Lecture 4: Principles of Bayesian inference and hierarchical modeling

Alan E. Gelfand
Duke University

Bayes Theorem

- Everyone knows Bayes' Theorem. In its simplest form

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- How did this become an inference paradigm (much less a controversial one!)?
- Suppose we move to random variables:

$$P(X \in A|Y \in B) = \frac{P(Y \in B|X \in A)}{P(Y \in B)}$$

- Then, it is a simple step to densities:

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}$$

cont.

- And finally, letting $\mathbf{Y}$ denote what you observe and replacing X with θ denoting what you don't know (didn't observe):

$$f(\theta|\mathbf{Y}) = \frac{f(\mathbf{Y}|\theta)\pi(\theta)}{f(\mathbf{Y})}$$

or

$$f(\theta|\mathbf{Y}) \propto f(\mathbf{Y}|\theta)\pi(\theta)$$

- So, we have a natural inference paradigm. **YOU INFER ABOUT WHAT YOU DON'T KNOW GIVEN WHAT YOU HAVE SEEN**
- Compare with classical inference approach using sampling distributions, $T(\mathbf{Y})$ given θ: **IMAGINE WHAT YOU MIGHT SEE GIVEN WHAT YOU DON'T KNOW**

cont.

- So, really just thinking about two ways of writing a joint distribution

$$f(\mathbf{Y}, \boldsymbol{\theta}) = f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\mathbf{Y})f(\mathbf{Y})$$

- Generative - likelihood (aleatory), prior (epistemic); Inferential - posterior and marginal for model checking
- Posterior inference - benefits of an entire distribution
- For parameters: $f(\boldsymbol{\theta}|\mathbf{Y})$
- For prediction: $f(Y_0|\mathbf{Y})$ arises from:

$$f(Y_0|\mathbf{Y}) = \int f(Y_0|\mathbf{Y}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta}$$

Priors

- What does the prior do mathematically? $f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$
- What does it mean practically?
- Proper? Improper? Subjective? Objective? Elicitation? Reference priors?
- Vague, weak, noninformative
- Improper for a location - $f(\theta) = c$; for a scale - $f(\sigma) = \frac{1}{\sigma}$ or perhaps $\frac{1}{\sigma^2}$
- Inference device vs. a model that could have generated what you have observed
- Nearly improper for a variance - $IG(\epsilon, \epsilon)$. Implications
- Play it safe. Go proper.

Basics of Bayesian Inference

- Since $\boldsymbol{\lambda}$ will not be known, a second stage (hyperprior) distribution $h(\boldsymbol{\lambda})$ will be required, so that

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y},\boldsymbol{\theta})}{p(\mathbf{y})} = \frac{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda})\, d\boldsymbol{\theta} d\boldsymbol{\lambda}} .$$

- Alternatively, we might replace $\boldsymbol{\lambda}$ in $p(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\lambda})$ by an estimate $\hat{\boldsymbol{\lambda}}$; this is called empirical Bayes analysis
- So, $p(\boldsymbol{\theta}|\mathbf{y}) \neq \pi(\boldsymbol{\theta})$
  This is referred to as Bayesian learning (the change in the posterior distribution compared with the prior).

Illustration of Bayes' Theorem

- Suppose $f(y|\theta) = N(y|\theta, \sigma^2)$, $\theta \in \Re$ and $\sigma > 0$ known
- If we take $\pi(\theta|\boldsymbol{\lambda}) = N(\theta|\mu, \tau^2)$ where $\boldsymbol{\lambda} = (\mu, \tau)'$ is fixed and known, then it is easy to show that

$$p(\theta|y) = N\left(\theta \,\middle|\, \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}y\,,\ \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)\ .$$

- Note that
  - The posterior mean $E(\theta|y)$ is a weighted average of the prior mean μ and the data value y, with weights depending on our relative uncertainty
  - the posterior precision (reciprocal of the variance) is equal to $1/\sigma^2 + 1/\tau^2$, which is the sum of the likelihood and prior precisions.

A linear model example

- Let $\mathbf{Y}$ be an $n \times 1$ data vector, X an $n \times p$ matrix of covariates, and adopt the likelihood and prior structure,

$$\mathbf{Y}|\boldsymbol{\beta} \sim N_n\left(X\boldsymbol{\beta}, \Sigma\right) \quad \text{and} \quad \boldsymbol{\beta} \sim N_p\left(A\boldsymbol{\alpha}, V\right)$$

- Then the posterior distribution of $\boldsymbol{\beta}|\mathbf{Y}$ is

$$\boldsymbol{\beta}|Y \sim N\left(D\mathbf{d}, D\right) \ , \quad \text{where}$$

$D^{-1} = X^T\Sigma^{-1}X + V^{-1}$ and $\mathbf{d} = X^T\Sigma^{-1}\mathbf{Y} + V^{-1}A\boldsymbol{\alpha}$.

- $V^{-1} = 0$ delivers a "flat" prior; if $\Sigma = \sigma^2 I_p$, we get

$$\boldsymbol{\beta}|Y \sim N\left(\hat{\boldsymbol{\beta}}\ , \sigma^2(X'X)^{-1}\right) \ , \quad \text{where}$$

$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y} \Longleftrightarrow$ usual likelihood analysis!

Bayesian updating

- Often referred to as "crossing bridges as you come to them"
- Simplifies sequential data collection
- Simplest version: Y_1, Y_2 indep given θ. So joint model is

$$p(y_2|\theta)p(y_1|\theta)\pi(\theta) \propto p(y_2|\theta)\pi(\theta|y_1),$$

  i.e., Y_1 updates $\pi(\theta)$ to $\pi(\theta|y_1)$ before Y_2 arrives
- Works for more than two updates, for updating in blocks, for dependent as well as independent data

Posterior inference

- Measures of centrality
- Credible intervals - equal tail, highest posterior density (HPD)
- Probability statements
- Bayesian hypothesis testing $\equiv$ model comparison
- Formal Bayesian model selection framework: M-closed, M-open, M-complete
- Say models M_1, M_2, M_k with prior probability of being correct, $p_1, p_2, \ldots, p_k$

cont.

- Then apply Bayes' Theorem: With data $\mathbf{Y}$,

$$P(M_j|\mathbf{Y}) = \frac{P(\mathbf{Y}|M_j)p_j}{\sum_{j=1}^{k} P(\mathbf{Y}|M_j)p_j}$$

- **BUT**: Calculating $P(\mathbf{Y}|M_j) = \int P(\mathbf{Y}|\boldsymbol{\theta}_j; M_j)\pi(\boldsymbol{\theta}_j|M_j)d\boldsymbol{\theta}_j$
- **BUT**: Where do the p_j's come from? equally likely?
- Rewarding parsimony? sparseness?

Variable Selection

- An important current problem - variable selection in a regression model, parsimony in model dimension, i.e., number of variables -
- For number of variables: Poisson(λ), i.e., probability of k variables is $\lambda^k e^{-\lambda}/k!$ with a prior on λ
- More useful: prior inclusion probability p with exchangeable Bernoulli selection. A priori, P(variable X_k selected) is p, i.e., $P(I_k = 1) = p$ with a prior on p. So, binomial model for number of variables selected.
- Seek posterior probability of $I_k = 1$. Reparametrize to $\theta_k = I_k \beta_k$.
- So, a "spike and slab" prior for θ_k using the auxiliary variable I_k; point nulls
- Large model spaces; search algorithms

Bayesian Modeling Averaging

- Bayesian model averaging (BMA) fits well with the general Bayesian model selection framework
- With a collection of models, can we choose a meaningful average one?
- Note that we can not consider model averaging with regard to parameters
- How about with regard to prediction? $f(y_0|\mathbf{Y})$?
- So, $f_{BMA}(y_0|\mathbf{Y}) = \sum_{j=1}^{k} f(y_0|\mathbf{Y}, M_j)P(M_j|\mathbf{Y})$
- Here, as above,
  $f(Y_0|\mathbf{Y}, M_j) = \int f(Y_0|\mathbf{Y}, \boldsymbol{\theta}_j; M_j)f(\boldsymbol{\theta}_j|\mathbf{Y}, M_j)d\boldsymbol{\theta}$
- Optimality for $E_{BMA}(Y_0|\mathbf{Y}) = \sum_j E(Y_0|\mathbf{Y}; M_j)P(M_j|\mathbf{Y})$ under SEL
- Note: For Y_0, best model over j is not necessarily the highest probability model

Bayes factors

- Can we avoid specifying the p_j's?
- Look at models in pairs (so, only good for a small number of models)
- Bayes Factor: Model M_0 vs M_1, $B_{01} = \frac{P(\mathbf{Y}|M_0)}{P(\mathbf{Y}|M_1)}$.
- Nice interpretation:

$$\frac{P(M_0|\mathbf{Y})}{P(M_1|\mathbf{Y})} = \frac{P(\mathbf{Y}|M_0)}{P(\mathbf{Y}|M_1)} \times \frac{p_o}{p_1}$$

- Density ordinate; proper prior
- In fact, problem with *nearly* improper prior: $Y \sim f(y|\theta)$. $M_0 \equiv H_0 : \theta = 0$, $M_1 \equiv H_A : \theta \neq 0$, i.e., $M_0 \subset M_1$. And with Uniform$(-K, K)$ prior on θ, $B_{01} = 2Kf(Y|0)$

Bayes Factors cont.

- Calculation is a challenge - marginalization over $\boldsymbol{\theta}$
- Monte Carlo calculation ideas
- Computing $c_M = f(\mathbf{Y}|M)$: Sample prior and MC integration
- Sample $g(\boldsymbol{\theta})$ with MC integration based on $\frac{f(\mathbf{Y}|\boldsymbol{\theta},M)\pi(\boldsymbol{\theta}|M)}{g(\boldsymbol{\theta})}$

cont.

- With posterior samples, $f(\mathbf{Y}|M) = \frac{f(\mathbf{Y}|\boldsymbol{\theta},M)\pi(\boldsymbol{\theta}|M)}{f(\boldsymbol{\theta}|\mathbf{Y},M)}$
- Again, with posterior samples, $\int \frac{\tau(\boldsymbol{\theta})}{f(\mathbf{Y}|\boldsymbol{\theta},M)\pi(\boldsymbol{\theta}|M)} f(\boldsymbol{\theta}|\mathbf{Y},M)d\boldsymbol{\theta} = 1/c_M$ for any density $\tau(\boldsymbol{\theta})$
- So, MC integration is a harmonic mean
- Using $\tau(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|M)$ not a good idea

Lindley paradox

- Suppose $Y_1, Y_2, ..., Y_n, i.i.d. N(\theta, 1)$ with prior $\pi(\theta) = N(0, 1)$. Consider $H_0 : \theta = 0$ vs $H_A : \theta \neq 0$. Then, $B_{01} = \frac{N(\bar{Y}|0, \frac{1}{n})}{N(\bar{Y}|0, 1+\frac{1}{n})}$.
- The usual test statistic is $Z = \sqrt{n}\bar{Y}$ in which case

$$B_{01} = \sqrt{n+1}\left(|Z|e^{-\frac{n}{n+1}Z^2/2}\right)$$

- So, regardless of $|Z|$, as $n \to \infty$, $B_{01} \to \infty$, i.e., we choose M_0.
- So, Bayes factor is "too large!", too much support for M_0 (or B_{10} too small)
- Lindley paradox

Classical hypothesis testing

- Again, point nulls
- Nested hypotheses
- Penalized likelihood - what does this mean?
- Consider the usual likelihood ratio test which, for $H_0 : \theta \in \Theta_0$ vs. $H_A : \theta \in \Theta - \Theta_0$
- Reject if

$$\lambda(\mathbf{Y}) = \frac{\sup_{\theta \in \Theta_0} L(\theta; \mathbf{Y})}{\sup_{\theta \in \Theta} L(\theta; \mathbf{Y})} < c$$

- More generally, for $M_1 \subset M_2$
- Under weak conditions, under M_1, $-2\log\lambda \approx \chi^2_{p_2 - p_1}$ where p_2 is the dimension of model M_2 and p_1 is the dimension of model M_1

cont.

- So, $P(\lambda < c|M_1) = P(-2\log\lambda > c') \approx P(\chi^2_{p_2-p_1} > c') > 0$
- So, $P(\texttt{reject}M_1|M_1\texttt{true}) \nrightarrow 1$ as $n \to \infty$
- $\lambda \Leftrightarrow \log\lambda$ is too large; too much support for M_2
- Penalty function; "penalized likelihood"
- Form: A penalty function of only model dimension, p_j, e.g., kp_j (AIC)
- A function of sample size, n and model dimension, p_j, e.g., $\log n(p_j)$ (BIC)

cont.

- BIC results in $-2\log\lambda_n - (p_2 - p_1)\log n$
- Then, c smaller, $c' + \log n(p_2 - p_1)$ and now $P(\text{rej}M_1 | M_1 \text{true}) \to 0$
- Moreover, we saw that the Bayes factor was too large and $\log\lambda$ is too small. A relationship between them?

$$\log BF_{01} = -\log\lambda_n + \log n \frac{p_2 - p_1}{2} + O(1)$$

- Really only works for "fixed" effects so not so interesting in today's hierarchical modeling landscape

The problems with P-values

- Point null can't be true; spike and slab priors
- Badly distorted with point nulls
- Why does an unobserved region criticize the null hypothesis?
- **NOT** $P(H_0\texttt{true}|\texttt{Data})$
- Violates the likelihood principle

To see the distortion

- Consider $Y_1, Y_2, ..., Y_n \sim N(\theta, \sigma^2)$ with $\theta \sim N(\mu, \tau^2)$
- Consider again $H_0 : \theta = 0$ vs. $H_A : \theta \neq 0$
- Fix variances and let $p_0 = p_1 = .5$
- Run standard "Z" test (PHoD is $P(H_0|\texttt{data})$)

n	=	10	=	50	=	100	
Z	Pval	BF	PHoD	BF	PHoD	BF	PHoD
1.645	.10	.89	.47	1.86	.65	2.57	.72
1.96	.05	.59	.37	1.08	.52	1.50	.60

Model adequacy and model comparison; Big picture

- Since the Bayesian framework is so *liberating*, we often explore many models
- Again, the old bromide, "All models are wrong but some models are useful" applies
- But necessitates assessing adequacy of models and comparison of models
- With tools to fit Bayesian models, we face the issue of "overfitting" (more often than underfitting) - specifying models that are richer than the data is capable of explaining
- Model adequacy is an *absolute criterion*. Does the model meet certain performance standards?
- Various criteria available but often difficult to calibrate

Model comparison; big picture

- Regardless, many models may be adequate so we need comparison criteria. These criteria are *relative* to order models
- Model comparison can be developed formally (as we illustrated above). But practically, model development is evolutionary which can contaminate probabilistic assessment of model selection
- Never agreement on a "best" model selection criterion
- Depends on utility for a model
- Further concern: reducing a model to a single number
- Parameter space vs. predictive space; posterior distribution for a parameter, posterior predictive distribution for an "observation"
- Again, issues with hierarchical models and parameter space
- So, perhaps consider model checking in predictive space
- Fitting or training dataset. Hold out or validation dataset (cross-validation), k-fold cross validation

For model adequacy/checking

- The formal Bayesian measure is is the marginal density ordinate of the data, $f(\mathbf{Y}_{obs})$
- Nowadays not of interest; impossible to calibrate especially when dimension of $\mathbf{Y}$ is large
- Single point deletion; conditional probability ordinate (CPO) - For $Y_{i,obs}$, calculate $f(Y_{i,obs}|\mathbf{Y}_{(-i)})$. Easy MCMC for $f(\boldsymbol{\theta}|\mathbf{Y}_{(-i)})$
- Only for simple models, small datasets, discrepant observations
- Posterior predictive checks vs. Prior predictive checks
- If you want to generate samples from the model to compare with the observed data in some fashion,
  (i) posterior predictive approach says generate $\mathbf{Y}_{rep}$ from $f(\mathbf{Y}_{rep}|\text{model}, \mathbf{Y}_{obs}) = \int f(\mathbf{Y}_{rep}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{Y}_{obs})d\boldsymbol{\theta}$.
  (ii) prior predictive approach says generate $\mathbf{Y}_{rep}$ from $f(\mathbf{Y}_{rep}|\text{model}) = \int f(\mathbf{Y}_{rep}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$.

cont.

- Introduce a discrepancy function $D(\mathbf{Y}_{rep}, \mathbf{Y}_{obs})$ and then consider its posterior distribution or its prior distribution.
- With a hierarchical model can introduce second stage (latent) variables to consider first stage or second stage discrepancies
- The debate: Generate under actual model or with a distribution for $\boldsymbol{\theta}$ that you are more comfortable with. Use the data twice, less critical of the model. However, is the goal model rejection or finding where model does not fit the data well?
- Connection to Monte Carlo tests, generate samples under the model, compute a function of the observed data, $T(\mathbf{Y}_{obs})$ and compare with a set $T(\mathbf{Y}_{rep,b}), b = 1, 2, B$ from model
- Again, in-sample vs. out-of-sample

Empirical coverage

- For a variable of interest, say $T(\mathbf{Y})$, obtain posterior predictive distribution, $f(T(\mathbf{Y})|\mathbf{Y}_{obs})$.
- Obtain say a 90% predictive interval.
- Compare observation with interval.
- Do this for many T's. Obtain empirical coverage
- Interpretation

Model comparison

- Perhaps, AIC, BIC and variations as above
- In parameter space, posterior log likelihood, $\pi(L(\boldsymbol{\theta}; \mathbf{Y}_{obs})|\mathbf{Y}_{obs})$
- Deviance information criterion (DIC)
- A generalization of AIC to hierarchical models based on the posterior distribution of the deviance statistic,

$$D(\boldsymbol{\theta}) = -2logf(\mathbf{y}|\boldsymbol{\theta}) + 2logh(\mathbf{y});,$$

  where $f(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood and $h(\mathbf{y})$ is any standardizing function of the data alone
- Summarize the fit of a model by the posterior expectation of the deviance, $\bar{D} = E_{\boldsymbol{\theta}|y}(D)$
- Summarize the complexity of a model by the effective number of parameters, $p_D = E_{\boldsymbol{\theta}|y}(D) - D(E_{\boldsymbol{\theta}|y}(\boldsymbol{\theta})) = \bar{D} - D(\bar{\boldsymbol{\theta}})$

cont.

- The *Deviance Information Criterion* (DIC) is then

$$DIC = \bar{D} + p_D = 2\bar{D} - D(\bar{\boldsymbol{\theta}}) \ ,$$

  with smaller values indicating preferred models.
- Both $E_{\theta|y}(D)$ and $D(E_{\theta|y}(\boldsymbol{\theta}))$, are easily estimated from MCMC samples, in fact, automatic in WinBUGS.
- While p_D has a scale (effective model size), DIC does not, so only differences in DIC across models matter.
- DIC can be sensitive to parametrization and "focus" (what is the likelihood?)
- Comparative explanatory performance. DIC tends to select "bigger" models

Criteria in predictive space

- Posterior predictive loss criterion
- When looking at predictive performance, again we need to penalize for model complexity
- We need a loss function that rewards goodness of fit to the observed data as well as predictive performance for new or replicate data
- We introduce a *balanced* loss function
- For the squared error loss case, we obtain $D_k = \frac{k}{k+1} G + P$ where $G = \sum_l (E(Y_{l,new}|\mathbf{y}) - y_{l,obs})^2$ and $P = \sum_l \mathrm{Var}(Y_{l,new}|\mathbf{y})$
- Posterior predictive mean and variance readily computed
- G is a goodness of fit term, P is a penalty
- Comparative predictive performance. Small values of D_k are preferred

cont.

- Predictive MSE - $\sum_{\ell=1}^{L}(E(Y_\ell|\mathbf{Y}_{obs}) - Y_{\ell,obs})^2$, predictive MAE (replace square with absolute value, length of predictive intervals - again, out of sample
- Aggregate over hold out observations
- Continuous rank probability score (CRPS) - compare an entire (predictive) distribution with an observation
- The more concentrated the predictive distribution is around the held out observation the better
- For any continuous distribution/cdf F, CRPS $= \int(F(y) - 1(y > y_{obs}))^2 dy$ (RPS, sum, discrete distribution)
- Hard to compute but a convenient MC integration using posterior predictive samples from
- Under posterior predictive distribution for Y_ℓ, CRPS$= E|Y_\ell - Y_{\ell,obs}| - \frac{1}{2}E|Y_\ell - Y_m|$
- Aggregate over hold out observations

Hierarchical Modeling

Alan E. Gelfand
Duke University

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 $[data|process, parameters][process|parameters][parameters]$.
- **STOCHASTIC MODELING**
- Role of the statistician. An exciting new world for modern statistics

cont.

- 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 nonBayesian 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 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 relationship). 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:

$$\text{First stage} : \mathbf{Y}|\mathbf{X}, \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \Sigma_{\mathbf{Y}})$$

$$\text{Second stage} : \boldsymbol{\beta}|\mathbf{Z}, \boldsymbol{\alpha} \sim N(\mathbf{Z}\boldsymbol{\alpha}, \Sigma_{\boldsymbol{\beta}})$$

$$\text{Third stage} : \boldsymbol{\alpha} \sim N(\boldsymbol{\alpha}_0, \Sigma_{\boldsymbol{\alpha}}).$$

- 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 $\Pi_i[\mathbf{Y}_i|\boldsymbol{\theta}_i]\Pi_i[\boldsymbol{\theta}_i|\boldsymbol{\eta}][\boldsymbol{\eta}]$
- Exchangeable $\boldsymbol{\theta}_i$ are assumed. If $\boldsymbol{\eta}$ is fixed, fit separate models for each i.
- With unknown $\boldsymbol{\eta}$, 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:

$$Y_{ij} = X_{ij}^T \boldsymbol{\beta} + \phi_i + \epsilon_{ij}.$$

- At the second stage, $\boldsymbol{\beta}$ has a Gaussian prior while the ϕ_i are i.i.d. $\sim N(0, \sigma_\phi^2)$. The ϵ_{ij} are i.i.d. $\sim N(0, \sigma_\epsilon^2)$.
- The variance components become the third stage hyperparameters. Care with prior specifications for σ_ϕ^2, σ_ϵ^2. 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: $\mathbf{Y}_i \sim N(\boldsymbol{\mu}_i, \Sigma)$ (components of $\boldsymbol{\mu}_i$ may have regression forms). Some components of some of the $\mathbf{Y}_i$'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 $[\mathbf{Y}|\mathbf{Z}][\mathbf{Z}|\boldsymbol{\theta}][\boldsymbol{\theta}]$. Here, Y's are observed, Z's are latent and the "regression" modeling is moved to the second stage
- An elementary example: suppose $Y_i \sim \text{Bernoulli}(p(\mathbf{X}_i))$
- Let $\Phi^{-1}(p(\mathbf{X}_i)) = \mathbf{X}_i\boldsymbol{\beta}$ with a prior on $\boldsymbol{\beta}$
- Awkward to sample $\boldsymbol{\beta}$ using the likelihood in this form so, introduce $Z_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, 1)$. Immediately, $P(Y_i = 1) = \Phi(\mathbf{X}_i\boldsymbol{\beta}) = 1 - \Phi(-\mathbf{X}_i\boldsymbol{\beta}) = P(Z_i \geq 0)$.
- Now, a routine Gibbs sampler: update the Z's given $\boldsymbol{\beta}, \mathbf{y}$ (sampling from a truncated normal), update $\boldsymbol{\beta}$ given the Z's and $\mathbf{y}$ (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:

$$\mathbf{Y} \sim \sum_{l=1}^{L} p_l f_l(\mathbf{Y}|\boldsymbol{\theta}_l)$$

- Often f_l are normal densities, whence a normal mixture.

cont.

- If L is specified and we observe $\mathbf{Y}_i, i = 1, 2, ..., n$, then a latent *label*, L_i, for each $\mathbf{Y}_i$, i.e., if $L_i = l$, then $\mathbf{Y}_i \sim f_l(\mathbf{Y}|\boldsymbol{\theta}_l)$
- With labeling variables, hierarchical model becomes:

$$\Pi_i[\mathbf{Y}_i|L_i, \boldsymbol{\theta}][\Pi_i[L_i|\{p_l\}][\boldsymbol{\theta}][\{p_l\}]$$

- Again, Gibbs sampling is routine. Update $\boldsymbol{\theta}, \{p_l\}$ given the L's and the data. To update the L_i's given $\boldsymbol{\theta}, \{p_l\}$ 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 jth measurement for ith individual, let

$$Y_{ij} = g(\mathbf{X}_{ij}, \mathbf{Z}_i, \boldsymbol{\beta}_i) + \epsilon_{ij}$$

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

cont.

- At second stage, we set $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \eta_i$ where the η_i have mean 0 (or perhaps replace $\boldsymbol{\beta}$ with a regression in the $\mathbf{Z}_i$).
- The $\boldsymbol{\beta}_i$ (or the η_i) are the random effects. They provide the individual curves with $\boldsymbol{\beta}$ 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:

$$\mathbf{Y}_t = g(\mathbf{X}_t, \boldsymbol{\theta}_1) + \boldsymbol{\epsilon}_t, \text{observation equation with}$$

$$\mathbf{X}_t = h(\mathbf{X}_{t-1}; \boldsymbol{\theta}_2) + \boldsymbol{\eta}_t, \text{transition equation.}$$

- 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 ϵ 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 δ's being pure error.

cont.

- 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

An important remark

- Hierarchical models usually introduce latent variables in addition to parameters
- Recalling our general hierarchical specification, these will often be variables associated with the process, e.g., true environmental exposures
- However, often they are introduced either to facilitate computation or explanation
- This raises the opportunity to introduce them at the first stage or at the second stage.
- At the first stage, they imply that the observations are a function of them; at the second stage, they imply that they are explaining the mean of the function

cont.

- The simplest example: Suppose the data, Y_i's are Bernoulli trials and suppose the latent Z_i's are normal variables.
- In the first case, we set say $Y_i = g(Z_i) = 1(Z_i \geq 0)$
- In the second case, we set $E(Y_i) = P(Y_i = 1) = P(Z_i \geq 0)$, a probit model
- Another example is to handle positive random variable using the Tobit, e.g., $Y_i = \max(0, Z_i)$ vs. $Z_i^* = \max(0, Z_i)$ and $E(Y_i) = Z_i^*$
- Other possibilities include Poisson, ordinal categorical data, and compositional data

Caveats with hierarchical modeling

- An extremely powerful modeling tool **BUT**
- Identifiability of parameters; parameters vs. prediction
- Multiple modes, convergence of MCMC
- Models grow very big; can data support the model?
- random effects vs. fixed effects
- **COMMENT**: Hierarchical models vs. Graphical models