Research Statement

Jesse Wheeler

Recent advancements in data collection and storage have led to an abundance of complex datasets. Often, these datasets include time-dependent measurements, resulting in interdependent observations that render many traditional statistical modeling approaches ineffective. This issue is particularly prevalent in nonlinear dynamic systems that are frequently of scientific or industrial interest.

A modern approach to modeling time-dependent observations is the use of state-space (or mechanistic) models [1]. These models involve proposing 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 allow researchers to fit models that are scientifically accurate rather than those that are statistically convinient [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 (for instance, [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 desireable outcome.

One simplifying assumption that can often be made for high-dimensional, nonlinear dynamic systems is approximate independence bewtween measurement units. Data collected from these systems are sometimes called panel or longitudinal data. 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 inferential 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 of model treat panel state-space models similar to their lower-dimensional equivalents [11], and can therefore still suffer from particle depletion [9], resulting in the need for expensive numerical calculations.

One of my current research projects aims to reduce the computational burden associated with fitting state-space models to panel data. An algortihm I have designed, called the block panel iterated filter (BPIF), does this by accounting for the indepdenence 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 proven to be an improvement over existing approaches on simulated data and in some practical examples that have been attempted by undergraduate students I have mentored [12, 13].

One weakness of maximum likelihood estimation is that estimated parameters may not make sense in the context of the model and data in question [14]. 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 inplausible regions of the parameter space increases [15]. 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.

2 Additional Research

State-space models have additional uses beyond the class of mechanistic models that are popular in Ecology. 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 [17, 1]. Despite the wide use of ARMA models across various scientific disciplines, current algorithms for maximum likelihood estimation of ARMA models have weaknesses that are not well known. 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—maxima of the likelihood surface. To rectify this, I propose a random restart algorithm that frequently results in higher likelihoods than standard alternatives [18].

While my current research is focused on state-space modeling of dynamic systems, I am interested in many other statistical topics and am open to new research topics and collaborations. Some themes common to each of my research interests are transparency, reproducibility and simplicity. I believe that many of the challenges faced by various scientific disciplines are a direct result of the poor practice of these three principles. I also believe that statisticians must play a more active role in promoting and executing these principles. As such, one of my primary research goals is to positively impact the scientific community by writing papers and developing software that encourage and enable other researchers to incorporate these principles in their own work.

3 Software

The development of theory and methods for likelihood based inference of state-space models requires development of new software. This section briefly describes some of the open-source software packages to which I have contributed.

- `panelPomp`: This package provides a framework for developing high-dimensional state-space models under the assumption of independence between panel units. This package is publically available on both CRAN and GitHub. While I was not the original creator of the package, I am currently the primary package maintainer and developer.

- `arima2`: This package provides useful functions for fitting auto-regressive, integrated, moving-average (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 [18]. I am the creator and primary contributor of this package, which is publically available on both CRAN and GitHub.

- `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] L. Le Cam, "Maximum Likelihood: An Introduction," *International Statistical Review / Revue Internationale de Statistique*, vol. 58, no. 2, pp. 153–171, 1990.

[15] J. Li, E. L. Ionides, A. A. King, M. Pascual, and N. Ning, "Inference on spatiotemporal dynamics for coupled biological populations," *Journal of the Royal Society Interface*, vol. 21, no. 216, p. 20240217, 2024.

[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] 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.

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