Gaussian Processes for Inference with Implicit Likelihoods

Murali Haran
Department of Statistics
Pennsylvania State University.

School of Statistics 40th Reunion
University of Minnesota, May 2011.

This Talk

- I will discuss two of my current interdisciplinary research projects.
- The talk will be interspersed with quotes from U. of Minnesota professors. Some wise, some offbeat (but still enlightening.)
- Disclaimer: I will paraphrase from memory. I may misquote, misattribute quotes.

The first time you give a talk, you learn a lot and the audience learns nothing; the second time, the audience learns a lot, you learn nothing; after that, nobody learns anything. – Brad Carlin

Complex Scientific Models

You could do worse than to talk to scientists – Charlie Geyer

- Scientists working in the physical and natural sciences are often interested in learning about the mechanisms or "laws" and processes underlying physical phenomena.
- These models may be useful for predictions/projections.
- Critical to work with the model provided by the scientists.
- These scientific models may be
  - Numerical solutions of mathematical (deterministic) models or stochastic models that reflect scientific processes.
  - Translated into computer code to study simulations of the physical processes for different parameters/conditions.

Some Challenges Posed by Complex Models

When you give a talk, to a first approximation nobody knows what you are talking about. And nobody cares (as much as you.) – Bill Sudderth.

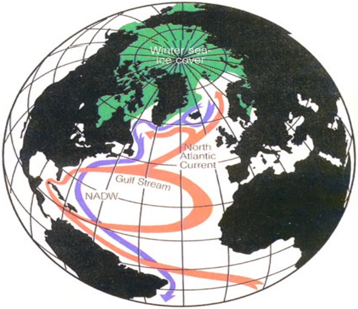
- More scientifically plausible models are typically more complex. Challenges:
  - Computationally expensive simulations.
  - May not be possible to write closed-form expressions relating input/parameters to output.
  - (When stochastic) The likelihood function may be very expensive to evaluate: hard to optimize or use Monte Carlo methods.
  - There are non-ignorable discrepancies between the model and reality.
- Likelihood is often *implicit* or has to be treated as such.

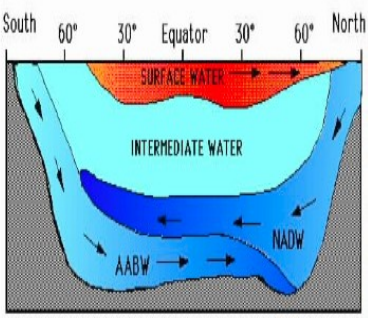
Two Examples

I. Climate: An Earth System Model of Intermediate Complexity (EMIC) for projecting the behavior of global ocean circulation systems.
   - Deterministic
   - Model runs are expensive
   - High-dimensional multivariate spatial process

II. Disease Dynamics: A Gravity Time Series Susceptible-Infected-Recovered (TSIR) model for the spread of infectious disease (measles).
   - Stochastic
   - Likelihood is expensive to evaluate
   - Space-time process with large number of “no incidence” observations (0s)

The Meridional Overturning Circulation (MOC) and Climate Change





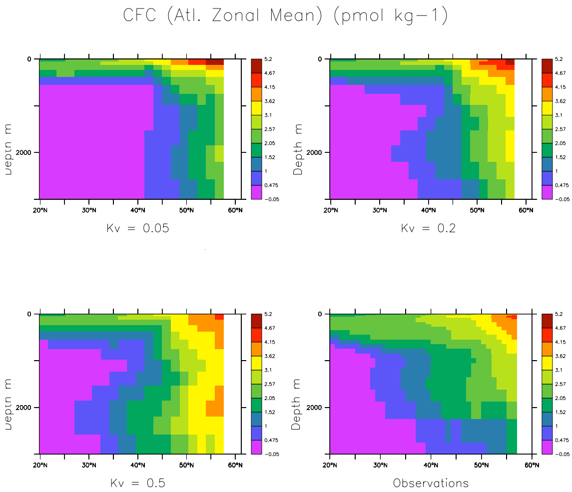
(plots: Rahmstorf (Nature, 1997) and Behl and Hovan)

Climate Models: Learning About $\mathbf{K_v}$

"Collapse" of MOC may result in dramatic climate change. $\mathbf{K_v}$ is a key climate model parameter that influences the MOC.

- $\mathbf{K_v}$ is a model parameter which quantifies the intensity of vertical mixing in the ocean, cannot be measured directly.
- Two sources of indirect information on $\mathbf{K_v}$:
  - Observations of two ocean "tracers", both provide information about $\mathbf{K_v}$: Carbon-14 (^{14}C) and Trichlorofluoromethane (CFC11).
  - Climate model output of these two tracers at different values of $\mathbf{K_v}$ from the University of Victoria(UVic) Earth System Climate Model (Weaver et. al. 2001): $\mathbf{Y}_1(\mathbf{K_v}), \mathbf{Y}_2(\mathbf{K_v})$

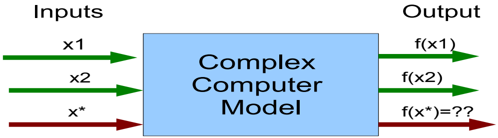
CFC-11 Example



- Bottom right: observations
- Remaining plots: climate model output at 3 settings of K_v.

Computer Models and Emulation

Statistical interpolation



Green inputs/output = training data.

Red = the input where predictions are desired.

Input and output are typically multivariate.

Computer Model Emulation

- Fit an emulator ("meta model") to a training set of runs from the complex computer model.
- Advantages:
  - Fast approximate simulator.
  - Uncertainties associated with interpolation (predictions), for example greater uncertainty where there is less training data information.
  - "Without any quantification of uncertainty, it is easy to dismiss computer models." (A.O'Hagan)
  - Now have a probability model.

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.

Gaussian Process Model Basics

- Process at location $\mathbf{s} \in D$, $D \subset \mathbb{R}^d$ is $Z(\mathbf{s}) = \mu_{\boldsymbol{\beta}}(\mathbf{s}) + w(\mathbf{s})$. Location $\mathbf{s}$ may be physical or from “input space”.
- Model dependence among spatial random variables by modeling $\{w(\mathbf{s}) : \mathbf{s} \in D\}$ as a Gaussian process.
- For any n locations, $\mathbf{s}_1, \ldots, \mathbf{s}_n$, $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T$ is multivariate normal.
- Parametric covariance function with parameters Θ. E.g. exponential covariance:
  $\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) = \kappa \exp(-\|\mathbf{s}_i - \mathbf{s}_j\|/\phi)$, $\kappa > 0, \phi > 0$.
  Here, $\Theta = (\kappa, \phi)$.
- Let $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$, so

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

GP Linear Model Inference

- Inference and prediction can be done via ML or Bayes.
- ML: maximize likelihood with respect to $\Theta, \boldsymbol{\beta}$.
- Bayes: prior on $\Theta, \boldsymbol{\beta}$; learn about $\pi(\Theta, \boldsymbol{\beta} \mid \mathbf{Z})$ using Markov chain Monte Carlo (MCMC).

  What's that four-letter word again? – Seymour Geisser

GP Linear Model Prediction

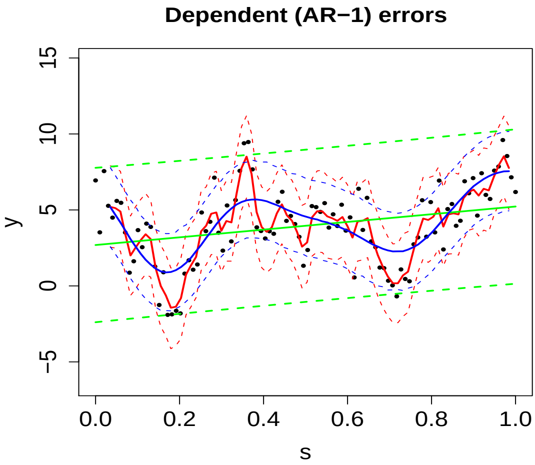
- Process at new locations $\mathbf{s}_1^*, \ldots, \mathbf{s}_m^* \in D$ be $\mathbf{Z}^* = (Z(\mathbf{s}_1^*), \ldots, Z(\mathbf{s}_m^*))^T$.
- Under the GP assumption ($\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma$s depend on $\boldsymbol{\beta}, \Theta$):

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

  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*.

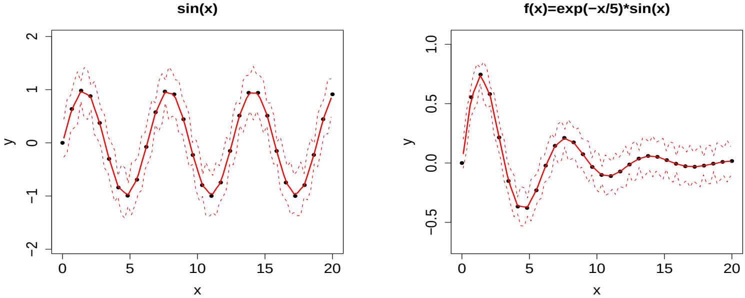
GP Model for Dependence: Toy 1-D Example



It's always a good idea to have a toy example. – Galin Jones

Toy examples don't teach you anything. – Charlie Geyer

GP for Function Approximation: Toy 1-D Example



The red curves are interpolations using *the same, simple GP model* with constant mean μ:

$$y(x) = \mu + w(x),$$

where $\{w(x), x \in (0, 20)\}$ is a zero-mean GP.

Summary of Inferential Problem

Let parameter of interest be θ (here θ=$\mathbf{K_v}$).

Statistical problem:

- Model output is a bivariate spatial process at each θ: $\mathbf{Y} = ((\mathbf{Y}_1(\psi_1), \mathbf{Y}_2(\psi_1)), (\mathbf{Y}_1(\psi_2), \mathbf{Y}_2(\psi_2)), \dots, (\mathbf{Y}_1(\psi_K), \mathbf{Y}_2(\psi_K))$, where $\{\psi_1, \psi_2, \dots, \psi_K\}$ 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:

- There is usually real prior information about θ.
- Access to the full posterior distribution is useful with potential multimodality and identifiability issues.
- If θ is multivariate, important to look at bivariate and marginal distributions (easier w/ sample-based approach).

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

You shouldn't be reading papers, you should be writing them. (Yeah!) – Jim Dickey

Even if you don't know the literature and reinvent something, you will probably end up doing something cool(er), so it doesn't matter. – Charlie Geyer

Two-stage Approach to Inference

1. Find probability model for **Z** (data) using **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:

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

     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. Prior for $\boldsymbol{\theta}$ + likelihood (based on above model) provides posterior distribution for $\boldsymbol{\theta}$.

Need (i) flexible model for relationship between $\mathbf{Y}_1$ and $\mathbf{Y}_2$ (two spatial fields), (ii) computational tractability.

Step 1: Details

People don't care too much about the details. – Glen Meeden

- Model $(\mathbf{Y}_1, \mathbf{Y}_2)$ as a hierarchical model: $\mathbf{Y}_1|\mathbf{Y}_2$ and $\mathbf{Y}_2$ as Gaussian processes (cf. Royle and Berliner, 1999.)

$$\mathbf{Y}_1 \mid \mathbf{Y}_2, \boldsymbol{\beta}_1, \boldsymbol{\xi}_1, \gamma \sim N(\mu_{\boldsymbol{\beta}_1}(\boldsymbol{\theta}) + \mathbf{B}(\gamma)\mathbf{Y}_2, \Sigma_{1.2}(\boldsymbol{\xi}_1))$$

$$\mathbf{Y}_2 \mid \boldsymbol{\beta}_2, \boldsymbol{\xi}_2 \sim N(\mu_{\boldsymbol{\beta}_2}(\boldsymbol{\theta}), \Sigma_2(\boldsymbol{\xi}_2))$$

- $\mathbf{B}(\gamma)$ is a matrix relating $\mathbf{Y}_1$ and $\mathbf{Y}_2$, with parameters γ.
- The covariances of the Gaussian processes depend on both $\mathbf{s}$ (spatial distance) and $\boldsymbol{\theta}$ (distance in parameter space).
- $\boldsymbol{\beta}$s, $\boldsymbol{\xi}$s are regression, covariance parameters.

Step 2: Details

- Emulation: Fit GP via maximum likelihood, then obtain predictive distribution at locations of observations.
- Add model discrepancy and measurement error:

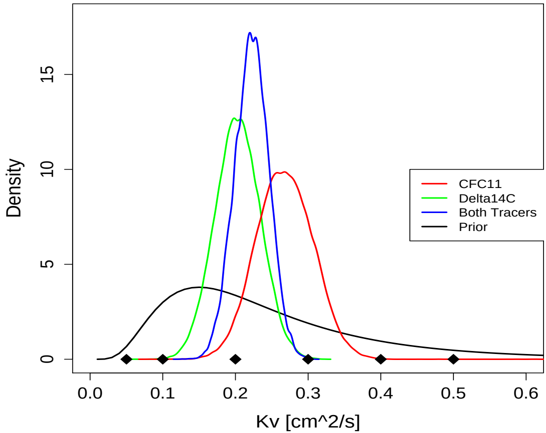
$$\mathbf{Z} = \boldsymbol{\eta}(\mathbf{Y}, \boldsymbol{\theta}) + \boldsymbol{\delta}(\mathbf{Y}) + \boldsymbol{\epsilon}$$

  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.
- Model discrepancy term can make crucial adjustment to $\boldsymbol{\theta}$ estimates (Bayarri, Berger et al. 2007; Bhat et al., 2010).
- Use Markov chain Monte Carlo (MCMC) to estimate $\pi(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{Y})$, integrating out remaining parameters.
- Separating stages: 'modularization' (e.g. Liu, Bayarri, Berger, 2009). Computational advantages + reduce identifiability issues.

Computational Issues

- Matrix computations are $\mathcal{O}(N^3)$, where N is the number of observations. Here: $N \approx$ tens of thousands.
- Markov chain mixes slowly so need long MCMC runs.
- We use a reduced rank approach based on kernel mixing (Higdon, 1998): continuous process created by convolving a discrete white noise process with a kernel function.
- Special structure + Sherman-Woodbury-Morrison identity + Sylvester's Theorem used to reduce matrix computations.

Results for K_v Inference



posteriors: only CFC-11, only $\Delta^{14}C$, both CFC-11 & $\Delta^{14}C$.

Result: $\mathbf{K_v}$ pdf suggests weakening of MOC in the future.

Summary of Climate Model Inference

Two-stage approach:

1. Obtain a probability model connecting CFC-11, $\Delta^{14}C$ tracer observations to $\mathbf{K_v}$ by fitting a flexible Gaussian process model to climate model runs. Hierarchical model for multiple spatial processes + patterned covariances $\Rightarrow$ flexible and computationally tractable.
2. Using this probability model, infer a posterior density for $\mathbf{K_v}$ from the observations.

We can use inferred $\mathbf{K_v}$ in the climate model to project the MOC. We find that the MOC weakens over the next 50 years.

II. Infectious Disease Models

You should not try to communicate more than one idea in a single talk. – Glen Meeden

- Gravity-TSIR model: Space-time model for the spread of measles. Unknown parameters of this model control the dynamics of the spread of this disease e.g. how the disease spreads as a function of distance between locations.
- Thousands of latent variables e.g. number of immigrants moving from one location to another.
- Rich space-time data set from England and Wales. Time points$\times$ locations = 546$\times$952= 519,792.
  Potential for learning about parameters, but also poses computational challenges.

Inference for Gravity TSIR Model Parameters

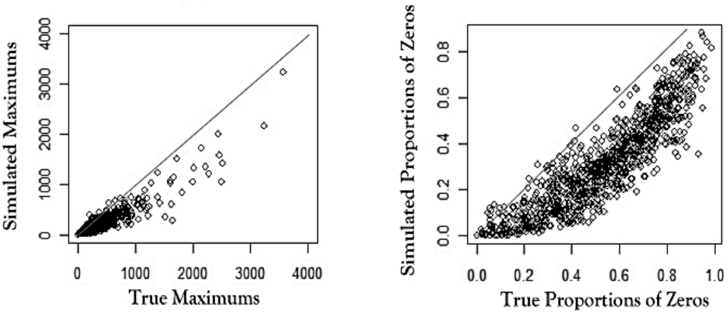
- An approximate grid-based Markov chain Monte Carlo approach provides a way out of the computational challenges.

  Sounds good. – Luke Tierney

- However, traditional likelihood-based/Bayesian inference does not result in a fitted model that reproduces scientifically relevant features of the data.

Traditional Likelihood-Based Approach

Simulations from fitted model (Bayes/ML) do not match up well with the data for important characteristics of the process.



Just because it sounds like a good idea, that doesn't mean it's a good idea. – Charlie Geyer.

Most of the time when people think they are screwed, they aren't really screwed. Occasionally they *are* screwed. – Glen M.

Inference for Gravity TSIR Model Parameters

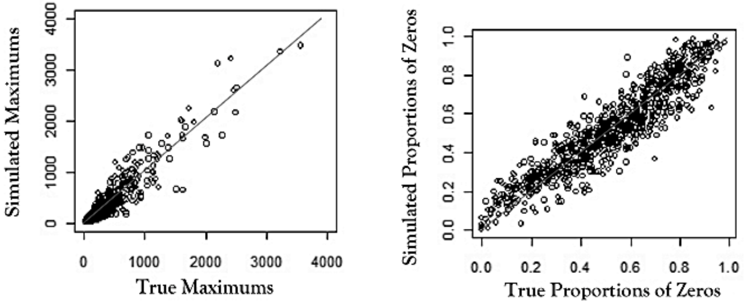
You are thinking about it *all wrong*. – Charlie Geyer.

- Likelihood-based approaches do not take into account features that are of scientific interest.
- Instead, fit GP to *summary statistics* of model runs where summaries are based on scientifically relevant features.
- Inference based on using this GP with the data results in improved inference.

(Skipping lots of details, computational issues etc. . . .)

GP-based Inference Using Key Summaries

Simulations from fitted model are a much better match.



Summary

- Gaussian processes are a powerful tool for problems where the likelihood is implicit and simulating from the model is expensive.

  Try it, you'll like it. – Frank Martin.
- GPs are useful for deterministic and stochastic models.
- GP-based approach can be used to take into account the scientifically important features of the data; may be preferable to traditional likelihood-based approaches.
- Limitation: computationally intractable when the number of parameters of interest (dimensionality of θ) is large.

Collaborators

- K. Sham Bhat, Los Alamos National Laboratories.
- Roman Tonkonojenkov, Dept of Geosciences, Penn State University.
- Klaus Keller, Dept of Geosciences, Penn State University.
- Roman Jandarov, Dept of Statistics, Penn State University.
- Ottar Bjørnstad, Center for Infectious Disease Dynamics, Penn State University.

Woof. – Charlie Geyer

Go have a beer. – Frank Martin

References

- Kennedy, M.C. and O'Hagan, A.(2001), Bayesian calibration of computer models, *J. of Royal Statistical Society (B)*.
- Grenfell, B.T., Bjørnstad, O. N. and Kappey, J. (2001), "Traveling waves and spatial hierarchies in measles epidemics." *Nature*.
- Bhat, K.S., Haran, M., Tonkonojenkov, R., and Keller, K. (2011), "Inferring likelihoods and climate system characteristics from climate models and multiple tracers."
- 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. (2011) "Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease."

II. Infectious Disease Models

- Infectious disease models are useful for investigating key questions in biology. They are of practical use in the management and control of infectious diseases, including immunization and epidemic control strategies.
- Here: focus on statistical inference for the Gravity-TSIR model, which models spatiotemporal dynamics. This model presents several inferential and computational challenges.

Simple SIR models

Basic SIR models classify individuals as one of **susceptible** (S), **infected** (I) or **recovered** (R).

- Individuals are born into the susceptible class.
- Susceptible individuals have never come into contact with the disease and are able to catch the disease, after which they move into the infected class.
- Infected individuals spread the disease to susceptibles, and remain in the infected class (the infected period) before moving into the recovered class.
- Individuals in the recovered class are assumed to be immune for life.

Gravity T-SIR model

- Extension of the discrete time-series SIR (T-SIR) model (Bjornstad et al.2002; Grenfell et al. 2002) with explicit formulation of the spatial transmission between different host communities.
- Notation:
  - $I_{k,t}$ - number of **infected** individuals in city k at time t.
  - $S_{k,t}$ - number of **susceptible** individuals in city k at time t.
  - $d_{k,j}$ - **distance** between cities k and j.
  - $N_{k,t}$ - **population** of city k at time t.
  - $B_{k,t}$ - local number of new hosts (**births**) in city k at time t.
  - $L_{k,t}$ - number of infected people moved (**immigrants**) to city k at time t.
  - T cities, K time points.

Modeling incidences

Following Xia, Bjornstad and Grenfell (2004):

- Number of incidences of a disease at time $t+1$ for city k,

  $$I_{k,t+1} = \text{Poisson}(\lambda_{k,t+1}), \text{ where } \lambda_{k,t+1} = \beta_t S_{k,t}(I_{k,t} + L_{k,t})^\alpha.$$

- $\alpha, \{\beta_t\}$ are local transmission parameters.

Modeling susceptibles

- Number of susceptible individuals at time $t+1$ for city k is then modeled via balance equation (Bartlett, 1957):

$$S_{k,t+1} = S_{k,t} + B_{k,t} - I_{k,t+1}$$

- Finally, unobserved number of infected immigrants moved to city k at time t is modeled as:

$$L_{k,t} = \text{Gamma}(m_{k,t}, 1),$$

where

$$m_{k,t} = \theta N_{k,t}^{\tau_1} \sum_{j=1, j \neq k}^{K} \frac{(I_{jt})^{\tau_2}}{d_{k,j}^{\rho}}, \quad \theta, \tau_1, \tau_2, \rho > 0.$$

Statistical inference for measles

- **Measles data**
  - The UK Registrar General's data for 952 cities in England and Wales for years 1944-1966 of biweekly incidences of measles. Very rich spatio-temporal data.
  - Data for number of susceptibles from standard susceptible reconstruction algorithms (cf. Fine and Clarkson, 1982)
- **Parameters of the model**:
  - Reliable estimates of local transmission parameters α and $\{\beta_t\}$ are assumed known from previous work (Bjornstad et al. 2001).
- **Goal**: Infer unknown gravity parameters: $\theta, \tau_1, \tau_2, \rho$.

Challenges with likelihood-based inference

- Dimensions of the data (TK): 546*952 = 519,792.
- Number of infected immigrants $\{L_{k,t}\}$ are unobserved.
- The likelihood function is complicated:
  - Involves integrating over 519,792 latent variables.
  - Very expensive calculations per iteration.
- Approximate Bayesian computation (ABC) approaches are infeasible since simulating draws from this model is computationally expensive.

A simplified model and gridded MCMC

Simplify the model by fixing the number of immigrants (latent variables) at their means.

- Likelihood evaluations are still very expensive.
- Studying likelihood surface, learning about variability of estimates is computationally infeasible.

Gridded Metropolis-Hastings:

- We evaluate expensive parts of the likelihood on a grid of parameter values (can use parallel processors for this) and store these in a look-up table.
- M-H algorithm on discretized parameter space (on grid). M-H ratio evaluation is now much faster.

Results

- The gridded MCMC algorithm produces posterior distributions similar to a non-gridded MCMC algorithm, but *much* faster.
- Conclusions based on a simulation study:
  - Serious identifiability issues. Can only infer 2 of the 4 parameters.
  - In simulation studies: posterior (and likelihood) surface is peaked away from the true parameter values. There's a significant shift (bias) in parameter estimates.

Alternative approach

- Instead of likelihood-based approach, focus on important biological 'signatures' of the process. E.g. proportion of zeros (# of times no disease incidences in a city).
- Borrow ideas from computer model emulation, calibration (cf. Sacks et al. , 1989.)
  1. Simulate realizations from the gravity model at different parameter values.
  2. Use the signatures to define summary statistics.
  3. Find distance between summary statistics for the simulated process and the observations.
  4. Fit a Gaussian process to this distance, as a function of the parameters.
  5. Can obtain a likelihood and perform Bayesian inference for the gravity model parameters using the observations.

Inferential approach outline

- Gravity parameters, $\Theta = (\theta, \tau_1, \tau_2, \rho)$.
- Summary statistics (distance to observations) based on simulations at $\Theta_i, i = 1, \ldots, n$ parameter settings, $\mathbf{Y} = (\mathbf{Y}(\Theta_1), \ldots, \mathbf{Y}(\Theta_n))$.
- Model stochastic model output $\mathbf{Y}$ using a Gaussian process: $\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\xi} \sim N(\mu_{\boldsymbol{\beta}}(\Theta), \Sigma(\boldsymbol{\xi}, \Theta))$. Infer $\boldsymbol{\beta}, \boldsymbol{\xi}$: regression, covariance parameters.
- Model summary statistic for real data set $\mathbf{Z}$:
- $\mathbf{Z} = \boldsymbol{\eta}(\mathbf{Y}, \theta) + \boldsymbol{\delta}_{\Psi}(\mathbf{Y}, \Theta) + \boldsymbol{\epsilon}_{\sigma^2}(\mathbf{Y})$
  where $\boldsymbol{\eta}$ is a random variable with predictive distribution derived above. $\boldsymbol{\delta}$ is a discrepancy function, modeled as Gaussian process, and $\boldsymbol{\epsilon}$ is a vector of i.i.d. errors.
- Infer posterior $\pi(\Theta, \Psi, \sigma^2 \mid \mathbf{Z}, \mathbf{Y})$ using MCMC.

Conclusions

- Our GP-based emulation approach appears to produce unbiased estimates of the parameters.
- With estimated parameters, the model is able to reproduce well the signatures of the disease process.
- This is the first statistically rigorous approach to this problem: estimates of uncertainty, joint distributions of parameters, predictions/variability from fitted model.

Caveats and future work:

- Our statistical approach unearths serious identifiability issues: can still only learn about 2 parameters at most.
- Computational concerns only allow for a limited number of model forward runs.

Key references

- Xia, Y. C., Bjørnstad, O. N. and Grenfell, B. T. (2004), Measles Metapopulation Dynamics: A gravity model for epidemiological coupling and dynamics, *American Naturalist*.
- Bjørnstad, O. N., Finkenstädt, B. and Grenfell, B. T.(2001), Dynamics of measles epidemics. I. estimating scaling of transmission rates using a time series SIR model, *Ecological Monographs*.
- Bhat, K.S., Haran, M., Tonkonojenkov, R., and Keller, K. (2011), Inferring likelihoods and climate system characteristics from climate models and multiple tracers.
- Bhat, K.S., Haran, M., and Goes, M. (2010),

Acknowledgment: Support from Bill and Melinda Gates Foundation