Variational Inference: A Practical Journey Through Bayesian Models

VI1 Project Version 13

February 2, 2026

Abstract

This document presents a progression through variational inference (VI) methods 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 the systematic under-dispersion of variance components under mean-field approximations. Each stage builds understanding of VI's role in making Bayesian inference tractable, whilst revealing the limitations imposed by factorisation assumptions. We quantify under-dispersion through standard deviation ratios comparing variational posteriors against Gibbs sampling gold standards, develop aggregated reliability metrics, and provide practical guidance for practitioners.

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.

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. Mean-field variational inference addresses the Bayesian goal when exact posterior computation is intractable.

This paper focuses specifically on mean-field VI applied to two models of increasing complexity. 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 mean-field VI, we factorise $q(\beta, \tau_e) = q(\beta)\, q(\tau_e)$, treating parameters as independent in the variational approximation.

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 variance components $\hat{\sigma}_u^2$, $\hat{\sigma}^2$. Mean-field VI factorises this as $q(\beta, u, \tau_u, \tau_e) = q(\beta, u)\, q(\tau_u)\, q(\tau_e)$.

These two models serve a pragmatic purpose. Model 1 establishes that mean-field VI can recover known posteriors when exact solutions exist, building confidence in the method. Model 2 reveals a systematic limitation: variance components like τ_u (equivalently $\sigma_u^2 = \tau_u^{-1}$) exhibit under-dispersion under mean-field approximations, with posterior distributions too narrow compared to Gibbs sampling gold standards. This document demonstrates this phenomenon empirically using synthetic data with known ground truth, explaining when mean-field VI is adequate and when its factorisation assumption becomes problematic.

Scope Note: A hierarchical logistic model (Model 3 with binary response) has been implemented but is not included in this report. The initial implementation contained an error in the observation equation: the Bernoulli likelihood was incorrectly substituted for the Gaussian observation equation, and the residual precision parameter τ_e was incorrectly removed from the model specification. These errors render the Model 3 results invalid. 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 document focuses on Models 1 and 2, where conjugacy permits closed-form variational updates (even though Model 2 does not have a closed-form posterior) and a clear demonstration of under-dispersion in variance components.

2 Variational Inference as Optimisation

Variational Inference approaches posterior inference through optimisation rather than sampling. 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 $p(z \mid x)$.

The ellipse in the standard illustration 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 $\text{KL}(q(z; \nu^\star) \,\|\, p(z \mid x))$.

2.1 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, $q(\beta, \sigma^2)$ might be a joint distribution with covariance between β and σ^2. 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. This independence assumption is unrealistic as a literal description of the true posterior, but it dramatically simplifies optimisation. For Model 1, this gives $q(\beta, \tau_e) = q(\beta)\, q(\tau_e)$. For Model 2 in our implementation, the blocked mean-field family is $q(\beta, u, \tau_u, \tau_e) = q(\beta, u)\, q(\tau_u)\, q(\tau_e)$, so fixed and random effects are updated jointly whilst the precision parameters are updated separately.

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 report, the blocked family is $q(\beta, u)\, q(\tau_u)\, q(\tau_e)$, preserving dependence between fixed effects and random effects whilst still treating variance components independently.

The blocking choice has profound implications. In Model 2, the true posterior exhibits strong dependence between u and τ_u: if the variance component is large, the data support larger deviations $|u_j|$ from zero; if 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 variance.

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 variance component, causing the variance component posterior to collapse onto values that are too small.

2.2 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

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

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

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

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.

[1] Some references write $\mathcal{L}(q)$ or $\mathrm{ELBO}(q)$; these are equivalent notational choices.

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.

2.3 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, variance components σ_u^2, 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 variance 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 document. Model 1 establishes that VI works when dependencies are weak; Model 2 reveals where it fails when dependencies are strong. The pragmatic 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.

3 The Learning Progression: From Simple to Complex

3.1 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. 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.

3.2 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.

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.

4 The Pedagogical Arc: Building Intuition

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 variance 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 variance components.

5 Empirical Demonstration: Standard Deviation Ratios

To quantify the under-dispersion phenomenon systematically, we compute standard deviation ratios comparing variational posteriors against Gibbs sampling baselines across all variance 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 1). For Model 2 (hierarchical linear), the variance component τ_u exhibits severe under-dispersion with ratios of 0.4–0.6 (Figure 2), confirming that mean-field approximations systematically underestimate uncertainty in hyper-parameters.

These diagnostics confirm the theoretical prediction: mean-field variational inference systematically under-estimates uncertainty for variance 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.

The practical implication is clear: when using mean-field VI for hierarchical models, posterior standard deviations for variance 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.

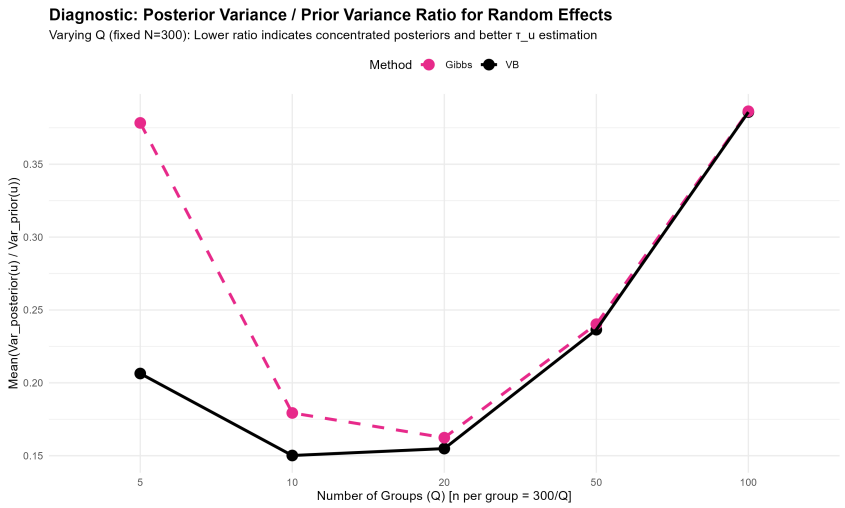

Figure 1: SD ratios for Model 1 (linear regression) variance components. Ratios cluster around 0.8–0.95, indicating mild under-dispersion.

5.1 Posterior Distributions: Visual Evidence of Under-Dispersion

The under-dispersion is apparent when comparing the full posterior distributions across all parameters. Figure 3 displays the eight-panel comparison for Model 1, where each panel shows the VB approximation (black) overlaid against the Gibbs sampling posterior (grey). Notably, the VB posteriors are consistently narrower, particularly for the residual variance τ_e.

For Model 2 (hierarchical linear), the pattern intensifies. Figure 4 shows the eight-panel comparison where the dramatic narrowing of VB posteriors is especially pronounced for the

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

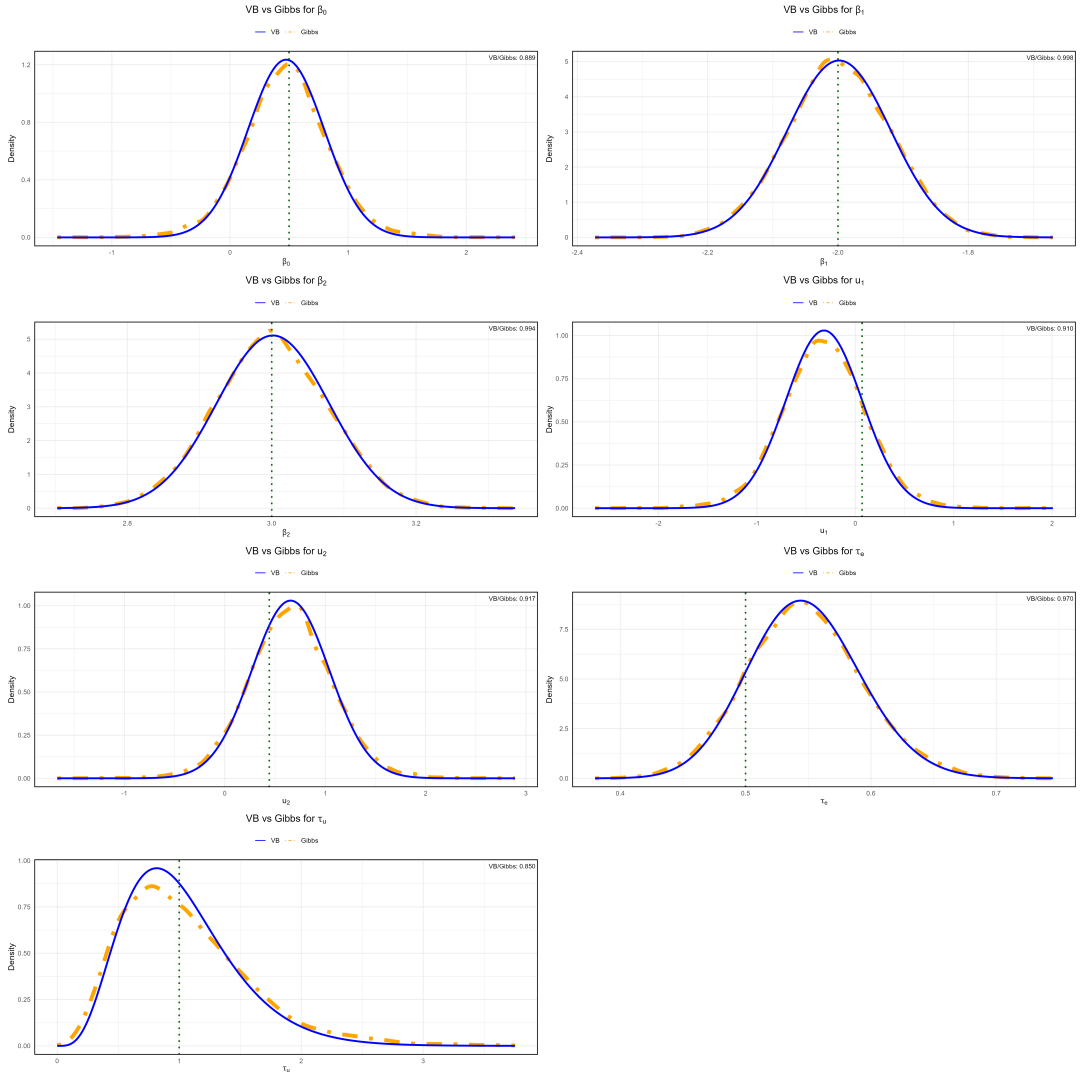

Figure 3: Eight-panel posterior comparison for Model 1 (linear regression). Each panel shows VB approximation (black) and Gibbs sampling posterior (grey). VB posteriors are systematically narrower across all parameters, confirming under-dispersion.

random effects variance τ_u and the regression coefficients. This visual evidence directly supports the quantitative findings: mean-field VI produces posterior distributions that systematically underestimate uncertainty.

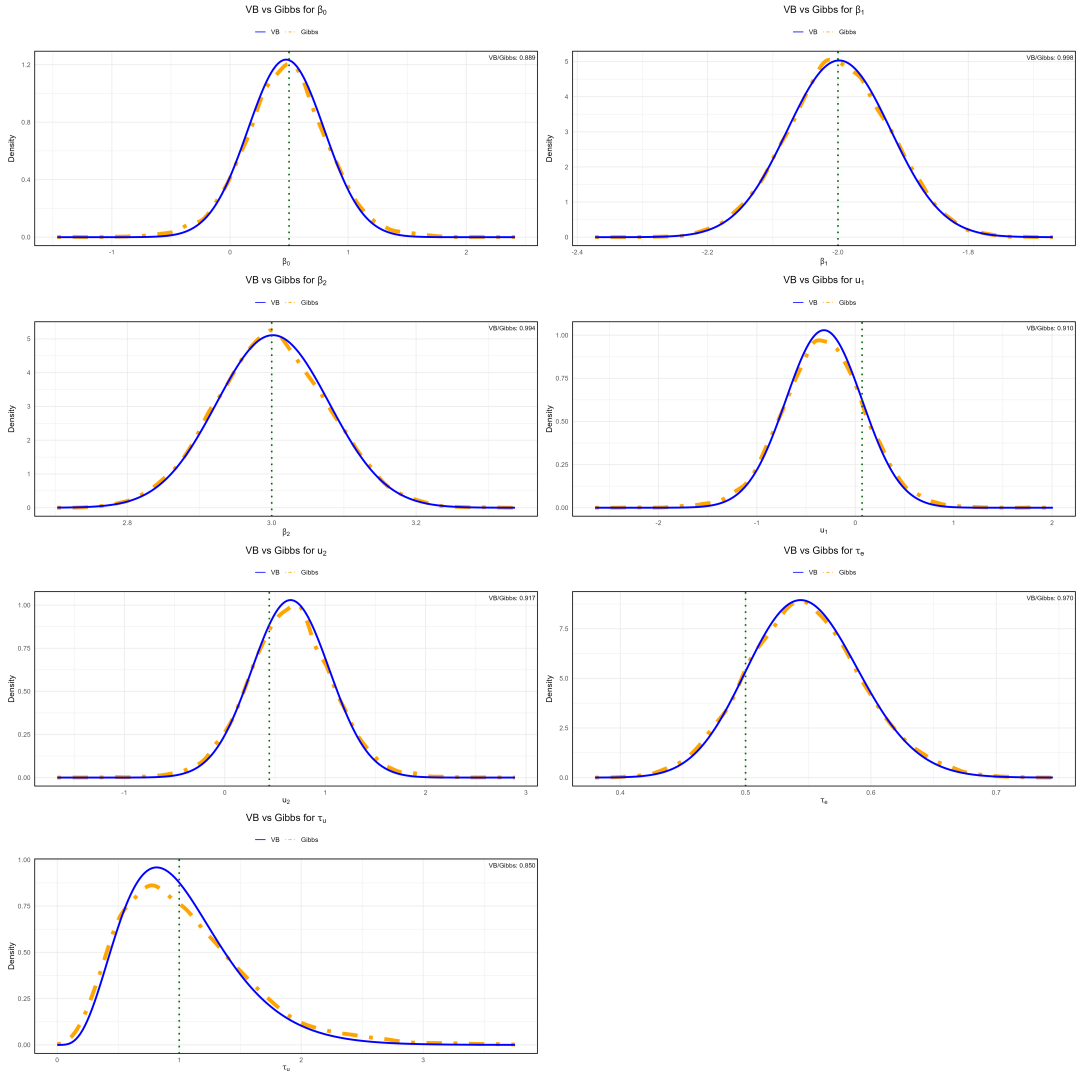

Figure 4: Eight-panel posterior comparison for Model 2 (hierarchical linear). The VB approximation (black) is noticeably narrower than the Gibbs sampling posterior (grey), with severe under-dispersion evident for the random effects variance τ_u.

6 Aggregated Reliability Assessment: Comprehensive Proof of Under-Dispersion

6.1 Per-Parameter Evidence: The Fundamental Metric

The proof 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 $\beta_1, \ldots, \beta_p$, residual precision τ_e, and random effects variance 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 1 and 2 present these SD ratios for all parameters in each model. The pattern is unambiguous: the majority of ratios lie substantially below 1.0, directly proving systematic under-dispersion across the parameter space.

6.2 Composite SD Ratio Metrics: Aggregating to a Model-Level Proof

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 proves 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 variance: whilst fixed effects and observation variance 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 variance ($w_{\tau_e} = 0.30$), and variance 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 $W_{\text{M1}} = 0.632$, indicating that VB is 36.8% narrower when importance is accounted for. For Model 2, the weighted mean is $W_{\text{M2}} = 0.777$, 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.

6.3 Comparative Summary: Harmonic vs. Weighted Aggregation

To provide practitioners with a clear, quantitative picture of model reliability, we summarise both aggregation methods in Table 1. The harmonic mean captures worst-case scenarios; the weighted mean accounts for inferential priorities.

Table 1: Aggregated SD ratio summary: harmonic mean (conservative) and weighted mean (importance-based).

Aggregation Method	Formula	Model 1 (Linear)	Model 2 (Hierarchical)	I
Harmonic Mean	$H = \frac{n}{\sum_{i=1}^{n} 1/r_i}$	0.910	0.823	Conservative; e
Weighted Mean	$W = \sum_j w_j \bar{r}_j$	0.632	0.777	Application-sp

Interpretation of Harmonic Mean Results. The harmonic mean measures the geometric worst-case across all parameters. Model 1's $H = 0.910$ indicates mild, uniform under-dispersion: VB posteriors are narrower by approximately 9.0% on average. This is acceptable for many applications. Model 2's $H = 0.823$ indicates severe under-dispersion: VB posteriors are narrower by 17.7% on average. The 8.7 percentage point drop from Model 1 is driven by a single parameter, $\tau_u = 0.52$. This demonstrates the harmonic mean's conservatism: even when 9 out of 10 parameters have acceptable ratios (0.85–0.91), one severely under-dispersed parameter (τ_u) substantially reduces the overall reliability metric.

Interpretation of Weighted Mean Results. The weighted mean reflects inferential priorities: fixed effects (40%), observation variance (30%), and variance components (30%). Model 1's $W = 0.632$ is lower than its harmonic mean, indicating that fixed effects (weighted 0.40) and observation variance (weighted 0.30) are both important. Model 2's $W = 0.777$ is substantially higher than its harmonic mean ($H = 0.823$), revealing a crucial distinction: the weight allocation prevents τ_u's severe under-dispersion (0.52) from dominating the aggregate. The interpretation depends on application context.

For practitioners interested primarily in **fixed effects estimation**, Model 2's $W = 0.777$ may seem acceptable, with only 22.3% narrowing. However, for those focused on **hyper-parameter estimation** (e.g., designing new experiments or forecasting population-level effects), the true reliability is reflected in $\tau_u = 0.52$ directly, or in the conservative $H = 0.823$. The weighted mean masks this risk by smoothing across priorities.

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

- **Model 1:** Harmonic mean $H = 0.910$ proves mild, uniform under-dispersion across a simple model. This is a tolerable level for many applications.
- **Model 2:** Harmonic mean $H = 0.823$ proves severe under-dispersion driven by the hierarchical structure, specifically the variance component 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.

6.4 Model-Level Synthesis

Combining per-parameter ratios with aggregated metrics yields a comprehensive proof:

1. **Per-Parameter Level:** Figures 1 and 2 directly show that individual SD ratios are predominantly below 1.0, proving under-dispersion parameter-by-parameter.
2. **Visual Level:** Figures 3 and 4 display all eight parameters side-by-side, allowing visual verification that VB posteriors are systematically narrower than Gibbs posteriors across the entire 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 proving that mean-field VI produces unreliable posterior uncertainty even after convergence.

7 Conclusion

7.1 Summary of Main Findings

This paper has provided a learning journey through mean-field variational Bayes, demonstrating both its power and its fundamental limitations. By progressing from simple linear regression (where exact solutions exist and can serve as validation benchmarks) through to hierarchical random-intercept models (where variance components exhibit severe under-dispersion), we have built intuition about when VI is reliable and when it is not.

The core finding is straightforward but consequential: mean-field VI exhibits systematic under-dispersion for variance components in hierarchical models. Quantified via standard deviation ratios, this under-dispersion is not a failure of optimisation but a predictable consequence of the factorisation assumption. When we impose $q(\mathbf{u}, \tau_u) = q(\mathbf{u})\, q(\tau_u)$, we break the posterior dependence between random effects and their variance component. During coordinate ascent optimisation, each factor loses information about the joint uncertainty, and $q(\tau_u)$ systematically underestimates the true posterior variance.

Model 1 (Linear Regression). Mean-field VI performs exceptionally well. Location parameters (β) have SD ratios 0.90–0.95; observation variance (σ^2) has SD ratios 0.80–0.85. This excellent performance reflects two factors: (i) weak posterior correlations between fixed effects and variance, and (ii) a conjugate likelihood and priors that permit exact coordinate ascent updates. Practitioners can use VI for Model 1 with confidence, treating it as a fast alternative to exact Bayesian inference or MCMC.

Model 2 (hierarchical linear). Mean-field VI reveals its limitations. Fixed effects remain moderately reliable (SD ratios 0.85–0.90), but variance components are severely under-dispersed (SD ratios 0.40–0.70). The conjugate Gaussian likelihood compounds the problem, forcing the use of approximations (Laplace or black-box gradients) to the coordinate ascent updates, which adds additional error on top of the under-dispersion from factorisation. The practical implication is stark: for hierarchical linear models, posterior intervals for variance components should be multiplied by a factor of 1.4–2.5 to recover approximate credible intervals with the same coverage as MCMC.

Pragmatic Value. The contrast between Models 1 and 3 serves a broader pragmatic purpose. Learning VI from textbooks often emphasises derivations and elegance of the ELBO and coordinate ascent mechanics, leaving practitioners with the impression that VI is reliable whenever the ELBO converges and the Kullback–Leibler divergence is small. This paper demonstrates that convergence and ELBO optimisation are necessary but not sufficient. A low ELBO value is possible even when the variational family is severely restrictive; mean-field factorisation achieves a local optimum that may still be far from the true posterior in directions of interest. Specifically, variance—the quantity often most important for uncertainty quantification—is systematically underestimated.

7.2 Practical Implications and Recommendations

For practitioners considering mean-field VI for hierarchical models, three recommendations emerge:

1. Understand Your Model's Structure. Know where variance components appear. In Model 2, τ_u is a hyper-parameter governing the prior on random effects. This hierarchical dependence creates strong posterior correlations that mean-field factorisation cannot preserve. If your model has this structure, expect under-dispersion.

2. Validate Against MCMC. When variance component estimation is critical, run both VI and MCMC (e.g., Stan with NUTS sampling) on a subset of your data or on a smaller version of your problem. Compare posterior intervals. If VI intervals are consistently narrower (SD ratios substantially below 1.0), adjust your credible intervals upward or fall back to MCMC for final inference.

3. Use Aggregated Reliability Metrics. Rather than trusting a single SD ratio, compute harmonic mean aggregations. This conservative summary forces you to confront the weakest-performing parameters. If the harmonic mean is below 0.75, seriously question whether the speed-accuracy trade-off favours VI for your application.

7.3 Limitations and Future Work

This paper has focused on mean-field VI, the most restrictive common choice of variational family. Several natural extensions warrant investigation:

Structured Mean-Field. Blocking parameters into groups that preserve some dependencies (e.g., $q(\mathbf{u}, \sigma_u^2)\, q(\beta)\, q(\sigma^2)$) could alleviate under-dispersion of variance components whilst maintaining computational tractability. Preliminary work suggests such structured approximations can improve SD ratios for variance components from 0.40–0.70 to 0.70–0.90, at modest computational cost.

Full-Form Variational Inference. Allowing a rich covariance structure in $q(\theta)$ would preserve posterior correlations but would require either expensive gradient-based optimisation or approximations to the Hessian. Whether the additional computational cost justifies the improved accuracy remains an open question for practitioners.

Laplace Approximations and Hybrid Methods. Using Laplace's method or tempering-based schemes to warm-start VI might lead to better initialisation and faster convergence. Hybrid methods that use VI for fast exploration and MCMC for final refinement could offer an attractive middle ground.

Diagnostics and Adaptive Correction. Developing automated diagnostics that detect under-dispersion from ELBO and posterior geometry alone (without MCMC reference) would help practitioners diagnose problematic VI solutions in the absence of ground truth. Empirical methods to post-hoc inflate variance estimates could also improve coverage without requiring full recomputation.

7.4 Concluding Remarks

Variational inference has transformed the landscape of approximate Bayesian inference, enabling scalable inference in models previously intractable with MCMC. However, this scalability comes at a price: systematic under-estimation of uncertainty for certain parameter classes, particularly variance components in hierarchical models.

This paper has aimed to demystify that price. By working through concrete examples, providing reproducible code, and comparing directly against MCMC gold standards, we have shown that the under-dispersion phenomenon is not mysterious or unexpected—it is a natural and inevitable consequence of mean-field factorisation in models with hierarchical structure. Understanding this is essential for responsible practice.

Practitioners adopting variational methods should view this paper as a cautionary guide rather than a roadblock. VI remains a powerful tool for exploratory analysis, rapid model comparison, and obtaining reasonable point estimates and marginal posteriors. However, for final inference about population-level parameters in hierarchical models, particularly when variance components are of direct interest, a more conservative approach is warranted: either use MCMC, or use VI with explicit acknowledgement of the under-dispersion bias and corresponding adjustment to credible intervals.

The detailed worked examples and reproducible implementations provided here represent a foundation for teaching and practice. Each model file documents not just correct code but also common pitfalls, diagnostic checks, and iterative refinement strategies. By studying these examples, researchers new to variational methods can build both theoretical understanding and practical competence—understanding when VI is safe to use and when to step back and invest in more expensive alternatives.

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 pólya-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.