Innovations in Likelihood-Based Inference for State Space Models

Oral Defense

Jesse Wheeler

May 29, 2025

Department of Statistics, University of Michigan

Table of contents

1. Introduction
2. Likelihood Maximization for ARMA models
3. Informing Policy via Dynamic Models: Cholera in Haiti
4. The Marginalized Panel Iterated Filter (MPIF)
5. Conclusion and Future Directions

1. Introduction

Synonyms and Definitions

State Space Models (SSMs) are used for time series analysis (Durbin and Koopman, 2012).

- Highly interpretable, used to aid computations, and useful for modeling real-world mechanisms.

There are several terms that have been used as synonyms with SSMs:

- Mechanistic model
- Hidden Markov model (HMM)
- Partially observed Markov process (POMP) model

I will begin by defining how these terms are used in my thesis.

I follow the definition used by Durbin and Koopman (2012) for a SSM.

- $Y_1, Y_2, \ldots, Y_N$ are observed time series. Observations occur at time points $t_1, \ldots, t_N$, and can be vector valued.
- A SSM introduces unobservable (latent) states $X_1, \ldots, X_N$ at the same observation times. These latent variables are connected to the observations, in a way defined by the model.
- We often include an initial value for the latent states, X_0.

I will adopt the shorthand $t_{1:N} = (t_1, \ldots, t_N)$, $Y_{1:N} = (Y_1, \ldots, Y_N)$, and $X_{0:N} = (X_0, \ldots, X_N)$.

We assume that the random variables $Y_{1:N}$, $X_{0:N}$ have a joint probability density $f_{X_{0:N},Y_{1:N}}(x_{0:N}, y_{1:N};\, \theta)$ with respect to some dominating measure (Lebesgue or counting).

θ is a parameter vector $\theta \in \mathbb{R}^{d_\theta}$ that indexes the model (unknown).

Because only $Y_{1:N}$ is observable, the likelihood function involves a high-dimensional integral:

$$\mathcal{L}(\theta; y^*) = f_{Y_{1:N}}\left(y^*_{1:N};\, \theta\right) = \int f_{X_{0:N},Y_{1:N}}\left(x_{0:N}, y^*_{1:N};\, \theta\right) dx_{0:N}. \tag{1}$$

A common approach is to treat SSMs as partially observed Markov process (POMP) models. We make the following assumptions:

- We assume that the latent variables are a Markov process

$$f_{X_n|X_{1:n-1}}(x_n|x_{1:n-1};\,\theta) = f_{X_n|X_{n-1}}(x_n|x_{n-1};\,\theta).$$

- Measurements are conditionally independent

$$f_{Y_n|X_{1:N},Y_{-n}}(y_n|x_{0:N},y_{-n};\,\theta) = f_{Y_n|X_n}(y_n|x_n;\,\theta).$$

With these assumptions, we can express the joint density as

$$f_{X_{0:N},Y_{1:N}}\left(x_{0:N},y_{1:N};\,\theta\right) = f_{X_0}\left(x_0;\,\theta\right)\prod_{n=1}^{N} f_{X_n|X_{n-1}}\left(x_n|x_{n-1};\,\theta\right) f_{Y_n|X_n}\left(y_n|x_n;\,\theta\right). \tag{2}$$

$$f_{X_0}\left(\boldsymbol{x}_0;\,\theta\right)\prod_{n=1}^{N} f_{X_n|X_{n-1}}\left(\boldsymbol{x}_n|\boldsymbol{x}_{n-1};\,\theta\right)f_{Y_n|X_n}\left(\boldsymbol{y}_n|\boldsymbol{x}_n;\,\theta\right)$$

- The initialize density (or initializer): $f_{X_0}\left(\boldsymbol{x}_0;\,\theta\right)$.
- The transition density (or process model): $f_{X_n|X_{n-1}}\left(\boldsymbol{x}_n|\boldsymbol{x}_{n-1};\,\theta\right)$.
- The measurement density (or model): $f_{Y_n|X_n}\left(\boldsymbol{y}_n|\boldsymbol{x}_n;\,\theta\right)$.

The latent states can be defined as a continuous time process with values in-between measurement times.

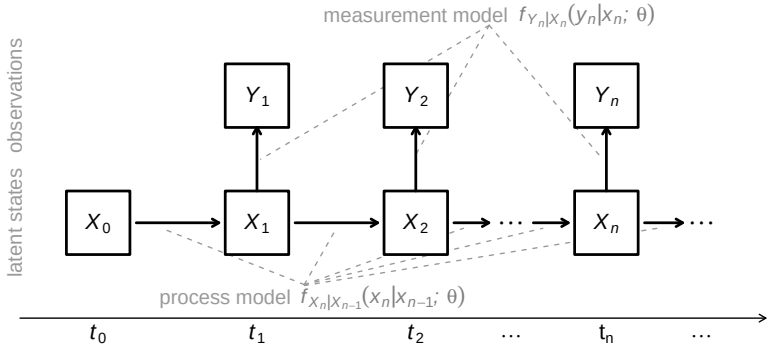


Figure 1: A flow diagram representing an arbitrary POMP model. Modified figure from SBIED course (King, Ionides).

Each of the SSMs considered in this thesis are POMP models.

Other synonyms and definitions

Other common terms that are sometimes used as synonyms are used for special cases

Mechanistic Model

A SSM (or POMP) where the evolution of latent variables is dictated by equations that mimic real-world mechanisms.

Hidden Markov Model (HMM)

A SSM (or POMP) where the latent variables take values in a discrete and finite space.

I chose SSM for the title as it is the terminology that potential collaborators are most familiar with.

Remaining Chapters and Outline

- Chapter 2: Inference for ARMA models.
  - ARMA models are a special type of linear Gaussian SSMs, and are an important part of modern science (In review).
- Chapter 3: Mechanistic models for modeling cholera (Wheeler et al., 2024).
  - This case study discusses the strengths and weaknesses of using mechanistic models to inform government policy, using a retrospective analysis of the 2010-2019 cholera outbreak in Haiti.
- Chapter 4: The marginalized panel iterated filter (MPIF) algorithm.
  - A new inference algorithm is proposed for a large collections of related POMP models, called PanelPOMPs, and theory for existing algorithms for these models is also presented.
- Chapter 5: Summary and future directions.

2. Likelihood Maximization for ARMA models

ARMA models are the most frequently used approach to modeling time series data.

ARMA model definition

A time series $Y_{1:N}$ is called $\mathrm{ARMA}(p, q)$ if it is (weakly) stationary and

$$Y_n = \phi_1 Y_{n-1} + \cdots + \phi_p Y_{n-p} + w_n + \varphi_1 w_{n-1} + \ldots + \varphi_q w_{n-q}, \tag{3}$$

with $\{w_n; n = 0, \pm 1, \pm 2, \ldots\}$ denoting a mean zero white noise (WN) processes with variance $\sigma_w^2 > 0$, and $\phi_p \neq 0$, $\varphi_q \neq 0$.

p and q of Eq. (3) as the autoregressive (AR) and moving average (MA) orders, respectively.

How are they relevant? Inference methodology treats ARMA models as *non-mechanistic* SMMs. Let $r = \max(p, q+1)$, and assume $w_n \overset{\text{iid}}{\sim} N(0, \sigma_w^2)$. Define

$$X_n = \begin{pmatrix} Y_n \\ \phi_2 Y_{n-1} + \ldots + \phi_r Y_{n-r+1} + \varphi_1 w_n + \ldots + \varphi_{r-1} w_{n-r+2} \\ \phi_3 Y_{n-1} + \ldots + \phi_r Y_{n-r+2} + \varphi_2 w_n + \ldots + \varphi_{r-1} w_{n-r+3} \\ \vdots \\ \phi_r Y_{n-1} + \varphi_{r-1} w_n \end{pmatrix} \in \mathbb{R}^r$$

$$T = \begin{pmatrix} \phi_1 & 1 & 0 & \ldots & 0 \\ \phi_2 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & \ddots & \\ \phi_{r-1} & 0 & \ldots & & 1 \\ \phi_r & 0 & \ldots & & 0 \end{pmatrix} \in \mathbb{R}^{r \times r}, \qquad Q = \begin{pmatrix} 1 \\ \varphi_1 \\ \vdots \\ \varphi_{r-1} \end{pmatrix} \in \mathbb{R}^r$$

State space formulation (ii)

We can then recover the ARMA model using the following state space formulation:

$$X_n = TX_{n-1} + Qw_n$$
$$Y_n = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} X_n$$

This results in a linear-Gaussian SSM, and the likelihood function $\mathcal{L}(\theta)$ can be evaluated exactly using the Kalman filter (Kalman, 1960).

- The likelihood can be maximized by combining this with a numeric optimizer (Gardner et al., 1980).

This approach has been the standard method for fitting ARIMA models since the early 2000's due to modern computing capabilities (Ripley, 2002).

This existing approach frequently results in sub-optimal parameter estimates. Simple example in **R**:

- Generate data from an $\mathrm{ARMA}(2,2)$ model.
- Fit both $\mathrm{ARMA}(2,1)$ and $\mathrm{ARMA}(2,2)$ models to simulated data.

The $\mathrm{ARMA}(2,1)$ is formally a special case of an $\mathrm{ARMA}(2,2)$ model, with $\varphi_2 = 0$.

Simulation details:

- $(\phi_1, \phi_2, \varphi_1, \varphi_2) = (0.2, -0.1, 0.4, 0.2)$.
- $w_n \overset{\text{iid}}{\sim} N(0,1)$.
- $N = 100$ observations with intercept $\mu = 13$ so that $E[Y_n] \neq 0$.

This existing approach frequently results in sub-optimal parameter estimates. Simple example in **R**:

- Generate data from an $\mathrm{ARMA}(2,2)$ model.
- Fit both $\mathrm{ARMA}(2,1)$ and $\mathrm{ARMA}(2,2)$ models to simulated data.

The $\mathrm{ARMA}(2,1)$ is formally a special case of an $\mathrm{ARMA}(2,2)$ model, with $\varphi_2 = 0$. Simulation details:

- $(\phi_1, \phi_2, \varphi_1, \varphi_2) = (0.2, -0.1, 0.4, 0.2)$.
- $w_n \overset{\text{iid}}{\sim} N(0,1)$.
- $N = 100$ observations with intercept $\mu = 13$ so that $E[Y_n] \neq 0$.

The Gardner et al. (1980) is the standard method for fitting ARMA model parameters. It is implemented in the base **stats** package in R, as well as the **statsmodels** module in Python.

```r
mod1 <- stats::arima(y, order = c(2, 0, 1))
mod2 <- stats::arima(y, order = c(2, 0, 2))
```

The likelihood of **mod1** is -141.2, and the likelihood of **mod2** is -144.3. The smaller model has a log-likelihood that is 3.1 units higher than the larger model, which is mathematically impossible under proper optimization (could estimate $\hat{\varphi}_2 = 0$).

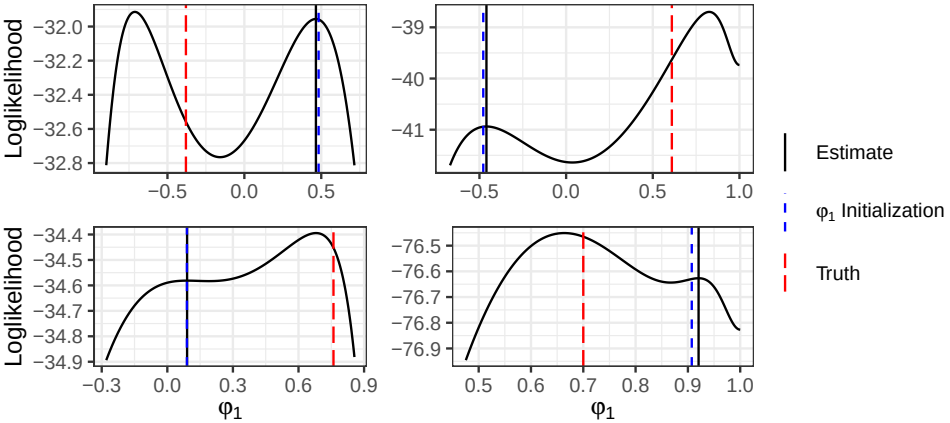
The Gardner et al. (1980) is the standard method for fitting ARMA model parameters. It is implemented in the base **stats** package in R, as well as the **statsmodels** module in Python.

```
mod1 <- stats::arima(y, order = c(2, 0, 1))
mod2 <- stats::arima(y, order = c(2, 0, 2))
```

The likelihood of **mod1** is -141.2, and the likelihood of **mod2** is -144.3. The smaller model has a log-likelihood that is 3.1 units higher than the larger model, which is mathematically impossible under proper optimization (could estimate $\hat{\varphi}_2 = 0$).

Convergence to local optima

Why? Likelihood surface is often multimodal, and existing procedures run the risk of converging to a local solution (Ripley, 2002). Example $\mathrm{MA}(1)$ profiles:



Multiple parameter initializations

In other contexts with multi-model loss functions, the optimization is often repeated using multiple initializations.

However, I have seen no instances of this for ARIMA models. There are a few challenges:

- Most users don't know about the possibility of converging to local solutions.
- Their are complex constraints on parameter initializations.
  - Constraints are on the roots of polynomials formed by model parameters, not directly on parameters themselves; transformations not readily available.

The roots of the polynomials $\Phi(x) = 1 - \phi_1 x - \phi_2 x^2 - \ldots - \phi_p x^p$ and $\Psi(x) = 1 + \varphi_1 x + \varphi_2 x^2 + \ldots + \varphi_q x^q$ must lie outside the complex unit circle.

Multiple parameter initializations

In other contexts with multi-model loss functions, the optimization is often repeated using multiple initializations.

However, I have seen no instances of this for ARIMA models. There are a few challenges:

- Most users don't know about the possibility of converging to local solutions.
- Their are complex constraints on parameter initializations.
  - Constraints are on the roots of polynomials formed by model parameters, not directly on parameters themselves; transformations not readily available.

The roots of the polynomials $\Phi(x) = 1 - \phi_1 x - \phi_2 x^2 - \ldots - \phi_p x^p$ and $\Psi(x) = 1 + \varphi_1 x + \varphi_2 x^2 + \ldots + \varphi_q x^q$ must lie outside the complex unit circle.

For parameters to be real, the roots need to be sampled as real or conjugate pairs.

We cannot sample all roots as conjugate pairs (or real), as this would result in specific parameters being all one sign.

Our approach for each root is the following:

- Sample inverted-root magnitudes uniformly $U(\gamma, 1-\gamma)$.
- With probability $p = \sqrt{1/2}$, sample inverted-root pairs as real.
  - If real, assign the same sign with probability p.
  - If complex, sample angle from $U(0, \pi)$, and use to assign conjugate pairs of inverted-roots.
- With roots sampled, calculate corresponding coefficients and perform optimization routine.
- Repeat until convergence.

Revisiting toy example

Algorithm allows for minimal changes to user interface. Using the same data, models fit using the **arima2** package:

```
mod1v2 <- arima2::arima(y, order = c(2, 0, 1))
mod2v2 <- arima2::arima(y, order = c(2, 0, 2))
```

With this new algorithm and software, the likelihood of **mod1v2** is -141.173, and the likelihood of **mod2v2** is -141.172.

The likelihood of the smaller model was unchanged, but the larger model had an increase in log-likelihood of 3.1. The likelihoods of the nested models are now consistent.

Revisiting toy example

Algorithm allows for minimal changes to user interface. Using the same data, models fit using the **arima2** package:

```
mod1v2 <- arima2::arima(y, order = c(2, 0, 1))
mod2v2 <- arima2::arima(y, order = c(2, 0, 2))
```

With this new algorithm and software, the likelihood of **mod1v2** is -141.173, and the likelihood of **mod2v2** is -141.172.

The likelihood of the smaller model was unchanged, but the larger model had an increase in log-likelihood of 3.1. The likelihoods of the nested models are now consistent.

Summary

ARMA models are not necessarily state-of-the art. Why should we care?

- ARMA models are among the most frequently used approaches in all of statistics, so even small improvements are worth the effort.
- Software that claims to maximize model likelihoods fails to do so in a large number of cases ($> 20\%$).
- ARMA models are often used in conjunction with linear regression. Likelihood ratio tests are common for testing the inclusion / significance of regression parameters.
  - Typical improvements in log-likelihood in the range $(0.22, 1.46)$. This shortcoming in one or both model is large enough to change the outcome of these tests.
- Even if outcomes are unchanged, confidence that software / algorithms will reliably do what you expect is important.

Summary

ARMA models are not necessarily state-of-the art. Why should we care?

- ARMA models are among the most frequently used approaches in all of statistics, so even small improvements are worth the effort.
- Software that claims to maximize model likelihoods fails to do so in a large number of cases ($> 20\%$).
- ARMA models are often used in conjunction with linear regression. Likelihood ratio tests are common for testing the inclusion / significance of regression parameters.
  - Typical improvements in log-likelihood in the range $(0.22, 1.46)$. This shortcoming in one or both model is large enough to change the outcome of these tests.
- Even if outcomes are unchanged, confidence that software / algorithms will reliably do what you expect is important.

Summary

ARMA models are not necessarily state-of-the art. Why should we care?

- ARMA models are among the most frequently used approaches in all of statistics, so even small improvements are worth the effort.
- Software that claims to maximize model likelihoods fails to do so in a large number of cases ($> 20\%$).
- ARMA models are often used in conjunction with linear regression. Likelihood ratio tests are common for testing the inclusion / significance of regression parameters.
  - Typical improvements in log-likelihood in the range $(0.22, 1.46)$. This shortcoming in one or both model is large enough to change the outcome of these tests.
- Even if outcomes are unchanged, confidence that software / algorithms will reliably do what you expect is important.

Summary

ARMA models are not necessarily state-of-the art. Why should we care?

- ARMA models are among the most frequently used approaches in all of statistics, so even small improvements are worth the effort.
- Software that claims to maximize model likelihoods fails to do so in a large number of cases ($> 20\%$).
- ARMA models are often used in conjunction with linear regression. Likelihood ratio tests are common for testing the inclusion / significance of regression parameters.
  - Typical improvements in log-likelihood in the range $(0.22, 1.46)$. This shortcoming in one or both model is large enough to change the outcome of these tests.
- Even if outcomes are unchanged, confidence that software / algorithms will reliably do what you expect is important.

Summary

ARMA models are not necessarily state-of-the art. Why should we care?

- ARMA models are among the most frequently used approaches in all of statistics, so even small improvements are worth the effort.
- Software that claims to maximize model likelihoods fails to do so in a large number of cases ($> 20\%$).
- ARMA models are often used in conjunction with linear regression. Likelihood ratio tests are common for testing the inclusion / significance of regression parameters.
  - Typical improvements in log-likelihood in the range $(0.22, 1.46)$. This shortcoming in one or both model is large enough to change the outcome of these tests.
- Even if outcomes are unchanged, confidence that software / algorithms will reliably do what you expect is important.

3. Informing Policy via Dynamic Models: Cholera in Haiti

One of the most scientifically interesting types of SSMs are *mechanistic models*.

- Used when we have some understanding of how a dynamic system evolves over time.
- Useful in modern science, and have some advantages over machine learning models (Baker et al., 2018; Hogg and Villar, 2024):
  - Accounting for known (but unobserved) features can improve model performance.
  - More intepretable.
  - Facilitates predictions of interventions and other counter-factuals.

In this chapter, I demonstrate these capabilities by fitting mechanistic models to the 2010-2019 cholera outbreak in Haiti (Wheeler et al., 2024).

One of the most scientifically interesting types of SSMs are *mechanistic models*.

- Used when we have some understanding of how a dynamic system evolves over time.
- Useful in modern science, and have some advantages over machine learning models (Baker et al., 2018; Hogg and Villar, 2024):
  - Accounting for known (but unobserved) features can improve model performance.
  - More intepretable.
  - Facilitates predictions of interventions and other counter-factuals.

In this chapter, I demonstrate these capabilities by fitting mechanistic models to the 2010-2019 cholera outbreak in Haiti (Wheeler et al., 2024).

Cholera in Haiti

- Haiti experienced a cholera outbreak following the devastating 2010 earthquake.
- From 2010-2019, more than 800,000 recorded cases, making it one of the largest recorded outbreaks.
- Oral cholera vaccination (OCV) is available, but in limited supply.
- Image credit: UNICEF (2022).

Proposing interventions

A group of researchers built three mechanistic models to estimate the potential impacts of various vaccination strategies (Lee et al., 2020).

- Distinct groups each predicted large cholera incidence from Feb 2019–Feb 2024.
- There were no confirmed cases from Feb 2019 - Sep 2022 (Trevisin et al., 2022).
- Though there were some cases recently recorded, not near the predicted scale (Pan American Health Organization, 2023).

Questions: What are strengths and weaknesses of mechanistic models? What are common mistakes researchers make? How can we improve outcomes in the future?

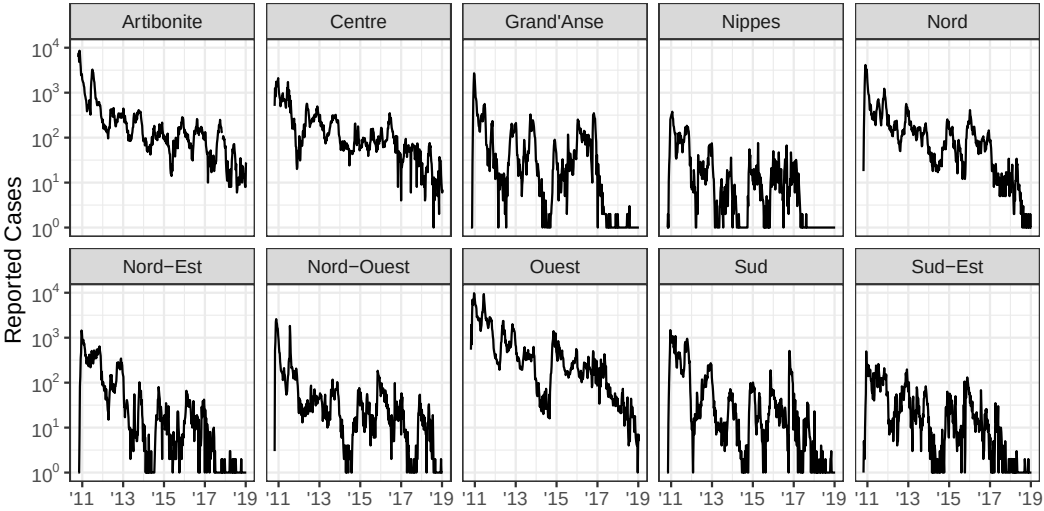
Proposing interventions

A group of researchers built three mechanistic models to estimate the potential impacts of various vaccination strategies (Lee et al., 2020).

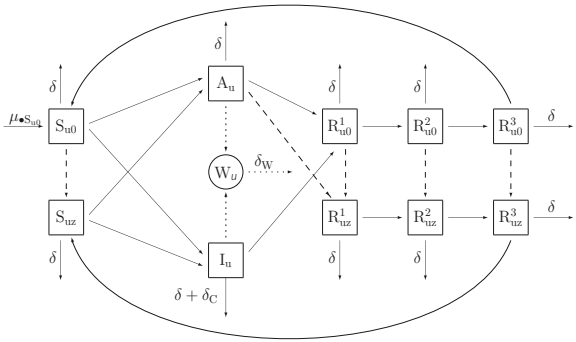
- Distinct groups each predicted large cholera incidence from Feb 2019–Feb 2024.
- There were no confirmed cases from Feb 2019 - Sep 2022 (Trevisin et al., 2022).
- Though there were some cases recently recorded, not near the predicted scale (Pan American Health Organization, 2023).

Questions: What are strengths and weaknesses of mechanistic models? What are common mistakes researchers make? How can we improve outcomes in the future?

Available Data



Spatially explicit model: Model 3



- Spatial Dependence between units.
- Stochastic transmission rates.
- Overdispersed Markov counting system.
- Rainfall driven transmission.
- Environmental reservoir of bacteria.
- Adjustments for Hurricane Mathew (Oct 2016).

In total, there are 34 parameters to estimate.

Same model used by Lee et al. (Group 3 of 2020), who fit using an independence approximation.

We used the the iterated block particle filter (IBPF) to fit the model (Ionides et al., 2024).

	Our Fit	Original Fit	Benchmark
Log-likelihood	-17332.9	-33832.6	-17932.6
AIC	34733.9	67723.2	35945.0

Table 1: Comparison of our fitted model to original parameters used to inform vaccination policy, compared to log $\mathrm{ARMA}(2,1)$ benchmark.

Key findings

Key findings of this study include:

- Confirmed importance of rainfall and reduced transmission over time.
- Importance of proper model diagnostics.
  - Comparing to benchmarks.
  - Checking results against features of the system.
- Reproducibility and Extendability.
- Confirmed previous findings: stochastic models are better descriptions of the system, and over-dispersed models are best.

4. The Marginalized Panel Iterated Filter (MPIF)

Often we have a collection of related time series called, called *panel data*. We want to make inference using the entire collection, not just on each time series.

Examples:

- Model for disease outbreaks of the same disease, different locations (hospitals / cities) (Lee et al., 2020).
- Experiments / observational studies on ecological populations (Searle et al., 2016).
- Longitudinal studies using within-host dynamic models (Ranjeva et al., 2017).

Mechanistic models are routinely fitted to time series data but seldom to panel data, despite its widespread availability.

Measurements for unit u taken at times $t_{u,1:N_u}$. Observed and latent process at these times denoted $Y_{u,n}$ and $X_{u,n}$, respectively.

Each unit u defines an independent POMP model, the entire collection of models is a PanelPOMP.

$$\mathcal{L}(\theta; \boldsymbol{y}^*) = \int \prod_{u=1}^{U} f_{X_{u,0}}\big(x_{u,0};\, \theta\big) \prod_{n=1}^{N_u} f_{X_{u,n}|X_{u,n-1}}\big(x_{u,n}|x_{u,n-1};\, \theta\big) f_{Y_{u,n}|X_{u,n}}\big(y_{u,n}|x_{u,n};\, \theta\big)\, dx_{1:U,0:N_u}.$$

The parameter vector θ has shared components ϕ, and unit specific components $\psi_{1:U}$.

$$\theta = (\phi, \psi_{1:U})$$

The problem

Independent models, why not do inference independently?

- Each model may share features (or parameters), and we want to estimate using all of the data.

Common inference procedures in low dimensions rely on particle filters (Arulampalam et al., 2002).

✓ Particle filters work in low-dimensions, can be applied independently to units.

✗ Iterated filtering (IF) is an extension used to perform maximum likelihood estimation (Ionides et al., 2015).

- IF introduces dependence because of shared θ, making it a high-dimensional problem.

IF is a special type of Data cloning (Lele et al., 2007).

Denote $\pi_i(\theta)$ as the posterior distribution of the parameter vector θ after the ith Bayesian update, and $\mathcal{L}(\theta; \boldsymbol{y}^*)$ as the likelihood

$$\begin{aligned}
\pi_1(\theta) &\propto \mathcal{L}(\theta; \boldsymbol{y}^*)\pi_0(\theta), \\
\pi_2(\theta) &\propto \mathcal{L}(\theta; \boldsymbol{y}^*)\pi_1(\theta) \propto \mathcal{L}(\theta; \boldsymbol{y}^*)^2\pi_0(\theta), \\
&\vdots \\
\pi_m(\theta) &\propto \mathcal{L}(\theta; \boldsymbol{y}^*)^m\pi_0(\theta).
\end{aligned}$$

If we let $m \to \infty$, the effect of the initial prior distribution diminishes, and the mth posterior has all of its mass centered at the MLE.

Iterated filtering

Loosely speaking, iterated filtering is just data cloning with the additional pieces:

1. Likelihood cannot be evaluated exactly, it's approximated using particle filters.
2. At each time-step, the parameter particles are perturbed.
3. Parameter particles reweighted using conditional log-likelihoods.

✓ The perturbation of parameters is necessary to avoid particle depletion, a known problem with particle filters + Bayesian inference (Chen et al., 2024).

✗ The perturbations introduce a loss of information (Liu, 2001), so are decreased over the cloning iteration.

Iterated filtering

Loosely speaking, iterated filtering is just data cloning with the additional pieces:

1. Likelihood cannot be evaluated exactly, it's approximated using particle filters.
2. At each time-step, the parameter particles are perturbed.
3. Parameter particles reweighted using conditional log-likelihoods.

✓ The perturbation of parameters is necessary to avoid particle depletion, a known problem with particle filters + Bayesian inference (Chen et al., 2024).

✗ The perturbations introduce a loss of information (Liu, 2001), so are decreased over the cloning iteration.

Iterated Filtering for Panel Models

for $m \in 1:M$ **do**

  Set $\Theta^{F,m}_{0,j} = \Theta^{m-1}_{j} = (\Phi^{m-1}_{j}, \Psi^{m-1}_{1:U,j})$ for $j \in 1:J$;

  for $u \in 1:U$ **do**

    Set $\Theta^{F,m}_{u,0,j} = (\Phi^{F,m}_{u,0,j}, \Psi^{F,m}_{1:U,0,j}) \sim h_{u,0}\big(\,\cdot\,|\Theta^{F,m}_{u-1,j}; \sigma_{u,m}\big)$;

    Initialize $X^{F,m}_{u,0,j} \sim f_{X_{u,0}}(x_{u,0}; \Phi^{F,m}_{u,0,j}, \Psi^{F,m}_{u,0,j})$ for $j \in 1:J$;

    for $n \in 1:N_u$ **do**

      Set $\Theta^{P,m}_{u,n,j} = (\Phi^{P,m}_{u,n,j}, \Psi^{P,m}_{1:U,n,j}) \sim h_{u,n}\big(\,\cdot\,|\Theta^{F,m}_{u,n-1,j}; \sigma_{u,m}\big)$ for $j \in 1:J$;

      $X^{P,m}_{u,n,j} \sim f_{X_{u,n}|X_{u,n-1}}\big(x_{u,n}|X^{F,m}_{u,n-1,j}; \Phi^{P,m}_{u,n,j}, \Psi^{P,m}_{u,n,j}\big)$ for $j \in 1:J$;

      $w^{m}_{u,n,j} = f_{Y_{u,n}|X_{u,n}}\big(y^{*}_{u,n}|X^{P,m}_{u,n,j}; \Phi^{P,m}_{u,n,j}, \Psi^{P,m}_{u,n,j}\big)$ for $j \in 1:J$;

      Draw $k_{1:J}$ with $P(k_j = i) = w^{m}_{u,n,i} / \sum_{v=1}^{J} w^{m}_{u,n,v}$ for $i, j \in 1:J$;

      Set $X^{F,m}_{u,n,j} = X^{P,m}_{u,n,k_j}$, and $(\Phi^{F,m}_{u,n,j}, \Psi^{F,m}_{u,n,j}) = (\Phi^{P,m}_{u,n,k_j}, \Psi^{P,m}_{u,n,k_j})$ for $j \in 1:J$;

      if MARGINALIZE **then**

        $\Psi^{F,m}_{\tilde{u},n,j} = \Psi^{P,m}_{\tilde{u},n,j}$ for all $\tilde{u} \neq u, j = 1:J$

      else

        $\Psi^{F,m}_{\tilde{u},n,j} = \Psi^{P,m}_{\tilde{u},n,k_j}$ for all $\tilde{u} \neq u, j = 1:J$

      end

    end

    Set $\Theta^{F,m}_{u,j} = (\Phi^{F,m}_{u,N_u,j}, \Psi^{F,m}_{u,N_{1:u},j})$ for $j \in 1:J$;

  end

  Set $\Theta^{(m)}_{j} = \Theta^{F,m}_{U,j}$ for $j \in 1:J$;

end

Function Inputs:

- Initializer: $f_{X_{u,0}}(x_{u,0}; \theta)$.
- Process Simulator: $f_{X_{u,n}|X_{u,n-1}}(x_{u,n}|x_{u,n-1}; \theta)$.
- Measurement Model: $f_{Y_{u,n}|x_{u,n}}(y_{u,n}|x_{u,n}; \theta)$.
- Data $y^{*}_{u,n}$.
- Iterations, Particles: (M, J).
- Starting parameter swarm, $\Theta^{0}_{j} = \big(\Phi^{0}_{j}, \Psi^{0}_{1:U,j}\big)$.
- Perturbation simulator: $h_{u,n}(\cdot|\varphi; \sigma)$.
- Variance sequence: $\sigma_{1:U,1:M}$.

The panel iterated filter (PIF) is a type of IF, mitigating information loss (Bretó et al., 2020). It has been successfully used to estimate the MLE previously (e.g., Domeyer et al., 2022).

Theorem (Chapter 4, Theorem 1)

Extends theory of Chen et al. (2024) to panel models. Denote the output of the PIF algorithm as $\Theta_{1:J}^{(M)}$. *Then there exists some positive sequences* $\{C_M\}_{M\geq 1}$ *and* $\{\epsilon_M\}_{M\geq 1}$ *where* $\lim_{M\to\infty} \epsilon_M = 0$ *such that for all* $(J, M) \in \mathbb{N}^2$,

$$E\left[\left|\frac{1}{J}\sum_{i=1}^{J}\Theta_j^{(M)} - \hat{\theta}\right|_2\right] \leq \frac{C_M}{\sqrt{J}} + \epsilon_M$$

Proof sketch and discussion

Conditions:

- On parameter space Θ: Compact, corners not too "sharp" (regular compact set, Def 1 of Chen et al., 2024).
- Regularity conditions on the model (positive and finite likelihood, densities are smooth functions of θ).
- Conditions on the parameter perturbations (type of perturbations, and cooling schedule).

Proof Sketch:

- Chen et al. (2024) provides theory for convergence of IF of POMP models; we write a general PanelPOMP as a POMP model, and PIF is a special case of IF for these models.

Iterated Filtering for panel models: a tradeoff

Iterated filtering is done one observation at a time. In the panel setting, this means we also process one unit u at a time.

Ignoring perturbations, we have do *unit* data cloning, iterating over (m, u):

$$\pi_{m,u}(\theta) \propto \mathcal{L}(\theta; y_u^*)\, \pi_{m,u-1}(\theta) = \mathcal{L}_u(\phi, \psi_u; y_u^*)\, \pi_{m,u-1}(\theta), \tag{4}$$

Using $\pi_{0,0}(\theta)$ as the initial prior distribution. Parameter dependence in posteriors introduced by iterating Eq. 4 over u.

Two options of iterated filtering:

- Perturb all parameters at each time step (IF2, high loss of information).
- When using data from unit u, only perturb ϕ and ψ_u (PIF, high particle depletion).

If the previous prior $\pi_{m,u-1}(\theta)$ has parameter independence: $\pi_{m,u-1}(\theta) = f(\psi_{-u})g(\phi, \psi_u)$, then there is no need to resample particles ψ_{-u}.

This would avoid the particle depletion *and* loss-of-information, and motivates marginalized data cloning (repeating Eqs. 5–6).

$$\tilde{\pi}_{m,u}(\theta) \propto \mathcal{L}_u(y_u^*;\, \phi, \psi_u)\, \pi_{m,u-1}(\theta) \tag{5}$$

$$\pi_{m,u}(\theta) \propto \int \tilde{\pi}_{m,u}(\theta)\, d\phi\, d\psi_u \;\times \int \tilde{\pi}_{m,u}(\theta)\, d\psi_{-u}. \tag{6}$$

Marginalization can happen in various ways. Next slide is a figure with bivariate Gaussian densities, marginalized over all parameters.

If the previous prior $\pi_{m,u-1}(\theta)$ has parameter independence: $\pi_{m,u-1}(\theta) = f(\psi_{-u})g(\phi, \psi_u)$, then there is no need to resample particles ψ_{-u}.

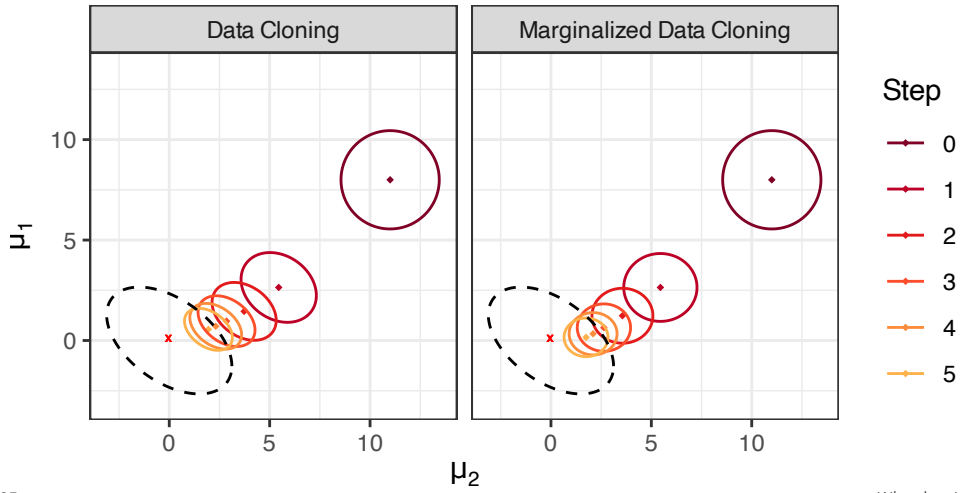
This would avoid the particle depletion *and* loss-of-information, and motivates marginalized data cloning (repeating Eqs. 5–6).

$$\tilde{\pi}_{m,u}(\theta) \propto \mathcal{L}_u(y_u^*;\, \phi, \psi_u)\, \pi_{m,u-1}(\theta) \tag{5}$$

$$\pi_{m,u}(\theta) \propto \int \tilde{\pi}_{m,u}(\theta)\, d\phi\, d\psi_u \;\times \int \tilde{\pi}_{m,u}(\theta)\, d\psi_{-u}. \tag{6}$$

Marginalization can happen in various ways. Next slide is a figure with bivariate Gaussian densities, marginalized over all parameters.

Marginalized Bayes: Guassian figure



Just as IF extends data cloning (perturbations + particle approximation), the MPIF algorithm extends marginalized data cloning.

- Existing theory for IF algorithms rely heavily on the data cloning principle Ionides et al. (2015); Chen et al. (2024).
- The non-linearity introduced by the marginalization step invalidates these approaches.
- A natural first question is whether or not marginalized data cloning converges.
  - Unfortunately, a few toy examples suggests not always (computationally and analytically).

Convergence is explored via Gaussian likelihoods.

The properties of this special case is relevant to the broader class of models that is well approximated by Gaussian models, (e.g., local asymptotic normality (Le Cam and Yang, 2000)).

Theorem (Chapter 4, Theorem 2)

Let $\mathcal{L}_u(y_u^*;\, \phi, \psi_u)$ be the likelihood that corresponds to a Gaussian distribution with mean (ϕ^*, ψ_u^*) and precision $\Lambda_u^* \in \mathbb{R}^{2\times 2}$. Under suitable conditions on the matrices Λ_u^*, then if the initial prior density is Gaussian, then the density of the mth iteration of Eq. 6 converges to a point mass at the MLE $(\phi^*, \psi_1^*, \ldots, \psi_U^*)$ as $m \to \infty$.

- Gaussian priors + Gaussian likelihoods imply Gaussian posteriors.
- Transform data so likelihood is centered at zero.
- The marginalization step only modifies the covariance, setting off-diagonal terms to zero. Conditions ensure this loss of information is not too large.
- Diagonals of covariance shrink to zero asymptotically at rate $1/m$.
- Each unit-iteration updates the ϕ and ψ_u components of μ_m.
- $\mu_m = \left(\prod_{i=1}^{m}\prod_{u=1}^{U} A_{u,i}\right)\mu_0 = \left(\prod_{i=1}^{m} P_m\right)\mu_0$, with $\|P_m\|_2 = 1 - \epsilon_m/m + o(1/m)$, with ϵ_m positive, bounded.

Theorem (Chapter 4, Corollary 1)

Consider the same setup as Theorem 2 (Chapter 4). If parameters are perturbed prior performing the Bayes update at each step using Gaussian additive noise with mean 0 and covariance $\sigma_m^2 \Sigma_0$, then if $\sigma_m^2 = o(1/m)$, the algorithm still converges to the MLE as $m \to \infty$.

- Useful heuristic: a more dispersed prior typically results in a posterior distribution closer to the likelihood function.
- Adding Gaussian noise at each step results in larger updates towards to MLE at each step.
- Heuristically, convergence of unperturbed case implies convergence of perturbed case, if perturbation variance decreases to zero.

The MPIF advantage: improved particle representation

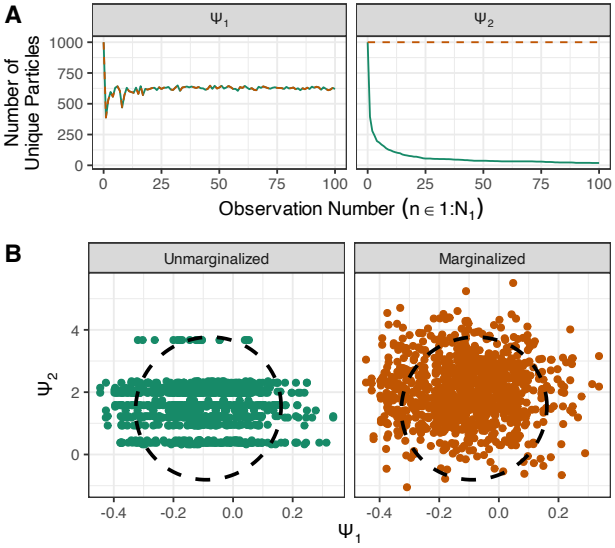
One perspective is that the marginalization adds a small amount of bias in each intermediate posterior distribution.

- Theorem 2 (Chapter 4) gives formal conditions where the bias at each step is small, and algorithm converges to MLE (no bias).
- The advantage is improved particle representations. In this case, MPIF may still be preferable even if there is small bias.

Thus, we can think of MPIF as improving a bias-variance tradeoff.

- To demonstrate this, perform a single unit-iteration of PIF and MPIF on Gaussian data and model.

Bias-Variance tradeoff

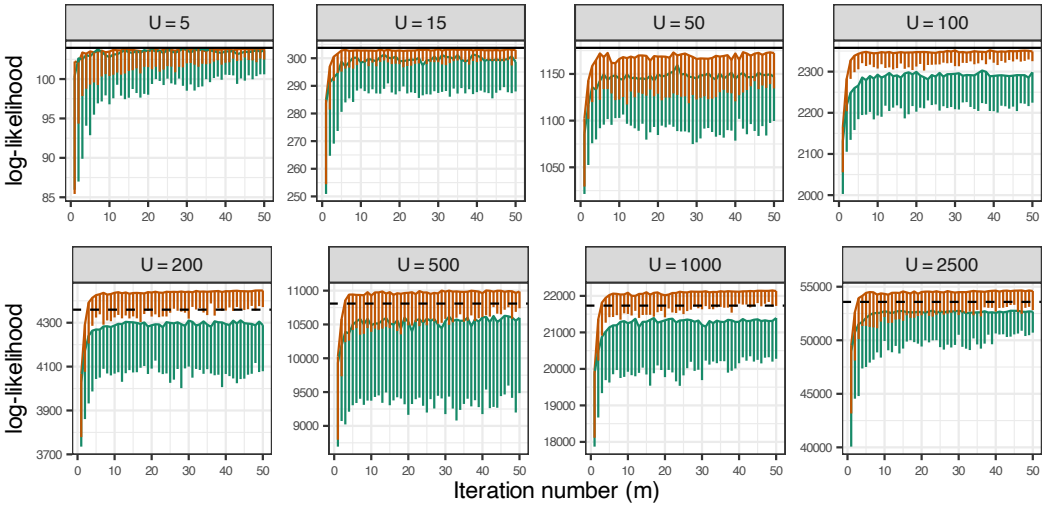


Gompertz population model: high-dimensional space

The MPIF algorithm is demonstrated on data simulated from a stochastic population model. We have U ranging from $5 - -2500$, and $N_u = N \in \{20, 50, 100\}$.

- $X_{u,n+1} = K_u^{1-\exp r_u} \epsilon_{u,n}$, $\epsilon_{u,n} \sim N(0, \sigma_u^2)$.
- $Y_{u,n} | X_{u,n} \sim N\big(\log X_{u,n}, \tau_u^2\big)$
- Fix $K_u = 1$, $X_{u,0} = 1$. Estimate $\sigma_u^2 = \sigma^2 = 0.01$, $r_u = r = 0.1$, and $\tau_u^2 = 0.01$ for all u.

This is log-linear Gaussian, so exact likelihood can be computed (Kalman, 1960). Parameter choices match Bretó et al. (2020).


U = 5
U = 15
U = 50
U = 100
U = 200
U = 500
U = 1000
U = 2500
log-likelihood
log-likelihood
Iteration number (m)
Marginalized
TRUE
FALSE

Discussion

- In all tested models, the MPIF algorithm outperforms PIF, especially as the number of units U is large, and number of unit-specific parameters is large.
- Even poor performing replicates of MPIF often outperform PIF, despite having weaker theoretical guarantees.
- Improved performance is a result of reduced particle depletion.
- Stronger theory for MPIF is available in some special cases, each covered by Theorem 1 of Chapter 4:
  - No shared parameters: there is no parameter dependence, and the algorithm is equivalent to performing IF2 independently to each model.
  - No unit-specific parameters: the algorithm is the same as PIF, as resampling unit-specific parameters is not needed.

5. Conclusion and Future Directions

Summary

Likelihood based inference of SSMs is a challenging task. This thesis presents novel research related to various aspects of this problem:

- Chapter 2 proposes new methodology for ARIMA models, perhaps the most used type of SSM. It demonstrates existing algorithms fail to properly maximize likelihoods, and fixes this with limited additional computational effort.
- Chapter 3 is an application that proposes new standards for using mechanistic models to inform policy. It demonstrates some strengths and weaknesses of existing approaches, and how recent methodological advances can be used to aid this task.
- Chapter 4 proposes new methodology to help perform maximum likelihood estimation for high-dimensional panel models, and provides theory for the proposed approach in special cases. It also extends existing theory for iterating filtering on panel models.

The importance of ARIMA models to modern science additional developments related to this chapter may be worth the effort. Particularly well-suited for an undergraduate research project(s):

- Building a Python library to implement the existing algorithm.
- Consider stratified sampling of root initializations (i.e., sample from each quadrant rather than randomly).
- Develop theoretical bounds on the number of local optima, resulting in improved stopping criteria.
- Leveraging iterated filtering to do non-Gaussian ARMA models and/or in panel models.

The lessons learned from this retrospective analysis may be relevant in other disciplines. Additionally, there are a number of things I learned that didn't make the final paper:

- The importance of initialization and measurement models. These are often overlooked, but are key to fitting insightful models.
- Uncertainty propagation: the natural approach is perform Monte Carlo adjusted profile confidence intervals (Ionides et al., 2017), giving a large number of parameter values, and sample these using their likelihoods (Empirical Bayes). If dimensions are high, the Monte Carlo variance is high, so you don't get to resample many parameters this way.

The theory developed for both PIF and MPIF gives rise to a few extensions:

- Current theory for MPIF ignores particle approximations, and is limited to Gaussian densities.
- Can Gaussian bias be bounded? Bias negated via perturbations?
- Nested levels of shared and unit-specific parameters.

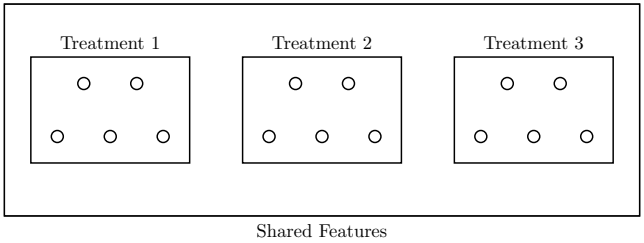


Figure 3: A diagram of a controlled experiment with replications.

Future Directions: General

Parameter perturbations introduced in IF result in loss-of-information. Recent work in automatic differentiation can help avoid this issue, though appears to require IF-algorithms to initialize (Tan et al., 2024).

MPIF can be particularly useful for this algorithm in panel settings. Related to this extension:

- Software for panel models with auto-diff (pypomp, work ongoing with other students).
- Using methodology to build shallow neural networks to model unknown mechanisms (e.g., Noordijk et al., 2024).

Future Directions: Applications

My interest in methodology in SSMs is in part driven by my interest in the types of problems they solve. Some potential applications include:

- Fisheries; long history of SSM, but particle filter / IF approach hasn't caught on. Interesting to know if this approach is useful in fisheries, or if there are lessons there to be learned.
  - Concrete examples include: modeling disease progression in native cutthroat species, modeling reproductive dynamics of native cutthroat with stocked rainbow trouts (threatening native reproduction) (Rosenthal et al., 2022).
- Disease progression in farms: plants and livestock (Skølstrup et al., 2022).

Acknowledgements

There are so many people that I would like to thank, impossible to thank everyone who has helped and supported me.

I would like to give a special thanks to my advisor (Edward L. Ionides), dissertation committee (Aaron A. King, Kerby Shedden, Jeffery Regier), my family, classmates, and friends.

Section 6

References

Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp (2002). A tutorial on particle filters for online nonlinear, non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing 50*, 174 – 188.

Baker, R. E., J.-M. Pena, J. Jayamohan, and A. Jérusalem (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology letters 14*(5), 20170660.

Bretó, C., E. L. Ionides, and A. A. King (2020). Panel data analysis via mechanistic models. *Journal of the American Statistical Association 115*(531), 1178–1188.

References (ii)

Chen, Y., M. Gerber, C. Andrieu, and R. Douc (2024). Self-organizing state-space models with artificial dynamics. *arXiv preprint arXiv:2409.08928*.

Domeyer, J. E., J. D. Lee, H. Toyoda, B. Mehler, and B. Reimer (2022). Driver-pedestrian perceptual models demonstrate coupling: Implications for vehicle automation. *IEEE Transactions on Human-Machine Systems 52*(4), 557–566.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods*, Volume 38. OUP Oxford.

Gardner, G., A. C. Harvey, and G. D. A. Phillips (1980). 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) 29*(3), 311–322.

Hogg, D. W. and S. Villar (2024). Is machine learning good or bad for the natural sciences? *arXiv preprint arXiv:2405.18095*.

References (iii)

Ionides, E. L., C. Breto, J. Park, R. A. Smith, and A. A. King (2017). Monte Carlo profile confidence intervals for dynamic systems. *Journal of the Royal Society Interface 14*, 1–10.

Ionides, E. L., D. Nguyen, Y. Atchadé, S. Stoev, and A. A. King (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences of the USA 112*(3), 719––724.

Ionides, E. L., N. Ning, and J. Wheeler (2024). An iterated block particle filter for inference on coupled dynamic systems with shared and unit-specific parameters. *Statistica Sinica 34*, 1241–1262.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*(1), 35–45.

Le Cam, L. and G. L. Yang (2000). *Asymptotics in Statistics* (2nd ed.). New York: Springer.

References (iv)

Lee, E. C., D. L. Chao, J. C. Lemaitre, L. Matrajt, D. Pasetto, J. Perez-Saez, F. Finger, A. Rinaldo, J. D. Sugimoto, M. E. Halloran, I. M. Longini, R. Ternier, K. Vissieres, A. S. Azman, J. Lessler, and L. C. Ivers (2020). Achieving coordinated national immunity and cholera elimination in Haiti through vaccination: A modelling study. *The Lancet Global Health 8*(8), e1081–e1089.

Lele, S. R., B. Dennis, and F. Lutscher (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters 10*(7), 551–563.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Noordijk, B., M. L. Garcia Gomez, K. H. W. J. ten Tusscher, D. de Ridder, A. D. J. van Dijk, and R. W. Smith (2024, August). The rise of scientific machine learning: A perspective on combining mechanistic modelling with machine learning for systems biology. *Frontiers in Systems Biology 4*.

Pan American Health Organization (2023). Cholera epidemic in Hispaniola 2023 - situation report 19.

Ranjeva, S. L., E. B. Baskerville, V. Dukic, L. L. Villa, E. Lazcano-Ponce, A. R. Giuliano, G. Dwyer, and S. Cobey (2017). Recurring infection with ecologically distinct HPV types can explain high prevalence and diversity. *Proceedings of the National Academy of Sciences 114*(51), 13573–13578.

Ripley, B. D. (2002). Time series in R 1.5.0. *The Newsletter of the R Project Volume 2*, 2.

Rosenthal, W. C., J. M. Fennell, E. G. Mandeville, J. C. Burckhardt, A. W. Walters, and C. E. Wagner (2022). Hybridization decreases native cutthroat trout reproductive fitness. *Molecular Ecology 31*(16), 4224–4241.

References (vi)

Searle, C. L., M. H. Cortez, K. K. Hunsberger, D. C. Grippi, I. A. Oleksy, C. L. Shaw, S. B. de la Serna, C. L. Lash, K. L. Dhir, and M. A. Duffy (2016). Population density, not host competence, drives patterns of disease in an invaded community. *The American Naturalist 188*(5), 554–566.

Skølstrup, N. K., D. B. Lastein, L. V. de Knegt, and A. R. Kristensen (2022). Using state space models to monitor and estimate the effects of interventions on treatment risk and milk yield in dairy farms. *Journal of Dairy Science 105*(7), 5870–5892.

Tan, K., G. Hooker, and E. L. Ionides (2024). Accelerated inference for partially observed Markov processes using automatic differentiation.

Trevisin, C., J. C. Lemaitre, L. Mari, D. Pasetto, M. Gatto, and A. Rinaldo (2022). Epidemicity of cholera spread and the fate of infection control measures. *Journal of the Royal Society Interface 19*(188), 20210844.

References (vii)

UNICEF (2022). With UNICEF support, Haiti kickstarts campaign to immunize about 1.7 million people against cholera. https://www.unicef.org/lac/en/press-releases/with-unicef-support-haiti-kickstarts-campaign-to-immunize-about-1.7-million-people-against-cholera.

Wheeler, J., A. Rosengart, Z. Jiang, K. Tan, N. Treutle, and E. L. Ionides (2024). Informing policy via dynamic models: Cholera in Haiti. *PLOS Computational Biology 20*(4), 1–31.