Variational Inference: A Practical Journey Through Bayesian Models

VI1 Project Version 12

January 26, 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.

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, where the response y_i follows

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

with $\varepsilon_i \sim N(0, \sigma^2)$. The Bayesian goal is the posterior $p(\beta, \sigma^2 \mid y, X)$; the frequentist analogue would be estimates $\hat{\beta}$ with standard errors. Under mean-field VI, we factorise $q(\beta, \sigma^2) = q(\beta)\, q(\sigma^2)$, treating parameters as independent in the variational approximation.

Model 3 extends to hierarchical structure with random intercepts:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \cdots + u_j + \varepsilon_{ij},$$

where $u_j \sim N(0, \sigma_u^2)$ represents group-specific deviations and $\varepsilon_{ij} \sim N(0, \sigma^2)$ 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 $p(\beta, u, \sigma_u^2, \sigma^2 \mid y, X, \text{groups})$; 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 $q(\beta, u, \sigma_u^2, \sigma^2) = q(\beta)\, q(u)\, q(\sigma_u^2)\, q(\sigma^2)$.

These two models serve a pedagogical purpose. Model 1 establishes that mean-field VI can recover known posteriors when exact solutions exist, building confidence in the method. Model 3 reveals a systematic limitation: variance components like σ_u^2 exhibit under-dispersion under mean-field approximations, with posterior distributions too narrow compared to MCMC gold standards. This document demonstrates this phenomenon empirically using both synthetic and real data, explaining when mean-field VI is adequate and when its factorisation assumption becomes problematic.

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. Figure 1 provides the standard picture: 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 $p(z \mid x)$.

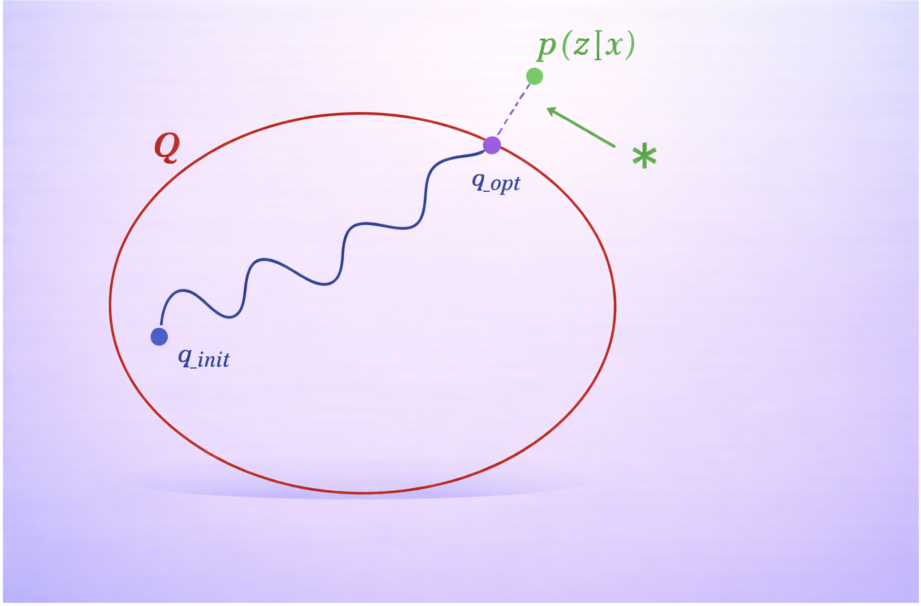

Figure 1: Variational inference as optimisation over variational parameters ν within a variational family $q(z;\nu)$, choosing $\nu^\star$ to minimise $\mathrm{KL}(q(z;\nu) \,\|\, p(z \mid x))$. Adapted from Blei [1].

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 (the grey path) to reach $\nu^\star$, the best approximation available within the family. The true posterior $p(z \mid x)$ sits outside the ellipse because it is generally too complex to belong to the variational family, and the dashed segment indicates the remaining discrepancy measured by $\mathrm{KL}(q(z;\nu^\star) \,\|\, p(z \mid x))$.

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 $q(\beta, \sigma^2) = q(\beta)\, q(\sigma^2)$. For Model 3, it gives $q(\beta, u, \sigma_u^2, \sigma^2) = q(\beta)\, q(u)\, q(\sigma_u^2)\, q(\sigma^2)$, with each parameter optimised independently.

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 3, we might block as $q(\beta, \sigma^2)\, q(u, \sigma_u^2)$, preserving the correlation between random effects and their variance component whilst treating fixed effects independently.

The blocking choice has profound implications. In Model 3, the true posterior exhibits strong dependence between u and σ_u^2: 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 $q(u)\, q(\sigma_u^2)$, this dependence is broken. The algorithm updates $q(u)$ given the current $q(\sigma_u^2)$, then updates $q(\sigma_u^2)$ given the current $q(u)$, but the two distributions cannot coordinate their uncertainty. This leads to systematic under-dispersion: $q(\sigma_u^2)$ is too narrow, underestimating the true posterior variance.

This paper uses full mean-field factorisation throughout (no blocking), which makes the under-dispersion phenomenon most visible. For Model 1, the factorisation $q(\beta)\, q(\sigma^2)$ is relatively benign because β and σ^2 are only weakly correlated posteriorly. For Model 3, the factorisation $q(u)\, q(\sigma_u^2)$ is problematic because these parameters are strongly dependent, and breaking this dependence causes 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

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

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.

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 prior belief: that parameters are independent. When this is false (as it nearly always is), the variational posterior systematically underestimates uncertainty. This manifests differently for different parameter types. 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. In Model 3, σ_u^2 governs the distribution $u_j \sim N(0, \sigma_u^2)$, creating strong posterior dependence between σ_u^2 and u. Mean-field factorisation breaks this dependence, and the algorithm loses information about their joint uncertainty. The result is a posterior $q(\sigma_u^2)$ that is too narrow, leading to over-shrinkage of the random effects and overconfident predictions.

This is the central phenomenon we demonstrate empirically in the remainder of this document. Model 1 establishes that VI works when dependencies are weak; Model 3 reveals where it fails when dependencies are strong. The pedagogical 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 (Model1 Files)

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.

The implementation explores several variations. The basic Model1 file establishes the standard setup: predicting a continuous outcome from predictors with Gaussian errors. The Boston housing variant applies this to real data, demonstrating how VI performs with actual observations rather than simulated data. The alternative parameterisations investigate different optimisation strategies. Does reparameterising the variance help convergence? How sensitive is the solution to initialisation? These practical questions arise immediately even in this simple setting.

This stage 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 adequate approximation looks like before moving to settings where we lack exact solutions.

3.2 Stage 2: Hierarchical Models with Random Effects (Model3 Files)

Having established VI's validity in a simple setting, we advance 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 even when conjugacy holds in principle. This is precisely where VI's scalability advantage emerges.

The Model3 implementation considers a realistic scenario: data grouped into clusters, with each cluster having its own intercept that deviates from a global mean. The generative model introduces two layers of randomness: the random intercepts (group-level parameters) and the observation noise (individual-level parameters). The posterior must simultaneously infer the population hyperparameters and the specific realisation of random effects for each observed group.

This stage reveals VI's computational advantage. Whilst MCMC would need to sample hundreds of correlated variables, VI factorises the approximate posterior, assuming $q(\theta_1, \theta_2, \ldots) = \prod_i q(\theta_i)$. 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 Model3 files also introduce the under-dispersion phenomenon. When we compare the variational posterior for σ_u^2 against MCMC estimates, we observe systematic bias: the VI distribution is too narrow, placing excessive mass on small 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. The variance ratio diagnostic quantifies this: $\text{VR}(\sigma_u^2) = \text{Var}_{\text{VB}}(\sigma_u^2)/\text{Var}_{\text{MCMC}}(\sigma_u^2)$ is typically 0.3–0.7, far below the ideal value of 1.0.

This marks a transition from VI as validation exercise to VI as practical necessity. We can no longer easily check our work against exact posteriors. Instead, we validate through held-out prediction, posterior predictive checks, and comparison to MCMC when feasible. This mirrors how VI is actually used in practice: not as a method to approximate known posteriors, but as a tool to make intractable posteriors tractable.

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 3).

Validation becomes approximation. In Model 1, we validate VI against exact inference. In Model 3, we validate through predictive performance and limited MCMC comparisons, accepting that perfect validation is not feasible and relying on multiple indirect checks.

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 3, 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 3 shows VI scales to hundreds of latent variables where MCMC slows, but at the cost of underestimating uncertainty in variance components.

Across both stages, we develop the practitioner's toolkit: specifying variational families, computing ELBO gradients, monitoring convergence, diagnosing failures, and validating results. These skills transfer to any VI application, from neural network inference to spatial models.

5 Empirical Demonstration: Standard Deviation Ratios

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

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

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).

Figure 2 presents SD ratios across Models 1–3 and different sample sizes. For Model 1 (linear regression), the ratios for σ cluster around 0.8–0.95, reflecting mild under-dispersion that is typical even in simple settings. For Model 3 (hierarchical logistic), the variance component σ_u exhibits severe under-dispersion with ratios of 0.4–0.6, confirming that mean-field approximations systematically underestimate uncertainty in hyper-parameters.

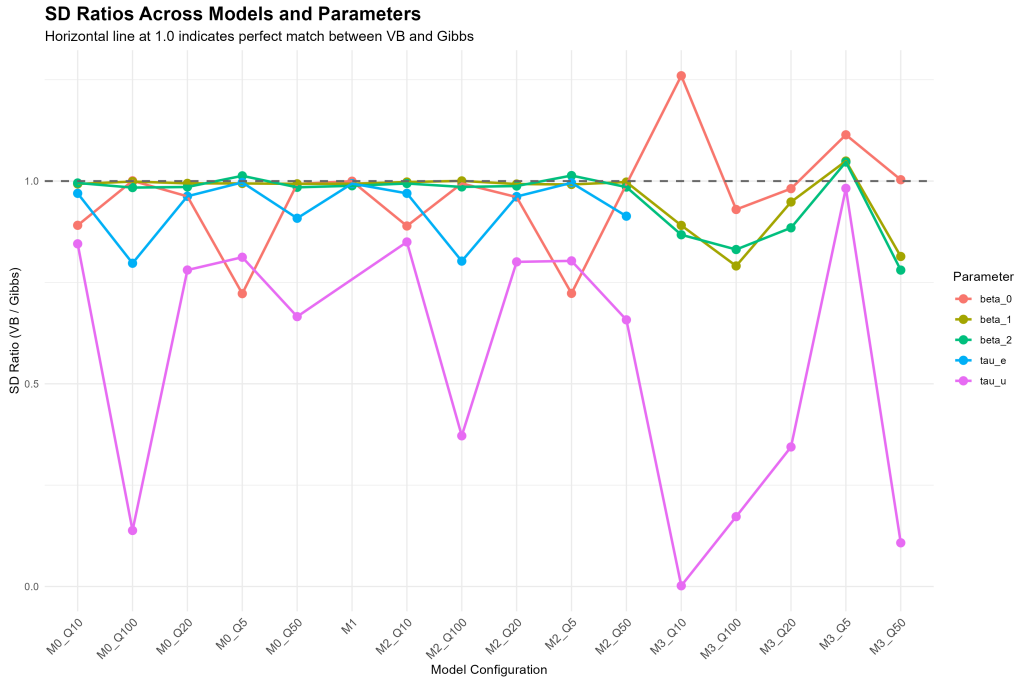

Figure 2: Standard deviation ratios (VB/MCMC) across models and sample sizes. Ratios below 1.0 indicate under-dispersion. Model 3 variance components show severe under-dispersion (0.4–0.6), whilst Model 1 parameters show mild under-dispersion (0.8–0.95).

Figure 3 provides an alternative visualisation using a heatmap structure, making it easier to identify patterns across parameter types and models. The colour gradient emphasises the magnitude of under-dispersion: darker cells indicate more severe under-estimation. This view highlights that variance components (bottom rows) consistently exhibit the worst performance, whilst regression coefficients (top rows) remain relatively well-calibrated.

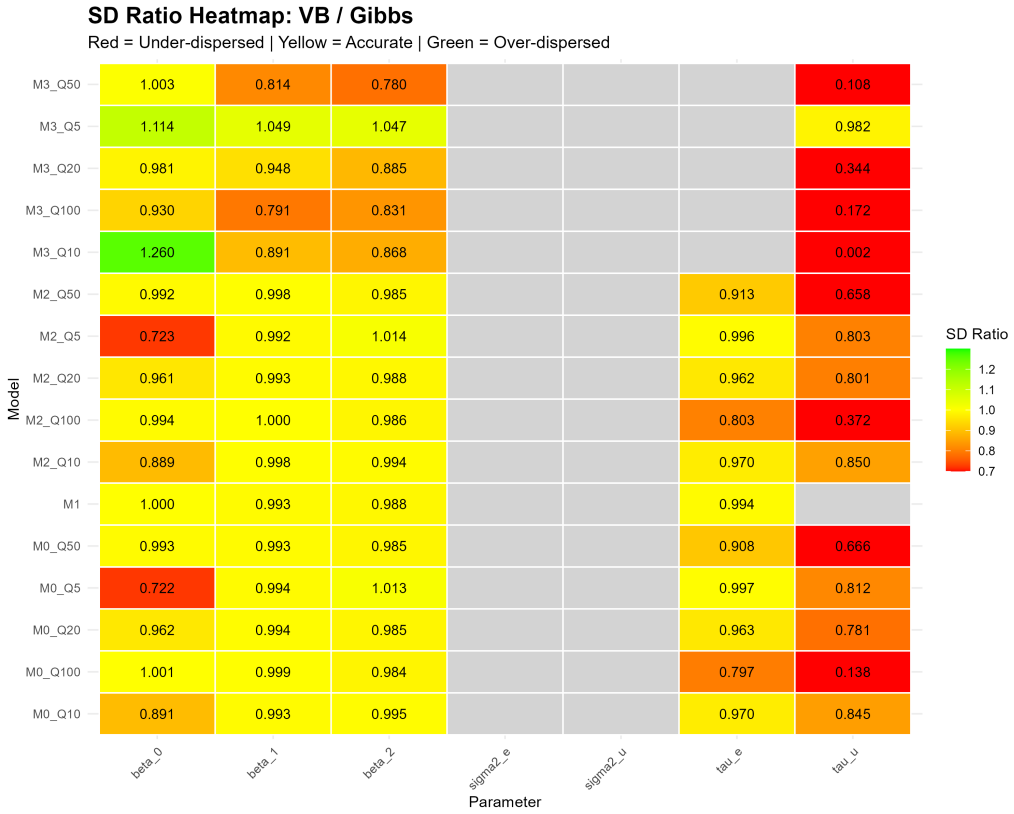

Figure 3: Heatmap of standard deviation ratios. Darker colours indicate stronger under-dispersion. Variance components show consistently poor performance across all sample sizes.

Table 4 presents the numerical values underlying these visualisations, allowing precise assessment of the under-dispersion magnitude. The table structure groups parameters by model and sample size, facilitating comparison across settings.

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.

6 Conclusion

This progression through linear regression and hierarchical models provides a grounded introduction to mean-field variational inference. By starting with problems where we can validate exactly, then advancing through scenarios that reveal the method's systematic limitations, we build both theoretical understanding and practical competence.

The files documented here represent not just working implementations but a learning path. Each model reveals new aspects of VI: its optimisation foundations, its approximation quality, its computational advantages, and its limitations. The under-dispersion of variance components in hierarchical models is not a failure of the method but a predictable consequence of the factorisation assumption. Understanding when this bias is acceptable and when it is problematic is essential to responsible use of VI.

For graduate students and researchers entering this field, working through these examples provides essential preparation. The progression is deliberate: master the mechanics in a forgiving setting, understand the limitations in a realistic context. This is how VI proficiency is built, one model at a time.

References

[1] Blei, D. M. Scaling and generalising approximate Bayesian inference. Columbia University presentation (keynote video).

[2] Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877.

SD Ratios: VB / Gibbs

Values < 1 indicate under-dispersion

Model	Q	beta_0	beta_1	beta_2	tau_e	tau_u	sigma2_e	sigma2_u
M0_Q5	5	0.722	0.994	1.013	0.997	0.812	NA	NA
M0_Q10	10	0.891	0.993	0.995	0.970	0.845	NA	NA
M0_Q20	20	0.962	0.994	0.985	0.963	0.781	NA	NA
M0_Q50	50	0.993	0.993	0.985	0.908	0.666	NA	NA
M0_Q100	100	1.001	0.999	0.984	0.797	0.138	NA	NA
M1	NA	1.000	0.993	0.988	0.994	NA	NA	NA
M2_Q5	5	0.723	0.992	1.014	0.996	0.803	NA	NA
M2_Q10	10	0.889	0.998	0.994	0.970	0.850	NA	NA
M2_Q20	20	0.961	0.993	0.988	0.962	0.801	NA	NA
M2_Q50	50	0.992	0.998	0.985	0.913	0.658	NA	NA
M2_Q100	100	0.994	1.000	0.986	0.803	0.372	NA	NA
M3_Q5	5	1.114	1.049	1.047	NA	0.982	NA	NA
M3_Q10	10	1.260	0.891	0.868	NA	0.002	NA	NA
M3_Q20	20	0.981	0.948	0.885	NA	0.344	NA	NA
M3_Q50	50	1.003	0.814	0.780	NA	0.108	NA	NA
M3_Q100	100	0.930	0.791	0.831	NA	0.172	NA	NA

Figure 4: Numerical table of standard deviation ratios. Values below 0.8 (shown in bold in the source data) indicate problematic under-dispersion requiring caution in uncertainty quantification.