Hierarchical modeling

Alan Gelfand\textsuperscript{1} and Andrew O. Finley\textsuperscript{2}

\textsuperscript{1} Department of Statistical Science, Duke University, Durham, North Carolina.

\textsuperscript{2} Department of Forestry & Department of Geography, Michigan State University, Lansing, Michigan.

September 9, 2014
A changing world

- The statistical landscape has changed substantially.
- Remarkable growth in data collection, with datasets now of enormous size.
- Also a change toward examination of observational data, rather than being restrict to carefully-collected experimentally designed data.
- Also, an increased examination of complex systems using such data, requiring synthesis of multiple sources of information (empirical, theoretical, physical, etc.), necessitating the development of multi-level models.
- The general hierarchical framework
  \[ \text{data} | \text{process, parameters} | \text{process} | \text{parameters} | \text{parameters} \].

**STOCHASTIC MODELING**

- Role of the statistician. An exciting new world for modern statistics.
The range of applications runs the scientific gamut, e.g., biomedical and health sciences, economics and finance, environment and ecology, engineering and natural science, political and social science.

Again, hierarchical modeling has taken over the landscape in contemporary stochastic modeling.

Though analysis of such modeling can be attempted through non-Bayesian approaches, the Bayesian paradigm enables exact inference and proper uncertainty assessment within the given specification.

Computation: MCMC and Gibbs sampling but also sequential importance sampling, particle filters and particle learning, and now, INLA, ABC, and variational Bayes.
What are hierarchical models?

“Hierarchical model” is a very broad term that refers to a wide range of model specifications:

- Multilevel models
- Random effects models
- Random coefficient models
- Variance-component models
- Mixed effect models
- Latent variable models
- Missing data models
- State space models

Key feature: Hierarchical models are statistical models - a formal framework for analysis with a complexity of structure that matches the system being studied.
Four important notions

- Modeling data with a complex structure - large range of structures that can be handled routinely using hierarchical models, e.g. pupils nested in schools, houses nested in neighborhoods

- Modeling heterogeneity - standard regression “averages” (i.e. the general relation ship). Hierarchical models additionally model variances, e.g., variability in house prices varies from neighborhood to neighborhood

- Modeling dependent data - potentially complex dependencies in the outcome over time, over space, over context, e.g. house prices within a neighborhood tend to be similar

- Modeling contextuality - micro and macro relations, e.g., individual house prices depend on individual property characteristics and on neighborhood characteristics
Fitting hierarchical models

- Gibbs sampling and MCMC are ideally suited to fit such models.

- The overarching building block is the notion of latent variables, e.g., random effects, missing data, labels.

- These variables introduce unobservable process features which will be of interest, as well as facilitating model fitting.

- For fitting, Gibbs sampling loops become natural - update other parameters given the values of the latent variables and then update the latent variables given the values of the other parameters.
The basics

- The standard hierarchical linear model:
  First stage: \( y \mid X, \beta \sim N(X\beta, \Sigma_Y) \)
  Second stage: \( \beta \mid Z, \alpha \sim N(Z\alpha, \Sigma_{\beta}) \)
  Third stage: \( \alpha \sim N(\alpha_0, \Sigma_{\alpha}) \).

- Inverse Gamma or Wishart priors at the third stage

- Routine fitting within the Bayesian framework. Due to the conjugacy, a *vanilla* Gibbs sampler

- NonGaussian first stage (exponential family distribution, link function), a hierarchical generalized linear model.

- Conjugacy between the first and second stages is lost. Metropolis-Hastings updating would likely be used with adaptive tuning of the acceptance rates.
CIHM’s

- Early work with conditionally independent hierarchical models (CIHM’s) at Carnegie Mellon University using Laplace approximation
- Preceded Gibbs sampling and MCMC as Bayesian computation tools.
- Now enjoying a revival through the recent development of integrated nested Laplace approximation (INLA).
- The CIHM takes the basic form $\prod_i[y_i|\theta_i]\prod_i[\theta_i|\eta][\eta]$
- Exchangeable $\theta_i$ are assumed. If $\eta$ is fixed, fit separate models for each $i$.
- With unknown $\eta$, shrinkage or borrowing strength across the $i$’s
- The CIHM includes the hierarchical GLM, also natural extension to ARMA time series models
Random Effects

- Random under both Bayesian and frequentist modeling, usually normal with a variance component.
- Effects can be at different levels of the modeling but usually assumed exchangeable, in fact i.i.d.
- A typical linear version with i.i.d. effects takes the form:

\[
y_{ij} = X_{ij}^T \beta + \phi_i + \epsilon_{ij}.
\]

- At the second stage, \( \beta \) has a Gaussian prior while the \( \phi_i \) are i.i.d. \( \sim N(0, \sigma^2_\phi) \). The \( \epsilon_{ij} \) are i.i.d. \( \sim N(0, \sigma^2_\epsilon) \).
- The variance components become the third stage hyperparameters. Care with prior specifications for \( \sigma^2_\phi, \sigma^2_\epsilon \). Avoid \( IG(\epsilon, \epsilon) \); a protective recommendation is an \( IG(1, b) \) or \( IG(2, b) \).
Missing data; imputation

- In collecting information on, e.g., individuals, often vectors of data with one or more components missing.
- Don’t want to analyze only the complete data cases.
- To use the individuals with missing data, we must complete them, so-called imputation.
- Fully model-based imputation in the Bayesian setting results in latent variables and Gibbs looping. Extends the E-M algorithm to provide full posterior inference.
- A simple example: $y_i \sim N(\mu_i, \Sigma)$ (components of $\mu_i$ may have regression forms). Some components of some of the $y_i$’s are missing.
- Gibbs sampling to perform the imputation: update the parameters given values for the missing data, then update missing data given values for parameters.
Latent variables

- Again, latent variables are at the heart of most hierarchical modeling.
- Can envision beyond random effects or missing data
- Customarily, a hierarchical specification of the form
  \[ y \mid Z \mid \theta \mid \theta \]. Here, \( y \)'s are observed, \( Z \)'s are latent and the “regression” modeling is moved to the second stage
- An elementary example: suppose \( y_i \sim \text{Bernoulli}(p(X_i)) \)
- Let \( \Phi^{-1}(p(X_i)) = X_i \beta \) with a prior on \( \beta \)
- Awkward to sample \( \beta \) using the likelihood in this form so, introduce \( Z_i \sim N(X_i \beta, 1) \). Immediately,
  \[ P(y_i = 1) = \Phi(X_i \beta) = 1 - \Phi(-X_i \beta) = P(Z_i \geq 0). \]
- Now, a routine Gibbs sampler: update the \( Z \)'s given \( \beta, y \)
  (sampling from a truncated normal), update \( \beta \) given the \( Z \)'s and \( y \) (usual conjugate normal updating)
Errors in variables models

- Errors in variables models, another latent variables setting
- Usual objective is to learn about the relationship between say $y$ and $X$. Unfortunately, $X$ is not observed. Rather, we observe say $W$ instead of $X$
- $W$ may be a version of $X$, subject to measurement error, i.e., $W$ may be $X_{obs}$ while $X$ may be $X_{true}$.
- $W$ may be a variable (variables) that play the role of a surrogate for $X$
- Conceptually, we may condition in either direction. A model for $W|X$: a measurement error model; a model for $X|W$: a Berkson model
cont.

- In fact, a further errors in variables component - perhaps we observe $Z$, a surrogate for $y$.
- Altogether a hierarchical model with latent $X$’s, possibly $y$’s. For the measurement error case:

$$\Pi_i[Z_i|y_i, \gamma][y_i|X_i, \beta][W_i|X_i, \delta][X_i|\alpha]$$

while for the Berkson case we have:

$$\Pi_i[Z_i|y_i, \gamma][y_i|X_i, \beta][X_i|W_i, \delta]$$

- Usually, have some *validation* data to inform about the components of the specification.
- With a full Bayesian specification, can learn about the relationship between $y$ and $X$ without ever observing $X$ (and, possibly, without observing $y$ as well)
Mixture models

- Mixture models now widely used due to (i) their flexibility for distributional shapes and (ii) their representation of a population in terms of unidentified groups.
- Mixture models - parametric or nonparametric, incorporating discrete (finite, countable) or continuous mixing
- Basic finite mixture version:

\[ y \sim \sum_{l=1}^{L} p_l f_l(y | \theta_l) \]

- Often \( f_l \) are normal densities, whence a normal mixture.
If $L$ is specified and we observe $y_i, i = 1, 2, ..., n$, then a latent label, $L_i$, for each $y_i$, i.e., if $L_i = l$, then $y_i \sim f_l(y | \theta_l)$.

With labeling variables, hierarchical model becomes:

$$
\Pi_i [y_i | L_i, \theta] \Pi_i [L_i | \{p_l\}] [\theta] [\{p_l\}]
$$

Again, Gibbs sampling is routine. Update $\theta, \{p_l\}$ given the $L$’s and the data. To update the $L_i$’s given $\theta, \{p_l\}$ and the data, sample from an $L$-valued discrete distribution.

If $L$ is unknown with a prior specification, model dimension changes with $L$ - Reversible jump MCMC or model choice across a set of $L$’s.

Identifiability is a challenge.
Back to random effects

- Consider individual level longitudinal data with interest in growth curves
- Model individual level curves centered around a population level curve
- Population level curve to see *average* behavior of the process; individual level curves, for example, to prescribe *individual* level treatment
- If $y_{ij}$ is $j$th measurement for $i$th individual, let

  $$y_{ij} = g(X_{ij}, Z_i, \beta_i) + \epsilon_{ij}$$

  where $\epsilon_{ij} \sim N(0, \sigma_i^2)$.
- The form for $g$ depends upon the application
cont.

- At second stage, we set $\beta_i = \beta + \eta_i$ where the $\eta_i$ have mean 0 (or perhaps replace $\beta$ with a regression in the $Z_i$).

- The $\beta_i$ (or the $\eta_i$) are the random effects. They provide the individual curves with $\beta$ providing the global curve.

- Evidently, a CIHM as well. Learning with regard to any individual curve will borrow strength from the information about the other curves.
Dynamic models

- Dynamic models now a standard formulation for a wide variety of processes (also called Kalman filters, state space models and hidden Markov models)

- A first stage (or observational model), a second stage (or transition model), with third stage hyperparameters

- The first stage provides the data model while the second stage provides a latent dynamic process model

- The basic dynamic model takes the form:

  \[ y_t = g(X_t, \theta_1) + \epsilon_t, \text{ observation equation with} \]

  \[ X_t = h(X_{t-1}; \theta_2) + \eta_t, \text{ transition equation.} \]

- Time is discrete with dynamics in the mean. Bayesian model fitting using the forward filter, backward sample (ffbs) algorithm
Data fusion

Data assimilation/fusion/melding has only recently received serious attention in the statistics community.

In the spatial setting we would be fusing a dataset consisting of measurements at monitoring stations with the output of a computer model.

The former is associated with point referenced locations, is accurate but only sparsely available, often with missingness. The latter is supplied for grid cells, is uncalibrated, but is available everywhere.

Envision a latent true exposure surface informed by both the station data and the computer model data.
cont.

- The two data sources provide the first stage model. The latent true model is at the second stage, a process specification, with hyperparameters at the third stage.

- Let the $y(s_i)$ be the observed station data at $s_i$, let $X(B_j)$ be the computer model output for grid cell $B_j$ and let $Z(s)$ be the true exposure surface.

- Model the station data as a measurement error model, $y(s_i) = Z(s_i) + \epsilon(s_i)$ where the $\epsilon$ are pure errors.

- Model the computer output as a calibration specification, $X(B_j) = \int_{B_j} (a(s) + b(s)Z(s) + \delta(s))ds$ where $a(s)$ and $b(s)$ are Gaussian processes with the $\delta$'s being pure error.
Finally, we have the second stage process model,
\[ Z(s) = \mu(s) + \eta(s). \]

- \( \mu(s) \), captures the large scale structure, perhaps through covariates or a trend surface

- \( \eta(s) \) captures the small scale structure or second order dependence through a Gaussian process

Approach is called Bayesian melding. Has a stochastic integration challenge, infeasible to do for a large number of grid cells and/or with dynamics

Fully model-based alternatives, so-called downscalers, can address these limitations