Research Statement

Jesse Wheeler

A modern approach to modeling time series data is the use of state-space (or mechanistic) models [1]. These models require the proposal of mathematical equations that describe how unobservable states of the system evolve over time and how measurements of the system variables are obtained. State-space models are appealing because they allow researchers to incorporate mechanisms that reflect our current scientific understanding of the system in question. This capability enables researchers to estimate the effects of potential interventions on the system, provides a framework for statistical testing of our current understanding against alternative explanations, and facilitates inference on unobservable variables that may be of scientific interest.

Despite the utility of these models, statistical inference of a dynamic system using state-space models remains a challenging task [2]. The primary goal of my current research is to expand the current state-of-the-art for state-space modeling. In particular, I develop methodology, theory, and software for likelihood based inference of state-space models with the *plug-and-play* property, which is that one only needs the ability to simulate from the model in order to perform inference [3]. Algorithms with the plug-and-play property enable researchers to fit models that align with their scientific objectives instead of being limited to those that are statistically convenient [4].

1 State-space modeling in higher dimensions

The nonlinear, stochastic nature of many dynamic systems of scientific interest make the state-space modeling approach challenging. Contemporary approaches that have been successfully used for nonlinear, low-dimensional systems are based on sequential Monte Carlo (SMC) techniques [5, 6, 7]. As the dimensionality of the system increases, however, the approximation error of SMC increases exponentially [8, 9]. This necessitates the development of algorithms that can be used to perform inference on high-dimensional nonlinear state-space models. Recent progress in this area has been made—including my own research projects [10, 4]—yet several open problems remain. Furthermore, existing plug-and-play algorithms used to model these systems are computationally expensive, making any improvements in computational speed a desirable outcome.

A simplifying assumption that can often be made for high-dimensional, nonlinear dynamic systems is approximate independence between measurement units. Data collected from these systems are sometimes called panel or longitudinal data. Examples of this type include controlled ecological experiments, or systems where we are interested in modeling within-host dynamics. Each time-series in the panel may be relatively short, so that inference on an underlying model must combine information across the units. Differences between units may be of direct interest or may be a nuisance for studying the commonalities. One example includes fitting parameters that are both shared across units in the panel and also those that are specific to each unit. Current state-of-the-art plug-and-play algorithms available for fitting models of this type treat panel state-space models similar to their lower-dimensional equivalents [11], and can therefore still suffer from particle depletion [9], making these approaches less applicable in high dimensions.

One of my current research projects aims to reduce the computational burden associated with fitting state-space models to panel data, and allow for the ability to fit higher-dimensional models. An algorithm I have designed, called the marginalized panel iterated filter (MPIF), does this by accounting for the independence between the likelihood of a given unit in a panel and the unit-specific parameters of other units. While the theory for this algorithm is still being developed, it has demonstrated to be an improvement over existing approaches on simulated data and on examples that have been attempted by undergraduate students I have mentored [12, 13]. I am also

the current maintainer of a CRAN R package that is designed to aid in fitting mechanistic models to panel data. The current version of this package emphasizes plug-and-play methodology, though it is built to support other methodological developments for this class of models [14].

A weakness of maximum likelihood estimation is that estimated parameters may not make sense in the context of the model and data in question [15]. For mechanistic models, this is particularly troublesome because estimated parameters should provide a quantitative description of the data while retaining a meaningful mechanistic interpretation [4]. In high dimensional systems, the chance of finding parameters that maximize the likelihood in implausible regions of the parameter space increases. For this reason I plan on extending existing iterated filtering algorithms [7, 10] to enable the maximization of penalized likelihood [16], which helps address issues of parameter instability that arise in likelihood maximization. Additionally, penalized likelihood may enable the estimation of random effects in high-dimensional state-space models, an open research question. This work may be facilitated by recent methodological advances in automatic differentiation by some of my collaborators [17] and software projects which I am involved in (`https://github.com/pypomp`).

2 Revisiting inference for ARMA models

State-space models have additional uses beyond the class of mechanistic models that are popular for modeling dynamic systems. Likelihood maximization for auto regressive moving average (ARMA) models, for instance, is traditionally done by reparameterizing the model into an equivalent linear Gaussian state-space form [18, 1]. In this case, the state-space model serves primarily as a computational tool rather than being of direct interest itself.

Despite the widespread use of ARMA models across various scientific disciplines, existing algorithms for maximum likelihood estimation of ARMA models possess under-recognized weaknesses. In one of my working papers, I demonstrate how existing optimization strategies for ARMA models often result in parameter estimates corresponding to local—rather than global—maximum of the likelihood surface [19].

In other optimization contexts, the issue of converging to a local solution is addressed through multiple runs of the algorithm with different starting values. However, this approach is not readily applicable to ARMA models due to the intricate geometry of their likelihood surfaces [20]. To rectify this, I propose a random restart algorithm that accounts for the geometry of the likelihood surface and can frequently result in higher likelihoods than current standards [19]. This included the creation of a now-popular R package, `arima2`. These results represent a significant advance of statistical practice given the importance of ARMA models in both scientific and industrial applications. I believe there are various potential extensions of this package and the implemented methodology that are particularly suitable for introductory research projects at the undergraduate and master's levels.

3 Scientific applications

Developing a state-space model to make inference on a dynamic system requires close collaboration with experts from pertinent scientific disciplines. For example, conducting a statistical analysis of an infectious disease outbreak relies heavily on expert opinion regarding the specific disease [4]. One of my ongoing projects is the investigation of population dynamics in freshwater *Daphnia* species. In this endeavor, I closely collaborate with colleagues in the Ecology and Evolutionary Biology

department, leveraging their knowledge of *Daphnia*. This collaboration has honed my skills as an effective statistical collaborator, a competency I look forward to further refine in future research projects.

Currently, most of my applied research has been in the modeling of infectious diseases, though I am interested in developing collaborations in other scientific disciplines. In future work, I plan to apply recent advancements in automatically differentiable particle filters [17] to investigate how machine learning tools can enhance mechanistic models for nonlinear dynamic systems. One example is examining how changes in human behavior may affect disease dynamics. Although similar ideas have been explored recently [21, 22], these studies required fitting an ODE model to cumulative case counts using summary statistics, an approach that has elsewhere been found to be problematic [23]. The `pypomp` software project, currently under development (`https://github.com/pypomp`), may enable fitting scientifically motivated machine learning models via maximum likelihood. This could lean to novel insights into an infectious disease system. Similarly, this approach has the potential to yield new understanding in other nonlinear dynamic systems by incorporating the scientific benefits of mechanistic models [24] with the statistical efficiency of likelihood-based inference and the flexibility of neural networks.

4 Software

Central to my research interests is the emphasis on good statistical practices, such as transparency, reproducibility, and effective communication. Statisticians must take a proactive role in promoting and implementing these principles. Consequently, one of my primary research goals is to positively impact the scientific community by writing papers and developing software that facilitate and promote good statistical practices. This commitment has guided me in the development of open-source software related to statistical modeling and inference. This section briefly describes some of the software projects I am currently involved in.

- `arima2`: This package provides a novel computational approach for fitting ARIMA models in `R`. The most important function of this package is the `arima` function, which fits ARIMA models using a multiple restart algorithm that frequently results in models with higher likelihoods than other alternatives [19]. I am the creator and primary contributor of this package, which is publicly available on both CRAN and GitHub.

- `panelPomp`: This package provides a framework for developing high-dimensional state-space models under the assumption of independence between panel units. This package is publicly available on both CRAN and GitHub. After joining the development team for panelPomp in May, 2023, I became the most active contributor and am now the package maintainer. Together with collaborators, I wrote a tutorial demonstrating a typical use case of this package that is available on ArXiv [25].

- `pypomp`: This is a new Python project with the goal of providing a framework for developing arbitrary POMP models in Python. The code currently focuses on supporting the automatic differentiation methodology explored by Tan, K., Ionides, E. L. and Hooker, G. [17]. I am one of the core developers on this project.

- `pomp` and `spatPomp`: These are both popular packages that enable inference for mechanistic models. The `pomp` package in particular is used by thousands of researchers across a variety of disciplines. I am an active contributor to both of these open source projects.

References

[1] J. Durbin and S. J. Koopman, *Time series analysis by state space methods*, vol. 38. OUP Oxford, 2012.

[2] M. Auger-Méthé, C. Field, C. M. Albertsen, A. E. Derocher, M. A. Lewis, I. D. Jonsen, and J. Mills Flemming, "State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems," *Scientific reports*, vol. 6, no. 1, p. 26677, 2016.

[3] C. Bretó, D. He, E. L. Ionides, and A. A. King, "Time series analysis via mechanistic models," *The Annals of Applied Statistics*, vol. 3, no. 1, pp. 319 – 348, 2009.

[4] J. Wheeler, A. Rosengart, Z. Jiang, K. Tan, N. Treutle, and E. L. Ionides, "Informing policy via dynamic models: Cholera in Haiti," *PLOS Computational Biology*, vol. 20, no. 4, p. e1012032, 2024.

[5] E. L. Ionides, C. Bretó, and A. A. King, "Inference for nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 103, no. 49, pp. 18438–18443, 2006.

[6] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 72, no. 3, pp. 269–342, 2010.

[7] E. L. Ionides, D. Nguyen, Y. Atchadé, S. Stoev, and A. A. King, "Inference for dynamic and latent variable models via iterated, perturbed Bayes maps," *Proceedings of the National Academy of Sciences*, vol. 112, no. 3, pp. 719–724, 2015.

[8] T. Bengtsson, P. Bickel, and B. Li, "Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems," in *Probability and statistics: Essays in honor of David A. Freedman*, vol. 2, pp. 316–335, Institute of Mathematical Statistics, 2008.

[9] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson, "Obstacles to high-dimensional particle filtering," *Monthly Weather Review*, vol. 136, no. 12, pp. 4629–4640, 2008.

[10] E. L. Ionides, N. Ning, and J. Wheeler, "An iterated block particle filter for inference on coupled dynamic systems with shared and unit-specific parameters," *Statistica Sinica. pre-published online*, pp. pre–published online, 2022.

[11] C. Bretó, E. L. Ionides, and A. A. King, "Panel data analysis via mechanistic models," *Journal of the American Statistical Association*, 2020.

[12] B. Yang, "Analysis of panel data via mechanistic models in a panelPomp framework," *University of Michigan, Undergraduate Honors Thesis*, 2023.

[13] W. Sun, "Model based inference of stochastic volatility via iterated filtering," *University of Michigan, Undergraduate Honors Thesis*, 2024.

[14] C. Bretó, J. Wheeler, A. A. King, and E. L. Ionides, "panelpomp: Analysis of panel data via partially observed markov processes in r," *arXiv preprint arXiv:2410.07934*, 2024.

[15] L. Le Cam, "Maximum likelihood: An introduction," *International Statistical Review / Revue Internationale de Statistique*, vol. 58, no. 2, pp. 153–171, 1990.

[16] S. R. Cole, H. Chu, and S. Greenland, "Maximum likelihood, profile likelihood, and penalized likelihood: A primer," *American Journal of Epidemiology*, vol. 179, pp. 252–260, 10 2013.

[17] K. Tan, G. Hooker, and E. L. Ionides, "Accelerated inference for partially observed Markov processes using automatic differentiation," *arXiv preprint arXiv:2407.03085*, 2024.

[18] G. Gardner, A. C. Harvey, and G. D. A. Phillips, "Algorithm AS 154: An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 29, no. 3, pp. 311–322, 1980.

[19] J. Wheeler and E. L. Ionides, "Likelihood based inference of ARMA models," *ArXiv preprint*, 2023.

[20] B. Ripley, "Time series in R 1.5.0. R News, 2/2, 2–7." https://www.r-project.org/doc/Rnews/Rnews_2002-2.pdf, June 2002.

[21] R. Dandekar, C. Rackauckas, and G. Barbastathis, "A machine learning-aided global diagnostic and comparative tool to assess effect of quarantine control in COVID-19 spread," *Patterns*, vol. 1, no. 9, 2020.

[22] S. Kim, W. Ji, S. Deng, Y. Ma, and C. Rackauckas, "Stiff neural ordinary differential equations," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 31, no. 9, 2021.

[23] A. A. King, M. Domenech de Cellès, F. M. Magpantay, and P. Rohani, "Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to ebola," *Proceedings of the Royal Society B: Biological Sciences*, vol. 282, no. 1806, p. 20150347, 2015.

[24] R. E. Baker, J.-M. Pena, J. Jayamohan, and A. Jérusalem, "Mechanistic models versus machine learning, a fight worth fighting for the biological community?," *Biology letters*, vol. 14, no. 5, p. 20170660, 2018.

[25] C. Breto, J. Wheeler, A. A. King, and E. L. Ionides, "A tutorial on panel data analysis using partially observed Markov processes via the R package panelPomp," *arXiv preprint arXiv:2409.03876*, 2024.