Inference with Implicit Likelihoods and High-dimensional Data

Murali Haran

Department of Statistics
Pennsylvania State University

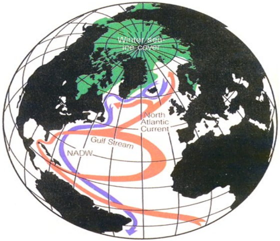
joint with:
Won Chang (Statistics, Penn State)
Sham Bhat (Los Alamos National Labs)
Roman Jandarov (University of Washington Biostatistics)
Roman Olson (Geosciences, Penn State)
Klaus Keller (Geosciences, Penn State)
Ottar Bjornstad (Center for Infectious Disease Dynamics, Penn State)

University of Minnesota, Biostatistics. November 2012

What This Talk is About

- Models for complex physical systems can be used to inform science and policy
  - Climate models: projections about future climate
  - Infectious disease models: design intervention strategies
- These models are based on the dynamics underlying the systems. Complicated and involve unknown parameters
- I will discuss "calibration" methods: how to use high-dimensional multivariate (spatial/space-time) observations of the system to infer unknown parameters

The Atlantic Meridional Overturning Circulation (MOC)



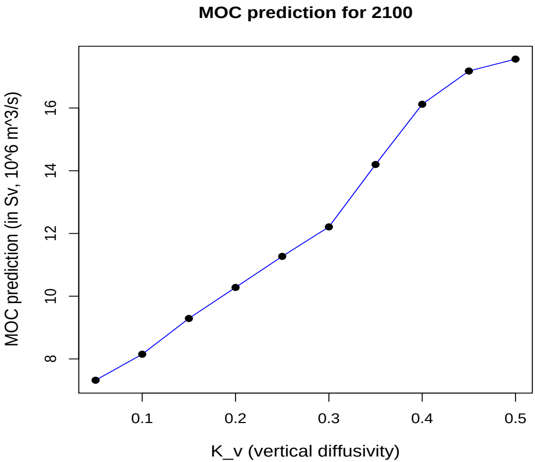
Rahmstorf (1997)

Global conveyor belt: carries warm upper waters into far-northern latitudes and returns cold deep waters southward across the equator

The MOC and Climate Change

- Its heat transport makes a substantial contribution to the moderate climate of Europe (cf. Bryden et al., 2005)
- Any slowdown in the overturning circulation would have profound implications for climate change
- Climate scientists use climate models to make projections about the MOC

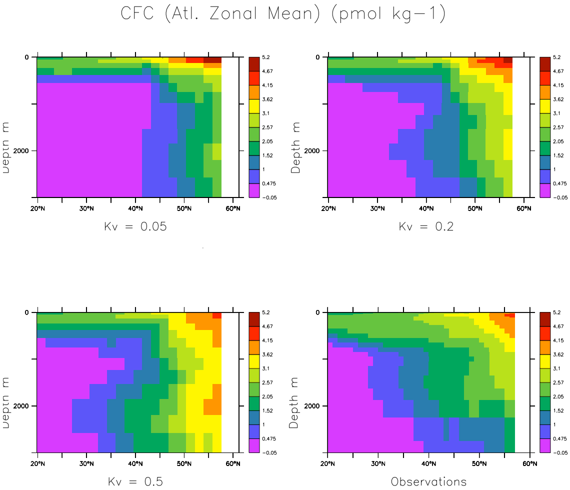
MOC Predictions and Model Parameter K_v



Learning about K_v

- K_v is a model parameter that quantifies the intensity of vertical mixing in the ocean. Cannot be measured directly
- Two sources of indirect information:
  - **Observations** of ocean "tracers" that provide information about K_v. Examples: $\Delta^{14}C$ and trichlorofluoromethane (CFC11) collected in the 1990s
  - **Climate model output** at different values of K_v from University of Victoria (**UVic**) Earth System Climate Model (Weaver et. al., 2001)
- Each tracer has
  - 2D spatial observations: 3706 locations
  - 2D model output: 5926 locations at *at each parameter setting*
- (Later) 3D spatial observations: 61,000 locations

CFC-11 Example: 2-D



Bottom right corner: observations

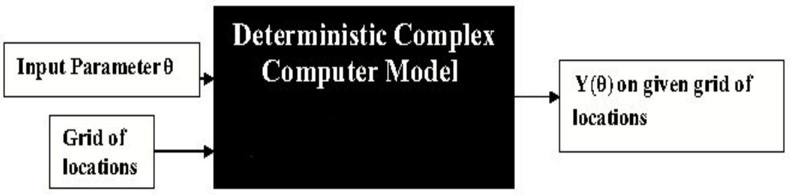
Other plots: climate model output at 3 settings of K_v

Challenges

This is a computer model calibration problem

1. The climate model is computationally intensive: can only be run at a few different settings
2. Output/observations are in the form of multivariate spatial data. (Toy e.g. was scalar!) Poses modeling, computational challenges
3. Combining information from tracers CFC-11, $\Delta^{14}C$: need a computationally tractable model for flexible relationships *between* the spatial fields.

Computer Model Emulation

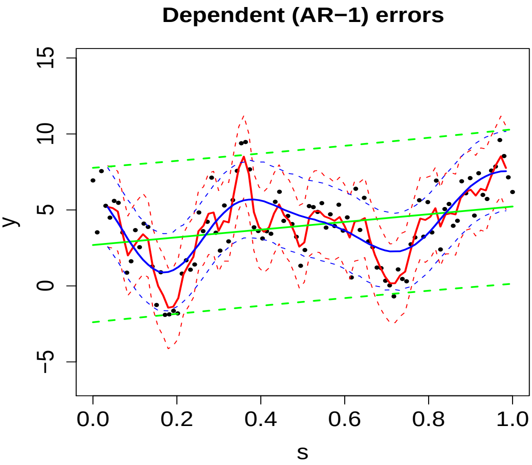


- Replace complicated computer model with a stochastic approximation: Gaussian process (Sacks et al., 1989)
- Gaussian processes (GPs) are infinite-dimensional spatial process. Joint distribution at any finite set of locations is multivariate normal

  For computer models "location" = parameter (input) setting

Currin et al. (1991); Bayarri, Berger et al. (2007); Sanso et al. (2008)

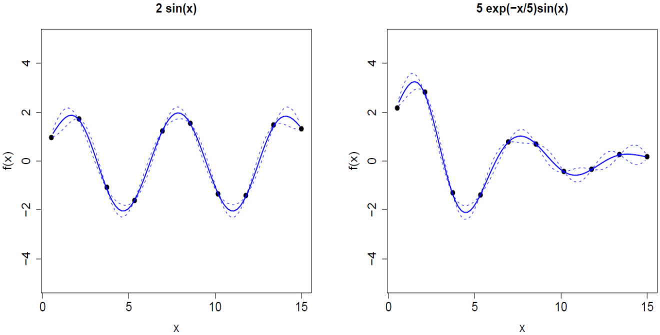
GP Model for Dependence: Toy 1-D Example



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

Red: GP with exponential, Blue: GP with gaussian covariance.

GP Model for Emulation: Toy 1-D Example



Same simple model for both, $f(x) = \alpha + w(x)$ where $\{w(x),\ x \in (0, 15)\}$ is a Gaussian process

Notation

- $Z_1(\mathbf{s}), Z_2(\mathbf{s})$: tracer 1 and 2 at location **s**=(latitude, depth). Let $\mathbf{Z}_1, \mathbf{Z}_2$ be the two spatial fields
- $Y_1(\mathbf{s}, \theta)$, $Y_2(\mathbf{s}, \theta)$: model output at **s**, θ
  Let $\mathbf{Y}_1, \mathbf{Y}_2$ be the model output for the two tracers, spatial fields across multiple parameter settings

Goal: Inference for climate parameter $\boldsymbol{\theta}$ using $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Y}_1, \mathbf{Y}_2$.
We will exploit the fact that GPs can be used to model complicated functions *and* spatial data

Two-Stage Computer Model Calibration

Our approach

1. **Emulation**: Model relationship between $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ and $\boldsymbol{\theta}$ via emulation of model output.
   - i An approximation to the computer model using $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$: $f(\mathbf{Y} \mid \boldsymbol{\theta})$
   - ii Take above approximation + systematic model-data discrepancy + measurement error. This gives a model for the observations $\mathbf{Z}$: $f(\mathbf{Z} \mid \boldsymbol{\theta})$
2. **Calibration**: obtain posterior distribution of $\boldsymbol{\theta}$,

$$\pi(\boldsymbol{\theta} \mid \mathbf{Z}) \propto f(\mathbf{Z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Step 1: Emulation with Multiple Spatial Fields

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

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

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

- $\mathbf{B}(\gamma)$ relates $\mathbf{Y}_1$ and $\mathbf{Y}_2$, with parameters γ
- *Covariance is a function of spatial distance and distance in parameter space*
- $\boldsymbol{\beta}$s, $\boldsymbol{\xi}$s are regression, covariance parameters

Flexible relationship between $\mathbf{Y}_1$ and $\mathbf{Y}_2$

Step 2: Calibration with Multiple Spatial Fields

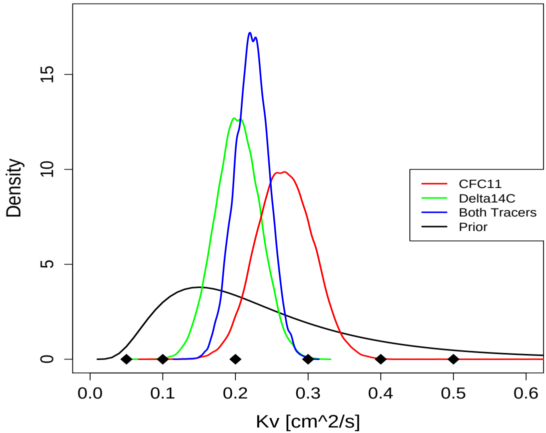
- Fit GP via maximum likelihood, then obtain predictive distribution at locations of observations
- Model observations by adding measurement error and a model discrepancy term to the GP emulator:

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

  where $\boldsymbol{\delta}(\mathbf{Y}) = (\boldsymbol{\delta}_1 \ \boldsymbol{\delta}_2)^T$ is the model discrepancy,
  $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1 \ \boldsymbol{\epsilon}_2)^T$ is the observation error
  Discrepancy can make crucial adjustments to $\boldsymbol{\theta}$ inference (Bayarri et al. 2007; Bhat et al., 2010)
- Markov chain Monte Carlo (MCMC) to obtain $\pi(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{Y})$

Details: kernel mixing + patterned covariances for fast matrix operations; discrepancy function; MCMC algorithm

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.

Alternate Sources of Information

Can also learn about $\mathbf{K_v}$ via sea temperatures

- Scientific interest: how does aggregation affect inference? At what spatial scale should we be looking at information?
- Statistical question: compare calibration based on 1-D, 2-D versus 3-D information
- Methodological issue: existing approaches (ours, Higdon et al. (2008); Sanso et al. (2008); Bayarri et al. (2008) etc.) do not apply to this 3D spatial data with 61,051 data points $\times$ 250 parameter settings

Fast Approach for High-dimensional Calibration

- Construct low-dimensional representation of model output $\mathbf{Y}$ and observations $\mathbf{Z}$
  - Find eigenvectors $\mathbf{K}_Y$ and corresponding principal components of model output. Low-dimensional representation of model output: $\mathbf{Y}_R$
  - Project $\mathbf{Z}$ on space spanned by $\mathbf{K} = [\mathbf{K}_y \ \mathbf{K}_d]$ where $\mathbf{K}_d$ is kernel basis for discrepancy. Low-dimensional representation: $\mathbf{Z}_R$, still accounting for discrepancy
- Emulation and calibration as before, but with $\mathbf{Y}_R, \mathbf{Z}_R$
- Crucial details: determining discrepancy basis, # of PCs, . . .

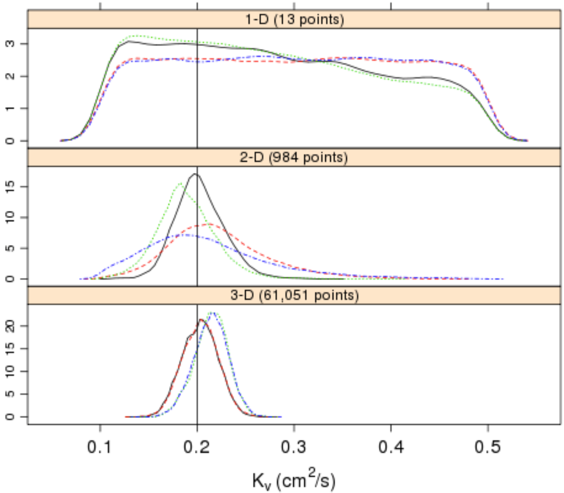
Simulated Example

Studied several simulated examples. Most challenging:

- Synthetic truth: 3-D model output at $\mathbf{K_v}$ = 0.2
- Pseudo-residual= averaged residuals between data and model at a few settings. This is more sensible, realistic, challenging than simulating from various error models (cf. Jim Hodges' recent work)
- Pseudo observational data in 3D= synthetic truth + pseudo-residual
- Aggregate 3-D pseudo observations into 2-D and 1-D.

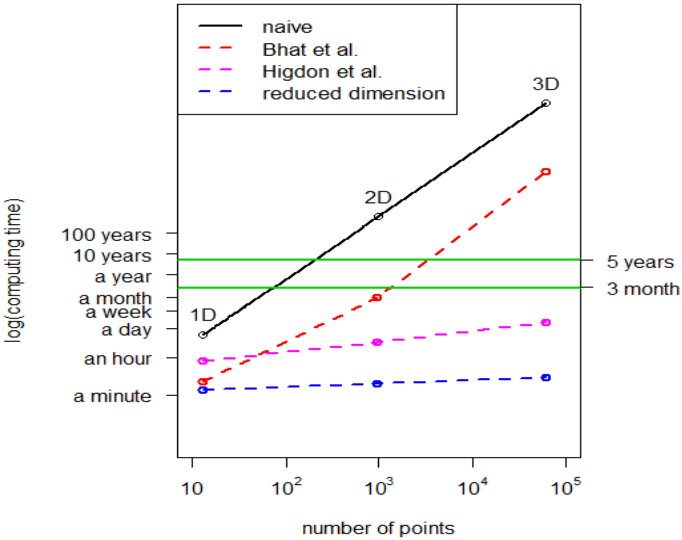
Compare 1D versus 2D versus 3D inference

Effect of Aggregation on Inference



Simulated example: Unaggregated 3-D data (1) has sharpest posterior pdf and (2) most robust to changes in prior

Computational Cost



Summary

1. Our approach:
   - Multivariate spatial data model via flexible hierarchical structure
   - Kernel mixing/patterned covariances and matrix identities (e.g. Sherman-Woodbury-Morrison) for fast computing
   - Reliability of approach was studied extensively
2. For high-dimensional spatial output: dimension-reduction approach for emulation and calibration. Very fast and our study suggests that it works well in practice. Allows for the first time an analysis based on 3D tracers
3. Regardless of tracers, aggregation, model or methods: MOC projected to weaken in the future

Collaborators

- Sham Bhat, Los Alamos National Laboratories
- Won Chang, Statistics, Penn State University
- Roman Olson, Department of Geosciences, Penn State University
- Klaus Keller, Department of Geosciences, Penn State University

Calibration with Large Spatial Data

- Basis-representation approaches (Higdon et al., 2008, and Bayarri et al., 2008) are very effective but do not extend in obvious fashion to our problem but have some shortcomings
- Higdon et al.(JASA, 2008): May become computationally expensive if number of parameter settings and/or required number of principal components are too large (requires inversion of $(J_y + J_d) + p(J_y)$ matrix) where $J_y =$ number of principal components, $J_d =$ number of kernel basis.
- Bayarri et al. (Annals, 2007):
  - For ultra high dimensional data, their representation is not parsimonious enough.
  - Requires a dyadic(a power of 2) grid for data.

PCA-based Approach for High-dimensional Calibration

Outline of approach:

- **Dimension Reduction:** Summarize the model output $\mathbf{Y}$ and the observation $\mathbf{Z}$ using PCA and kernel basis.
    1. Find the first J_y eigenvectors $\mathbf{K}_y = (k_1, \ldots, k_{J_y})$ and the corresponding principal components $\mathbf{W}$ of the model output.
    2. Project $\mathbf{Z}$ on the space spanned by $\mathbf{K} = [\mathbf{K}_y \ \mathbf{K}_d]$ where $\mathbf{K}_d$ is the matrix of kernel basis with J_d knots. Denote the projected vector by $\mathbf{Z}_{red}$.
- **Emulation:** Construct an emulator for each of the principal components in $\mathbf{W}$ separately. Computation reduces to $\mathcal{O}((J_y + J_d)^3)$ instead of $\mathcal{O}(n^3p^3)$. E.g. 4,913,000 flops vs 1.5×10^{16} flops.
- **Calibration:** Estimate $\boldsymbol{\theta}$ based on the likelihood function

$$|\Sigma_{\mathbf{Z}_{red}|\mathbf{W}}|^{-\frac{1}{2}} \exp[-\frac{1}{2}\mathbf{Z}_{red}^T(\Sigma_{\mathbf{Z}_{red}|\mathbf{W}} + (\mathbf{K}^T\mathbf{K})^{-1})^{-1}\mathbf{Z}_{red}.$$

PCA-based Approach for High-dimensional Calibration

Climate parameter calibration with sea temperature:

- Climate model output: 250 UVic ensembles (1D: 13, 2D: 988, 3D: 61,051 spatial points for each).
- Observation data: World Ocean Atlas 2009.