Narrative and Mathematical Structure for Variational Inference Presentation

Probabilistic Model

We begin by specifying a probabilistic model through a joint distribution over observed data and unknown parameters.

Let

$$\theta = (\beta, u, \tau_e, \tau_u)$$

denote the collection of unknown parameters, where:

- β are regression coefficients,
- $u = (u_1, \ldots, u_Q)$ are group-level random effects,
- τ_e is the residual precision,
- τ_u is the random-effects precision.

The joint distribution is defined as

$$p(y, \theta) = p(y \mid \theta)\, p(\theta),$$

where:

- $p(y \mid \theta)$ is the *likelihood*,
- $p(\theta)$ is the *prior distribution*.

Posterior Distribution

Bayesian inference targets the posterior distribution

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}.$$

The numerator,

$$p(y \mid \theta)\, p(\theta),$$

is the *joint distribution* of data and parameters.

The denominator,

$$p(y) = \int p(y \mid \theta)\, p(\theta)\, d\theta,$$

is called the *marginal likelihood* or *evidence*.

This integral is typically intractable for hierarchical models and is the primary obstacle to exact posterior inference.

Exact Inference versus Approximation

If the marginal likelihood were available in closed form, exact inference would be possible. However, for hierarchical models, this is generally not the case.

As a result, posterior inference must be approximated.

Exact Posterior Inference (Conceptual)

Conceptually, exact inference would correspond to finding an approximation $q(\theta)$ such that

$$q(\theta) = p(\theta \mid y).$$

In practice, this equality is not achievable for the model under consideration.

Variational Inference

Variational inference reframes posterior approximation as an optimisation problem.

A family of distributions $\mathcal{Q}$ is selected, and the optimal approximation is defined as

$$q^*(\theta) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(\theta) \,\|\, p(\theta \mid y)),$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence.

Mean-Field Factorisation

The most common variational family is the mean-field family, which assumes full independence among parameter blocks:

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

This assumption removes all posterior dependencies between parameters and enables tractable optimisation.

Implications of the Mean-Field Assumption

The mean-field assumption implies

$$\mathrm{Cov}_q(\theta_i, \theta_j) = 0 \quad \text{for } i \neq j.$$

While this simplifies computation, it is well known to produce *under-dispersed* posterior distributions, particularly in hierarchical settings.

Coordinate Ascent Interpretation

Under mean-field factorisation, variational inference proceeds via coordinate ascent.

Each factor is updated sequentially:

$$q(\theta_i) \leftarrow \arg\min_{q(\theta_i)} \mathrm{KL}(q(\theta) \,\|\, p(\theta \mid y)),$$

holding all other factors fixed at their current expectations.

Gibbs Sampling as a Reference Method

Gibbs sampling targets the true posterior distribution asymptotically.

Each parameter is sampled from its full conditional distribution:

$$\theta_i \sim p(\theta_i \mid \theta_{-i}, y).$$

This preserves posterior dependencies and provides a reference against which variational approximations can be evaluated.

Comparison Strategy

We compare Variational Bayes (VB) to Gibbs sampling with respect to:

- posterior means,
- posterior variances,

- posterior distributional shape,

for β, u_i, τ_e, and τ_u.

Regression Coefficients

For regression coefficients β, VB and Gibbs posteriors are generally similar.

This is expected, as β is strongly informed by the data and exhibits weak posterior dependence with other parameters.

Random Effects

The random effects u_i behave differently.

Each u_i is weakly informed when the group size is small, and its posterior uncertainty depends strongly on τ_u.

Under mean-field factorisation, this dependency is broken.

Random-Effects Precision

The precision parameter τ_u governs the variability of the random effects:

$$u_i \mid \tau_u \sim \mathcal{N}(0, \tau_u^{-1}).$$

If the posterior distribution of τ_u is overly concentrated, the posterior distributions of all u_i are artificially tightened.

Observed Behaviour

Empirically, we observe that:

- Gibbs posteriors for τ_u are stable across configurations,
- VB posteriors for τ_u become increasingly concentrated as group size decreases.

This induces systematic underestimation of uncertainty in the u_i.

Explanation

Variational inference minimises

$$\mathrm{KL}(q(\theta) \,\|\, p(\theta \mid y)),$$

which penalises mass where q assigns probability but p does not, but not the reverse.

This asymmetry favours concentrated approximations and explains the observed under-dispersion.

Key Takeaway

The primary issue is not bias in posterior means, but mis-calibration of posterior uncertainty.

Mean-field Variational Bayes systematically underestimates variance in hierarchical components.

Conclusion

Variational Bayes provides fast and scalable inference.

However, for hierarchical models with weak group-level information, mean-field assumptions distort posterior uncertainty.

Gibbs sampling remains the appropriate reference when accurate uncertainty quantification is required.