Innovations in Likelihood-Based Inference for State Space Models

Oral Defense

Jesse Wheeler

May 23, 2025

Department of Statistics, University of Michigan

Table of contents

1. Introduction

Synonyms and Definitions

There are several terms that have been used as synonyms with state space models (SSMs):

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

I chose SSM as it is the terminology often preferred by practitioners.

State Space Models (SSM)

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

- Let $Y_1, Y_2, \ldots, Y_N$ be random variable representing the observed time series. These 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.

I will adopt the shorthand $t_{1:N} = (t_1, \ldots, t_N)$, $Y_{1:N} = (Y_1, \ldots, Y_N)$, and $X_{1:N} = (X_1, \ldots, X_N)$.

When defining a SSM, we often want to include an initial value for the latent states, X_0.

Likelihood function

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 (typically Lebesgue or a counting measure), where θ is a parameter vector $\theta \in \mathbb{R}^{d_\theta}$ that indexes the model.

The difficulty in likelihood-based inference for these models is a result of only $Y_{1:N}$ being observable, and thus the likelihood function involves a high-dimensional integral:

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

POMP models

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

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

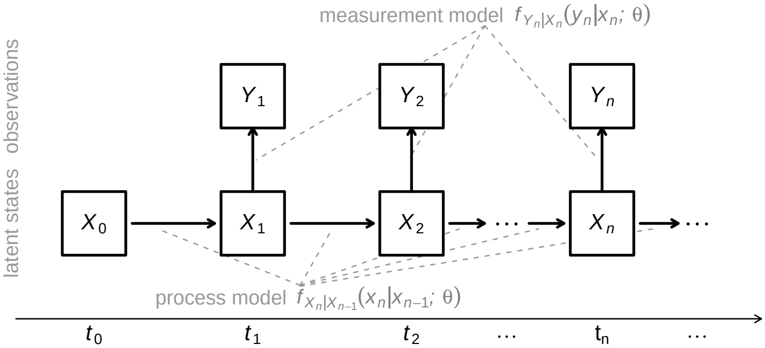


Figure 1: A flow diagram representing an arbitrary POMP model. Modified figure from SBIED course (King, Ionides).

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 mimicing real-world mechanisms.

Hidden Markov Model (HMM)

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

Remaining Chapters and Outline

- Inference for ARMA models.
- Mechanistic models for modeling cholera outbreak in Haiti.
- The marginalized panel iterated filter (MPIF) algorithm.

2. Likelihood Maximization for ARMA models

ARMA models are the most frequently used approach to modeling time series data. ARMA models are as foundational to time series analysis as linear models are to regression analysis, and they are often used in conjunction for regression with ARMA errors.

ARMA model definition

A time series $Y_{1:N}$ is called 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 $\phi_p \neq 0$, $\varphi_q \neq 0$.

We refer to the positive integers p and q of Eq. (3) as the autoregressive (AR) and moving average (MA) orders, respectively.

State space formulation (i)

For practitioners, ARMA models do not appear to be SSMs. However, inference methodology treats ARMA models as *non-mechanistic* SMMs. Let $r = \max(p, q+1)$, and we now define

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

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

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. To demonstrate this, we fit an ARMA$(2, 2)$ and an ARMA$(2, 1)$ model to data generated from an ARMA$(2, 2)$ model. The ARMA$(2, 1)$ is formally a special case of an ARMA$(2, 2)$ model, with $\varphi_2 = 0$.

In **R**, we draw a single instance from Model class 2: $y^*_{1:100} \sim \mathrm{ARMA}(2, 2)$ with:

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

Fitting ARMA models

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.

Fitting ARMA models

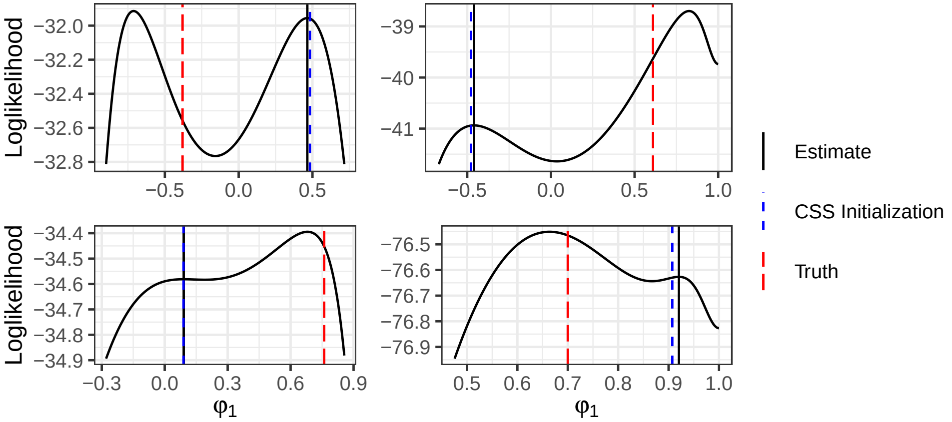
The Gardner et al. (1980) is the standard method for fitting ARMA model parameters. It is implemented in the base **stats** package in R, as well as the **statsmodels** module in Python.

```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.

Convergence to local optima

The difficulty is that the likelihood surface is often multimodal, and the existing procedure runs the risk of converging to a local solution (Ripley, 2002).



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 of possible initialization.
  - Constraints are on the roots of polynomials formed by model parameters, not directly on parameters themselves.

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 of possible initialization.
  - Constraints are on the roots of polynomials formed by model parameters, not directly on parameters themselves.

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

Now we'll fit the exact same models using the **arima2** package:

```r
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 -141.2, and the likelihood of **mod2v2** is -141.2.

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

Now we'll fit the exact same models 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 -141.2, and the likelihood of **mod2v2** is -141.2.

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 statistical models. Why does this project matter?

- 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 statistical models. Why does this project matter?

- 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 statistical models. Why does this project matter?

- 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 statistical models. Why does this project matter?

- 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 statistical models. Why does this project matter?

- 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).

Introduction

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

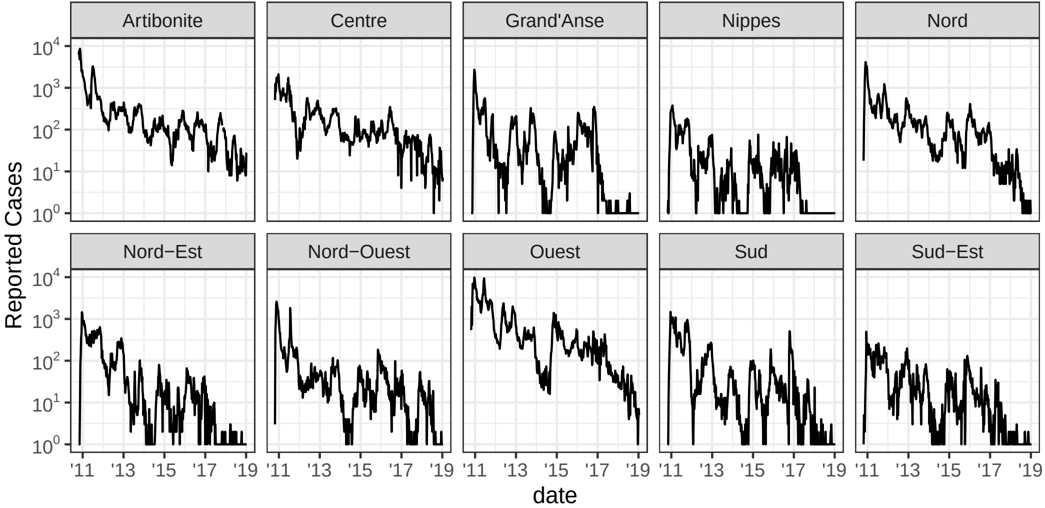
- Haiti experienced a cholera outbreak following the devastating 2010 earthquake.
- From 2010-2019, more than 800,000 recorded cases, making it one of the largest recorded outbreaks.
- Oral cholera vaccination (OCV) is available, but in limited supply.
- Image credit: UNICEF (2022).

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

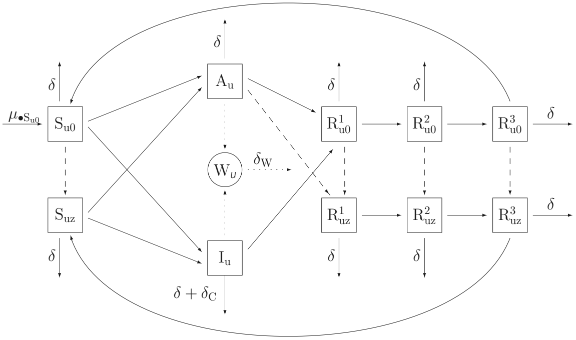
- The distinct teams each concluded predicted cholera resurgence from Feb 2019 - Feb 2024.
- There were no confirmed cases from Feb 2019 - Sep 2022 (Trevisin et al., 2022).
- Though there were some cases recently recorded, not near the predicted scale (Pan American Health Organization, 2023).

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).

Model Fitting

The iterated block particle filter (IBPF) was used 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.

Key findings

- 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)

Test

5. Concluding Remarks

Summary

Get the source of this theme and the demo presentation from

`https://gitlab.com/RomainNOEL/beamertheme-gotham`

Section 6

References

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.

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*.

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.

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.

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

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

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.

UNICEF (2022). 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.