Key Insights from Meeting

16 January 2026

David Ewing

16 January 2026

Overview

This document extracts the key insights and decision implications from Dr John Holmes' supervision meeting regarding the variational Bayes project. Each insight is presented with its context and practical implications for the report and implementation.

1 Methodological Insights

Insight 1: Laplace vs VB – Different Approaches to the Same Approximation

[0:00:20] "You're replacing the posterior with a normal... The difference is how you work out the mean and variance of that normal distribution."

Decision implication: Both Laplace and VB use Gaussian approximations, but:

- **Laplace**: Mean = where first derivative = 0 (point estimate)
- **VB**: Mean = where *expected value* of first derivative = 0 (expectation under q)
- Result: Different variances $\rightarrow$ different uncertainty estimates

Insight 2: Laplace Has Maximum Under-Dispersion

[0:01:53] "Laplace has the highest density at the peak... it's the one that's most heavily underestimating the true uncertainty."

Decision implication:

- Laplace is *more spiked* than VB
- Because it uses point estimates for derivatives (no averaging)
- **Decision: Skip Laplace in your report** – not needed for the hierarchical model comparison

Insight 3: Focus on Variance, Not Modes

[0:04:10] "What you're more interested in is showing the changes in the posterior variance compared to the Gibbs sampling... focus is much more on the variance."

Decision implication:

- VB gets modes approximately correct
- VB gets *shape* correct (Gammas look like Gammas)
- **The problem is the spread (variance) – that's what to demonstrate**
- This is the core pedagogical message

Insight 4: Gibbs is the Gold Standard

[0:06:36] "That's why you're having to run these Gibbs samples alongside it... if you take enough samples, you will get draws from the posterior distribution."

Decision implication:

- Gibbs (MCMC) = ground truth when no analytical posterior exists
- Slow but accurate
- **Decision: Use HMC/Stan as baseline for all comparisons**

2 Variance Components – The Core Challenge

Insight 5: Variance Components Are Hardest

[0:07:37] "The thing that's hardest to do in this model is the variance components. It's not the betas and the u's. It's the variance components (τ_u, τ_e)."

Decision implication:

- βs and u's: VB does fine
- **Precisions/variances: VB struggles significantly**
- This is *especially* true with small sample sizes
- **Core finding for your report**

Insight 6: Experimental Design – Vary Group Sizes, Fix Total N

[0:07:37] "Do this model with multiple groups, changing the group sizes, but leaving the total number of observations unchanged."

Decision implication:

- **Design decision:** Keep N constant (e.g., $N = 500$)
- Vary number of groups (Q): 6, 10, 15, 25
- This changes observations per group level
- Controls for "more data = better results" criticism
- **This is your Model 3 experimental design**

Insight 7: Shrinkage – The Core VB Weakness

[0:09:57] "The more times you see the same random effect level, the better you will do at estimating the variance components... variational Bayes doesn't take shrinkage into account."

Decision implication:

- Few observations per level $\rightarrow$ heavy shrinkage in Bayesian estimation
- **VB's approximate posteriors don't capture this shrinkage**
- More observations per level $\rightarrow$ less shrinkage $\rightarrow$ VB does better
- **This explains the relationship between sample size and VB performance**

3 Mathematical Understanding

Insight 8: ELBO Measures Distribution Distance

[0:10:29] "The ELBO is measuring... minimised distance between two distributions... that log distribution has information about both the mean and variance."

Decision implication:

- ELBO isn't just about parameter means
- It's about the entire distribution (all moments)
- Optimising ELBO = making $q(\theta)$ close to $p(\theta|y)$
- But mean-field approximation breaks dependencies

Insight 9: Conditional vs Unconditional Posteriors

[0:10:29] "In Gibbs, you work out conditional posterior distributions. In variational Bayes these are your independent posterior distributions... The conditional posterior doesn't take into account the shrinkage."

Decision implication:

- Gibbs: Cycles through $p(\beta|u, \tau, y)$, $p(u|\beta, \tau, y)$, $p(\tau|\beta, u, y) \rightarrow$ captures dependencies
- VB: Assumes $q(\beta)q(u)q(\tau) \rightarrow$ **ignores** how estimating β and u affects degrees of freedom for τ
- **This is the mathematical reason VB underestimates variance of τ_u**

Insight 10: Focus on Sample Size Effects

[0:13:50] "Emphasise that this sort of model, variational Bayes approximations will do a lot better with larger sample sizes."

Decision implication:

- **Pedagogical message:** VB isn't universally bad
- With enough data (large n per group), VB converges to truth
- **Frame VB as data-hungry, not fundamentally broken**

4 Diagnostic Metrics

Insight 11: Diagnostic Ratio – Posterior Var / Prior Var of u's

[0:15:49] "Plot the posterior variance of the u's over the prior variance... When you do badly with τ_u, this ratio will be high. When you do well, this ratio will be low."

Decision implication:

- **New diagnostic metric:** $\text{Var}_{\text{posterior}}(u_i)/\text{Var}_{\text{prior}}(u_i)$
- Ratio $\rightarrow 0$ means posterior is narrow (lots of information learnt)
- **Narrow posteriors for $u \rightarrow$ accurate τ_u estimation in VB**
- This explains the mechanism: VB succeeds when posterior concentrates

Insight 12: The Mathematical Connection

[0:17:42] "The narrower the posterior distributions for the u's are, the better you will do at getting the posterior for τ_u."

Decision implication:

- Small sample per group $\rightarrow$ wide posterior for $u_i \rightarrow$ bad τ_u estimate
- Large sample per group $\rightarrow$ narrow posterior for $u_i \rightarrow$ good τ_u estimate
- **This is because:** VB doesn't account for lost degrees of freedom when estimating u's
- Mathematical: $q(\tau_u)$ has $Q/2$ in shape parameter (assumes all Q levels independent)

5 Practical Implementation Decisions

Insight 13: Speed vs Accuracy Trade-off

[0:19:39] "The intention of variational Bayes is it goes fast... there's a trade-off there... but it's still going to always be faster than MCMC."

Decision implication:

- Larger samples help VB accuracy
- Larger samples slow down both VB and MCMC
- **But VB remains computationally advantageous**
- Question becomes: "How close do you need to be?"

Insight 14: Blocking Strategy Matters

[0:21:52] "There are two ways to mess up variational Bayes. One is to ignore interdependencies... if every single scalar parameter was given its own approximate posterior, that would completely screw it up."

Decision implication:

- Your current code: $q(\beta)q(u)q(\tau_u)q(\tau_e) \leftarrow$ **as blocked as analytically feasible**
- Worse alternative: $q(\beta_1)q(\beta_2)\cdots q(u_1)q(u_2)\cdots \leftarrow$ ignores known correlations
- **Your blocking choice is optimal given analytical constraints**
- Remaining under-dispersion is inherent to VB, not your implementation

Insight 15: β and u Estimation is Fine

[0:22:30] "It probably doesn't matter very much for β and u... once sample size is about 30–50, you can't tell the difference between a t and a normal distribution."

Decision implication:

- **Don't focus report on β estimates** – VB does well here
- βs have enough data (N observations total)
- u's benefit from shrinkage but VB handles them adequately
- **The drama is in τ_u and τ_e** – these are hyper-parameters

Insight 16: Degrees of Freedom Problem

[0:22:30] "Conditional posteriors don't take into consideration the loss of degrees of freedom in estimating some parameters when updating the precisions."

Decision implication:

- True posterior: $N - p$ effective observations after estimating β
- VB posterior for τ: **assumes full** N (doesn't subtract p)
- Few obs per level $\rightarrow$ large proportional loss $\rightarrow$ big error
- **Mathematical root cause of τ_u under-dispersion**

6 Report Structure and Presentation

Insight 17: Experimental Plan for Report

[0:25:24] "This is going to be the example that goes into the report... total number of observations stay unchanged, vary the number of random effect levels."

Decision implication:

- **Report structure:**
  1. Theory section (existing)
  2. Model 3: Random intercept logistic
  3. Show τ_u posterior for $Q = \{5, 10, 20, 50\}$ observations per level
  4. Demonstrate: VB $\rightarrow$ HMC as observations per level increases
  5. Explain via posterior variance ratio diagnostic

Insight 18: $Q/2$ Parameter

[0:26:28] "Our approximate posterior for τ_u has $Q/2$ in it... which is what you get in the conditional posterior, but that means it's not taking out the loss of degrees of freedom."

Decision implication:

- VB: $a_n = a_0 + Q/2$ (number of random effect levels)
- True: Should be smaller to account for estimating β and u
- **This is the smoking gun in your equations**
- Point this out in report derivations

Insight 19: Report Simplification

[0:27:17] "Take out the Exact and the Laplace... it's just variational Bayes versus Gibbs. Make sure you run multiple chains."

Decision implication:

- **Model 1**: Keep Exact (pedagogical calibration)
- **Model 3**: Only VB vs HMC (no Laplace needed)
- Run 4 chains for convergence diagnostics ($\hat{R}$)
- Focus plots on τ_u **distributions** at different Q levels

Insight 20: Don't Vary Priors

[0:28:36] "Don't change the priors, because that doesn't help you illustrate what you want to illustrate, which is variational Bayes tends to struggle with shrinkage."

Decision implication:

- **Fix priors** across all experiments
- Variation in results = VB's handling of data, not prior choice
- Stronger priors would confound the message
- **Core message: VB + hierarchical structure + small group sizes = under-dispersion**

Summary of Key Decisions

1. **Model focus**: Random intercept model (Model 3)
2. **Comparison**: VB vs HMC (skip Laplace)
3. **Experimental design**: Vary Q (group levels), fix N (total observations)
4. **Key metric**: Posterior distributions of τ_u
5. **Diagnostic**: $\text{Var}_{\text{post}}(u)/\text{Var}_{\text{prior}}(u)$ ratio
6. **Mathematical explanation**: $Q/2$ in shape parameter doesn't account for degrees of freedom loss
7. **Pedagogical message**: VB struggles with variance components in hierarchical models when observations per level are small

This is your roadmap for completing the project.