Gaussian processes for inference with implicit likelihoods

Murali Haran[1]

Collaborators: K. Sham Bhat[2], Roman Tonkonojenkov [3], Klaus Keller [3], Roman Jandarov[1] and Ottar Bjornstad[4]

[1]Department of Statistics, The Pennsylvania State University

[2]Los Alamos National Laboratories

[3]Department of Geosciences, The Pennsylvania State University

[4]Center for Infectious Disease Dynamics, The Pennsylvania State University

The Fourth Erich L. Lehmann Symposium

Rice University, May 2011.

Complex Scientific Models

- 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. Purely statistical approaches may not answer questions of interest or permit sound predictions.
- 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

- As models become more scientifically plausible, they typically become more complex. Inference for unknown parameters may be challenging for the following reasons:
  - Simulations from the model may be computationally expensive.
  - May not be possible to write closed-form expressions relating input (parameters) to output.
  - 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 so even a perfectly calibrated model will not match observations well.
- The likelihood function 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

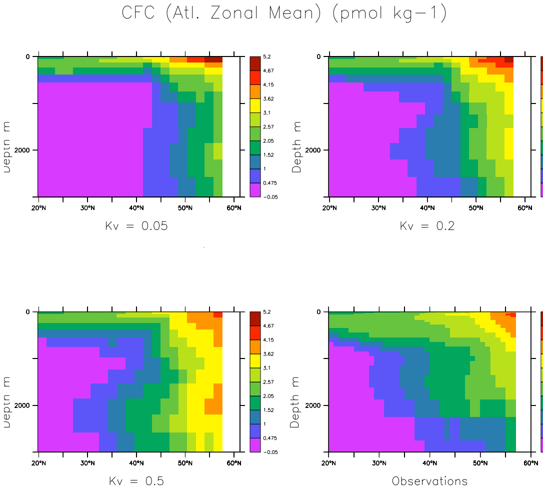
II. Disease Dynamics: A Gravity Time Series Susceptible-Infected-Recovered (TSIR) model for the spread of infectious disease (measles).
   - Stochastic
   - Likelihood is expensive to evaluate
   - Space-time process with large number of “no incidence” observations (0s)

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

The meridional overturning circulation (MOC) has a strong influence on global climate. Collapse of this ocean "conveyor belt" 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) collected in the 1990s (latitude, longitude, depth), zonally averaged: $\mathbf{Z}_1, \mathbf{Z}_2$
  - 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): $\mathbf{Y}_1(\mathbf{K_v}), \mathbf{Y}_2(\mathbf{K_v})$

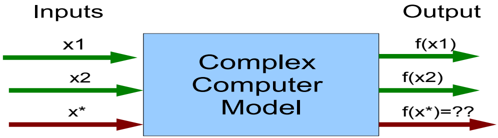
CFC-11 Example



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

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 an emulator ("meta model") to a training set of runs from the complex computer model.
- The emulator serves as a surrogate for the computer model, and is much faster/simpler. Hence, it is possible to simulate output from the emulator very quickly.
- Advantages of doing it in a probabilistic framework:
  - 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.
- Convenient to specify covariance by a 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 $\Theta, \boldsymbol{\beta}$, and MCMC to learn about $\pi(\Theta, \boldsymbol{\beta} \mid \mathbf{Z})$.

GP Linear Model Prediction

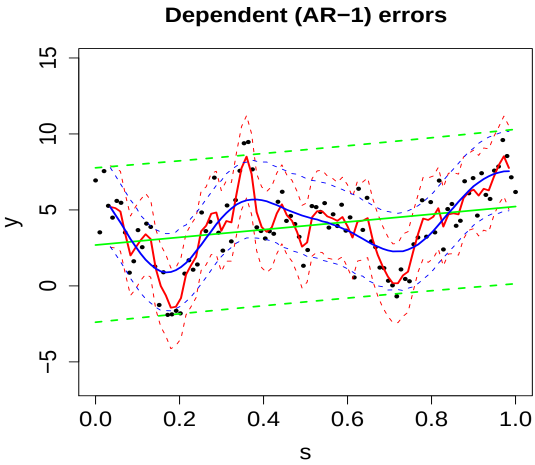
- Let the predictions at the new locations $\mathbf{s}_1^*, \ldots, \mathbf{s}_m^* \in D$ be $\mathbf{Z}^* = (Z(\mathbf{s}_1^*), \ldots, Z(\mathbf{s}_m^*))^T$.
- Under the GP assumption ($\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma$ depend on $\boldsymbol{\beta}, \Theta$):

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

  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*.

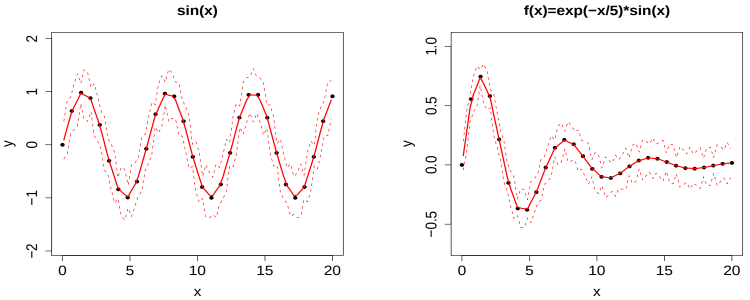
GP Model for Dependence: Toy 1-D Example



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

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

GP for Function Approximation: Toy 1-D Example



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

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

Summary of Inferential ("Calibration") Problem

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

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 $\boldsymbol{\theta}$ 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).

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

Two-stage Approach to Calibration

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 emulation of model output $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$.
   - $\mathbf{Z}$ is assumed to be a realization of computer model at "true" $\boldsymbol{\theta}$ + model-data discrepancy + measurement error.
   - Discrepancy term accounts for "structural uncertainty", i.e., understanding that the the model will have systematic errors and biases. Emulation done via a Gaussian process model. Add a discrepancy term and additional source of error to this model. This provides a probability model for $\mathbf{Z}$ in terms of $\boldsymbol{\theta}$.
2. Specify a prior for $\boldsymbol{\theta}$ and use observations $\mathbf{Z}$ to infer $\boldsymbol{\theta}$ (parameter of interest) based on above probability model.

Inference with Multiple Spatial Fields: Step 1

Need (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).
- $\boldsymbol{\beta}$s, $\boldsymbol{\xi}$s 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:

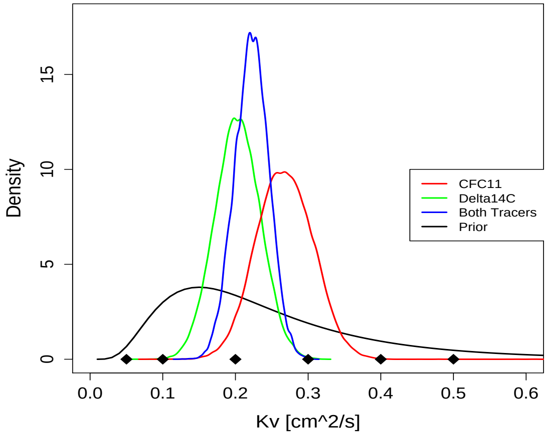
$$\mathbf{Z} = \boldsymbol{\eta}(\mathbf{Y}, \boldsymbol{\theta}) + \boldsymbol{\delta}(\mathbf{Y}) + \boldsymbol{\epsilon}$$

  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 is in tens of thousands. Computationally intractable without some dimension reduction.*
- Need long MCMC runs since there may be multimodality issues, and the chain mixes slowly.
- We use 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.
- In MLE step: take advantage of structure of hierarchical model to reduce 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

1. Our approach is to perform inference in two stages:
   - Obtain a probability model connecting CFC-11, $\Delta^{14}C$ tracer observations to $\mathbf{K_v}$ by fitting a Gaussian process model to climate model runs.
   - Using this probability model, infer a posterior density for $\mathbf{K_v}$ from the observations.
2. We model multivariate spatial data via a flexible hierarchical structure.
3. We use kernel mixing to obtain patterned covariances, making computations tractable for large data sets.

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

- 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.

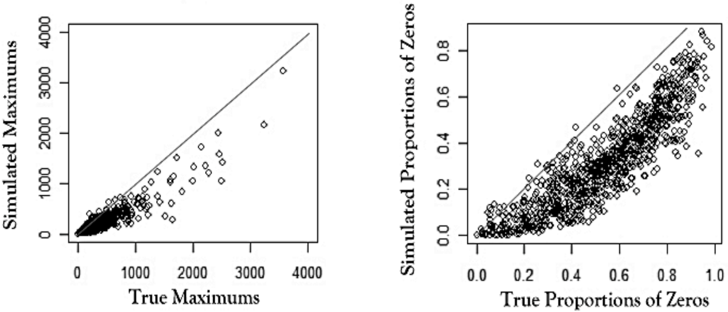
  Potential for learning about parameters, but also poses computational challenges.

Inference for Gravity TSIR Model Parameters

- An approximate grid-based Markov chain Monte Carlo provides a way out of the computational challenges.
- However, traditional likelihood-based/Bayesian inference does not result in a fitted model that reproduces scientifically relevant features of the data.
- Instead, fit GP to *summary statistics* of model runs where summaries are based on scientifically relevant features.
- Calibration based on using this GP with the data results in improved inference.

Traditional Likelihood-Based Approach

Simulations from fitted model do not match up well with the data for important characteristics of the process.

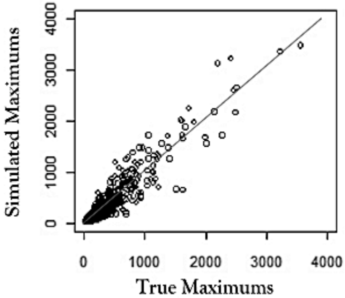


Inference for Gravity TSIR Model Parameters

- Instead, fit GP to *summary statistics* of model runs where summaries are based on scientifically relevant features.
- Calibration based on using this GP with the data results in improved inference.

GP-based Calibrations 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.
- GPs are useful for deterministic and stochastic models.
- Important to take computational complexity into account and devise strategies to expedite computing.
- When traditional likelihood-based approaches are unsatisfactory because they ignore scientifically important features of the data, GP-based approaches can provide an alternative that directly uses the important features of the data.
- Limitation: computationally intractable when the number of parameters of interest (dimensionality of θ) is large.

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:

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

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