction rules that operate on the log without relying on a pre-mined model. These rules form the core of a framework known as the Toothpaste Miner, which constructs Probabilistic Process Trees. These trees are then systematically transformed into corresponding GSPNs.

A different line of work approaches stochastic process discovery as an optimization problem. The main challenge lies in estimating the probability of each trace generated by the model, an algorithmically non-trivial task due to the potentially large or infinite size of the model's language. In [4], the authors proposed optimizing the earth mover's stochastic conformance [5] (EMSC) score of a discovered stochastic Petri net via subgradient ascent. Their method consists of two steps: (i) computing the stochastic language of the model by analysing its structure, and (ii) performing subgradient optimization on the conformance loss, propagating gradients to update the model's transition weights.

Leemans et al. [6] examined the stochastic process discovery problem from a broader perspective, emphasising how the choice of modelling formalism (representational bias) can affect the discovery of an optimal stochastic model. They showed that a perfectly fitting control-flow model can be less conformant than a less fitting one when trace probabilities are considered. In the same work, they proposed two strategies:

1. Direct discovery of an optimal stochastic Petri net from the log (one go).
2. Indirect discovery, where an existing non-stochastic model is enriched with optimal weights (two stages).

For the second approach, they introduced a weight optimisation scheme based on analytical probability expressions. Given N distinct traces in the log, these expressions are obtained from N absorbing-state probability problems on N discrete-time Markov chains, each derived from the cross-product of the model's stochastic reachability graph and a deterministic finite automaton representing a log trace. The resulting formulas are fed to an optimisation engine to maximise either the unit Earth Mover's Stochastic Conformance [5] (uESMC) or the inverted entropic relevance [7] (ER^{-1}). While effective, this approach suffers from scalability issues: the analytical expressions grow rapidly with trace length, making the method impractical for even moderately complex models.

3. Contribution and results

3.1. Stochastic process discovery via optimization

We address the problem of discovering stochastic process models by formulating them as optimization problems. Given a *parametric* stochastic Petri net, we aim to identify parameters (e.g., transition firing probabilities) that maximize the model's alignment with the observed behaviour recorded in an event log.

A central challenge is that optimizing a stochastic model requires evaluating its stochastic language, i.e., the probability distribution over all possible traces generated by the model. This computation is algorithmically non-trivial for workflow nets due to the potentially large or infinite size of the underlying language.

To address this, we proposed in [8] a log-driven *unfolding-based procedure* to compute the probability of each trace in the log without enumerating the entire model language. This approach avoids the need to derive and store large symbolic formulas, as required in previous methods, and directly exploits the structure of the model to identify relevant execution paths. We further improved this procedure by introducing memoization, preventing redundant recalculations for common prefixes shared across log traces. These enhancements yield substantial performance gains, particularly in iterative contexts such as optimization.

Once the stochastic language can be computed efficiently, we define objective functions to quantify stochastic conformance between the model and the log. In this thesis, we consider two such measures: the *Kullback–Leibler divergence* (KLD) and the *restricted Earth Mover's Distance* (rEMD), both of which compare probability distributions over traces. When the objective function is differentiable, as with KLD, we exploit analytical expressions of its derivatives to enable the use of gradient-based optimization methods such as L-BFGS-B [9] and TNC [10]. We employ derivative-free optimizers such as Powell's method [11] or Nelder–Mead [12] for non-differentiable objectives.

The resulting framework allows us to iteratively adjust the model's stochastic parameters to improve its conformance to the observed behaviour, balancing computational efficiency with conformance quality. Empirical evaluation on real-life event logs from the BPI Challenge demonstrates that this optimization-based approach can yield models with improved stochastic accuracy compared to existing state-of-the-art methods.

3.2. Stochastic process discovery via Bayesian inference

While the optimisation-based approach seeks a single set of parameters that maximises a stochastic conformance measure, a Bayesian perspective aims to estimate an entire *posterior distribution* over the model parameters given the observed event log. This allows for estimating and quantifying the uncertainty in the inferred parameters, providing deeper insights into how each parameter influences the model's stochastic behaviour.

This work [13] employs the *Approximate Bayesian Computation Sequential Monte Carlo* (ABC-SMC) algorithm to perform Bayesian inference on stochastic process models. ABC-SMC is particularly well-suited for this setting because it does not require the explicit likelihood function of the observed data under the model, which is intractable in most cases. Instead, it relies on simulating the model under candidate parameter values, comparing the resulting stochastic language to the log using a suitable distance measure, and progressively refining the parameter population across successive generations.

Our ABC-SMC framework first starts with an initial population of parameter vectors, sampled from a prior uniform distribution. For each parameter vector, the model is simulated, using a model checker, COSMOS, relying on the HASL logic to approximate its stochastic language. This simulation approach helps us deal with larger logs. A *distance function* (e.g., restricted Earth Mover's Distance) compares the simulated stochastic language with the event log. Only parameter vectors producing a distance below a predefined *threshold* are retained. The threshold is gradually reduced in subsequent generations, leading to increasingly accurate parameter estimates. New parameter vectors are generated by perturbing previously accepted ones, forming the next population generation.

The procedure's outcome is an empirical approximation of the posterior distribution over the parameters. This provides several benefits over pure optimisation as it captures multiple plausible parameter configurations rather than a single optimum and helps understand each parameter's impact on the overall model.

Experimental results on real-life BPI Challenge logs show that ABC-SMC can discover models with high stochastic conformance while offering richer interpretability and robustness than optimisation-only methods. Parallelisation of the particle search further improves scalability, making the approach applicable to realistically sized event logs.

3.3. A new formalism for stochastic process discovery: stochastic process trees

While stochastic Petri nets offer a powerful and expressive formalism for representing stochastic behaviours, they often become large and complex when modelling real-life processes, making them harder to operate and interpret. This complexity can hinder both the optimisation of their parameters and the communication of results to process analysts and domain experts.

To address these limitations, we introduce, in *stochastic process trees* (SPTs), an extension of classical process trees in which control-flow operators are enriched with stochastic semantics. Process trees provide a hierarchical and block-structured representation of processes, where each internal node corresponds to a control-flow operator (*sequence*, *choice*, *parallel*, *loop*) and leaves correspond to activities. Their structured nature ensures soundness by construction and facilitates interpretability. In our stochastic extension, each operator, but not the sequence, is associated with parameters that govern the probabilistic behaviour of its execution:

- For *choice* nodes, probabilities determine the likelihood of selecting each branch.
- For *parallel* nodes, probabilities influence the ordering of interleaved activities.
- For *loop* nodes, probabilities model the likelihood of repeating the loop body versus exiting it.

We formalise the semantics of SPTs by defining their associated *stochastic language*, extending the deterministic process tree semantics to account for execution probabilities. This formalisation enables a direct computation of the probability of any tree-generated trace, either by simulating the tree or by a log-driven exploration of it and applying optimization techniques to fit the tree's parameters to event log data.

The SPT formalism is particularly well-suited for parameter discovery, as the block-structured representation reduces the number of stochastic parameters. Moreover, its interpretability makes it a practical choice for interactive analysis, allowing analysts to understand not only the structure of the process but also the likelihood of alternative behaviours.

We demonstrate the applicability of SPTs in both optimisation-based stochastic discovery settings, showing that they can achieve competitive conformance while offering significant advantages in terms of scalability, interpretability, and ease of use.

4. Conclusion and Future works

In the context of this Ph.D. project, we have developed different approaches aimed at providing answers to the research questions outlined in Section 1, namely: (1) the efficient computation of the exact stochastic language issued by a stochastic workflow net, (2) the optimisation of stochastic nets with respect to a given stochastic conformance criterion, and (3) the alleviation of the computational hardness associated with stochastic language calculation. Specifically, we introduced a log-driven unfolding procedure for exact probability computation, extended with memoization to reuse intermediate results and reduce redundant calculations. We proposed optimisation schemes based on differentiable and non-differentiable conformance measures and a Bayesian inference framework to estimate complete posterior distributions over model parameters. Finally, we extended the process tree formalism with stochastic semantics, enabling scalable, interpretable, and structurally guaranteed sound stochastic models.

In the short-term future, we plan to refine the optimisation and inference procedures developed in this thesis and extend our series of experiments to additional datasets. More specifically, we are improving the unfolding procedure's efficiency, aiming to further reduce computation time and enhance scalability without compromising accuracy. In the same spirit, we are working on the traversal algorithm for stochastic process trees to compute their stochastic language efficiently. We are also developing mapping rules that would allow us to convert sPTs into other stochastic formalisms, which, so far, have proven to be non-trivial.

In the longer term, we envision extending our methods to support richer behavioural features, such as time distributions. We aim to model activity durations, waiting times, and temporal correlations between events by exploiting the temporal information in event logs through their timestamps. This would allow the discovery of timed stochastic process models capable of jointly capturing control-flow, probabilistic behaviour, and temporal dynamics. This would enable such models to support novel forms of analysis, such as performance evaluation, bottleneck detection, and predictive monitoring, and to derive deeper insights from event logs.

Declaration on Generative AI

During the preparation of this work, the author used Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Rogge-Solti, W. Aalst, van der, M. Weske, Discovering stochastic petri nets with arbitrary delay distributions from event logs, in: N. Lohmann, M. Song, P. Wohed (Eds.), Business Process Management Workshops : BPM 2013 International Workshops, Beijing, China, August 26, 2013, Revised Papers, Lecture Notes in Business Information Processing, Springer, Germany, 2014, pp. 15–27. doi:10.1007/978-3-319-06257-0_2, 9th International Workshop on Business Process Intelligence (BPI 2013), BPI 2013 ; Conference date: 26-08-2013 Through 26-08-2013.

[2] A. Burke, S. J. J. Leemans, M. T. Wynn, Stochastic process discovery by weight estimation, in: S. J. J. Leemans, H. Leopold (Eds.), Process Mining Workshops - ICPM 2020 International Workshops, Padua, Italy, October 5-8, 2020, Revised Selected Papers, volume 406 of *Lecture Notes in Business Information Processing*, Springer, 2020, pp. 260–272. URL: https://doi.org/10.1007/978-3-030-72693-5_20. doi:10.1007/978-3-030-72693-5_20.

[3] A. Burke, S. J. J. Leemans, M. T. Wynn, Discovering stochastic process models by reduction and abstraction, in: D. Buchs, J. Carmona (Eds.), Application and Theory of Petri Nets and Concurrency - 42nd International Conference, PETRI NETS 2021, Virtual Event, June 23-25, 2021,

Proceedings, volume 12734 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 312–336. URL: https://doi.org/10.1007/978-3-030-76983-3_16. doi:10.1007/978-3-030-76983-3_16.

[4] T. Brockhoff, M. S. Uysal, W. M. Van Der Aalst, Wasserstein weight estimation for stochastic petri nets, in: 2024 6th International Conference on Process Mining (ICPM), 2024, pp. 81–88. doi:10.1109/ICPM63005.2024.10680664.

[5] S. Leemans, A. Syring, W. Aalst, Earth Movers' Stochastic Conformance Checking, 2019, pp. 127–143. doi:10.1007/978-3-030-26643-1_8.

[6] S. J. J. Leemans, T. Li, M. Montali, A. 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, Springer-Verlag, Berlin, Heidelberg, 2024, p. 36–52. URL: https://doi.org/10.1007/978-3-031-61057-8_3. doi:10.1007/978-3-031-61057-8_3.

[7] A. Polyvyanyy, A. Moffat, L. García-Bañuelos, An entropic relevance measure for stochastic conformance checking in process mining, in: 2020 2nd International Conference on Process Mining (ICPM), IEEE, 2020, pp. 97–104.

[8] P. Cry, A. Horváth, P. Ballarini, P. Le Gall, A framework for optimisation based stochastic process discovery, in: J. Hillston, S. Soudjani, M. Waga (Eds.), Quantitative Evaluation of Systems and Formal Modeling and Analysis of Timed Systems, Springer Nature Switzerland, Cham, 2024, pp. 34–51.

[9] R. H. Byrd, P. Lu, J. Nocedal, C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM Journal on Scientific Computing 16 (1995) 1190–1208. URL: https://doi.org/10.1137/0916069. doi:10.1137/0916069. arXiv:https://doi.org/10.1137/0916069.

[10] S. G. Nash, Newton-type minimization via the lanczos method, SIAM Journal on Numerical Analysis 21 (1984) 770–788. URL: https://doi.org/10.1137/0721052. doi:10.1137/0721052. arXiv:https://doi.org/10.1137/0721052.

[11] M. J. D. Powell, An efficient method for finding the minimum of a function of several variables without calculating derivatives, The Computer Journal 7 (1964) 155–162. URL: https://doi.org/10.1093/comjnl/7.2.155. doi:10.1093/comjnl/7.2.155. arXiv:https://academic.oup.com/comjnl/article-pdf/7/2/155/959784/070155.pdf.

[12] F. Gao, L. Han, Implementing the nelder-mead simplex algorithm with adaptive parameters, Computational Optimization and Applications 51 (2012) 259–277. doi:10.1007/s10589-010-9329-3.

[13] P. Cry, P. Ballarini, A. Horváth, P. Le Gall, Statistical Bayesian Inference for Stochastic Process Discovery, in: Proceedings of the International Conference on Quantitative Evaluation of SysTems (QEST), Aarhus (Danemark), Denmark, 2025. URL: https://hal.science/hal-05134848.