Latent Variable Compartmental Models for Infectious Diseases

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

Joint Statistical Meetings, Montreal. August 2013

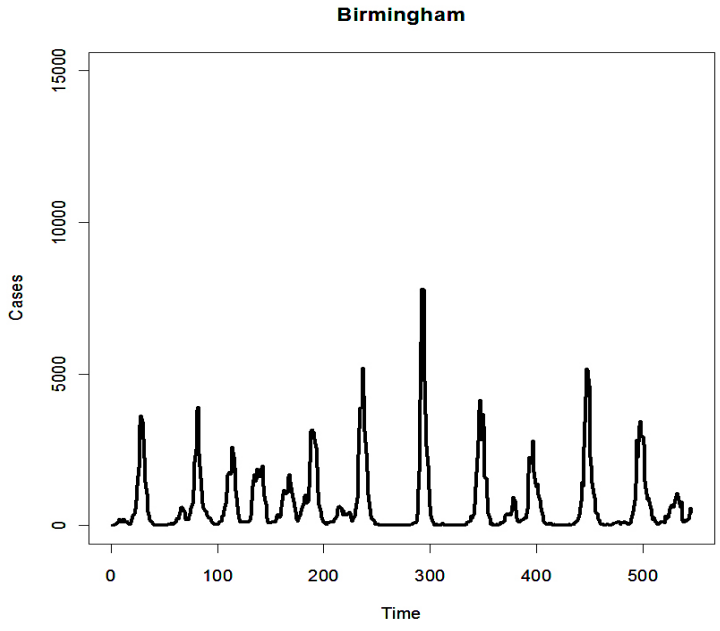
What This Talk is About

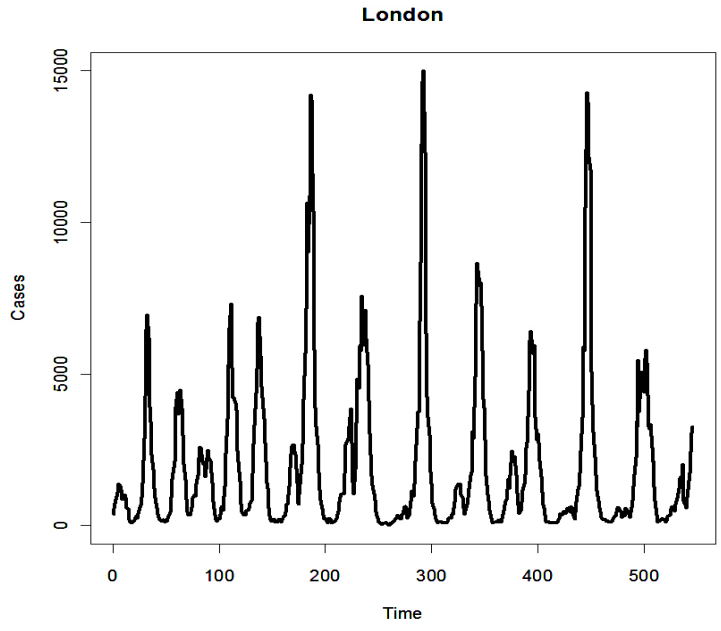
- Two flavors of spatial models for infectious diseases
  - Spatial generalized linear mixed models (SGLMMs)
  - Compartmental models for spatial transmission
- Both cases:
  - Random effects/latent variables are often key; large number of them
  - Expensive/complicated likelihood functions
- I will focus on compartmental models, describing
  - simulation-based computational approach
  - capability for answering a rich set of scientific questions

Measles Data

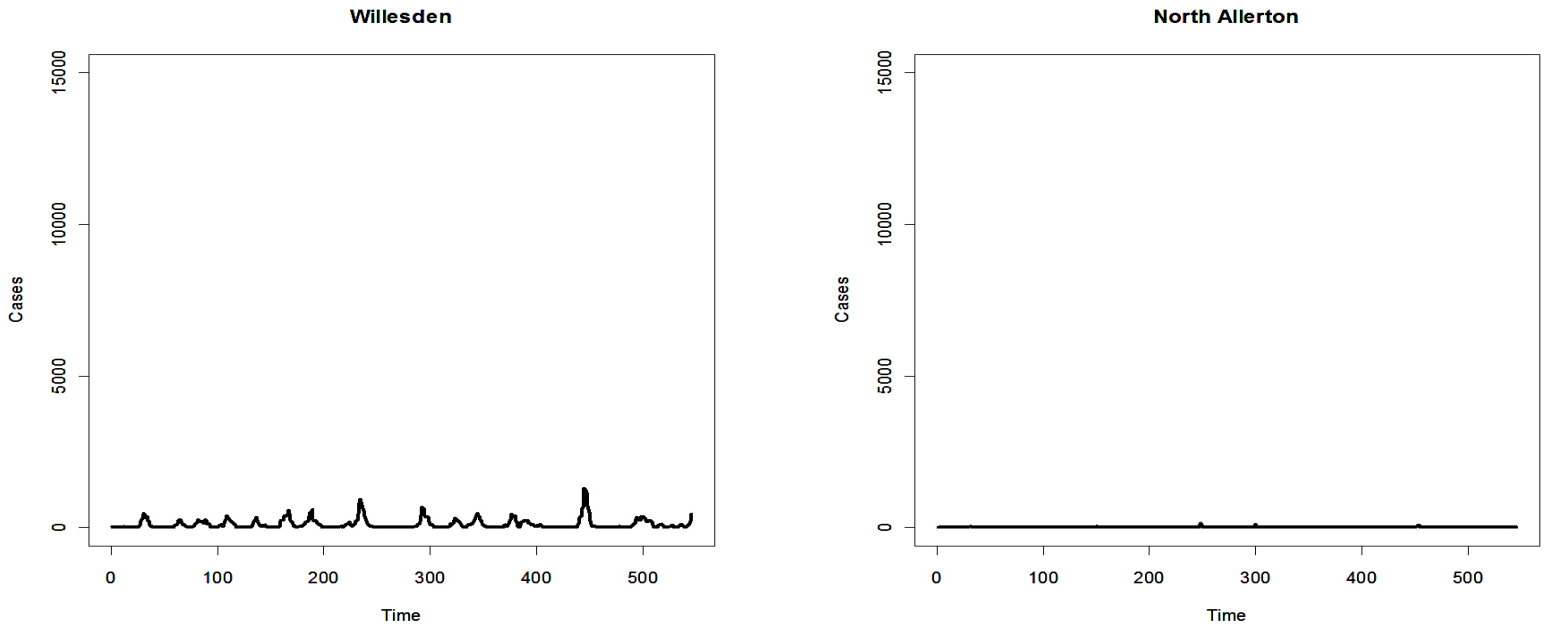
The UK Registrar General's data for 952 cities in England and Wales for years 1944-1966 of biweekly incidences of measles.

Measles Data: London and Birmingham





Measles Data: Willesden and North Allerton



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

Observations about SGLMMs

Spatial generalized linear mixed models, SGLMMs, (cf. Besag et al., 1991; Diggle et al., 1998) are popular and flexible models for such data:

- Prioritize: (1) space-time smoothing while accounting for variability appropriately; (2) adjusting for space-time dependence when performing regression
- Direct interpretation of spatial random effects may be problematic (cf. Reich et al. (2006); Hughes and Haran, 2012). They are only a convenient modeling device.
- Difficult to translate results into a “spatial transmission story”. Learning about spatial transmission is useful for public health policy, e.g. vaccination programs, controling outbreaks

Compartmental Models

- Instead of just capturing dependence, what if the goal were to directly capture how the disease spreads?
- Formulate models differently. But latent variables will be key here too

Basic 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.

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

Blue=latent variables

Gravity TSIR Model

- Incidences of a disease at time $t+1$ for city k,
  $I_{k(t+1)} \mid L_{kt} \sim \text{Poisson}(\lambda_{k(t+1)})$,
  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

- Reliable estimates of local transition parameters α and β are known (Bjørnstad et al. 2001).
- Spatial transmission parameters $\boldsymbol{\Theta} = (\theta, \tau_1, \tau_2, \rho)$ are unknown.
- **Goal**: Infer $\boldsymbol{\Theta}$ given observations

Challenges

MLE or Bayesian inference is simple in principle

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

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

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

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.

Simulation-based Approach

- Idea: instead of classical likelihood-based approach, build inferential approach that focuses on summaries, **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.

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

Two steps:

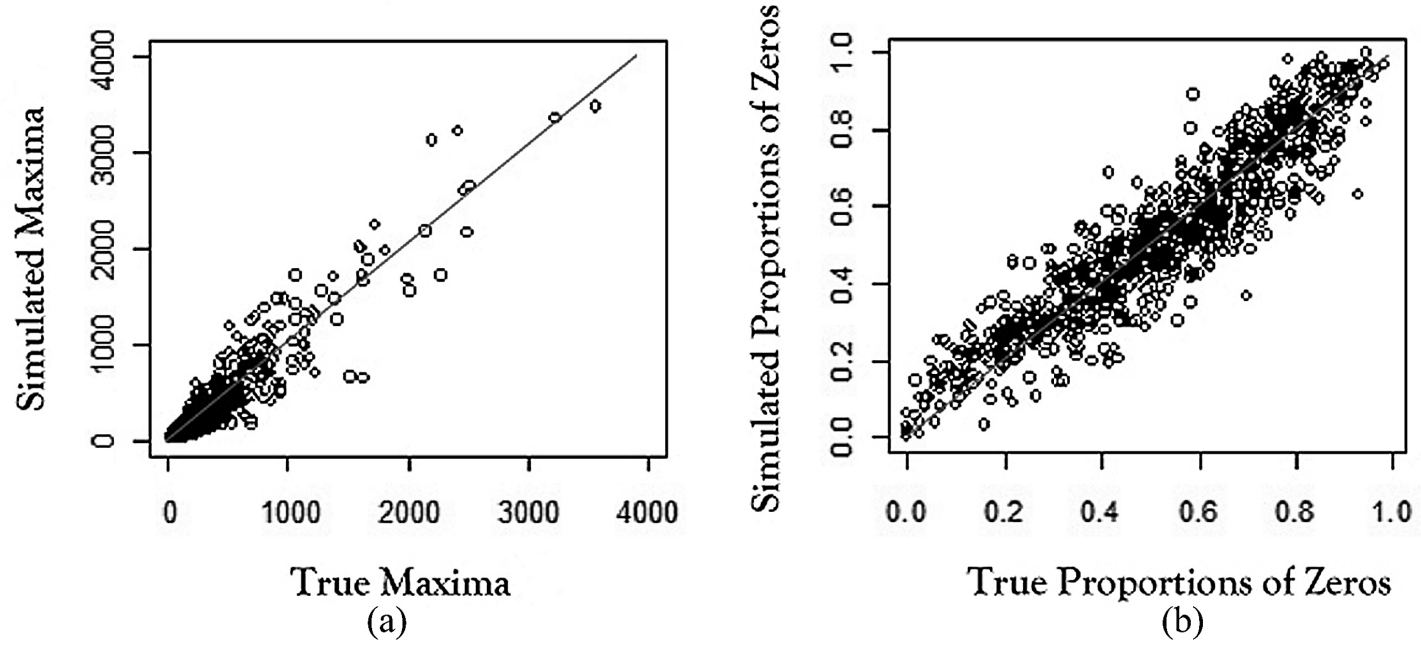
1. Gaussian process fit to simulations provides a probability model for observations **Z**.
   - Emulator (approximate) likelihood, $\mathcal{L}^*(\{I_{k,t}\} \mid \mathbf{\Theta})$
2. Bayesian inference for $\mathbf{\Theta}$
   - Original approach:

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

     $\pi^*(\mathbf{\Theta} \mid \{I_{k,t}\}) \propto \mathcal{L}^*(\{I_{k,t}\} \mid \mathbf{\Theta}) \times p(\mathbf{\Theta})$
   - MCMC to learn about $\pi^*(\mathbf{\Theta} \mid \{I_{k,t}\})$

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

Fitting Biological Characteristics using GP-approach



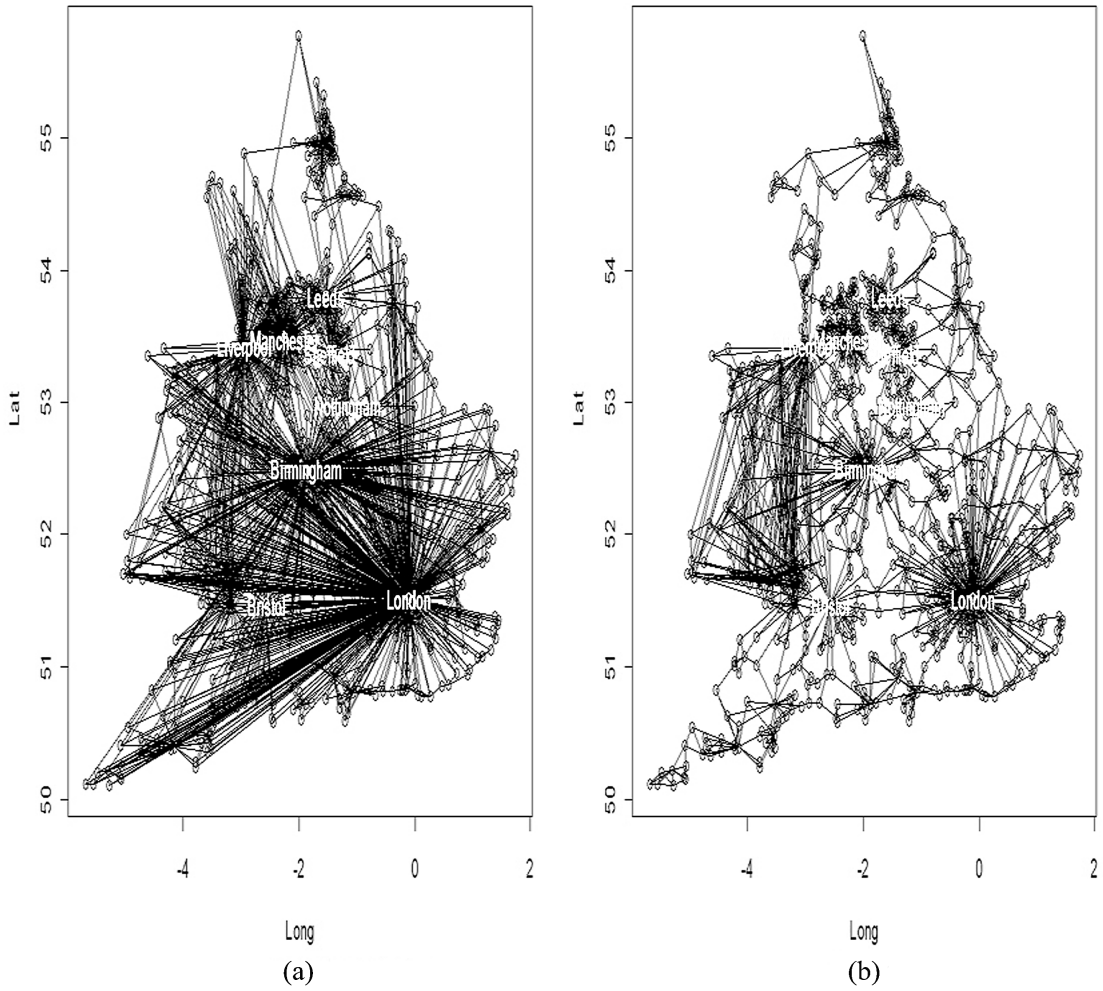
- **Fitted model captures well important characteristics of the data.**
- Also fits the original data (time series plots) well

Scientific Conclusions

The model and sample-based inferential approach allow us to answer a variety of scientifically interesting questions

- Matrix $M = \{m_{kj}\}$, where $m_{kj} = \theta' N_{kt}^{\tau_1} \sum_{t=1}^{T} \frac{(Ijt)^{\tau_2}}{d_{kj}^{\rho}}$ may be interpreted as the amount of movement
  - k-th row sum of M is # of infected individuals leaving city k
  - k-th column sum is # of infected people coming to city k
- Using samples for $(\theta, \rho, \tau_1, \tau_2)$, easy to obtain a sample for the spatial flux of infection for each city

Learning about Movement Networks



(a) outgoing infections, (b) incoming infections

Inference about Movement Networks

- Outgoing: big cities are important in the dynamics of measles for smaller communities where the infection may become locally extinct. Distances between big and small cities do not seem to matter much.
- Incoming infections are mostly dependent on distances between cities since edges connecting different cities in this graph are shorter.
- Big cities are main factors in starting outbreaks elsewhere, excluding the possibility of re-introduction of the disease from neighboring cities with small population sizes.

Other interesting questions: pre-vaccination versus post-vaccination movement, holiday versus non-holiday etc.

Summary

- Our approach provides insights into spatial transmission and allows for a rich set of scientific conclusions
- Our Gaussian process-based inferential approach
  - Focuses directly on scientifically relevant characteristics.
  - Circumvents challenges posed by latent variables.
- Caveats:
  - Will not readily apply when Θ is high-dimensional
  - Open questions: choice of summary statistics if scientists have multiple criteria, design of simulations

Collaborators

Jandarov, R., Haran, M., Bjornstad, O.N. and Grenfell, B. (2013) "Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease," *Journal of the Royal Statistical Society (C), in press.*

- 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

Latent Variables in SGLMMs

Example: Model # incidences of disease in region $\mathbf{s}$, $Z(\mathbf{s})$

1. $Z(\mathbf{s}_i) \mid \beta, \Theta \sim$Poisson$(\exp(X(\mathbf{s}_i)\beta + w(\mathbf{s}_i))$, conditionally independent for $i = 1, \ldots, n$
2. Latent variables $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T \sim$ Gaussian Markov random field (GMRF),

$$\mathbf{w}|\Theta, \beta \sim N(0, Q^{-1}(\Theta))$$

   Precision matrix, $Q(\Theta)$ captures dependence according to the adjacency structure (graph) specified for $\mathbf{s}_1, \ldots, \mathbf{s}_n$

3. Priors for Θ, β.

(cf. Besag, York, Mollié (1991); Diggle et al. (1998))

SGLMMs for Space-Time Incidences Data

- Natural extension to space-time modeling
- # incidences in region $\mathbf{s}$ at time t, $Z(\mathbf{s}, t)$
- Conditionally independent Poisson again
- Latent variables $w(\mathbf{s}, t)$ have a space-time dependence model, autoregressive in time as well as space
- # of $w(\mathbf{s}, t)$s for measles example: 546×952=519,792

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 $\kappa, \phi, \boldsymbol{\beta} \mid \mathbf{Z}$. 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).

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 (μ_1, μ_2, Σ depend on β, Θ):

$$\begin{bmatrix} \mathbf{Z} \\ \mathbf{Z}^* \end{bmatrix} \mid \Theta, \beta \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right), \tag{1}$$

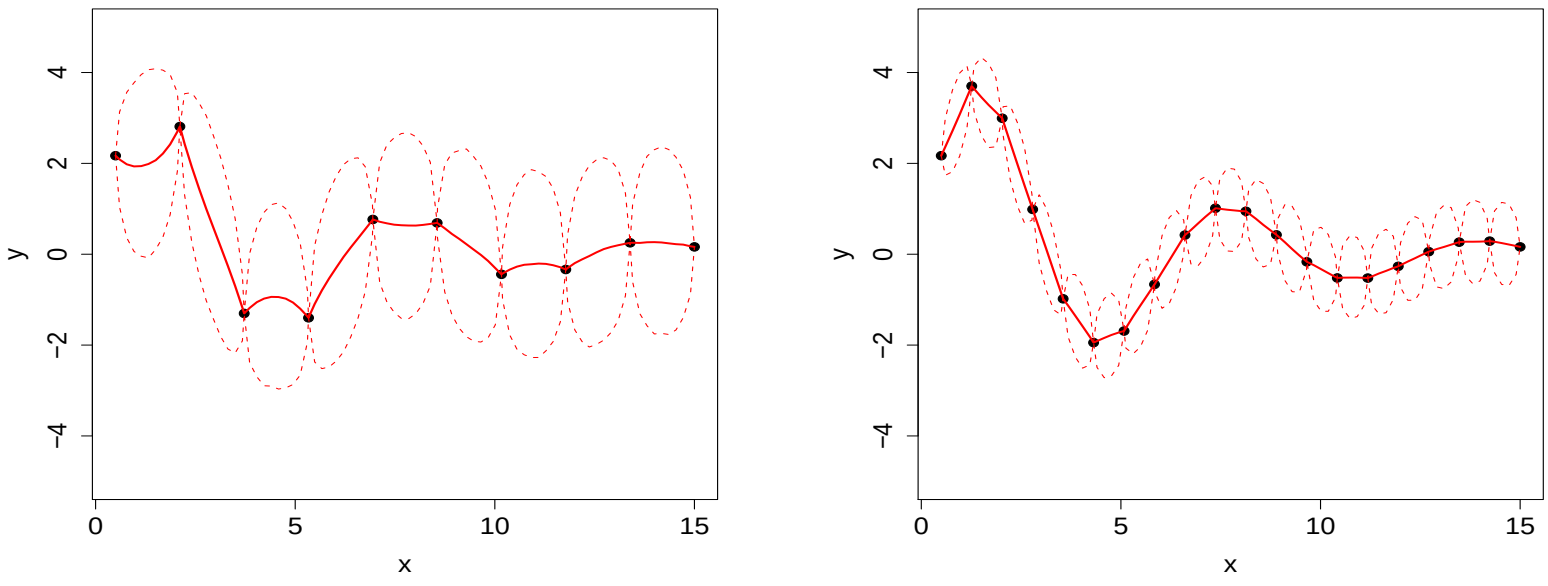
ML: use above with ML estimates plugged-in.

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

GP Model Emulation

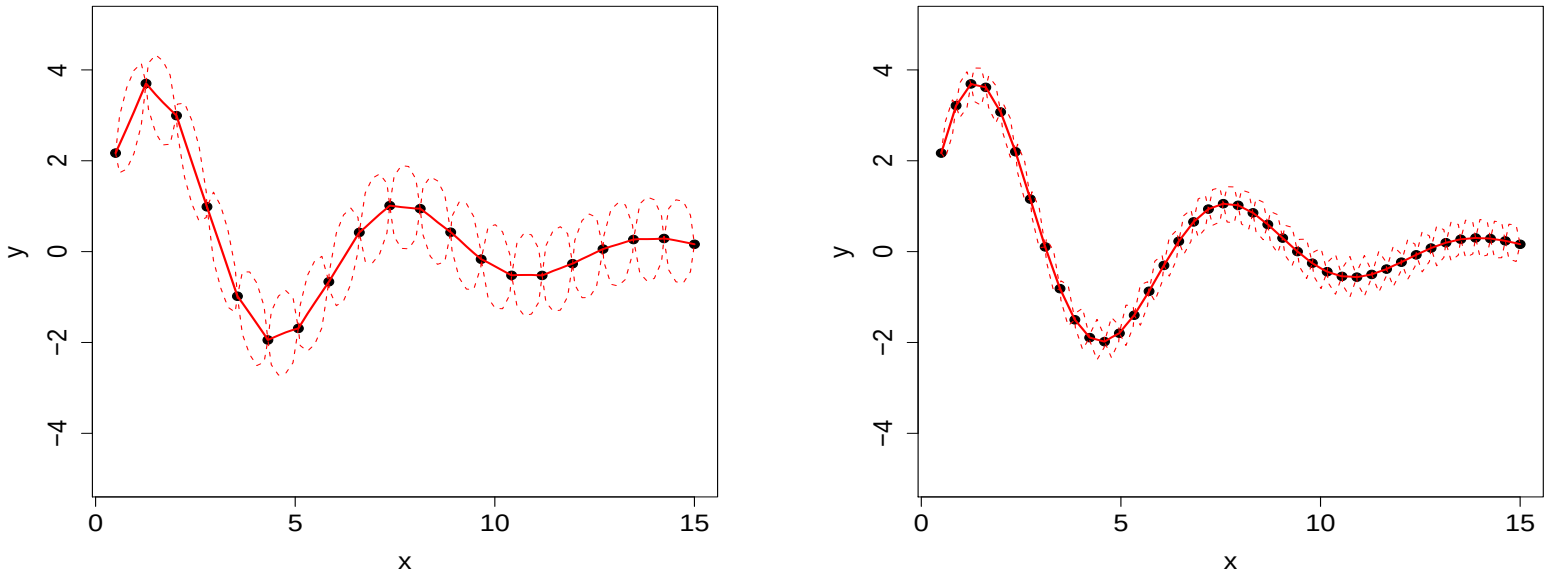
Interpolations using simple GP random effects model:

$y(x) = \mu + w(x)$, $\{w(x), x \in (0, 20)\}$ 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 $\boldsymbol{\theta}$ (here $\boldsymbol{\theta}$=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 $\boldsymbol{\theta}$ given $\mathbf{Z}, \mathbf{Y}$?

Bayesian Approach

A Bayesian framework is useful for computer model calibration:

- There is usually real prior information about $\boldsymbol{\theta}$.
- The likelihood surface for $\boldsymbol{\theta}$ may often be highly multimodal and there may be identifiability issues; useful to have easy access to the full posterior distribution.
- If $\boldsymbol{\theta}$ 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:

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

   where $\boldsymbol{\delta}(\mathbf{Y}) = (\boldsymbol{\delta}_1,\ \boldsymbol{\delta}_2)^T$ is the model discrepancy, also modeled as a GP. $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1,\ \boldsymbol{\epsilon}_2)^T$ 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.