Inference with Implicit Likelihoods

Murali Haran[1]

Roman Jandarov[1], Ottar Bjørnstad[2], and Bryan Grenfell[3]

[1]Department of Biostatistics, University of Washington

[2]Center for Infectious Disease Dynamics, Penn State University

[3]Ecology and Evolutionary Biology, Princeton University

Bayesian Inference in Stochastic Processes

CNR-IMATI, Milano. June 2013

What This Talk is About

- Some richly parameterized statistical models pose inferential challenges
  - Large number of latent variables
  - Expensive likelihood function
  - Likelihood-based inference obtains poor results
- I will use the gravity TSIR model for measles dynamics as a motivating example
- I will describe an approach that addresses these issues in some contexts, borrowing from methods used in computer model emulation and calibration

Basic SIR Model

- A model to explain and predict the spread of an infectious disease.
- SIR model: The population is subdivided into a set of distinct classes: individuals are either susceptible (S), infectious (I) or recovered (R).
- The SIR model describes the dynamics of the sizes of each group.

Gravity TSIR Model

- Models the number of incidences of measles in K different communities (cities)
- Time series SIR (TSIR) model for local dynamics (Bjørnstad et al., 2002; Grenfell et al. 2002) + explicit formulation for spatial transmission between different communities

Notation

- I_{kt} : number of infected individuals in city k at time t
- S_{kt} : number of susceptible individuals in city k at time t
- L_{kt} : number of infected people moved to city k at time t
- d_{kj} : distance between cities k and j
- N_{kt}, B_{kt} : size and birth rate of city k at time t

Gravity TSIR Model

- Number of incidences of a disease at time $t+1$ for city k
  $I_{k(t+1)} \sim \text{Poisson}(\lambda_{k(t+1)})$, where $\lambda_{k(t+1)} = \beta_t S_{kt}(I_{kt} + L_{kt})^{\alpha}$
- $I_{k(t+1)}$ increases with I_{kt}, S_{kt}, and L_{kt} (number of infected immigrants coming to city k at time t)
- $\{\beta_t\}$: seasonal transmission

(Xia, Bjørnstad and Grenfell, 2004)

Gravity TSIR Model

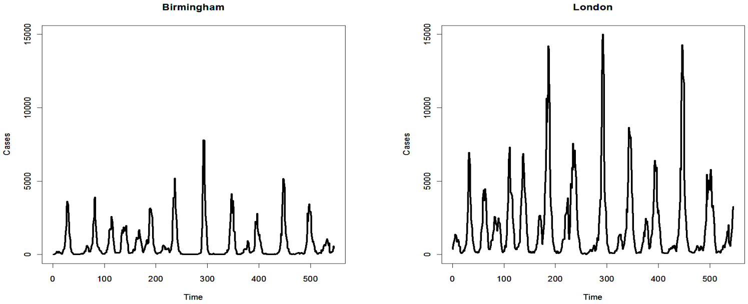
- Number of susceptible individuals at time $t+1$ for city k
  $S_{k(t+1)} = S_{kt} + B_{kt} - I_{k(t+1)}$
- Infected immigrants (latent) at time t for city k
  $L_{kt} \sim \text{Gamma}(m_{kt}, 1)$, where $m_{kt} = \theta N_{kt}^{\tau_1} \sum_{j=1, j\neq k}^{K} \frac{(Ijt)^{\tau_2}}{d_{kj}^{\rho}}$
- L_{kt} increases with size of city k, number of infected people in all other cities, taking into account distances

Inference for Measles Dynamics

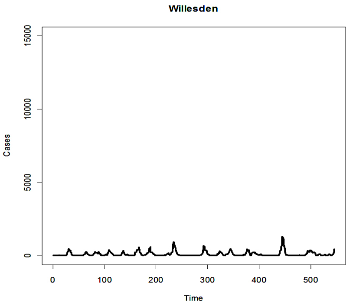
- Parameters of the model:
  - Reliable estimates of local transition parameters α and β are known (Bjørnstad et al. 2001).
  - Gravity parameters θ, τ_1, τ_2 and ρ are unknown.
- Sources of information:
  - The UK Registrar General's data for 952 cities in England and Wales for years 1944-1966 of biweekly incidences of measles.
  - Susceptibles from standard reconstruction algorithms (cf. Fine and Clarkson 1982a, Finkenstadt and Grenfell 2000).

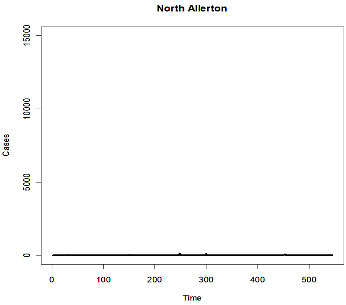
Goal: Infer spatial transmission parameters $\mathbf{\Theta} = (\theta, \tau_1, \tau_2, \rho)$

Measles Data: London and Birmingham



Measles Data: Willesden and North Allerton





Notice: 952 cities of varying sizes and levels of infecteds.

Challenges

MLE or Bayesian inference is simple in principle

- MLE: $\hat{\mathbf{\Theta}} = \arg\max \int \mathcal{L}(\mathbf{\Theta}, \{L_{k,t}\}; \{I_{k,t}\}) dL$
- Bayesian inference,

  $\pi(\mathbf{\Theta}, \{L_{k,t}\} \mid \{I_{k,t}\}) \propto \mathcal{L}(\{I_{k,t}\} \mid \{L_{k,t}\}, \mathbf{\Theta}) \times p(L, \mathbf{\Theta})$

But:

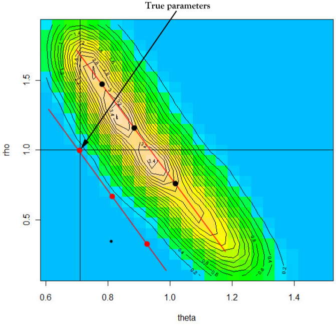
- Dimensions $K \times T$= 546*952 = 519,792
- Therefore:
  - Expensive calculations per iteration of optimizer or MCMC
  - Involves integrating over 519,792 latent variables

Simplifications and Gridded MCMC

- A simple solution:
  1. Fix number of immigrants (latent variables) at expected values. Likelihood function is still expensive $\approx$ 72 hours to find MLE alone.
  2. Discretize parameter space, parallelize pre-calculation of expensive parts of the likelihood.
- Good news: Greatly speeds up computing, permits maximum likelihood and Bayesian inference

Problems . . .

True Θ = ($\theta = 0.71$, $\tau_1 = 0.5$, $\tau_2 = 1$,$\rho = 1$).



Posterior surface for (θ, ρ). (τ_1, τ_2 fixed at true values)

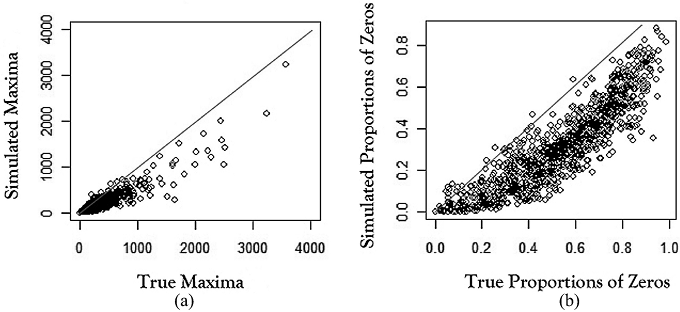
Poor inference for Θ.

Important Biological Characteristics

What do the biologists care about? “Signatures” of the process:

- Maximum number of incidences. $\mathbf{M} = (M_1, \cdots, M_K)$, where M_i is the maximum number of incidences for i-th city.
- Proportions of biweeks without any cases of infection. $\mathbf{P} = (P_1, \cdots, P_K)$, where P_i is the proportion of incidence free bi-weeks for i-th city.

Problems with Fitting Key Characteristics



Fitted model does not capture well important characteristics of the observations.

Back to the Drawing Board

- Likelihood-based approaches apparently do not give enough importance to features that are of scientific interest
- A careful study confirms that these issues are not due to our simplifications or gridded MCMC

New Approach

- Idea: instead of classical likelihood-based approach, build inferential approach that focuses on **fitting scientifically relevant features** of the data.
- Modeling/inference using summary statistics (features).
- Approximate Bayesian computing (ABC) (Pritchard et al., 1999; Beaumont et al. 2002; Marjoram et al., 2002) seems appropriate but is infeasible since simulating draws from this model is also time consuming.

Gaussian Process Emulation and Calibration

- Gaussian processes are useful for emulating (approximating) complex computer models (Sacks et al., 1989; Kennedy and O'Hagan, 2001 etc.) May be useful here.

Gaussian Process Model Basics

- Process at location $\Theta \in D \subset \mathbb{R}^d$ is $Z(\Theta) = \mu_{\boldsymbol{\beta}}(\Theta) + w(\Theta)$.
  Here:"Location" Θ is a parameter setting
- Model dependence among random variables by modeling $\{w(\Theta) : \Theta \in D\}$ as a Gaussian process
- Infinite-dimensional process. If $\Theta_1, \ldots, \Theta_n \in D$, $\mathbf{w} = (w(\Theta_1), \ldots, w(\Theta_n))^T$ is multivariate normal
- Parametric covariance, decays with distance. E.g. $\text{Cov}(Z(\Theta_i), Z(\Theta_j)) = \kappa \exp(-\|\Theta_i - \Theta_j\|/\phi)$, $\kappa > 0, \phi > 0$.
- Let $\mathbf{Z} = (Z(\Theta_1), \ldots, Z(\Theta_n))^T$, so

$$\mathbf{Z}|\kappa, \phi, \boldsymbol{\beta} \sim N(\mu_{\boldsymbol{\beta}}, \Sigma(\kappa, \phi))$$

GP Linear Model Prediction

- Can predict the process at any new parameter setting (Θ) by using simple multivariate normal theory
  - MLE plug-in to get predictive distribution
  - Bayes: same, but averaging over $\kappa, \phi, \boldsymbol{\beta} \mid$ **Z**. This is the *posterior predictive distribution*.
  - This is a stochastic emulator/interpolator
- This provides a distribution for the observations at any given parameter setting. Parametric family!
- For a given data set, therefore, can carry out likelihood-based inference (ML or Bayes).

An Emulation-Based Solution

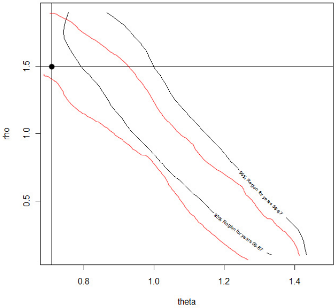
- Let vector of summary statistics from observations be $\mathbf{Z}$. Example: Maximum number of incidences for ith city.
- Simulate realizations of the gravity TSIR model at various parameter settings $\Theta_1, \Theta_2, \ldots, \Theta_p$.
- Let $\mathbf{Y}(\Theta)$ be the vector of summary statistics obtained at parameter setting Θ.
- Consider: $(\Theta_1, \mathbf{Y}(\Theta_1)), \ldots, (\Theta_p, \mathbf{Y}(\Theta_p))$.
- Stochastic emulation: Fit a Gaussian Process (GP) to above simulations.
  - Thus for any new parameter setting Θ^*, we have a predictive distribution for the process $\mathbf{Y}(\Theta^*)$.

New Inferential Approach

1. Gaussian process emulation provides a probability model for observations **Z**. Emulator likelihood, $\mathcal{L}^*(\{I_{k,t}\} \mid \mathbf{\Theta})$
2. Bayesian inference to infer $\mathbf{\Theta}$
   - Original approach:
     $\pi(\mathbf{\Theta}, \{L_{k,t}\} \mid \{I_{k,t}\}) \propto \mathcal{L}(\{I_{k,t}\} \mid \{L_{k,t}\}, \mathbf{\Theta}) \times p(L, \mathbf{\Theta})$
   - New approach:
     $\pi^*(\mathbf{\Theta} \mid \{I_{k,t}\}) \propto \mathcal{L}^*(\{I_{k,t}\} \mid \mathbf{\Theta}) \times p(\mathbf{\Theta})$

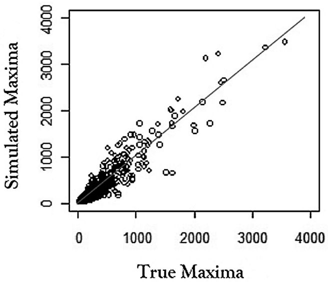
Skipping lots of important details: computational issues, data-model discrepancy, design points ...

Improved Inference for Θ



95% C.I.'s for (θ, ρ): Solid black line: the likelihood-based method; Solid red line: the Gaussian process emulator.

Fitting Biological Characteristics using GP-approach



(a)

(b)

Fitted model better captures important characteristics of the data.

Summary

- Our Gaussian process-based inferential approach focuses directly on scientifically relevant characteristics.
- Improves inference, model fit, addresses computational challenges, circumvents latent variable issues.
- We are able to apply our approach to the England-Wales data set and obtain scientific conclusions.
- Caveats:
  - Will not readily apply when Θ is high-dimensional
  - Open questions: choice of summary statistics if scientists have multiple criteria; design of simulations etc.

Collaborators

- Roman Jandarov, Postdoctoral fellow, University of Washington
- Ottar Bjørnstad, Center for Infectious Disease Dynamics, Penn State University
- Bryan Grenfell, Ecology and Evolutionary Biology, Princeton University

Support from Bill & Melinda Gates Foundation

References

- Grenfell, B.T., Bjørnstad, O. N. and Kappey, J. (2001), "Traveling waves and spatial hierarchies in measles epidemics." *Nature*.
- Bhat, K.S., Haran, M., Olson, R., and Keller, K. (2012), "Inferring likelihoods and climate system characteristics from climate models and multiple tracers," *Environmetrics*.
- Bhat, K.S., Haran, M. and Goes, M. (2010) "Computer model calibration with multivariate spatial output."
- Jandarov, R., Haran, M., Bjornstad, O.N. and Grenfell, B. (2013) "Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease."

Gaussian Process Prediction/Interpolation

- Let the predictions at the new locations $\mathbf{s}_1^*, \ldots, \mathbf{s}_m^* \in D$ be $\mathbf{Z}^* = (Z(\mathbf{s}_1^*), \ldots, Z(\mathbf{s}_m^*))^T$.
- Under the GP assumption ($\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma$ depend on $\boldsymbol{\beta}, \Theta$):

$$\begin{bmatrix} \mathbf{Z} \\ \mathbf{Z}^* \end{bmatrix} \mid \Theta, \boldsymbol{\beta} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right), \tag{1}$$

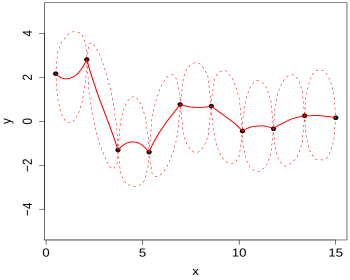
ML: use above with ML estimates plugged-in.

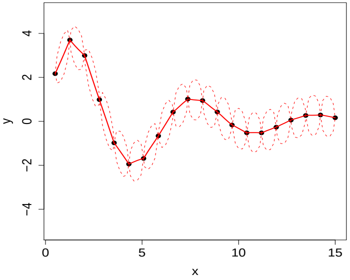
Bayes: use above, while averaging over $\Theta, \boldsymbol{\beta} \mid \mathbf{Z}$. This is the *posterior predictive distribution*.

GP Model Emulation

Interpolations using simple GP random effects model:

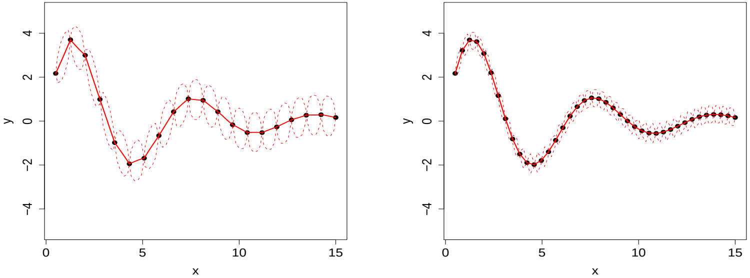
$y(x) = \mu + w(x)$, $\{w(x), x \in (0, 20)\}$ is a zero-mean GP.





Increase data from 10 to 20 points

GP Model Emulation



Increase data from 20 to 40 points

Modeling with Gaussian Processes

- Gaussian processes (GPs) are useful models for dependent processes, e.g. time series, spatial data.
- GPs are also very useful for modeling complicated functions.

  Key idea: dependence (spatial random effects) adjusts for non-linear relationships between input and output.

Summary of Inferential Problem

Let parameter of interest be θ (here θ=K_v).

Statistical problem:

- Model output is a bivariate spatial process at each θ: $\mathbf{Y} = ((\mathbf{Y}_1(\psi_1), \mathbf{Y}_2(\psi_1)), (\mathbf{Y}_1(\psi_2), \mathbf{Y}_2(\psi_2)), \ldots, (\mathbf{Y}_1(\psi_K), \mathbf{Y}_2(\psi_K)),$ where $\{\psi_1, \psi_2, \ldots, \psi_K\}$ is a set of plausible θ values.
- Observations: $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$.
- What can we learn about θ given $\mathbf{Z}, \mathbf{Y}$?

Bayesian Approach

A Bayesian framework is useful for computer model calibration:

- There is usually real prior information about θ.
- The likelihood surface for θ may often be highly multimodal and there may be identifiability issues; useful to have easy access to the full posterior distribution.
- If θ is multivariate, important to look at bivariate and marginal distributions: easier w/ sample-based approach.
- Amenable to hierarchical specification: we will exploit this for multivariate spatial process model.

Kennedy and O'Hagan (2001); Bayarri, Berger et al. (2007, 2008).

Latter provides wavelets-based approach for functional output.

Two-stage Approach to Inference

1. Find probability model for $\mathbf{Z}$ (data) using $\mathbf{Y}$ (simulations.)
   - Model relationship between $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ and $\boldsymbol{\theta}$ via flexible emulator for model output $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$.
   - Add model discrepancy and measurement error:

   $$\mathbf{Z} = \boldsymbol{\eta}(\mathbf{Y}, \boldsymbol{\theta}) + \boldsymbol{\delta}(\mathbf{Y}) + \boldsymbol{\epsilon}$$

   where $\delta(\mathbf{Y}) = (\delta_1,\ \delta_2)^T$ is the model discrepancy, also modeled as a GP. $\epsilon = (\epsilon_1,\ \epsilon_2)^T$ is the observation error.
2. Posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{Z})$ derived from prior on $\boldsymbol{\theta}$ and likelihood based on above model.