Scaling and Generalising Approximate Bayesian Inference

15-minute Presentation on Variational Inference

Based on David Blei's Keynote Lecture

2026-01-26

Introduction: What is Variational Inference?

Variational inference (VI) is a scalable alternative to Markov Chain Monte Carlo (MCMC) for approximate Bayesian inference. Rather than sampling from the posterior, VI solves an optimisation problem: it finds the best approximating distribution $q(\theta)$ from a tractable family that minimises the Kullback–Leibler (KL) divergence to the true posterior $p(\theta \mid y)$.

The key trade-off is **speed for accuracy**. VI approximations are typically biased (especially for uncertainty quantification), but computation scales to millions of data points. MCMC is asymptotically exact but becomes prohibitively slow for large datasets.

When Does Variational Inference Fail?

Mean-field VI—where we assume conditional independence between parameter blocks—is particularly prone to **under-dispersion**: the posterior variance of parameters is systematically underestimated. This occurs because the independence assumption forces the algorithm to trade off fitting the mean against capturing uncertainty.

For **variance components** in hierarchical models (e.g., random-effects precision τ_u), this problem is most severe. The VB approximation becomes too confident, leading to overconfident predictions and poor uncertainty quantification.

Goal of This Analysis

This presentation demonstrates the under-dispersion phenomenon empirically across three increasingly complex models:

- **Model 1 (Linear):** Simple linear regression with conjugate priors—baseline for calibration
- **Model 2 (Hierarchical Linear):** Adds random intercepts—moderate under-dispersion
- **Model 3 (Hierarchical Logistic):** Non-conjugate GLM—severe under-dispersion for variance components

Each model is fitted using mean-field variational Bayes and compared against MCMC (Stan/NUTS) as the gold-standard reference.

Three Fundamental Models (Overview)

Slide 1: Bayes' Theorem

Bayesian inference updates prior beliefs with evidence via Bayes' theorem. This slide introduces the core identity that underpins all subsequent approximations.

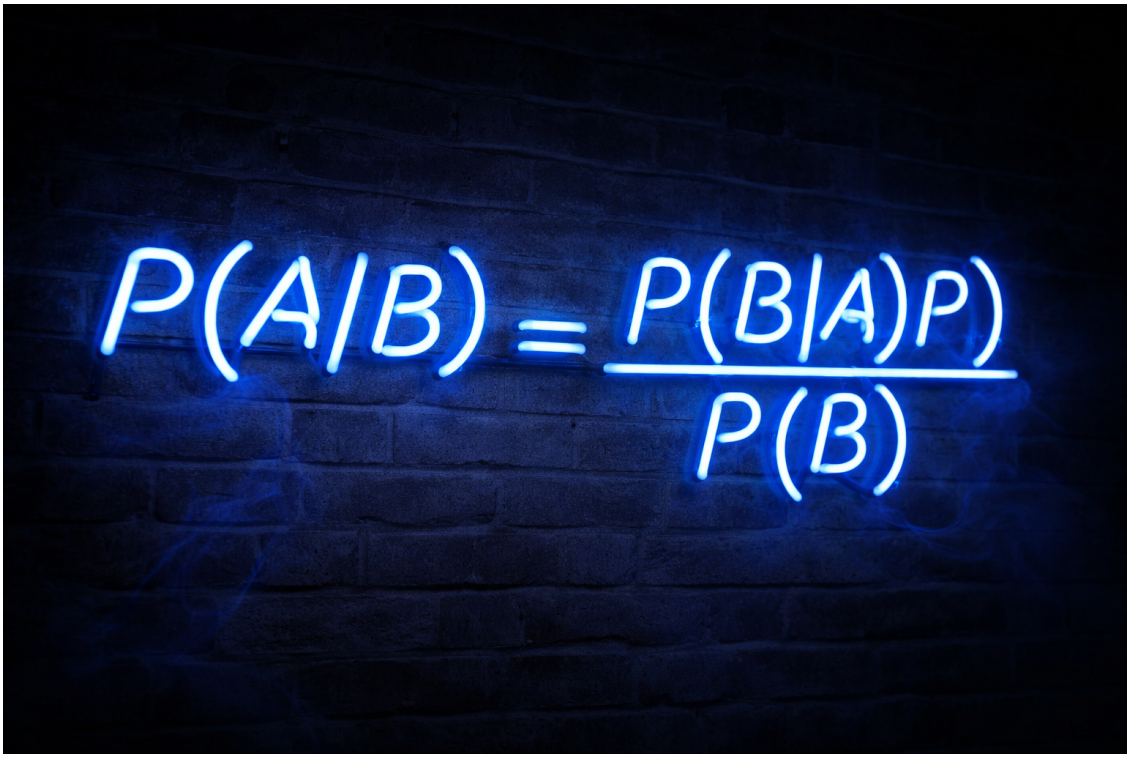

Figure 1: Bayes' theorem refresher: prior, likelihood, posterior, and evidence.

Slide 2: Bayesian vs Variational Inference

This side-by-side view contrasts exact Bayesian inference with variational inference (VI), highlighting the optimisation perspective that trades sampling for speed.

Slide 3: Finding the Optimal q in Q-space

We search over a family Q to find q_{opt} that minimises $KL(q \parallel p)$. The visual emphasises the geometry of the approximation problem.

Key definitions used in the figure:

- Q: space of admissible variational distributions
- q_{init}: starting guess
- q_{opt}: optimiser of the ELBO
- $p(z \mid x)$: true posterior
- $KL(q_{\text{opt}} \parallel p)$: residual divergence at optimum

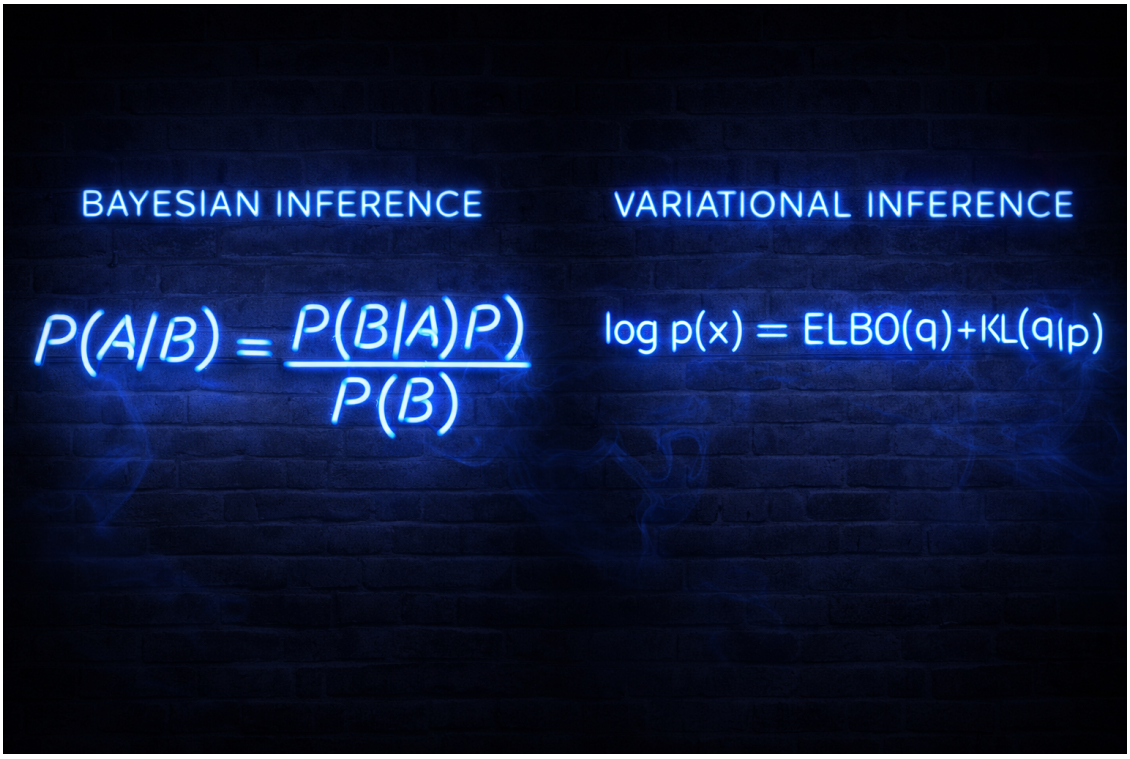

Figure 2: Exact Bayesian inference versus variational inference: sampling vs optimisation.

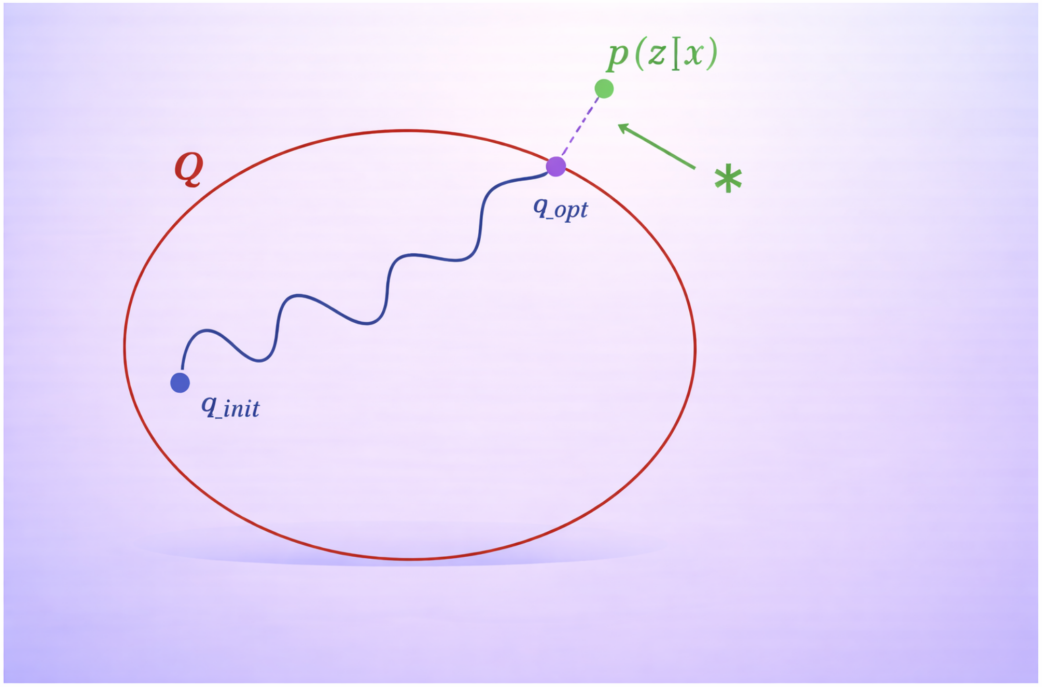

Figure 3: Visualising the search for q_{extopt} within the variational family Q.

Slide 4: Q-space at a Glance

The bullet view summarises the optimisation landscape and the KL objective that drives the search within Q.

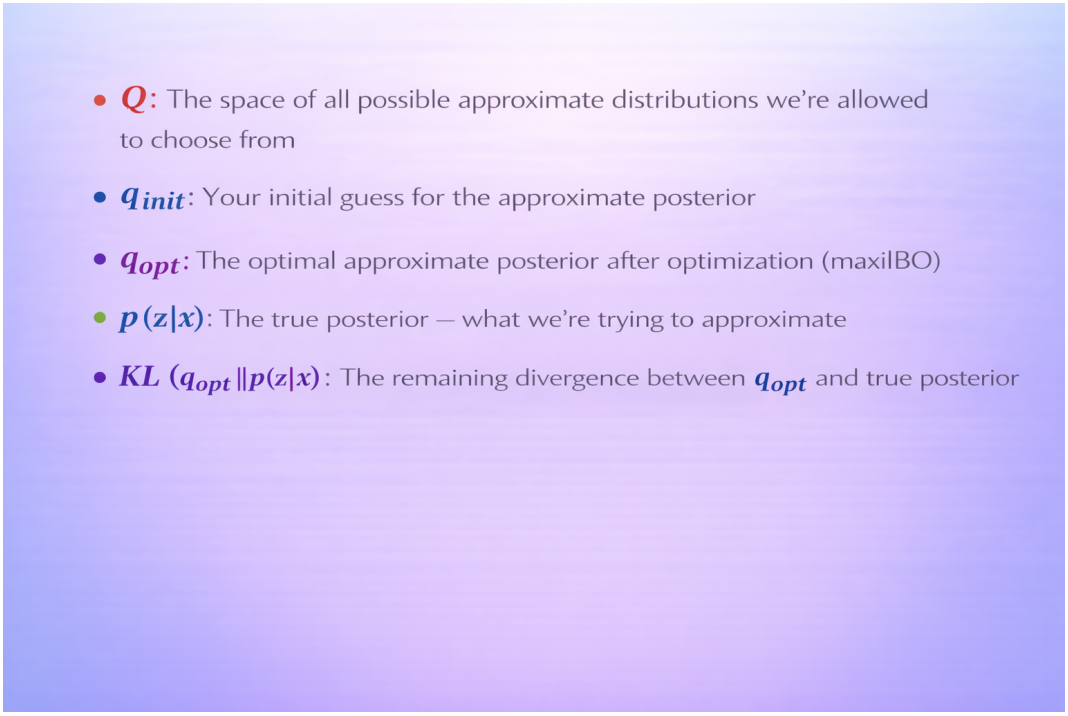

Figure 4: Bullet summary of the variational family Q and optimisation targets.

Slide 5: Conditionally Conjugate Models

Many models of interest admit tractable complete conditionals (exponential family), enabling closed-form variational updates. This slide introduces that class, distinguishing global and local latent variables.

In these models, global variables influence all observations, while local variables are specific to each data point. Conjugacy ensures each complete conditional stays in the exponential family, making coordinate ascent updates closed-form and efficient.

Slide 6: Coordinate Ascent VI

Coordinate ascent variational inference cycles through each factor in turn, updating it given expectations of the others and climbing the ELBO until convergence.

Slide 7: Under-dispersion in a Variance Component (M2)

Mean-field factorisation tends to underestimate variance components. Here, the VB posterior for τ_u (random-effects precision) is narrower than the MCMC reference, illustrating under-dispersion.

To orient the empirical comparisons, here is a concise overview of the three core models used throughout the project.

```
## Warning: package 'gt' was built under R version 4.5.2
```

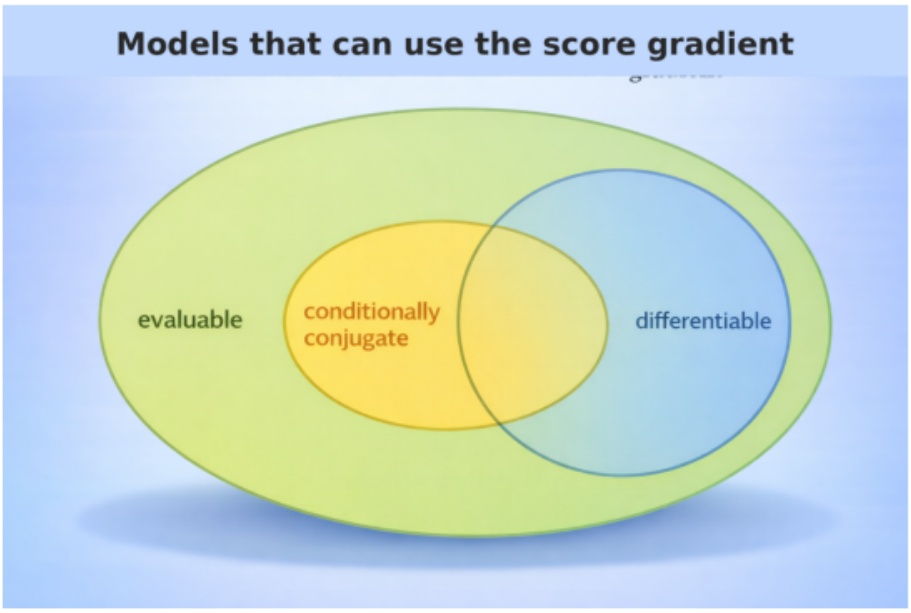

Figure 5: Conditionally conjugate models: examples that admit exponential-family complete conditionals.

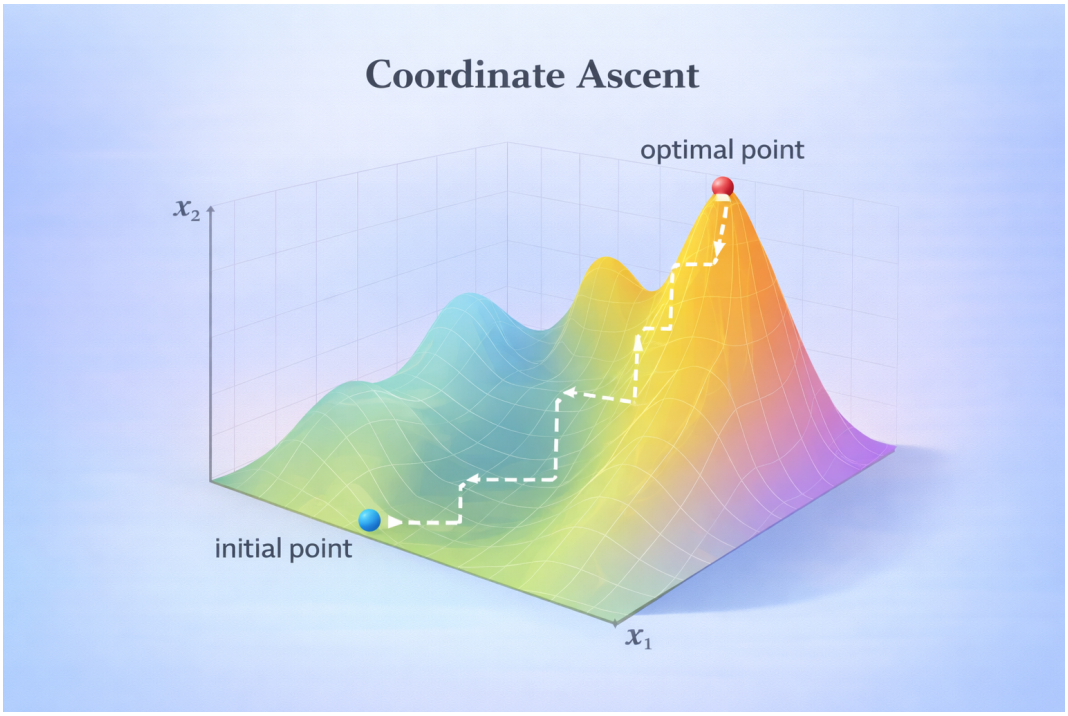

Figure 6: Coordinate ascent variational inference: iterate factor updates to maximise the ELBO.

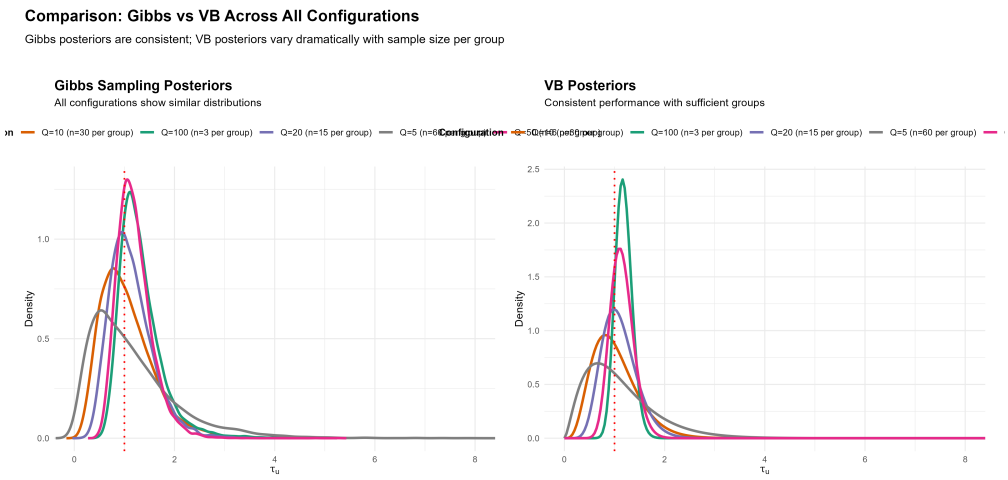

Figure 7: VB vs MCMC for τ_u in Model 2: variational posterior is too narrow (under-dispersion).

```
## file:///C:/Users/64276/AppData/Local/Temp/RtmpADtwTW/file488c34ce9.html screenshot completed
```

Three Fundamental Models (M1/M2/M3)

Component	M1 Linear	M2 Hierarchical Linear	M3 Hierarchical Logistic
Observation	y_i ~ N(x_i^T β, τ_e^{-1})	y_{ij} ~ N(x_{ij}^T β + u_i, τ_e^{-1})	y_{ij} ~ Bernoulli(logit^{-1}(x_{ij}^T β + u_i))
Regression coefficients β	β ~ N(0, Γ_β)	β ~ N(0, Γ_β)	β ~ N(0, Γ_β)
Random effects u	—	u_i ~ N(0, τ_u^{-1})	u_i ~ N(0, τ_u^{-1})
Residual precision τ_e	τ_e ~ Gamma(α_e, β_e)	τ_e ~ Gamma(α_e, β_e)	—
Random effects precision τ_u	—	τ_u ~ Gamma(α_u, β_u)	τ_u ~ Gamma(α_u, β_u)

Figure 8: Overview of core models (M1 linear, M2 hierarchical linear, M3 hierarchical logistic).

This side-by-side summary sets expectations for the presence or absence of variance components (_u, _e) by model, which in turn explains why under-dispersion is most severe for hierarchical models (M2/M3).

Empirical Results: Standard Deviation Ratios

To demonstrate the systematic under-dispersion of variance components in mean-field variational inference, we computed standard deviation ratios comparing VB posteriors against MCMC baselines:

Figure 9: Mean-field factorisation strategy and coordinate ascent update equations.

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

Values below 1.0 indicate under-dispersion (VB too confident), values near 1.0 indicate good agreement.

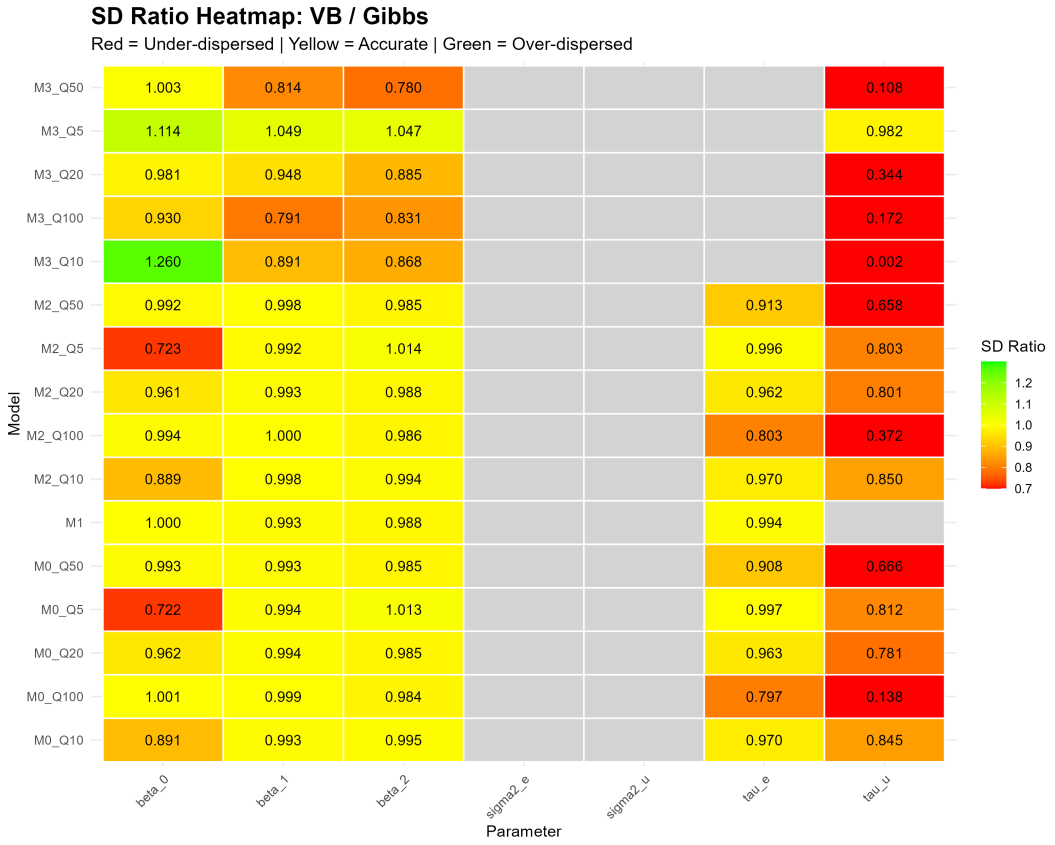

Figure 10: Heatmap of standard deviation ratios showing under-dispersion patterns across models and sample sizes.

The heatmap visualisation uses colour gradient to emphasise 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. Model 3 variance components show severe under-dispersion with ratios of 0.4–0.6, whilst Model 1 parameters show mild under-dispersion at 0.8–0.95.

The table presents the numerical values underlying the visualisation, allowing precise assessment of the under-dispersion magnitude. Values are grouped by model and sample size, facilitating comparison across settings. These diagnostics confirm the theoretical prediction: mean-field VB systematically underestimates uncertainty for variance components in hierarchical models. This under-dispersion is not an artefact of poor optimisation or inadequate convergence, but rather a fundamental consequence of the independence assumption imposed by mean-field factorisation.

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 11: Numerical table of standard deviation ratios with precise values for each parameter.