

# Gaussian processes, Monte Carlo, and an application to climate science

Murali Haran

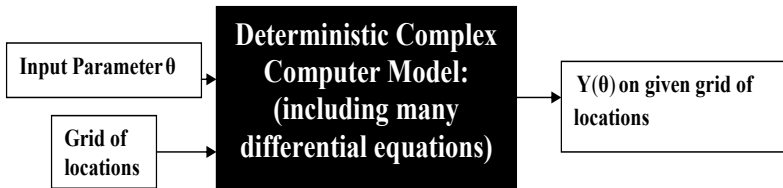
Department of Statistics  
Penn State University

Penn State Computational Science Lectures  
February 2009

# Spatial Models

- ▶ Lots of scientific questions involve analyzing data that are **spatially dependent**: Data points close together are more closely related (dependent) than data further away.
- ▶ Examples:
  - ▶ Concentrations of PM2.5 (air pollutants) across the U.S.
  - ▶ Disease rates by county.
  - ▶ Abundance of plant/animal species across Pennsylvania.
- ▶ ‘Space’ does not always mean physical distance. Similar ideas are applicable to many other research areas:
  - ▶ Machine learning: two objects may be close in ‘feature space’.
  - ▶ Approximations to computationally expensive computer models.

# Computer model emulation



- ▶ An example of above: climate model runs (collaborative work between Statistics and Geosciences at PSU).
- ▶ **Emulation** involves replacing a complicated computer model with a simpler (usually stochastic) approximation.
- ▶ Sacks et. al. (1989) introduced a Gaussian process (GP or 'kriging') model as an emulator for a complicated function.
- ▶ Using model: Obtain approximate output at any parameter setting along with associated uncertainty.

## Application to climate science

- ▶ Want to learn about important unobservable characteristics of climate system, e.g. ‘climate sensitivity’. Learning about these characteristics helps understand climate system today and project into the future.
- ▶ Sources of information:
  - ▶ Instrumental observations of ‘tracers’ e.g. CFC (chlorofluorocarbon) in ocean at different depths, lat/long. Space-time data:  $\mathbf{Z}$ .
  - ▶ Have climate model runs at different input values of climate parameter (e.g. climate sensitivity  $\theta$ ). This produces output of tracers at different depths, lat/long, not necessarily same as above. Call this:  $\mathbf{Y}(\theta)$ .
- ▶ Combine  $\mathbf{Y}$ ,  $\mathbf{Z}$  to learn about  $\theta$ : want *distribution* on  $\theta$ .
- ▶ Can use Gaussian processes and Monte Carlo to do this.

## GP Models: modeling a surface

- ▶ Want a spatial process over a region — this works both for statistically interpolating over space (conventional spatial modeling), as well as interpolating a surface over a parameter space (GP models for computer experiments).
- ▶ Simplest way to define a joint distribution is to assign a joint normal distribution to a finite set of points in that region. The set of points=location of observations and locations where we want to predict.
- ▶ Note: want this to be true (want a joint normal distribution) for *any* finite set of points in this region. This is therefore an infinite dimensional distribution called a Gaussian *process* (**GP**).

## Spatial Models: General Ideas (Contd.)

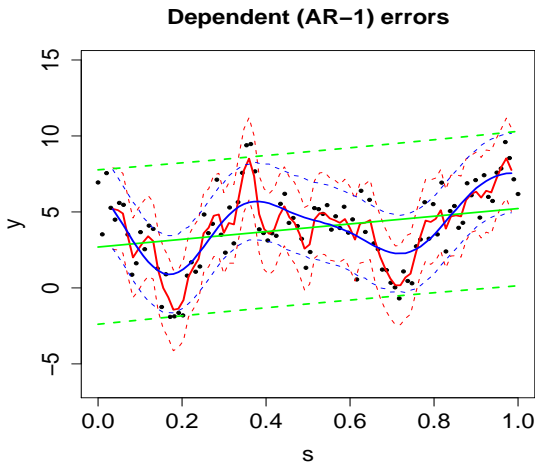
Consider a joint distribution for 3 locations:

- ▶ Multivariate Normal:

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \sim N \left( \begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \right).$$

- ▶ We want the dependence (characterized by covariance matrix) to be related to the distance between the locations.
- ▶ E.g. covariance:  $\Sigma_{ij} = \psi \exp(-(\|s_i - s_j\|)/\phi)$ ,  $\phi > 0$ ,  $\psi > 0$ , where  $s_i$  is the location of the  $i$ th observation.
- ▶ If the distance between  $i$ th and  $j$ th locations is large,  $\Sigma_{ij}$  will be small. If the locations are close  $\Sigma_{ij}$  will be large and hence  $z_i$  and  $z_j$  will be strongly dependent.

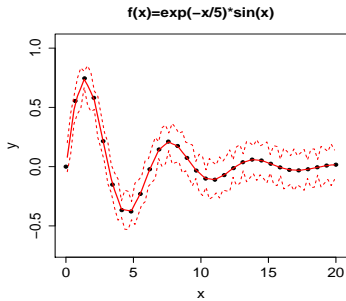
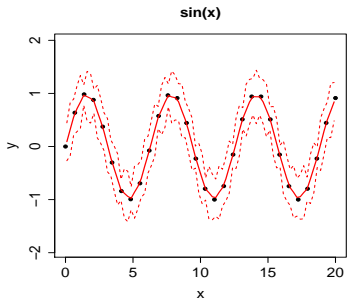
# GP model for dependence: toy 1-D example



Black: 1-D AR-1 process simulation. Green: independent error.  
Red: GP with exponential, Blue: GP with gaussian covariance.

# GP model for emulation

Want to fit (interpolate) a non-linear function.



Functions:  $f(x) = \sin(x)$  and  $f(x) = \exp(-x/5) \sin(x)$ .

Both were fit with linear GP model,  $f(x) = \alpha + \epsilon(x)$ , where

$\{\epsilon(x), x \in (0, 20)\}$  is a GP,  $\alpha$  is just a constant mean

(surprising: did not know form of function; used dependence instead!)

## Likelihood-based Inference

- ▶ Back to our problem: have  $\mathbf{Z}$  and  $\mathbf{Y}$ , want to learn about  $\theta$ .
- ▶ **Likelihood**,  $\mathcal{L}(\mathbf{Z}; \theta)$  connects observations ( $\mathbf{Z}$ ) to parameters ( $\theta$ ).
- ▶ Maximum likelihood (ML/'frequentist') approach: maximize  $\mathcal{L}(\mathbf{Z}; \theta)$  w.r.t.  $\theta$  to obtain  $\hat{\theta}$ , ML estimate.
- ▶ Bayesian approach: treat  $\theta$  as random variable(s) and specify **prior distribution**  $f(\theta)$ .
- ▶ Inference is based on **posterior distribution**,

$$\pi(\theta|\mathbf{Z}) = \frac{\mathcal{L}(\mathbf{Z} | \theta)f(\theta)}{\int \mathcal{L}(\mathbf{Z} | \theta)f(\theta)d\theta} \propto \mathcal{L}(\mathbf{Z} | \theta)f(\theta)$$

(This is just an application of Bayes' rule)

- ▶ We don't have  $\mathcal{L}$ , i.e., a connection between  $\theta$ ,  $\mathbf{Z}$ ! Need to use  $\mathbf{Y}(\theta)$  to learn about it.

## Inference for climate science

We carry out inference in stages (Bhat, Haran, Tonkonojenkov, Keller, 2009):

1. Use GP model to simultaneously do emulation — model nonlinear relationship between  $\theta$  and  $\mathbf{Y}$ , as well as model space-time dependence in  $\mathbf{Y}$ . Fit this model via maximum likelihood. This involves optimization, and provides  $\hat{\mathcal{L}}(\mathbf{Z}, \theta)$ ,
2. Impose a prior on  $\theta$ , say  $f(\theta)$ , based on expert knowledge.

Can try multiple priors if experts disagree. Now use Bayesian inference in obtain posterior distribution

$\pi(\theta | \mathbf{Z}) \propto \frac{\hat{\mathcal{L}}(\mathbf{Z}|\theta)f(\theta)}{\int \hat{\mathcal{L}}(\mathbf{Z}|\theta)f(\theta)d\theta}$ . This posterior distribution is complicated. Need Monte Carlo methods.

# Monte Carlo for Inference

All inference for the model is based on the posterior ( $\pi$ ).  
For e.g.  $E_{\pi}(\theta|\mathbf{Z})$ , the posterior expectation of parameter  $\theta$ .  
In general we are interested in expectations of the form:

$$E_{\pi}g = \int g(x)\pi(x)dx$$

Integral is too hard, so use **sample based inference**.

- ▶ We simulate  $X_1, \dots, X_N$  from the distribution  $\pi$ . Use sample average:  $\sum_{i=1}^N g(X_i)/N$ .
- ▶ In principle, if we have enough samples (large enough  $N$ ), we can answer any question of interest.
- ▶ Example: What is the probability that  $\theta > 2$ ? Estimate: Count the proportion of times sampled  $\theta > 2$ .

## Monte Carlo: basic theory

- ▶ Assume  $X_1, \dots, X_N \stackrel{iid}{\sim} \pi$ .
- ▶ Strong Law holds: If  $E_\pi |g| < \infty$  then

$$\bar{g}_N = \sum_{i=1}^N g(X_i)/N \rightarrow E_\pi g \text{ as } N \rightarrow \infty.$$

- ▶ Central Limit Theorem: If  $E_\pi g^2 < \infty$  we have

$$\sqrt{N}(\bar{g}_N - E_\pi g) \rightarrow N(0, \sigma^2)$$

- ▶ Easy to estimate  $\sigma^2$  using sample variance ( $\hat{\sigma}^2$ ).
- ▶ Estimate accuracy of our estimate by  $\hat{\sigma}^2/N$ .
- ▶  $N$  is large enough when  $\hat{\sigma}^2/N$  is small enough.
- ▶ Note that  $X_i$ s can be multidimensional (accuracy is unaffected by the dimension of the problem here.)

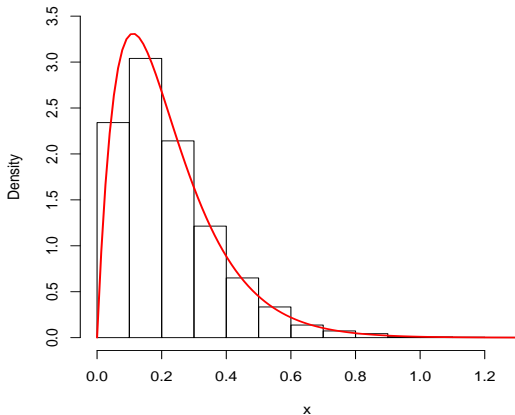
## Monte Carlo: Toy posterior distribution

Suppose  $\pi(\theta | \mathbf{Z}) = \text{Gamma}(2, 9)$ . Sample-based estimates:

$\Pr(\theta > 0.5) \approx 0.0613$ . (error = 0.0005, truth = 0.0610).

Mean  $\approx 0.222$  (error = 0.001, truth = 0.222)

Histogram with true density overlaid



# Markov chain Monte Carlo

- ▶ Life is simple with i.i.d. Monte Carlo.
- ▶ Generally very difficult to draw i.i.d. samples from  $\pi$ .
- ▶ More general approach: **Metropolis-Hastings algorithm**.
  - ▶ Start with an initial value  $X_0$ . For  $i = 2$  to  $N$ :
  - ▶ Propose a value  $X^*$  for  $X_i$  based on  $X_{i-1}$ .
  - ▶ Set  $X_i = X^*$  with M-H probability depending on  $X_i, X^*$ , the proposal distribution and the target distribution ( $\pi$ ).
- ▶ The Markov chain  $X_1, \dots, X_N$  has stationary distribution  $\pi$  (roughly: for large values of  $N$ ,  $X_N$  is approximately distributed according to  $\pi$ .)

## Markov chain Monte Carlo (contd.)

- ▶ Use  $X_1, \dots, X_N$  as before to estimate  $E_\pi g$ .
- ▶ Strong Law holds if  $E_\pi(|g|) < \infty$ .

$$\bar{g}_N = \sum_{i=1}^N g(X_i)/N \rightarrow E_\pi g \text{ as } N \rightarrow \infty$$

We also need technical conditions on the Markov chain but these typically hold by construction.

- ▶ It appears as if we have the same situation as in the i.i.d. case, except we need to be *very careful* about assessing accuracy of estimates and determining how long to run the chain. One approach: Flegal, Haran, Jones (2008) — assess s.error of estimates and stop when they attain a threshold.

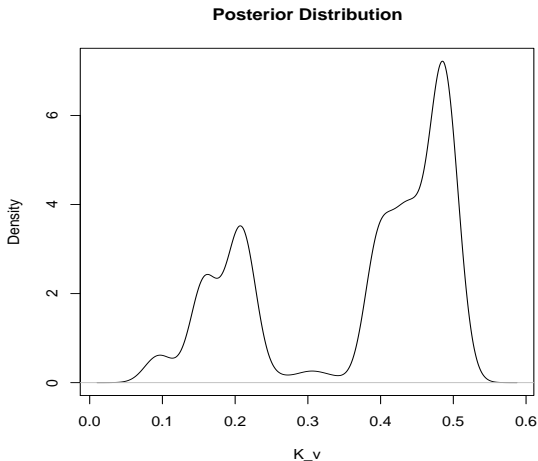
## Climate science problem

Example:  $K_V$  inference based on CFC Data.

- ▶ One climate parameter  $\theta$ , which is  $K_V$  ( $\text{cm}^2\text{s}^{-1}$ ). An EMIC (Earth Model of Intermediate Complexity) was run at different settings of  $K_V$  from 0.05 to 0.5 to produce CFC tracer output at each of these settings.
- ▶ Have instrumental data on CFC.
- ▶  $K_V$  is of scientific interest because it is an important characteristic related to the ocean circulation system, specifically the MOC (meridional overturning circulation) which plays a major role in global climate. Changes in MOC may result in disruptions to equilibrium state in the climate. Specifically, small  $K_V \Rightarrow$  low MOC, and higher sensitivity to anthropogenic (human-induced) forcings.

## Putting it all together for $K_v$ inference

Inferred a likelihood from climate model runs (using GP model).  
Then used this likelihood and Markov chain Monte Carlo to estimate distribution of  $K_v$ . After heavy computations....



## What this talk was about

- ▶ Using a climate science problem as motivation, provided a whirlwind tour of several (vast) topics:
  - ▶ Use of **Gaussian processes** to model spatial dependence and computer experiments.
  - ▶ **Likelihood and Bayesian inference**: If we can connect data ( $\mathbf{Z}$ ) to some parameters ( $\theta$ ) via a statistical model, can learn about  $\theta$  via likelihood/Bayes approach.
  - ▶ **Monte Carlo**/Markov chain Monte Carlo methods: powerful computational approach to help perform inference.
- ▶ Skipped lots of details. Example: expensive matrix operations involved — order  $n^3$  at each iteration of algorithm, where  $n$  is the number of data points. We actually use a different version of Gaussian process model to solve this computational problem.

## Some references

- ▶ Bayesian inference and Monte Carlo methods: Carlin and Louis (2009) “Bayesian Methods for Data Analysis.”
- ▶ Flegal, J.M., Haran, M. and Jones, G.L. (2008) “Markov chain Monte Carlo: Can We Trust the Third Decimal Place?”. R code available from my webpage.
- ▶ Gaussian processes:
  - ▶ Rasmussen and Williams (2006) “Gaussian processes for machine learning”.
  - ▶ Schabenberger and Gotway (2005) “Statistical methods for spatial data analysis”.
- ▶ Bhat, K.S., Haran, M., Tonkonojenkov, J., and Keller, K. “Inferring likelihoods and climate system characteristics from climate models and spatio-temporal tracer data” (in preparation).