Topics in hierarchical Bayesian modeling

Instructor: Kiona Ogle
Northern Arizona University
Informatics & Computing Program

Nutrient Network Bayesian Modeling Workshop
Sunday August 2, 2015

1

Lecture content

- Simple Bayesian formulation and graphical models
- Simple example of tree mortality
  - Non-hierarchical vs hierarchical versions
  - Borrowing of strength
  - Model complexity & effective # of parameters
- Detailed case study of ecosystem respiration data
  - Hierarchical model with fixed and random effects
  - Treatment-specific fixed effect, plot random effects
  - Considerations when modeling random effects
  - Modeling building process and choice of priors
  - Implementation and derived quantities (treatment contrasts, Bayesian p-values)
- Additional topics relevant to hierarchical Bayesian modeling

2

Motivation for hierarchical models

- Represent variation among individuals, plots, times, etc.
- Representation of processes occurring at multiple scales:
  - Spatial scales (e.g., plots, watersheds, basins, regions, etc.)
  - Temporal scales (e.g., hour, day, season, year, etc.)
  - Biological/ecological scales (e.g., cells, tissues, individuals, populations, species, communities, etc.)
- Accommodate different sources of uncertainty
  - E.g., separation of observation and process error
- Representation of latent processes
  - E.g., true, hidden, or unobservable processes

3

Motivation for hierarchical models

- Ability to factor complicated models into easy to "think about" conditional pieces
  - Think globally, model locally
  - Graphical models help with model building
- Accommodate experimental / sampling design
  - Often have hierarchical or nested structure (e.g., individuals within plots within sites)
- Facilitates integration of multiple data types
  - Of different spatial, temporal, and biological resolutions
  - Measured with varying precision

4

Motivation for hierarchical models

- Flexibility and generality, can accommodate:
  - Unbalanced experimental designs
  - Missing data (*obtain posterior predictive distributions*)
  - Multiple data sources (*including prior sources of information*)
  - Non-linear processes (*example in this lecture*)
  - Non-Gaussian data or processes
  - Nested and non-nested effects
  - Spatial, temporal, individual effects
  - Multiple sources (or levels) of uncertainty
- Requires greater involvement of "user"
  - One must choose appropriate distributions
  - One must pick appropriate deterministic models
  - One must quantify underlying assumptions, processes, models
  - Good and bad…

5

First: Some notation & terminology

Probability density or probability of x represented as:
$f(x)$, $P(x)$, $Pr(x)$, or $[x]$

Joint probability of x and y:
$[x, y]$

Conditional probability of x given y:
$[x \mid y]$

Marginal probability of x ("ignoring" y):
$[x]$

Second: Bayesian foundation

Unknown parameters

$$[\theta \mid y] = \frac{[y,\theta]}{[y]} = \frac{[y \mid \theta][\theta]}{[y]} = \frac{[y \mid \theta][\theta]}{\int_{\theta}[y \mid \theta][\theta]d\theta}$$

Observed data

Normalizing constant $= [y]$

$$[\theta \mid y] \propto [y,\theta]$$

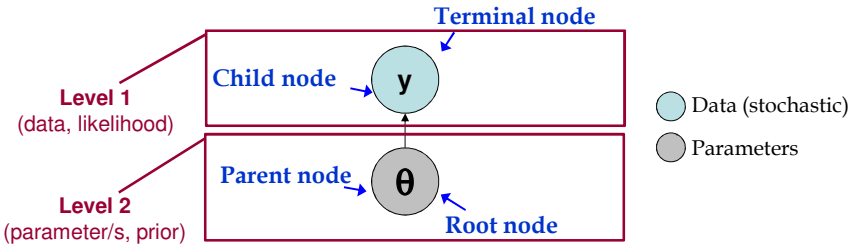
$$[\theta \mid y] \propto [y \mid \theta][\theta]$$

posterior = likelihood * the prior!

Posterior distribution | Likelihood of observed data | Prior dist'n

Graphical model

Directed Acyclic Graph (DAG)



$$[\theta, y] = [y \mid \theta][\theta]$$

$$[\theta \mid y] \propto [\theta, y]$$

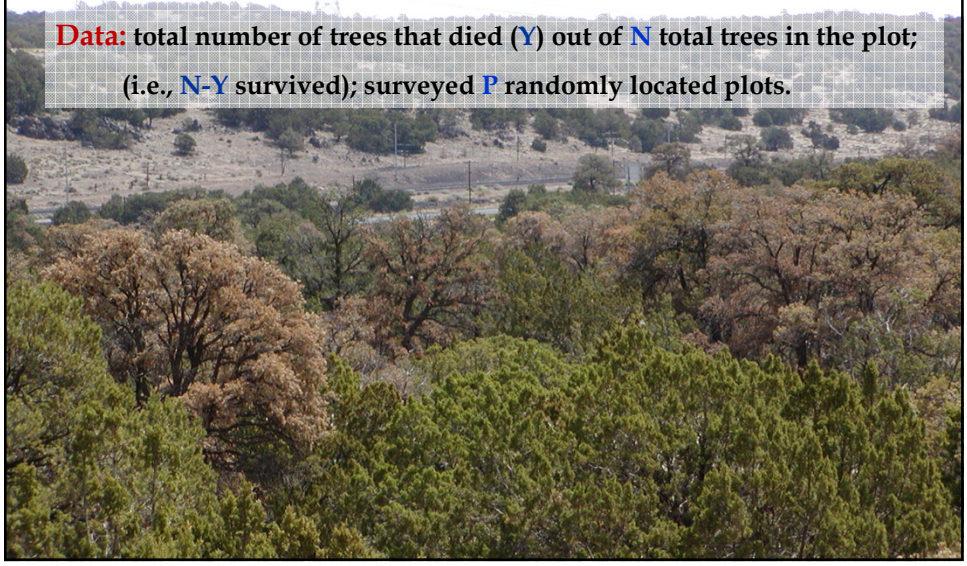
$$[\theta \mid y] \propto [y \mid \theta][\theta]$$

posterior = likelihood * the prior!

Simple example

Objective: estimate the probability of a tree dying during a drought (θ)

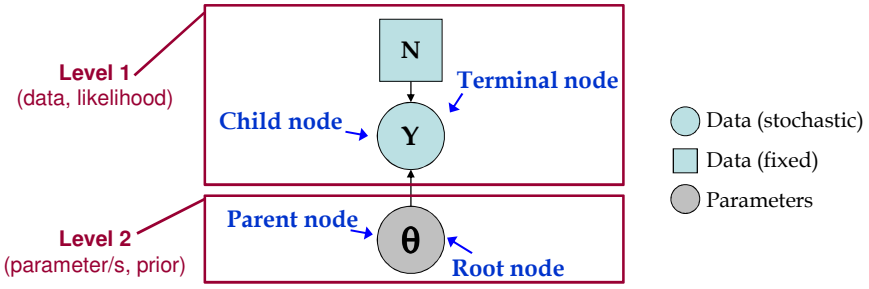
Data: total number of trees that died (Y) out of N total trees in the plot; (i.e., N-Y survived); surveyed P randomly located plots.



The DAG

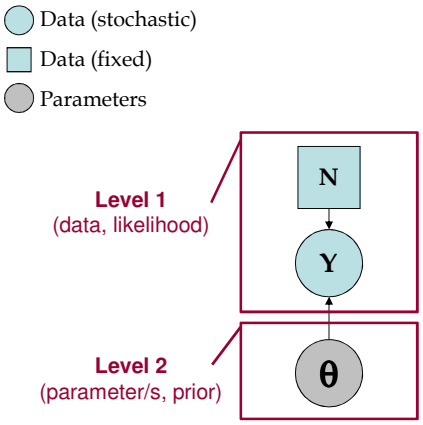
Parameter: θ = probability of a tree dying during a drought

Data: Y_i = number of trees that died in plot i (i = 1, 2, …, P)
N_i = total number of trees in plot i



$$[\theta \mid Y, N] \propto [\theta, Y \mid N] \neq [\theta, Y, N]$$
$$\propto [Y \mid \theta, N][\theta \mid N]$$
$$\propto [Y \mid \theta, N][\theta]$$ Prior for θ often doesn't depend on fixed "covariates"
$$\propto [Y \mid \theta][\theta]$$ Ignore fixed "covariates"

The statistical model

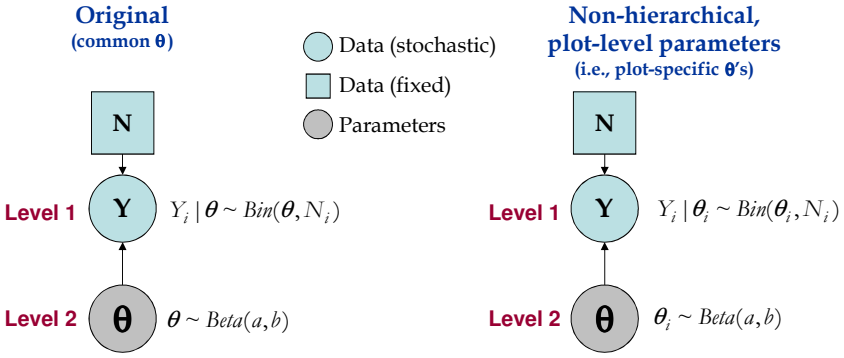


$$[Y_i \mid \theta, N_i] = Binomial(\theta, N_i)$$
$$Y_i \mid \theta \sim Bin(\theta, N_i)$$
$$[\theta] = Beta(a, b)$$
$$\theta \sim Beta(a, b)$$

theta can vary by plot

conjugate prior = special prior with the same functional form as the posterior; makes the algebra of this easier

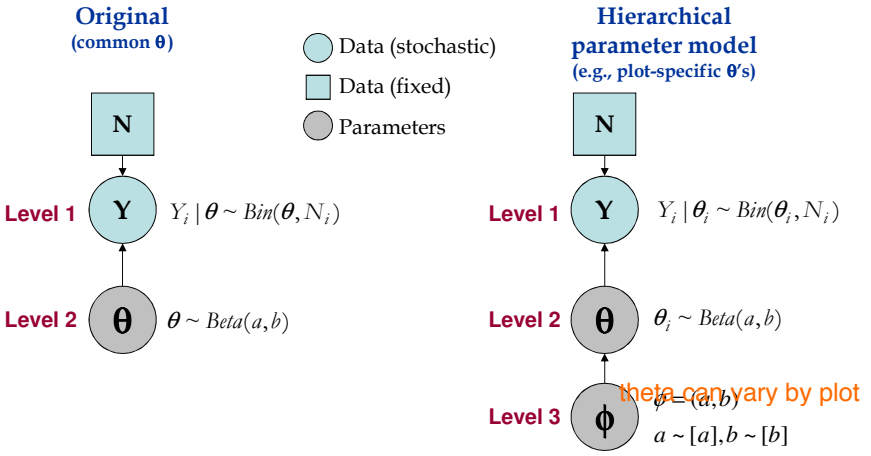
Modification: Plot-level parameters



$$[\theta \mid Y] \propto [Y, \theta]$$
$$[\theta \mid Y] \propto [Y \mid \theta][\theta]$$

$$[\theta \mid Y] \propto ??$$

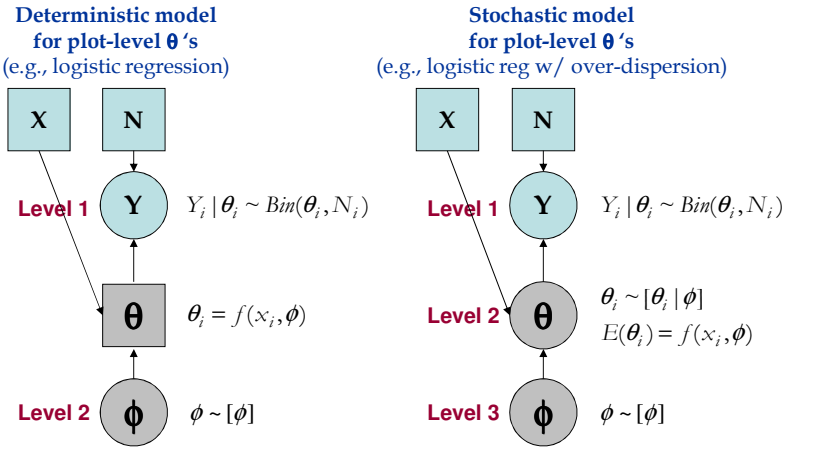
Modification: Hierarchical version



$$[\theta \mid Y] \propto [Y, \theta]$$
$$[\theta \mid Y] \propto [Y \mid \theta][\theta]$$

$$[\theta, \phi \mid Y] \propto ??$$

Alternative approaches



$[\phi \mid Y] \propto$??

$[\theta, \phi \mid Y] \propto$??

X = stand density

Example data

Plot	Y	N	X	Proportion
1	13	15	35.23	0.87
2	8	10	46.20	0.80
3	12	17	38.11	0.71
4	9	13	36.06	0.69
5	12	14	48.23	0.86
6	10	11	65.25	0.91
7	10	12	48.42	0.83
8	13	16	50.75	0.81
9	14	15	64.60	0.93
10	1	4	12.00	0.25

Y = number of dead trees
N = total number of trees
X = stand density
Proportion = proportion dead = Y/N

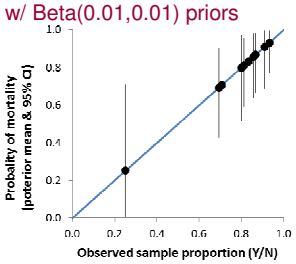
Borrowing of strength

- Sharing of information among unknown quantities (e.g., plot-level θ's) in a hierarchical model
- Useful for informing quantities associated with low replication (low sample sizes)
- Results in *shrinkage* towards population mean (e.g., individual θ's pulled towards E(θ))
- Degree of shrinkage depends on sample sizes and within versus between group variability

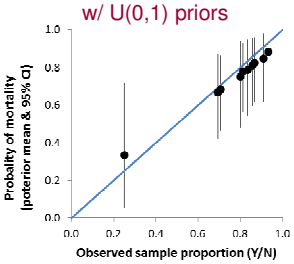
if there's a plot with low sample size, then it can learn from the other plots you're more confident in ("pulled" towards overall mean)

Borrowing of strength

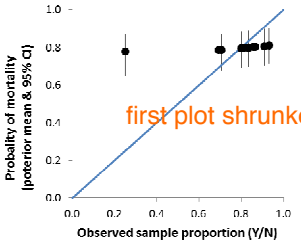
Non-hierarchical θ
w/ Beta(0.01,0.01) priors



Non-hierarchical θ
w/ U(0,1) priors

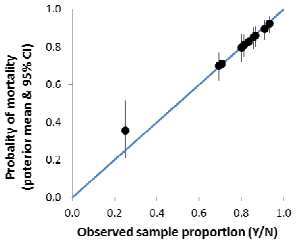


Hierarchical θ

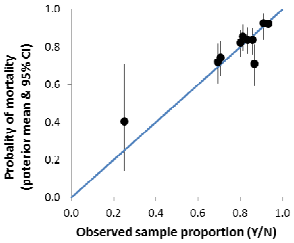


first plot shrunken towards the overall mean

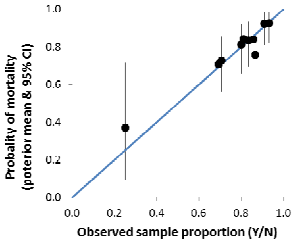
Hierarchical θ
w/ larger sample size (x10)



Deterministic logistic
regression for θ



Stochastic logistic
regression for θ



Effective # parameters

Model	Countable parameters	Effective parameters
Common θ	1	10
Non-hierarchical plot-level θ	10 (10 plots)	10.9
Non-hierarchical plot-level θ with U(0,1) priors	10	7.4
Hierarchical plot-level θ	12 (10 theta + a and b)	1.8
Hierarchical plot-level θ, larger sample size (10x)	12	9.2
Deterministic logistic regression for plot-level θ	2	2.0
Stochastic regression for plot-level θ	13	4.1

??

Case study:

Synthesis of ecosystem respiration data from a long-term experiment

Global Change Biology

Global Change Biology (2015) **21**, 2588–2602, doi: 10.1111/gcb.12910

Antecedent moisture and temperature conditions modulate the response of ecosystem respiration to elevated CO_2 and warming

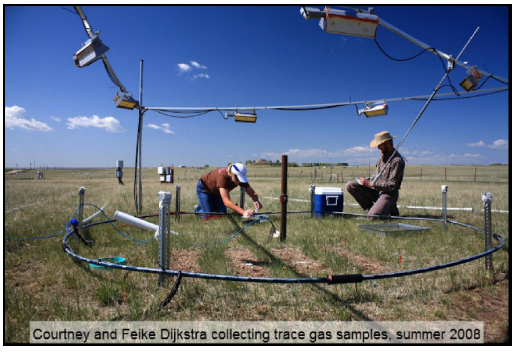
EDMUND M. RYAN[1], KIONA OGLE[1], TAMARA J. ZELIKOVA[2], DAN R. LECAIN[3], DAVID G. WILLIAMS[2], JACK A. MORGAN[3] and ELISE PENDALL[2,4]

[1]*School of Life Sciences, Arizona State University, Tempe, AZ, USA,* [2]*Department of Botany, University of Wyoming, Laramie, WY, USA,* [3]*USDA-ARS, Fort Collins, CO, USA,* [4]*Hawkesbury Institute for the Environment, University of Western Sydney, Penrith, NSW, Australia*

19

Prairie Heating & CO_2 Enrichment Experiment (PHACE)

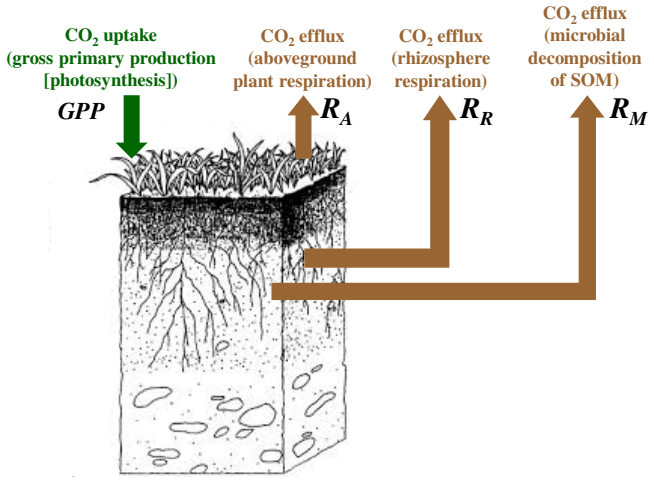
- Semi-arid mixed-grass prairie near Cheyenne, WY
- Global change experiment, incomplete factorial:
  - Atmospheric CO_2 (2 levels)
  - Temperature (2 levels)
  - Water (3 levels)
- Over 6 years of experimental data (2007-2012)
- Data sets compiled into large relational database



Courtney and Feike Dijkstra collecting trace gas samples, summer 2008

https://sites.google.com/site/pendalllab/gallery

Ecosystem respiration



- Soil respiration (R_{soil})
  - $R_{soil} = R_R + R_M$
- Ecosystem respiration (R_{eco})
  - $R_{eco} = R_A + R_R + R_M$

"Process" model framework

Mathematical model

Data model

$$\log(R) \sim Normal\left(\log(\mu_R), \sigma^2\right)$$

Process model

Respiration rate = *base rate* × *temperature sensitivity*

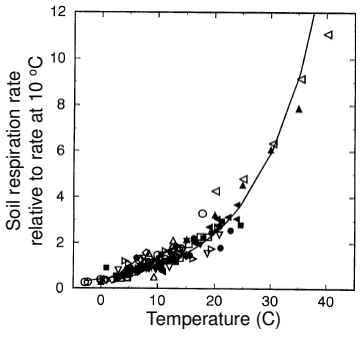
$$\mu_R = R_{base} \cdot \exp\left(Eo\left[\frac{1}{T_{base} - To} - \frac{1}{T - To}\right]\right)$$

$$\log(\mu_R) = \log(R_{base}) + Eo\left(\frac{1}{T_{base} - To} - \frac{1}{T - To}\right)$$

Symbol key:

R = respiration rate (observed)
T = temperature (observed)
μ_R = expected respiration rate (predicted)
R_{base} = base respiration rate at a reference temperature of T_{base}
Eo = temperature sensitivity parameter related to *energy of activation*
To = temperature sensitivity parameter

General pattern across biomes



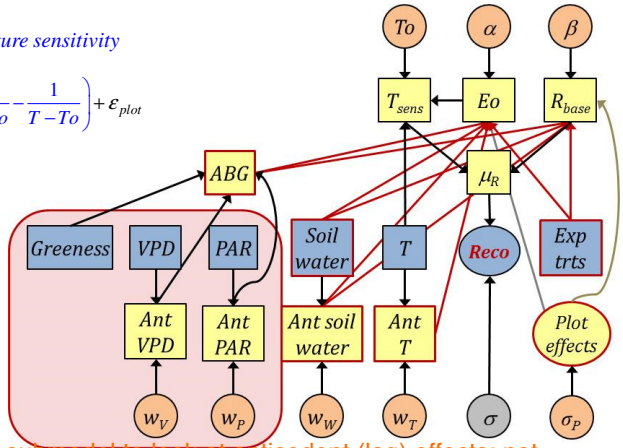
Lloyd & Taylor (1994) On the temperature dependence of soil respiration. *Functional Ecology* 8:315-323

Extension to ecosystem respiration data

Process model:

predicted = *base rate* × *temperature sensitivity*

$$\log(\mu_R) = \log(R_{base}) + Eo\left(\frac{1}{T_{base} - To} - \frac{1}{T - To}\right) + \varepsilon_{plot}$$



Symbol key:
Reco = respiration rate
T = temperature
μ_R = expected respiration rate
R_{base} = base respiration rate at temperature of T_{base}
Eo = temperature sensitivity related to *energy of activation*
To = temperature sensitivity parameter

submodel to look at anticedent (lag) effects; not going to estimate those parameters today

N observations:

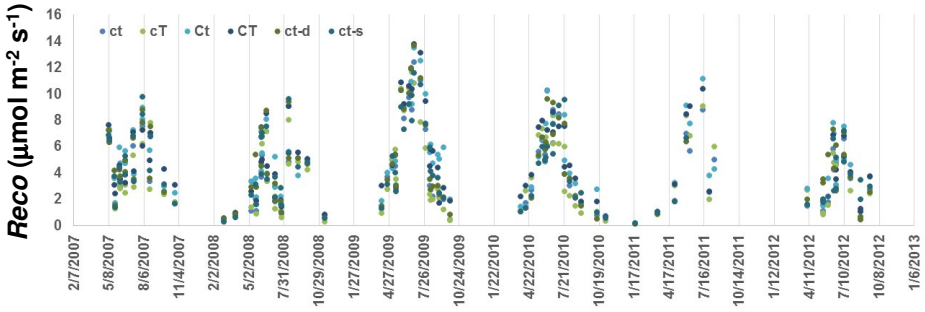
R_{eco} = 3570 (96 days)	N years = 6 (2007-2012)
Soil water = 200,000	N plots = 30
Soil temp = 3.1 million	N trts = 6
Micromet = 52000	

Ryan, Ogle, et al. (2015) *Global Change Biology*

blue boxes = data

Data: Time-series of observed *Reco*

Mean observed *Reco* by treatment over the 6-year study period



24

Posterior $\propto$ Likelihood $\times$ Prior

Simple Bayesian expression:

$$[\theta \mid y] \propto [y \mid \theta][\theta]$$

Modified to accommodate *Reco* analysis:

$$[To, \alpha, \beta, \varepsilon, \gamma, \sigma, \sigma_\varepsilon, \sigma_\gamma \mid Reco] \propto$$

$$[Reco \mid To, \alpha, \beta, \varepsilon, \gamma, \sigma][To][\alpha][\beta][\varepsilon \mid \sigma_\varepsilon][\gamma \mid \sigma_\gamma][\sigma][\sigma_\varepsilon][\sigma_\gamma]$$

blue = heirarchical priors, which are almost always used for random effects

25

Likelihood

$$[To, \alpha, \beta, \varepsilon, \gamma, \sigma, \sigma_\varepsilon, \sigma_\gamma \mid Reco] \propto [Reco \mid To, \alpha, \beta, \varepsilon, \gamma, \sigma][To][\alpha][\beta][\varepsilon \mid \sigma_\varepsilon][\gamma \mid \sigma_\gamma][\sigma][\sigma_\varepsilon][\sigma_\gamma]$$

For **observation** i = 1, 2, …, 3567:

$$Reco_i \sim Normal(\mu_i, \sigma)$$

Vary by treatment?

Likelihood of all *Reco* data: in JAGS / OpenBUGS, this part happens in the bg

$$[Reco \mid To, \alpha, \beta, \varepsilon, \gamma, \sigma] = \prod_{i=1}^{3567} Normal(Reco_i \mid \mu_i, \sigma)$$

26

Process (mean) model

$$[To, \alpha, \beta, \varepsilon, \gamma, \sigma, \sigma_\varepsilon, \sigma_\gamma \mid Reco] \propto [Reco \mid To, \alpha, \beta, \varepsilon, \gamma, \sigma][To][\alpha][\beta][\varepsilon \mid \sigma_\varepsilon][\gamma \mid \sigma_\gamma][\sigma][\sigma_\varepsilon][\sigma_\gamma]$$

For **observation** i = 1, 2, …, 3567 and **treatment** t = 1, 2, …, 6 associated with observation i, $t\{i\}$:

$$Reco_i \sim Normal(\mu_i, \sigma)$$

soil temperature (covariate)

$$\log(\mu_i) = LR_{base,i} + Eo_i \left(\frac{1}{T_{base} - To_{t\{i\}}} - \frac{1}{T_i - To_{t\{i\}}} \right)$$

$LR_{base,i} = \log(R_{base,i})$

constant set by "user" (e.g., 10 °C or 283.15 Kelvin)

treatment-specific temperature sensitivity parameter

Second-level process model

For **treatment** t and **plot** p = 1, 2, …, 30 associated with observation i:

$$\log(\mu_i) = LR_{base,i} + Eo_i \left(\frac{1}{T_{base} - To_{t\{i\}}} - \frac{1}{T_i - To_{t\{i\}}} \right)$$

Base-rate sub-model:

$$LR_{base,i} = \alpha_{1,t\{i\}} + \alpha_{2,t\{i\}} SWC_i + \alpha_{3,t\{i\}} SWC_{ant,i} + \alpha_{4,t\{i\}} SWC_i SWC_{ant,i} + \left(\alpha_{5,t\{i\}} + \alpha_{6,t\{i\}} PAR_{ant,i} + \alpha_{7,t\{i\}} VPD_{ant,i} + \alpha_{8,t\{i\}} VPD_{ant,i} PAR_{ant,i}\right) Greenness_i + \gamma_{p\{i\}}$$

***Eo* ("energy-of-activation") sub-model:**

$$Eo_i = \beta_{1,t\{i\}} + \beta_{2,t\{i\}} SWC_i + \beta_{3,t\{i\}} SWC_{ant,i} + \beta_{4,t\{i\}} SWC_i SWC_{ant,i} + \beta_{5,t\{i\}} T_{ant,i} + \left(\beta_{6,t\{i\}} + \beta_{7,t\{i\}} PAR_{ant,i} + \beta_{8,t\{i\}} VPD_{ant,i} + \beta_{9,t\{i\}} VPD_{ant,i} PAR_{ant,i}\right) Greenness_i + \varepsilon_{p\{i\}}$$

aboveground / plant effects

treatment-specific covariate ***fixed*** effects parameters (α_1, α_2, …, α_8, β_1, β_2, …, β_9)

plot ***random*** effects

Employ covariance centering

Hierarchical priors (plot random effects)

$$[To, \alpha, \beta, \varepsilon, \gamma, \sigma, \sigma_\varepsilon, \sigma_\gamma \mid Reco] \propto [Reco \mid To, \alpha, \beta, \varepsilon, \gamma, \sigma][To][\alpha][\beta][\varepsilon \mid \sigma_\varepsilon][\gamma \mid \sigma_\gamma][\sigma][\sigma_\varepsilon][\sigma_\gamma]$$

Option 1: Treat random effects like "additive error terms"

$$LR_{base,i} = \alpha_{1,t\{i\}} + \alpha_{2,t\{i\}} SWC_i + ... + aboveground\ effects + \gamma_{p\{i\}}$$

$$\gamma_p \sim Normal(0, \sigma_\gamma)$$ ← zero-centered hierarchical prior

Option 2: Hierarchically center random effects (nest in treatment)

$$LR_{base,i} = \gamma_{p\{i\}} + \alpha_{2,t\{i\}} SWC_i + ... + aboveground\ effects$$

For **treatment** t associated with **plot** p, $t\{p\}$:

$$\gamma_p \sim Normal(\alpha_{1,t\{p\}}, \sigma_\gamma)$$

29

Aside: which option?

Option 1: Treat random effects like "additive error terms"
When to use: Multiple additive random effects that vary at different scales

E.g., for plot p, day d, and observer o:

Overall (global) intercept

$$\mu_i = \alpha_1 + \alpha_2 X_i + ... + \gamma_{p\{i\}} + \varepsilon_{y\{i\}} + \lambda_{o\{i\}}$$

$$\gamma_p \sim Normal(0, \sigma_\gamma),\ \varepsilon_d \sim Normal(0, \sigma_\varepsilon),\ \lambda_o \sim Normal(0, \sigma_\lambda)$$

Option 2: Hierarchically center random effects
When to use: For nested random effects

E.g., for sup-plot s nested in plot p, with plot random effects:

$$\mu_i = \gamma_{s\{i\}} + \alpha_2 X_i + ...$$

$$\gamma_s \sim Normal(\varepsilon_{p(s)}, \sigma_\gamma)$$
$$\varepsilon_p \sim Normal(\alpha_1, \sigma_\varepsilon)$$

Overall (global) intercept

30

Issues / problems?

Option 1: Treat random effects like "additive error terms"
Problem: Non-identifiable intercept and random effects

- Intercept and random effects are added together; we can only identify (estimate) their sum.

$$\mu_i = \alpha_1 + \alpha_2 X_i + ... + \gamma_{p\{i\}} + \varepsilon_{y\{i\}} + \lambda_{o\{i\}}$$

- The zero-centered hierarchical priors are ***priors***, the posterior distribution of each group of random effects will not have a mean exactly equal to zero.

$$\gamma_p \sim Normal(0, \sigma_\gamma),\ \ \varepsilon_d \sim Normal(0, \sigma_\varepsilon),\ \ \lambda_o \sim Normal(0, \sigma_\lambda)$$

go to "Lecture 12" examples

Solution: Impose sum-to-zero constraints on random effects, thus forcing each (posterior) group of random effects to have a mean of zero.

Option 2: Hierarchically center random effects
Problem: Doesn't work for non-nested random effects
Solution: Use Option 1

Priors for root nodes

$$[To, \alpha, \beta, \varepsilon, \gamma, \sigma, \sigma_\varepsilon, \sigma_\gamma \mid Reco] \propto [Reco \mid To, \alpha, \beta, \varepsilon, \gamma, \sigma][To][\alpha][\beta][\varepsilon \mid \sigma_\varepsilon][\gamma \mid \sigma_\gamma][\sigma][\sigma_\varepsilon][\sigma_\gamma]$$

Informative priors:

$$\log(\mu_i) = LR_{base,i} + Eo_i\left(\frac{1}{T_{base} - To_{t\{i\}}} - \frac{1}{T_i - To_{t\{i\}}}\right)$$

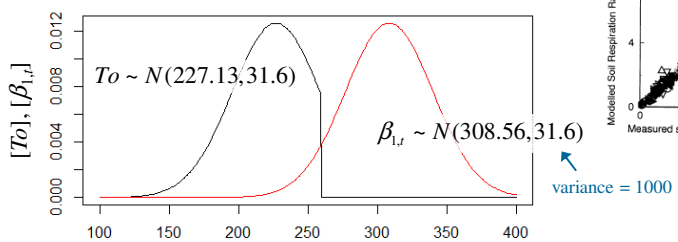
"reference" *Eo* at average or reference covariate values and no plants (*Greenness* = 0)

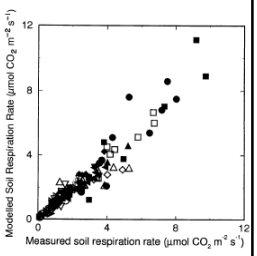
$$Eo_i = \beta_{1,t\{i\}} + \beta_{2,t\{i\}} SWC_i + ... + aboveground\ effects + \varepsilon_{p\{i\}}$$

Functional Ecology 1994 **8,** 315–323

On the temperature dependence of soil respiration

J. LLOYD and J. A. TAYLOR*





Priors for root nodes

$$[To, \alpha, \beta, \varepsilon, \gamma, \sigma, \sigma_\varepsilon, \sigma_\gamma \mid Reco] \propto [Reco \mid To, \alpha, \beta, \varepsilon, \gamma, \sigma][To][\alpha][\beta][\varepsilon \mid \sigma_\varepsilon][\gamma \mid \sigma_\gamma][\sigma][\sigma_\varepsilon][\sigma_\gamma]$$

Vague, relative non-informative priors:

Treatment-specific effects:

$$\alpha_{k,t}, \beta_{k+1,t} \sim N(0, 316.2) \qquad \text{for } k = 1, 2, ..., 8$$

variance = 100,000

Standard deviation terms:

$$\sigma \sim U(0,10),\ \sigma_\varepsilon \sim U(0,150),\ \sigma_\gamma \sim U(0,150)$$

describes *observation error*, informed by lots of data (large "group" size), could use a standard, conjugate gamma prior for precision

for variance terms that describe variability associated with "small" group sizes, use uniform or folded-t priors for standard deviations

33

Connection to "classical" terminology

The *Reco* model is an example of a:

- Non-linear regression
- Mixed effects model
  (fixed effects [treatments], random effects [plots])
- Multi-level model

34

Implementation

Software for implementing Bayesian models:

- OpenBUGS (Windows, Linux)
- WinBUGS (Windows, Linus)
- JAGS (Windows, Mac, Linux?)
- Stan (Windows, Mac, Linux?)

Use numerical routines to sample from posterior:

- Markov chain Monte Carlo (MCMC)
- Metropolis-Hastings
- Gibbs sampling
- Slice sampler
- Hamiltonian (Stan)
- Many others

Implementation

***Reco* model implemented in OpenBUGS and JAGS**

- Example, simplified code and output
- Treatment differences (compute various contrasts)
  - E.g., pairwise treatment differences
  - *Does the effect of soil water on R_{base} differ among treatment levels* 1 *and* 2*:*

$$LR_{base,i} = \alpha_{1,t\{i\}} + \alpha_{2,t\{i\}} SWC_i + ... + \gamma_{p\{i\}}$$
$$\Delta_{2,\{1,2\}} = \alpha_{2,1} - \alpha_{2,2}$$

- Bayesian p-values for treatment differences
- Obtain posteriors for *derived* quantities
  - E.g., overall greenness effect (*GE*)

$$LR_{base,i} = \alpha_{1,t\{i\}} + \alpha_{2,t\{i\}} SWC_i + \alpha_{3,t\{i\}} SWC_{ant,i} + \alpha_{4,t\{i\}} SWC_i SWC_{ant,i} +$$
$$\left(\alpha_{5,t\{i\}} + \alpha_{6,t\{i\}} PAR_{ant,i} + \alpha_{7,t\{i\}} VPD_{ant,i} + \alpha_{8,t\{i\}} VPD_{ant,i} PAR_{ant,i}\right) Greenness_i + \gamma_{p\{i\}}$$
$$GE_i = \left(\alpha_{5,t\{i\}} + \alpha_{6,t\{i\}} PAR_{ant,i} + \alpha_{7,t\{i\}} VPD_{ant,i} + \alpha_{8,t\{i\}} VPD_{ant,i} PAR_{ant,i}\right)$$

Implementation

Example code, execution of code, and output

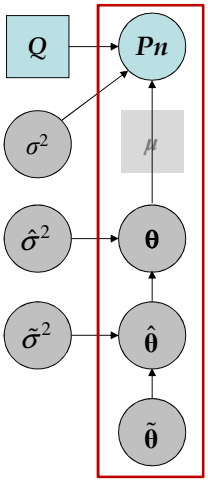
Additional topics

- Building hierarchical models with multiple levels
- Experimental design considerations
- Linkages to "fixed" and "random" effects terminology
- The "process sandwich" and process error
- Identifiability issues
- Extensions to multiple datasets

39

Hierarchical model w/ multiple levels

- Note dimension reduction when moving from level L to L+1
- Photosynthesis example: assuming full factorial design with N_c observations per curve (N total), C_p curves per plant (C total), P_s plants per species (P total), and S species



$N_c \times C_p \times P_s \times S$ observations **Level 1** (data/likelihood)

$C_p \times P_s \times S$ curves

Level 2 (priors; parameters directly linked to likelihood)

$P_s \times S$ plants

Level 3 (hyperpriors; hyper-parameters linked to priors)

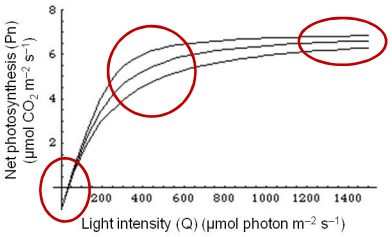
S species

Level 4 (hyper-hyperpriors; linked to hyperpriors)

40

Experimental design considerations

- Model specification and existing theory help to inform experimental design
- **Photosynthesis example:**
  - To separate curve-, plant-, and species-level effects
    - Must collect multiple observations for each level
    - Multiple curves per plant, multiple plants per species, 2+ species
  - If wish to use data to inform all photosynthesis parameters (*Pmax*, *Rd*, α, β)
    - Need to collect data that span different portions of curve (process model) (at a range of appropriate light levels)



41

Fixed vs random effects

	Random	Fixed
Levels	selected at random from "infinite" # of possibilities	finite # of possibilities
Another study	would likely use different levels from same population	would use same levels of the factor
Goal	estimate variance components (i.e., don't care about parameters or means associated with each level)	estimate means for each level (*and quantify *variability between levels*)
Inference	for population from which levels are selected	for levels actually selected, and to some extent, *for *population* from which levels are selected

*not applicable to classical approaches

- In Bayesian, the distinction between random and fixed effects is blurry
- But, the notion of fixed and random effects can facilitate model building and inference
- Consider the species-level parameters in the previous example; if we had modeled them hierarchically (vs root nodes), how would this affect our interpretations?
- Convenient to use random effects concepts and modeling approaches when separating observation error and process error…

42

Example: Fixed vs random effects

- Say we are interested in a parameter (θ) that varies by species identity, or experiment level (*j*)
- Assign prior that treats θ like a *fixed effect* (θ is typically a **root node** in the DAG).
  - For example, may chose relatively non-informative prior (mean = 0, small precision [large variance (as shown here)]):

$$\theta_j \sim Normal(0, 100000)$$

- Assign prior that treats θ like a *random effect* (θ typically has parent nodes in the DAG).
  - Treated θ like an "error" term (e.g., mean = 0; precision [variance] unknown and often a root node):

$$\theta_j \sim Normal(0, \sigma^2)$$

  - Treated θ as coming from a population with an unknown mean and unknown precision [variance]; μ root node or modeled hierarchically, σ often a root node :

$$\theta_j \sim Normal(\mu, \sigma^2)$$

43

The "process sandwich"

Recall "simple Bayesian" formulation:

$$[\theta \mid Data] \propto [Data \mid \theta] \cdot [\theta]$$

Extension to hierarchical priors (parameter models):

$$[\theta, \phi \mid \mathbf{y}] \propto [\mathbf{y} \mid \theta][\theta \mid \phi][\phi]$$

$$[\theta_1, \theta_2, ..., \theta_K \mid \mathbf{y}] \propto [\mathbf{y} \mid \theta_1, \theta_2, \theta_3 ..., \theta_K][\theta_1 \mid \theta_2, \theta_3 ..., \theta_K][\theta_2 \mid \theta_3 ..., \theta_K] ... [\theta_K]$$

Extension to the hierarchical Bayesian (process sandwich) formulation:

Unknown quantities; Observed data; Latent (or true) process; Process parameters; Data parameters

$$\underbrace{[\theta_D, \theta_P, Process \mid Data]}_{\text{Posterior}} \propto \underbrace{[Data \mid Process, \theta_D]}_{\text{Likelihood}} \underbrace{[Process \mid \theta_P]}_{\text{Process model}} \underbrace{[\theta_D, \theta_P]}_{\text{Prior(s)}}$$

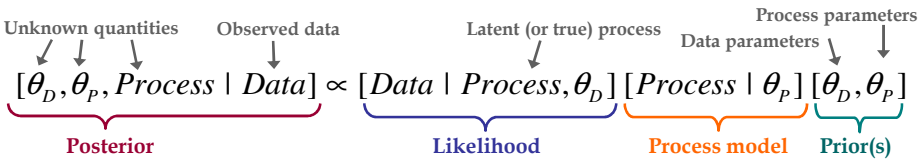
…with hierarchcial parameter model (priors)

$$[\theta_D, \theta_P, \phi_D, \phi_P \mid Process \mid Data] \propto$$

$$[Data \mid Process, \theta_D]\,[Process \mid \theta_P]\,[\theta_D \mid \phi_D][\theta_P \mid \phi_P][\phi_D][\phi_P]$$

44

The "process sandwich"



Data model (likelihood)

$[Data \mid Process, \theta_D]$:

$Data = Latent\ process + observation\ error$

Process model

$[Process \mid \theta_P]$:

$Latent\ process = Expected\ process + process\ error$

The "ecological process model"
(mathematical or simulation model)

45

Sources of uncertainty

- Measurement error
  - Instrument precision or analytical error
  - Instrument bias or drift
  - Different observers
  - Method or instrument used
  - Rigor of measurement protocol
- Other sampling or observational error
  - Aggregation, extrapolation, or scaling errors
- Process error
  - Didn't measure or account for all "key" drivers/covariates or potential interactions
  - Process model overly simplifies underlying true process
  - Process model inappropriate or lacking key functional relationships
  - Often spatially or temporally structured

46

Potential solutions?

- Incorporate "structure" into observation error model
  - E.g., temporal or spatial correlation
- Introduce process errors at different level
  - E.g., as an explicit random effect at coarser level (not observation level)
  - See next few slides…
- Employ informative priors for one of the error variances
  - E.g., most likely for measurement or observation error variance (motivated by other studies / existing information)

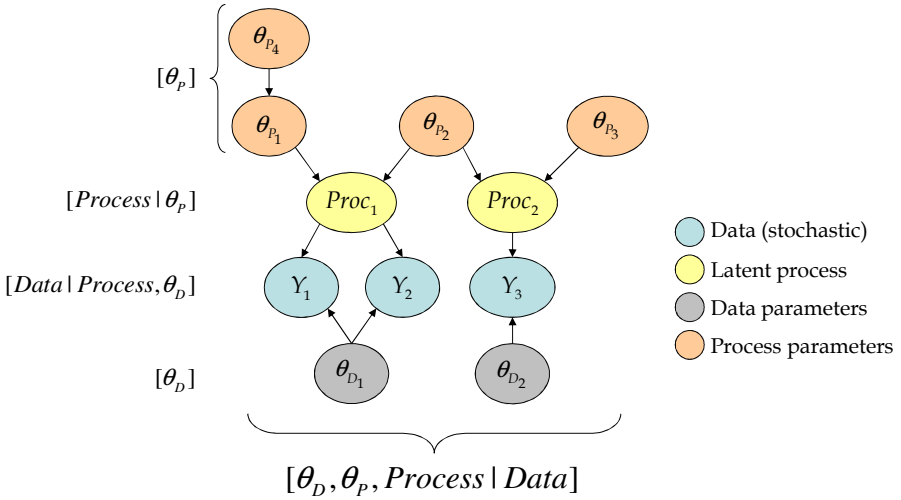
47

Experimental design considerations

- Experimental design is also key to separating observation or measurement error and process error
- Potential options:
  - Obtain "calibration" datasets to estimate measurement error variance (i.e., where true values may be known)
  - Collect "replicate" observations of y_i at each level (or a subset of levels) of the covariate(s) for all or a subset of experimental groups
- Collect data to facilitate more precise and accurate estimates of process parameters
  - ANOVA-based designs generally not appropriate for fitting process models
  - Response-surface type designs potentially more useful if constructed to sample along range of covariates
  - Sample across "range of process model" (i.e., need data to inform different parts of our process model / mean function)
- Model results can inform future experiment design
  - E.g., targeted measurements to reduce uncertainty in certain parameters

48

Extension to multiple datasets & processes



49

Extension to multiple datasets & processes

- **Advantages of hierarchical Bayesian methods**
  - Propagate uncertainty among model components
    - Avoid piece-wise, *ad hoc* methods for propagating uncertainty between model components
  - Link datasets by latent, shared processes or process parameters
  - Explicit accounting / modeling of correlations between shared processes and parameters
    - Prior(s) may assumed independence among components
    - Posterior reveals potential correlations between components
  - Can align data that are misaligned in time or space (often misalignment leads to missing data)
  - Can combined data from different types of experiments
    - Manipulative, observational, etc.

50

References

- Priors for variance terms in hierarchical models, or how to model variances terms hierarchically:
  - Gelman (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:515-533.
- Parameter expansion to address "zero variance trap" and to improving mixing in hierarchical models:
  - Gelman & Hill (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York.
    - Chapter 19: Debugging and speeding convergence
    - Section 19.4: Redundant parameters and intentionally nonidentifiable models
  - Gelman (2004) Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99:537-545.

51

References

- Gelman & Hill (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
- Gilks and Roberts (1996) Strategies for improving MCMC, in *Markov Chain Monte Carlo in Practice*, edited by Gilks, Richardson, and Spiegelhalter. Chapman & Hall/CRC. Boca Raton, 486 pages.
  - Section on random effects most relevant to this lecture
- Berry and Hochberg (1999) Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference,* 82: 215-227
  - I haven't read this paper yet…

52

Some approaches to parameterizing hierarchical priors

- **Option 1**: Use hierarchical normal priors for parameters (e.g., θ), regardless of their support.
- **Option 2**: Transform θ to the real-line; e.g., if $\theta > 0$, use $\phi = \log(\theta)$; assign normal hierarchical priors to ϕ; back-transform to obtain $\theta = \exp(\phi)$.
  - I do this frequently
- **Option 3**: Use prior that obeys support of θ (e.g., lognormal, gamma, etc. if $\theta > 0$; beta if $0 < \theta < 1$), assign hierarchical priors to $E(\theta)$, use moment matching to compute parameters (e.g., shape and scale) of the prior distributions.
  - See next slide

54

Option 3

- Could choose a different *parameterization* that obeys support of original scale, for convenience or ease of interpretation
- Could use "moment matching" combined with distributions defined on the positive real-line to assign hierarchical priors to the curve-level α, β, *Pmax*, and *Rd*
- Instead, let θ represent a parameter on the *original* scale (i.e., α, *Pmax*, *Rd*) defined on positive real-line. We might assume $\theta \sim Gamma(a,b)$ or $\theta \sim LogNormal(m,v)$, e.g.:

$$\theta_c \sim Gamma\left(\hat{a}_{p(c)}, \hat{b}_{p(c)}\right)$$
$$\hat{\theta}_p = E(\theta_c) \text{ and } \hat{\sigma}^2 = Var(\theta_c)$$
$$\hat{a}_p = \ldots \qquad \hat{b}_p = \ldots$$

$$\hat{\theta}_p \sim Gamma\left(\tilde{a}_{s(p)}, \tilde{b}_{s(p)}\right)$$
$$\tilde{\theta}_s = E\left(\hat{\theta}_p\right) \text{ and } \tilde{\sigma}^2 = Var\left(\hat{\theta}_p\right)$$
$$\tilde{a}_s = \ldots \qquad \tilde{b}_s = \ldots$$

Assign priors to root node parameters:

$\hat{\sigma}, \tilde{\sigma}, \tilde{\theta}_s$

(will do in lab)

$$y \sim Gamma(a,b)$$
$$E(y) = \frac{a}{b} \qquad Var(y) = \frac{a}{b^2}$$

$$a = \frac{[E(y)]^2}{Var(y)} \qquad b = \frac{E(y)}{Var(y)}$$

What distribution would you choose for β?

Recall, $0 < \beta < 1$

55

Advantages/disadvantages of each approach

- **Option 1**: Use hierarchical normal priors for θ, regardless of support.
  - **Advantages**:
    - easy to specify/interpret
    - conjugates available at higher levels
    - usually works if the posterior for θ lies from away from unrealistic values
    - can implement "tricks" to improving mixing (e.g., parameter expansion to solve "zero variance traps" and "flat-lining")
    - works well for hierarchical models with many levels
    - specification of starting values (usually) fairly straightforward
  - **Disadvantages**:
    - may cause errors if θ not defined on entire real-line
      - e.g., MCMC might sample a negative value when we know $\theta > 0$
      - this could result in numerical errors
    - requires careful specification of initial values

56

Advantages/disadvantages of each approach

- **Option 2**: Transform θ to the real-line; e.g., if θ > 0, use ϕ = log(θ); assign normal hierarchical priors to ϕ; back-transform to obtain θ = exp(ϕ).
  - **Advantages**:
    - conjugates available at higher levels
    - restricts θ to appropriate support
    - if assign prior to ϕ and compute θ = *f*(ϕ)
      - then modeling *median* hierarchically if *f* is monotonic
      - may be preferred if posterior for θ is highly skewed
    - can implement "tricks" to improving mixing (as in Option 1)
    - works well for hierarchical models with many levels
  - **Disadvantages**:
    - requires extra coding/calculations to compute θ
      - thus, may slow simulation time or result in errors (e.g., numerical overflow)
    - may want to model E(θ) hierarchically and not its median
    - specification of starting values more challenging

57

Advantages/disadvantages of each approach

- **Option 3**: Use prior that obeys support of θ (e.g., lognormal, gamma, etc. if θ > 0; beta if 0 < θ < 1), assign hierarchical priors to E(θ), use moment matching to compute parameters (e.g., shape and scale) of the prior distributions.
  - **Advantages**:
    - restricts θ to appropriate support
    - allows you to model the mean (or any other moment) hierarchically
    - works best for hierarchical models with relatively few levels (as in the lab)
    - specification of starting values (usually) fairly straightforward
  - **Disadvantages**:
    - conjugates often not available at higher levels
    - requires extra coding/calculations to compute the distribution parameters defining θ's prior
      - may slow simulation time or result in errors (e.g., numerical overflow)
    - can't easily use convergence/mixing "tricks" such as parameter expansion

58