Gaussian Processes for Inference with Implicit Likelihoods

Murali Haran
Department of Statistics
Pennsylvania State University

Microsoft Research
Redmond, Washington
January 2012

Complex Scientific Models

- Scientists are often interested in mechanisms 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

- Computationally expensive simulations
- May not be possible to write closed-form expressions relating input/parameters to output.
- (When stochastic) The likelihood function may be expensive to evaluate: hard to optimize or use Monte Carlo methods
- Non-ignorable discrepancies between model and reality.

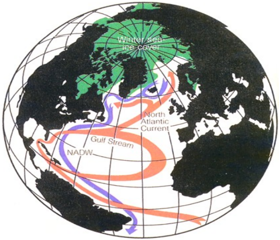
Likelihood is often *implicit* or has to be treated as such

Two Examples

I Climate: An Earth System Model of Intermediate Complexity (EMIC) for projecting the behavior of global ocean circulation systems.

II Disease Dynamics: A space-time model for the spread of infectious disease (measles).

The Meridional Overturning Circulation (MOC)



The Atlantic meridional overturning circulation (MOC) carries warm upper waters into far-northern latitudes and returns cold deep waters southward across the Equator.

Rahmstorf (Nature, 1997)

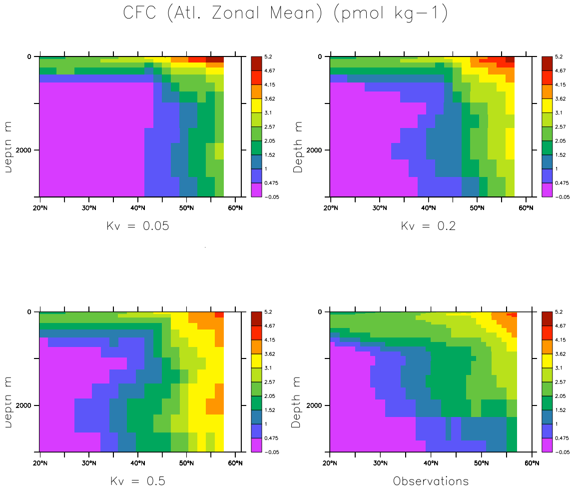
Climate Models: Learning About K_v

"Collapse" of MOC may result in dramatic climate change.

K_v is a key climate model parameter that influences the MOC.

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

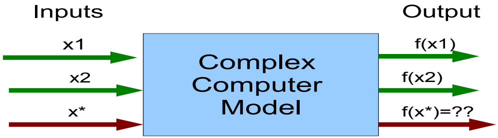
CFC-11 Example



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

Deterministic 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 emulator to a training set from complex model
- Advantages:
  - Fast approximate simulator
  - Uncertainties associated with prediction: greater uncertainty where there is less training data
    "Without any quantification of uncertainty, it is easy to dismiss computer models." (A.O'Hagan)
  - This provides a probability model

Modeling with Gaussian Processes

- Gaussian processes (GPs) are useful models for dependent processes, e.g. time series, spatial data.
- Also 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 $\mathbf{s} \in D \subset \mathbb{R}^d$ 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.
- Infinite-dimensional process. If $\mathbf{s}_1, \ldots, \mathbf{s}_n \in D$, $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T$ is multivariate normal.
- Parametric covariance, e.g. $\mathrm{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) = \kappa \exp(-\|\mathbf{s}_i - \mathbf{s}_j\|/\phi)$, $\kappa > 0, \phi > 0$. 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 $\Theta, \boldsymbol{\beta}$, and MCMC to learn about $\pi(\Theta, \boldsymbol{\beta} \mid \mathbf{Z})$.

GP Linear Model Prediction

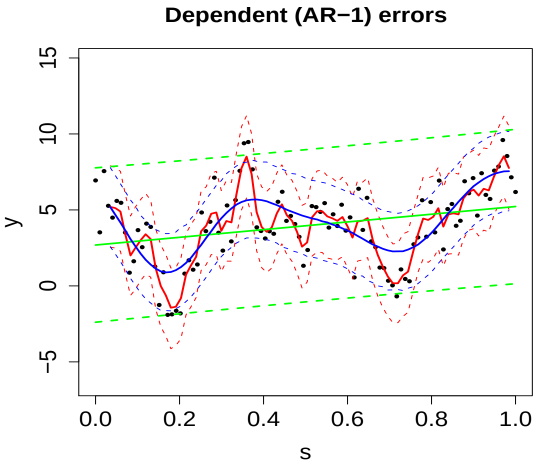
- Let the predictions at the new locations $\mathbf{s}_1^*, \ldots, \mathbf{s}_m^* \in D$ be $\mathbf{Z}^* = (Z(\mathbf{s}_1^*), \ldots, Z(\mathbf{s}_m^*))^T$.
- Under the GP assumption ($\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma$ depend on $\boldsymbol{\beta}, \Theta$)

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

  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*.

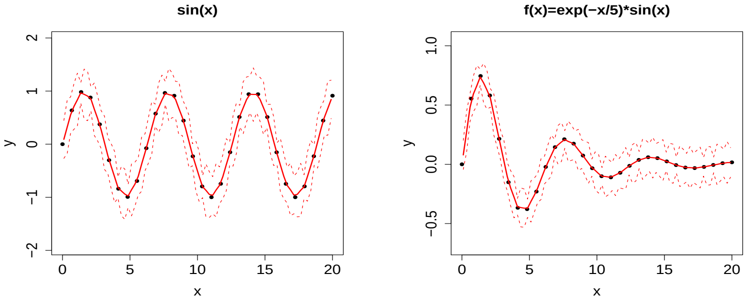
GP Model for Dependence: 1-D Example



Black: 1-D AR-1 process simulation. Green: independent error.

(Red, blue): GP with (exponential, gaussian) covariances.

GP for Function Approximation: 1-D Example



Same GP model used for both:

$$y(x) = \mu + w(x),\ \{w(x), x \in (0, 20)\}$$

Real data: bivariate spatial process at each input

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: $\mathbf{Y} = ((\mathbf{Y}_1(\psi_1), \mathbf{Y}_2(\psi_1)), (\mathbf{Y}_1(\psi_2), \mathbf{Y}_2(\psi_2)), \dots, (\mathbf{Y}_1(\psi_K), \mathbf{Y}_2(\psi_K)),$ 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:

- Usually real prior information about θ
- Likelihood surface for θ often multimodal; issues with identifiability. Nice to have 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 et al. (2007, 2008).

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 $\boldsymbol{\delta}(\mathbf{Y}) = (\boldsymbol{\delta}_1,\ \boldsymbol{\delta}_2)^T$ 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.

Inference with Multiple Spatial Fields: Step 1

Goals: (i) flexible model for relationship between $\mathbf{Y}_1$ and $\mathbf{Y}_2$, (ii) computational tractability.

- 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).
- $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2$ are regression, covariance parameters.

Inference with Multiple Spatial Fields: Step 2

- Emulation: Fit GP via maximum likelihood, then obtain predictive distribution at locations of observations
- Add model discrepancy and measurement error
- Model discrepancy term can make crucial adjustment to $\boldsymbol{\theta}$ estimates (Bayarri, Berger et al. 2007; Bhat et al., 2010).
- Separating stages: ‘modularization’ (e.g. Liu, Bayarri, Berger, 2009). Computational advantages + reduce identifiability issues.
- Use Markov chain Monte Carlo (MCMC) with slice sampler to estimate $\pi(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{Y})$ (integrating out rest)

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: $\mathcal{O}(J^3)$ where J ($\approx$ 300 here) is dimensionality of latent white noise process.

Kernel Mixing

- Model spatial dependence terms ($w(\mathbf{s})$) via kernel mixing of white noise process (Higdon, 1998, 2001).
- New process created by convolving a continuous white noise process with a kernel, k, which is a circular normal.

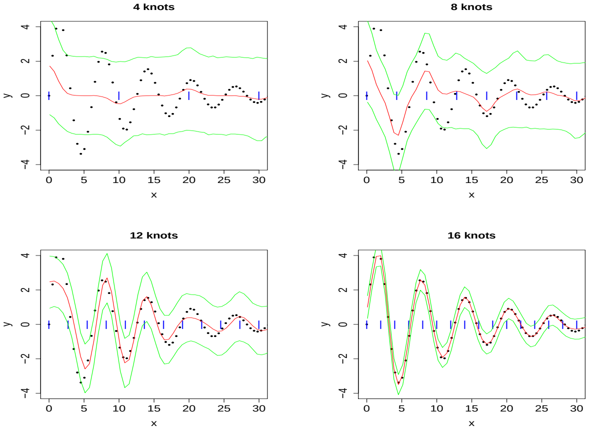
$$w(\mathbf{s}) = \int_D k(\mathbf{u} - \mathbf{s}) z(\mathbf{u}) d\mathbf{u}.$$

- Replace original GP by a finite sum approximation $\mathbf{z}$ defined on a lattice $\mathbf{u}_1, \ldots, \mathbf{u}_J$ (knot locations).

$$w(\mathbf{s}) = \sum_{j=1}^{J} k(\mathbf{u}_j - s) z(\mathbf{u}_j) + \mu(\mathbf{s}),$$

- Flexible: easily allows for non-stationarity and nonseparability. e.g. if k varies in space, have non-stationary process.

Kernel Mixing: Toy Example



- Dimension reduction: Computation involves only the J random variables $z_1, \ldots, z_J$ at the locations $\mathbf{u}_1, \ldots, \mathbf{u}_J$.
- Figures are for 4, 8, 12, and 16 knots.

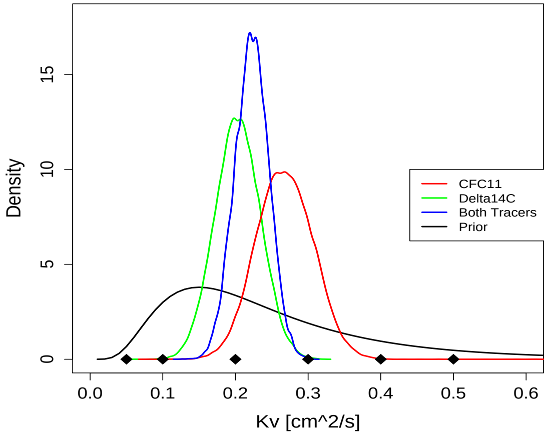
Matrix Identities

- Can use kernel mixing to obtain special covariance structure
- Sherman-Woodbury-Morrison identity: Suppose matrix is of form $A + UCV$, where A is $N \times N$, U is $N \times J$, V is $J \times N$, and C is $J \times J$. Its inverse:

  $$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

  Inversions involve $J \times J$ rather than $N \times N$ matrices (our e.g. $J = 190$ versus $N =$4,500.)

Results for K_v Inference



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

Result: K_v pdf suggests weakening of MOC in the future.

Summary of Climate Model Inference

Two-stage approach:

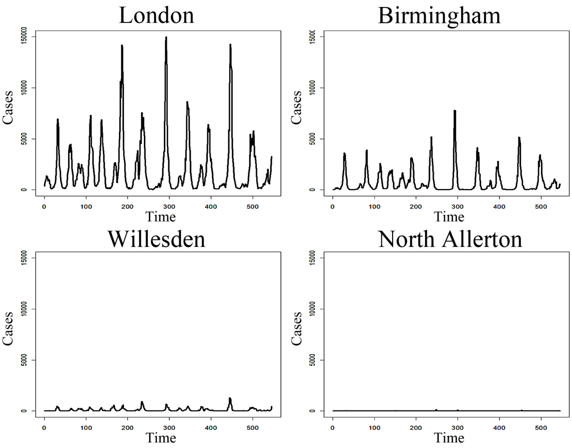
1. Obtain a probability model connecting CFC-11, $\Delta^{14}C$ tracer observations to K_v: hierarchical model using GPs + + patterned covariances so flexible and computationally tractable.
2. Using above probability model, infer K_v from observations

We can use inferred 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

- Gravity-TSIR model: Space-time model for spread of measles. Here θ=parameters controlling the dynamics of the spread of this disease e.g. how 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.

Measles Data



Notice: 952 cities of varying sizes and levels of infecteds

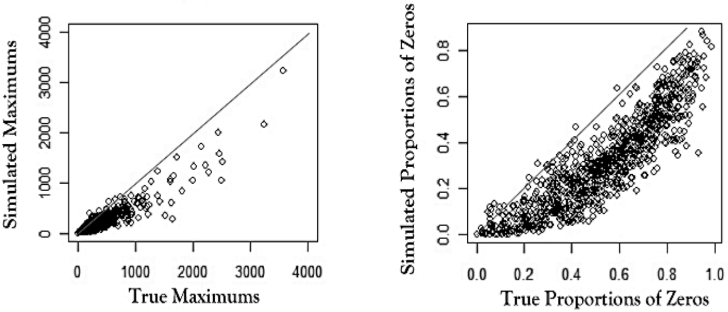
Complicates likelihood-based inference

Inference for Gravity TSIR Model Parameters

- Stochastic model: expensive to evaluate likelihood
- ABC (approximate Bayesian computing) approaches (Pritchard et al., 1999): infeasible due to simulation time.
- We develop approximate grid-based Markov chain Monte Carlo approach that is computationally tractable.
- However, traditional likelihood-based/Bayesian inference even with tractable computing is problematic:
  - Does not fit scientifically relevant features of the data
  - Simulations reveal: do not recover θ

Traditional Likelihood-Based Approach

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



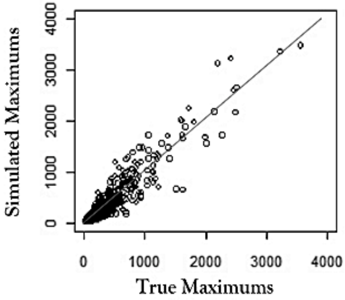
Inference for Gravity TSIR Model Parameters

- 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 match data well



Simulations show: can recover θ using this approach

Summary

- Gaussian processes are a powerful tool when likelihood is implicit and simulating from the model is expensive
- GPs are useful for deterministic and stochastic models
- Can perform inference based on scientifically important features of the data
  - Computationally expedient
  - May improve inference and prediction

Collaborators

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

Support:

- Bill & Melinda Gates Foundation
- U.S. Geological Survey
- National Science Foundation (NSF-HSD)

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., 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:

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

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