Running head: DEMYSTIFYING POSTERIOR DISTRIBUTIONS: A TUTORIAL ON THEIR DERIVATION

Demystifying Posterior Distributions: A Tutorial on Their Derivation

Han Du, Fang Liu, Zhiyong Zhang, and Craig Enders

Department of Psychology, University of California, Los Angeles

Department of Applied and Computational Mathematics and Statistics, University of Notre Dame

Department of Psychology University of Notre Dame

Department of Psychology, University of California, Los Angeles

Correspondence should be addressed to Han Du, Pritzker Hall, 502 Portola Plaza, Los Angeles, CA 90095.

Email: hdu@psych.ucla.edu.

Abstract

Bayesian statistics have gained significant traction across various fields over the past few decades. Bayesian statistics textbooks often provide both code and the analytical forms of parameters for simple models. However, they often omit the process of deriving posterior distributions or limit it to basic univariate examples focused on the mean and variance. Additionally, these resources frequently assume a strong background in linear algebra and probability theory, which can present barriers for researchers without extensive mathematical training. This tutorial aims to fill that gap by offering a step-by-step guide to deriving posterior distributions. We aim to make concepts typically reserved for advanced statistics courses more accessible and practical. This tutorial will cover two models: the univariate normal model and the multilevel model. The concepts and properties demonstrated in the two examples can be generalized to other models and distributions.

Demystifying Posterior Distributions: A Tutorial on Their Derivation

Over the past few decades, Bayesian statistics have gained widespread use, primarily as a result of the development of Markov chain Monte Carlo (MCMC) sampling techniques and advances in computational power (e.g., Van de Schoot et al., 2017). These methods have been applied in various fields, including cognitive psychology (e.g., Lee, 2008), developmental psychology (e.g., Van de Schoot et al., 2014; Walker et al., 2007), and social psychology (e.g., Marsman et al., 2017). In the Bayesian framework, parameters are considered random variables, and requires the specification of prior distributions. The posterior distributions are subsequently derived, and posterior point estimates and credible intervals are constructed based on a set of plausible parameter values sampled from these posterior distributions.

While software such as Stan (Carpenter et al., 2017), BUGS (Spiegelhalter et al., 1996), jags (Plummer et al., 2003), Mplus (Asparouhov & Muthén, 2010), and Blimp (Keller & Enders, 2021) enables users to sample from posterior distributions without requiring analytical derivations, these tools are often used as black boxes and rely on computationally intensive methods such as MCMC. Once samples from the posterior are obtained, users can estimate various properties of the parameters, such as posterior means and standard deviations. The use of MCMC typically involves iterative sampling, convergence diagnostics, and parameter tuning, which yield an approximation of the posterior under ideal conditions. As a result, MCMC methods can introduce both computational overhead and interpretive uncertainty. In contrast, for certainly models, posterior distributions can be derived analytically in a simple form as well-known distributions. They are faster to compute, easier to interpret, and more transparent in showing how the prior and data combine. Such solutions also facilitate downstream tasks such as computing posterior expectations or predictive distributions. It also serves as a valuable benchmark for evaluating the accuracy and efficiency of MCMC implementations. If MCMC samples deviate from the known posterior, this can indicate issues such as poor convergence, inadequate mixing, or prior misspecification.

Bayesian statistics textbooks often provide both code and the analytical forms of parameters for simple models. However, they often omit the process of deriving posterior distributions or limit it to basic univariate examples focused on the mean and variance. Furthermore, textbooks often assume a strong foundation in linear algebra and probability theory, which may pose challenges for researchers lacking such a background. As a result, methodological researchers without systematical training in statistics may still have limited intuition about how posterior distributions are actually constructed. As far as we know, no existing textbook or tutorial offers a step-by-step derivation of posterior distributions for multilevel models. The goal is to present material usually reserved for advanced mathematical statistics courses in a way that helps motivated researchers gain a deeper understanding of the essential features of Bayesian posterior distributions. With this foundation, researchers no longer need to rely solely on software. They can derive posterior distributions for their own models, directly code them for faster and more accurate inference, and gain clearer insight into how prior assumptions and observed data interact by studying the analytical form of the posterior. This also helps researchers better diagnose issues in their code when using existing software programs such as JAGS and Mplus.

This tutorial will cover two models: the univariate normal model and the multilevel model. If readers can master the derivation process for these two models, they will be well-equipped to extend the approach to more complex models. We focus on cases involving conjugate priors, which yield closed-form posteriors and make the derivation process more transparent and accessible. Conjugate priors are commonly used in applied Bayesian analysis and offer an ideal entry point for learning, as they allow us to demonstrate the core logic of posterior construction in a mathematically tractable way.

To familiarize readers with the properties of matrix operations, we provide a summary in the Appendix. Throughout the derivation process, key algebraic properties and equalities will be emphasized, with a preview presented in Table 1.

[Table 1]

Foundational Concepts

Bayes' theorem is a foundational concept in Bayesian modeling. Let $\boldsymbol{\theta}$ denote the unknown parameter(s), and $\boldsymbol{x}$ denote the observed data. The specified model determines the function $f\left(\boldsymbol{x}|\boldsymbol{\theta}\right)$, commonly referred to as the *likelihood*, where $\boldsymbol{x}$ on the left side of the vertical sign is random, and $\boldsymbol{\theta}$ on the right of the vertical sign is fixed or known. That is, $f\left(\boldsymbol{x}|\boldsymbol{\theta}\right)$ describes the probability of the data given one set of fixed parameters. The likelihood function is also often written as $L\left(\boldsymbol{\theta}|\boldsymbol{x}\right)$ to highlight that it represents the likelihood of observing a particular $\boldsymbol{x}$ when the true value of the parameter is $\boldsymbol{\theta}$. This expression is mathematically equivalent to $f\left(\boldsymbol{x}|\boldsymbol{\theta}\right)$. The goal of Bayes' theorem is to compute the *posterior distribution* $f\left(\boldsymbol{\theta}|\boldsymbol{x}\right)$, where $\boldsymbol{\theta}$ on the left side of the vertical sign is random, and $\boldsymbol{x}$ on the right of the vertical sign is fixed.

$$\begin{aligned} f\left(\boldsymbol{\theta}|\boldsymbol{x}\right) &= \frac{f\left(\boldsymbol{x}|\boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}\right)}{f\left(\boldsymbol{x}\right)} \\ &\propto f\left(\boldsymbol{x}|\boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}\right) \end{aligned} \tag{1}$$

where $f\left(\boldsymbol{\theta}\right)$ is the prior distribution, $f\left(\boldsymbol{x}|\boldsymbol{\theta}\right)$ is the likelihood, and $f\left(\boldsymbol{x}\right)$ is a normalizing constant that ensures that the area under the distribution curve $f\left(\boldsymbol{\theta}|\boldsymbol{x}\right)$ equals 1. It is not necessary to focus extensively on $f\left(\boldsymbol{x}\right)$ since it is constant that does not depend on $\boldsymbol{\theta}$. Therefore, we can use the proportionality sign to indicate that the posterior distribution is proportional to the product of the prior distribution and the likelihood.

Focusing on the terms that only involve $\boldsymbol{\theta}$ in Equation (1) allows us to employ an important concept for deriving posterior distributions, known as the *kernel*. A kernel refers to the part of a probability density or mass distribution that characterizes the shape of the distribution. In many cases, the kernel is easier to work with to derive the corresponding probability density or mass function. We can think of the kernel as the 'DNA' of each distribution. Each component in Equation (1), including $f\left(\boldsymbol{\theta}\right)$, $f\left(\boldsymbol{x}|\boldsymbol{\theta}\right)$, and $f\left(\boldsymbol{\theta}|\boldsymbol{x}\right)$, has a kernel. Thus, by deriving and identifying the kernel of $f(\boldsymbol{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})$, we can infer the form of the posterior distribution. For example, for a normal distribution $x \sim N\left(\mu, \sigma^2\right)$ where μ and σ^2 are unknown, its full probability density function is

$$\begin{aligned} f\left(x|\mu, \sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &\propto \frac{1}{\sqrt{\sigma^2}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \end{aligned} \tag{2}$$

In this equation, μ and σ^2 are the parameters, and $\frac{1}{\sqrt{2\pi}}$ is a normalizing constant that ensures the cumulative density is 1. Line 2 of Equation (2) represents the kernel of $f\left(x|\mu, \sigma^2\right)$, which captures the essence of the normal distribution's bell shape. In Figure 1, we plot the density function with $\mu = 0$ and $\sigma^2 = 1$ as the black line, and the kernel as the green line. The kernel preserves the normal distribution's shape, centered around 0, with a bell curve and the majority of values falling between -2 and 2. However, unlike the density function, the kernel does not ensure that the total area under the curve adds up to 1.

[Figure 1]

Another key concept in posterior distribution derivation is the *conjugate prior*. When specifying prior distributions such as $f(\boldsymbol{\theta})$ in Equation (2), various options are available. We need to choose both the family of the prior distribution and the values for the parameters in the prior distribution, known as *hyper-parameters*. In this paper, we will make use of a particular type of prior, referred to as a conjugate prior. A conjugate prior is a prior distribution that, when combined with certain types of likelihood, results in a posterior distribution that belongs to the same distributional family as the prior. We focus on conjugate priors in this tutorial primarily for pedagogical reasons. Conjugate prior–posterior pairs allow us to derive closed-form solutions step by step, which helps readers build intuition about how Bayesian updating works. These cases make the mathematical structure of Bayes' theorem fully transparent, enabling readers to see clearly how the prior and likelihood combine to form the posterior. While conjugate priors are limited in flexibility and may not always be appropriate for real data analysis, they provide a simple and analytically tractable starting point for learning. By working through these derivations, readers gain the mathematical foundation and conceptual understanding needed to later engage with more complex models with non-conjugate priors and require numerical methods like MCMC. We will also briefly discuss non-conjugate priors for comparison in the following sections.

Univariate Normal Model

Textbooks on Bayesian statistics frequently restrict the step-by-step derivations of posterior distributions to the univariate normal model with unknown mean and variance. We will begin with a review of this model, which will provide a foundational basis for exploring more complex models.

Consider a sample of n students who independently completed a mathematical test. Their scores can be denoted as $\boldsymbol{x} = \{x_1, ..., x_n\}$. We are interested in estimating the population mean and variance of mathematical ability based on this sample. We assume that the population of mathematical abilities follows a normal distribution $N\left(\mu, \sigma^2\right)$, where both the population mean μ and population variance σ^2 are unknown. Thus, we denote the likelihood function as $f\left(\boldsymbol{x}|\mu, \sigma^2\right)$, where $\boldsymbol{x}$ is a function of μ and σ^2 (which can also be expressed as $L\left(\mu, \sigma^2|\boldsymbol{x}\right)$). The likelihood alone provides the requisite information for frequentist inference. In a conditional probability function, it is important to note that the vertical line separates two components: the part before the vertical line represents the variables being determined and considered random, while the part after the vertical line represents the variables considered fixed. In Line 2 of Equation (3), we employ the proportionality sign to emphasize the kernel, as $\frac{1}{\sqrt{2\pi}}$ serves solely as a normalizing constant.

$$\begin{aligned} f\left(\boldsymbol{x}|\mu, \sigma^2\right) =& \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\ &\propto \left(\sigma^2\right)^{-n/2} exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right\} \end{aligned} \tag{3}$$

When dealing with a product of individual likelihoods, exponent rules allow us to rewrite the product across individuals as a sum within the exponential function. To draw Bayesian inferences regarding the parameters, it is necessary to determine their posterior distributions, as represented on the left side of Bayes' theorem in Equation (2). This will allow us to sample from or summarize the posterior distributions of μ and σ^2.

Prior Distributions

We first consider conjugate priors. For simplicity, we specify independently prior distributions for μ and σ^2, respectively. Strictly speaking, this tutorial discusses semi-conjugate priors, meaning only the conditional posterior distribution belongs to the same family as the prior. In contrast, fully conjugate priors for this example would require using a normal-inverse-gamma prior: a normal conditional prior for μ ($f\left(\mu|\sigma^2\right)$) and an inverse gamma prior for σ^2 ($f\left(\sigma^2\right)$).

The (semi-)conjugate prior for μ in the Gaussian (i.e., normal) likelihood is a normal distribution $N(\mu_0, \sigma_0^2)$, resulting in the conditional posterior distribution of μ also being normal. μ_0 and σ_0^2 are the mean and variance of the prior distributions, respectively. μ_0 serves as a prior guess about the population mean μ. σ_0^2 reflects our confidence in this initial estimate: a smaller σ_0^2 suggests strong confidence that μ is close to μ_0, while a larger σ_0^2 introduces greater uncertainty into the prior distribution. When σ_0^2 is sufficiently large, the prior distribution has minimal influence on the posterior estimation and inference because the normal curve is essentially flat over the plausible parameter space. In such cases, the prior is considered *non-informative* or *diffuse*. The determination of what value of σ_0^2 makes a prior non-informative depends on the sample size of the data. Researchers can conduct sensitivity analyses by varying σ_0^2, but the details of setting hyper-parameters are beyond the scope of this tutorial.

We begin by writing out the density function of the normal prior, utilizing the proportionality sign to focus on the kernel. Note that in the kernel, we omit $\frac{1}{\sqrt{2\pi\sigma_0^2}}$ since it is not a function of the unknown parameter μ. Again, the difference between Lines 1 (density function) and 2 (kernel) in Equation (4) is illustrated in Figure 1. Getting rid of $\frac{1}{\sqrt{2\pi}}$ does not affect the essential properties in a normal distribution shape.

$$f(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \tag{4}$$
$$\propto exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

The conjugate prior for σ^2 in the Gaussian likelihood is an inverse gamma distribution $IG(a, b)$, which results in the conditional posterior distribution of σ^2 also being inverse gamma. $a > 0$ is a shape parameter and $b > 0$ is a scale parameter. Researchers often use small hyper-parameters such as $a = 0.001$ and $b = 0.001$ to specify a non-informative inverse gamma prior distribution. We utilize the proportionality sign to focus on the kernel of the density function where we omit $\frac{b^a}{\Gamma(a)}$ since it is not a function of the unknown parameter σ^2.

$$f\left(\sigma^2\right) = \frac{b^a}{\Gamma\left(a\right)}\left(\sigma^2\right)^{-(a+1)} exp\left\{-\frac{b}{\sigma^2}\right\} \tag{5}$$
$$\propto \left(\sigma^2\right)^{-(a+1)} exp\left\{-\frac{b}{\sigma^2}\right\}$$

In Figure 2, we plot the density function with $a = 1$ and $b = 2$ as the black line, and the kernel as the green line. The kernel preserves the right-skewed shape of the density function; however, unlike the density function, its total area under the curve is less than 1.

[Figure 2]

For illustrative purposes, we assume μ and σ^2 are independent. Therefore, the joint prior distribution is $f\left(\mu, \sigma^2\right) = f\left(\mu\right) f\left(\sigma^2\right)$. After selecting and simplifying the prior distributions, we can proceed with the step-by-step derivation of each parameter's conditional posterior distribution.

Joint Posterior Distribution

We begin by writing the joint posterior distribution $f\left(\mu, \sigma^2|\boldsymbol{x}\right)$ by applying Equation (1). Specifically, we multiply the kernel in the likelihood (Equation 3), the kernel in the prior for μ (Equation 4), and the kernel in the prior for σ^2 (Equation 5) together. We first group the two terms outside the exponential functions together. Then, we apply **Property 1:** $\boldsymbol{exp\left(a+b\right) = exp\left(a\right) exp\left(b\right)}$

(summarized in Table 1). By utilizing Property 1, we combine the three exponential functions by adding the terms inside the three sets of curly braces to obtain a single set of curly braces.

$$\begin{aligned} f\left(\mu,\sigma^2|\boldsymbol{x}\right) &\propto f\left(\boldsymbol{x}|\mu,\sigma^2\right) f\left(\mu\right) f\left(\sigma^2\right) \\ &\propto \left(\sigma^2\right)^{-n/2} exp\left\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right\} \times exp\left\{-\frac{\left(\mu-\mu_0\right)^2}{2\sigma_0^2}\right\} \\ &\quad \times \left(\sigma^2\right)^{-(a+1)} exp\left\{-\frac{b}{\sigma^2}\right\} \\ &\propto \left(\sigma^2\right)^{-(n+2a+2)/2} exp\left\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2} - \frac{\left(\mu-\mu_0\right)^2}{2\sigma_0^2} - \frac{b}{\sigma^2}\right\} \text{(Property 1)} \end{aligned} \tag{6}$$

Conditional Posterior Distribution of μ

We have two options here. One approach is to use integration to compute the marginal posterior distributions as $f\left(\mu|\boldsymbol{x}\right) = \int f\left(\mu,\sigma^2|\boldsymbol{x}\right) d\sigma^2$ and $f\left(\sigma^2|\boldsymbol{x}\right) = \int f\left(\mu,\sigma^2|\boldsymbol{x}\right) d\mu$. Although the analytical derivation of these integrals is manageable in this case, in many problems, analytical and closed-form marginal distributions do not exist. A more widely used approach is to compute conditional posterior distributions $f\left(\mu|\boldsymbol{x},\sigma^2\right)$ and $f\left(\sigma^2|\boldsymbol{x},\mu\right)$. A full conditional posterior distribution describes the distribution of a parameter given the observed data and all other parameters. For example, in $f\left(\mu|\boldsymbol{x},\sigma^2\right)$, μ is treated as the unknown parameter while σ^2 is treated as known. With the conditional posterior distributions, we can use the Gibbs sampling algorithm to sample each variable one at a time sequentially (Gelfand & Smith, 1990). For example, for the tth iteration, sample $\mu^{(t)}$ based on $\sigma^{2(t-1)}$ in the previous iteration using $f\left(\mu|\boldsymbol{x},\sigma^2\right)$, and sample $\sigma^{2(t)}$ based on updated $\mu^{(t)}$ using $f\left(\sigma^2|\boldsymbol{x},\mu\right)$. In this way, we keep updating μ and σ^2 based on each other iteratively.

Step 1. Identify and Retain the Terms Involving μ*.* We derive the conditional posterior distribution of μ first. In this case, we assume that σ^2 is known and fixed, treating it as a constant. During estimation, the Gibbs sampler will plug in the current value of σ^2 into the conditional posterior distribution of μ for each iteration. Consequently, we need only to retain the terms involving μ from Equation (6) and write down them after the proportionality sign. The first two lines of Equation (7) use Property 1**:** $exp\left(a+b\right)=exp\left(a\right)exp\left(b\right)$. After separating the two exponential components, we can eliminate $exp\left\{-\frac{\beta}{\sigma^2}\right\}$ because it does not contain μ, resulting in the final line of Equation (7). If we plan to use MCMC methods to draw posterior samples, we can stop here, as the kernel in Equation (7) is sufficient.

$$\begin{aligned} f\left(\mu|\boldsymbol{x},\sigma^2\right) &\propto exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}-\frac{\left(\mu-\mu_0\right)^2}{2\sigma_0^2}-\frac{b}{\sigma^2}\right\} \\ &\propto exp\left\{-\frac{b}{\sigma^2}\right\}exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}-\frac{\left(\mu-\mu_0\right)^2}{2\sigma_0^2}\right\}\text{(Property 1)} \\ &\propto exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}-\frac{\left(\mu-\mu_0\right)^2}{2\sigma_0^2}\right\} \end{aligned} \tag{7}$$

Step 2. Expand Parentheses and Remove Components. Since we are utilizing a conjugate prior, the conditional posterior distribution $f\left(\mu|\boldsymbol{x},\sigma^2\right)$ should be normal. Our primary objective is to rearrange the components to piece together a normal kernel. In this kernel, μ is the random variable, while $\boldsymbol{x}$ and σ^2 are considered fixed. Hence, we want to combine the terms inside the curly braces into one normal kernel where μ appears in only one location. To achieve this, we must expand all parentheses to evaluate whether any components can be removed for simplification.

$$\begin{aligned} f\left(\mu|\boldsymbol{x},\sigma^2\right) &\propto exp\left\{-\frac{\sum_{i=1}^{n}\left(x_i-\mu\right)^2}{2\sigma^2}-\frac{\left(\mu-\mu_0\right)^2}{2\sigma_0^2}\right\} \\ &\propto exp\left\{-\frac{\sum_{i=1}^{n}\left(x_i^2+\mu^2-2x_i\mu\right)}{2\sigma^2}-\frac{\left(\mu^2+\mu_0^2-2\mu\mu_0\right)}{2\sigma_0^2}\right\} \\ &\propto exp\left\{-\frac{\sum_{i=1}^{n}x_i^2+n\mu^2-2\mu\sum_{i=1}^{n}x_i}{2\sigma^2}-\frac{\mu^2+\mu_0^2-2\mu\mu_0}{2\sigma_0^2}\right\} \end{aligned} \tag{8}$$

In the last line of Equation (8), we can see that $\sum_{i=1}^{n}x_i^2$ and μ_0^2 do not involve μ, and can be omitted. In addition, we replace $\sum_{i=1}^{n}x_i$ with $n\bar{x}$ where $\bar{x}$ is the sample mean to simplify the notation. The top line of Equation (9) below shows these simplifications.

$$\begin{aligned} f\left(\mu|\boldsymbol{x},\sigma^2\right) &\propto exp\left\{-\frac{n\mu^2-2\mu n\bar{x}}{2\sigma^2}-\frac{\mu^2-2\mu\mu_0}{2\sigma_0^2}\right\} \\ &\propto exp\left\{-\frac{n\mu^2\sigma_0^2-2\mu n\bar{x}\sigma_0^2+\mu^2\sigma^2-2\mu\mu_0\sigma^2}{2\sigma^2\sigma_0^2}\right\} \\ &\propto exp\left\{-\frac{\mu^2\left(n\sigma_0^2+\sigma^2\right)-2\mu\left(n\bar{x}\sigma_0^2+\mu_0\sigma^2\right)}{2\sigma^2\sigma_0^2}\right\} \\ &\propto exp\left\{-\frac{\mu^2-2\mu\left(n\bar{x}\sigma_0^2+\mu_0\sigma^2\right)/\left(n\sigma_0^2+\sigma^2\right)}{2\sigma^2\sigma_0^2/\left(n\sigma_0^2+\sigma^2\right)}\right\} \end{aligned} \tag{9}$$

Notice that the two terms in the top line of Equation (9) have different denominators. In Line 2, we introduce a common denominator so components inside the exponential function can be combined. In Line 3, we group the terms involving μ^2 and μ respectively. Then, in Line 4, we divide both the numerator and the denominator by $\left(n\sigma_0^2+\sigma^2\right)$ to ensure that μ^2 does not multiply any additional factors in the numerator, to be consistent with the posterior kernel of μ in Step 3.

Step 3. Compare with the Normal Kernel. At this point, Equation (9) represents the most simplified form we can achieve. Because we are using a conjugate prior, we know that the posterior distribution must have a kernel of the form $exp\left\{-\frac{(\mu-A)^2}{2B}\right\}\propto exp\left\{-\frac{\mu^2-2A\mu}{2B}\right\}$ where A is the posterior mean and B is the posterior variance. We can now compare $exp\left\{-\frac{\mu^2-2A\mu}{2B}\right\}$ with the last line of Equation (9) to identify A and B. Hence, $A=\left(n\bar{x}\sigma_0^2+\mu_0\sigma^2\right)/\left(n\sigma_0^2+\sigma^2\right)$ and $B=\sigma^2\sigma_0^2/\left(n\sigma_0^2+\sigma^2\right)$. Based on A and B, the conditional posterior distribution is as follows.

$$\begin{aligned} f\left(\mu|\boldsymbol{x},\sigma^2\right) &= N\left(\frac{n\bar{x}\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2},\frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}\right) \\ &= N\left(\frac{\frac{n}{\sigma^2}\bar{x}+\frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}},\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right) \end{aligned} \tag{10}$$

$\frac{\frac{n}{\sigma^2}\bar{x}+\frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}$ is the posterior mean and $\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}$ is the posterior variance for μ. Equations (10) and (9) differ by a normalizing constant, $\sqrt{\frac{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}{2\pi}}exp\left\{-\frac{\left(n\bar{x}\sigma_0^2+\mu_0\sigma^2\right)^2/\left(n\sigma_0^2+\sigma^2\right)}{2\sigma^2\sigma_0^2}\right\}$. Since this constant does not involve μ, we can focus directly on the kernel, as excluding it does not affect the posterior distribution derivation. The posterior variance is a function of data population variance (σ^2) and prior variance of μ (σ_0^2). The posterior mean is a weighted average of sample mean ($\bar{x}$) and prior mean of μ (μ_0). When $\sigma_0^2 \to \infty$, $\frac{1}{\sigma_0^2} \to 0$ and posterior mean converges to $\bar{x}$. In this case, the posterior mean coincides with the maximum likelihood estimation (MLE) of μ. This is why the prior distribution with large σ_0^2 is considered *non-informative.*

Conditional Posterior Distribution of σ^2

At each iteration, once the Gibbs sampler has completed the estimation of μ by drawing a random sample from the normal curve in Equation (10), it proceeds to estimate σ^2 with treating the sampled μ as a known constant in the posterior distribution of σ^2. Next, we derive the conditional posterior distribution of σ^2, assuming μ is known and treated as a constant.

Step 1. Identify and Retain the Terms Involving σ^2. We begin by selecting all functions involving σ^2 in the final line of Equation (6) and writing them down after the proportionality sign. In Line 2, we identify the common denominator to combine all terms. If we plan to use MCMC methods to draw posterior samples, we can stop here, as the kernel in Equation (11) is sufficient.

$$\begin{aligned} f\left(\sigma^2|\boldsymbol{x},\mu\right) &\propto \left(\sigma^2\right)^{-(n+2a+2)/2} exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}-\frac{b}{\sigma^2}\right\} \\ &\propto \left(\sigma^2\right)^{-(n+2a+2)/2} exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2+2b}{2\sigma^2}\right\} \end{aligned} \tag{11}$$

Deriving the posterior of μ required several simplification steps because μ appeared in multiple terms of the Equation (7). For σ^2, however, we do not need to expand the parentheses and remove components because σ^2 is by itself in the denominator inside the exponential function.

2. Compare with the Inverse Gamma Kernel. With the conjugate prior, we know that the posterior distribution must be inverse gamma and possess a kernel of the form $\left(\sigma^2\right)^{-(A+1)} exp\left\{-\frac{B}{\sigma^2}\right\}$. Now we can compare $\left(\sigma^2\right)^{-(A+1)} exp\left\{-\frac{B}{\sigma^2}\right\}$ with the last line of Equation (11) to identify A and B. Therefore, $A = \frac{n+2a}{2}$ and $B = \frac{\sum_{i=1}^{n}(x_i-\mu)^2+2b}{2}$. Based on A and B, the conditional posterior distribution is as follows.

$$f\left(\sigma^2|\boldsymbol{x},\mu\right) = IG\left(\frac{n}{2}+a, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}+b\right) \tag{12}$$

When $a \to 0$ and $b \to 0$, the prior becomes non-informative, and the posterior distribution is dominated by the likelihood. In this case, the posterior mode of σ^2 approaches MLE.

Non-Conjugate Priors

Conjugate priors are limited to specific distributions that pair with certain likelihoods. They may impose unrealistic assumptions, such as symmetry or unbounded parameter ranges, which may not align with our prior beliefs. In these cases, we will consider non-conjugate priors. For example, there are various potential non-conjugate priors for μ and σ^2 in the Gaussian likelihood. Here, we illustrate one such prior for σ^2. In hierarchical models with a small number of groups, Gelman (2006) suggested that a half-Cauchy prior might be more effective and inverse gamma prior, as it avoids distorting inferences in regions of high likelihood: $\sigma \sim half-Cauchy\left(\gamma\right)$. This prior leads to the following form for $f\left(\sigma^2|\boldsymbol{x},\mu,\right)$:

$$\begin{aligned} f\left(\sigma^2|\boldsymbol{x},\mu,\right) &\propto f\left(\boldsymbol{x}|\mu,\sigma^2\right) f\left(\sigma^2\right) \\ &\propto \left(\sigma^2\right)^{-n/2} exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right\} \times \frac{1}{1+\sigma^2/\gamma^2} \end{aligned} \tag{13}$$

Even after identifying the kernel of the conditional posterior distribution of σ^2, it is still challenging to identify which distribution family $f\left(\sigma^2|\boldsymbol{x},\mu,\right)$ belongs to. However, with the kernel in Equation (13), we can apply the Metropolis-Hastings (MH) algorithm to sample the posterior distribution (Gilks et al., 1996; Hastings, 1970). However, compared to the analytical posterior derived from a conjugate prior, it is less straightforward to assess the influence of the prior when using a non-conjugate prior.

Multilevel Model

A multilevel model is appropriate when the data have a hierarchical structure where observations are nested within higher-level groups, such as students within schools or repeated measures within individuals. For example, suppose there are N students from J different schools, and we are interested in analyzing their mathematics achievement scores. In this case, student-level data are nested within schools. The dataset includes several covariates: parents' social and economic status (ses), whether the school is private or public (pri), and the average social and economic status of the schools ($mses$). A multilevel model for this dataset can be specified as follows.

$$y_{ij} = b_{1j} + b_{2j} \times ses_{ij} + \varepsilon_{ij} \tag{14}$$

$$b_{1j} = \beta_{11} + \beta_{12} \times mses_j + \beta_{13} \times pri_j + u_{1j}$$

$$b_{2j} = \beta_{21} + \beta_{22} \times mses_j + \beta_{23} \times pri_j + u_{2j}$$

where y_{ij} and ses_{ij} indicate the mathematics score and ses of the ith student from the jth school respectively, b_{1j} and b_{2j} are the random effects of the jth school, and $mses_j$ and pri_j are the average ses and school type of the jth school. u_{1j} and u_{2j} are level-2 residuals and follow a multivariate normal distribution: $\boldsymbol{u_j} = (u_{1j}, u_{2j})' \sim MN\,(0, \boldsymbol{\Sigma})$. ε_{ij} is the level-1 residual and follows a univariate normal distribution: $\varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right)$.

To derive the posterior distribution of all fixed effects simultaneously, we employ a matrix notation for notation convenience and simplicity.

$$\boldsymbol{y_j} = \boldsymbol{X_j\beta} + \boldsymbol{Z_j u_j} + \boldsymbol{\varepsilon_j} \tag{15}$$

$\boldsymbol{y_j}$ is a $n_j \times 1$ vector of the outcome (mathematics scores) for n_j student in the jth school.

$\boldsymbol{X_j} = \begin{pmatrix} 1 & mses_j & pri_j & ses_{ij} & mses_j ses_{ij} & pri_j ses_{ij} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & mses_j & pri_j & ses_{n_j j} & mses_j ses_{n_j j} & pri_j ses_{n_j j} \end{pmatrix}$ is a $n_j \times p$ design matrix for the fixed effects, corresponding to the p fixed effects ($p = 6$ in this example).

$\boldsymbol{\beta} = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix}$ is a $1 \times p$ vector of fixed effects coefficients.

$\boldsymbol{Z_j} = \begin{pmatrix} 1 & ses_{ij} \\ \vdots & \vdots \\ 1 & ses_{n_j j} \end{pmatrix}$ is a $n_j \times q$ design matrix for the random effects, corresponding to the q random effects ($q = 2$ in this example). $\boldsymbol{u_j} = \begin{pmatrix} u_{1j} \\ \\ u_{2j} \end{pmatrix}$ is a $q \times 1$ vector of random effects in the jth school. Equation (15) can be applied to clustered and longitudinal data, regardless of whether the design is balanced.

There are three sets of unknown parameters: $\boldsymbol{\beta}$ (fixed effects), $\boldsymbol{\Sigma}$ (level-2 covariance matrix), and σ^2_ε (level-1 residual variance). The random effects $\boldsymbol{u_j}$ are known as the latent variables in the *data augmentation* framework. Alternatively, we can model b_{1j} and b_{2j}, which is equivalent to modeling $\boldsymbol{u_j}$. Assuming independence across J schools, the joint distribution of $\boldsymbol{u}$ and $\boldsymbol{y}$ is as follows, where $f\left(\boldsymbol{y_j}|\boldsymbol{X_j}, \boldsymbol{\beta}, \boldsymbol{Z_j}, \boldsymbol{u_j}\right)$ is based on Equation (15), and $f\left(\boldsymbol{u_j}|\boldsymbol{\Sigma}\right)$ is based on $\boldsymbol{u_j} \sim MN\left(0, \boldsymbol{\Sigma}\right)$. We further keep the kernel that only involving $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and σ^2_ε to simplify the equation.

$$\begin{aligned}
f\left(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{\Sigma}, \sigma^2_\varepsilon\right) =& \prod_{j=1}^{J} f\left(\boldsymbol{y_j}|\boldsymbol{X_j}, \boldsymbol{\beta}, \boldsymbol{Z_j}, \boldsymbol{u_j}, \sigma^2_\varepsilon\right) f\left(\boldsymbol{u_j}|\boldsymbol{\Sigma}\right) \qquad (16) \\
=& \prod_{j=1}^{J} \Bigg((2\pi)^{-nj/2} \left|\sigma^2_\varepsilon \boldsymbol{I_{nj}}\right|^{-\frac{1}{2}} exp\Bigg\{ -\frac{1}{2} \left(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta} - \boldsymbol{Z_j}\boldsymbol{u_j}\right)' \left(\sigma^2_\varepsilon \boldsymbol{I_{nj}}\right)^{-1} \\
& \left(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta} - \boldsymbol{Z_j}\boldsymbol{u_j}\right)\Bigg\} \times (2\pi)^{-q/2} \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} exp\left\{ -\frac{1}{2} \boldsymbol{u_j}' \boldsymbol{\Sigma}^{-1} \boldsymbol{u_j} \right\} \Bigg) \\
\propto& \prod_{j=1}^{J} \Bigg(\left|\sigma^2_\varepsilon \boldsymbol{I_{nj}}\right|^{-\frac{1}{2}} exp\Bigg\{ -\frac{1}{2} \left(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta} - \boldsymbol{Z_j}\boldsymbol{u_j}\right)' \left(\sigma^2_\varepsilon \boldsymbol{I_{nj}}\right)^{-1} \\
& \left(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta} - \boldsymbol{Z_j}\boldsymbol{u_j}\right)\Bigg\} \times \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} exp\left\{ -\frac{1}{2} \boldsymbol{u_j}' \boldsymbol{\Sigma}^{-1} \boldsymbol{u_j} \right\}
\end{aligned}$$

In deriving Equation (16), we used the data augmentation method to get the joint distribution of $\boldsymbol{u}$ and $\boldsymbol{y}$. An alternative approach is to integrate out $\boldsymbol{u}$ and use the marginal likelihood of $\boldsymbol{y}$ directly. Information criteria based on the marginal likelihood have generally shown higher detection rates compared to those based on the conditional likelihood (Du et al. 2024; Merkle et al. 2019; Tong et al. 2022; Zhang et al. 2019). However, since the marginal covariance matrix depends nonlinearly on both $\boldsymbol{\Sigma}$ and σ^2_ε, their conditional posterior distributions no longer have simple expressions with well-known distributions, although $\boldsymbol{\beta}$ still retains an analytical conditional posterior. We will illustrate this in a later section.

Prior Distributions

For simplicity, we specify independent prior distributions for $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and σ_ε^2, respectively. Given the likelihood, the (semi-)conjugate prior for $\boldsymbol{\beta}$ is a multivariate normal distribution $f\left(\boldsymbol{\beta}\right) = MN\left(\boldsymbol{\beta_0}, \boldsymbol{\mathit{\Sigma}_0}\right)$, and the resulting conditional posterior distribution of $\boldsymbol{\beta}$ will also be multivariate normal. In this multivariate normal prior, $\boldsymbol{\beta_0}$ represents the prior mean, and $\boldsymbol{\mathit{\Sigma}_0}$ denotes the prior variance-covariance matrix. $\boldsymbol{\beta_0}$ indicates a prior estimate of $\boldsymbol{\beta}$, while $\boldsymbol{\mathit{\Sigma}_0}$ reflects our confidence regarding the values of $\boldsymbol{\beta}$. The density function of the multivariate normal prior distribution of $\boldsymbol{\beta}$ is provided in Equation (17), where $|\mathit{\Sigma}_0|$ indicates the determinant of $\mathit{\Sigma}_0$. In Line 2, we remove $(2\pi)^{-\frac{p+1}{2}}\left|\mathit{\Sigma}_0\right|^{-\frac{1}{2}}$ to retain the kernel as it does not depend on $\boldsymbol{\beta}$.

$$\begin{aligned} f\left(\boldsymbol{\beta}\right) &= (2\pi)^{-\frac{p+1}{2}}\left|\mathit{\Sigma}_0\right|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)' \mathit{\Sigma}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right\} \\ &\propto exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)' \mathit{\Sigma}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right\} \end{aligned} \tag{17}$$

The conjugate prior for $\boldsymbol{\Sigma}$ is an inverse Wishart distribution $\boldsymbol{\Sigma} \sim IW(m, \boldsymbol{V})$ where m is the degrees of freedom and V is the scale matrix, since the resulting conditional posterior distribution of $\boldsymbol{\Sigma}$ is also inverse Wishart. The inverse Wishart distribution is a multivariate generalization of inverse Gamma distribution. If individuals have strong prior knowledge, they can set V close to the covariance matrix they expect, and choose a high degree of freedom m, indicating high confidence. If individuals less prior knowledge, they might set m lower. The inverse Wishart density function is in Equation (18). In addition to m and V in the density function, p is the dimension of $\boldsymbol{\Sigma}$, $\mathit{\Gamma}\left(\right)$ is the multivariate gamma function, and $tr\left(\boldsymbol{V}\boldsymbol{\mathit{\Sigma}}^{-\mathbf{1}}\right)$ indicates the trace of $\boldsymbol{V}\boldsymbol{\mathit{\Sigma}}^{-\mathbf{1}}$. On Line 2 of Equation (18), we remove the terms not involving $\boldsymbol{\Sigma}$ (i.e., $|\boldsymbol{V}|^{m/2}$ and $2^{mp/2}\mathit{\Gamma}\left(m/2\right)$) and keep the kernel for this prior distribution.

$$\begin{aligned} f\left(\boldsymbol{\Sigma}\right) &= \frac{|\boldsymbol{V}|^{m/2}|\boldsymbol{\mathit{\Sigma}}|^{-(m+p+1)/2} exp\left(-tr\left(\boldsymbol{V}\boldsymbol{\mathit{\Sigma}}^{-\mathbf{1}}\right)/2\right)}{2^{mp/2}\mathit{\Gamma}\left(m/2\right)}. \\ &\propto |\boldsymbol{\mathit{\Sigma}}|^{-(m+p+1)/2} exp\left(-tr\left(\boldsymbol{V}\boldsymbol{\mathit{\Sigma}}^{-\mathbf{1}}\right)/2\right) \end{aligned} \tag{18}$$

The conjugate prior for σ_ε^2 is an inverse gamma distribution $\sigma_\varepsilon^2 \sim IG(a, b)$, which is identical to the univariate normal case in Equation (5). We do not need to specify a prior distribution for $\boldsymbol{u_j}$, however, it is necessary to derive its posterior distribution. This is because the posterior distributions of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and σ_ε^2 depend on $\boldsymbol{u_j}$ as a data augmentation approach.

Joint Posterior Distribution

We write out the joint posterior distribution $f\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_\varepsilon^2, \boldsymbol{u} | \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{\Sigma}, \boldsymbol{y}\right)$ by applying Equation (1) (multiplying Equations 16, 17, 18, and 5 together).

$$
\begin{aligned}
f\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma_\varepsilon^2, \boldsymbol{u} | \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{\Sigma}, \boldsymbol{y}\right) \propto & \left\{\prod_{j=1}^{J} f\left(\boldsymbol{y_j}, \boldsymbol{u_j} | \boldsymbol{X_j}, \boldsymbol{\beta}, \boldsymbol{Z_j}, \boldsymbol{\Sigma}\right)\right\} f\left(\sigma_\varepsilon^2\right) f\left(\boldsymbol{\beta}\right) f\left(\boldsymbol{\Sigma}\right) \qquad (19) \\
\propto & \prod_{j=1}^{J} \left(\left|\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\left(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta} - \boldsymbol{Z_j}\boldsymbol{u_j}\right)' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1}\right.\right. \\
& \left.\left.\left(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta} - \boldsymbol{Z_j}\boldsymbol{u_j}\right)\right\} \times \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}\right\}\right) \\
& \times \left(\sigma_\varepsilon^2\right)^{-(a+1)} exp\left\{-\frac{b}{\sigma_\varepsilon^2}\right\} \times exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)' \Sigma_0^{-1}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)\right\} \\
& \times \left|\boldsymbol{\varSigma}\right|^{-(m+q+1)/2} exp\left(-tr\left(\boldsymbol{V}\boldsymbol{\varSigma^{-1}}\right)/2\right)
\end{aligned}
$$

The Gibbs sampler is applied to sample $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, σ_ε^2, and $\boldsymbol{u_j}$ iteratively: it first samples $\boldsymbol{\beta}$ from its conditional posterior distribution with treating $\boldsymbol{\Sigma}$, σ_ε^2, and $\boldsymbol{u_j}$ as known constants, samples $\boldsymbol{\Sigma}$ with treating the $\boldsymbol{\beta}$, σ_ε^2, and $\boldsymbol{u_j}$ as known, draws σ_ε^2 with treating the $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{u_j}$ as known, and draws $\boldsymbol{u_j}$ with treating the $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and σ_ε^2 as known.

Conditional Posterior Distribution of $\boldsymbol{\beta}$

First, we derive the conditional posterior distribution of $\boldsymbol{\beta}$.

1. Identify and Retain the Terms Involving $\boldsymbol{\beta}$. To simplify the notation, we denote $f\left(\boldsymbol{\beta} | \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}, \boldsymbol{\Sigma}, \sigma_\varepsilon^2, \boldsymbol{u}\right)$ as $f\left(\beta | \cdot\right)$, where $\cdot$ indicates all parameters that $\boldsymbol{\beta}$ is conditional on. By focusing on the terms involving $\boldsymbol{\beta}$ in Equation (19), we can significantly simplify the equation. We identify all functions involving $\boldsymbol{\beta}$ of Equation (19) and list them following the proportionality sign.

$$\begin{aligned} f\left(\boldsymbol{\beta}|\cdot\,\right) \propto & \prod_{j=1}^{J}\left(exp\left\{-\frac{1}{2}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)\right\}\right) \\ & \times exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)'\Sigma_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right\} \\ \propto & \, exp\left\{-\frac{1}{2}\sum_{j=1}^{J}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)\right\} \\ & \times exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)'\Sigma_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right\} \end{aligned} \tag{20}$$

In the first two lines of Equation (20), we retain the terms involving $\boldsymbol{\beta}$. In the bottom two lines, we move the product function $\prod\limits_{j=1}^{J}$ inside the exponential function, thereby transforming the product function into a sum function $\sum_{j=1}^{J}$.

2. Expand Parentheses and Remove Components. Since we utilize a conjugate prior, we know the conditional posterior distribution $f\left(\boldsymbol{\beta}|\cdot\,\right)$ follows a multivariate normal distribution. The goal is to rearrange the components to construct a multivariate normal kernel in which $\boldsymbol{\beta}$ is the random variable. To achieve this goal, the following derivation utilizes Properties 2 and 3 from Table 1. Another essential operation underlying the computational steps is the *distributive property for matrix multiplication*. This property is reviewed in the appendix. By applying the distributive property for matrix multiplication, we expand the parentheses of $\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)$ and

$(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\,\Sigma_0^{-1}\,(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$ from Equation (20) as Lines 1-2 and 3 in Equation (21), respectively.

$$
\begin{aligned}
f\left(\boldsymbol{\beta}|\cdot\right) \propto exp & \left\{-\frac{1}{2}\sum_{j=1}^{J}\Big[(\boldsymbol{X_j\beta})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j\beta})+(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})\right. \qquad (21)\\
& \left.-(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j\beta})-(\boldsymbol{X_j\beta})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})\Big]\right\}\\
& \times exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta}+\boldsymbol{\beta_0}'\Sigma_0^{-1}\boldsymbol{\beta_0}-\boldsymbol{\beta_0}'\Sigma_0^{-1}\boldsymbol{\beta}-\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta_0}\right)\right\}\\
\propto exp & \left\{-\frac{1}{2}\sum_{j=1}^{J}\Big[(\boldsymbol{X_j\beta})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j\beta})-2\,(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j\beta})\Big]\right\}\text{(Property 2)}\\
& \times exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}_0'\Sigma_0^{-1}\boldsymbol{\beta}\right]\right\}\text{(Property 2)}\\
\propto exp & \left\{-\frac{1}{2}\sum_{j=1}^{J}\Big[(\boldsymbol{X_j\beta})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j\beta})-2\,(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j\beta})\Big]\right.\\
& \left.-\frac{1}{2}\left[\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}_0'\Sigma_0^{-1}\boldsymbol{\beta}\right]\Big]\right\}\text{(Property 1)}
\end{aligned}
$$

In Line 1, we can get rid of $(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})$ since it is not a function $\boldsymbol{\beta}$, further simplifying the expression in Line 4.

In Line 2, we notice that both $(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j\beta})$ and $(\boldsymbol{X_j\beta})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})$ lead to scalar result. To simplify the expression, we can combine these two terms by applying **Property 2: $\boldsymbol{a'Bc}=\boldsymbol{c'Ba}$ if both sides yield a scalar** (summarized in Table 1). Hence, treating $(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})$ as $\boldsymbol{a}$, $\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}$ as $\boldsymbol{B}$, and $(\boldsymbol{X_j\beta})$ as $\boldsymbol{c}$, we find that $(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j\beta})=(\boldsymbol{X_j\beta})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})$. These duplicate terms are combined as a single term, resulting in $2\,(\boldsymbol{y_j}-\boldsymbol{Z_ju_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j\beta})$ in Line 4.

In Line 3, we can similarly combine $\boldsymbol{\beta_0}'\Sigma_0^{-1}\boldsymbol{\beta}$ and $\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta_0}$. With treating $\boldsymbol{\beta_0}$ as $\boldsymbol{a}$, $\boldsymbol{\Sigma_0^{-1}}$ as $\boldsymbol{B}$, and $\boldsymbol{\beta}$ as $\boldsymbol{c}$, we find that $\boldsymbol{\beta_0}'\Sigma_0^{-1}\boldsymbol{\beta}=\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta_0}$. These duplicate terms are combined into a single term, resulting in $2\boldsymbol{\beta_0}'\Sigma_0^{-1}\boldsymbol{\beta}$ in Line 5. Additionally, we can eliminate $\boldsymbol{\beta_0}'\Sigma_0^{-1}\boldsymbol{\beta_0}$ in Line 5, as it does not depend on $\boldsymbol{\beta}$, further simplifying the expression.

By applying Property 1: $exp\left(a+b\right)=exp\left(a\right)exp\left(b\right)$, we combine the two exponential functions in Lines 4-5, leading to the expressions in Lines 6 and 7.

Equation (21) is still not succinct enough. To refine it further, we will rearrange the terms to construct a multivariate normal kernel. First, we move the summation $\sum_{j=1}^{J}$ from outside the square brackets in Equation (21) to inside. Next, we rearrange the quadratic forms $(\sum_{j=1}^{J}(\boldsymbol{X_j}\boldsymbol{\beta})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j}\boldsymbol{\beta})$ and $\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta})$ at the end of Equation (21) into Line 1 of Equation (22). Additionally, we rearrange the terms $-2\sum_{j=1}^{J}(\boldsymbol{y_j}-\boldsymbol{Z_j}\boldsymbol{u_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\boldsymbol{\beta}$ and $-2\boldsymbol{\beta}_0'\Sigma_0^{-1}\boldsymbol{\beta}$ into Line 2 of Equation (22).

$$
\begin{aligned}
f\left(\boldsymbol{\beta}|\cdot\right) \propto exp & \left\{-\frac{1}{2}\left[\sum_{j=1}^{J}(\boldsymbol{X_j}\boldsymbol{\beta})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j}\boldsymbol{\beta})+\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta}\right.\right. \\
& \left.\left. -2\sum_{j=1}^{J}(\boldsymbol{y_j}-\boldsymbol{Z_j}\boldsymbol{u_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j}\boldsymbol{\beta})-2\boldsymbol{\beta}_0'\Sigma_0^{-1}\boldsymbol{\beta}\right]\right\} \\
\propto exp & \left\{-\frac{1}{2}\left[\boldsymbol{\beta}'\sum_{j=1}^{J}\left(\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\right)\boldsymbol{\beta}+\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta}\right.\right. \\
& \left.\left. -2\sum_{j=1}^{J}(\boldsymbol{y_j}-\boldsymbol{Z_j}\boldsymbol{u_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\boldsymbol{\beta}-2\boldsymbol{\beta}_0'\Sigma_0^{-1}\boldsymbol{\beta}\right]\right\} \\
\propto exp & \left\{-\frac{1}{2}\left[\boldsymbol{\beta}'\left(\sum_{j=1}^{J}\left(\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\right)+\Sigma_0^{-1}\right)\boldsymbol{\beta}\ \text{(Property 3)}\right.\right. \\
& \left.\left. -2\left(\sum_{j=1}^{J}(\boldsymbol{y_j}-\boldsymbol{Z_j}\boldsymbol{u_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}+\boldsymbol{\beta}_0'\Sigma_0^{-1}\right)\boldsymbol{\beta}\right]\right\}
\end{aligned}
\tag{22}
$$

In Lines 1, we move $\sum_{j=1}^{J}$ to the middle of $(\boldsymbol{X_j}\boldsymbol{\beta})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}(\boldsymbol{X_j}\boldsymbol{\beta})$ by expanding the parentheses of $(\boldsymbol{X_j}\boldsymbol{\beta})$ and $(\boldsymbol{X_j}\boldsymbol{\beta})'$. This leads to $\boldsymbol{\beta}'\sum_{j=1}^{J}\left(\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\right)\boldsymbol{\beta}$ in Line 3. Line 4 is the same as Line 2.

In Line 3, we can combine $\boldsymbol{\beta}'\sum_{j=1}^{J}\left(\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\right)\boldsymbol{\beta}$ and $\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta}$, as both are quadratic terms. Specifically, we can apply **Property 3:** $\boldsymbol{a'Ba+a'Ca=a'\left(B+C\right)a}$ (summarized in Table 1). By treating $\boldsymbol{\beta}$ as $\boldsymbol{a}$, $\sum_{j=1}^{J}\left(\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\right)$ as $\boldsymbol{B}$, and $\boldsymbol{\Sigma_0^{-1}}$ as $\boldsymbol{c}$, we have $\boldsymbol{\beta}'\sum_{j=1}^{J}\left(\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\right)\boldsymbol{\beta}+\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta}=\boldsymbol{\beta}'\left(\sum_{j=1}^{J}\left(\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\right)+\Sigma_0^{-1}\right)\boldsymbol{\beta}$ in Line 5.

Similarly, in Line 4, $\sum_{j=1}^{J}(\boldsymbol{y_j}-\boldsymbol{Z_j}\boldsymbol{u_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0'\Sigma_0^{-1}\boldsymbol{\beta}$ are summarized as a single term $\left(\sum_{j=1}^{J}(\boldsymbol{y_j}-\boldsymbol{Z_j}\boldsymbol{u_j})'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}+\boldsymbol{\beta}_0'\Sigma_0^{-1}\right)\boldsymbol{\beta}$ in Line 6.

3. Compare with the Multivariate Normal Kernel. As a multivariate normal distribution, the posterior distribution must have a kernel of the form $exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{A}\right)' B^{-1}\left(\boldsymbol{\beta}-\boldsymbol{A}\right)\right\} \propto exp\left\{-\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{B}^{-1}\boldsymbol{\beta}+A'\boldsymbol{B}^{-1}\boldsymbol{\beta}\right\}$. By comparing with the last two lines of Equation (22),we can get two relationships:

$$\boldsymbol{B}^{-1} = \left(\sum_{j=1}^{J}\left(\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\right)+\mathit{\Sigma}_0^{-1}\right)$$

$$\boldsymbol{A}'\boldsymbol{B}^{-1} = \left(\boldsymbol{y_j}-\boldsymbol{Z_j}\boldsymbol{u_j}\right)'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}+\boldsymbol{\beta}_0'\mathit{\Sigma}_0^{-1}$$

Solving these two equations leads to $\boldsymbol{B} = \left(\sum_{j=1}^{J}\left(\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\right)+\mathit{\Sigma}_0^{-1}\right)^{-1}$, $\boldsymbol{A}' = \left(\sum_{j=1}^{J}\left(\boldsymbol{y_j}-\boldsymbol{Z_j}\boldsymbol{u_j}\right)'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}+\boldsymbol{\beta}_0'\mathit{\Sigma}_0^{-1}\right)\boldsymbol{B}$, $\boldsymbol{A} = \boldsymbol{B}\left(\sum_{j=1}^{J}\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\left(\boldsymbol{y_j}-\boldsymbol{Z_j}\boldsymbol{u_j}\right)+\mathit{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right)$ ($\boldsymbol{B}$ is a symmetric matrix so $\boldsymbol{B}' = \boldsymbol{B}$). Based on $\boldsymbol{A}$ and $\boldsymbol{B}$, the conditional posterior distribution is as follows.

$$f\left(\boldsymbol{\beta}|\cdot\right) = MN\left(\boldsymbol{B}\left(\sum_{j=1}^{J}\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\left(\boldsymbol{y_j}-\boldsymbol{Z_j}\boldsymbol{u_j}\right)+\mathit{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right), \boldsymbol{B} = \left(\sum_{j=1}^{J}\left(\boldsymbol{X_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{X_j}\right)+\mathit{\Sigma}_0^{-1}\right)^{-1}\right) \tag{23}$$

We can see when the diagonal elements of $\mathit{\Sigma}_0$ growing large, or equivalently, the precision matrix $\mathit{\Sigma}_0^{-1}$ approaching the zero matrix. In this limit, the influence of the prior diminishes, and the posterior mean of β converges to the generalized least squares estimate, which is also equivalent to MLE.

Conditional Posterior Distribution of σ_ε^2

Next, we derive the conditional posterior distribution of σ_ε^2, treating $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}$, and $\boldsymbol{u_j}$ as constants.

1. Identify and Retain the Terms Involving σ_ε^2. Driving the conditional posterior distribution of σ_ε^2 is not different from the process in the univariate normal model. We denote the conditional posterior distribution as $f\left(\sigma_\varepsilon^2|\cdot\right)$ for brevity. Focusing on the terms involving σ_ε^2 in Equation (19), we express them after the proportionality sign:

$$\begin{aligned} f\left(\sigma_{\varepsilon}^{2}|\cdot\right) \propto & \prod_{j=1}^{J}\left(\left|\sigma_{\varepsilon}^{2}\boldsymbol{I_{nj}}\right|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)'\left(\sigma_{\varepsilon}^{2}\boldsymbol{I_{nj}}\right)^{-1}\right.\right. \\ & \left.\left.\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)\right\}\right) \times \left(\sigma_{\varepsilon}^{2}\right)^{-(a+1)} exp\left\{-\frac{b}{\sigma_{\varepsilon}^{2}}\right\} \\ & \propto \left(\sigma_{\varepsilon}^{2}\right)^{-\left(\frac{N}{2}+a+1\right)} exp\left\{-\frac{1}{2}\sum_{j=1}^{J}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)'\left(\sigma_{\varepsilon}^{2}\boldsymbol{I_{nj}}\right)^{-1}\right. \\ & \left.\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)-\frac{b}{\sigma_{\varepsilon}^{2}}\right\} \text{(Property 1)} \\ & \propto \left(\sigma_{\varepsilon}^{2}\right)^{-\left(\frac{N}{2}+a+1\right)} exp\left\{-\frac{\sum_{j=1}^{J}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)'\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)-2b}{2\sigma_{\varepsilon}^{2}}\right\} \end{aligned} \tag{24}$$

In Lines 1-2, we retain the terms involving σ_{ε}^{2}. We notice that there are multiple σ_{ε}^{2} terms that can be combined. In Line 1, $\prod_{j=1}^{J}\left|\sigma_{\varepsilon}^{2}\boldsymbol{I_{nj}}\right|^{-\frac{1}{2}} = \prod_{j=1}^{J}\left(\sigma_{\varepsilon}^{2}\right)^{-\frac{n_j}{2}} = \left(\sigma_{\varepsilon}^{2}\right)^{-\frac{\sum_{j=1}^{J} n_j}{2}} = \left(\sigma_{\varepsilon}^{2}\right)^{-\frac{N}{2}}$ where N is the total sample size across all schools. Thus, we combine this with $\left(\sigma_{\varepsilon}^{2}\right)^{-\frac{N}{2}} \times \left(\sigma_{\varepsilon}^{2}\right)^{-(a+1)}$, resulting in $\left(\sigma_{\varepsilon}^{2}\right)^{-\left(\frac{N}{2}+a+1\right)}$ in Line 3. Additionally, we merge all terms inside the exponential functions from Lines 1-2 to Lines 3-4 by applying Property 1. In the final line, we identify a common denominator of $2\sigma_{\varepsilon}^{2}$ inside the exponential function.

2. Compare with the Inverse Gamma Kernel. Similar to the univariate normal model, with the conjugate prior, we know that the posterior distribution must follow an inverse gamma distribution. Its kernel takes the form $\left(\sigma^{2}\right)^{-(A+1)} exp\left\{-\frac{B}{\sigma^{2}}\right\}$. By comparing this with the last line of Equation (24), we can get $A = \frac{N}{2}+a$ and $B = \frac{\sum_{j=1}^{J}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)'\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)+2b}{2}$. Based on these values of A and B, the conditional posterior distribution of σ^{2} is given by:

$$f\left(\sigma_{\varepsilon}^{2}|\cdot\right) = IG\left(\frac{N}{2}+a, \frac{\sum_{j=1}^{J}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)'\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j u_j}\right)}{2}+b\right) \tag{25}$$

Similar to the univariate normal example, when $a \to 0$ and $b \to 0$, the posterior distribution becomes dominated by the likelihood, and the posterior mode converges to MLE.

Conditional Posterior Distribution of $\boldsymbol{\Sigma}$

When deriving the conditional posterior distribution of $\boldsymbol{\Sigma}$, we treat $\boldsymbol{\beta}$, σ_{ε}^{2}, and $\boldsymbol{u_j}$ as constants.

1. Identify and Retain the Terms Involving Σ*, and Add Trace Function.* We select all functions involving Σ in Equation (19) and write down them after the proportionality sign.

$$\begin{aligned} f\left(\boldsymbol{\Sigma}|\cdot\right) &\propto \prod_{j=1}^{J}\left(|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\, exp\left\{-\frac{1}{2}\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}\right\}\right)|\boldsymbol{\Sigma}|^{-(m+q+1)/2}exp\left(-tr\left(\boldsymbol{V\Sigma^{-1}}\right)/2\right) \qquad (26)\\ &\propto |\boldsymbol{\Sigma}|^{-(J+m+q+1)/2}\, exp\left\{-\frac{1}{2}\sum_{j=1}^{J}\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}-\frac{1}{2}tr\left(\boldsymbol{V\Sigma^{-1}}\right)\right\}\\ &\propto |\boldsymbol{\Sigma}|^{-(J+m+q+1)/2}\, exp\left\{-\frac{1}{2}\sum_{j=1}^{J}tr\left(\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}\right)-\frac{1}{2}tr\left(\boldsymbol{V\Sigma^{-1}}\right)\right\} \text{(Property 4)}\\ &\propto |\boldsymbol{\Sigma}|^{-(J+m+q+1)/2}\, exp\left\{-\frac{1}{2}\sum_{j=1}^{J}tr\left(\boldsymbol{u_j}\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\right)-\frac{1}{2}tr\left(\boldsymbol{V\Sigma^{-1}}\right)\right\} \text{(Property 5)}\\ &\propto |\boldsymbol{\Sigma}|^{-(J+m+q+1)/2}\, exp\left\{-\frac{1}{2}tr\left(\sum_{j=1}^{J}\boldsymbol{u_j}\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\right)-\frac{1}{2}tr\left(\boldsymbol{V\Sigma^{-1}}\right)\right\} \text{(Property 6)}\\ &\propto |\boldsymbol{\Sigma}|^{-(J+m+q+1)/2}\, exp\left\{-\frac{1}{2}tr\left(\left(\sum_{j=1}^{J}\boldsymbol{u_j}\boldsymbol{u_j}'+\boldsymbol{V}\right)\boldsymbol{\Sigma}^{-1}\right)\right\} \text{(Property 6)} \end{aligned}$$

In Line 1, there are multiple $|\boldsymbol{\Sigma}|$ terms that can be combined as $|\boldsymbol{\Sigma}|^{-(J+m+q+1)/2}$ in Line 2. In addition, we can move the product function $\prod_{j=1}^{J}$ inside the exponential function, transforming the product function into a sum function $\sum_{j=1}^{J}$ in Line 2. In addition, the two exponential terms in Line 1 are combined as one term in Line 2.

In Line 2, we notice that $\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}$ yields a scalar. For a scalar, adding a trace function does not change anything. **Property 4: $\boldsymbol{tr\left(A\right)=A}$ is A is a scalar** (summarized in Table 1). That is, $\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}=tr\left(\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}\right)$ in Line 3.

Now, in Line 3, there are two trace expressions: $tr\left(\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}\right)$ and $tr\left(\boldsymbol{V\Sigma^{-1}}\right)$. However, in the first expression, Σ^{-1} appears in the middle, while in the second, it is positioned at the end. Our target kernel is of the form $|\boldsymbol{\Sigma}|^{-(A+p+1)/2}exp\left(-tr\left(\boldsymbol{B\Sigma^{-1}}\right)/2\right)$. To align with this form, we must place Σ^{-1} at the end in both expressions. This can be accomplished using the *cyclic property of the trace*. **Property 5: $\boldsymbol{tr\left(ABC\right)=tr\left(CAB\right)=tr\left(BCA\right)}$** (summarized in Table 1). With treating $\boldsymbol{A=u_j'}$, $\boldsymbol{B=\Sigma^{-1}}$, and $\boldsymbol{C=u_j}$, we conclude that $tr\left(\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}\right)=tr\left(\boldsymbol{u_j}\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\right)$, which is reflected in Line 4.

From Lines 4 to 5, we move the summation symbol $\sum_{j=1}^{J}$ inside the trace function. In this step, we apply **Property 6: $\boldsymbol{tr\left(A\right)+tr\left(A\right)=tr\left(A+B\right)}$** (summarized in Table 1).

Now in Line 5, there are two trace functions within the curly brackets: $tr\left(\sum_{j=1}^{J}\boldsymbol{u_j}\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\right)$ and $tr\left(\boldsymbol{V}\boldsymbol{\Sigma^{-1}}\right)$. To combine them, we apply Property 6 again, which leads to $tr\left(\left(\sum_{j=1}^{J}\boldsymbol{u_j}\boldsymbol{u_j}'+\boldsymbol{V}\right)\boldsymbol{\Sigma}^{-1}\right)$ in the final line.

2. Compare with the Inverse Wishart Kernel. With a conjugate prior, the posterior kernel takes the form as $|\boldsymbol{\Sigma}|^{-(A+p+1)/2}exp\left(-tr\left(\boldsymbol{B\Sigma^{-1}}\right)/2\right)$. By comparing with the last line in Equation (26), We can get $A=J+m$ and $B=\sum_{j=1}^{J}\boldsymbol{u_j}\boldsymbol{u_j}'+\boldsymbol{V}$. Based on A and B, the conditional posterior distribution is as follows.

$$f\left(\boldsymbol{\Sigma}|\cdot\right)=IW\left(J+m,\sum_{j=1}^{J}\boldsymbol{u_j}\boldsymbol{u_j}'+\boldsymbol{V}\right) \tag{27}$$

By studying Equation (27), we can see as $m\to q-1$ (where q is the dimension of Σ) and the diagonal elements of $\boldsymbol{V}\to\infty$, the prior contributes less information, allowing the data to dominate the posterior inference. This reflects a diffuse prior for the level-2 covariance structure.

Conditional Posterior Distribution of $\boldsymbol{u_j}$

As the final step, we derive the conditional posterior distribution of $\boldsymbol{u_j}$, treating $\boldsymbol{\beta}$, σ_ε^2, and $\boldsymbol{\Sigma}$ as constants.

1. Identify and Retain the Terms Involving $\boldsymbol{u_j}$. We extract all functions involving $\boldsymbol{u_j}$ in Equation (19) and present them after the proportionality sign of Equation (28). We use $f\left(\boldsymbol{u_j}|\cdot\right)$ to denote $f\left(\boldsymbol{u_j}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\Sigma},\sigma_\varepsilon^2\right)$ where $\cdot$ indicates all parameters that $\boldsymbol{u_j}$ is conditional on.

$$\begin{aligned} f\left(\boldsymbol{u_j}|\cdot\right)\propto exp&\left\{-\frac{1}{2}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j}\boldsymbol{u_j}\right)'\left(\sigma_\varepsilon^2\boldsymbol{I_{n_j}}\right)^{-1}\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}-\boldsymbol{Z_j}\boldsymbol{u_j}\right)\right\} \\ \times\, exp&\left\{-\frac{1}{2}\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}\right\}\end{aligned} \tag{28}$$

2. Expand Parentheses and Remove Components. Deriving the conditional posterior distribution of $\boldsymbol{u_j}$ follows a similar process as for $\boldsymbol{\beta}$. The key step is to expand the parentheses and eliminate terms that do not involve $\boldsymbol{u_j}$. First, we expand the parentheses of $(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta} - \boldsymbol{Z_j}\boldsymbol{u_j})' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta} - \boldsymbol{Z_j}\boldsymbol{u_j})$ in Equation (28), which leads to Lines 1-2 in Equation (29).

$$
\begin{aligned}
f\left(\boldsymbol{u_j}|\cdot\right) \propto exp&\left\{-\frac{1}{2}\left[(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta})' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta}) + (\boldsymbol{Z_j}\boldsymbol{u_j})' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{Z_j}\boldsymbol{u_j})\right.\right. \\
&\left.\left. - (\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta})' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{Z_j}\boldsymbol{u_j}) - (\boldsymbol{Z_j}\boldsymbol{u_j})' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta})\right]\right\} \\
&\times exp\left\{-\frac{1}{2}\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}\right\} \\
\propto exp&\left\{-\frac{1}{2}\left[(\boldsymbol{Z_j}\boldsymbol{u_j})' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{Z_j}\boldsymbol{u_j}) - 2\left(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta}\right)' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{Z_j}\boldsymbol{u_j})\right]\right\} (\text{Property } 2) \\
&\times exp\left\{-\frac{1}{2}\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}\right\}
\end{aligned}
\tag{29}
$$

In Line 1, we observe that both $(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta})' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{Z_j}\boldsymbol{u_j})$ and $(\boldsymbol{Z_j}\boldsymbol{u_j})' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta})$ result in a scalar. Therefore, by applying Property 2 ($\boldsymbol{a'Bc} = \boldsymbol{c'Ba}$ if both sides yield a scalar), we can combine these duplicate terms, yielding $2\left(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta}\right)' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{Z_j}\boldsymbol{u_j})$ in Line 4. And Lines 3 and 5 are identical. In this process, we also get rid of $(\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta})' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta})$ in Line 1 as it does not contain $\boldsymbol{u_j}$.

We can further simplify terms in Equation (29) by utilizing Property 1: $exp\left(a+b\right) = exp\left(a\right) exp\left(b\right)$. This allows us to rearrange the quadratic forms $((\boldsymbol{Z_j}\boldsymbol{u_j})' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{Z_j}\boldsymbol{u_j})$ and $\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j})$ in the bottom two lines of Equation (29) into Line 1 of

Equation (30).

$$
\begin{aligned}
f\left(\boldsymbol{u_j}|\cdot\right) \propto exp\Bigg\{-\frac{1}{2}\Big[&\left(\boldsymbol{Z_j u_j}\right)'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\left(\boldsymbol{Z_j u_j}\right)+\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j} \\
&-2\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}\right)'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\left(\boldsymbol{Z_j u_j}\right)\Big]\Bigg\} \text{ (Property 1)} \\
\propto exp\Bigg\{-\frac{1}{2}\Big[&\boldsymbol{u_j}'\boldsymbol{Z_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{Z_j u_j}+\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j} \\
&-2\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}\right)'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{Z_j u_j}\Big]\Bigg\} \\
\propto exp\Bigg\{-\frac{1}{2}\Big[&\boldsymbol{u_j}'\left(\boldsymbol{Z_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{Z_j}+\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{u_j} \text{ (Property 3)} \\
&-2\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}\right)'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{Z_j u_j}\Big]\Bigg\}
\end{aligned}
\tag{30}
$$

In Lines 1, we can expand the parentheses of $(\boldsymbol{Z_j u_j})$ and $(\boldsymbol{Z_j u_j})'$, which leading to $\boldsymbol{u_j}'\boldsymbol{Z_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{Z_j u_j}$ in Line 3. Line 4 is the same as Line 2.

In Line 3, the quadratic term $\boldsymbol{u_j}'\boldsymbol{Z_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{Z_j u_j}$ and $\boldsymbol{u_j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u_j}$ are combined using Property 3: $\boldsymbol{a'Ba}+\boldsymbol{a'Ca}=\boldsymbol{a'}\left(\boldsymbol{B}+\boldsymbol{C}\right)\boldsymbol{a}$. Treating $\boldsymbol{a}$ as $\boldsymbol{u_j}$, this yields $\boldsymbol{u_j}'\left(\boldsymbol{Z_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{Z_j}+\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{u_j}$ in Line 5. Line 6 remains identical to Line 4.

3. Compare with the Multivariate Normal Kernel. The posterior distribution appears to have a multivariate normal kernel, which might be readily recognized by researchers familiar with density functions. Even if one is not, consider this approach: the posterior distribution of $\boldsymbol{u_j}$ is proportional to the product of two normal density functions, $f\left(\boldsymbol{u_j}|\cdot\right)\propto f\left(\boldsymbol{y_j}|\boldsymbol{X_j},\boldsymbol{\beta},\boldsymbol{Z_j},\boldsymbol{u_j},\sigma_\varepsilon^2\right)f\left(\boldsymbol{u_j}|\boldsymbol{\Sigma}\right)$. Given this, it is highly likely that $f\left(\boldsymbol{u_j}|\cdot\right)$ is multivariate normal. Thus, we aim to approximate this distribution with a multivariate normal kernel.

A multivariate normal kernel has the form $exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{A}\right)'B^{-1}\left(\boldsymbol{\beta}-\boldsymbol{A}\right)\right\}\propto exp\left\{-\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{B}^{-1}\boldsymbol{\beta}+A'\boldsymbol{B}^{-1}\boldsymbol{\beta}\right\}$. By comparing with the bottom two lines of Equation (30), we can get two relationships.

$$\boldsymbol{B}^{-1}=\boldsymbol{Z_j}'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{Z_j}+\boldsymbol{\Sigma}^{-1}$$

$$\boldsymbol{A}'\boldsymbol{B}^{-1}=\left(\boldsymbol{y_j}-\boldsymbol{X_j\beta}\right)'\left(\sigma_\varepsilon^2\boldsymbol{I_{nj}}\right)^{-1}\boldsymbol{Z_j}$$

Solving these two equations leads to $\boldsymbol{B} = \left(\boldsymbol{Z_j}' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} \boldsymbol{Z_j} + \boldsymbol{\Sigma}^{-1}\right)^{-1}$,

$\boldsymbol{A}' = (\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta})' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} \boldsymbol{Z_j}\boldsymbol{B}$, $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{Z_j}' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta})$. Based on $\boldsymbol{A}$ and $\boldsymbol{B}$, the conditional posterior distribution is as follows.

$$f\left(\boldsymbol{u_j}|\cdot\right) = MN\left(\boldsymbol{B}\boldsymbol{Z_j}' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} (\boldsymbol{y_j} - \boldsymbol{X_j}\boldsymbol{\beta}), \boldsymbol{B} = \left(\boldsymbol{Z_j}' \left(\sigma_\varepsilon^2 \boldsymbol{I_{nj}}\right)^{-1} \boldsymbol{Z_j} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right) \tag{31}$$

Sampling Procedure

The full cadre of step-by-step Gibbs sampler procedure is given below. Interested researchers can follow the outlined steps and posterior distributions to implement their own Bayesian code for drawing posterior samples.

0. Initialization step: set initial values for $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$, $\sigma_\varepsilon^{2(0)}$, and $\boldsymbol{u_j}^{(0)}$ (for the individuals who have missing outcome). For the individuals who have missing outcome y_{ij}, set initial values for $y_{ij}^{(0)}$.

1. In the tth iteration, given $\boldsymbol{\Sigma}^{(t-1)}$, $\sigma_\varepsilon^{2(t-1)}$, and $\boldsymbol{u_j}^{(t-1)}$, sample $\boldsymbol{\beta}^{(t)}$ from Equation (23).

2. Given $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\Sigma}^{(t-1)}$, and $\boldsymbol{u_j}^{(t-1)}$, sample $\sigma_\varepsilon^{2(t)}$ from Equation (25).

3. Given $\boldsymbol{\beta}^{(t)}$, $\sigma_\varepsilon^{2(t)}$, and $\boldsymbol{u_j}^{(t-1)}$, sample $\boldsymbol{\Sigma}^{(t)}$ from Equation (27).

4. Given $\boldsymbol{\beta}^{(t)}$, $\sigma_\varepsilon^{2(t)}$, and $\boldsymbol{\Sigma}^{(t)}$, sample $\boldsymbol{u_j}^{(t)}$ from Equation (30).

5. If y_{ij} is missing, sample $y_{ij}^{(t)}$ from $f\left(\boldsymbol{y_j}|\boldsymbol{X_j}, \boldsymbol{\beta}, \boldsymbol{Z_j}, \boldsymbol{u_j}\right)$ given by Equation (15), which is $y_{ij} \sim N\left(\boldsymbol{X_{ij}}\boldsymbol{\beta}^{(t)} + \boldsymbol{Z_{ij}}\boldsymbol{u_j}^{(t)}, \sigma_\varepsilon^{2(t)}\right)$ where $\boldsymbol{X_{ij}}$ and $\boldsymbol{Z_{ij}}$ are the ith row of the design matrices $\boldsymbol{X_j}$ and $\boldsymbol{Z_j}$ respectively.

6. Repeat steps 1 to 5 until a sufficient number of posterior samples have been obtained.

Marginal Likelihood

As mentioned earlier, we can integrate out $\boldsymbol{u}$ and use the marginal likelihood to derive the posteriors of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and σ_ε^2. This is feasible but $\boldsymbol{\Sigma}$ and σ_ε^2 no longer have simple known posterior distributions. Based on Equation (15), after integrating out $\boldsymbol{u}$, the distribution for the outcome becomes $\boldsymbol{y_j} \sim MN\left(\boldsymbol{X_j}\boldsymbol{\beta}, \boldsymbol{V_j}\right)$

where $\boldsymbol{V_j} = \boldsymbol{Z_j}\boldsymbol{\Sigma}\boldsymbol{Z_j}' + \sigma_\varepsilon^2\boldsymbol{I}$. The marginal likelihood is as follows:

$$\begin{aligned} f\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{Z},\boldsymbol{\Sigma},\sigma_\varepsilon^2\right) =& \prod_{j=1}^{J} f\left(\boldsymbol{y_j}|\boldsymbol{X_j},\boldsymbol{\beta},\boldsymbol{V_j}\right) \\ =& \prod_{j=1}^{J}\left((2\pi)^{-nj/2}\left|\boldsymbol{V_j}\right|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}\left(\boldsymbol{y_j}-\boldsymbol{X_j}\boldsymbol{\beta}\right)'\left(\boldsymbol{V_j}\right)^{-1}\left(\boldsymbol{y_j}-\boldsymbol{X_j}\boldsymbol{\beta}\right)\right\}\right) \end{aligned} \tag{32}$$

We still assume a multivariate normal distribution prior for $\boldsymbol{\beta}$, $f\left(\boldsymbol{\beta}\right) = MN\left(\boldsymbol{\beta_0},\boldsymbol{\Sigma_0}\right)$. To derive the posterior of $\boldsymbol{\beta}$, we follow the same steps as in the conditional likelihood approach. Specifically, we (1) identify and retain the terms involving $\boldsymbol{\beta}$, (2) expand the quadratic form and remove constants, and (3) match the resulting expression to the multivariate normal kernel. This yields the multivariate normal posterior distribution:

$$f\left(\boldsymbol{\beta}|\cdot\right) = MN\left(\boldsymbol{B}\left(\sum_{j=1}^{J}\boldsymbol{X_j}'\boldsymbol{V_j^{-1}}\boldsymbol{y_j} + \Sigma_0^{-1}\boldsymbol{\beta}_0\right), \boldsymbol{B} = \left(\sum_{j=1}^{J}\left(\boldsymbol{X_j}'\boldsymbol{V_j^{-1}}\boldsymbol{X_j}\right) + \Sigma_0^{-1}\right)^{-1}\right) \tag{33}$$

The conditional posterior distributions for $\boldsymbol{\Sigma}$ and σ_ε^2 are as follows.

$$f\left(\boldsymbol{\Sigma}|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{Z},\sigma_\varepsilon^2\right) \propto f\left(\boldsymbol{\Sigma}\right)\prod_{j=1}^{J} f\left(\boldsymbol{y_j}|\boldsymbol{X_j},\boldsymbol{\beta},\boldsymbol{V_j}\right) \tag{34}$$

$$f\left(\sigma_\varepsilon^2|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{Z},\boldsymbol{\Sigma}\right) \propto f\left(\sigma_\varepsilon^2\right)\prod_{j=1}^{J} f\left(\boldsymbol{y_j}|\boldsymbol{X_j},\boldsymbol{\beta},\boldsymbol{V_j}\right) \tag{35}$$

Since as illustrated in Equation (32), $f\left(\boldsymbol{y_j}|\boldsymbol{X_j},\boldsymbol{\beta},\boldsymbol{V_j}\right)$ contains $|\boldsymbol{V_j}|$ and an exponential function of $\boldsymbol{V_j}$. These are nonlinear functions of $\boldsymbol{\Sigma}$ and σ_ε^2 that cannot be algebraically separated into standard conjugate kernel forms. More specifically, the conditional posterior of $\boldsymbol{\Sigma}$ cannot be written in inverse Wishart form because $\boldsymbol{\Sigma}$ is embedded within $\boldsymbol{Z_j}\boldsymbol{\Sigma}\boldsymbol{Z_j}'$ inside the covariance matrix $\boldsymbol{V_j}$, and thus also appears inside both the determinant and inverse of $\boldsymbol{V_j}$. And the conditional posterior of σ_ε^2 cannot be written in inverse Gamma form because $\boldsymbol{V_j}$ includes σ_ε^2 as part of a matrix sum, as $\boldsymbol{V_j} = \boldsymbol{Z_j}\boldsymbol{\Sigma}\boldsymbol{Z_j}' + \sigma_\varepsilon^2\boldsymbol{I}$. In this form, σ_ε^2 does not appear as a simple scalar factor applied to the entire covariance matrix, which breaks the algebraic structure needed for constructing a conjugate inverse gamma posterior. Without a known simple form, we can still use the MH algorithm to sample posteriors with Equations (34) and (35). Although this approach is feasible, it no longer allows us to directly observe how the prior contributes to the posterior or assess its influence as clearly as in the conjugate case.

Summary

To summarize, we have covered the fundamental concepts of Bayesian statistics and the derivation of posterior distributions. The key steps are as follows.

1. Identify Parameters and Specify Priors. Determine which parameters are unknown and define the prior distributions for these identified parameters. The model you specify for the data will determine the likelihood function of the parameters.

2. Formulate the Joint Posterior Distribution. Write out the joint posterior distribution for all unknown parameters. To simplify the derivation and minimize errors, we suggest focusing on the terms in the likelihood function and prior distributions that only involve the unknown parameters (i.e., kernel). Keeping too many terms can complicate the derivation are error-prone.

3. Compute Conditional Posterior Distribution. For each unknown parameter, we select the corresponding kernel in the joint posterior distribution to compute the conditional posterior distribution.

4. Simplify Conditional Posterior Distribution. When deriving the conditional posterior distribution, we suggest the following steps: (1) expand all parentheses to eliminate terms that do not involve the parameter of interest; (2) add or remove trace functions as needed; (3) reorder the components to facilitate easier manipulation; and (4) group similar components together to streamline the expression and highlight key features.

5. Compare with Kernel. We compare the kernel of the posterior distribution with the target kernel we suspect it might resemble. If we use conjugate priors, this process is more straightforward, as we can anticipate the general shape of the posterior distribution. However, if we are unsure whether conjugate priors are used, we need to determine which family the posterior distribution belongs to, which often relies on individual experience. In many cases, the posterior distribution may not fit any familiar family. In such situations, we can use the Metropolis-Hastings (MH) algorithm to sample from the posterior distribution based solely on the kernel.

For methodological researchers with knowledge in linear algebra but struggle to connect it with Bayesian derivation, this tutorial offers both conceptual clarity and practical skills. By walking through the full derivation of posterior distributions for the univariate normal and multilevel models, the step-by-step algebra fosters intuition about how priors and likelihoods combine, how conditional distributions are constructed, and how matrix operations such as transposition, inversion, and trace properties are applied in derivations. Readers learn not only how to recognize familiar distributional forms (e.g., normal, inverse gamma, inverse Wishart) from algebraic expressions but also how to manipulate and simplify complex joint and conditional posteriors. This empowers them to (1) better understand how Bayesian methods work, (2) identify potential issues in their software code, and (3) improve algorithm efficiency when working with more complex models or large datasets by utilizing opportunities to replace computational approximations with analytical solutions. By bridging the gap between abstract linear algebra and applied Bayesian inference, this tutorial fills a pedagogical void by offering the kind of hands-on mathematical reasoning often assumed but rarely shown in existing resources.

We hope this tutorial can be helpful to researchers who wish to learn how to derive posterior distributions on their own. The concepts and properties demonstrated in the two examples can be generalized to other models and distributions.

References

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis using mplus: Technical implementation (version 3).* Citeseer.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1), 1–32.

Du, H., Keller, B., Alacam, E., & Enders, C. (2024). Comparing dic and waic for multilevel models with missing data. *Behavior Research Methods*, *56*(4), 2731–2750.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, *1*(3), 515–534.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain monte carlo in practice* (pp. 339 – 357). London: Chapman & Hall.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Keller, B. T., & Enders, C. K. (2021). Blimp userâs guide (version 3). Retrieved from `www.appliedmissingdata.com/multilevel-imputation.html`

Lee, M. D. (2008). Three case studies in the bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15. doi: 10.3758/pbr.15.1.1

Marsman, M., Schönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A bayesian bird’s eye view of âReplications of important results in social psychology’. *Royal Society open science*, *4*(1), 160426. doi: 10.1098/rsos.160426

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *psychometrika*, *84*(3), 802–829.

Plummer, M., et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, pp. 1–10).

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W., & Lunn, D. (1996). Bugs: Bayesian inference using gibbs sampling. *Version 0.5,(version ii) http://www. mrc-bsu. cam. ac. uk/bugs*, *19*.

Tong, X., Kim, S., & Ke, Z. (2022). Impact of likelihoods on class enumeration in bayesian growth mixture modeling. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), *Quantitative psychology* (pp. 111–120). Cham: Springer International Publishing.

Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2014). A gentle introduction to bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 842–860. doi: 10.2196/10873

Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of "bayesian" articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217. doi: 10.1037/met0000100

Walker, L. J., Gustafson, P., & Frimer, J. A. (2007). The application of bayesian analysis to issues in developmental research. *International Journal of Behavioral Development*, *31*(4), 366–373. doi: /10.1177/0165025407077763

Zhang, X., Tao, J., Wang, C., & Shi, N.-Z. (2019). Bayesian model selection methods for multilevel irt models: A comparison of five dic-based indices. *Journal of Educational Measurement*, *56*(1), 3–27.

Appendix

Properties of Matrix Operations

Here are some key properties of matrix operations relevant to this tutorial. Assume r and s are scalars, and the matrices A, B, and C are of appropriate dimensions to ensure each operation is valid.

Multiplication and Addition

$$A + B = B + A$$

$$(A + B) + C = A + (B + C)$$

$$r\left(A + B\right) = rA + rB$$

$$\left(r + s\right)A = rA + sA$$

$$A\left(BC\right) = \left(AB\right)C$$

$$A\left(B + C\right) = AB + AC$$

$$\left(B + C\right)A = BA + CA$$

$$r\left(AB\right) = \left(rA\right)B = A\left(rB\right)$$

Transpose

$$(A + B)' = A' + B'$$

$$(AB)' = B'A'$$

$$(ABC)' = C'B'A'$$

$$(rA)' = rA'$$

$$\left(A'\right)' = A$$

Inverse

$$A^{-1}A = AA^{-1} = I$$

$$(AB)^{-1} = B^{-1}A^{-1}$$

$$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$$

$$(rA)^{-1} = r^{-1}A^{-1}$$

$$\left(A^{-1}\right)^{-1} = A$$

Trace

$$tr\left(A\right) = tr\left(A'\right)$$

$$tr\left(AB\right) = tr\left(BA\right)$$

$$tr\left(ABC\right) = tr\left(BCA\right) = tr\left(CAB\right)$$

$$tr\left(A + B\right) = tr\left(A\right) + tr\left(B\right)$$

$$tr\left(A+\right) = tr\left(A\right) + tr\left(B\right)$$

Determinants

$$\left|A\right|^{r}\left|A\right|^{s} = \left|A\right|^{r+s}$$

Table 1: properties and equalities

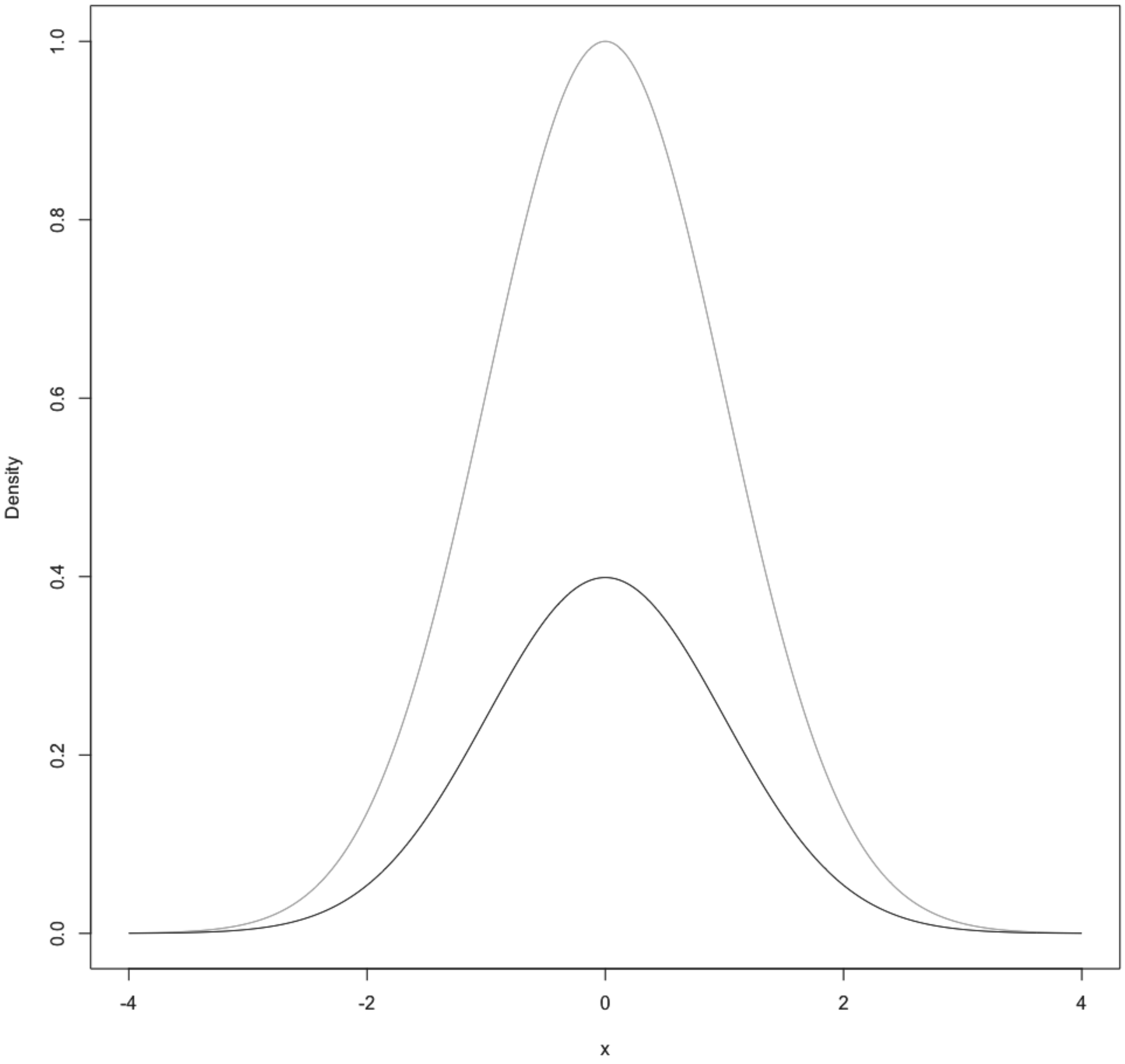
Property 1	$exp\left(a+b\right)=exp\left(a\right)exp\left(b\right)$ It is a property of exponential functions that allows the sum of exponents to be represented as the product of their individual exponentials.
Property 2	$a'Bc=c'Ba$ if both side yield a scalar If both sides yield a scalar, the transpose of the product of vector a with matrix B and vector c is equal to the transpose of the product of vector c with matrix B and vector a.
Property 3	$a'Ba+a'Ca=a'\left(B+C\right)a$ The sum of the quadratic forms $a'Ba$ and $a'Ca$ is equivalent to the quadratic form obtained by multiplying vector a by the sum of matrices B and C.
Property 4	$tr\left(A\right)=A$ if A is a scalar The trace of a scalar A is equal to the scalar itself, as the trace operation sums the diagonal elements of a matrix and a scalar can be viewed as a 1×1 matrix.
Property 5	$tr\left(ABC\right)=tr\left(CAB\right)=tr\left(BCA\right)$ This is the cyclic property of the trace function. The trace of a product of matrices remains invariant under cyclic permutations of the matrices.
Property 6	$tr\left(A\right)+tr\left(B\right)=tr\left(A+B\right)$ This is the additive property of the trace function, indicating that the trace of the sum of two matrices equals the sum of their individual traces.

Figure Captions

Figure 1. Density and Kernel of a Normal Distribution

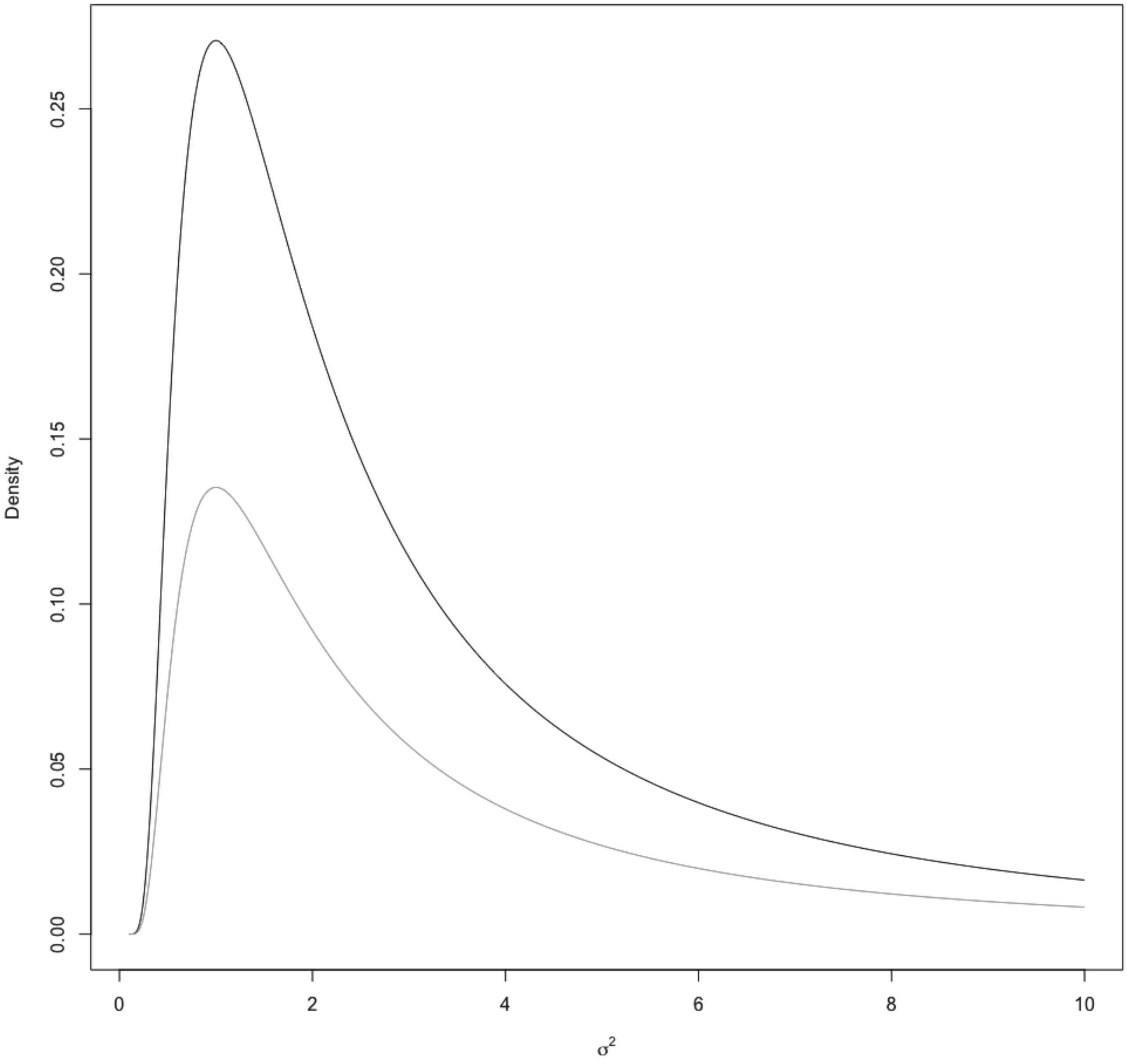
Figure 2. Density and Kernel of a Inverse Gamma Distribution

Figure 1: Density and Kernel of a Normal Distribution



We plot the density function with $\mu = 0$ and $\sigma^2 = 1$ as the black line, and the kernel as the green line.

Figure 2: Density and Kernel of a Inverse Gamma Distribution



We plot the density function with $a = 0$ and $b = 1$ as the black line, and the kernel as the green line.