Bayesian hierarchical models for interpreting geoscience data

Douglas Nychka
National Center for Atmospheric Research

National Science Foundation
NCAR April 2015
Outline

- Inverse problems
- Statistical and hierarchical model (Data, process, priors)
- Inverting Ice Cores
- Blending total precipitation

Main ideas:
Solve a complex data problem by
- separating what is observed from what is to be estimated.
- using Bayesian statistics as a strategy for inference.
- Simpler models for geophysical variables that use stochastic components in place on complex nonlinear physics
Data Science and Data Analytics is the extraction of knowledge from data

- Includes statistics but many other fields.
- This is a skill for everyone!

Some IMAGE/NCAR activities where we will teach
- May 18 - 21 Data assimilation summer school for statistics students
- June 22 - 26 Data analytics boot camp for high school students
- July 6-17 Climate, space climate, and couplings between
- July 22- 24 Regional climate tutorial
- July 27 - 31 Data analytics for ecology
- Aug 3-7 Frontiers in data assimilation (DART tutorial)
- Fall 2015 Data analysis course through Applied Math CU.

Hands on, substantive NCAR data sets, and fun for all.
Law Dome, East Antarctica

Ice core drill and glaciologists from the Australian Antarctic Division and Antarctic Climate and Ecosystems CRC, Law Dome, East Antarctica

What are the past concentrations for atmospheric CO$_2$ for the past 2000 years?

Eugene Wahl, Wendy Gross, David Robinson, (NOAA) and Catherine Trudinger (CSIRO).
Observations

$\text{CO}_2$ concentrations as a function depth.

Conceptually: Depth is related to the time air was trapped in the core.

\[ y(\text{depth}) = \mathcal{F}(c(\text{time})) \]

$\mathcal{F}$ developed by Trudinger et al. (2013)

Inverse problem: Map the concentrations from depths to concentrations in time.

Firn Ice Inverse problem harder for upper ice layers that have not completely consolidated.
Indirectly observing CO$_2$.

**Observations:**
$y_1, \ldots y_{100}$ concentrations at depths

**Goal:**
$c = c_1, c_2, \ldots c_{2000}$ yearly concentrations

$y = Wc + \text{measurement error}$

Rows of $W$ matrix:
A Hierarchical (geophysical) model

The goal: Estimate a continuous geophysical process: $c(t)$ atmospheric concentration at time $t$.

Data level: Observations given the process of interest

Distribution of the depth measurements given CO$_2$ time series.

Process level: A statistical/physical model for the process – does not depend on the observations but possibly on other parameters.

Distribution of the CO$_2$ time series given statistical parameters.

Priors: Probability distribution that indicates likely ranges for parameters.
Why this structure?

- Each piece is easier to formulate because an important part is fixed.
- Each level is based on conditional distributions.
  e.g. $y = Wc + \text{measurement error}$
- Straight forward to find most likely values of the process given the data. (Posterior distribution from Bayes Theorem)
More on the process level model

Annual CO$_2$ concentrations based on a linear trend, human emissions covariate, and correlated “noise”.

\[ c_t = \text{trend} + \alpha \text{Emissions}(t) + u_t \]

\[ u_t = \beta u_{t-1} + v_t \]

$v_t$ are uncorrelated normals, $u_t$ has variance $\rho$

This can be written as $c$ is $N(X\alpha, \rho \Sigma(\beta))$

$X$ is $2000 \times 3$ and $\Sigma$ $2000 \times 2000$

**Recent carbon emissions:** Emissions$(t)$
More on the correlated noise

**AR 1 process:**
\[ u_t = \beta u_{t-1} + \text{white noise} \]
- same as \( \text{corr}(u_t, u_s) = e^{-|t-s|/\delta} \)

\( \beta = .8 \)

\( \beta = .99 \)
• Maximize probabilities over $c$ and parameters to get “best” estimate.

• Sample from the posterior density to quantify uncertainty.

  e.g. in general Markov Chain Monte Carlo (MCMC)
Reconstruction of CO$_2$ from Law Dome.
Minima and locations of the posterior reconstructions help to pin down the timing and size of the decrease during this cooler period of Northern Hemisphere climate.
Onset of industrialization.

The distribution of the 400 minima for an ensemble of possible reconstructions. Suggests that the acceleration in concentrations occurs after 1790.
An industrial strength example

Parallel inference for massive distributed spatial data using low-rank models

Matthias Katzfuss (Texas A & M) Dorit Hammerling (NCAR)
Blending total precipitable water

Currently a NOAA blended product is used for NWS forecasting support

Hourly observations from three sources:
- GPS – ground network
- GOES – infrared sounder
- MIRS – microwave sounder

Goal: Create a consistent total precipitable field without boundary artifacts and with measures of uncertainty – \( P(\text{grid, time}) \)
A Hierarchical model

Data level:

\[
\begin{align*}
\text{GPS}( \text{location}, \text{time} ) &= P(\text{location}, \text{time}) + \text{error}1 \\
\text{GOES}( \text{grid1}, \text{time} ) &= P(\text{grid1}, \text{time}) + \text{error}2 \\
\text{MIRS}(\text{grid2}, \text{time}) &= P(\text{grid2}, \text{time}) + \text{error}3
\end{align*}
\]

*P is the common (true) total precipitable water field*

Process level:

- \( P(\text{location}, \text{time}) = \sum B_j(\text{location}) \eta_j(\text{time}) \), with \( B_j(\text{location}) \) basis functions
- \( \eta_j(\text{time}) \) correlated over space and time

Priors: Parameters of the space-time model and also error variances.
One time frame

(a) GPS (n=218)  
(b) SNDR (n=10586)  
(c) MIRS (n=50106)

(d) Posterior mean

(e) Posterior standard deviation

D. Nychka Hierarchical models
Summary

- Bayesian hierarchical models separate the specification of observations from the process of interest.

- Easier to handle complex data analysis problems.

- Conditional statistical models facilitate uncertainty quantification and computing.

- Flexible and useful methods where the assumptions made about the data are transparent.
Thank you!

Questions?