Scaling and Generalising Approximate Bayesian Inference

15-minute Presentation on Variational Inference

Based on David Blei's Keynote Lecture

2026-01-26

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.

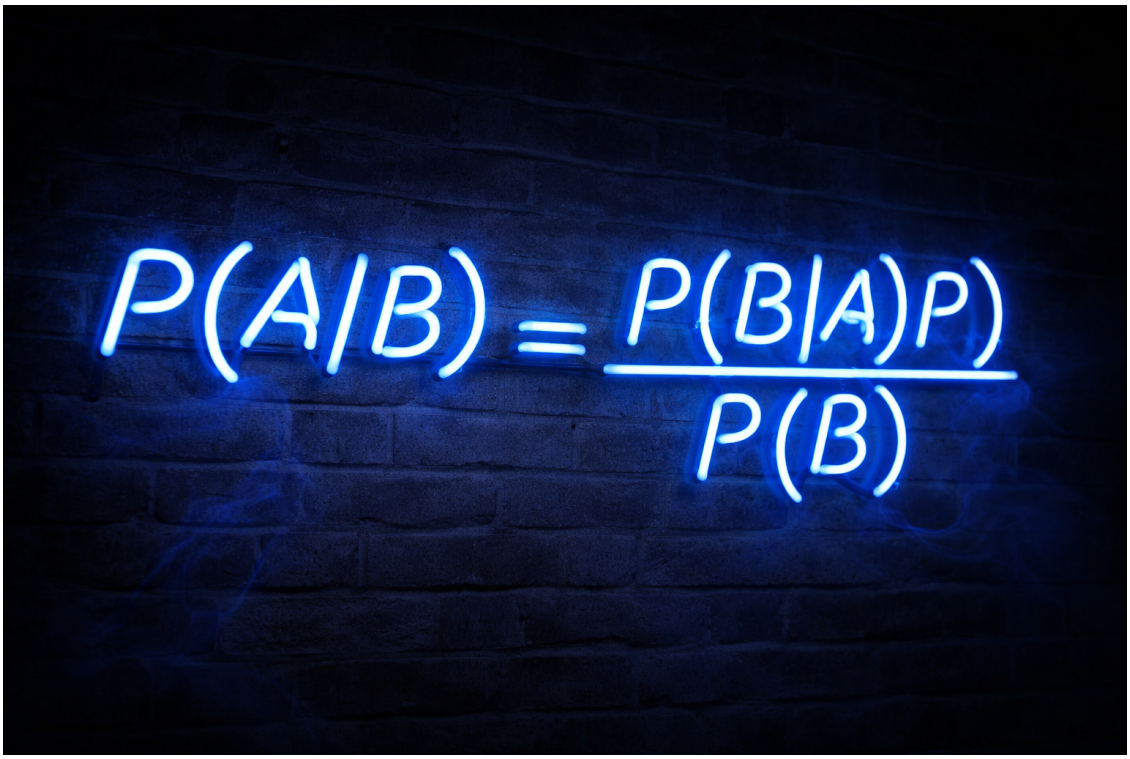

Figure 1: Bayes' theorem refresher: prior, likelihood, posterior, and evidence.

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.

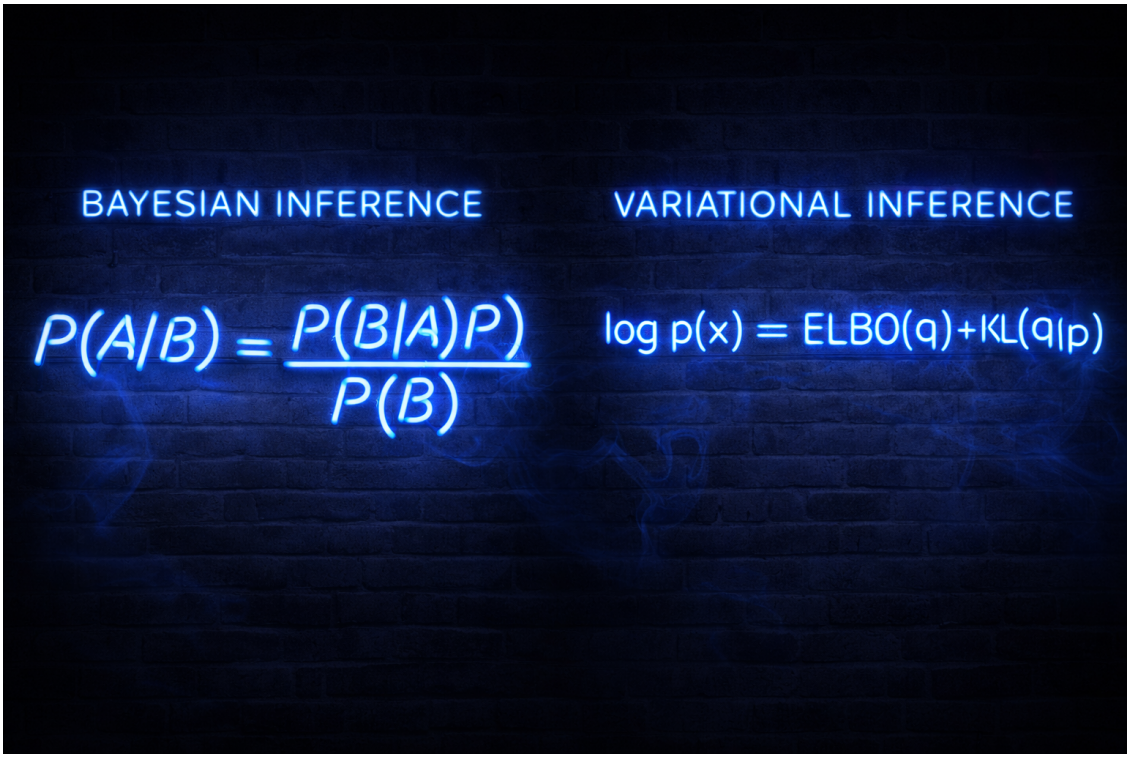

Figure 2: Exact Bayesian inference versus variational inference: sampling vs optimisation.

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

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.

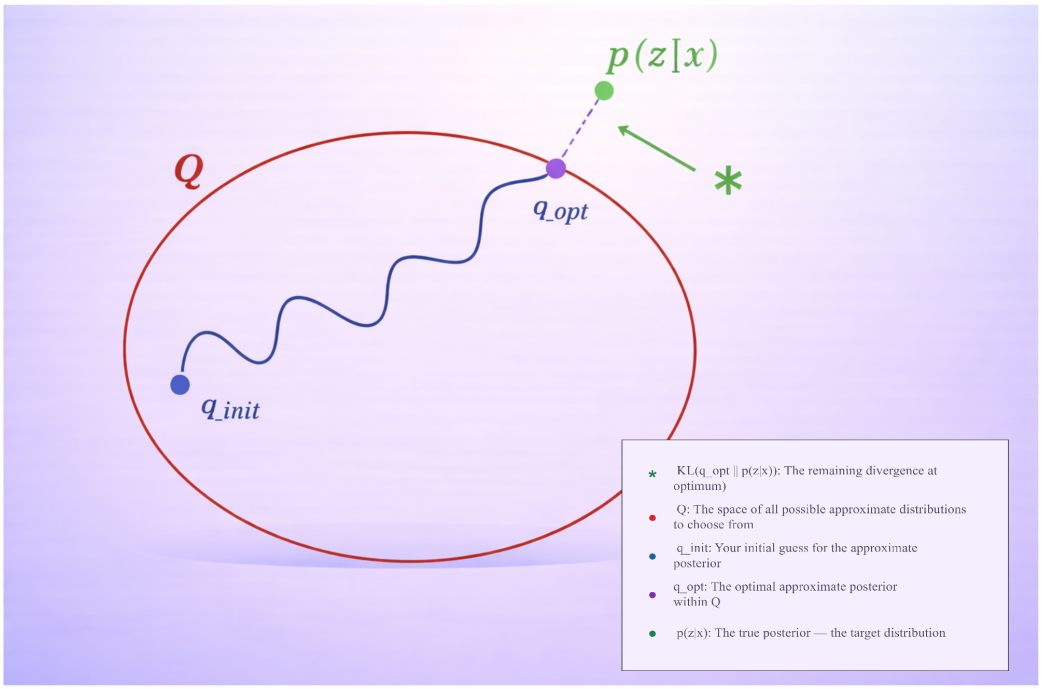

Figure 3: Visualising the search for q_{extopt} within the variational family Q (with bullet summary).

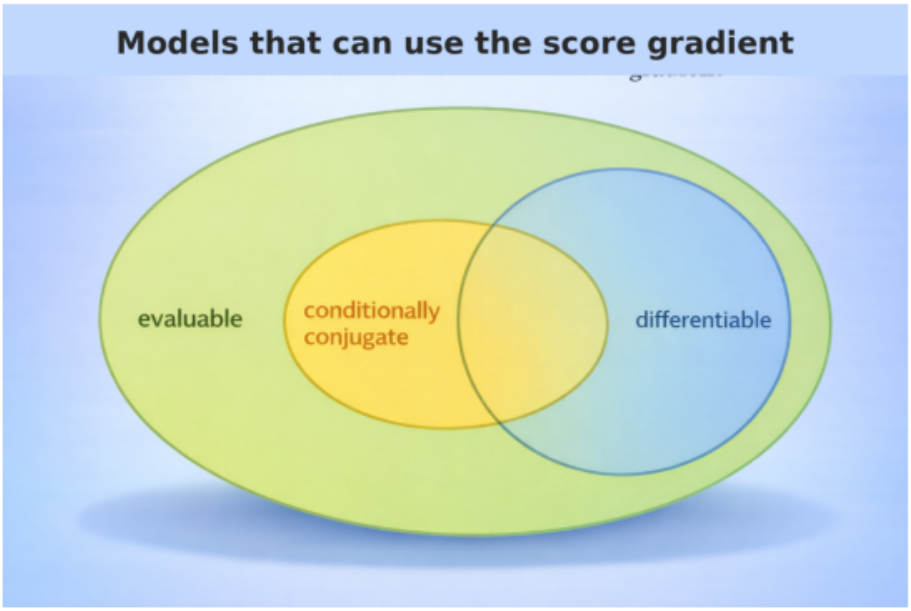

Figure 4: Conditionally conjugate models: examples that admit exponential-family complete conditionals.

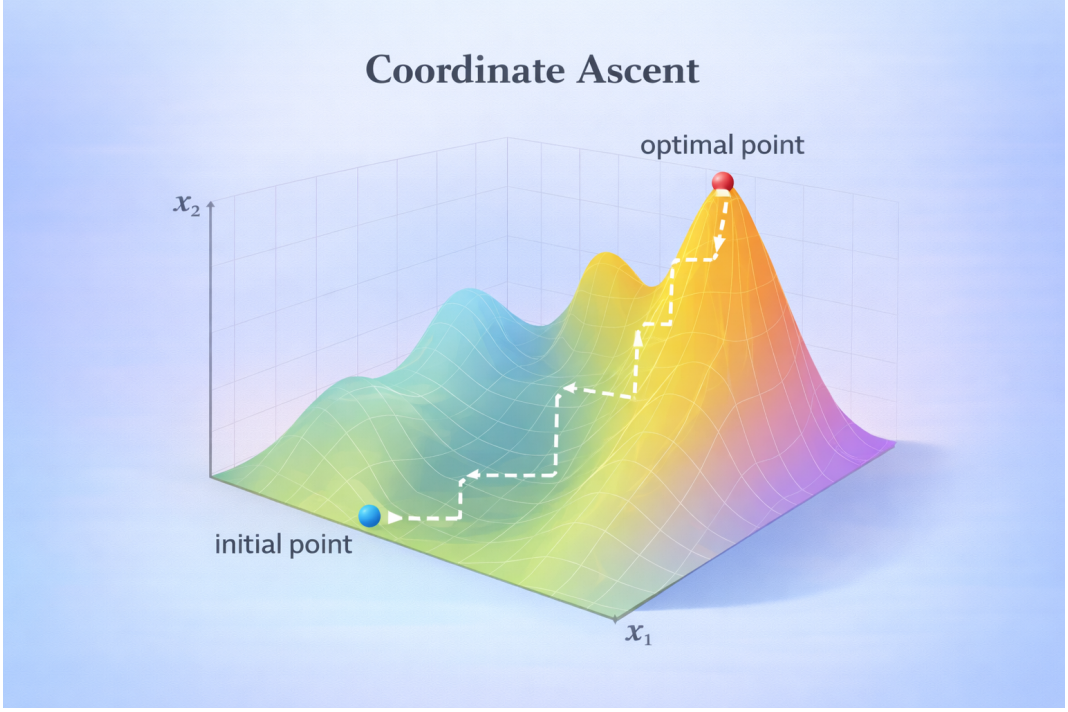

Figure 5: Coordinate ascent variational inference: iterate factor updates to maximise the ELBO.

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

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 Illustration: Under-dispersion in M2 Variance Component

Mean-field factorisation enables efficient inference through conditional independence, but this independence induces systematic under-dispersion in hyper-parameters. The posterior for the random-effects precision τ_u in Model 2 exemplifies this effect:

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 Gibbs baselines:

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

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

The table presents standard deviation ratios grouped by model and Q. Rows are ordered with M1 first, followed by M2 rows sorted by Q, then M3 rows sorted by Q (M0 removed). Parameter columns follow the order: , , , ^{2}_u, ^{2}_e, _u, _e. Values below 1.0 indicate under-dispersion; values near 1.0 indicate good agreement. This empirical pattern confirms the theoretical prediction: mean-field VB systematically underestimates uncertainty for variance components in hierarchical models.

Three Fundamental Models

Component	M1 Linear	M2 Hierarchical Linear	M3 Hierarchical Logistic
Observation	$y_i \sim N(x_i^T\beta, \tau_e^{-1})$	$y_{ij} \sim N(x_{ij}^T\beta + u_i, \tau_e^{-1})$	$y_{ij} \sim \text{Bernoulli}(\text{logit}^{-1}(x_{ij}^T\beta + u_i))$
Regression coefficients β	$\beta \sim N(0, \Gamma_\beta)$	$\beta \sim N(0, \Gamma_\beta)$	$\beta \sim N(0, \Gamma_\beta)$
Random effects u	—	$u_i \sim N(0, \tau_u^{-1})$	$u_i \sim N(0, \tau_u^{-1})$
Residual precision τ_e	$\tau_e \sim \text{Gamma}(\alpha_e, \beta_e)$	$\tau_e \sim \text{Gamma}(\alpha_e, \beta_e)$	—
Random-effects precision τ_u	—	$\tau_u \sim \text{Gamma}(\alpha_u, \beta_u)$	$\tau_u \sim \text{Gamma}(\alpha_u, \beta_u)$

Figure 6: Overview of core models (M1 linear, M2 hierarchical linear, M3 hierarchical logistic).

Mean-Field Factorisation Strategy

Mean-field variational inference imposes conditional independence on the posterior to enable tractable inference. The key is to factor parameters according to their coupling structure in the likelihood.

Full joint posterior:

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

Mean-field factorisation:

$$q(\beta, u, \tau_u, \tau_e) = q(\beta, u) \cdot q(\tau_u) \cdot q(\tau_e)$$

This factorisation preserves the coupling between β and u in the likelihood, while separating the precision parameters to admit conjugate Gamma updates.

Coordinate Ascent Updates

Regression block: $p(\beta, u \mid y, \tau_u, \tau_e)$ is Gaussian, so $q(\beta, u)$ is Gaussian:

$$\begin{aligned}
\Sigma_{\beta u}^{\text{new}} &= \left[X^T X\, \mathbb{E}[\tau_e] + \text{diag}(\Gamma_\beta^{-1}, \mathbb{E}[\tau_u]\, \mathbf{1}_u)\right]^{-1} \\
\mu_{\beta u}^{\text{new}} &= \Sigma_{\beta u}^{\text{new}}\, X^T y\, \mathbb{E}[\tau_e]
\end{aligned}$$

Residual precision: $p(\tau_e \mid y, \beta, u)$ is Gamma, so $q(\tau_e)$ is Gamma:

$$\begin{aligned}
a_e^{\text{new}} &= a_e + \tfrac{n}{2} \\
b_e^{\text{new}} &= b_e + \tfrac{1}{2}\mathbb{E}[(y - X\beta - Zu)^T (y - X\beta - Zu)]
\end{aligned}$$

Random-effects precision: $p(\tau_u \mid u)$ is Gamma, so $q(\tau_u)$ is Gamma:

$$\begin{aligned}
a_u^{\text{new}} &= a_u + \tfrac{Q}{2} \\
b_u^{\text{new}} &= b_u + \tfrac{1}{2}\mathbb{E}[u^T u]
\end{aligned}$$

All expectations are computed using current variational parameters. The loop repeats until ELBO convergence.

Figure 7: Mean-field factorisation strategy and coordinate ascent update equations.

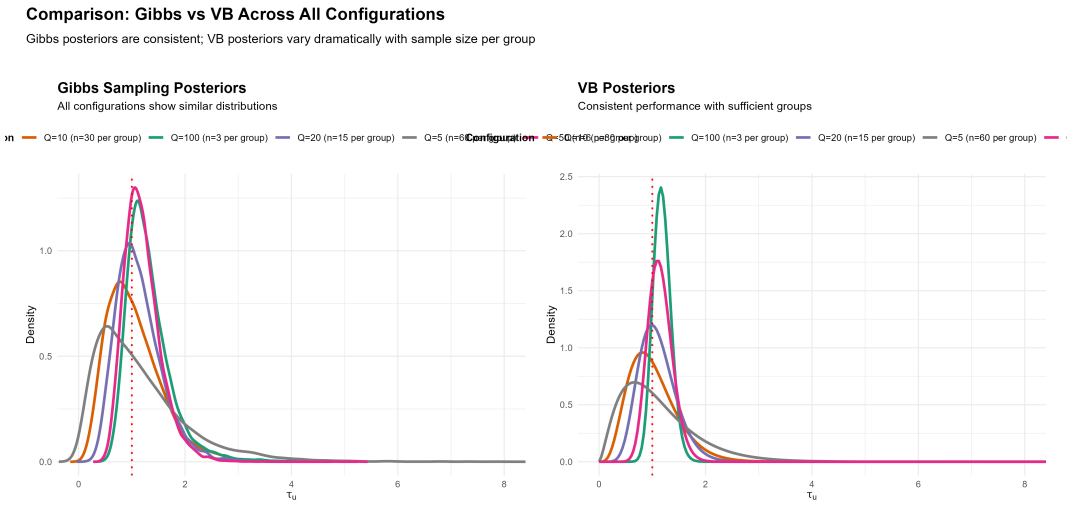

Figure 8: VB vs Gibbs for τ_u in Model 2: variational posterior is too narrow (under-dispersion).

SD Ratios: VB / Gibbs

Values < 1 indicate under-dispersion

Model	Q	β_0	β_1	β_2	τ_e	τ_u	σ^2_e	σ^2_u
M1	—	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 9: Standard deviation ratios (VB / Gibbs) across all model configurations.