On Variational Bayesian Methods

David Ewing

2026-01-27

Slide 1: Bayes' Theorem

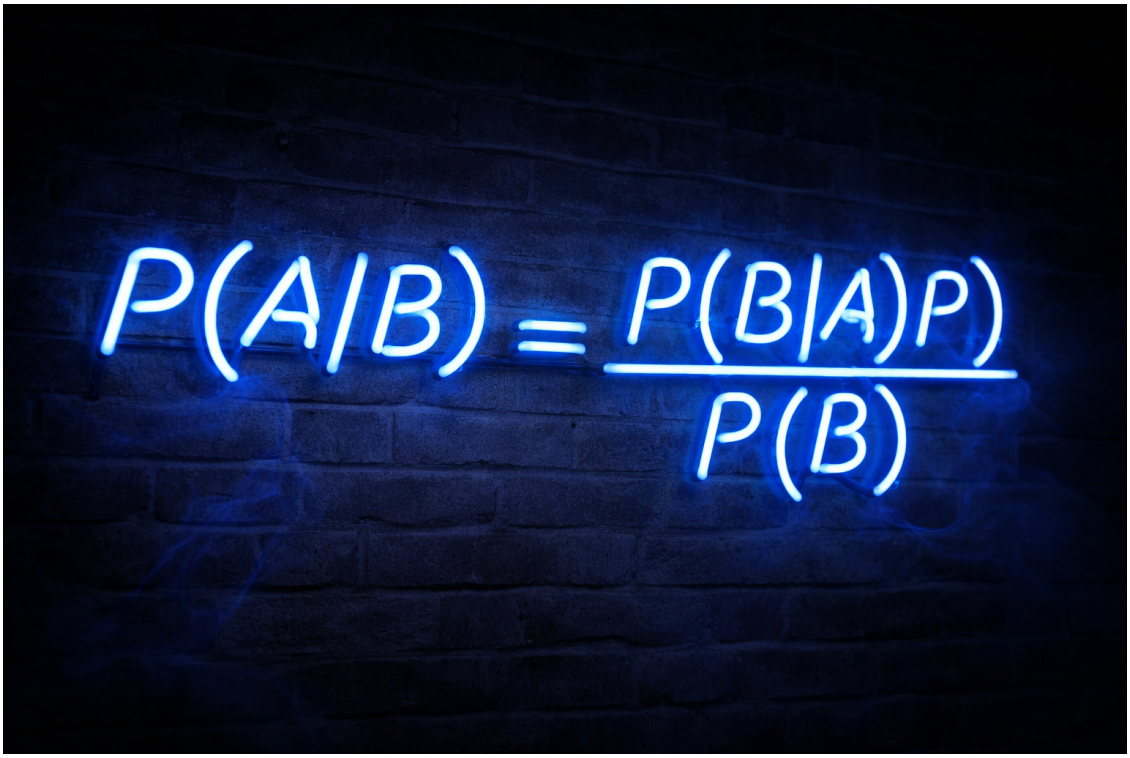

Figure 1: Bayes' theorem : prior, likelihood, posterior, and evidence.

(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

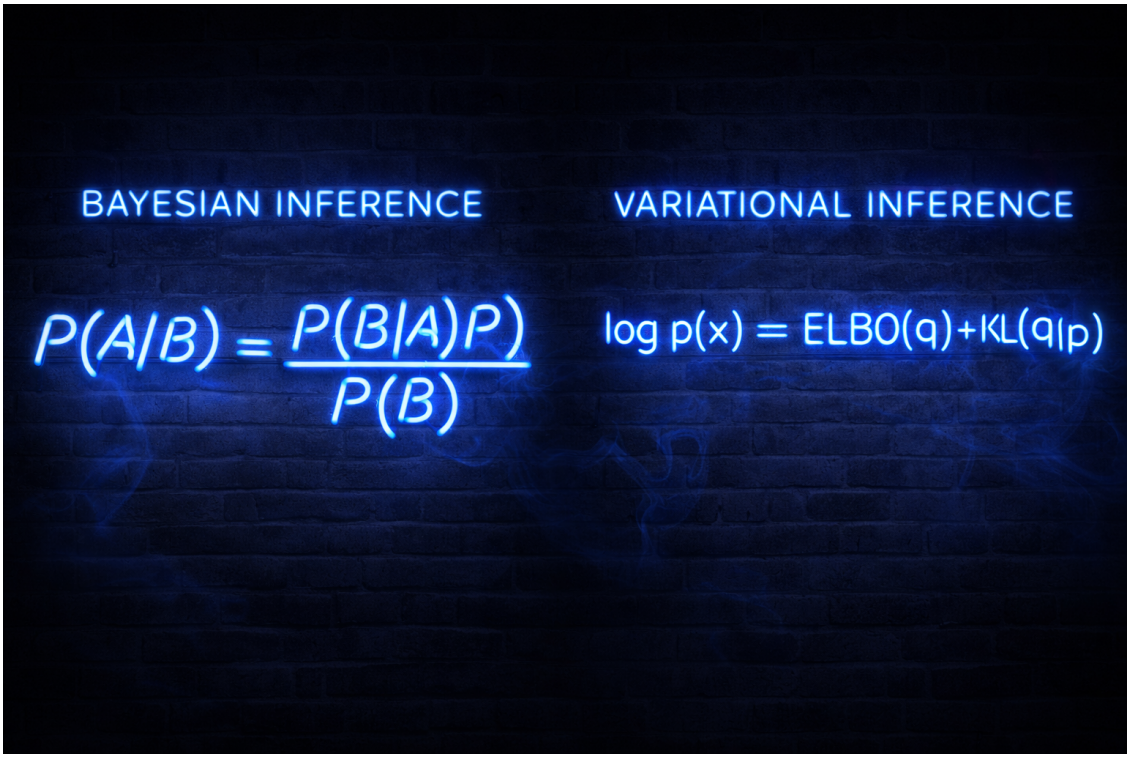

Figure 2: Exact Bayesian inference versus variational inference: sampling vs optimisation.

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

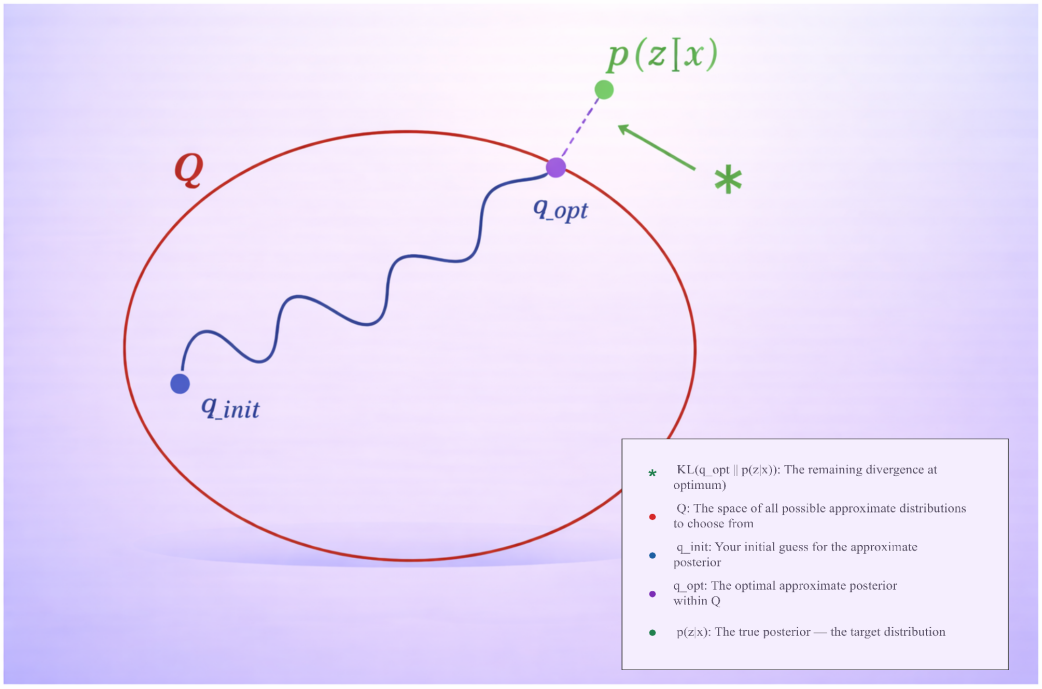

Figure 3: Visualising the search for q_{opt} within the variational family Q.

"This diagram, adapted from Dr David Blei of Columbia University, 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 6: Coordinate Ascent VI

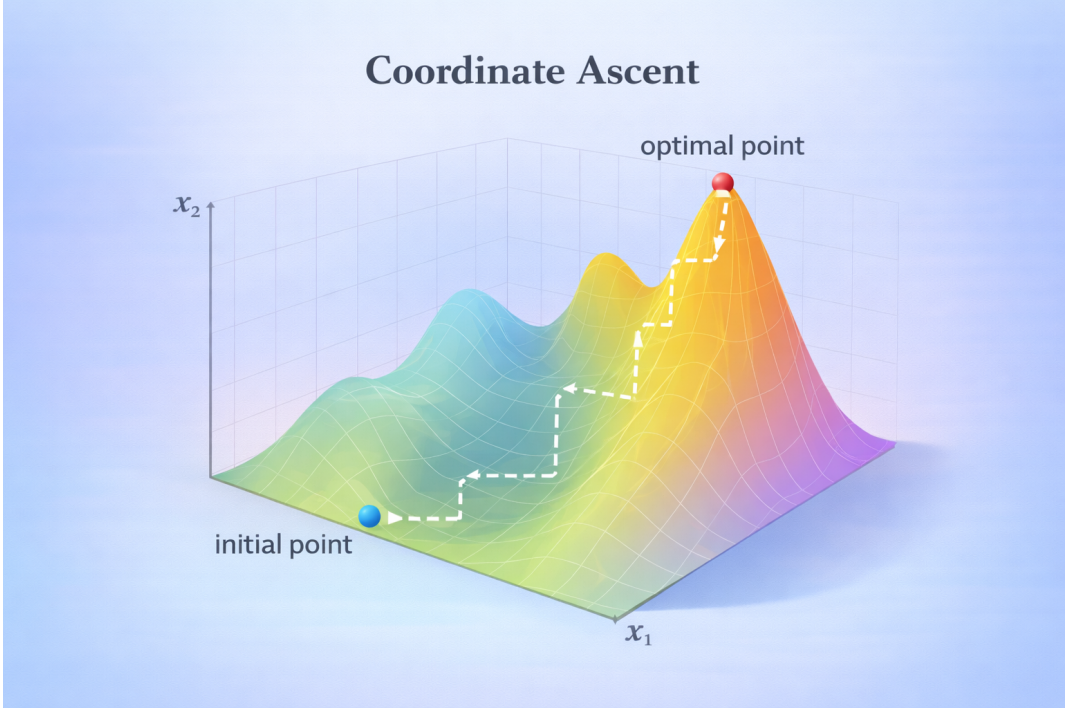

Figure 4: Coordinate ascent variational inference: iterate factor updates to maximise the ELBO.

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. | This slide illustrates the optimisation strategy behind variational inference in conditionally conjugate models. We use coordinate ascent: we cycle through each variational factor in turn, updating it while holding the others fixed. Each update improves the ELBO — the evidence lower bound — and we continue until convergence. The path shown here is schematic: we start at an initial point, and climb step by step along each coordinate direction until we reach the optimum. In the models we focus on, these updates are closed-form — thanks to conjugacy — which makes the procedure fast and scalable.

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

To orient the empirical comparisons, this slide introduces the three core models used throughout the project. Model 1 is a simple linear regression — no hierarchy, no random effects. Model 2 introduces hierarchical structure with Gaussian random effects. Model 3 is a hierarchical logistic model — structurally similar to M2, but with a Bernoulli likelihood. Each model includes regression coefficients , and where applicable, random effects u, residual precision _e, and random-effects precision _u. This side-by-side summary sets expectations for which variance components are present — and that's crucial, because under-dispersion is most severe in the hierarchical cases, especially for _u and _e.

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 5: 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).

This slide shows how we structure the variational approximation and derive the update equations. We begin with the full joint posterior — which couples all parameters through the likelihood. To make inference tractable, we impose a mean-field factorisation: we preserve the coupling between and u, but factor out the precision parameters _u and _e. This structure admits closed-form updates for each factor. The regression block — q(,u) — is Gaussian, and its mean and covariance are updated using expectations of the precision parameters. The precision parameters — q(_e) and q(_u) — are Gamma distributions, updated using expectations of squared residuals and squared random effects. All expectations are computed using the current variational parameters, and the loop repeats until the ELBO converges.

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 6: 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:

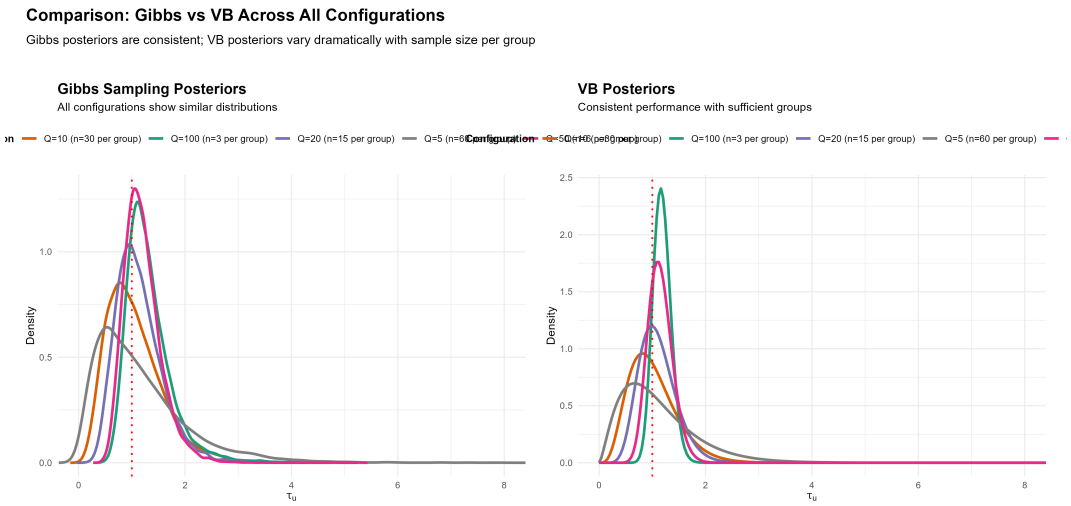

Figure 7: VB vs Gibbs for τ_u in Model 2: variational posterior is too narrow (under-dispersion).

This slide illustrates a key limitation of mean-field variational inference: under-dispersion. We're looking at the posterior for the random-effects precision _u in Model 2. On the left, we see Gibbs sampling posteriors — consistent across configurations. On the right, we see VB posteriors — noticeably narrower, especially when the number of groups is small. This narrowing reflects overconfidence: the variational posterior underestimates uncertainty. It's a direct consequence of the mean-field assumption — which imposes conditional independence and breaks the coupling that would otherwise inflate variance. The effect is systematic, and it's most severe in hierarchical models with limited group structure.