# HW2

Sam Olson

## Q1

Discuss whether you believe it would be better to view this problem as one involving a Bayesian analysis of a mixture model, or as one we should approach with a model having several levels of prior distributions. *Hint: Read Assignment 2 – Background carefully before developing your answer to this question.*

### Answer

#### Direct Answer

Because the goal is inference about **specific lakes and their condition parameters**, the correct conceptual framework is with a model having several levels of prior distributions, rather than as a Bayesian mixture model.

Following the principle in Chapter 15, the interpretation should be based on **what quantities are scientifically meaningful to estimate**. Here, the lake-specific parameters $\beta_i$ represent the health of identifiable lakes, and inference is required for each of these individual units.

Management questions such as ("Is Lake $i$ exhibiting poor condition?", "Which lakes appear healthiest?", "Which lakes should be targeted for intervention?", etc.), all depend directly on the posterior distributions $p(\beta_i \mid y)$.

#### Added Context

**Ch. 15 Notes** Based on the 520 text, the modelling choice between a Bayesian analysis of a mixture model vs. a model having several levels of prior distributions hinges upon *what the quantity of interest is*.

If the goal is inference about *specific units*, such that the group parameters $\theta_i$ represent persistent, interpretable characteristics and one could plausibly observe additional data generated under the same $\theta_i$ (e.g., the same hospital, school, or lake), then treat the hierarchy as a **multi-stage prior**, with quantities:

$$f(y \mid \theta), \quad \pi_1(\theta \mid \lambda), \quad \pi_2(\lambda),$$

and focus on the unit-level posteriors $p(\theta_i \mid y)$ with partial pooling.

If instead the goal of inference is about the *population distribution of effects across groups*, and the observed units are best viewed as exchangeable draws from that population so that the individual $\theta_i$ are not themselves of intrinsic interest, then one should treat the model as a **mixture**, integrate out $\theta$ to obtain a posterior of the form:

$$f(y \mid \lambda) = \int f(y \mid \theta)\, g(\theta \mid \lambda)\, d\theta,$$

and focus on $p(\lambda \mid y)$ or predictive distributions $p(\theta^* \mid y)$.

**Relevant Background Context** As described in the background material, biologists at the Iowa Department of Natural Resources (IDNR) are concerned with assessing the *health ("condition") of fish populations within specific lakes*, where condition reflects how heavy fish are for their length. Although individual fish exhibit natural variability, condition is largely determined by *lake-level environmental factors*. Consequently, and rather explicitly noted, condition is interpreted as a *persistent lake-level characteristic*, not an individual-level trait.

Importantly, the scientific objective is to evaluate the health (condition) of the *particular lakes sampled (the 92 lakes in the study)*. From a management perspective, each lake represents a meaningful unit for decision-making, and additional data could plausibly be collected from the same lake under the same underlying ecological conditions. Therefore, lake-specific parameters represent scientifically interpretable quantities of direct interest rather than nuisance or latent variables.

# Q2:

Based on your answer to question 1, what would you include in a summary of your inferences associated with the problem. Be specific about types of summary information (e.g., five number summary or table of quantiles) and/or graphs and plots (e.g., scatterplot of $y$ versus $x$, plot of empirical distribution of $f$). Indicate how these inferences can be used to address the goals of the Iowa DNR.

## Answer

Because the problem is appropriately framed as a **hierarchical (multi-level) model**, inference should focus primarily on the **lake-specific condition parameters** $\beta_i$. These represent persistent, interpretable ecological characteristics of identifiable lakes and are the direct quantities needed for management decisions. The population-level hyperparameters $(\lambda, \tau^2)$ provide context by describing between-lake variability and enabling partial pooling, but they are secondary to the lake-level effects.

Accordingly, summaries should prioritize **posterior inference for each lake**, followed by **comparisons across lakes**, with **population summaries and diagnostics** used to support interpretation.

### Lake-specific summaries (primary targets)

The main inferential objects are the posteriors $p(\beta_i \mid y)$.

For each lake $i$, report:

- Posterior mean or median of $\beta_i$
- Posterior standard deviation
- 80%, 90%, and 95% credible intervals
- Posterior probability of being below a biologically meaningful threshold

$$P(\beta_i < c \mid y)$$

Tabular outputs:

- Table of posterior summaries and credible intervals for all lakes
- Ranked table (lowest to highest condition)
- Posterior ranks or probabilities of being among the worst/best lakes

Graphs:

- Caterpillar plot of $\beta_i$ with credible intervals (sorted)
- Boxplot or density of $\{\beta_i\}$ across lakes
- Optional map/heatmap for spatial interpretation

These directly allow the Iowa DNR to identify poorly conditioned lakes, prioritize interventions, compare lakes objectively, and quantify uncertainty.

**Population-level summaries (secondary context)**

To describe statewide patterns and heterogeneity, report:

- Posterior means and 95% credible intervals for $(\lambda, \tau^2)$
- Density or histogram of posterior draws of $\beta_i$

These quantify:

- typical statewide condition (location $\lambda$),
- lake-to-lake variability (spread $\tau^2$).

**Model fit and diagnostics**

To assess adequacy of the growth model:

- Scatterplot of $Y_{i,j}$ versus $x_{i,j}$ with fitted curves

$$\hat{\mu}_{i,j} = \hat{\beta}_i x_{i,j}^{\hat{\alpha}}$$

- Posterior predictive checks (simulated vs observed data)
- Residual plots

These ensure that inference for $\beta_i$ is biologically and statistically credible.

**Interpretation for the Iowa DNR**

Together, these summaries enable the DNR to:

- estimate condition for each lake with uncertainty,
- identify unusually poor or healthy lakes,
- allocate management resources efficiently,
- understand statewide variability,
- and make predictions for future or unsampled lakes.

Thus, inference appropriately emphasizes **lake-by-lake assessment**, with population summaries providing supporting ecological context.

# Q3:

A statistical issue in the use of the hierarchical model developed in Assignment 2 – Background is the fixed value of the power $\theta = 1.0$ used in the analysis. Describe how you would conduct an assessment of this modeling choice. In particular, there are two immediate alternatives to our choice. One is that a single value of $\theta$ should be adequate to reflect data behavior, but its value should be something other than $\theta = 1.0$. In this case, we might assign a prior (similarly to $\alpha$), but this question is not about what we might choose for that prior, only how we might assess the output of analysis of the hierarchical model in files `LMBmcmc1.txt` and `LMBmcmc2.txt` and the actual data (in file `LMBdat_for601.txt`) to determine whether we are motivated to include a prior for $\theta$.

The other possibility is that we might allow this power to have different values in different lakes, and make use of parameters $\{\theta_i : i = 1, \ldots, n\}$. These quantities would then need to be assigned a distribution, the parameters of which will be assigned a prior but, again, don't worry about what any of those distributions might be, only how we could determine whether there is evidence that a single value of $\theta$, be that $\theta = 1$ or some other value, appears adequate or inadequate to reflect the behavior of the actual data.

## Answer

The parameter $\theta$ determines how the variance scales with the mean through

$$\mathrm{Var}(Y_{i,j} \mid \cdot) = \sigma_i^2 \mu_{i,j}^{2\theta}.$$

Thus, fixing $\theta = 1$ imposes a specific mean–variance relationship, namely that variability is proportional to $\mu_{i,j}^2$. To assess whether this assumption is adequate, I would evaluate whether this scaling produces approximately homoscedastic, Gaussian residuals and whether the data suggest a different or heterogeneous exponent.

### Diagnostic strategy

Using the posterior output from the existing hierarchical fit together with the observed data, I would proceed in three steps.

**1. Residual diagnostics under $\theta = 1$.**
Compute standardized residuals

$$r_{i,j}^{(1)} = \frac{y_{i,j} - \mu_{i,j}}{\sigma_i \mu_{i,j}},$$

which should resemble iid $N(0, 1)$ if the variance scaling is correct. I would examine:

- residuals versus fitted means $\mu_{i,j}$ to check for remaining heteroscedasticity,
- absolute residuals versus $\mu_{i,j}$ to detect systematic changes in spread,
- normal Q–Q plots to assess departures from Gaussianity.

If $\theta = 1$ is appropriate, these plots should show no systematic patterns and roughly constant variance.

**2. Estimating a common exponent.**
To assess whether a single value of $\theta$ other than 1 is more appropriate, I would exploit the implied relationship

$$\log\left(\frac{(y_{i,j} - \mu_{i,j})^2}{\sigma_i^2}\right) = c + 2\theta \log(\mu_{i,j}) + \varepsilon.$$

A regression of the log residual variance on $\log(\mu_{i,j})$ provides an estimate of $2\theta$. If this estimated slope differs substantially from 2, this would indicate that a common $\theta \neq 1$ better describes the data, motivating inclusion of $\theta$ as an unknown parameter with a prior.

**3. Assessing lake-specific heterogeneity.**
To determine whether the scaling differs across lakes, I would repeat the same regression separately for each lake to obtain lake-specific estimates $\theta_i$. If these estimates are similar across lakes, a single common $\theta$ is adequate. If they vary widely or show systematic differences, this would suggest modeling $\{\theta_i\}$ hierarchically.

**Decision criteria**

- If residuals appear homoscedastic and the estimated slope is close to 2, then $\theta = 1$ is adequate.
- If residuals show systematic mean–variance patterns and the estimated slope differs from 2, then a single unknown $\theta$ should be estimated.
- If there is substantial between-lake variability in the estimated exponents, then allowing lake-specific $\theta_i$ may be justified.

In this way, the adequacy of the fixed value $\theta = 1$ can be evaluated using graphical and regression-based diagnostics without specifying any additional priors in advance.