Computational Discovery of Quantitative Process Models

Pat Langley
Center for Design Research, Stanford University, Stanford, CA 94305

The field of computational scientific discovery aims to construct interpretable laws and models that are stated in established scientific formalisms. Early research in this area emphasized induction of algebraic equations from quantitative data (e.g., Langley, 1981; Langley & Żytkow, 1989), but mature disciplines move beyond such descriptive summaries to explain observations in deeper terms. One important class of such accounts involves linked differential equations that predict multivariate trajectories when simulated, and there is a substantial body of work on inducing them from time series (e.g., Džeroski & Todorovski, 1995). However, scientific publications also refer to underlying *processes* that these models instantiate, which suggests an additional discovery challenge.

In previous work (Langley, Sanchez, Todorovski, & Džeroski, 2002; Bridewell, Langley, Todorovski, & Džeroski, 2008), we have defined a *quantitative process model* as a collection of processes, each having one or more differential equations and a set of numeric parameters. Examples include process accounts of population dynamics and chemical reaction networks. The effects of processes are additive, so one can compile any such model into a set of differential equations and simulate its behavior over time. Process models are *causal* in that they specify how some changes in variables depend on others (Langley, 2019) and they are *interpretable* in that they draw on a formalism that is familiar to scientists in many different disciplines.

We have also defined *inductive process modeling* as the task of discovering such models from multivariate time series and background knowledge about the domain (Langley et al., 2002). We can state this problem in terms of inputs and outputs:

- *Given:* A set of typed variables and observed trajectories for their values over time;
- *Given:* A subset of these variables whose values one wants to explain;
- *Given:* Knowledge about types of processes that might explain the observations;
- *Find:* A process model that reproduces the observed trajectories, explains them in terms of known processes, and predicts future values.

Background knowledge includes *generic processes*, each of which specifies a set of variable *types*, one or more differential equation *forms*, and ranges for their parameter values. Knowledge can also provide constraints on how to combine processes into models (Bridewell & Langley, 2010).

We can characterize the computational discovery of quantitative process models in terms of a constrained search through two distinct but connected spaces:

- A discrete space of model *structures*, each of which comprises a set of processes that specify variables and equations that relate them but not the latter's parameter values.
- A continuous space of model *parameters* for a given structure, typically with bounds on values specified in the generic versions of its processes.

The modular character of process models affords the automated construction of candidate model structures, but their evaluation requires some metric such as quantitative fit to target observations. This in turn requires estimation of parameters for these candidates, as they are needed to simulate the models and generate predicted trajectories.

We have developed a number of systems for process model induction and tested them on observational (nonexperimental) data from a variety of domains. These have included natural data sets from ecology, hydrology, and biochemistry, as well as challenging synthetic data for chemical reactions and population dynamics (Bridewell et al., 2008). Extensions have mitigated overfitting with ensembles, estimated missing observations with iterative optimization (Bridewell et al., 2006), and explained spatio-temporal phenomena with partial differential equations (Park et al., 2010). However, early systems used exhaustive search through the structure space, so they did not scale well to models with many variables or structures. They also relied on gradient descent through the parameter space, which often halted at local optima and required random restarts.

More recent work has reformulated the task by assuming that each process has an algebraic rate expression and that derivatives are proportional to this rate (Langley & Arvay, 2015). This assumption lets one carry out heuristic search through the space of model structures and use multiple linear regression to estimate parameters. These features make process model induction far more efficient, scaling well to many variables and processes, and more likely to find good-fitting candidates. One impressive result has been the reconstruction of a food chain with 20 organisms from time series of their populations (Langley, 2019). The new framework also makes it easy, when the environment changes, to identify and repair the responsible processes (Arvay & Langley, 2016).

Inductive process modeling is a promising paradigm for computational scientific discovery that incorporates a formalism familiar to scientists, takes advantage of knowledge about the problem domain, produces meaningful results from moderate amounts of data, and generates interpretable models that explain, not just describe, observations. Future research should develop systems that devise experiments to discriminate among competing models, discover forms of entirely new processes, and induce multi-scale models that operate at different temporal resolutions. Promising applications include elucidating metabolic pathways in biochemistry, understanding ecological dynamics of human microflora, and generating designs for chemical production and synthetic biology.

References

Arvay, A., & Langley, P. (2016). Heuristic adaptation of quantitative process models. *Advances in Cognitive Systems*, *4*, 207–226.

Bridewell, W. & Langley, P. (2010). Two kinds of knowledge in scientific discovery. *Topics in Cognitive Science*, *2*, 36–52.

Bridewell, W., Langley, P., Racunas, S., & Borrett, S. R. (2006). Learning process models with missing data. *Proceedings of the Seventeenth European Conference on Machine Learning* (pp. 557–565). Berlin: Springer.

Bridewell, W., Langley, P., Todorovski, L., & Džeroski, S. (2008). Inductive process modeling. *Machine Learning*, *71*, 1–32.

Džeroski, S., & Todorovski, L. (1995). Discovering dynamics: From inductive logic programming to machine discovery. *Journal of Intelligent Information Systems*, *4*, 89–108.

Langley, P. (1981). Data-driven discovery of physical laws. *Cognitive Science*, 5, 31–54.

Langley, P. (2019). Scientific discovery, causal explanation, and process model induction. *Mind & Society*, *18*, 43–56.

Langley, P., & Arvay, A. (2015). Heuristic induction of rate-based process models. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 537–544). Austin: AAAI Press.

Langley, P., Sanchez, J., Todorovski, L., & Džeroski, S. (2002). Inducing process models from continuous data. *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 347–354). Sydney: Morgan Kaufmann.

Langley, P., & Żytkow, J. M. (1989). Data-driven approaches to empirical discovery. *Artificial Intelligence*, *40*, 283–312.

Park, C., Bridewell, W., & Langley, P. (2010). Integrated systems for inducing spatio-temporal process models. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (pp. 1555–1560). Atlanta, GA: AAAI Press.