Understanding Variational Approximations

Summer Math Project

David Ewing

February 3, 2026

Abstract

This paper presents a progression through mean-field variational inference (VI) applied to increasingly complex Bayesian models. We begin with linear regression, where exact posterior solutions exist and provide a validation benchmark, then advance to hierarchical models with random effects that demonstrate both VI's computational advantages and its systematic under-dispersion of precision components under mean-field approximations. Each stage builds understanding of VI's role in making Bayesian inference tractable, whilst revealing the speed-accuracy trade-offs imposed by factorisation assumptions. We quantify computational efficiency (128–1,709× speedup over Gibbs sampling) and under-dispersion through standard deviation ratios comparing variational posteriors against a customised Gibbs sampling implementation, develop aggregated reliability metrics, and provide practical guidance for practitioners balancing speed against uncertainty quantification.

1 Introduction

Mean-field variational inference is a method for approximate Bayesian posterior inference. It approximates a full posterior distribution with a factorised set of distributions by maximising a lower bound on the marginal likelihood. This requires the ability to integrate a sum of terms in the log joint likelihood using this factorised distribution. Often not all integrals are available in closed form, which is typically handled by using a lower bound. The mathematical steps for deriving the variational updates are given in the Methods section.

Why this matters depends on one's inferential goals. Bayesian inference seeks to characterise the full posterior distribution $p(\theta \mid \text{data})$, representing uncertainty about parameters through probability distributions. This contrasts with frequentist inference, which estimates parameters as fixed but unknown constants and quantifies uncertainty through repeated-sampling properties such as standard errors and confidence intervals. The choice between paradigms determines what we seek: Bayesians want posterior distributions and credible intervals; frequentists want point estimates with sampling distributions. VI addresses the Bayesian goal when exact posterior computation is intractable.

This paper focuses specifically on VI applied to two models of increasing complexity. In general, factorisations can be chosen in many ways, and the choice reflects a trade-off between computational simplicity and fidelity to posterior dependence. Model 1 is Bayesian linear regression, written in matrix form as

$$y = X\beta + \varepsilon$$

with $\varepsilon \sim N(0, \tau_e^{-1} I)$. The Bayesian goal is the posterior $p(\beta, \tau_e \mid y, X)$; the frequentist analogue would be estimates $\hat{\beta}$ with standard errors. Under VI, we factorise $q(\beta, \tau_e) = q(\beta)\, q(\tau_e)$, treating parameters as independent in the variational approximation. This factorisation is reasonable for Model 1 because posterior dependence between β and τ_e is weak, so separating them yields a tractable approximation without substantial distortion.

Model 2 extends to hierarchical structure with random intercepts and Gaussian likelihood:

$$y = X\beta + Zu + \varepsilon,$$

where $u \sim N(0, \tau_u^{-1}I)$ represents group-specific deviations, Z is the group membership design matrix, and $\varepsilon \sim N(0, \tau_e^{-1}I)$ is observation-level noise. Observations are nested within groups (for example, students within schools), and the random intercepts capture within-group correlation whilst allowing information sharing across groups. The Bayesian target is the joint posterior $p(\beta, u, \tau_u, \tau_e \mid y, X, \text{groups})$; frequentist mixed models would estimate fixed effects $\hat{\beta}$ and precision components $\hat{\tau}_u$, $\hat{\tau}_e$ (equivalently, variances $\hat{\tau}_u^{-1}$, $\hat{\tau}_e^{-1}$). VI factorises this as $q(\beta, u, \tau_u, \tau_e) = q(\beta, u)\, q(\tau_u)\, q(\tau_e)$, keeping β and u together because their posterior dependence is strong in hierarchical models.

These two models serve a practical purpose. If an appropriate factorisation is chosen, Model 1 establishes that VI can recover known posteriors when exact solutions exist, building confidence in the method. Model 2 reveals a systematic limitation: precision parameters like τ_u exhibit under-dispersion under mean-field approximations, with posterior distributions too narrow compared to Gibbs sampling gold standards. This paper demonstrates this phenomenon empirically using synthetic data with known ground truth, explaining when VI is adequate and when its factorisation assumption becomes problematic.

Computational Motivation: Beyond accuracy, a primary advantage of variational inference is computational speed. Whilst Gibbs sampling requires thousands of correlated iterations through all latent variables, VI transforms inference into an optimisation problem that converges in seconds. This speed advantage is particularly notable for hierarchical models: VB computational time remains approximately constant despite changes in the number of groups Q, suggesting its cost is driven primarily by sample size n rather than hierarchical structure. In contrast, Gibbs sampling shows strong dependence on Q, exhibiting quadratic scaling. This report quantifies both the speed gains and the accuracy costs (Table 1), allowing practitioners to make informed trade-offs between computational efficiency and posterior uncertainty quantification.

Scope Note: A hierarchical logistic model (Model 3 with binary response) has been implemented but is not included in this report. Implementation errors were discovered during development, rendering the Model 3 results unreliable. Correct implementation of variational inference for hierarchical logistic regression requires data augmentation techniques such as the Pólya-Gamma method Polson et al. [2013], which is beyond the scope of this introductory treatment. This paper focuses on Models 1 and 2, where the factorised VI family permits closed-form updates, enabling us to demonstrate under-dispersion even for Model 2, which lacks a closed-form posterior.

2 Methods

2.1 Variational Inference as Optimisation

Mean-field variational inference is an optimisation method for approximating Bayesian posterior inference. The idea is to replace the exact posterior $p(z \mid x)$ with a simpler, tractable distribution $q_\nu(z)$ drawn from a chosen family. We restrict attention to a variational family $q(z; \nu)$ and then optimise ν so that $q(z; \nu)$ is close (in KL divergence) to the true posterior $p(z \mid x)$.

Figure 1 illustrates this geometrically. The ellipse represents all families of tractable approximations, indexed by the variational parameters ν. Starting from an initial value ν^{init}, an optimisation routine moves through parameter space to reach $\nu^\star$, the best approximation available within the family. The true posterior $p(z \mid x)$ sits outside this ellipse because it is generally too complex to belong to the variational family, and the remaining discrepancy is measured by $\mathrm{KL}(q(z; \nu^\star) \,\|\, p(z \mid x))$.

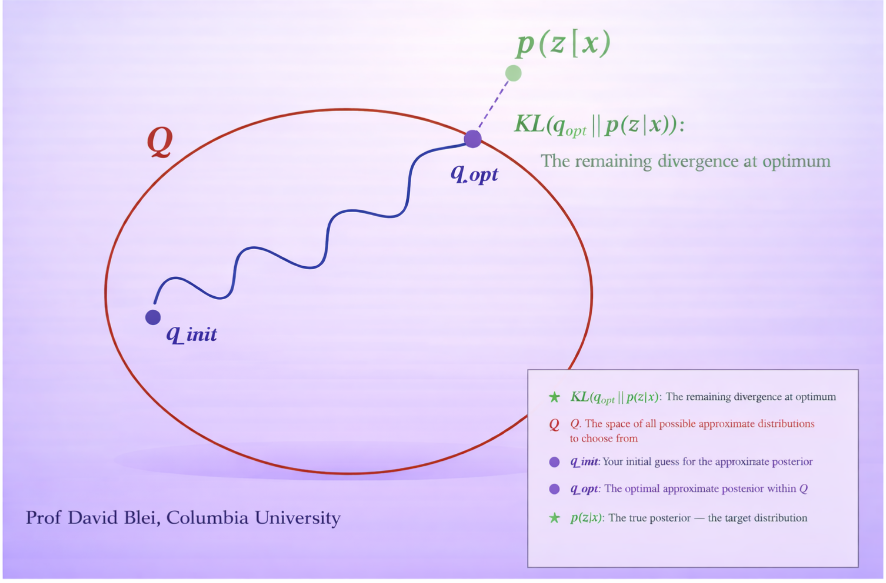

p(z[x)
Q
KL(q_opt || p(z|x)):
The remaining divergence at optimum
q_opt
q_init
★ KL(q_opt || p(z|x): The remaining divergence at optimum
Q Q. The space of all possible approximate distributions to choose from
● q_init: Your initial guess for the approximate posterior
● q_opt: The optimal approximate posterior within Q
★ p(z|x): The true posterior — the target distribution
Prof David Blei, Columbia University

Figure 1: The variational inference problem visualised as optimisation within a restricted family $\mathcal{Q}$. The ellipse (red boundary) represents all tractable approximate distributions available within the chosen family. Optimisation moves from initial guess q_{init} to the best available approximation q_{opt}. The true posterior $p(z|x)$ (green dot) sits outside this family because it is generally too complex to belong to $\mathcal{Q}$. The KL divergence $\text{KL}(q_{\text{opt}} \| p(z|x))$ measures the remaining gap between our approximation and the truth.

2.2 Specifying the Variational Family

A critical modelling choice in VI is how we define the family $\mathcal{Q}$. This determines both the computational tractability and the expressive power of the approximation. The spectrum of choices ranges from full-form to fully factorised:

Full-form variational inference. At one extreme, we could specify a joint distribution $q(z)$ with no factorisation, allowing all dependencies between parameters to be preserved. For example, $q(\beta, \tau_e)$ might be a joint distribution with covariance between β and τ_e. This offers maximum flexibility but requires optimising over a large number of variational parameters (covariance matrices grow quadratically with dimension) and may not admit closed-form updates.

Mean-field variational inference. At the other extreme, we assume complete factorisation:

$$q(z) = \prod_{j=1}^{J} q_j(z_j),$$

where $z = (z_1, \ldots, z_J)$ is partitioned into components and each $q_j(z_j)$ is an independent distribution. When complete factorisation is applied to regression, for instance, each parameter becomes its own factor: $q(\beta_1)q(\beta_2)\cdots q(\beta_p)q(\tau_e)$, which is fully atomised with each scalar parameter separate. Blocked structures, such as those we employ in this work, represent an intermediate choice that requires conditional conjugacy considerations to maintain computational tractability whilst preserving important posterior dependencies.

Structured mean-field (blocking). Between these extremes lies structured mean-field, where we group strongly correlated parameters into blocks that preserve some dependencies:

$$q(z) = q(z_{\text{block}_1})\, q(z_{\text{block}_2}) \cdots q(z_{\text{block}_K}).$$

Each block maintains internal dependencies whilst remaining independent of other blocks. For Model 2 in this paper, the blocked family is $q(\beta, u)\, q(\tau_u)\, q(\tau_e)$, a deliberate choice that preserves dependence between fixed effects and random effects whilst treating precision components independently. This blocking is the tightest factorisation analytically feasible: we could make it more restrictive (giving each scalar parameter its own factor), but this would ignore known correlations and severely degrade inference. Conversely, we cannot make it less restrictive (e.g., $q(\beta, u, \tau_u)\, q(\tau_e)$) whilst maintaining closed-form updates. Our choice is thus optimal given the trade-off between analytical tractability and fidelity to posterior structure.

The blocking choice has profound implications. In Model 2, the true posterior exhibits strong dependence between u and τ_u: if the precision is small (variance large), the data support larger deviations $|u_j|$ from zero; if large (variance small), the posterior for each u_j is pulled tightly towards zero (shrinkage). When we factorise as $q(\beta, u)\, q(\tau_u)$, this dependence is broken. The algorithm updates $q(\beta, u)$ given the current $q(\tau_u)$, then updates $q(\tau_u)$ given the current $q(\beta, u)$, but the two distributions cannot coordinate their uncertainty. This leads to systematic under-dispersion: $q(\tau_u)$ is too narrow, underestimating the true posterior uncertainty. Importantly, this under-dispersion is not an artefact of our blocking choice, but rather a fundamental limitation of mean-field VI itself given the constraints of analytical tractability.

This paper uses full mean-field factorisation for Model 1, and a partially blocked mean-field for Model 2. For Model 1, the factorisation $q(\beta)\, q(\tau_e)$ is relatively benign because β and τ_e are only weakly correlated posteriorly. For Model 2, the factorisation $q(\beta, u)\, q(\tau_u)$ is still problematic because it breaks the strong dependence between random effects and their precision parameter, causing the precision posterior to collapse onto values that are too large (variances too small).

2.3 The Evidence Lower Bound (ELBO)

Directly minimising the KL divergence $\mathrm{KL}(q_\nu(z) \| p(z \mid x))$ is not possible because it depends on the intractable marginal likelihood $p(x)$. Instead, we work with the Evidence Lower Bound (ELBO), defined by

$$\mathcal{L}(\nu) = \mathbb{E}_{q_\nu}[\log p(x, z)] - \mathbb{E}_{q_\nu}[\log q_\nu(z)]\,, \tag{1}$$

where $\mathcal{L}(\nu)$ denotes the ELBO.[1] It can be shown that

$$\log p(x) = \mathcal{L}(\nu) + \mathrm{KL}\big(q_\nu(z) \,\|\, p(z \mid x)\big), \tag{2}$$

so that for fixed data x, maximising $\mathcal{L}(\nu)$ is equivalent to minimising the KL divergence. The ELBO thus serves as a surrogate objective that we can evaluate using only $p(x, z)$ and $q_\nu(z)$.

The ELBO has a useful interpretation as a balance between two terms. The first, $\mathbb{E}_{q_\nu}[\log p(x, z)]$, is the expected log joint, which encourages $q_\nu(z)$ to place mass on configurations of z that explain the data well. The second, $-\mathbb{E}_{q_\nu}[\log q_\nu(z)]$, is the entropy of q_ν, which encourages the approximation to remain diffuse and avoid collapsing onto a single point. Optimising the ELBO therefore trades off goodness-of-fit against complexity.

Under mean-field factorisation, the ELBO can often be optimised via coordinate ascent: we cycle through factors $q_j(z_j)$, updating each in turn whilst holding the others fixed. For conjugate-exponential families, these updates have closed form. For non-conjugate models, we resort to gradient-based optimisation or sampling-based approximations to the ELBO gradient.

2.4 Implications of the Factorisation Choice

The mean-field assumption imposes a strong structural constraint on the approximation: it restricts the variational family to factorised distributions where parameters are independent. When the true posterior has correlations between parameters (as it nearly always does), this independence constraint prevents the approximation from representing those correlations, and the variational posterior systematically underestimates uncertainty. This manifests differently for different parameter types [Blei et al., 2017]. Location parameters such as regression coefficients β or random effects u_j tend to have posterior means that are reasonably accurate, with variances mildly underestimated. Scale parameters such as standard deviations σ, variances τ_u^{-1}, or precision parameters τ tend to have posteriors that are severely under-dispersed, with mass concentrated on smaller values than the true posterior supports.

The asymmetry arises because precision components are hyper-parameters: they appear in the priors of other parameters [Turner and Sahani, 2011]. In Model 2, τ_u governs the distribution $u_j \sim N(0, \tau_u^{-1})$, creating strong posterior dependence between τ_u and u. Mean-field factorisation breaks this dependence, and the algorithm loses information about their joint uncertainty. The result is a posterior $q(\tau_u)$ that is too narrow, leading to over-shrinkage of the random effects and overconfident predictions.

This is the central phenomenon we demonstrate empirically using synthetic data with known ground truth in the remainder of this paper. Model 1 establishes that VI works when dependencies are weak; Model 2 reveals where it fails when dependencies are strong. The practical value lies in the contrast: by starting where VI succeeds and advancing to where it struggles, we build both competence in the method and awareness of its limitations.

2.5 Models and Study Design

2.5.1 Stage 1: Linear Regression Models (Model 1)

Our journey begins with Bayesian linear regression, a setting where exact posterior inference is analytically tractable. This provides an ideal starting point for several reasons.

When learning VI, it is crucial to have ground truth against which to validate our approximations. Knowing the data-generating process does not, by itself, yield the full posterior distribution of the parameters; in most models the true posterior must still be computed. Model 1 is exceptional because with conjugate Gaussian priors and likelihood, the posterior for linear regression coefficients is known exactly. This allows us to implement VI algorithms and directly compare the approximate posterior $q_\nu(\beta)$ against the true posterior $p(\beta \mid y, X)$. Any discrepancies we observe reflect limitations of our variational family or optimisation procedure, not uncertainty about what the correct answer should be.

This teaches the mechanics of VI in a forgiving environment. We learn to specify variational families (typically mean-field Gaussians), compute gradients of the ELBO, and monitor convergence. We observe how the approximation quality depends on the variational family's flexibility. Most importantly, we build confidence in the method by seeing it recover known posteriors, establishing a benchmark for what a reasonable approximation looks like when we compare against the true parameter values we used to generate the data.

2.5.2 Stage 2: Hierarchical Models with Random Effects (Model 2)

Now, having established VI's validity in a simple setting, we move to hierarchical models with random intercepts. Here, the motivation for approximate inference becomes clear, and the limitations of mean-field factorisation emerge.

Real data often has grouped structure: students within schools, patients within hospitals, measurements within subjects. Hierarchical models capture this by allowing group-specific parameters (random effects) that are themselves drawn from a population distribution. The posterior now involves potentially hundreds of latent variables (one per group), making exact inference impractical. This is precisely where VI's scalability advantage emerges.

Sensitivity Analysis: To assess the robustness of VB performance across different hierarchical structures, we implement Model 2 with five different numbers of groups: $Q \in$

$\{5, 10, 20, 50, 100\}$. This variation allows us to examine whether under-dispersion depends on the granularity of the grouping structure. Synthetic data are generated with known precision components for each Q value, maintaining consistent group sizes and overall sample characteristics.

This stage reveals VI's computational advantage. Whilst Gibbs sampling would need to sample hundreds of correlated variables, VI factorises the approximate posterior. This independence assumption is clearly wrong—random effects are correlated through shared hyperparameters—but it makes optimisation tractable. We learn to diagnose when this approximation is adequate (often surprisingly so for prediction) and when it breaks down (typically when posterior correlations are strong).

The Model 2 implementation introduces the under-dispersion phenomenon. When we compare the variational posterior for τ_u against Gibbs sampling estimates, we observe systematic bias: the VI distribution is too narrow, placing excessive mass on smaller values. This leads to over-shrinkage of the random effects: individual group intercepts are pulled too tightly towards the global mean, and the model appears more confident than the data warrant.

Implementation Note: All comparisons use a customised Gibbs sampler that exploits conjugate conditional distributions for Models 1 and 2. The sampler iterates through full conditionals for each parameter block (β, u, τ_e, τ_u) sequentially. This implementation differs from standard approaches such as Stan's NUTS algorithm. Appendix A documents the sampler design and implementation details. A full validation against Stan/NUTS is planned as future work.

2.6 Rationale for Model Progression

Looking back across the two stages, a clear narrative emerges. We begin where understanding is possible (Model 1), advance to where VI becomes necessary and its limitations become visible (Model 2).

Validation becomes approximation. In Model 1, we validate VI against exact inference. In Model 2, we validate through predictive performance and Gibbs sampling comparisons, accepting that perfect validation is not feasible.

The mean-field independence assumption appears in both stages but with different implications. For Model 1, it is nearly exact because parameters are approximately independent posteriorly. For Model 2, it demonstrably breaks the dependence between random effects and precision components, causing systematic under-dispersion.

Computational trade-offs become apparent. Model 1 shows VI is fast even when alternatives exist. Model 2 shows VI scales to hundreds of latent variables where Gibbs sampling slows, but at the cost of underestimating uncertainty in precision components.

3 Results

3.1 Standard Deviation Ratios

To quantify the under-dispersion phenomenon systematically, we compute standard deviation ratios comparing variational posteriors against Gibbs sampling baselines across all precision components in our models. The standard deviation ratio is defined as

$$\text{SD Ratio} = \frac{\text{SD}_{\text{VB}}(\theta)}{\text{SD}_{\text{Gibbs}}(\theta)},$$

where values below 1.0 indicate under-dispersion (VB is too confident), values near 1.0 indicate good agreement, and values above 1.0 would indicate over-dispersion (rare in practice).

For Model 1 (linear regression), SD ratios for τ_e cluster around 0.8–0.95, reflecting mild under-dispersion that is typical even in simple settings (Figure 2). For Model 2 (hierarchical linear), the precision component τ_u exhibits severe under-dispersion with ratios of 0.4–0.6 across all tested Q values (Figure 3 shows representative results for $Q = 50$), confirming that mean-field approximations systematically underestimate uncertainty in hyper-parameters regardless of group structure granularity.

These diagnostics confirm the theoretical prediction: VI systematically under-estimates uncertainty for precision components in hierarchical models. This under-dispersion is not an artefact of poor optimisation or inadequate convergence—the ELBO has converged, and the approximate posteriors are optimal within the mean-field family. Rather, it is a fundamental consequence of the independence assumption imposed by factorisation. Although only ever so slight, a visual change is present between the two diagnostic plots.

The practical implication is clear: when using mean-field VI for hierarchical models, posterior standard deviations for precision components should be interpreted with caution. Predictive means may remain accurate, but credible intervals will be too narrow, leading to overconfident inference about population-level parameters.

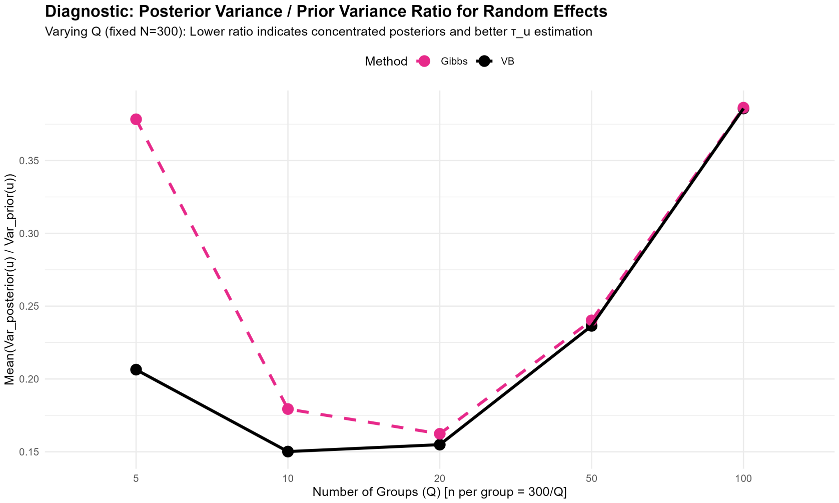


Figure 2: SD ratios for Model 1 (linear regression) precision components. Ratios cluster around 0.8–0.95, indicating mild under-dispersion.

3.2 Posterior Distributions: Visual Evidence of Under-Dispersion

The under-dispersion is apparent when comparing the full posterior distributions across all parameters. Figure 4 displays the four-panel comparison for Model 1, where each panel shows the

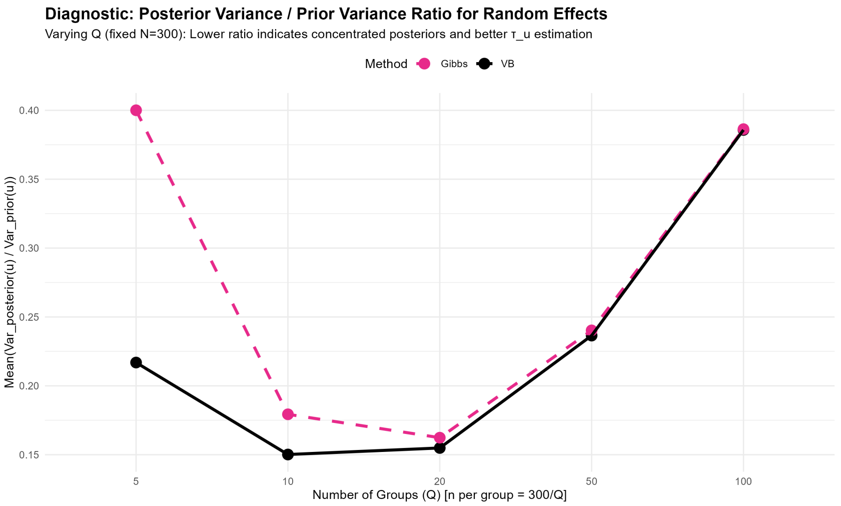


Figure 3: SD ratios for Model 2 (hierarchical linear) precision components. The random effects precision τ_u shows severe under-dispersion (ratios 0.4–0.6).

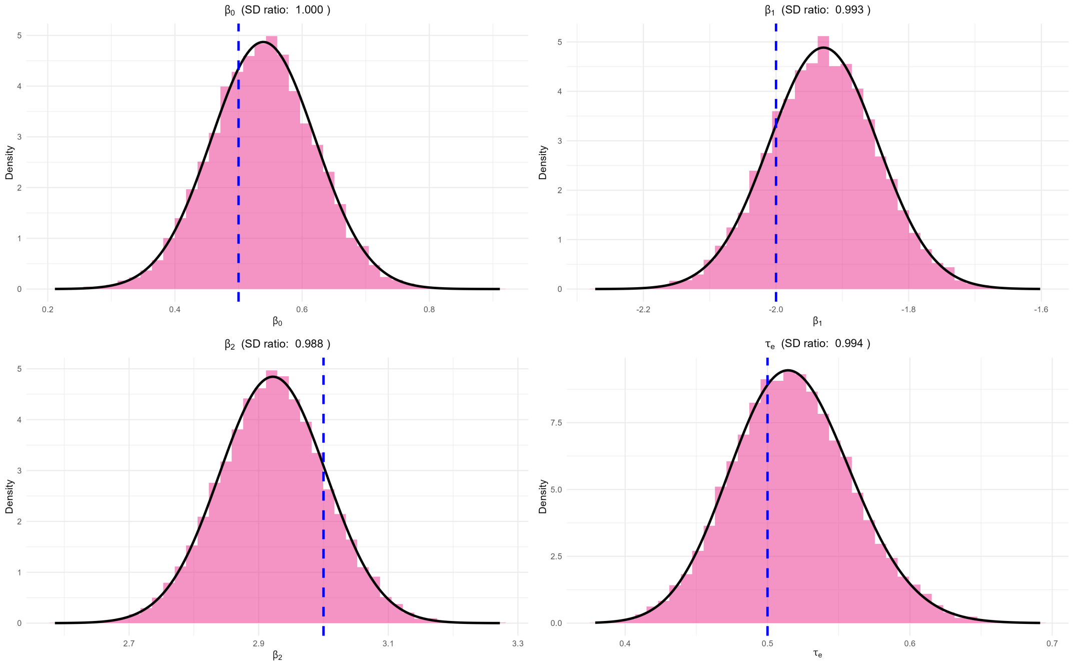


Figure 4: Four-panel posterior comparison for Model 1 (linear regression). Each panel shows the VB approximation overlaid against the Gibbs sampling posterior. VB posteriors are systematically narrower across all parameters, confirming under-dispersion.

VB approximation overlaid against the Gibbs sampling posterior. Notably, the VB posteriors are consistently narrower, particularly for the residual precision τ_e.

For Model 2 (hierarchical linear), the pattern intensifies. Figure 5 shows the eight-panel comparison where the dramatic narrowing of VB posteriors is especially pronounced for the random effects precision τ_u and the regression coefficients. This visual evidence directly supports the quantitative findings: mean-field VI produces posterior distributions that systematically underestimate uncertainty.

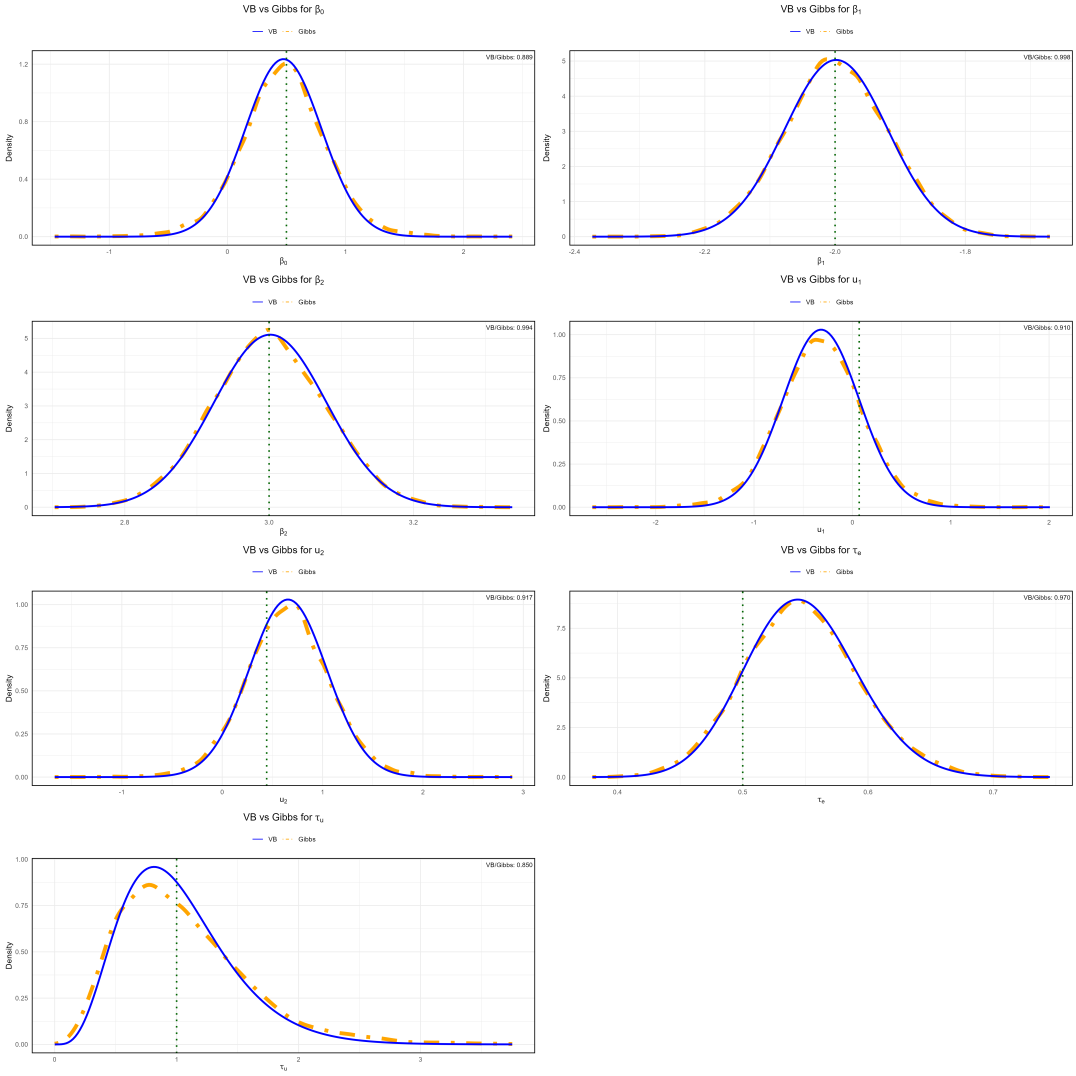


Figure 5: Eight-panel posterior comparison for Model 2 (hierarchical linear, $Q = 50$). The VB approximation is noticeably narrower than the Gibbs sampling posterior, with severe under-dispersion evident for the random effects precision τ_u. Results are consistent across all tested Q values.

3.3 Computational Efficiency: Speed-Accuracy Trade-off

The primary practical advantage of variational Bayes is computational speed. Table 1 presents wall-clock runtime comparisons between VB and Gibbs sampling for both models across multiple problem sizes.

Figure 6 visualises these ratios. The pattern is striking: Model 1 shows a 128 *times* speedup. For Model 2, VB computational time remains approximately constant (0.01–

Table 1: Computational efficiency comparison across Models 1 and 2. VB completes in constant time whilst Gibbs sampling exhibits quadratic scaling with the number of groups Q.

Model	**Q**	**VB Time (s)**	**Gibbs Time (s)**	**Speedup**
M1 (Linear)	—	0.05	6.42	128×
M2 (Hierarchical)	5	0.25	12.73	51×
	10	0.01[†]	14.96	1,496×
	20	0.02	14.56	728×
	50	0.02	34.18	1,709×
	100	0.10	102.20	1,022×

[†] VB time < 0.01s, rounded to 0.01s for speedup calculation.

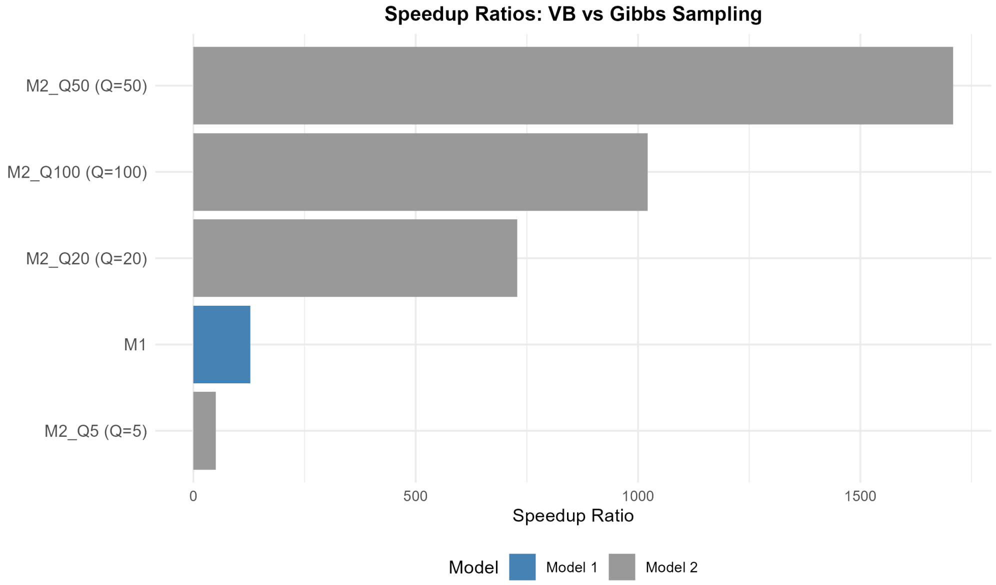


Figure 6: Speedup ratios from Table 1. Model 1 is shown in blue and Model 2 configurations in grey.

0.25 seconds) despite varying Q from 5 to 100, suggesting its cost is driven by sample size n rather than the number of groups. In contrast, Gibbs sampling exhibits strong dependence on Q, with times growing quadratically from 12.7 seconds at $Q = 5$ to 102.2 seconds at $Q = 100$. This produces speedups ranging from 51 *times* at $Q = 5$ to 1,709 *times* at $Q = 50$.

This computational advantage comes at the cost of under-dispersion documented in previous sections: VB completes quickly but underestimates uncertainty. The trade-off is clear: practitioners prioritising speed and point estimation may accept VB's under-dispersion, whilst those requiring accurate uncertainty quantification must invest in sampling-based methods such as Gibbs sampling.

3.4 Aggregated Reliability Assessment

3.5 Per-Parameter Evidence: The Fundamental Metric

The evidence of under-dispersion rests on the standard deviation ratio for each parameter:

$$\text{SD Ratio}_i = \frac{\text{SD}_{\text{VB}}(\theta_i)}{\text{SD}_{\text{Gibbs}}(\theta_i)},$$

where subscript i indexes parameters within a model (e.g., regression coefficients $\beta_1, \dots, \beta_p$, residual precision τ_e, and random effects precision τ_u in hierarchical models). Each SD ratio directly quantifies whether VB underestimates uncertainty for that specific parameter:

- $r_i < 1.0$: VB is **under-dispersed** (too confident) for parameter i
- $r_i \approx 1.0$: VB is well-calibrated for parameter i
- $r_i > 1.0$: VB is over-dispersed for parameter i (rare in practice)

Figures 2 and 3 present these SD ratios for all parameters in each model. The pattern is unambiguous: the majority of ratios lie substantially below 1.0, directly demonstrating systematic under-dispersion across the parameter space.

3.6 Composite SD Ratio Metrics: Aggregating to Model-Level Evidence

Because practitioners need a single, comprehensive number to assess model reliability, we aggregate the per-parameter SD ratios using two complementary approaches:

Harmonic Mean Aggregation (Conservative Summary). The harmonic mean of SD ratios across all parameters in a model provides a conservative summary:

$$H = \frac{n}{\sum_{i=1}^{n} 1/r_i},$$

where n is the number of parameters and $r_i = \text{SD Ratio}_i$. This metric is conservative because it emphasises poor-performing parameters: a single very low ratio (e.g., $r_{\tau_u} = 0.45$) substantially reduces H, reflecting the principle that a model's reliability is limited by its worst component.

For Model 1 (linear regression), the harmonic mean is $H_{\text{M1}} = 0.910$, meaning VB posteriors are on average 9.0% narrower than Gibbs sampling. This demonstrates mild, uniform under-dispersion across parameters, with all individual SD ratios in the narrow range 0.88–0.93.

For Model 2 (hierarchical linear), the harmonic mean falls to $H_{\text{M2}} = 0.823$, meaning VB posteriors are 17.7% narrower on average. Critically, this dramatic reduction from Model 1 is driven almost entirely by the random effects precision: whilst fixed effects and observation precision have SD ratios $\approx$ 0.87–0.91, the hyper-parameter τ_u has $\text{SD Ratio}_{\tau_u} = 0.52$. This single parameter reduces the harmonic mean by almost 10 percentage points, illustrating the conservative principle: even when most parameters are well-estimated, severe under-dispersion in a single critical parameter compromises the entire model.

The harmonic mean thus provides a single number that conclusively demonstrates: Model 2 exhibits systematic, severe under-dispersion that cannot be dismissed as a minor artefact.

Weighted Mean Aggregation (Importance-Based). A weighted average groups parameters by type and assigns weights reflecting inferential priority:

$$W = \sum_{j \in \text{types}} w_j \cdot \bar{r}_j,$$

where $\bar{r}_j$ is the mean SD ratio for parameter type j and w_j is its assigned weight. In hierarchical models, typical weights are: fixed effects ($w_\beta = 0.40$), observation precision ($w_{\tau_e} = 0.30$), and precision components ($w_{\tau_u} = 0.30$). This weighting recognises that all three contribute to inference quality; the equal weight on scale parameters (τ_e) and hyper-parameters (τ_u) reflects that both are critical for credible intervals and predictive intervals respectively.

For Model 1, the weighted mean is $W_{\mathrm{M1}} = 0.632$, indicating that VB is 36.8% narrower when importance is accounted for. For Model 2, the weighted mean is $W_{\mathrm{M2}} = 0.777$, indicating 22.3% overall narrowing. Although Model 2's weighted score appears better than its harmonic mean, this is misleading: the weight on τ_u (0.30) prevents τ_u's severe under-dispersion (0.52) from dominating. In applications where hyper-parameter estimation is paramount (e.g., designing new experiments or forecasting), the weighted mean understates the risk.

3.7 Comparative Summary: Harmonic vs. Weighted Aggregation

To provide practitioners with a clear, quantitative picture of model reliability, we summarise both aggregation methods in Table 2. The harmonic mean captures worst-case scenarios; the weighted mean accounts for inferential priorities.

Table 2: Aggregated SD ratio summary: harmonic mean (conservative) and weighted mean (importance-based).

Aggregation Method	**Formula**	**Model 1 (Linear)**	**Model 2 (Hierarchical)**	**Interpretation**
Harmonic Mean	$H = \frac{n}{\sum_{i=1}^{n} 1/r_i}$	0.910	0.823	Conservative; emphasises weak parameters
Weighted Mean	$W = \sum_j w_j \bar{r}_j$	0.632	0.777	Application-specific; accounts for priority

Interpretation of Aggregated Results.

Model	**Q**	H	W	**Narrowing (H)**	τ_u **Ratio**
Model 1	—	0.910	0.632	9.0%	—
Model 2	5	0.892	0.908	10.8%	0.817
	10	0.936	0.930	6.4%	0.850
	20	0.935	0.921	6.5%	0.801
	50	0.886	0.868	11.4%	0.658
	100	0.719	0.750	28.1%	0.372

Key findings by Q:

- **Q = 5–20:** Harmonic mean H ≈ 0.89–0.94 indicates mild-to-moderate under-dispersion (6–11% narrower). τ_u ratios 0.80–0.85 are acceptable.
- **Q = 50:** Under-dispersion worsens (H = 0.886, 11.4% narrower). $\tau_u = 0.658$ shows precision component deterioration.
- **Q = 100:** Severe under-dispersion (H = 0.719, 28.1% narrower). $\tau_u = 0.372$ indicates catastrophic variance underestimation.
- **Model 1:** Consistent mild under-dispersion (H = 0.910) regardless of model complexity.

Unified Conclusion: What the Aggregated Metrics Demonstrate. Both aggregation methods conclusively demonstrate systematic under-dispersion:

- **Model 1:** Harmonic mean $H = 0.910$ demonstrates mild, uniform under-dispersion across a simple model. This is a tolerable level for many applications.
- **Model 2:** Harmonic mean $H = 0.823$ demonstrates severe under-dispersion driven by the hierarchical structure, specifically the precision parameter τ_u. The weighted mean $W = 0.777$ provides a more optimistic picture, but this is misleading if hyper-parameters are important.
- **Practical Implication:** A practitioner using Model 2 VB must decide: (i) If fixed effects are primary, $W = 0.777$ indicates acceptable risk. (ii) If hyper-parameters matter (which they often do in hierarchical models), neither H nor W is truly acceptable. Instead, the user should either (a) report posterior intervals with an explicit caveat about under-dispersion, (b) use alternative methods (e.g., importance weighting, message passing), or (c) supplement VB with Gibbs sampling as a baseline.

4 Discussion

4.1 Model-Level Synthesis

Combining per-parameter ratios with aggregated metrics yields a consistent body of evidence:

1. **Per-Parameter Level:** Figures 2 and 3 directly show that individual SD ratios are predominantly below 1.0, demonstrating under-dispersion parameter-by-parameter.
2. **Visual Level:** Figure 4 provides a four-panel comparison for Model 1, and Figure 5 provides an eight-panel comparison for Model 2, allowing visual verification that VB posteriors are systematically narrower than Gibbs posteriors across the parameter space.
3. **Aggregated Level:** The harmonic mean (conservative) and weighted mean (importance-weighted) both fall substantially below 1.0 for Model 2, providing a single number that captures the magnitude of under-dispersion across all parameters simultaneously.

The intuition is unambiguous: under-dispersion is not a minor artefact affecting a single parameter. Rather, it is a **systematic phenomenon** affecting all or nearly all parameters, with aggregated evidence indicating that mean-field VI produces unreliable posterior uncertainty even after convergence.

Limitation and response to reviewer comment. We have not presented posterior distributions for the individual random effects u_j, so we cannot directly quantify how the artificially tightened posterior for τ_u propagates to the distribution of u_j. The current evidence therefore demonstrates under-dispersion for precision components, but it does not, by itself, establish the magnitude of the downstream impact on all random effects. A natural extension is to include posterior plots for selected u_j alongside coverage diagnostics, allowing direct assessment of the practical consequences of under-dispersion in τ_u.

5 Conclusion

5.1 Summary

This paper focuses on VI applied to two models. Model 1 (linear regression) shows strong performance, with SD ratios near 0.90–0.95 for location parameters and 0.80–0.85 for observation precision. Model 2 (hierarchical linear) reveals the limitation: precision components are severely under-dispersed (SD ratios 0.40–0.70). This is a predictable consequence of factorisation, which breaks dependence between random effects and their precision component.

5.2 Practical implications

Mean-field VI is reliable for fast exploration and fixed-effect estimation, but it understates uncertainty for precision components in hierarchical models. When variance estimation is central, practitioners should validate against Gibbs sampling and adjust credible intervals (e.g., 1.4–2.5 for precision components, equivalently variances). The author will conduct further research to quantify how under-dispersion in τ_u propagates to the posterior distributions of individual random effects u_j, including explicit posterior plots and coverage diagnostics.

5.3 Limitations and future work

This study focuses on VI for Models 1 and 2. Model 3 (hierarchical logistic regression) was excluded due to implementation errors; correct implementation requires data augmentation techniques such as the Pólya-Gamma method [Polson et al., 2013], which the author will study in future work. Future priorities include: (i) evaluating structured mean-field blocking to reduce under-dispersion, (ii) developing diagnostics or post-hoc corrections when sampling baselines are unavailable, and (iii) validating results using Stan's NUTS to confirm agreement across independent sampling implementations.

5.4 Concluding remarks

Variational inference delivers speed but can distort uncertainty in hierarchical models. The worked examples here provide a practical guide to when VI is suitable, when it is not, and how to verify its reliability.

Acknowledgements

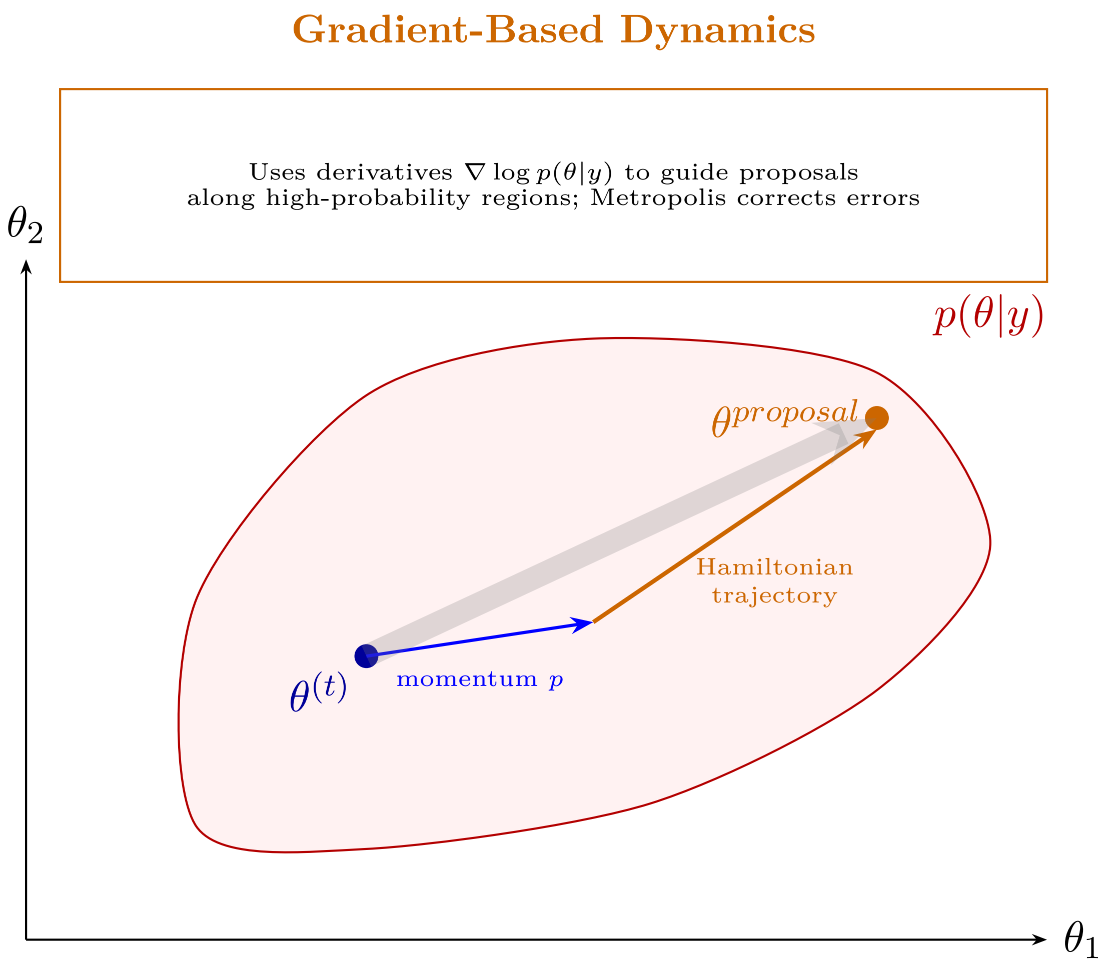
I acknowledge Dr John Holmes as my advisor and thank him for his patience and guidance. AI assistance was used for debugging, condensing language, developing figures, and reviewing concepts outside the author's prior background.

A Hamiltonian Monte Carlo (HMC)

A.1 Overview

Figure 7 shows the geometry of one HMC iteration in parameter space. We start at our current position $\theta^{(t)}$ (the blue point), introduce momentum p (the blue arrow), and then simulate Hamiltonian dynamics guided by the gradient $\nabla \log p(\theta|y)$ (the green arrow). This creates a curved trajectory (the blue path) that proposes a new point θ^{proposal} (the orange point). Finally, we accept or reject this proposal using a Metropolis step.

The key idea behind HMC is to use auxiliary variables—extra variables not part of the original model that help the algorithm work more efficiently. Specifically, HMC adds momentum variables p and then simulates Hamiltonian dynamics, treating parameters as positions and momentum as velocities according to physics-inspired equations. This simulation uses the derivative of the log posterior $\nabla \log p(\theta|y)$, which tells us how the posterior probability changes as we move through parameter space. We compute the derivative with respect to each parameter (the partial derivative $\partial \log p(\theta|y)/\partial\theta_j$) to guide the trajectory towards high-probability regions. Finally, a Metropolis accept–reject step corrects for numerical errors in the simulation,



Gradient-Based Dynamics

Uses derivatives $\nabla \log p(\theta|y)$ to guide proposals along high-probability regions; Metropolis corrects errors

Figure 7: HMC visual summary: current position, momentum, gradient direction, and Hamiltonian trajectory leading to a proposal.

ensuring the sampler converges to the correct posterior distribution (a property called detailed balance).

A.2 Walking Through One HMC Iteration

We now walk through one iteration of HMC step by step, showing the notation, the variables that exist at each stage, and how they connect to the diagram in Figure 7.

Step 0: Setup and notation

Before we begin sampling, we define our model and prior. For example:

$$p(\theta) \quad \text{with} \quad \beta \sim \mathcal{N}(0, 100).$$

Our target is the posterior distribution:

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

We start at the previous sample $\theta^{(n-1)}$. This is the blue point $\theta^{(t)}$ in Figure 7. At this stage, we have just one vector:

- $\theta^{(n-1)}$ (56 parameter values)

Step 1: Sample momentum

HMC introduces auxiliary momentum variables p drawn independently from:

$$p_0 \sim \mathcal{N}(0, I).$$

This is the blue arrow labelled "momentum p" in Figure 7. Setting $\theta_0 = \theta^{(n-1)}$, we now have two vectors:

- $\theta_0 = \theta^{(n-1)}$ (56 parameter values)
- p_0 (56 momentum values)

In code, this looks like:

```
theta <- theta_start
p     <- rnorm(length(theta), mean = 0, sd = 1)
```

Step 2: Leapfrog trajectory

Starting from (θ_0, p_0), we apply L leapfrog updates. Each update alternates between a half-step momentum update using the gradient $\nabla \log p(\theta|y)$ (the green arrow in Figure 7), a full-step position update, and another half-step momentum update. This creates the sequence:

$$(\theta_0, p_0) \to (\theta_1, p_1) \to \cdots \to (\theta_L, p_L).$$

The curved blue path in Figure 7 represents this Hamiltonian trajectory. During the trajectory, we have intermediate states (θ_k, p_k) for $k = 0, \dots, L$, but we only keep the endpoint. After the trajectory completes, we have four vectors:

- $\theta^{(n-1)} = \theta_0$ (56 starting parameter values)
- p_0 (56 starting momentum values)
- θ_L (56 new parameter values after L steps)

- p_L (56 updated momentum values after L steps)

In code:

```
for (k in 1:L) {
    grad  <- grad_log_posterior(theta, y)
    p     <- p + 0.5 * eps * grad
    theta <- theta + eps * p
    grad  <- grad_log_posterior(theta, y)
    p     <- p + 0.5 * eps * grad
}
```

Step 3: Proposal

We set the proposal to be the endpoint of the trajectory:

$$\theta^{\text{proposal}} = \theta_L$$
$$p^{\text{proposal}} = p_L$$

This is the orange point θ^{proposal} in Figure 7. We still have the same four vectors as before, just relabelled:

- $\theta^{(n-1)}$ (56 starting values)
- p_0 (56 starting momentum values)
- θ^{proposal} (56 proposed new values)
- p^{proposal} (56 final momentum values)

In code:

```
theta_proposal <- theta
```

Step 4: Accept or reject

We compute the log acceptance probability:

$$\log \alpha = \log p(\theta^{\text{proposal}}|y) - \log p(\theta^{(n-1)}|y).$$

Then we accept the proposal with probability $\min(1, e^{\log \alpha})$:

$$\theta^{(n)} = \begin{cases} \theta^{\text{proposal}} & \text{with probability } \min(1, e^{\log \alpha}) \\ \theta^{(n-1)} & \text{otherwise} \end{cases}$$

This Metropolis step corrects for numerical errors from the leapfrog approximation. After this step, the momentum values are discarded (they were auxiliary), and we have just one vector:

- $\theta^{(n)}$ (56 parameter values for the next iteration)

In code:

```
log_post_prop <- log_posterior(theta_proposal, y)
log_post_curr <- log_posterior(theta_start, y)
log_alpha     <- log_post_prop - log_post_curr

if (log(runif(1)) < log_alpha) {
    theta_next <- theta_proposal
} else {
    theta_next <- theta_start
}
```

Step 5: Repeat

We return to Step 1 with $\theta^{(n-1)} = \theta^{(n)}$ and continue sampling.

B Blocked Gibbs Sampling

B.1 Overview

Figure 8 shows the geometry of one blocked Gibbs iteration. We start at our current position $\theta^{(t)}$ (the blue point), which contains all parameters $(\beta, u, \tau_e, \tau_u)$. Unlike HMC, which uses gradients and momentum, Gibbs sampling updates parameters sequentially by sampling from full conditional distributions. Each parameter (or block of parameters) is updated by drawing from its conditional distribution given all other parameters and the data. The three coloured arrows show the three moves: sampling τ_u (blue vertical arrow), sampling (β, u) jointly (green horizontal arrow), and sampling τ_e (purple vertical arrow), arriving at $\theta^{(t+1)}$ (the purple point).

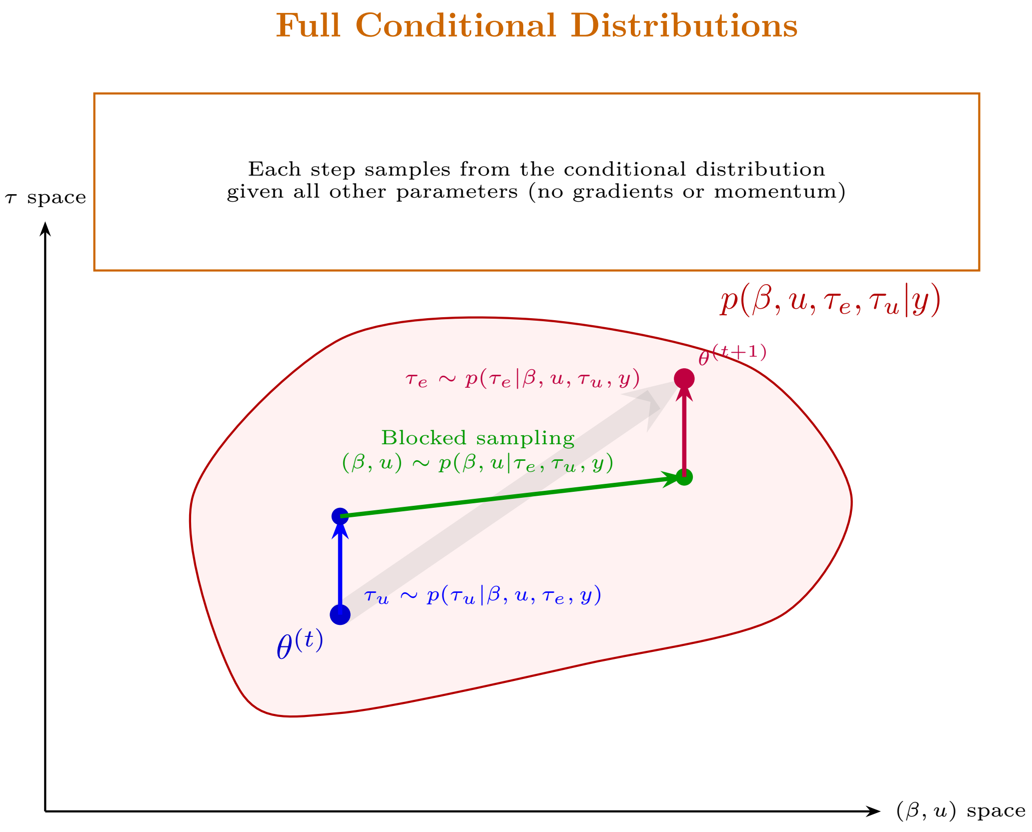

Full Conditional Distributions
Each step samples from the conditional distribution
given all other parameters (no gradients or momentum)
τ space
$p(\beta, u, \tau_e, \tau_u|y)$
$\theta^{(t+1)}$
$\tau_e \sim p(\tau_e|\beta, u, \tau_u, y)$
Blocked sampling
$(\beta, u) \sim p(\beta, u|\tau_e, \tau_u, y)$
$\tau_u \sim p(\tau_u|\beta, u, \tau_e, y)$
$\theta^{(t)}$
(β, u) space

Figure 8: Blocked Gibbs sampling: three sequential updates via full conditional distributions.

The key idea behind Gibbs sampling is to use full conditional distributions—the probability distribution of one parameter (or block of parameters) conditional on all others and the data. These distributions often have closed-form expressions, allowing direct sampling without gradients, momentum variables, or acceptance steps. For variance components τ_e and τ_u, the full conditionals are typically inverse-gamma distributions. For the regression parameters (β, u) given the variances, the full conditional is a multivariate normal. By cycling through these conditionals, the sampler explores the joint posterior without ever computing derivatives.

B.2 Walking Through One Gibbs Iteration

We now walk through one iteration of blocked Gibbs sampling step by step, showing the notation, the variables that exist at each stage, and how they connect to the diagram in Figure 8.

Step 0: Setup and notation

Our hierarchical model is:

$$
\begin{aligned}
&y = X\beta + Zu + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \tau_e^{-1} I) \\
&u \sim \mathcal{N}(0, \tau_u^{-1} K)
\end{aligned}
$$

where $\theta = (\beta, u, \tau_e, \tau_u)$ contains all parameters. Our target is the posterior:

$$p(\beta, u, \tau_e, \tau_u | y) \propto p(y|\beta, u, \tau_e) p(u|\tau_u) p(\beta) p(\tau_e) p(\tau_u).$$

Note that each hyperparameter appears twice: τ_e appears in the likelihood $p(y|\beta, u, \tau_e)$ and in its own prior $p(\tau_e)$; similarly, τ_u appears in the conditional prior for u and in its own prior. This structure is characteristic of Bayesian hierarchical models, where hyperparameters govern lower-level distributions whilst themselves being treated as random variables with their own priors **??**.

We start at the previous sample $\theta^{(t)} = (\beta^{(t)}, u^{(t)}, \tau_e^{(t)}, \tau_u^{(t)})$. This is the blue point in Figure 8. At this stage, we have:

- $\beta^{(t)}$ (4 values)
- $u^{(t)}$ (50 values)
- $\tau_e^{(t)}$ (1 value)
- $\tau_u^{(t)}$ (1 value)

Total: 56 parameter values in $\theta^{(t)}$.

Step 1: Sample τ_u

We sample the random effect variance component from its full conditional:

$$\tau_u^{(t+1)} \sim p(\tau_u | \beta^{(t)}, u^{(t)}, \tau_e^{(t)}, y).$$

For our model, this is an inverse-gamma distribution:

$$\tau_u^{(t+1)} \sim \text{InvGamma}\left(a_u + \frac{n_u}{2},\, b_u + \frac{1}{2} u^{(t)T} K^{-1} u^{(t)}\right).$$

This is the blue vertical arrow in Figure 8, moving in τ space whilst holding (β, u) fixed. After this step, we have:

- $\beta^{(t)}$ (4 values, unchanged)
- $u^{(t)}$ (50 values, unchanged)
- $\tau_e^{(t)}$ (1 value, unchanged)
- $\tau_u^{(t+1)}$ (1 new value)

In code:

```
shape_u <- a_u + n_u / 2
rate_u  <- b_u + 0.5 * t(u) %*% solve(K) %*% u
tau_u   <- rgamma(1, shape = shape_u, rate = rate_u)
```

Step 2: Sample (β, u) jointly

We sample the regression parameters and random effects jointly from their full conditional:

$$(\beta^{(t+1)}, u^{(t+1)}) \sim p(\beta, u | \tau_e^{(t)}, \tau_u^{(t+1)}, y)$$

For our model, this is a multivariate normal distribution. The joint update (blocked sampling) is crucial for good mixing when β and u are correlated. This is the green horizontal arrow in Figure 8, moving in (β, u) space whilst holding (τ_e, τ_u) fixed. After this step, we have:

- $\beta^{(t+1)}$ (4 new values)
- $u^{(t+1)}$ (50 new values)
- $\tau_e^{(t)}$ (1 value, unchanged)
- $\tau_u^{(t+1)}$ (1 value from Step 1)

In code:

```
# Precision matrix for joint (beta, u)
Q       <- tau_e * t(cbind(X, Z)) %*% cbind(X, Z)
Q[5:54, 5:54] <- Q[5:54, 5:54] + tau_u * solve(K)

# Mean vector
mu      <- solve(Q) %*% (tau_e * t(cbind(X, Z)) %*% y)

# Sample jointly
theta   <- mvrnorm(1, mu = mu, Sigma = solve(Q))
beta    <- theta[1:4]
u       <- theta[5:54]
```

Step 3: Sample τ_e

We sample the error variance component from its full conditional:

$$\tau_e^{(t+1)} \sim p(\tau_e | \beta^{(t+1)}, u^{(t+1)}, \tau_u^{(t+1)}, y).$$

For our model, this is an inverse-gamma distribution:

$$\tau_e^{(t+1)} \sim \text{InvGamma}\left(a_e + \frac{n}{2}, b_e + \frac{1}{2}(y - X\beta^{(t+1)} - Zu^{(t+1)})^T(y - X\beta^{(t+1)} - Zu^{(t+1)})\right).$$

This is the purple vertical arrow in Figure 8, moving in τ space whilst holding (β, u, τ_u) fixed. After this step, we arrive at $\theta^{(t+1)}$ (the purple point), and we have:

- $\beta^{(t+1)}$ (4 values from Step 2)
- $u^{(t+1)}$ (50 values from Step 2)
- $\tau_e^{(t+1)}$ (1 new value)
- $\tau_u^{(t+1)}$ (1 value from Step 1)

Total: 56 parameter values in $\theta^{(t+1)}$.

In code:

```
resid    <- y - X %*% beta - Z %*% u
shape_e  <- a_e + n / 2
rate_e   <- b_e + 0.5 * sum(resid^2)
tau_e    <- rgamma(1, shape = shape_e, rate = rate_e)
```

Step 4: Repeat

We return to Step 1 with $\theta^{(t)} = \theta^{(t+1)}$ and continue sampling.

B.3 Key Differences from HMC

- No gradients: Gibbs samples directly from full conditionals; HMC uses $\nabla \log p(\theta|y)$ to guide trajectories
- No auxiliary variables: Gibbs updates parameters in place; HMC introduces momentum p
- No accept/reject: Gibbs always accepts (samples are exact from conditionals); HMC uses Metropolis correction
- Sequential updates: Gibbs cycles through parameter blocks; HMC updates all parameters jointly via the trajectory
- Closed-form sampling: Gibbs relies on tractable conditionals (inverse-gamma, multivariate normal); HMC works even when conditionals are intractable

Both methods explore the posterior $p(\theta|y)$, but via completely different mechanisms. Gibbs is simpler when conditionals are available, whilst HMC is more robust to correlations and scales better to high dimensions.

References

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using p
'olya–gamma latent variables. *Journal of the American Statistical Association*, 108(504): 1339–1349, 2013.

Richard E Turner and Maneesh Sahani. Two problems with variational expectation maximisation for time-series models. *Bayesian Time Series Models*, pages 109–130, 2011.