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 health (condition) parameters*, it is better to view this problem as one with a model having several levels of prior distributions, rather than as a Bayesian mixture model.

Following the principles detailed in Chapter 15, the choice of interpretation should be based on *what quantities are scientifically meaningful to estimate*; another way to note this distinction is whether we are more interested in the data itself (multi-level) or more interested in the data-generating mechanism (mixture). Here, the lake-specific parameters β_i represent the health of identifiable lakes, we are interested in those particular lakes, and as such, the scope of our inference is for each of these individual units (lakes). What's more, using the other heuristic to identify the interpretation choice, the full population of lakes is known, we could hypothetically get another observation from any one of these lakes, and of interest to us the population of fish for the lake they come from.

Added 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 as explicitly noted, condition is interpreted as a *persistent lake-level characteristic*.

What's more, the scientific objective is to evaluate the health (condition) of the *particular lakes sampled (the 75 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

Based on my answer to Q1, inference should focus primarily on the *lake-specific condition parameters*, β_i. These represent persistent, interpretable ecological characteristics of identifiable lakes and are the quantities most directly tied to lake/fish management decisions. The population-level hyperparameters (λ, τ^2) primarily are still helpful though, as they provide context by describing between-lake variability and enabling partial pooling for inference, but they are secondary to lake-level effects.

Accordingly, summaries should prioritize posterior inferences for each lake, supplemented by cross-lake comparisons and statewide context (the latter-most of these being overall, non-inference based summaries). Together, these summaries should allow the Iowa DNR to estimate lake condition with quantified uncertainty, identify unusually poor or healthy lakes, understand statewide variability, and possibly even predict the condition of future or other (characteristically similar but unsampled) lakes.

Lake-specific summaries (primary)

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

For each lake i, I'd report:

- Posterior summary table for β_i including mean, median, standard deviation, and particular quantiles (e.g., 0.25 and 0.95).
- 80%, 90%, and 95% credible intervals.
- Biologically interpretable probability statements tied to management thresholds.

More detail on the last point: Because standard-weight benchmarks provide meaningful context, they apply to *individual fish* rather than entire lakes; as such, they should not be treated as a hard cutoff for β_i. Instead, we should consider using the benchmark to summarize the *population within the lake*. One such method is to define:

$$\theta_i = P(W_{ij} < c \mid \text{lake } i),$$

where W_{ij} is an individual fish condition metric and c is a biologically meaningful standard. Then report posterior summaries such as $E(\theta_i \mid y)$, credible intervals for θ_i, or probability statements like

$$P(\theta_i > p_0 \mid y),$$

where p_0 is a management-relevant tolerance (e.g., more than 30% of fish in lake i are below the standard).

Visuals For variety, and perhaps to illustrate the adage "a picture is worth a thousand words" (pun intended), some accompanying graphs/tables could include:

Tables:

- Table of posterior summaries and credible intervals for all β_i (and possibly θ_i as well).

- Ranked table of lakes from lowest to highest condition.
- Posterior ranks or probabilities of being among the worst/best lakes.

Graphs:

- Sorted plot of β_i with credible intervals (caterpillar plot).
- Boxplot or density of $\{\beta_i\}$ across lakes.
- Optionally, a spatial map of posterior lake condition could be included to identify geographic patterns or general condition across lake systems.

At any rate, the goal of these summaries is to allow the Iowa DNR ("biologists") to identify poorly conditioned lakes, prioritize interventions, and compare lakes quantitatively while incorporating estimates of uncertainty.

Population-level summaries (secondary)

To describe statewide patterns and heterogeneity, I'd also report:

- Posterior means and 95% credible intervals for (λ, τ^2) to summarize typical statewide condition (location λ) and lake-to-lake variability (spread τ^2).
- Density or histogram of posterior draws of λ, τ^2, and/or β_i.

The goal of these quantities is primarily context, i.e., to help interpret the strength of pooling and the degree of heterogeneity across lakes, in effect contexting for aforementioned lake-specific results.

Additional Note: Model assessment is typically beyond the scope, but important

The aforementioned inferences are only as good as the model that produced them. In real-world reporting, it's vital, responsible, ethical, and another word I struggle to identify to note model adequacy, even if it's sequestered to an appendix.

Examples of what that would look like include:

- Scatterplots of $Y_{i,j}$ versus $x_{i,j}$ with the fitted mean curve

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

to visually assess whether the assumed functional form captures the dominant signal.

- Posterior predictive checks, obtained by simulating replicate datasets from the fitted model and comparing observed and simulated distributional features such as quantiles, spread, or tail behavior. Care should be taken to examine aspects of the data not automatically enforced by the model fit.
- Residual plots versus fitted values or covariates, which may reveal heteroskedasticity, temporal dependence, or other systematic departures from model assumptions.

Including these materials strengthens confidence in the inferential summaries, graphs, and tables noted previously.

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 θ 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 α), 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 θ.

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 θ, be that $\theta = 1$ or some other value, appears adequate or inadequate to reflect the behavior of the actual data.

Answer

Generally, I would treat the question of θ as a *model adequacy* problem: Does the assumed scaling with $\theta = 1$ already explain the observed mean–variance behavior, or is additional flexibility warranted?

To that end, recall that θ determines how variability changes with the mean:

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

Thus, fixing $\theta = 1$ implies

$$\mathrm{Var}(Y_{i,j} \mid \cdot) \propto \mu_{i,j}^2,$$

so the residual variance should increase quadratically with the fitted mean. Using the existing posterior draws for $\mu_{i,j}$ and σ_i, I would directly check whether this relationship is consistent with the observed data before introducing any different θ parameter value or impose priors on θ.

Overall approach

Practically, I would proceed in stages, beginning with the current model and only adding complexity if diagnostics suggest lack of fit (or improvement in fit for the alternative).

- First assess whether $\theta = 1$ already appears adequate,
- If not, check whether a single (lake-wide) $\theta \neq 1$ captures the pattern,
- And only then consider lake-specific θ_i if there is clear, stable evidence of heterogeneity across lakes.

At each stage, the process would be based on residual behavior and predictive performance rather than formal hypothesis tests. Each step is noted in greater detail below.

Step 1: Model Adequacy for $\theta = 1$

Using posterior means or draws of $\mu_{i,j}$ and σ_i from the existing fit, I would standardize residuals via ("the typical"):

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

If $\theta = 1$ is appropriate, these residuals should behave approximately like iid $N(0,1)$ with no remaining dependence of their spread on $\mu_{i,j}$.

Specifically, I would examine:

- residuals versus fitted means $\mu_{i,j}$,
- absolute residuals versus $\mu_{i,j}$,
- normal Q–Q plots,
- posterior predictive checks comparing simulated and observed variability.

If these diagnostics show roughly constant variance and no systematic structure, then we'd have evidence to support that the fixed choice $\theta = 1$ is adequate.

Step 2: Model Adequacy for a single unknown (lake-wide) $\theta \neq 1$

If Step 1 suggests a systematic mean–variance trend that appears similar across lakes, I would next assess whether a single common exponent $\theta \neq 1$ better describes the observed relationship.

Conceptually, this step plays the same role as a Box–Cox–type diagnostic: However, Rather than transforming the response to stabilize variance, I would estimate the power in the variance model that makes residual variability approximately constant.

Starting from

$$(y_{i,j} - \mu_{i,j})^2 \approx \sigma_i^2 \mu_{i,j}^{2\theta},$$

taking logs gives

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

so the slope of a regression of log residual variance on $\log(\mu_{i,j})$ directly estimates 2θ.

Then, I would pool observations across lakes and run this regression to obtain an empirical estimate

$$\hat{\theta} = \frac{\text{slope}}{2}$$

- If the slope is near 2 (i.e., $\hat{\theta} \approx 1$), I would conclude that the current assumption is adequate.
- If the slope differs materially from 2, I would interpret this as evidence that a single common $\theta \neq 1$ better captures the global mean–variance relationship and would recommend refitting the model with θ treated as an unknown parameter (either using the value taken from the empirical estimate, or with its own prior, after careful consideration).

Afterward, I would recompute standardized residuals and perform posterior predictive checks to verify that the revised scaling meaningfully reduces heteroscedasticity and provides a better fit of the observed.

Step 3: Model Adequacy for lake-specific θ_i

If a single global θ still leaves a systematic lack of fit ($\theta = 1$ or for $\theta \neq 1$) would I consider lake-specific exponents $\{\theta_i\}$.

To explore this, I would repeat the same regression separately within each lake:

$$\log\left(\frac{(y_{i,j} - \mu_{i,j})^2}{\sigma_i^2}\right) = c_i + 2\theta_i \log(\mu_{i,j}) + \varepsilon_{i,j}.$$

The focus would then be on whether the slopes are stable and well identified, by assessing:

- Does each lake have enough data and spread in $\log(\mu_{i,j})$ to estimate a slope?
- Are the uncertainty intervals for $2\theta_i$ reasonable (e.g., are they especially wide, do they tend to be narrow)?
- Do differences appear systematic rather than white noise (random)?

If most estimates are imprecise and largely overlapping, I would prefer to retain a single common θ, since the apparent differences are likely attributable to sampling variability rather than meaningful between-lake heterogeneity. However, if several (ideally a majority of) lakes show stable, well-identified slopes and posterior predictive checks improve when allowing heterogeneous scaling, I would recommend a hierarchical extension such as

$$\theta_i \sim N(\theta, \kappa^2),$$

so that between-lake variability is modeled while still borrowing strength across lakes.

Importantly, I would use these lake-specific estimates only as *diagnostics* to assess whether heterogeneity is present. I would not use the empirical distribution of the $\hat{\theta}_i$'s to construct or tune the prior itself, since that would amount to using the data twice (first to set the prior and again to fit the model). Instead, the prior specification should be chosen independently, with the data informing only the model structure.