Stat 534: formulae referenced in lecture, week 12: Hierarchical modeling

Vocabulary:

- Lots of related terms
  - Multilevel model
  - Hierarchical model
  - Mixed model
- Often used interchangeably
  - But many make distinctions
  - between specific subtypes of models
- Confusing!
- Common feature: distribution / likelihood of the data includes an integral or sum over an unknown quantity

Examples:

- Mixed model for subsampled data
  - Plots assigned to treatments
  - 5 soil cores per plot
  - plot = eu
  - soil core(plot) = ou

$$\begin{aligned} Y_{ijk} &= \mu_i + \tau_{ij} + \varepsilon_{ijk} \\ \tau_{ij} &\sim N(0, \sigma^2_{plot}) \\ \varepsilon_{ijk} &\sim N(0, \sigma^2_{core}) \end{aligned}$$

  - Alternatively:

$$\begin{aligned} \mu_{ij} &\sim N(\mu_i, \sigma^2_{plot}) \\ Y_{ijk} \mid \mu_{ij} &\sim N(\mu_{ij}, \sigma^2_{core}) \end{aligned}$$

$$Y_{ijk} = \int_{\sigma^2_{plot}} f(Y_{ijk} \mid \mu_{ij})\, f(\mu_{ij}) d\,\mu_{ij}$$

- Tag loss problem

- Two groups of fish:
  - those with tag
  - those without tag
- Introduce a random variable $R = \#$ fish who retained tag
- Know $T = \#$ fish tagged last year
- $T - R = \#$ fish with only a fin clip
  - define $c = (t_{12},\ t_1,\ t_2, R - (t_{12} + t_1 + t_2))$
  - If R known, these are a multinomial sample of the R fish
  - and $T - R$ known
  - f_1 and f_2 are two independent binomial samples of the $T - R$ fish who lost a tag

$$
\begin{aligned}
R &\sim \text{Bin}(T, r) \\
c &\sim \text{Multinom}\left(R, \left(p^2, p(1-p), p(1-p), (1-p)^2\right)\right) \\
f_1 &\sim \text{Bin}(T - R, p) \\
f_2 &\sim \text{Bin}(T - R, p)
\end{aligned}
$$

- and (importantly) c, f_1, and f_2 are conditionally independent given the value of R

$$f(c, f_1, f_2) \sim \sum_R f(c \mid R) f(f_1 \mid R) f(f_2 \mid R) f(R)$$

Concepts

- Latent variable:
  - Find a random variable that, if you knew it, would simplify the problem
    - R for tag loss problem
    - μ_{ij} for the subsampling problem
    - N_t in grizzly bear problem
- to construct a model,
  - Write out the model for the latent variable
  - Write out the model for the observations given the latent variable

- fitting the model to data
  - need a likelihood for the data, $f(Y))$
    - the conditional distribution $f(Y|\text{latent})$ is not enough
  - Need to deal with the integral or the sum
  - When all random variables have normal distributions, added or substracted
    - Y has a multivariate normal distribution
    - Usually not independent
    - but covariance matrix is a function of the various variances
  - In general, need to numerically approximate that integral or evaluate that sum
  - Could use likelihood
  - Almost all applications shift to a Bayesian paradigm

Bayes in principle

- How it differs from frequentist inference
- Frequentist (e.g. likelihood)
  - parameters are fixed but known constants
  - Data are random variables
  - Inference by maximizing the likelihood function
- Bayesian inference
  - parameters are random variables
  - Inference is conditional on the observed data
  - so data are fixed values
  - need to identify prior distributions
    - what you want to say about the parameters before seeing the data
  - Inference by averaging the likelihood function w.r.t. the prior
  - Result is the posterior distribution of the parameters

Bayes in action

- Bayes rule - a mathematical statement about probabilities

$$f(X \mid Y) = \frac{f_Y(Y|X)f_X(X)}{\int_X f_Y(Y|X)f_X(X)dX}$$

- Allows you to go from one conditional distribution to the conditional distribution "the other way"
- A mathematical fact
- disagreements are about whether this is relevant to data analysis
- An example of a Bayes rule computation
  - Screening for rare diseases / terrorist activity
  - Prostate cancer and PSA tests
  - If you get a positive PSA test result, how likely are you to have prostate cancer?
  - Test is reasonably good at detecting cancer
    * sensitivity = 86%
    * P[positive test | have cancer] = 0.86
    * specificity = 33%
    * P[negative test | no cancer] = 0.33
  - need P[have cancer | positive test] to answer the Q
  - Answer about 2%, depending on prevalence of cancer!
  - How gotten:
    * US white men, 40-59 yr old, prevalence = 1.6%
    * assume 4,000,000 people

have cancer	PSA +	PSA -	# people
Yes			64,000
No			3,936,000
Total			4,000,000

- Use sensitivity and specificity to fill in test results - going across the rows

have cancer	PSA +	PSA -	# people
Yes	55,040		64,000
No	2,637,120		3,936,000
Total	2,692,160		4,000,000

- Compute P[cancer | + test] = 55,040 / 2,692,160 = 2.04%

- Applying Bayes rule directly

$$\begin{aligned} P[\text{cancer} \mid + \text{test}] &= \frac{P[+ \text{ test} \mid \text{cancer}] \times P[\text{cancer}]}{P[+ \text{ test} \mid \text{cancer}] \times P[\text{cancer}] + P[+ \text{ test} \mid \text{no cancer}] \times P[\text{no cancer}]} \\ &= \frac{0.86 \times 0.016}{0.86 \times 0.016 + (1 - 0.33) \times (1 - 0.016)} \\ &= \frac{0.0138}{0.0138 + 0.6593} \\ &= 0.0204 \end{aligned}$$

Bayes in data analysis

- θ, the parameters $\Rightarrow X$
- the data $\Rightarrow$ Y
- Bayes rule gives you a distribution for the parameters given the data, $f(X \mid Y)$
- Connection: likelihood = $f(Y \mid X)$
- But Bayes not "free"
- Need to specify $f(X)$: "the prior"
- Aside:
  - Fisher developed fiducial inference in the 1930's, 40's
  - attempted give f(parameter | data) without requiring a prior
  - Savage (1961): "enjoy the Bayesian omelet without breaking the Bayesian eggs"
  - almost never used today
- Types of Bayesians

  - Differ in their view of the prior
  - subjective Bayes
    * Heyday: 1950's, 1960's
    * The prior is **your** belief
    * People may/will have different priors
    * And reach different conclusions from the same data
    * led to vicious arguments
  - objective Bayes
    * ad hoc but useful collection of methods for learning from data
    * emphasizes weakly informative priors
    * now the most commonly used approach

What you get by being Bayesian

- More intuitive "intervals"
  - Credible interval gives P[parameter in a specified interval]
  - Posterior predictive interval gives P[new observation in a specified interval]
- Useful probabilities, e.g.,
  - P[$N > 100$]
  - P[yield increase > cost of treatment]
- Model fit information
  - Information criteria:
    * WAIC: Widely Applicable Information Criterion
    * replacing older DIC: Deviance Information Criterion
  - Cross-validation assessment of fit
    * LOO-CV: leave-one-out cv
    * sped up by some neat theory: PSIS: pareto-smoothed importance sampling
  - Model probabilities, P[model | data]
- Account for all modeled sources of uncertainty

- most models have "nuisance" parameters
- e.g. the variance in a t-test
  - simple problems: can account for that uncertainty
  - e.g., by using a T distribution
- much harder in more complicated problems
- e.g., σ^2_{plot} in the subsampling model
- non-Bayesian methods usually fix those at estimated values
- ignores uncertainty in those estimates
- Kenward-Roger degrees of freedom
- Bayesian methods account for that uncertainty
- in a principled manner, without problem-specific adjustments

What you don't get from a Bayesian analysis

- Confidence intervals
  - Replaced by credible intervals
  - Can choose priors so that credible interval = confidence interval
  - called matching or probability matching priors
- p-values
  - statements about probability of the data
  - Bayes conditions on the data
  - Bayes factors can be used as an alternative
- General shift from yes/no decisions to "how big", "how precise" questions

What you get by being a Bayesian in the 21'st century

- Bayes requires solving that integral/sum

$$f(\theta \mid \text{data}) = \frac{f_Y(\text{data}|\theta) f_X(\theta)}{\int_X f_Y(\text{data}|\theta) f_X(\theta) d\theta}$$

- For years, was a huge limitation
- Some combinations of $f(\text{data} \mid \theta)$ and $f(\theta)$ are "nice"
  - Analytical solution to that integral
  - Called "conjugate" priors
  - Beta distributions for binomial probabilities
  - Normal distributions for means
  - Gamma distributions for variances
- Not appropriate for most non-trivial problems
- late 1980's "the MCMC revolution"
  - Markov-Chain-Monte-Carlo
  - A collection of numerical methods to draw samples from the the posterior distribution without solving that integral
- 1997: BUGS/WinBUGS software
  - Allowed a data analyst to write out a model
  - software took care of all the computation
- Now: superseded by JAGS, STAN
- All have R interfaces to handle data management and graphing

Doing Bayes in practice

- MCMC is an iterative algorithm
  - requires initial values
  - want the stationary distribution
  - discard "burn-in" samples
  - how much to discard depends on problem
- Assess convergence to the stationary distribution
  - Use 3 or 4 sets of initial values
  - see whether they give similar distributions
  - Trace plots

    - Can you see the individual chains?
    - Hopefully not
  - Gelman-Rubin statistic: want close to 1
    - Ideally less than 1.05
    - hard problems have to accept up to 1.2
  - If you haven't converged
    - Increase # burning samples
    - Think about the problem:
      are your priors too loose?

- If prior is arbitrary:
  - How important is the choice of prior?
  - Rerun with different choices
  - Prior sensitivity analysis
  - If sufficient data, "data overwhelms the prior"
- Gives you samples from the posterior distribution for each parameter, from which you can compute:
  - median or mean estimate
  - standard errors
  - credible intervals
- for parameters or combinations / transformations of parameters

Common "weakly-informative" priors

- means, regression coefficients
  - $\theta \sim N(0, \sigma^2$, with large σ^2, e.g. 1000
  - but be careful if logit(θ)
- variances
  - canonical: $\Gamma(0.001, 0.001)$
    - mean = 1, variance = 1000
    - doesn't put enough probability close to 0
    - concern for hierarchical model variances, e.g., σ^2_{plot}
  - Exponential(1)

  - sd   Unif(0, 100) or sd   Unif(0, 10)
  - Half-Cauchy

Local linear trend model for population dynamics:

- A very useful hierarchical model
- Two latent variables
  - level: l, mean at time i
  - slope: s, non-random change in mean at time i

$$
\begin{aligned}
\text{process} & & \\
s_i &= s_{i-1} + \gamma_i \\
l_i &= l_{i-1} + s_{i-1} + \tau_i \\
\text{observation} & & \\
Y_i &\sim N(l_i, \sigma^2_{error}) \\
or: Y_i &\sim Pois(exp(l_i)) \\
or: Y_i &\sim NegBin(exp(l_i), \phi) \\
\\
\gamma_i &\sim N(0, \sigma^2_{slope}) \\
\tau_i &\sim N(0, \sigma^2_{level})
\end{aligned}
$$

- Simplifications give other useful models
  - $\sigma^2_{slope} = 0$: s_i constant
  - $\sigma^2_{slope} = 0$ and $\sigma^2_{level} = 0$:
    $l_i = l_{i-1} + s$
    $l_t = l_0 + st$
- An example of a state-space time series model