Bayes: All uncertainty is described using probability.

Let $\mathbf{w}$ be the data and θ be any unknown quantities.

- *Likelihood*.
  The probability model $\pi(\mathbf{w}|\theta)$ has θ fixed and $\mathbf{w}$ varying.
  The likelihood $L(\theta; \mathbf{w})$ is $\pi(\mathbf{w}|\theta)$ with $\mathbf{w}$ fixed and θ varying.

- *Prior Distribution*. $\pi(\theta)$, information about θ external to $\mathbf{w}$.

- *Bayes's Theorem*. $\pi(\theta|\mathbf{w}) = \pi(\mathbf{w}|\theta)\pi(\theta)/\pi(\mathbf{w})$,
  where $\pi(\mathbf{w}) = \int \pi(\mathbf{w}|\theta)\pi(\theta)d\theta$.

- *Posterior Distribution*. Apply Bayes's theorem to update the information in the prior distribution: $\pi(\theta|\mathbf{w}) \propto L(\theta; \mathbf{w})\pi(\theta)$.

- *Fundamental Principle*. *All* statements about θ depend on the data $\mathbf{w}$ *only* through the posterior distribution.

Bayes for mixed linear models

The mixed linear model in standard form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \text{ and } \mathbf{u} \sim N(\mathbf{0}, \mathbf{G}(\boldsymbol{\phi}_G)),\ \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R}(\boldsymbol{\phi}_R))$$

Bayes	**Mixed Linear Model**
Data $\mathbf{w}$	Outcome $\mathbf{y}$
Unknowns θ	Mean structure: $\boldsymbol{\beta}$, $\mathbf{u}$ Variance structure: $\boldsymbol{\phi}_G$, $\boldsymbol{\phi}_R$
Likelihood	$\pi(\mathbf{y}\|\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\phi}_R)$ is $N(\ \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}\ ,\ \mathbf{R}(\boldsymbol{\phi}_R)\ )$
Prior	$\boldsymbol{\beta} \sim$ usually multivariate normal, usually flat $\mathbf{u} \sim N(0, \mathbf{G}(\boldsymbol{\phi}_G))$ $\boldsymbol{\phi}_G, \boldsymbol{\phi}_R$: to follow

Some comments on Bayes as applied to MLMs

Bayes treats $\boldsymbol{\beta}$ and $\mathbf{u}$ the same, unlike conventional analyses.

Because of $\mathbf{u}$, the likelihood and prior cannot be cleanly distinguished.

- This causes confusion and ad-hockery in conventional analyses.
- BUT it causes no important difficulties in Bayesian analyses.
- Some people call $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ part of the prior; some call it part of the model.
- But once the model and prior are specified, this label has no effect.

Bayesian inference for MLMs

Fundamental Principle. *All* statements about θ depend on the data **w** *only* through the posterior distribution.

Estimates and intervals:

- Estimates: *Marginal* posterior median or mean (for $\boldsymbol{\beta}$, **u** only).
- Intervals: One-sided; Two-sided: most commonly equal-tailed.

Testing effects: Bayes does not handle this easily; common expedients:

- 95% posterior interval or region for the effect's coefficients.
- A beauty measure, most commonly DIC with vs. without the effect.
- The Bayes factor:

$$\frac{\int \pi(\mathbf{w}|\theta, \text{effect included})\pi(\theta)d\theta}{\int \pi(\mathbf{w}|\theta, \text{effect excluded})\pi(\theta)d\theta}$$

The latter two cheat on the Fundamental Principle, it seems to me.

Prior distributions for ϕ_R and ϕ_G

This question is not at all settled; individuals have preferences, but no consensus exists.

Broad generalizations:

- Unknowns in error covariance **R**: ϕ_R's posterior is not too sensitive to the prior.
- Unknowns in random-effect covariance **G**: ϕ_G's posterior *is* sensitive to the prior. (*Maybe not* for generalized linear mixed models.)
- Operating characteristics of commonly-used priors are not well understood.

I will show you some commonly used priors; *caveat emptor*.

Popular prior distributions for variances

- Gamma prior on the precision (1/variance)
  - Conjugate.
  - Γ(small, small) has variance = 1/small. Γ(0.001, 0.001) has mean 1 and 95th percentile 3×10^{-20} – it's a weird distribution.
- $U(\epsilon, \text{upper})$ on the standard deviation (Gelman 2006)
  - Estimates and intervals are sensitive to the choice of "upper".
- Spike and slab prior on the precision τ
  - $\pi(\log \tau) = p\ N(\text{big}, \text{small}) + (1 - p)\ N(0, \text{moderate})$
  - Through p, the data inform about whether variance = zero.
- Half-t on the standard deviation (Gelman 2006)
  - Has a conjugate form, if written with an auxiliary variable.
  - Allows informative priors.

Prior distributions for covariance matrices

- Wishart prior on the precision matrix (covariance^{-1})
  - Conjugate; generalizes the gamma on precisions.
  - Marginally, precisions $\sim$ gamma, correlations $\sim$ beta.
  - Prior correlations between parameters are ...

- Covariance = diag(**S**) **B** diag(**S**), for **B** a correlation matrix and **S** and **B** independent (Barnard et al 2000).
  - **S**: Multivariate normal on log **S**; priors on preceding slide.
  - **B**: Constrained Wishart or flat prior on legal matrices.

- Other proposals use Givens angles (Daniels & Kass 1999) or the Cholesky decomposition (e.g., Chen & Dunson 2003).

- Jeffreys and Berger-Bernardo priors do poorly in simulations. (Daniels & Kass 1999, 2001).

Computing for Bayes, before 1990

Explicit expressions for $\pi(\boldsymbol{\beta}, \mathbf{u}|\mathbf{y})$, $\pi(\boldsymbol{\beta}|\mathbf{y})$, $\pi(\mathbf{u}|\mathbf{y})$, or $\pi(\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}|\mathbf{y})$ were only tractable in simple or unrealistic special cases.

It's easy to derive $\pi(\boldsymbol{\phi}_G, \boldsymbol{\phi}_R|\mathbf{y})$ – it's the restricted likelihood.

If $\mathbf{R} = \sigma_e^2\mathbf{I}$ and $1/\sigma_e^2$ has a gamma prior, it's easy to derive the marginal posterior of the unknowns in $\mathbf{G}/\sigma_e^2$.

These marginal posteriors have nonstandard forms so it's hard to compute posterior regions unless $\boldsymbol{\phi}_G$ has one dimension.

Thus, these results have little practical utility.

Computing for Bayes, after Gelfand & Smith (1990)

WARNING: I am not close to being well-informed about this.

There are two broad approaches:

Simulating draws from posteriors

- Markov chain Monte Carlo (MCMC), by far the most popular.
- Particle filters, about which I know almost nothing, are commonly used for certain problems.

Approximations

- INLA (iterated nested Laplace approximations).
- VB (variational Bayes) aka ABC (approx Bayesian computation).

The hot problem now is scaling up to really big models/datasets.

Markov chain Monte Carlo: a tiny bit of theory

For a MLM, $\theta = (\boldsymbol{\beta}, \mathbf{u}, \phi)$ is the vector of unknowns.

MCMC produces a serially-correlated sequence of draws
from $\pi(\theta|\mathbf{y})$, $\theta_{(b)}, b = 1, \ldots, B$.

It's a Markov chain because $f(\theta_{(b)}|\theta_{(i)}, i < b) = f(\theta_{(b)}|\theta_{(b-1)})$

Markov chain draws from the posterior are usually easier than iid draws.

Extremely useful:

Given MCMC draws $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(B)}$, from $\pi(\theta|\mathbf{y})$, for *any* function g,
$g(\theta_{(1)}), g(\theta_{(2)}), \ldots, g(\theta_{(B)})$ are MCMC draws from $\pi(g(\theta)|\mathbf{y})$.

Markov chain Monte Carlo: Practicalities 1

How do you derive and compute an MCMC?

- WinBUGS, JAGS (rjags), stan, NIMBLE, others (?):
  You write the model/prior in the package's syntax and it does the rest.

- SAS and lme4 (in R) have limited (?) MCMC capability.

- Otherwise, you have to derive the chain and code it yourself.

Markov chain Monte Carlo: Practicalities 2

Have you run the chain long enough? Much confusion exists because

Convergence of an MCMC is used to mean two distinct things:

- Is the chain drawing from the stationary distribution $\pi(\theta|\mathbf{y})$?
  - Gelman & Rubin diagnostic addresses this (if anything does).
- Is the estimate of $E(g(\theta)|\mathbf{y})$ precise enough?
  - Jones, Geyer and their students: theory and methods for computing MCMC standard errors (R package mcmcse).
  - Results in this theory have the form:

    $\{$Sufficient conditions on a chain$\}$ imply

    $\{$A function of the chain's draws has an ergodic property $+$

    asymptotic distribution of the MCMC estimator$\}$

Approximations

WARNING: I know little about either of these.

Quandary: MCMC imprecision (it's slow) vs. approximation error.

Iterated nested Laplace approximations (INLA)

- Arises from Rue & Held syntax for mixed linear models.
- Software: R-INLA package and associated infrastructure.

Variational Bayes (aka ABC) – Intro in Blei et al *JASA* 2017, p. 859

- The approximation:
  - Partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$
  - Approximate $\pi(\boldsymbol{\theta}|\mathbf{y}) \approx q_1(\boldsymbol{\theta}_1) \times \cdots \times q_K(\boldsymbol{\theta}_K)$.
  - Yes, this forces $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ to be independent *a posteriori*.
- This lends itself to parallel computing (Wand 2017 *JASA*, p. 137).
- Packages exist; see Blei or Wand.

Bayes vs. conventional: Conventional

Advantages:

- Maximizing is simpler than integrating; it's often much faster than Bayesian analysis using MCMC.
- All major software packages offer flexible analyses.
- Compared to ML, it avoids known biases.

Disadvantages:

- It's effectively impossible to account for variation in $\hat{\phi}_G$ and $\hat{\phi}_R$.
- Even asymptotic results exist only for limited classes of models.
- Standard theory cannot handle zero estimates for variances in $\mathbf{G}$, which occur routinely.

Bayes vs. conventional: Bayes

Advantages:

- Naturally accounts for uncertainty about ϕ_G and ϕ_R.
- Does not give zero point estimates for variances in **G** or **R**, and naturally provides intervals of plausible values.

Disadvantages:

- Posterior dependence on the prior is not well understood.
- Existing software is slower and a lot harder to use.
- MCMC routines for mixed linear models are less well understood than routines for conventional analyses.
- As with any Bayesian analysis, there's no guarantee that intervals will have nominal coverage in a frequentist sense.

Example: Molecular structure of a virus

Peterson et al. (2001) estimated counts of each of six proteins in the prohead of the bacteriophage virus $\phi 29$.

For y_i the i^{th} measurement of the number of gp8 molecules

$$
\begin{aligned}
y_i &= \mu + \text{parent}_{j(i)} + \text{batch}_{k(i)} + \text{gel}_{l(i)} + \text{run}_{m(i)} + \epsilon_i \\
\text{parent}_{j(i)} &\overset{iid}{\sim} N(0, \sigma_p^2), \quad j = 1, \ldots, 4 \\
\text{batch}_{k(i)} &\overset{iid}{\sim} N(0, \sigma_b^2), \quad k = 1, \ldots, 9 \\
\text{gel}_{l(i)} &\overset{iid}{\sim} N(0, \sigma_g^2), \quad l = 1, \ldots, 11 \\
\text{run}_{m(i)} &\overset{iid}{\sim} N(0, \sigma_r^2), \quad m = 1, \ldots, 7 \\
\epsilon_i &\overset{iid}{\sim} N(0, \sigma_e^2), \quad i = 1, \ldots, 98
\end{aligned}
$$

Deterministic functions $j(i)$, $k(i)$, $l(i)$, and $m(i)$ map i to parent, batch, gel, and run indices.

Estimates and Intervals for Copy Numbers

Molecule	N Measurements	Method	Copies	95% Interval
gp7	31	REML	25.8	(19.0,32.6)
		Bayesian	26.9	(13.6,40.9)
gp8	98	REML	231.9	(200.5,263.3)
		Bayesian	232.4	(199.8,266.5)
gp8.5	46	REML	53.5	(37.7,69.3)
		Bayesian	52.6	(29.1,74.4)
gp9	40	REML	9.1	(7.9,10.3)
		Bayesian	9.1	(3.3,15.2)
gp11	40	REML	11.2	(10.3,12.1)
		Bayesian	11.2	(5.6,16.6)
gp12	40	REML	58.6	(51.9,65.3)
		Bayesian	58.2	(44.7,71.2)

Results for gp8, which had the *most* informative design

	σ_e	σ_p	σ_b	σ_g	σ_r
REML estimate	20.3	23.8	20.6	0	15.1
Posterior mean	20.6	15.5	26.4	5.3	11.9
Posterior median	20.6	9.7	25.5	3.3	10.4
Bayes interval	(17.7,24.1)	(0.6,63)	(4.0,53)	(0.5,20)	(0.7,34)

- Maximizing the restricted likelihood gives $\hat{\sigma}_g = 0$, which is wrong.
- The posterior mean and median often differ noticeably.
- The 95% intervals are wide even with a modestly informative prior.
  $\Rightarrow$ wider intervals for the estimated count.

. . . and here are some worse results

	σ_e	σ_p	σ_b	σ_g	σ_r
gp8.5					
REML estimate	4.3	9.1	7.6	3.6	0
Posterior mean	4.5	8.3	9.7	3.6	2.3
Posterior median	4.4	3.2	8.1	3.1	1.6
Bayes interval	(3.6,5.6)	(0.5,49)	(2.6,27)	(0.8,9.1)	(0.5,9.3)
gp9					
REML estimate	0.96	—	0.85	0	0
Posterior mean	1.2	—	2.7	1.3	1.4
Posterior median	1.2	—	1.4	0.95	1.1
Bayes interval	(0.99,1.6)	—	(0.5,12)	(0.4,4.1)	(0.4,4.9)