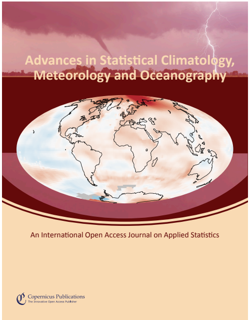
Statistical parameter estimation and inference for dynamical models: Disease modeling

Jennifer Hoeting, Colorado State University

Collaborators:
  Libo Sun, Johnson & Johnson
  Chihoon Lee, Stevens Institute of Technology
  N. T. Hobbs, Colorado State University

April 2017

Thank you Peter!



Advances in Statistical Climatology, Meteorology and Oceanography

An International Open Access Journal on Applied Statistics

Copernicus Publications

ASCMO

Open Access

NEW JOURNAL

Advances in Statistical Climatology, Meteorology, and Oceanography (ASCMO)

- ASCMO is an interdisciplinary journal that publishes cutting-edge scientific advances and statistical methods.
- ASCMO serves at the interface between statistics and the atmospheric and oceanic sciences.
- Journal Website: ascmo.net

Modern statistics is a multidisciplinary discipline. It works across boundaries.

P. Guttorp and G. Lindgren, 2016

Ecology of infectious diseases

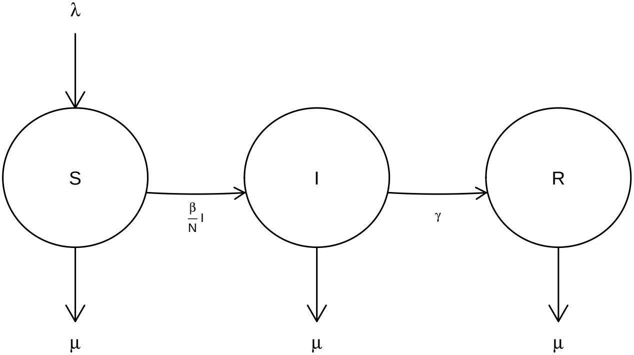
Ecological models can be deterministic or stochastic

Bolker, Ecological Models and Data in R, 2008

Classification of modeling approaches

Deterministic	Stochastic
Mathematical	Statistical
Mechanistic	Phenomenological
Process	Pattern

Models for infectious diseases



Susceptible-Infected-Recovered (SIR) model for disease transmission where the state variables are described by a set of differential equations.

Challenges in the ecology of infectious diseases

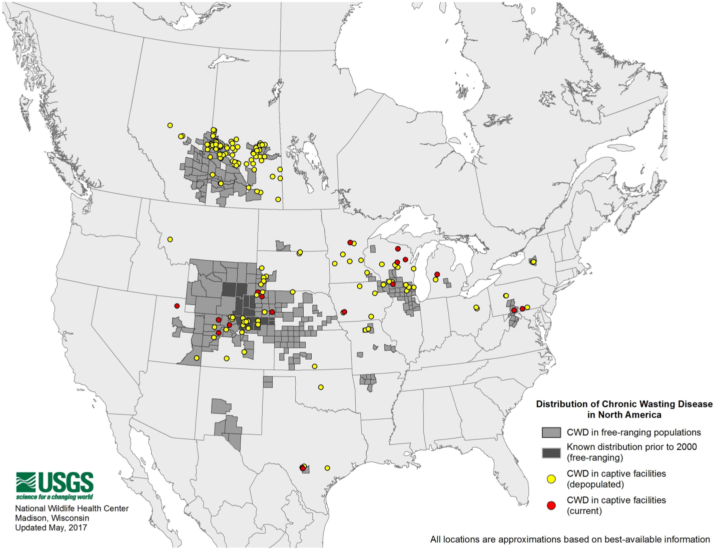
- Data can be messy and sparse
- You have to find the animals who are sick, they don't visit the nearest health clinic
- You need knowledge of mathematical biology
- You need to able to synthesize a broad range of statistical methods

Sounds like fun!

Chronic Wasting Disease

Deer (female) with
Chronic Wasting Disease

Healthy deer (male)



April 2016: The first detection of CWD in Europe

CWD Transmission

- Deer were held at the Colorado Division of Wildlife in Fort Collins, Colorado
- Annual observations of cumulative mortality from two CWD epidemics in captive mule deer
- Epidemic 1: 1974 to 1985
  Epidemic 2: 1992 to 2001 (in a new deer herd)
- 21 observations over time
- The dataset also includes the annual number of new deer added to the herd and the per capita losses due to natural deaths and removals

CWD Transmission

We develop a type of Susceptible-Infected-Recovered (SIR) model for disease transmission where the state variables are described by a set of differential equations.

Consider the state vector $\mathbf{X}(t) = (S(t), I(t), C(t))^T$, where

- S is the number of susceptible animals,
- I is the number of infected animals,
- C is the cumulative number of deaths from CWD over time.

Only C is observed and the other two state variables, S and I, are unobserved.

Direct transmission ODE model for CWD

$$\begin{aligned} dS &= [a - S(\beta I + m)]\, dt \\ dI &= [\beta S I - I(\mu + m)]\, dt \\ dC &= \mu I\, dt \end{aligned}$$

where

- β is the transmission coefficient
- μ is the per capita CWD mortality rate

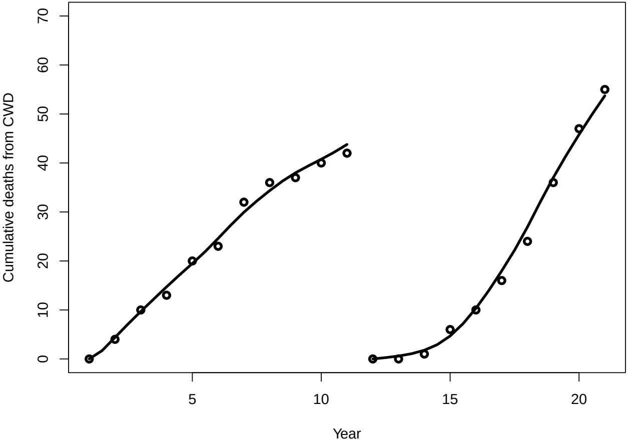
unknown

- a is the number of susceptible animals annually added to the population via births or importation
- m is the per capita natural mortality rate

known

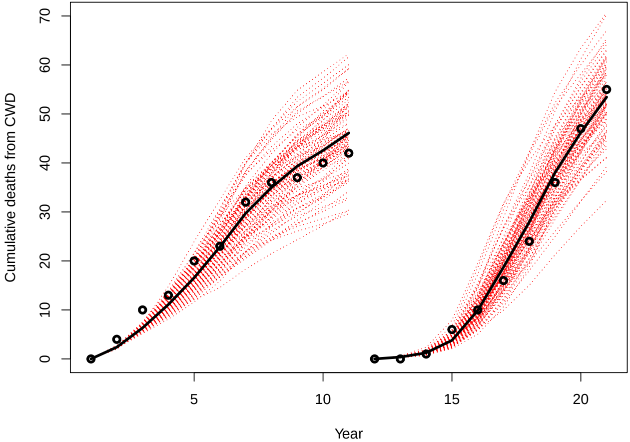
We assume $\mathbf{X}(0) = (S(0), I(0), C(0))^T$ are (un)known initial conditions.

Ordinary Differential Equation Model for CWD



Deterministic dynamical models can be used to determine whether or not transmission will occur.

Stochastic Differential Equation Model for CWD



Stochastic dynamical models

- Can be used to determine the probability of disease transmission between two individuals
- Allow more realistic description of the transmission of disease.

Direct transmission SDE Model for CWD

A SDE model for direct transmission of CWD is given by

$$\begin{aligned}
dS =& [a - S(\beta I + m)]dt + B_{11} dW_1 + B_{12} dW_2 + B_{13} dW_3, \\
dI =& [\beta SI - I(\mu + m)]dt + B_{21} dW_1 + B_{22} dW_2 + B_{23} dW_3, \\
dC =& \mu I dt + B_{31} dW_1 + B_{32} dW_2 + B_{33} dW_3,
\end{aligned}$$

where

- initial condition $\mathbf{X}(0) = (S(0), I(0), C(0))^T$ assumed known
- $\mathbf{W}$ is a k-dimensional standard Wiener process.
- $B = (B_{ij}) = \sqrt{\Sigma}$ with

$$\Sigma = \begin{bmatrix} a + S(\beta I + m) & -\beta SI & 0 \\ -\beta SI & \beta SI + I(\mu + m) & -\mu I \\ 0 & -\mu I & \mu I \end{bmatrix}.$$

Methods

Some methods for statistical inference for dynamical disease models

- Method 1: MLE for SDEs
- Method 2: Model selection for dynamical models
- Method 3: POMP models
- Method 4: Approximate an ODE via smoothing
- More methods: see talks by J. Xu, V. Minin

Method 1: Parameter estimation for SDEs

Goal: Statistical inference for SDE model parameters

Challenges:

- Multivariate state space
- Data are partially observed, discrete, sparse.
- The transition density between two observations is typically unknown.
- Likelihood functions involving SDEs are computationally expensive

Method 1: Parameter estimation for SDEs

Sun, Lee, Hoeting 2015 CSDA

- Automatic tuning of importance sampler via an auxiliary parameter
- Parameter estimation via
  penalized simulated maximum likelihood
  where we simultaneously
  - estimate dynamical model parameters $\boldsymbol{\theta}$ and
  - optimize the importance sampler using parameter ρ to minimize the variation of the Monte Carlo approximation of the transition density.

Method 1: Penalized simulated maximum likelihood (PSML)

The PSML estimator $(\hat{\boldsymbol{\theta}}, \hat{\rho})$ is defined by

$$(\hat{\boldsymbol{\theta}}, \hat{\rho}) = \text{arg max} \sum_{i=1}^{n} \log\left(\frac{1}{J}\sum_{j=1}^{J} h_\rho\right) \text{ subject to } \sum_{i=1}^{n} \widehat{\text{cv}}\,(h_\rho) \leq s,$$

where

- $s \geq 0$ is a tuning parameter
- h_ρ is the importance sampling ratio where ρ is an auxiliary parameter used to tune the importance sampling function.
- $\widehat{\text{cv}}(h_\rho)$ is the sample coefficient of variation of h_ρ

Method 2: Model selection for dynamical models

Hierarchical models

- **Stage 1:** Observation model
- **Stage 2:** Process model (dynamical model)
- **Stage 3:** Parameter model

Goal: Choose between 3 dynamical models used at stage 2

1. Ordinary differential equation (ODE) model
2. Stochastic differential equation (SDE) model
3. Continuous time Markov chain (CTMC) model

Libo, Sun, Hoeting (Envirometrics 2015) use an
Approximate Bayesian Computation method (ABC-SMC)

Method 2: Computational methods

Approximate Bayesian Computation (ABC):

- Method to estimate the model parameters when the likelihood is difficult to compute
- Basic idea: Simulate data from the model and compute a distance function between simulated data and the observed data

ABC-SMC (sequential Monte Carlo):
Improve ABC by simulating data through a sequence of intermediate distributions

Model selection:
Compare models using Bayes factors (Kass & Raftery 1995)

Method 3: POMP models

Ionides and co-authors (2006, 2015 . . .) have developed methodology to enable

- maximum likelihood inference on
- partially observed Markov process (POMP) models

Method 4: Approximate an ODE via smoothing

Ramsay et al. 2007, Cao et al. 2007

General idea:

$$\begin{aligned} y(t_j) &\sim N(X(t_j), \sigma^2), \quad j = 1, \ldots, n \\ \frac{dX}{dt} &= f(X|\theta) \end{aligned}$$

Two levels of nested optimization:

1. Approximate *X* with a non-parametric, smooth function conditional on θ
2. Minimize a sum of squared errors between the observed data and the smoothed curve

What’s next

- **Theory:** Theoretical properties of the methods
- **Methodology:**
  - Need methods that can be used for more complex disease models
  - Faster, accurate computation
  - More precise variance estimation
- **Application:**
  - Which method to use when
  - Need comprehensive and easy to use software

Thank you to

- This work was supported by
  - Sun & Hoeting: National Science Foundation (EF-0914489)
  - Lee: Army Research Office (W911NF-12-1-0494) & National Security Agency (H98230-12-1-0250)
  - We used the CSU ISTeC Cray HPS System, which is supported by NSF Grant CSN-0923386
- Thank you to
  - Tom Hobbs for introducing us to this problem.
  - Mike Miller and the Colorado Division of Parks and Wildlife for sharing the CWD data.