On Variational Bayesian Methods

David Ewing

2026-01-27

Slide 1: Bayes' Theorem

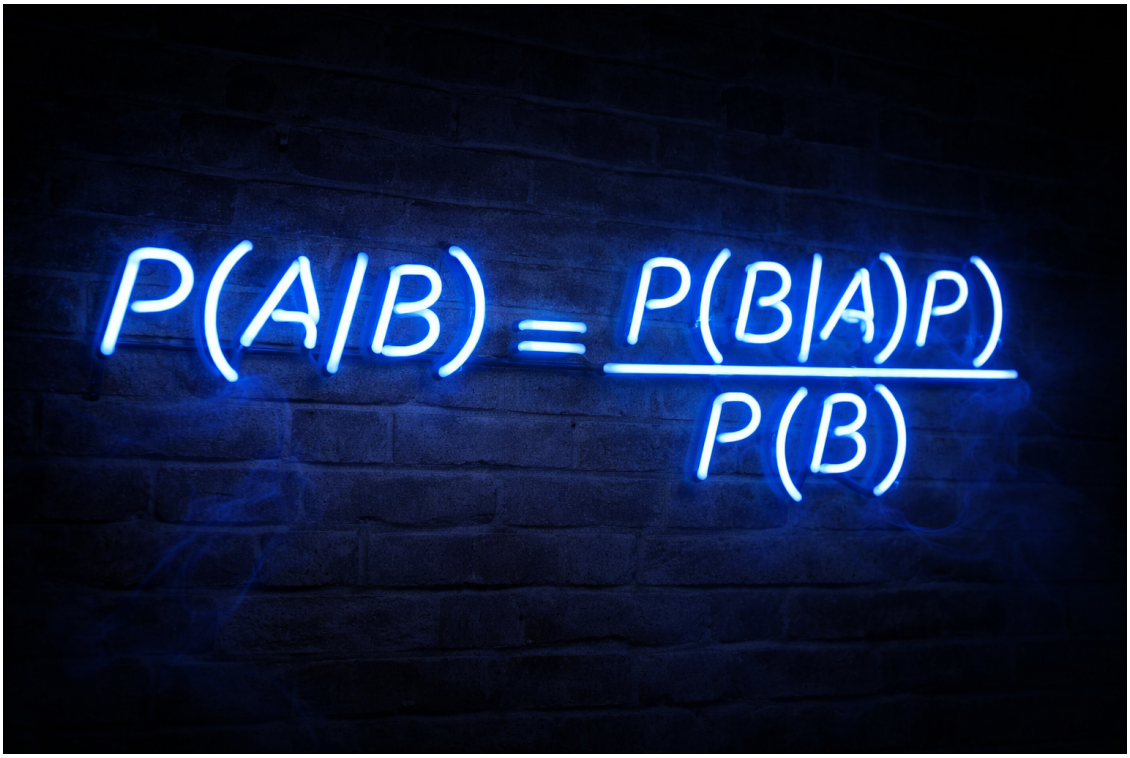

Figure 1: Bayes' theorem : prior, likelihood, posterior, and evidence.

(Silent slide while I make my way to the podium)

note: P(A) in the numerator needs to be fixed.

Slide 2: Bayesian vs Variational Inference

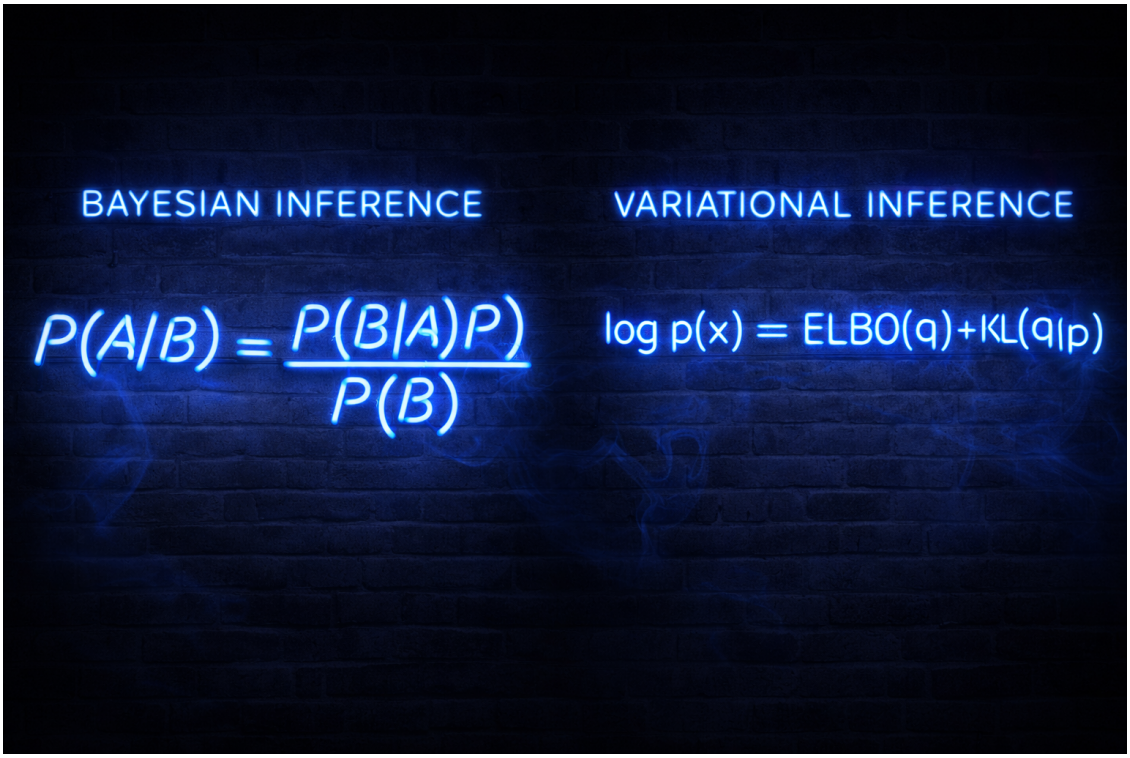

Figure 2: Exact Bayesian inference versus variational inference: sampling vs optimisation.

At the podium: Good afternoon. My name is David Ewing, and I'm with the MADS programme. My presentation is on Variational Bayes Methods, also known as Variational Inference. Bayes' Rule gives us the exact posterior — assuming we can compute it. But in practice, especially with complex models, that denominator — p(B) or p(x) — becomes intractable. This is where variational inference comes in. Rather than computing the true posterior directly, we approximate it — by reframing inference as an optimisation problem. Let me show you what that looks like geometrically.

note: P(A) in the numerator needs to be fixed.

Slide 3: Finding the Optimal q in Q-space

Visualising the variational family

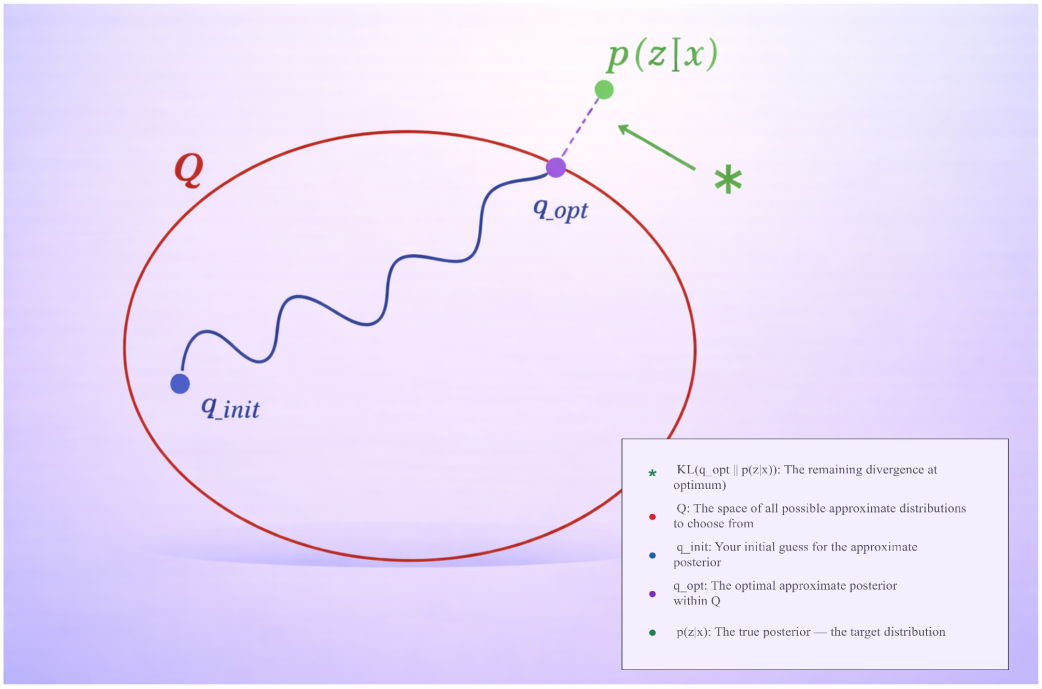

Figure 3: Visualising the search for q_{opt} within the variational family Q.

"This diagram, adapted from David Blei, shows the core idea behind variational inference. We begin with a space of candidate distributions — that's the red region labelled Q. Inside it, we choose an initial guess — q_init — and then optimise it to find q_opt, the best approximation we can achieve within that space. The true posterior — p(zx) — lies outside this space. It's what we'd ideally compute, but often cannot. The green asterisk marks the remaining divergence — the KL divergence between our best approximation and the true posterior. In essence, variational inference is about choosing a tractable family of distributions, and then finding the member of that family that gets us as close as possible to the truth. It's not exact, but it's fast, scalable, and surprisingly effective — especially in high-dimensional models."

Slide 5: Conditionally Conjugate Models

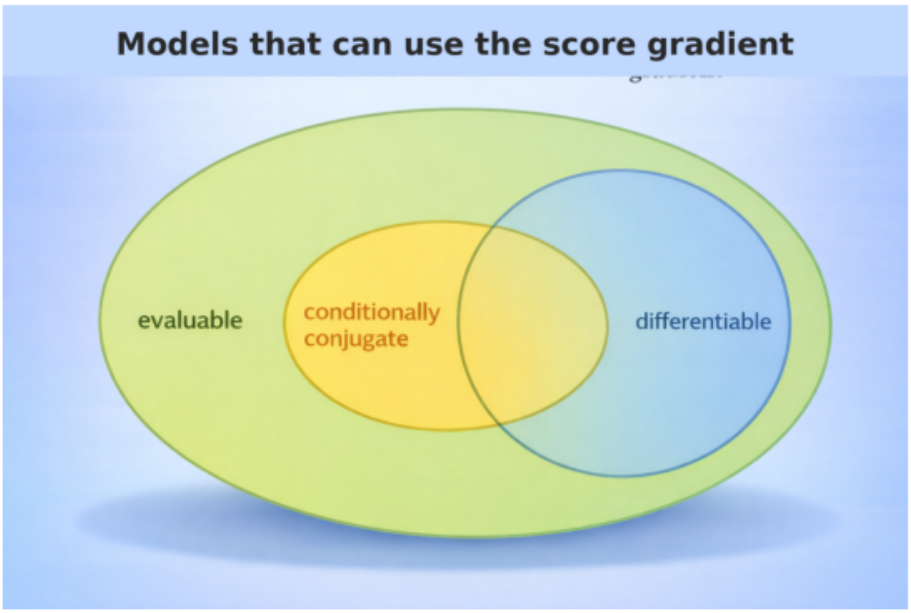

Figure 4: Conditionally conjugate models: examples that admit exponential-family complete conditionals.

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. Once inference is framed as optimisation, the availability of gradients becomes crucial. This slide highlights which models admit score-based gradients. When these gradients are available, optimisation-based inference becomes feasible and scalable. When they are not, alternative strategies or approximations are required. This distinction plays a central role in determining which variational methods are practical for a given model.

Slide 6: Coordinate Ascent VI

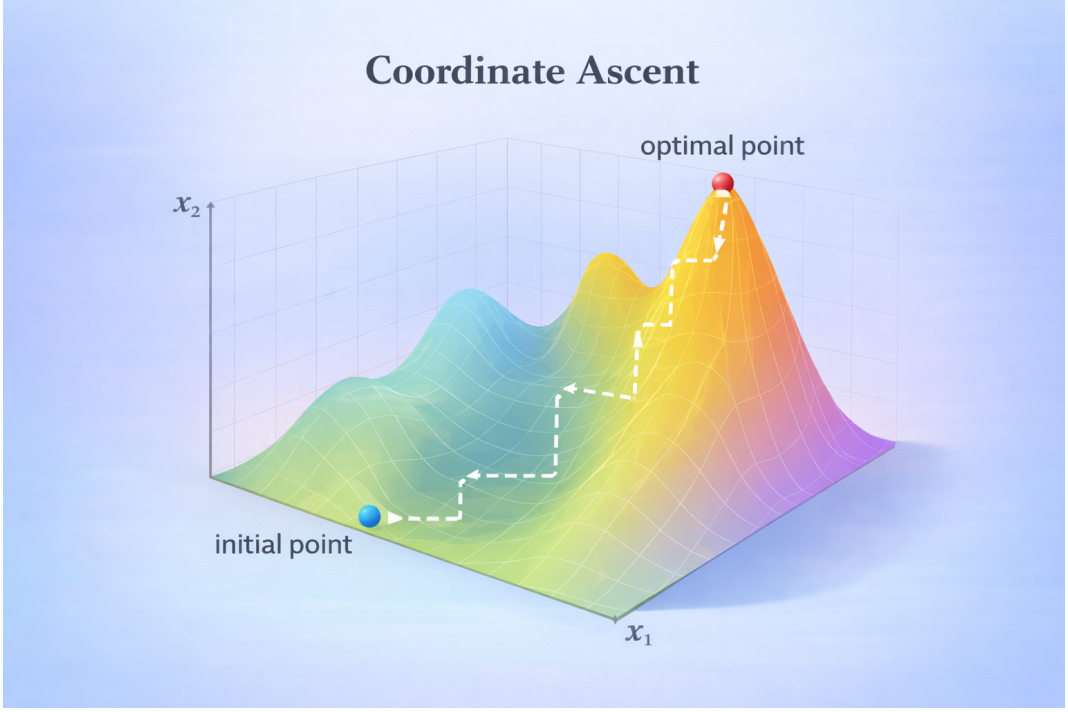

Figure 5: Coordinate ascent variational inference: iterate factor updates to maximise the ELBO.

Coordinate ascent provides a simple and intuitive optimisation strategy. By updating one component at a time whilst holding the others fixed, a complex optimisation problem is broken into simpler steps. Under certain factorisation assumptions, these updates can be derived in closed form. This idea underlies many classical variational inference algorithms and explains their computational efficiency.

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

Three Fundamental Models

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

Figure 6: 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).

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.

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:

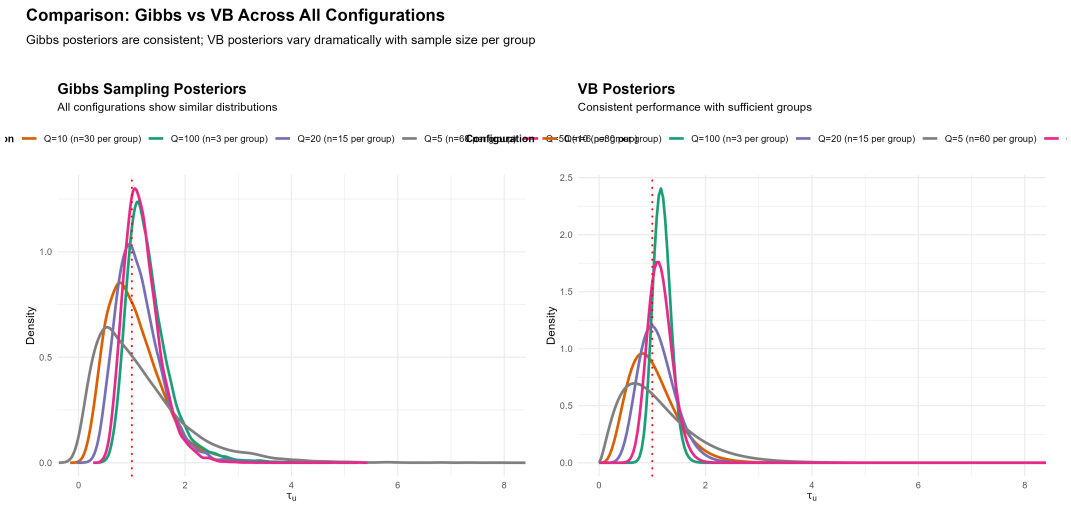

Figure 8: VB vs Gibbs for τ_u in Model 2: variational posterior is too narrow (under-dispersion).

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.

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.