SAND2022-9080 C

MLDL

Machine Learning and Deep Learning Conference 2022

Data-Driven Model-Form Uncertainty with Bayesian Statistics and Neural Differential Equations

Presenter: Erin C.S. Acquesta (5554)

Teresa Portone* (1463), Raj Dandekar (MIT), and Chris Rackauckas (MIT)

CIS LDRD: Trusted AI

*Originally presented by Teresa Portone at ECCOMAS 2022, June 9, Olso, Norway.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Abstract

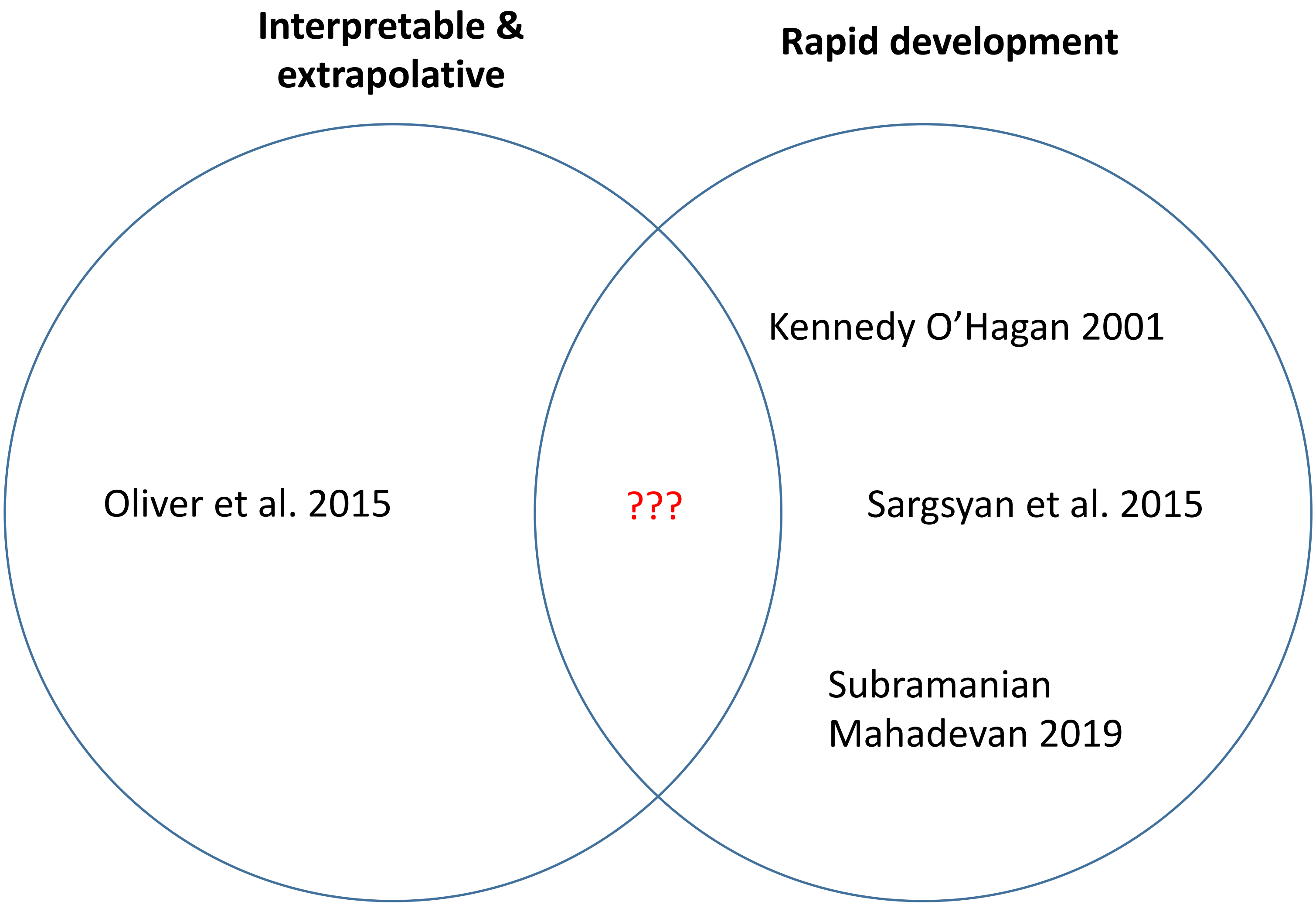
Modeling real-world phenomena to any degree of accuracy is a challenge that the scientific research community has navigated since its foundation. Insufficient knowledge, such as inability to observe or represent all the relevant phenomena, induces uncertainty in the appropriate model form. Characterizing this model-form uncertainty (MFU) is essential to understanding the reliability of predictions made with these models, especially when such predictions inform high-consequence decisions. Here we present a novel model-form uncertainty representation which combines Bayesian statistics with Universal Differential Equations [1], a powerful new approach to data-driven modeling wherein a universal function approximator (a neural network in this work) is embedded within a known differential-equation model at the source of MFU. The neural network is endowed with a probabilistic representation and is updated using available observational data in a Bayesian framework. By representing the MFU explicitly and deploying an embedded, data-driven model, this approach enables an agile, expressive, and interpretable method for representing MFU.

Motivation

Develop an

- Agile
- Interpretable
- Extrapolative

Method to represent model-form uncertainty (MFU) for predictions



Algorithmic Approach

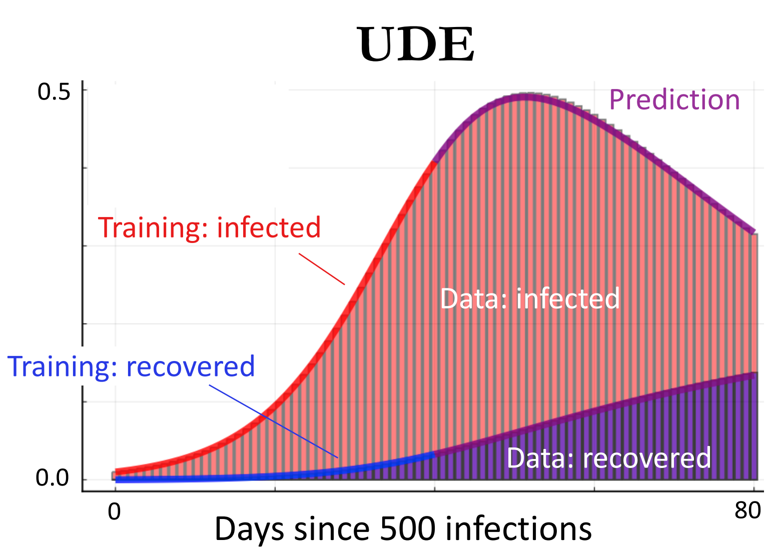
Combine universal differential equations (UDEs) with Bayesian statistics to represent MFU

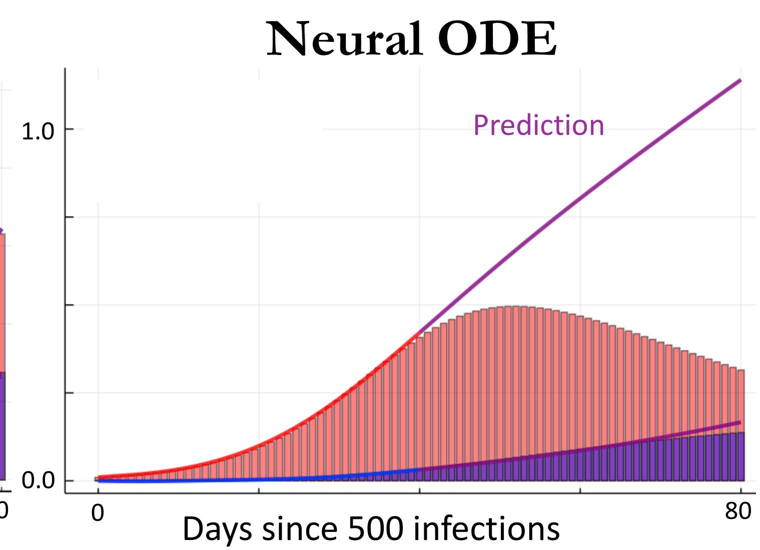
- UDEs embed ML models within existing scientific models:
- $\boldsymbol{u}' = F(\boldsymbol{u}, t, NN_\theta(\boldsymbol{u}))$

$$\min_\theta \|\boldsymbol{d} - \boldsymbol{u}(\theta)\|$$

- Data driven, **BUT**
- Time-independent parameterization
- Can respect physical principles by construction
- Can be more predictive than Neural ODEs:

$$\boldsymbol{u}' = NN_\theta(\boldsymbol{u})$$





Algorithmic Approach

Combine universal differential equations (UDEs) with Bayesian statistics to represent MFU

UDEs successfully used in a deterministic setting to find "model corrections" or "missing dynamics."

Data not always informative enough to identify a single "model correction."

By endowing UDEs with a Bayesian parameterization, can we use them to represent model-form uncertainty?

Bayesian UDEs

Challenges:

- NNs notoriously challenging to train even in deterministic setting.
- Traditional Bayesian methods computationally challenging & suffer from curse of dimensionality.

Are Bayesian UDEs a feasible approach to representing MFU?

- Explored this in context of Bayesian NN embedded in a compartment-based disease model.

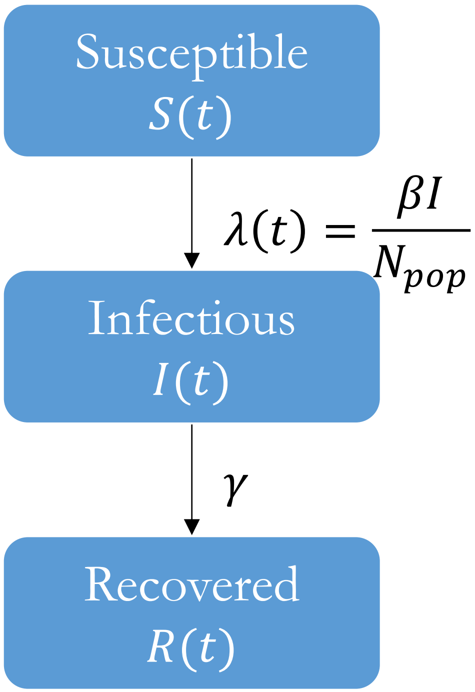
Exemplar: Compartmental Disease Models

$$\frac{dS}{dt} = -\frac{\beta IS}{N_{pop}}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N_{pop}} - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

$$N_{pop} = S(t) + I(t) + R(t)$$



- SIR a common, simple model of disease spread.
- Doesn’t account for infected population quarantine as we saw for COVID-19.
- Quarantine dynamics could depend nonlinearly on state variables.

Exemplar: Compartmental Disease Models

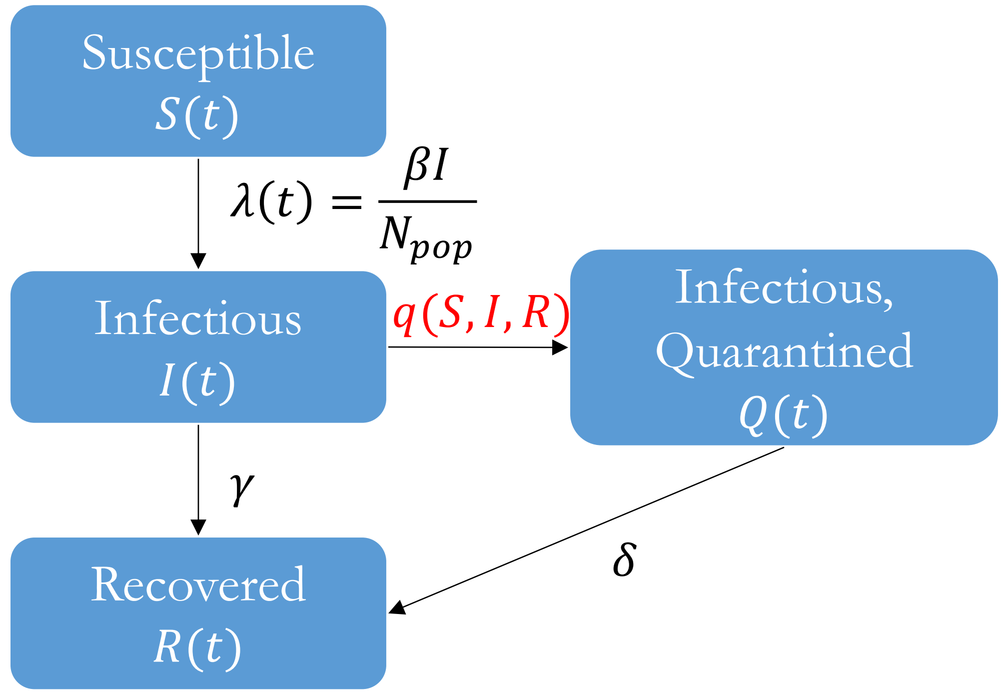
$$\frac{dS}{dt} = -\frac{\beta I S}{N_{pop}}$$

$$\frac{dI}{dt} = \frac{\beta I S}{N_{pop}} - \gamma I - q(S, I, R) I$$

$$\frac{dR}{dt} = \gamma I + \delta Q$$

$$\frac{dQ}{dt} = q(S, I, R) I - \delta Q$$

$$N_{pop} = S(t) + I(t) + R(t) + Q(t)$$



- Represent nonlinear transition into quarantine with a small neural network.
- Constrained to conserve population by construction.

Bayesian UDE Study

Inferring disease parameters $[\beta, \gamma, \delta]$ along with NN parameters

Prior

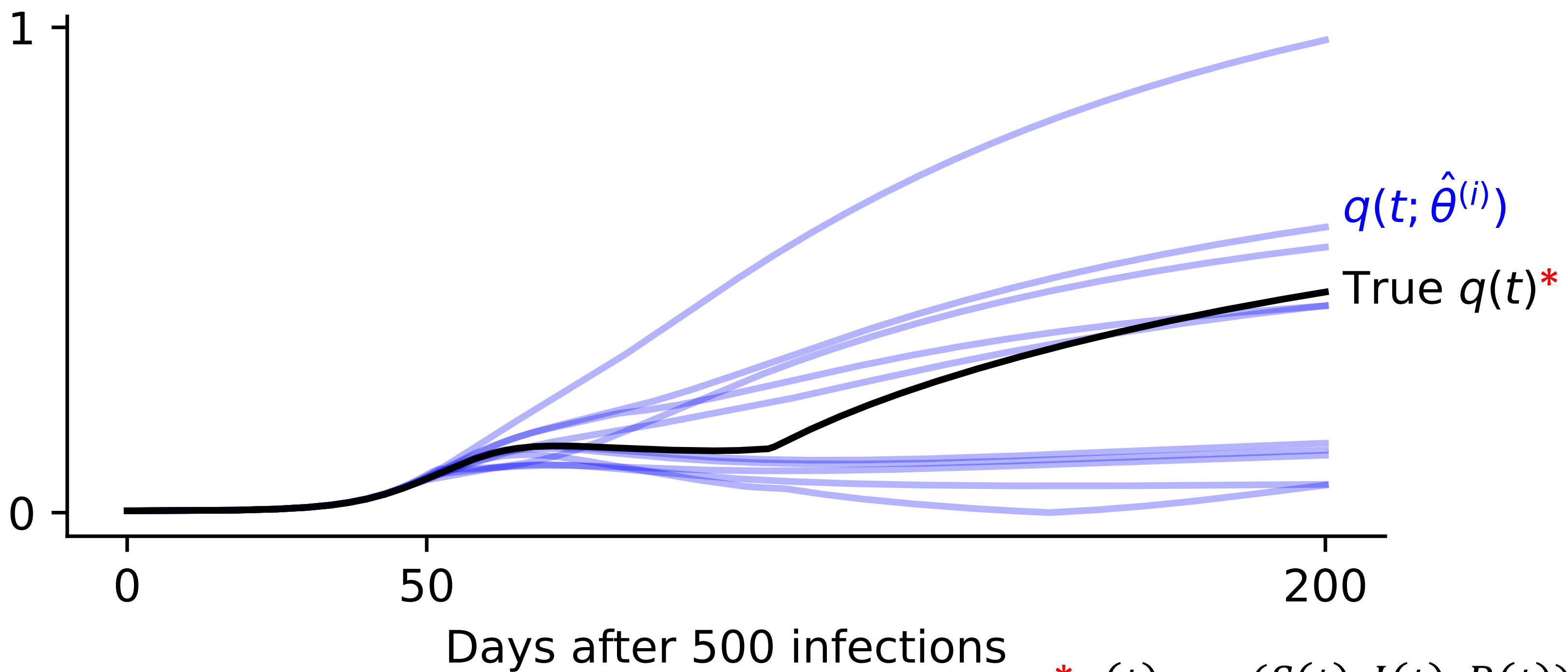
- Disease parameters ~ $U(0,2)$
- 51 NN parameters ~ $N(0, (50)^2)$

Likelihood

- Synthetic data generated from SIRQ model
- Calibration data = observations of I, R, Q first 50 days after 500 infections.
- No noise added; likelihood assumes 95% confidence bound of $\pm 10\%$ error, i.e.

$$d = s + \epsilon, \qquad \epsilon \sim N(0, \sigma^2), \qquad 2\sigma = \pm 0.1s$$

MAP Estimates Differ by Initial Guess



*$q(t) \equiv q(S(t), I(t), R(t))$

Multiple parameter combinations reproduce calibration data.

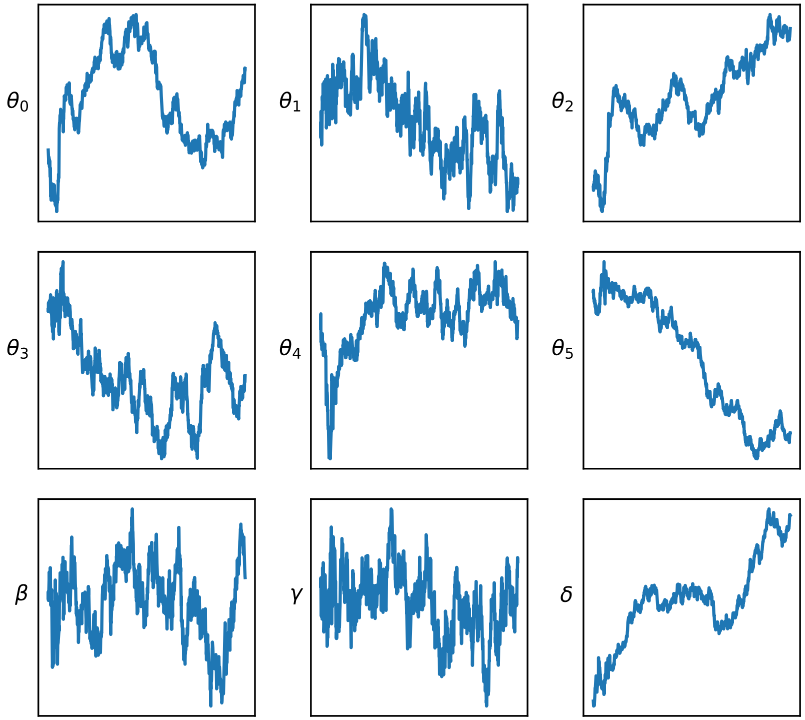
Posterior Approximation

Seeded posterior approximations at MAP point.

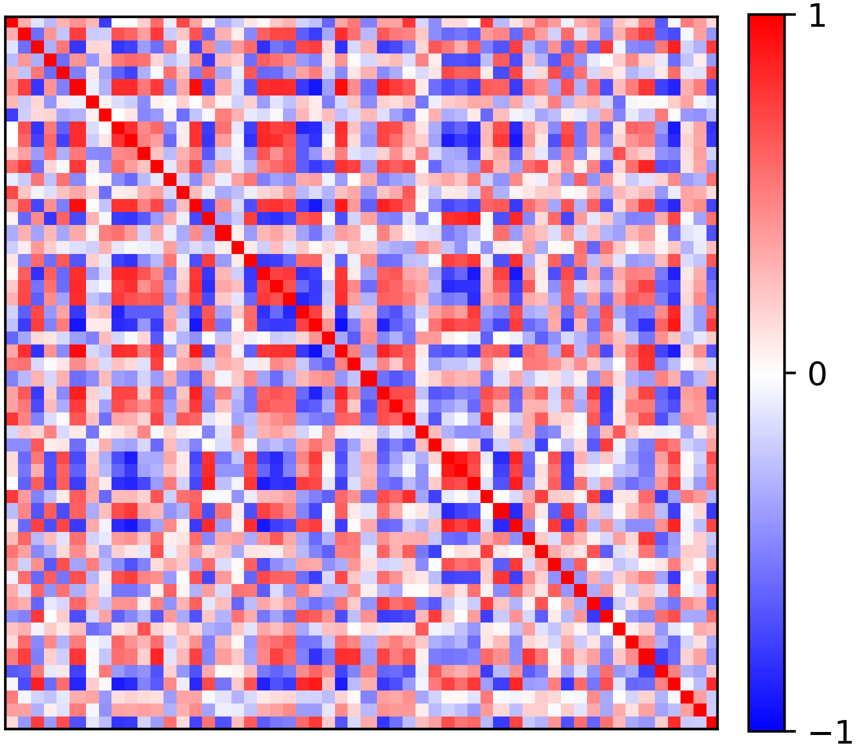
Method: NUTS

- HMC variant, derivative based

NUTS posterior chains

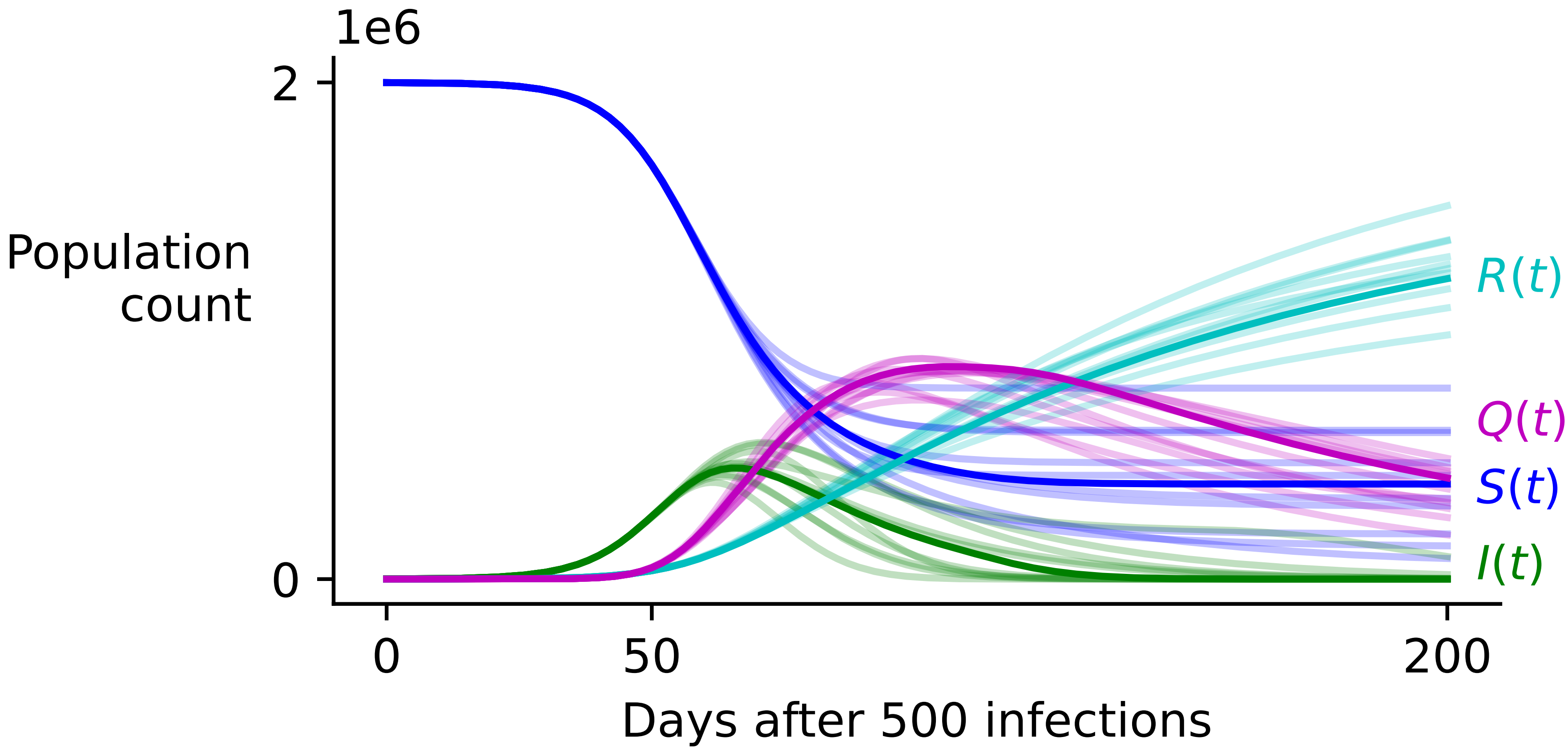


Correlations indicate complex posterior structure



2000 steps / Average acceptance rate: 0.86 / Adaptive step size

Ensemble of MAP Estimates



Conclusions

Despite lower-d NN, Bayesian inference challenging.

Posterior likely multimodal, non-Gaussian.

Next steps

- Sparsity-inducing priors
- Estimate posterior with Gaussian mixture model
- Goal-oriented Bayesian inference

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

© 2022