mathematics

Review

Bayesian Nonlinear Models for Repeated Measurement Data: An Overview, Implementation, and Applications

Se Yoon Lee

Department of Statistics, Texas A&M University, College Station, TX 77843, USA; stat_math@tamu.edu

Abstract: Nonlinear mixed effects models have become a standard platform for analysis when data is in the form of continuous and repeated measurements of subjects from a population of interest, while temporal profiles of subjects commonly follow a nonlinear tendency. While frequentist analysis of nonlinear mixed effects models has a long history, Bayesian analysis of the models has received comparatively little attention until the late 1980s, primarily due to the time-consuming nature of Bayesian computation. Since the early 1990s, Bayesian approaches for the models began to emerge to leverage rapid developments in computing power, and have recently received significant attention due to (1) superiority to quantify the uncertainty of parameter estimation; (2) utility to incorporate prior knowledge into the models; and (3) flexibility to match exactly the increasing complexity of scientific research arising from diverse industrial and academic fields. This review article presents an overview of modeling strategies to implement Bayesian approaches for the nonlinear mixed effects models, ranging from designing a scientific question out of real-life problems to practical computations.

Keywords: Bayesian nonlinear hierarchical model; Bayesian nonlinear mixed effects models; inter-individual variation; intra-individual variation; Markov chain Monte Carlo technique

MSC: 62-01; 62C10; 62F15; 70K75; 60J22

check for updates

Citation: Lee, S.Y. Bayesian Nonlinear Models for Repeated Measurement Data: An Overview, Implementation, and Applications. *Mathematics* **2022**, *10*, 898. https://doi.org/10.3390/math10060898

Academic Editors: Vitaly Schetinin, Livija Jakaite, Dayou Li and Manuel Alberto M. Ferreira

Received: 26 January 2022
Accepted: 9 March 2022
Published: 11 March 2022

1. Introduction

One of the common challenges in biological, agricultural, environmental, epidemiological, financial, and medical applications is to make inferences on characteristics underlying profiles of continuous, repeated measures data from multiple individuals within a population of interest [1–4]. By 'repeated measures data' we mean the data type generated by observing a number of individuals repeatedly under differing experimental conditions, where the individuals are assumed to constitute a random sample from a population of interest. A common type of repeated measures data is longitudinal data such that the observations are ordered by time [5,6].

Linear mixed effects models for repeated measures data have become popular due to their straightforward interpretations, flexibility allowing correlation structure among the observations, and utility accommodating unbalanced and multi-level data structure (i.e., clustered designs that vary among individuals) [7,8]. The modeling framework is also intuitively appealing: the central idea that individuals' responses are governed by a linear model with slope or intercept parameters that vary among individuals seems to be appropriate in many scientific problems (for, e.g., see [9,10]). It also allows practitioners to test and evaluate multivariate causal relationships by conducting regression analysis at the population level. By preserving the multi-level structure in a single model, estimation or prediction for the analyses can take advantage of information borrowing [11].

For many applications, researchers often want to theorize that time courses of individual response commonly follow a certain nonlinear function dictated by a finite number of parameters [12]. These nonlinear functions are based on reasonable scientific hypotheses, typically represented as a differential equation system. By tuning the parameters, the shape of the function in terms of curvature, steepness, scale, height, etc., may change, which is used as the rationale behind describing heterogeneity between subjects. Nonlinear mixed effects models, also referred to as hierarchical nonlinear models, have gained broad acceptance as a suitable framework for these purposes [13–15]. Analyses based on this model are now routinely reported in various industrial problems, which is, in part, enabled by the breakthrough development of software [16–20]. The excellent books and review papers were published by [14,15,21]. Although their works were published more than 20 years ago, they still provide statisticians, programmers, and researchers with many pedagogical insights about the modeling framework, implementations, and practical applications of using the nonlinear mixed effects models.

While frequentist analysis of nonlinear mixed effects models has a long history, Bayesian analysis for the models was a relatively dormant field until the late 1980s. This is due primarily to the time-consuming nature of the calculations required for Bayesian computation to implement a Bayesian model [22]. Since the early 1990s, Bayesian approaches began to re-emerge, motivated both by exploitation of rapid developments in computing power and by the growing desire to quantify the uncertainty associated with parameter estimation and prediction [23–25]. Since then, Bayesian nonlinear mixed effects models, also called Bayesian hierarchical nonlinear models, have been extensively used in diverse industrial and academic researches, endowed with new computational tools providing a far more flexible framework for statistical inference exactly matching the increasing complexity of scientific research [26–31].

The objective of this article is to present an updated look at the Bayesian nonlinear mixed effects models. Although the works of [14,15] discuss some of the Bayesian approaches for the nonlinear mixed effects models, the main perspective adopted in the works is much more oriented to the frequentist framework, and prior distributions and Bayesian computing strategy explained in the works are quite outdated. In the literature, it is striking that very few research works provide an updated overview of the Bayesian methodologies on the nonlinear mixed effects models. Motivated by this, in this article, we provide an overview of modeling strategies to implement Bayesian approaches for the nonlinear mixed effects models, ranging from designing a scientific question out of real-life problems to practical computations. The novelty of this paper is as follow:

I. Guidance for Bayesian workflow to solve a real-life problem is provided for domain experts to facilitate efficient collaboration with quantitative researchers;
II. Recently developed prior distributions and Bayesian computation techniques for a basic model and its extensions are illustrated for statisticians to develop more complex models built on the basic model;
III. Illustrated methodologies can be directly exploited in diverse applications, ranging from small data to big data problems, for quantitative researchers, modeling scientists, and professional programmers working in diverse industries.

This article is organized as follows. In Section 2, we explore trends and workflow on the use of Bayesian nonlinear mixed effects. In Section 3, we motivate readers to understand why it is necessary to use the Bayesian nonlinear mixed effects model by illustrating four real-life problems, which will be conceptualized as a statistical problem. To solve the statistical problem, we suggest a basic version of the Bayesian nonlinear mixed effects models in Section 4, and its likelihood is analyzed in Section 5 wherein frequentist computations are briefly discussed. Section 6 describes modern Bayesian computation strategies to implement the basic model. Popularly used prior distributions are presented in Section 7. Section 8 discusses model selection, and Section 9 reviews recent advances and extensions that build on the basic model. Finally, Section 10 concludes the article.

2. Trends and Workflow of Bayesian Nonlinear Mixed Effects Models

2.1. Rise in the Use of Bayesian Approaches for the Nonlinear Mixed Effects Models

As of January 1970 to December 2021, PubMed.gov (https://pubmed.ncbi.nlm.nih.gov/, accessed on 23 February 2022) searched of "nonlinear mixed-effect models" yielded 6288 publications. Among the published articles, nearly 94% of works used frequentist approaches (5929 articles), while only 6% of works adopted Bayesian approaches (359 articles). Figure 1 displays a bar plot based on the published articles, categorized by the frequentist and Bayesian approaches over time. In the panel, it is observed that, until the late 1980s, the Bayesian research was nearly dormant, but since the early 1990s, Bayesian works begin to re-emerge, and the gap between frequentist and Bayesian works becomes gradually narrower as time evolves.

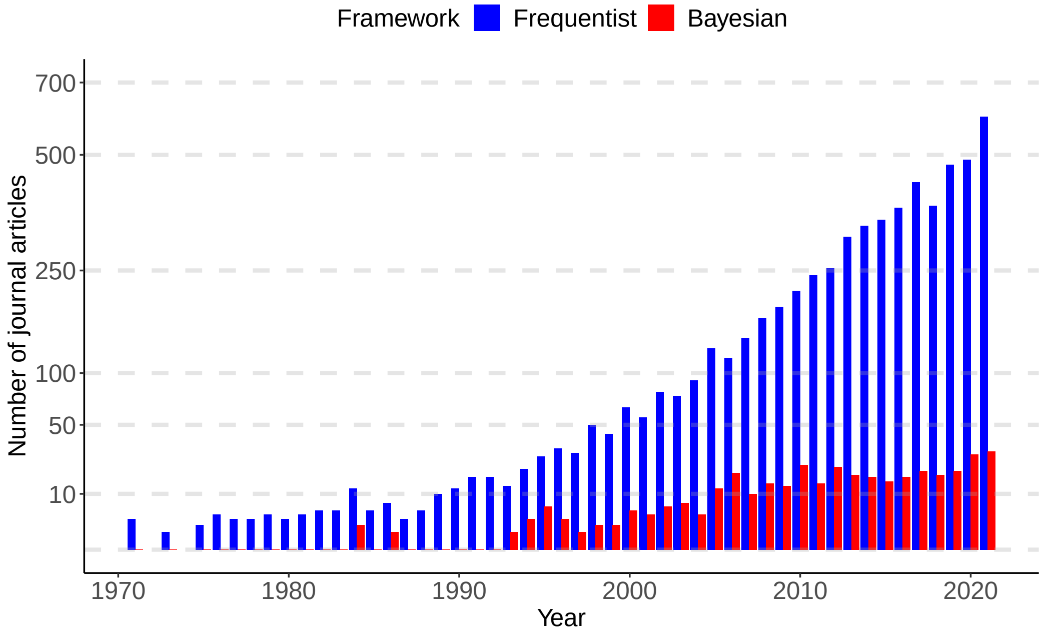

Figure 1. Publication trends of the nonlinear mixed-effect models categorized by frequentist and Bayesian frameworks. the x-axis represents the year from 1 January 1970 to 31 December 2021. The value on the y-axis is the number of published articles in each year (data sources: PubMed.gov, accessed on 23 February 2022).

The dormancy of the Bayesian approaches until the late 1980s is mainly due to the time-consuming nature of the calculations based on a sampling scheme, which was previously impossible by the limitation of computing power. Fortunately, a breakthrough in computer processors (for, e.g., Am386 in 1991, Pentium Processor in 1993, etc.) took place in the early 1990s, driving the computing revolution to solve computationally intense problems, and this has given the statistical community the ability to solve statistical questions by using Bayesian methods. This timeline is also aligned with the widespread Markov chain Monte Carlo (MCMC) sampling techniques in the Bayesian community [32,33]. Since then, the Bayesian community has been gradually gaining the momentum to leverage the rapidly growing developments of computing power, and now, assorted Bayesian software packages (e.g., JAGS [34], BUGS [35], and Stan [17]) are available for researchers to answer scientific questions arising from industrial and academic research.

To understand the rise of the Bayesian approaches, we first want to understand what will be some of the advantages of using Bayesian methods over frequentist methods in the context of nonlinear mixed effects models. As the primary focus of this review paper is to provide the readers with some insight on methodologies and practical implementation of using Bayesian approaches, our comparison and exposition below are described from an operational viewpoint. Table 1 summarizes the modeling strategies of using the frequentist and Bayesian approaches for the nonlinear mixed effects models. Broadly speaking, the usual estimation method of the frequentist computation is optimization, while that of the Bayesian computation is sampling. Normally, it is known that the former is much faster than the latter. This is not surprising because a sampling scheme, by its nature, needs to explore a wide range of the parameter space, whereas the optimization only needs to find the best point estimate, which is often described by the maximum likelihood estimate. In many practical problems, widely used frequentist optimization algorithms are the first-order approximation [36], Laplace approximation [37], and stochastic approximation of expectation-maximization algorithm [38]. They will be briefly discussed in Section 5.4. As for the Bayesian sampling algorithms, combinations of Gibbs sampler [39], Metropolis-Hastings algorithm [40], Hamiltonian Monte Carlo [41], and No-U-Turn sampler [42] are popularly used, among many others [43–45]. We explain these in detail in Section 6.

Table 1. Comparison of modeling strategies used in frequentist and Bayesian approaches for the nonlinear mixed effects models from an implementational viewpoint.

Characteristic	**Frequentist**	**Bayesian**
Estimation objective	Maximize a likelihood [14,15,21]	Sample from a posterior [22,28,30]
Computation algorithm	First-order approximation [36], Laplace approximation [37], and stochastic approximation of EM algorithm [38]	Gibbs sampler [39], Metropolis-Hastings algorithm [40], Hamiltonian Monte Carlo [41], and No-U-Turn sampler [42]
Software	SAS [46], NONMEM [47], MONOLIX [48], NLMIXR [18]	JAGS [49], BUGS [35], STAN [17], BRMS [50]
Advantages	Relatively fast computation speed, the objectivity of inference results, and widely available software packages to implement complex models	Inherent uncertainty quantification, better small sample performance, and utility of prior knowledge
Disadvantages	Needs large-sample theory for uncertainty quantification and cannot incorporate prior knowledge	Needs high computing power for big data and requires Bayesian expertise in prior elicitation

The stark difference between using frequentist and Bayesian approaches may be the procedure of describing an uncertainty underlying the parameter estimation for the nonlinear mixed effects models. Here, the parameter which is of primary interest is the population-level parameters (also called fixed effects), typical values for the individual-level parameters. In many cases, frequentist 95% confidence intervals for the parameters of the models are constructed by assuming that asymptotic normality of maximum likelihood estimator holds in a finite sample study, which is actually the most accurate in large sample scenario [51]. Most frequentist software packages, such as NONMEM [13,47], MONOLIX [48], and NLMIXR [18], by default, may print out a 95% confidence interval of the form, "$\textit{Estimate} \pm 1.96 \times \textit{ Standard Error}$", or some transformation of the lower and upper bounds, if necessary, such that the *Standard Error* is calculated by using (observed) Fisher information matrix [52–54]. Using such a scheme in small sample studies is highly likely to overlook the gap between the reality of the data and the idealistic asymptotic situation.

In contrast, as for the Bayesian approaches, the large-sample theory is not needed for the uncertainty quantification, and the procedure to obtain 95% posterior credible intervals is a lot easier than obtaining 95% confidence intervals (See Chapter 4 of [55]). Furthermore, Bayesian credible intervals based on percentiles of posterior samples allow for a strongly skewed distribution, wherein frequentist confidence intervals (based on large-sample theory) may induce a non-negligible approximating error due to the deviation from the asymptotic normality. Along with that, Bayesian methods are highly appreciated when researchers wish to incorporate prior knowledge from previous studies into the model so that posterior inference provides the researchers with an updated view on the problem, possibly with a more accurate estimation. Using prior information would be particularly useful in small-sample contexts [56,57]. For example, in medical device clinical trials, some opportunities and challenges in developing a new medical device are: (i) there is often a great deal of prior information for a medical device; (ii) a medical device evolves in relatively small increments from previous generations to a new generation; (iii) there are only a few numbers of patients for the trials; and (iv) companies need to make a rational decision promptly to reduce cost. In those settings, Bayesian methods have been demonstrated to be suitable, and their proper use is guided by Food and Drug Administration [58,59].

2.2. Bayesian Workflow

We outline the first two steps in the Bayesian workflow of using Bayesian nonlinear mixed effect models described in Figure 2. The panel includes some mathematical notations that are consistently used throughout the paper. These notations will be clearly understood later. The aim of our explanation at this point is to provide readers with a blueprinted plan to implement Bayesian modeling strategies for the nonlinear mixed effects models. We assume that readers are familiar with basic concepts and generic workflow in Bayesian statistics; see [55,60,61] for those basic concepts and refer to the review paper by [62] and references therein for detailed concepts and general terminologies used in workflow, such as prior and posterior predictive checks, and prior elicitation, etc.

The first step of the Bayesian research cycle is (a) standard research cycle [63,64]. Some early activities at this step involve reviewing literature, defining a problem, and specifying a research question and a hypothesis. After that, researchers specify which analytic strategy would be taken to solve the research question and suggest possible model types, followed by data collection. The data type arising in this process may include a response variable and some covariates that are grouped longitudinally, which then formulates repeated measures data of a population of interest. Furthermore, if there appears to be some nonlinear temporal tendency at each subject, then a possible model type for the analysis is a nonlinear mixed effects model [15,65].

The second step of the Bayesian research cycle is (b) Bayesian-specific workflow. Logically, the first thing to do at this step is to determine prior distributions (see Step (b)–(i) in Figure 2). The selection of priors is often viewed as one of the most crucial choices that a researcher makes when implementing a Bayesian model, as it can have a substantial impact on the final results [66]. As exemplified earlier in the context of Bayesian medical device trials, using a prior in small sample studies may improve the estimation accuracy, but an unthoughtful choice of priors would lead to a significant bias in estimation. Prior elicitation effort would require Bayesian expertise to formulate domain expert's knowledge in a probabilistic form [67]. Strategies for prior elicitation include asking domain experts to provide suitable values for the hyperparameters of the prior [68,69]. After prior is specified, one can check the appropriateness of the priors through prior predictive checking process [70]. For almost all practical problems, prior distribution of Bayesian nonlinear mixed effect models can be hierarchically represented as follow: (1) a prior for the parameters used in likelihood, often called 'population-level model' in the literature of mixed effects modeling; and (2) a prior for the parameters used in the population-level model and for the parameters describing the residual errors used in likelihood. It is important to note that the former type of prior distribution (that is, (1)) is also a requirement to implement frequentist approaches for the nonlinear mixed effects model, as a name of 'distribution for random effects'. Essentially, the defining factor of the Bayesian framework is the latter type of prior distribution (that is, (2)), which is fixed in the frequentist framework, as a name of 'fixed effects'. Some prior options of the latter type will be discussed in Section 7.

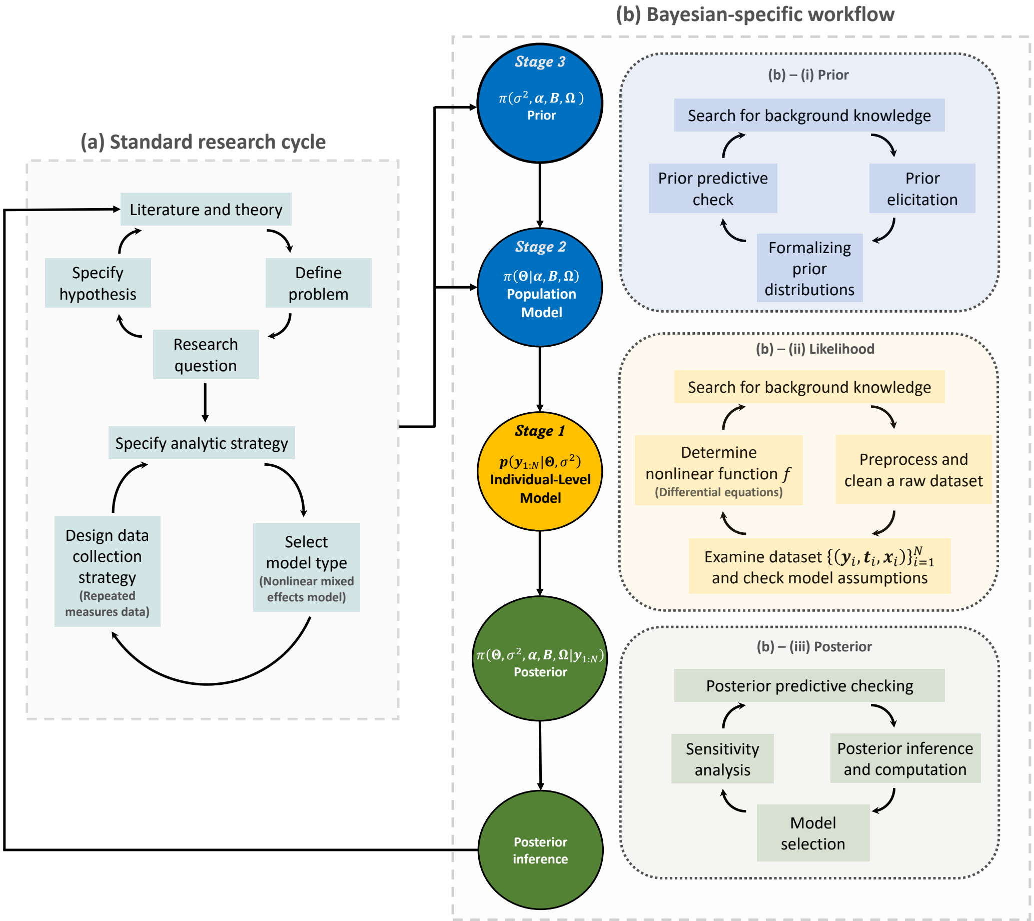
(b) Bayesian-specific workflow
(a) Standard research cycle
Literature and theory
Specify hypothesis
Define problem
Research question
Specify analytic strategy
Design data collection strategy
(Repeated measures data)
Select model type
(Nonlinear mixed effects model)
Stage 3
$\pi(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$
Prior
Stage 2
$\pi(\boldsymbol{\Theta}|\boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$
Population Model
Stage 1
$p(\boldsymbol{y}_{1:N}|\boldsymbol{\Theta}, \sigma^2)$
Individual-Level Model
$\pi(\boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}|\boldsymbol{y}_{1:N})$
Posterior
Posterior inference
(b) – (i) Prior
Search for background knowledge
Prior predictive check
Prior elicitation
Formalizing prior distributions
(b) – (ii) Likelihood
Search for background knowledge
Determine nonlinear function f
(Differential equations)
Preprocess and clean a raw dataset
Examine dataset $\{(\boldsymbol{y}_i, \boldsymbol{t}_i, \boldsymbol{x}_i)\}_{i=1}^{N}$ and check model assumptions
(b) – (iii) Posterior
Posterior predictive checking
Sensitivity analysis
Posterior inference and computation
Model selection

Figure 2. The Bayesian research cycle. A research cycle using Bayesian nonlinear mixed effects model comprises two steps: (**a**) standard research cycle and (**b**) Bayesian-specific workflow. Standard research cycle involves literature review, defining a problem and specifying the research question and hypothesis. Bayesian-specific workflow comprises three sub-steps: (b)–(i) formalizing prior distributions based on background knowledge and prior elicitation; (b)–(ii) determining the likelihood function based on a nonlinear function f; and (b)–(iii) making a posterior inference. The resulting posterior inference can be used to start a new research cycle. Distributions for prior, likelihood, and posterior are colored in blue, yellow, and green, respectively. $\boldsymbol{\Theta}$, model matrix; σ^2, error variance parameter; $\boldsymbol{\alpha}$, intercepts; $\boldsymbol{B}$, coefficient matrix; $\boldsymbol{\Omega}$, covariance matrix; $p(.)$, probability distribution; $\pi(.)$, prior or posterior probability distribution; $\{(\mathbf{y}_i, \mathbf{t}_i, \mathbf{x}_i)\}_{i=1}^{N}$, data.

The second task is to determine the likelihood function (see Step (b)–(ii) in the panel). At this time, the raw dataset collected in (a) standard research cycle should be cleaned and preprocessed. Before embarking on more serious statistical modeling, it is a common practice to get some insight about the research question via exploratory data analysis and have a discussion with domain experts such as clinical pharmacologists, clinicians, physicians, engineers, etc. To some extent, eventually, all these efforts are to determine a nonlinear function (denoted as f in this paper) that best describes the temporal profiles of all subjects. This nonlinear function is a *known* function because it should be specified by researchers. In other words, the branch of the nonlinear mixed effects models belongs to parametric statistics. However, one technical challenge is that, in many problems, such a nonlinear function is represented as a solution of a differential equation system [71,72], and therefore there is no guarantee that we can conveniently work with a closed-form expression of the nonlinear function. For example, if researchers wish to work with nonlinear differential equations [73,74], then some approximation via differential equation solver [75,76] may be needed to calculate the nonlinear function. As such, most software packages dedicated to implementing a nonlinear mixed effect model, or, more generally, a Bayesian hierarchical model, are equipped with several built-in differential equation solvers [17,47,77]. For instance, visit the website (https://mc-stan.org/docs/2_29/stan-users-guide/ode-solver.html, accessed on 20 February 2022) to see some functionality supported in Stan [17].

Finally, the likelihood is combined with the prior to form the posterior distribution (see Step (b)–(iii) in the panel). Given the important roles that the prior and the likelihood have in determining the posterior, this step must be conducted with care. The implementational challenge at this step is to construct an efficient MCMC sampling algorithm. The basic idea behind MCMC here is the construction of a sampler that simulates a Markov chain that is converging to the posterior distribution. One can use software packages if prior distributions to be implemented in Bayesian models exist in the list of prior options available in the packages. Otherwise, professional programmers and Bayesian statisticians are needed to make codes manually; this review paper will be useful for that purpose. Another activity important at this step is to compare multiple models with different priors and nonlinear functions, specified in Step (b)–(i) and (ii), and select the best model out of them. This topic is broadly called the model selection [78], which will be discussed in Section 8.

3. Applications of Bayesian Nonlinear Mixed Effects Model in Real-Life Problems

3.1. The Setting

To exemplify circumstances for which the nonlinear mixed effects model is a suitable modeling framework, we review challenges from several diverse applications. Table 2 summarize four real-life problems that will be illustrated in the next subsections.

Table 2. Summary of examples.

Research Field	Problem	Objective	References
Pharmaceutical industry	Pharmacokinetics analysis	Estimation of typical values of pharmacokinetics parameters	[79–82]
Oil and gas industry	Decline curve analysis	Prediction of estimated ultimate recovery	[31,83–85]
Financial industry	Yield curve modeling	Estimation of the interest rate parameters over time	[86–89]
Epidemiology	Epidemic spread prediction	Prediction of final epidemic size and finding risk factors	[30,90,91]

3.2. Example 1: Pharmacokinetics Analysis

Studies of the pharmacokinetics of drugs help us learn about the variability in drug disposition in a population [92]. Figure 3 shows theophylline concentration in the plasma as a function of time after oral administration of the same amount of anti-asthmatic theophylline for 12 subjects (the data considered here are courtesy of Dr. Robert A. Upton of the University of California, San Francisco.) As seen in the panel, concentration trajectories have a similar functional shape for all individuals. However, C_{max} and t_{max} (peak concentration and time when it is achieved), absorption, and elimination phases are substantially different across subjects. Clinical pharmacologists believe that these differences are attributable to between-subject variation in the underlying pharmacokinetic processes, explained by Absorption, Distribution, Metabolism, and Excretion (ADME), understanding of which is crucial in a new drug development in the pharmaceutical industry.

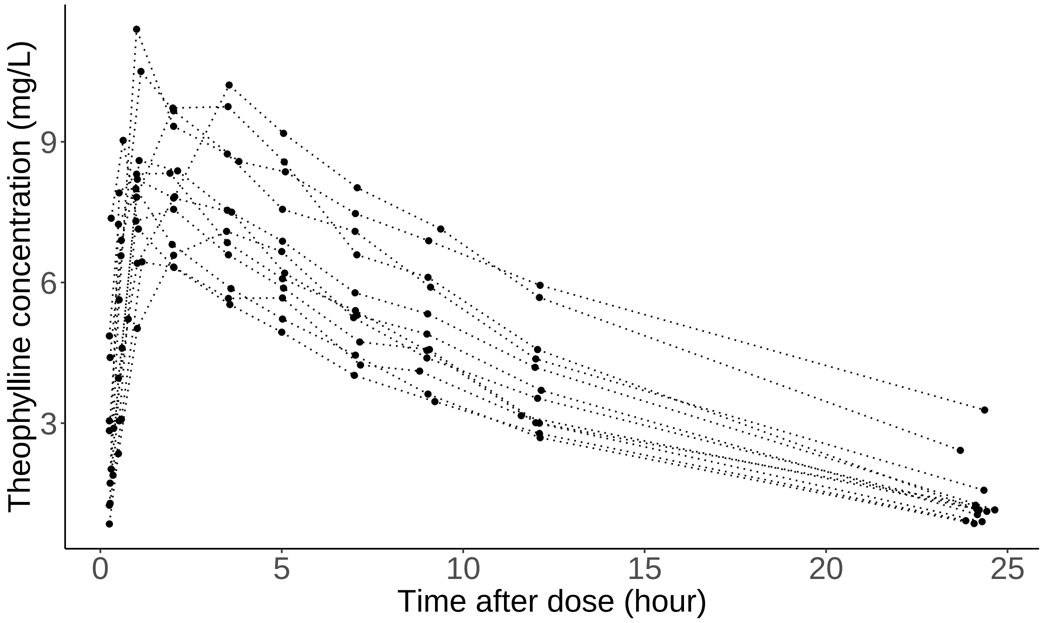

Figure 3. Theophylline concentrations for 12 subjects following an oral dose.

In pharmacokinetics analysis, often abbreviated by 'PK analysis', it is routine to use compartmental modeling to describe the amount of drug in the body by dividing the whole body into one or more compartments [93]. For theophylline, a one-compartment model is normally used, which assumes that the entire body acts like a single, uniform compartment; see page 30 from [94] for a detailed explanation about the model:

$$C(t) = \frac{DFk_a}{V(k_a - Cl/V)}\left\{\exp\left(-\frac{Cl}{V}t\right) - \exp(-k_a t)\right\}, \tag{1}$$

where $C(t)$ is drug concentration at time t for a single subject following oral dose D at $t = 0$. Here, F is the bioavailability which expresses the proportion of a drug that gains access to the systemic circulation. k_a is the absorption rate constant describing how quickly the drug is absorbed from the gut into the systemic circulation. V is the volume of the central compartment. Cl is the clearance rate representing the volume of plasma from which the drug is eliminated per unit time. Eventually, the pharmacokinetic processes for a given subject is summarized by the 4-dimensional vector with 'PK parameters' (F, k_a, V, Cl). Obviously, it is the modeler's discretion to proceed with a more complex PK model such as a three compartment models with nonlinear clearance to fit the data, but in this case, over-parameterization should be carefully examined [95].

Typically, the dataset collected in a drug development program includes demographic and clinical covariates obtained from each subject, for, e.g., body weight, height, age, sex, creatinine clearance, albumin, etc.; and furthermore, one can also involve genetic information in an individual's response to drugs. Most covariates are measured at baseline, before assigning the drug, while some covariates can be measured at every sampling time. One of the crucial goals of PK analysis is to illustrate the effect of such covariates on the PK parameters [96]. The causal relationship inferred by the covariate analysis can be used to support physicians in making the necessary judgments about the medicines that they prescribe, tailored to individual patients [97].

In a PK report for a new drug application to government authorities like U.S. Food and Drug Administration (FDA) or European Medicines Agency (EMA), the PK parameters are summarized by mean or median, and very importantly, estimates of parameter precision. Estimates of parameter precision can provide valuable information regarding the adequacy of the data to support those parameters [98]. Parameter uncertainty can be estimated through several methods, including bootstrap procedures [99], log-likelihood profiling [100], or using the asymptotic standard errors of parameter estimates, and recently, Bayesian approaches draw a lot of attention from the pharmaceutical industry [101]. Particularly, Bayesian approach for the population PK analysis can be very useful when there is prior knowledge about PK parameters learned from preclinical studies, published works, etc., and one wants to incorporate them into the prior specification for PK parameters [28].

3.3. Example 2: Decline Curve Analysis

The US shale boom—a product of technological advances in horizontal drilling and hydraulic fracturing that unlocked new stores of energy—has greatly benefited the growth in the US economy. Horizontal drilling is a directional drilling technology where a well is drilled parallel to the reservoir bedding plane [102]. Well productivity of a horizontal well is known to often be 3 to 5 times greater than that of a vertical well [103,104], but it also costs 1.5 to 2.5 times more than a vertical well [105]. Therefore, the eventual success of the drilling project of unconventional shale wells relies on a large degree of well construction costs [106]. Because of very low permeability, and a flow mechanism very different from that of conventional reservoirs, estimates for the shale well construction cost often contain high levels of uncertainty. For this reason, one of the crucial tasks of petroleum engineers is to quantify the uncertainty associated with the process of oil or gas production to reduce the extra initial risk for the projects.

Figure 4 shows monthly production rate trajectories of 360 shale oil wells completed in the Eagle Ford Shale of South Texas, studied by [31]. The declining pattern manifested in the trajectories is commonly observed in almost all oil production rate time series data following well completion. Here, the completion is terminology in petroleum engineering, meaning the process of transforming a well ready for the initial production [107]. Decline curve analysis (DCA), introduced by [83] around 100 years ago, is one of the most popularly utilized methods for petroleum engineers. Its purpose is to (i) theorize a curve describing the declining pattern, (ii) analyze the declining production rates, (iii) characterize the well-productivity, and (iv) forecast the future performance of oil and gas wells. Particularly, estimation and uncertainty quantification of estimated ultimate recovery (EUR) (here, EUR is a special jargon defined as an approximated quantity of oil from a well that is potentially recoverable by the end of its producing life [108]) is the utmost important task and a starting point in the decision-making process for future drilling projects. In addition, the oil and gas companies comply with financial regulations for EUR outlined by the U.S. Securities and Exchange Commission: see https://www.sec.gov/rules/final/2008/33-8995.pdf (accessed on 20 February 2022), for the regulations.

Most curves used in DCA are derived from solving certain differential equations that describe a hidden dynamic from production rate trajectory [109–114]. See [84,85,115,116] for an overview of such curves. Ref. [31] studied Arps' hyperbolic, stretched exponentiated decline, Duong, and Weibull curves to fit the trajectories shown in the Figure 4. Particularly, the Duong model was developed for unconventional reservoirs with very low permeability:

$$P(t) = q_1 t^{-m} \exp\left\{ \frac{a(t^{1-m} - 1)}{1 - m} \right\},$$

where $P(t)$ is the production rate at time t for a single well following completion. q_1 is the initial rate coefficient, and m and a are additional model parameters. We note that the parameters, q_1, m and a, have their own meanings in terms of well-productivity: see [114] for the interpretation. That being said, the well-productivity for a given well is summarized by the 3-dimensional parameter vector, (q_1, m, a). In a modeling perspective, the variation of the well-productivity across different wells is attributable to the different values for (q_1, m, a). To explain this variability, one can regress the values (q_1, m, a) on the well-design parameters such as true vertical depth, measure depth, etc. The causal relationship inferred by the covariate analysis will be used in a future drilling project. Geological information of wells can also be incorporated to make a spatial prediction for the EUR at a new location, as researched by [31].

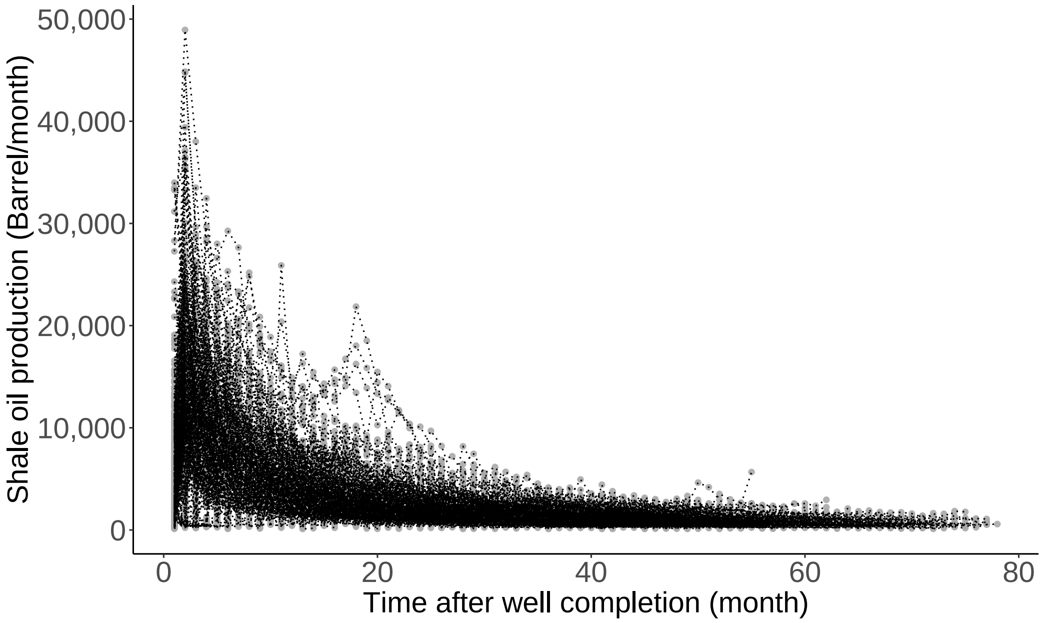

Figure 4. Production rates for 360 shale oil wells after completion.

3.4. Example 3: Yield Curve Modeling

Macroeconomists, financial economists, and market participants all attempt to build good models of the 'yield curve' [117]. The yield curve on a given day is a curve showing the interest rates across different maturity spans (one month, one year, five years, etc.) for a similar debt contract at a particular date. It determines the interest rate pattern (i.e., cost of borrowing), which can be used to calculate a bond's price [118]. Figure 5 shows daily treasury par yield curve rates spanning from 3 to 13 January 2022, with maturities up to 30 years. The data source is from the U.S. Department of the treasury (https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=yield, accessed on 20 February 2022). As seen from the panel, the shape of the yield curve displays a slightly delayed humped shape. Economists believe that such a shape of the yield curve has an important implication on the economic growth [119].

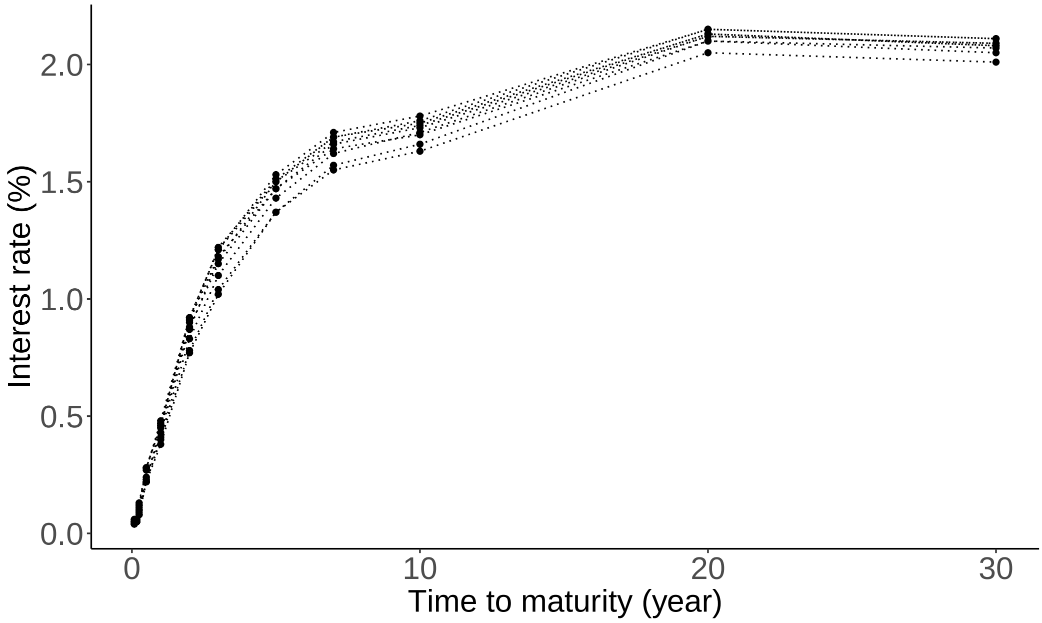

Figure 5. Daily Treasury par yield curve rates from 3 to 13 January 2022.

The Nelson–Siegel model [86] is a very popular model in the literature to fit the term structure:

$$Y(\tau) = \beta_0 + \beta_1 \left\{ \frac{1 - \exp(-\lambda\tau)}{\lambda\tau} \right\} + \beta_2 \left\{ \frac{1 - \exp(-\lambda\tau)}{\lambda\tau} - \exp(-\lambda\tau) \right\},$$

where $Y(\tau)$ denotes the (zero-coupon) yield evaluated at τ, and τ denotes the time to maturity. The model parameters have a specific financial meaning: β_0, β_1, and β_2 are related long-term, short-term, and mid-term effects on the interest rate, respectively, and λ is referred to as a decay factor [87]. Each of the yield curves is summarized by the 4-dimensional parameter $(\beta_0, \beta_1, \beta_2, \lambda)$, and it is known that the model can capture a wide range of possible shapes of the yield curve [86,87,120,121]. Therefore, the Nelson–Siegel model is extensively used by central banks and monetary policymakers [122]. For example, The Federal Reserve updates estimates of $(\beta_0, \beta_1, \beta_2, \lambda)$ once per week: visit the website (https://www.federalreserve.gov/data/yield-curve-tables/feds200628_1.html, accessed on 20 February 2022). In recent years, there has been a great deal of interest in the uncertainty quantification of the Nelson–Siegel parameters over time, and their relationship with macroeconomic variables such as inflation and real activity, etc, in financial applications: refer to [121,123–125] for some of those works.

3.5. Example 4: Early Stage of Epidemic

Novel coronavirus disease 2019 (COVID-19) is a big threat to global health. The rapid spread of the virus has created a pandemic, and countries all over the world are struggling with a surge in COVID-19 infected cases. Figure 6 displays the daily infection trajectories describing the cumulative numbers of infected cases for 40 countries, spanning from 22 January to 14 May 2020, studied by [30]. The data source is from COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University (https://coronavirus.jhu.edu/map.html, accessed on 20 February 2022). Refer to Table S.1 in [30] for the list of 40 countries. The time frame of the authors' research was the early stage of the pandemic when there was no drug or other therapeutics approved by the US FDA.

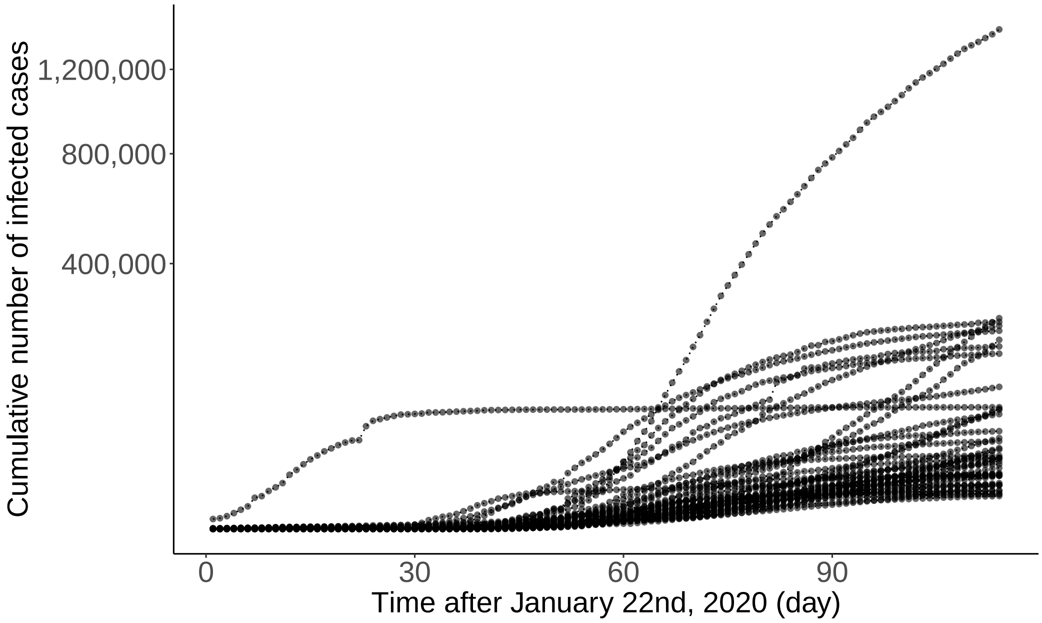

Figure 6. Daily trajectories for cumulative numbers of COVID-19 infections for 40 countries from 22 January to 14 May 2020.

In general, during an early phase of a pandemic, information regarding the disease is very limited and scattered, even if it exists. In spite of that, it is crucial to predict future cases of infection or death. In such a situation, one consideration is to use data integration (also called 'borrowing information'), combining data from diverse sources and eliciting useful information with a unified view of them. Additionally, it is very important to find risk factors relevant to the disease. Reliable and early risk assessment of a developing infectious disease outbreak allow policymakers to make swift and well-informed decisions that would be needed to ensure epidemic control. Quantifying uncertainty about the final epidemic size is also very important.

Richards growth curve [126], so-called the generalized logistic curve [127], is a popularly used growth curve for population studies in situations where growth is not symmetrical about the point of inflection [128,129]. There are variant reparamerized forms of the Richards curve in the literature [130–133], and one of the frequently used form is

$$I(t) = \frac{a}{[1 + \xi \exp\{-b(t-c)\}]^{1/\xi}},$$

where $I(t)$ is the cumulative number of infected cases at time t. Here, epidemiological meanings of the parameters, a, b, and c, are the final epidemic size, infection rate, and lag phase of the trajectory, respectively. The parameter ξ is the shape parameter, and there seems no clear epidemiological meaning [134]. Each infection trajectory in Figure 6 can be characterized by the 4-dimensional parameters (a, b, c, ξ) if the Richards curve is used. Due to its flexibility originating from the shape parameter ξ, Richards curve has been widely used in epidemiology for real-time prediction of outbreak of diseases, possibly at an early phase of the pandemic when there is no second wave. Examples include SARS [135,136], dengue fever [137,138], pandemic influenza H1N1 [139], and COVID-19 outbreak [30,140].

3.6. Statistical Problem

In the previous subsections, we presented a range of examples in which nonlinear mixed effects models can be exploited. They have their own challenges to solve the problems that are representative of issues many researchers have to deal with in other areas: for example, (1) how to describe a possible nonlinear clearance with a limited number of patients; (2) how to handle an enormously large number of shale oil wells and make a spatial prediction of EUR at a new location; (3) how to describe the dynamic of the financial parameters over time; and (4) how to integrate data from different sources to produce more accurate forecast on the epidemic size.

An emerging issue accompanied by these problems, requested from researchers, government agencies, domain experts, etc., is how to quantify the uncertainty associated with parameter estimation and prediction. Although the traditional nonlinear mixed effects models, based on the maximum likelihood method, can provide confidence intervals and statistical tests, calculations of those generally involve approximations that are most accurate for large sample sizes, as discussed in Section 2.1. On the other hand, in the Bayesian approach—in which the prior automatically imposes the parameter constraints—inferences about parameter values based on the posterior distribution usually require integration rather than maximization, and no further approximation is involved. For that reason, the Bayesian approach is often suggested as a viable alternative to the frequentist approach to solving the problems.

We now formulate these problems as a statistical problem. First, we summarize common features of the dataset for the analysis.

(1) There exist repeated measures of a continuous response over time for each subject;
(2) There exists a variation of individual observations over time;
(3) There exists a variation from subject-to-subject in trajectories;
(4) There exist covariates measured at baseline for each subject.

The subject of sampling units considered in the statistical analysis is quite comprehensive. We have seen that it can be a patient, a shale oil well, a particular date, and a country. As the unique identifier, we assign the index i to each individual. By denoting N as the number of individuals (i.e., the sample size), the index i will take an integer from 1 to N. The sample size N available for the data analysis substantially varies across different industrial problems as well as subfields within the same industry. For example, the number of shale oil wells on Eagle Ford Shale Play can be as large as 6000 [31]. As for the pharmaceutical industry, in phase I cancer clinical trials, the number of cancer patients N may be strictly confined to 25 [141], but for phase III trials for non-oncology drug studies, N can be as large as 2000 [142].

Here the term 'time' is meant in the broadest sense. It can be a calendar time, a nominal time, a time after some event (e.g., the time after dose from Figure 3 and the time after well completion from Figure 4), or a time to some event (e.g., the time to maturity from Figure 5). Essentially, time can be defined as a physical quantity that can be indexed with consecutive integers to produce a temporal record. Another important characteristic of the time is that each subject may have different time points where observations are measured. In this article, we use t_{ij} to represent the time point, where the integer $j = 1, 2, \cdots, M_i$ indexes the time point from the earliest to the last observations. Thus, M_i represents the number of repeated observations for the i-th individual. When M_i is relatively small (or large), we say the repeated measures are sparsely observed (or densely observed). For example, the theophylline and yield curve data shown in Figures 3 and 5 are sparse data, while the oil production and COVID-19 data shown in Figures 4 and 6 are dense data.

As for the repeated measures, y_{ij} denotes the continuous response of the i-th subject at the time point t_{ij}. We assume that y_{ij} has been already pre-processed so that it is ready to be used for statistical modeling. For most applications, it may be necessary first to transform the data into some new representation before training the model. For example, as seen from Figure 4, oil productions vary substantially across different wells. For that reason, the authors [31] take a logarithm on the productions to derive the response y_{ij}, followed by appropriate statistical modeling on the log-scale. To some extent, data pre-processing may enhance the performance of the model.

Suppose that researchers collected P number of covariates at the baseline from each subject i ($i = 1, \cdots, N$). Here, the baseline refers to the time point t_{i1} (or possibly right before the time point t_{i1}), where at the first response y_{i1} has not been observed yet. Let x_{ib} denote the b-th covariate of the i-th subject ($b = 1, \cdots, P$). In general, there are two types of covariates: time-invariant and time-varying covariates. This article mainly concerns the former type. As similar to N, the number of covariates P substantially varies across industries and specific problems. For instance, in pharmacogenetics analysis, the number of protein-coding genes P would be around 20,000 [143]. In the oil and gas industry, if we consider most of the covariates obtained from the well completion procedure, P could be at least 100 [31].

In conclusion, the dataset for the statistical analysis can be represented by the collection of the N triplets $\{(\mathbf{y}_i, \mathbf{t}_i, \mathbf{x}_i)\}_{i=1}^{N}$. Here, for each subject i ($i = 1, \cdots, N$), we formulated two M_i-dimensional vectors $\mathbf{y}_i = (y_{i1}, \cdots, y_{ij}, \cdots, y_{iM_i})^\top$ and $\mathbf{t}_i = (t_{i1}, \cdots, t_{ij}, \cdots, t_{iM_i})^\top$, and a P-dimensional vector $\mathbf{x}_i = (x_{i1}, \cdots, x_{ib}, \cdots, x_{iP})^\top$.

The data structures described up to this point are commonly encountered in longitudinal data studies [7]. Essentially, the feature of dataset motivating the use of nonlinear mixed effects models is that, for each subject i, the response vector $\mathbf{y}_i$ displays some nonlinear tendency over time $\mathbf{t}_i$, as seen in Figures 3–6. To explain this nonlinearity, a researcher needs to theorize some nonlinear function, denoted as f, such as one compartment, Duong, Nelson-Siegel, and Richards models, depending on the contexts. The construction of such functions relies on human modelers' abstraction of data into a suitable dynamical system, which is often represented by a differential equation. Such a differential equation has a finite number of parameters that control the dynamic of the solution of the system, the understanding of which is vital for causal inference for the nature of the system by associating with covariates $\mathbf{x}_i$.

Figure 7 displays a pictorial description about how a PK modeler would see the theophylline concentration trajectory from the modeling perspective, where she theorized that one compartment model (1) would be suitable to describe the trajectories $\mathbf{y}_i$ over time $\mathbf{t}_i$ for each subject i ($i = 1, \cdots, N$). Then the 10-dimensional vector $\mathbf{y}_i$ is summarized by a 4-dimensional PK parameter vector (F_i, k_{ai}, V_i, Cl_i); the dimension reduction is intrinsically embedded in this process. As each of the parameters F_i, k_{ai}, V_i, and Cl_i has an important clinical meaning, it is very natural to ask how they are related with P covariates $\mathbf{x}_i$ to induce a causal relationship. For the purpose of modeling, it may be necessary to transform the original PK parameters $(F_i, k_{ai}, V_i, Cl_i) \in [0,1] \times (0,\infty)^3$ to model parameters $(\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) \in \mathbb{R}^K$ ($K = 4$) so that elements θ_{li} ($l = 1, \cdots, K$) are supported on the real number by taking transformations $\theta_{1i} = \log\{F_i/(1-F_i)\}$, $\theta_{2i} = \log k_{ai}$, $\theta_{3i} = \log V_i$, and $\theta_{4i} = \log Cl_i$. As these transformations were taken only for the modeling purpose, interpretations on the PK parameter for the PK report should be carried out after transforming back to the original scale.

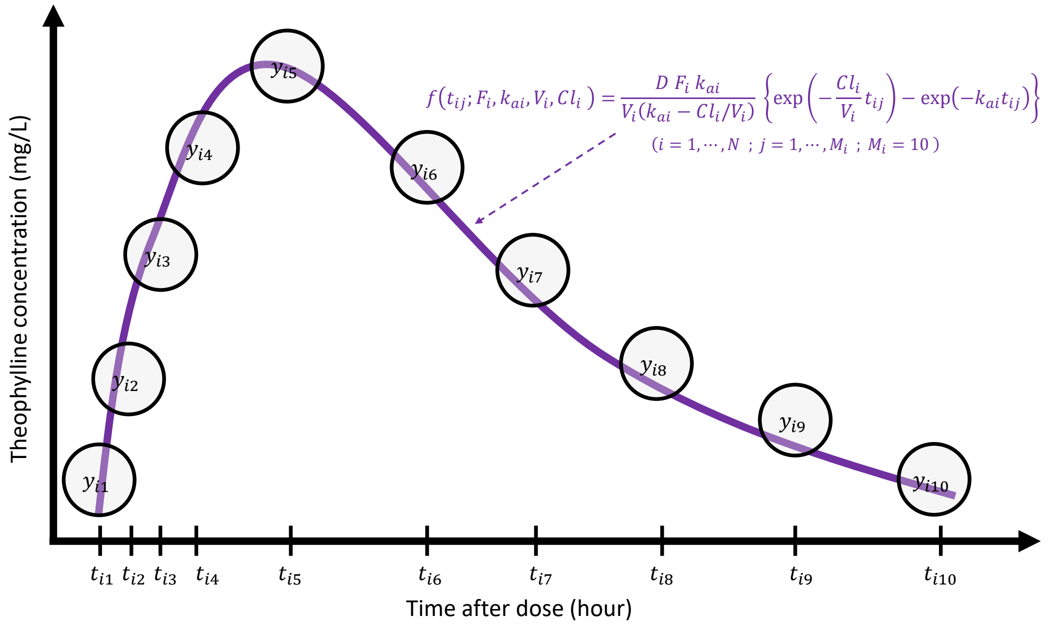
$f(t_{ij}; F_i, k_{ai}, V_i, Cl_i) = \frac{D\ F_i\ k_{ai}}{V_i(k_{ai} - Cl_i/V_i)} \left\{\exp\left(-\frac{Cl_i}{V_i} t_{ij}\right) - \exp(-k_{ai} t_{ij})\right\}$
$(i = 1, \cdots, N\ ;\ j = 1, \cdots, M_i\ ;\ M_i = 10)$
y_{i1}
y_{i2}
y_{i3}
y_{i4}
y_{i5}
y_{i6}
y_{i7}
y_{i8}
y_{i9}
y_{i10}
Theophylline concentration (mg/L)
t_{i1} t_{i2} t_{i3} t_{i4} t_{i5} t_{i6} t_{i7} t_{i8} t_{i9} t_{i10}
Time after dose (hour)

Figure 7. Pictorial illustration of PK modeling for the theophylline data.

4. The Model

4.1. Basic Model

Assume that we have dataset for a statistical analysis $\{(\mathbf{y}_i, \mathbf{t}_i, \mathbf{x}_i)\}_{i=1}^N$ from N subjects, as explained in Section 3.6. We consider a basic version of the model here. Extensions are discussed in Sections 7.3 and 9. The usual Bayesian nonlinear hierarchical model may then be written as a three-stage hierarchical model as follows:

- ***Stage 1: Individual-Level Model***

$$y_{ij} = f(t_{ij}; \boldsymbol{\theta}^i) + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad (i = 1, \cdots, N; j = 1, \cdots, M_i). \tag{2}$$

In (2), the conditional mean $\mathbb{E}[y_{ij}|\boldsymbol{\theta}^i, \sigma^2] = f(t_{ij}; \boldsymbol{\theta}^i)$ is a known function governing within-individual temporal behavior dictated by a K-dimensional parameter $\boldsymbol{\theta}^i = (\theta_{1i}, \theta_{2i}, \cdots, \theta_{li}, \cdots, \theta_{Ki})^\top \in \mathbb{R}^K$ specific to the subject i. We assume that the residuals, ϵ_{ij}, are normally distributed with mean zero and with an unknown variance, σ^2.

- ***Stage 2: Population Model***

$$\theta_{li} = \alpha_l + \mathbf{x}_i^\top \boldsymbol{\beta}_l + \eta_{li}, \quad \eta_{li} \sim \mathcal{N}(0, \omega_l^2), \quad (i = 1, \cdots, N; l = 1, \cdots, K). \tag{3}$$

In (3), the l-th model parameter θ_{li} is used as the response of an ordinary linear regression with predictor $\mathbf{x}_i$, with intercept $\alpha_l \in \mathbb{R}$ and coefficient vector $\boldsymbol{\beta}_l = (\beta_{l1}, \beta_{l2}, \cdots, \beta_{lP}) \in \mathbb{R}^P$. By letting $\boldsymbol{\eta}_i = (\eta_{1i}, \eta_{2i}, \cdots, \eta_{li}, \cdots, \eta_{Ki}) \in \mathbb{R}^K$, we assume that the $\boldsymbol{\eta}_i$ is distributed according a K-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ with covariance matrix $\boldsymbol{\Omega} = \text{diag}(\omega_1^2, \omega_2^2, \cdots, \omega_l^2, \cdots, \omega_K^2) \in \mathbb{R}^{K \times K}$. The diagonality in $\boldsymbol{\Omega}$ implies that each model parameter are uncorrelated across l.

- ***Stage 3: Prior***

$$\sigma^2 \sim \pi(\sigma^2), \quad \alpha_l \sim \pi(\alpha_l), \quad \boldsymbol{\beta}_l \sim \pi(\boldsymbol{\beta}_l), \quad \omega_l^2 \sim \pi(\omega_l^2), \quad (l = 1, \cdots, K). \tag{4}$$

Distributions in (4) are chosen to encapsulate any information or belief that have been formulated about the parameters. We suggest some popularly used prior options in Section 7.

Directed asymmetric graphical (DAG) model representation of the basic model (2)–(4) is depicted in Figure 8. Following the grammar of the graphical model (Chapter 8 of [144]), the circled variables indicate stochastic variables, while the observed ones are additionally colored in gray. Non-stochastic quantities are uncircled. The arrows indicate the conditional dependency between the variables.

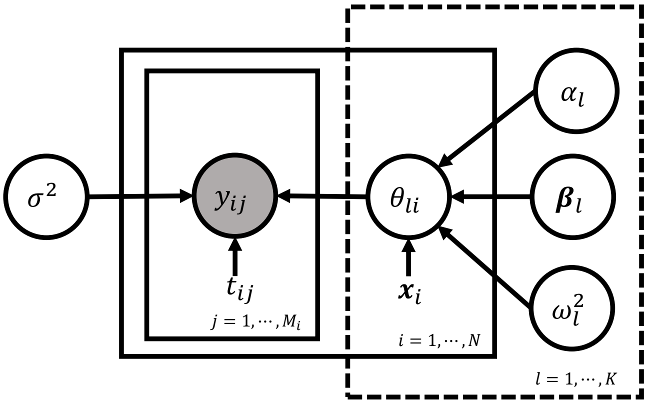

Figure 8. The basic model (2)–(4) as a graphical model.

4.2. Vectorized Form of the Basic Model

We will often wish to write the hierarchy (2)–(4) for the i-th individual's entire response vector and represent it with an equivalent vector-form. This turns out to be useful to develop relevant computational algorithms. We first introduce a $K \times N$ dimensional matrix frequently used throughout this article:

$$\boldsymbol{\Theta} = \begin{bmatrix} \theta_{11} & \cdots & \theta_{1i} & \cdots & \theta_{1N} \\ \vdots & & \vdots & & \vdots \\ \theta_{l1} & \cdots & \theta_{li} & \cdots & \theta_{lN} \\ \vdots & & \vdots & & \vdots \\ \theta_{K1} & \cdots & \theta_{Ki} & \cdots & \theta_{KN} \end{bmatrix} \in \mathbb{R}^{K \times N}. \tag{5}$$

The matrix (5) is referred to as *model matrix* because it comprises of scalar model parameters $\{\theta_{li}\}_{l=1,i=1}^{K,N}$ from all subjects. Indeed, most of the computational techniques either via frequentist or Bayesian setting in the literature have been developed to overcome an obstacle of a nonlinear association of the model matrix $\mathbf{\Theta}$ into the mean function f.

In (5), the subject index i is stacked column-wisely, while model parameter index l is stacked row-wisely, different from the conventional way adopted in most statistics. The column indexing for the subjects (i.e., stacking individual-based vector horizontally) shown in (5) is often adopted in modern computation theory of deep learning [145], and one of the main advantages of using this indexing is that it may give some pedagogical insights on the use of vectorization toward the entries $\{\theta_{li}\}_{l=1,i=1}^{K,N}$ to exploit parallel computations, stochastic updating, etc., in optimization or sampling techniques.

The model matrix $\mathbf{\Theta}$ (5) can be re-expressed as $\mathbf{\Theta} = [\boldsymbol{\theta}^1 \cdots \boldsymbol{\theta}^i \cdots \boldsymbol{\theta}^N] \in \mathbb{R}^{K\times N}$, obtained by stacking the individual model parameter vector in Stage 1 (2). Alternatively, we can represent the matrix with $\mathbf{\Theta} = [\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_l \cdots \boldsymbol{\theta}_K]^\top \in \mathbb{R}^{K\times N}$ by defining a N-dimensional vector corresponding the l-th model parameter across all subjects $\boldsymbol{\theta}_l = (\theta_{l1}, \theta_{l2}, \cdots, \theta_{lN})^\top \in \mathbb{R}^N$ $(l = 1, \cdots, K)$. Former and latter indexing method are referred to as i-indexing and l-indexing, respectively.

We are now in a position to re-write the hierarchy (2)–(4) using the vector notations:

- ***Stage 1: Individual-Level Model***

$$\mathbf{y}_i = \boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i) + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_{M_i}(\mathbf{0}, \sigma^2\mathbf{I}), \quad (i = 1, \cdots, N). \tag{6}$$

In (6), $\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i)$ is a M_i-dimensional vector whose elements are temporally stacked: $\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i) = (f(t_{i1}; \boldsymbol{\theta}^i), f(t_{i2}; \boldsymbol{\theta}^i), \cdots, f(t_{iM_i}; \boldsymbol{\theta}^i))^\top$ for the subject i. The vector $\boldsymbol{\epsilon}_i$ is distributed according to the M_i-dimensional Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$.

- ***Stage 2: Population Model (l-indexing)***

$$\boldsymbol{\theta}_l = \mathbf{1}\alpha_l + \mathbf{X}\boldsymbol{\beta}_l + \boldsymbol{\eta}_l, \quad \boldsymbol{\eta}_l \sim \mathcal{N}_N(0, \omega_l^2\mathbf{I}), \quad (l = 1, \cdots, K). \tag{7}$$

In (7), for each l, the N-dimensional model parameter vector $\boldsymbol{\theta}_l$ is used as the response vector of an ordinary linear regression: (i) N-by-P design matrix $\mathbf{X} = [\mathbf{x}_1\mathbf{x}_2 \cdots \mathbf{x}_N]^\top$; (ii) intercept α_l; (iii) coefficient vector $\boldsymbol{\beta}_l$, and (iv) isotropic Gaussian error vector $\boldsymbol{\eta}_l = (\eta_{l1}, \eta_{l2}, \cdots, \eta_{lN})^\top$ with variance ω_l^2. (Notation $\mathbf{1}$ in (7) represents an all-ones vector.).

- ***Stage 2′: Population Model (i-indexing)***

$$\boldsymbol{\theta}^i = \boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i + \boldsymbol{\eta}^i, \quad \boldsymbol{\eta}^i \sim \mathcal{N}_K(\mathbf{0}, \mathbf{\Omega}), \quad (i = 1, \cdots, N). \tag{8}$$

Equation (8) is derived by incorporating each of the N columns of the model matrix (5). Here, $\boldsymbol{\alpha}$ represents a K-dimensional vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_K)^\top$, and $\boldsymbol{B}$ represents a K-by-P matrix with rows $\boldsymbol{\beta}_l$ $(l = 1, \cdots, K)$. Here, the K-dimensional vector $\boldsymbol{B}\mathbf{x}_i$ in the right-hand side of (8) is the mathematically identical to $\mathbf{X}_i\boldsymbol{\beta}$, where $\mathbf{X}_i = \mathbf{I}_K \otimes \mathbf{x}_i^\top \in \mathbb{R}^{K\times KP}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_K) \in \mathbb{R}^{KP}$ ($\mathbf{I}_K$ is the K-by-K identity matrix and $\otimes$ represents the Kronecker matrix product.). The error vector $\boldsymbol{\eta}^i = (\eta_{1i}, \eta_{2i}, \cdots, \eta_{Ki})^\top$ is distributed according a K-dimensional Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Omega} = \mathrm{diag}(\omega_1^2, \omega_2^2, \cdots, \omega_K^2)$.

- ***Stage 3: Prior***

$$\sigma^2 \sim \pi(\sigma^2), \quad \boldsymbol{\alpha} \sim \pi(\boldsymbol{\alpha}), \quad \boldsymbol{B} \sim \pi(\boldsymbol{B}), \quad \mathbf{\Omega} \sim \pi(\mathbf{\Omega}). \tag{9}$$

Each of the parameter blocks in $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \mathbf{\Omega})$ is assumed to be independent a *priori*.

To summarize, we derived two equivalent vectorized formulations representing the basic model (2)–(4) according to how the model matrix $\mathbf{\Theta}$ (5) is vectorized:

- **Vector-form (a)**: Stage 1—(6), Stage 2—(7) (l-indexing), and Stage 3—(9);
- **Vector-form (b)**: Stage 1—(6), Stage 2′—(8) (i-indexing), and Stage 3—(9).

Figure 9 displays the DAG representations of the two vector forms of the basic model. In vector-form (a), K latent nodes $\{\boldsymbol{\theta}_l\}_{l=1}^K$ are fully connected toward the N response vectors $\{\mathbf{y}_i\}_{i=1}^N$. On the other hand, in vector-form (b), N latent nodes $\{\boldsymbol{\theta}^i\}_{i=1}^N$ are bijectively connected to the N response vectors $\{\mathbf{y}_i\}_{i=1}^N$ for each subject i. These two ways of looking at the framework of the Bayesian nonlinear hierarchical models complement each other for a more proper understanding of the modeling framework and provide modelers with a statistical insight. For example, vector-form (a) is useful to understand some mathematics underlying the linear regression using P regressors, while vector-form (b) makes it easy to comprehend the role of the covariance matrix $\mathbf{\Omega}$.

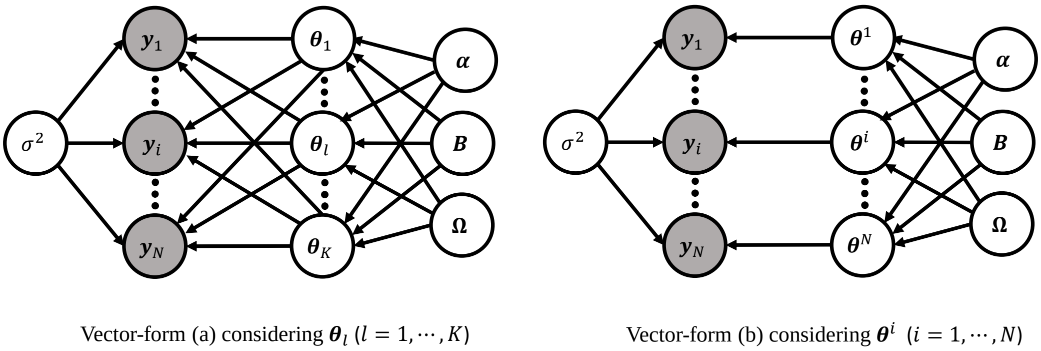

Figure 9. DAG representations of the basic model (2)–(4) in vector-form (a) (Stage 1—(6), Stage 2—(7) (l-indexing), and Stage 3—(9)) (**left**) and vector-form (b) (Stage 1—(6), Stage 2′—(8) (i-indexing), and Stage 3—(9)) (**right**). Two vector forms are equivalent, except for the way how the model matrix $\mathbf{\Theta}$ (5) is vectorized.

5. Likelihood

5.1. Outline

In this section, we investigate a likelihood function based on the basic model (2)–(4). As illustrated in Bayesian workflow in Section 2.2, the likelihood theory is fundamental of Bayesian inference (see Figure 2). Therefore, it is worth spending time to re-study the formulation of the likelihood function. Here, one caveat is that, due to the hierarchical nature of the nonlinear mixed effects model, a notion of the likelihood function depends on what part of the model specification is considered to be part of the likelihood, and what is not. Most papers directly consider the marginal likelihood that will be discussed in Section 5.4. In this paper, before marching there, we study other two formulations of the likelihood in Sections 5.2 and 5.3 to get some pedagogical insights. We will briefly discuss popularly used frequentist computing strategies in Section 5.4.

5.2. Likelihood Based on Stage 1

As in most of the statistical models, a natural starting point for inference is maximum likelihood estimation. We start with considering only Stage 1 from the basic model in Section 4.1 and ignore Stages 2 and 3 for now. Then the likelihood function for the i-th subject is

$$\mathcal{L}(\boldsymbol{\theta}^i, \sigma^2 | \mathbf{y}_i) = p(\mathbf{y}_i | \boldsymbol{\theta}^i, \sigma^2) = \mathcal{N}_{M_i}(\mathbf{y}_i | \boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2 \mathbf{I}), \quad (i = 1, \cdots, N).$$

Therefore, the likelihood function based on the N subjects $\mathbf{y}_{1:N} = \{\mathbf{y}_i\}_{i=1}^{N}$ is

$$\mathcal{L}(\mathbf{\Theta}, \sigma^2 | \mathbf{y}_{1:N}) = \prod_{i=1}^{N} \mathcal{N}_{M_i}(\mathbf{y}_i | \boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2 \mathbf{I}). \tag{10}$$

Now, we maximize the likelihood (10) with respect to the model matrix $\mathbf{\Theta}$ (5) given σ^2 fixed:

$$\begin{aligned}\widehat{\mathbf{\Theta}} &= \text{argmax}_{\mathbf{\Theta} \in \mathbb{R}^{K \times N}} \log\ \mathcal{L}(\mathbf{\Theta}, \sigma^2 | \mathbf{y}_{1:N}) \\ &= \text{argmin}_{\mathbf{\Theta} \in \mathbb{R}^{K \times N}} \sum_{i=1}^{N} \| \mathbf{y}_i - \boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i) \|_2^2 \\ &= \text{argmin}_{\mathbf{\Theta} \in \mathbb{R}^{K \times N}} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \Big(y_{it} - f(t_{ij}; \boldsymbol{\theta}^i) \Big)^2, \end{aligned} \tag{11}$$

where $\| \cdot \|_2$ is the Euclidean norm.

The estimator $\widehat{\mathbf{\Theta}} = [\widehat{\boldsymbol{\theta}}^1 \cdots \widehat{\boldsymbol{\theta}}^i \cdots \widehat{\boldsymbol{\theta}}^N] \in \mathbb{R}^{K \times N}$ of the model matrix $\mathbf{\Theta}$ (5) can be obtained by various optimization techniques such as Newton-Raphson method or Gradient descent method [146]. Noting from the summation across i in (11), N estimators $\{\widehat{\boldsymbol{\theta}}^i\}_{i=1}^{N}$ are independent. We can obtain an estimator of the variance σ^2 by plugging $\widehat{\mathbf{\Theta}}$ into the likelihood (10), and then maximize with respect to the σ^2. To investigate a denoised temporal tendency for the trajectory $\mathbf{y}_i$, we can simply plug $\widehat{\boldsymbol{\theta}}^i$ into the function $f(t_{ij}; \boldsymbol{\theta}^i)$ $(j = 1, \cdots, M_i)$. To see a future pattern, we can extrapolate the function by extending the time index beyond the last time point t_{iM_i}. Eventually, the illustrated approach is based on traditional least squares estimation.

Unfortunately, there are three major drawbacks in this approach. First, it forfeits the opportunity to use 'information borrowing' [30] to improve a predictive accuracy due to the ignorance of Stage 2. What happens in Stage 2 (3) is to borrow strength across N individuals to produce a better estimator for $\mathbf{\Theta}$ (5) than an estimator simply based on individual data. A similar issue can be found in the Clemente problem from [11] where the James–Stein estimator [147] predicts better than an individual hitter-based estimator. Another example applied to epidemic data can be found in [30]. Second, it is not well-aligned with the generic motivation to use the mixed effects models, whose primary purpose is to understand "typical" values for the model parameters in f, representing whole subjects, which should be addressed by making an inference about the parameters $\boldsymbol{\alpha}$, $\boldsymbol{B}$, and $\mathbf{\Omega}$. Third, it only produces point estimates for the parameters, failing to describe the underlying uncertainty.

A remedy of the first two drawbacks is the consideration of Stages 1 (2) and 2 (3) hierarchically in a single model, leading to a frequentist version of nonlinear mixed effects models, which will be discussed in Sections 5.3 and 5.4. To describe relevant uncertainty within the frequentist framework one may have to resort to bootstrap methods [99] or use a large-sample theory. Uncertainty quantification in frequentist analysis often needs to be done step-wisely. Another solution resolving all the three drawbacks at once is to incorporate Stage 1 (2), 2 (3), and 3 (4) into a single model in a fully Bayesian way, resulting in a Bayesian version of nonlinear mixed effects models, which is the main topic in this paper.

5.3. Likelihood Based on Stage 1 and 2 from Vector-Form (a)

A likehood function based on vector-form (a) is derived. More specifically, we examine the frequentist setting where the assumptions of Stage 1—(6) and Stage 2—(7) are considered, while the parameters introduced in Stage 3—(9) are fixed (i.e., no prior assumptions).

The individual model on Stage 1 (6) yields a conditional density $p(\mathbf{y}_i | \boldsymbol{\theta}^i, \sigma^2) = \mathcal{N}_{M_i}(\mathbf{y}_i | \boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2 I)$ for each subject $i = 1, \cdots, N$. Under the population assumption on Stage 2 (7), we have the density $p(\boldsymbol{\theta}_l|\alpha_l, \boldsymbol{\beta}_l, \omega_l^2) = \mathcal{N}_N(\boldsymbol{\theta}_l|\mathbf{1}\alpha_l + \mathbf{X}\boldsymbol{\beta}_l, \omega_l^2\mathbf{I})$ for each model parameter index $l = 1, \cdots, K$. The joint density of $(\mathbf{y}_{1:N}, \boldsymbol{\theta}_{1:K})$ given parameters $\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}$ and $\boldsymbol{\Omega}$ is a product-form distribution:

$$p(\mathbf{y}_{1:N}, \boldsymbol{\theta}_{1:K}|\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}) = \left\{ \prod_{i=1}^{N} \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2\mathbf{I}) \right\} \cdot \left\{ \prod_{l=1}^{K} \mathcal{N}_N(\boldsymbol{\theta}_l|\mathbf{1}\alpha_l + \mathbf{X}\boldsymbol{\beta}_l, \omega_l^2\mathbf{I}) \right\},$$

where $\mathbf{y}_{1:N} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$ and $\boldsymbol{\theta}_{1:K} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_K\}$. Now, the next step is to integrate out the latent model parameters $\boldsymbol{\theta}_{1:K}$ (or equivalently, the model matrix $\boldsymbol{\Theta}$ (5)) from the density above to get a likelihood for the $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$:

$$\mathcal{L}(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}|\mathbf{y}_{1:N}) = \int p(\mathbf{y}_{1:N}, \boldsymbol{\theta}_{1:K}|\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}) d\boldsymbol{\theta}_{1:K}. \tag{12}$$

In most cases, the integral (12) is not tractable due to the non-linearity of the function $\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i)$ with respect to the $\boldsymbol{\theta}^i$. Although it may be possible to use numerical techniques for the evaluation of the integral (12), this might require enormous computational effort, which is not really appreciated in the literature due to the high-dimensionality of the integral involving the KN dimensional model parameters $\boldsymbol{\theta}_{1:K}$.

5.4. Likelihood Based on Stage 1 and 2′ from Vector-Form (b)

A likehood function based on vector-form (b) adopting i-indexing is derived here. Similar to Section 5.3, we preserve the assumption of Stage 1—(6) and Stage 2′—(8), but work with fixing the parameters in Stage 3—(9). In these specifications, for each index $i = 1, \cdots, N$, the individual model on Stage 1 (6) and population model on Stage 2 (8) lead to densities $p(\mathbf{y}_i|\boldsymbol{\theta}^i, \sigma^2) = \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2 I)$ and $p(\boldsymbol{\theta}^i|\boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}) = \mathcal{N}_K(\boldsymbol{\theta}^i|\boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i, \boldsymbol{\Omega})$, respectively. Thus, the joint density of $(\mathbf{y}_i, \boldsymbol{\theta}^i)$ given parameters $\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}$ and $\boldsymbol{\Omega}$ is

$$\begin{aligned} p(\mathbf{y}_i, \boldsymbol{\theta}^i|\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}) &= p(\mathbf{y}_i|\boldsymbol{\theta}^i, \sigma^2) \cdot p(\boldsymbol{\theta}^i|\boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}) \\ &= \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2\mathbf{I}) \cdot \mathcal{N}_K(\boldsymbol{\theta}^i|\boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i, \boldsymbol{\Omega}). \end{aligned}$$

Given the parameters $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$, the ordered pairs in the collection $\{(\mathbf{y}_i, \boldsymbol{\theta}^i)\}_{i=1}^N$ are conditionally independent across individuals. Therefore, a likelihood for the $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ is based on the marginal density of $\mathbf{y}_{1:N} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$:

$$\begin{aligned} \mathcal{L}(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}|\mathbf{y}_{1:N}) &= \prod_{i=1}^{N} \int \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2\mathbf{I}) \cdot \mathcal{N}_K(\boldsymbol{\theta}^i|\boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i, \boldsymbol{\Omega}) d\boldsymbol{\theta}^i & (13) \\ &= \prod_{i=1}^{N} \int \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i + \boldsymbol{\eta}^i), \sigma^2\mathbf{I}) \cdot \mathcal{N}_K(\boldsymbol{\eta}^i|\mathbf{0}, \boldsymbol{\Omega}) d\boldsymbol{\eta}^i, & (14) \end{aligned}$$

where the last equality is derived by using the change of variable (8). The last expression (14) is a standard mathematical formulation that many frequentist computing strategies are constructed with: see Equation (3.2) from [14]. Essentially, the part that makes the MLE computation complicated is the mean vector $\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i + \boldsymbol{\eta}^i) \in \mathbb{R}^{M_i}$.

As the model parameter $\boldsymbol{\theta}^i$ in (13) (or similarly, $\boldsymbol{\eta}^i$ in (14) which is often called random effect in the frequentist framework) participates to the function f in a non-linear fashion, the integral generally cannot be obtained in a closed-form. Benefiting from a conditional independence [148], dimensionality of the N integrals (13) is much lower than that of the integral (12) based on vector-form (a). Analytically, the likelihood functions of the basic model (2)–(4) based on vector-form (a) (12) and vector-form (b) (13) may be equivalent. That being said, minimization of the two functions with respect to the parameters $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ yields the same solution, $(\widehat{\sigma^2}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Omega}})$, so-called maximum likelihood estimators (MLE).

We shall briefly discuss MLE computations. One approach would be to perform a multivariate numerical integration (e.g., Gauss-Hermite quadrature [149]) to each of the

N integrals (13), and then obtain the MLE by maximizing the product of the N numerical integrals with respect to the parameters $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ [150]. This approach turns out to be computationally expensive and may have poor converge properties due to the following two reasons [151]. First, the numerical integration necessitates increasingly expensive iterative procedures within a MLE algorithm as the correlation of the model parameters (or equivalently, random effects) increases. Second, convergence property may be highly deteriorated when the number of model parameters K is large (i.e., high-dimensional integral) and the number of sampling times M_i is small (i.e., sparse data) due to the 'curse of dimensionality' [152].

A class of common approaches for the MLE computations is based on analytical approximation to each of the N integrals (14) [13,36,153–155], and some of them have been successfully adopted to industrial software like NONMEM [47,101] and SAS [46]. Here, we illustrate a key idea of the first-order method attributed to [36]. Let us define a mapping $\boldsymbol{g}_i(\boldsymbol{\eta}^i) = \boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i + \boldsymbol{\eta}^i) : A \subset \mathbb{R}^K \to \mathbb{R}^{M_i}$ for each subject i, where A is an open set with $\mathbf{0} \in A$. Suppose that $\boldsymbol{g}_i(\boldsymbol{\eta}^i)$ is smooth on the set A: then, by Taylor's theorem (page 375 of [156]), we have the best linear approximation of the mapping $\boldsymbol{g}_i(\boldsymbol{\eta}^i)$ at the origin $\mathbf{0}$ given by $\boldsymbol{g}_i(\boldsymbol{\eta}^i) \approx \boldsymbol{g}_i(\mathbf{0}) + \mathbf{D}\boldsymbol{g}_i(\mathbf{0})\boldsymbol{\eta}^i$, where $\mathbf{D}\boldsymbol{g}_i(\mathbf{0}) \in \mathbb{R}^{M_i \times K}$ is the Jacobian matrix of $\boldsymbol{g}_i(\boldsymbol{\eta}^i)$ at $\mathbf{0}$. Now, we shall replace the function $\boldsymbol{g}_i(\boldsymbol{\eta}^i) = \boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i + \boldsymbol{\eta}^i)$ in integral (14) with the resulting approximation $\boldsymbol{g}_i(\mathbf{0}) + \mathbf{D}\boldsymbol{g}_i(\mathbf{0})\boldsymbol{\eta}^i$ for each i ($i = 1, \cdots, N$), leading to a closed-form expression

$$\begin{aligned} \widetilde{\mathcal{L}}(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega} | \mathbf{y}_{1:N}) &= \prod_{i=1}^{N} \int \mathcal{N}_{M_i}(\mathbf{y}_i | \boldsymbol{g}_i(\mathbf{0}) + \mathbf{D}\boldsymbol{g}_i(\mathbf{0})\boldsymbol{\eta}^i, \sigma^2 \mathbf{I}) \cdot \mathcal{N}_K(\boldsymbol{\eta}^i | \mathbf{0}, \boldsymbol{\Omega}) d\boldsymbol{\eta}^i \\ &= \prod_{i=1}^{N} \mathcal{N}_{M_i}(\mathbf{y}_i | \boldsymbol{g}_i(\mathbf{0}), \mathbf{D}\boldsymbol{g}_i(\mathbf{0}) \boldsymbol{\Omega} \mathbf{D}\boldsymbol{g}_i(\mathbf{0})^\top + \sigma^2 \mathbf{I}). \end{aligned} \tag{15}$$

To summarize, a linearization was used to convert the nonlinear mixed effects model to a linear mixed effects model, in some sense, equivalent to the Lindley–Smith form [157]. This enables us to integrate out the random vector $\boldsymbol{\eta}^i$ from the N integrals (15), deriving a marginal likelihood (15) to approximate the exact marginal likelihood (13). MLE $(\widehat{\sigma^2}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Omega}})$ can be obtained by jointly maximizing $\widetilde{\mathcal{L}}(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega} | \mathbf{y}_{1:N})$ (15) assuming the approximation is exact.

Another way to compute the MLE is through the use of expectation-maximization (EM) algorithm [158]. Borrowing terms from EM updating process [159], $\mathbf{y}_i$, $\boldsymbol{\theta}^i$, $(\mathbf{y}_i, \boldsymbol{\theta}^i)$, $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ and $p(\mathbf{y}_i, \boldsymbol{\theta}^i | \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ (i.e., the integrand in (13)) can be viewed as observable incomplete data, missing data, complete data, unknown parameters, and density of complete data, respectively, for the i-th subject. The goal is to maximize the exact marginal likelihood $\mathcal{L}(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega} | \mathbf{y}_{1:N})$ (13) by iterating E-step and M-step, leading to the MLE $(\widehat{\sigma^2}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Omega}})$. The E-step computes a conditional expected log-likelihood of $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ based on the hierarchy (2)–(3), followed by the M-step that maximizes the function with respect to $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$. The nonlinearity associated with the model matrix $\boldsymbol{\Theta}$ (5) makes the E-step intractable. As a remedy, variant versions of the EM algorithm are proposed; see [160–163] for a technical detail applied to a hierarchy similar to the basic model. Among them, the scheme of stochastic approximation EM algorithm proposed by [38], splitting the E-step into two steps, namely a simulation step and a stochastic approximation step, is widely used in many applications for its numerical stability, fast computation, and theoretical soundness [164,165], which has been successfully deployed as industrial software including MONOLIX [48] as well as open source software such as R package NLMIXR [18].

6. Bayesian Inference and Implementation

6.1. Bayesian Inference

We briefly overview two contrasting workflows of Bayesian and frequentist approaches for nonlinear mixed effects models before moving to a technical detail. Both settings allow the randomness in the model matrix $\boldsymbol{\Theta}$ (5), but then, they diverge when it comes to how parameters $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ are treated. Bayesians treat $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ as random, while frequentists regard it as fixed. To conceptualize a subtlety arising from this difference, let us recap frequentist computing strategies discussed in Section 5.4. There, the model matrix $\boldsymbol{\Theta}$ was eventually integrated out from the joint density of $(\mathbf{y}_{1:N}, \boldsymbol{\Theta})$, either approximately or exactly, to derive a marginal likelihood of $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ from which the MLE $(\widehat{\sigma^2}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Omega}})$ is computed via various optimization methods. After that, frequentists apply standard Bayesian formulas, such as posterior density, posterior mean, and so on, to estimate $\boldsymbol{\Theta}$ [7].

In contrast, to drive the Bayesian engine, one would need an appropriate prior $\pi(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$. After that, the entire collection of parameters $(\boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ will be updated through the Bayes' theorem post observing the data $\mathbf{y}_{1:N}$, leading to the posterior density $\pi(\boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}|\mathbf{y}_{1:N})$ [166] (See Figure 2). The essence of the Bayesian viewpoint is that there is no logical distinction between $\boldsymbol{\Theta}$ and $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$, which are associated with the random and fixed effects, respectively, from the frequentist perspective. In Bayesian framework, both $\boldsymbol{\Theta}$ and $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ are random quantities. It is important to point out that the likelihood principle is naturally incorporated in the Bayes' theorem [167]. Clearly, modern data complications such as enormous volume, large dimensionality, and multi-level structures may necessitate a sophistication on the prior specifications.

We are now in a position to describe the Bayesian analysis for the basic model (2)–(4), that assumed independence, a *priori*, for each parameter blocks σ^2, $\boldsymbol{\alpha}$, $\boldsymbol{B}$, and $\boldsymbol{\Omega}$, $\pi(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}) = \pi(\sigma^2) \cdot \pi(\boldsymbol{\alpha}) \cdot \pi(\boldsymbol{B}) \cdot \pi(\boldsymbol{\Omega})$. (Our logic below can be generalized to a more complex prior setting.) As discussed in Section 4, it is the discretion of the modeler of how she would treat the model matrix $\boldsymbol{\Theta}$ (5) with l-indexing $\boldsymbol{\theta}_{1:K} = \{\boldsymbol{\theta}_l\}_{l=1}^K$ or i-indexing $\boldsymbol{\theta}^{1:N} = \{\boldsymbol{\theta}^i\}_{i=1}^N$, leading to vector forms (a) and (b), respectively. For the sake of readability, we illustrate the Bayesian inference by using the vector-form (a), but we will sometimes use the vector-form (b) when this seems more understandable.

A central task in the application of the Bayesian nonlinear mixed effects models is to evaluate the posterior density, or indeed to compute expectation with respect to the density:

$$\begin{aligned}\pi(\boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}|\mathbf{y}_{1:N}) &= \frac{\pi(\mathbf{y}_{1:N}, \boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})}{p(\mathbf{y}_{1:N})} \\ &\propto \pi(\mathbf{y}_{1:N}, \boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}) \\ &= \underbrace{p(\mathbf{y}_{1:N}|\boldsymbol{\Theta}, \sigma^2)}_{\text{Stage 1}} \cdot \underbrace{\pi(\boldsymbol{\Theta}|\boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})}_{\text{Stage 2}} \cdot \underbrace{\pi(\sigma^2) \cdot \pi(\boldsymbol{\alpha}) \cdot \pi(\boldsymbol{B}) \cdot \pi(\boldsymbol{\Omega})}_{\text{Stage 3}},\end{aligned}$$

where the last equation can be detailed as follows

$$\left\{\prod_{i=1}^N \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2\mathbf{I})\right\} \cdot \left\{\prod_{l=1}^K \mathcal{N}_N(\boldsymbol{\theta}_l|\mathbf{1}\alpha_l + \mathbf{X}\boldsymbol{\beta}_l, \omega_l^2\mathbf{I})\pi(\alpha_l)\pi(\omega_l^2)\pi(\boldsymbol{\beta}_l)\right\} \cdot \pi(\sigma^2). \tag{16}$$

From a Bayesian perspective, all inferential problems regarding the parameter $(\boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ may be addressed in terms of the posterior distribution $\pi(\boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}|\mathbf{y}_{1:N})$ (16). Unfortunately, for almost all problems, the distribution is intractable. In such situations, we need to resort to approximation techniques, and these fall broadly into two classes, according to whether they rely on stochastic [41,45,168,169] or deterministic [170–173] approximations. See [174,175] for review papers of these techniques. In this article, we mainly focused on the stochastic approximation. The basic idea behind the methodology is to construct a Markov chain whose stationary distribution is the posterior distribution $\pi(\boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}|\mathbf{y}_{1:N})$ (16).

6.2. Gibbs Sampling Algorithm

We resort to MCMC technique [174] to sample from the full joint density $\pi(\boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}|\mathbf{y}_{1:N})$ (16). Among many MCMC techniques, we use the Gibbs sampling algorithm [39,168] to exploit the conditional independence [148] induced by the hierarchical formulation. A generic Gibbs sampler would cycle in turn through each of the conditional distributions for the parameter blocks $\boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}$, and $\boldsymbol{\Omega}$ as follows:

Step 1. Sample $\boldsymbol{\Theta}$ from its full conditional distribution

$$\pi(\boldsymbol{\Theta}|\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}, \mathbf{y}_{1:N}) \propto \left\{ \prod_{i=1}^{N} \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2\mathbf{I}) \right\} \cdot \left\{ \prod_{l=1}^{K} \mathcal{N}_N(\boldsymbol{\theta}_l|\mathbf{1}\alpha_l + \mathbf{X}\boldsymbol{\beta}_l, \omega_l^2\mathbf{I}) \right\}; \tag{17}$$

Step 2. Sample σ^2 from its full conditional distribution

$$\pi(\sigma^2|\boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}, \mathbf{y}_{1:N}) \propto \left\{ \prod_{i=1}^{N} \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2\mathbf{I}) \right\} \cdot \pi(\sigma^2); \tag{18}$$

Step 3. Sample $\boldsymbol{\alpha}$ from its full conditional distribution

$$\pi(\boldsymbol{\alpha}|\sigma^2, \boldsymbol{\Theta}, \boldsymbol{B}, \boldsymbol{\Omega}, \mathbf{y}_{1:N}) \propto \prod_{l=1}^{K} \mathcal{N}_N(\boldsymbol{\theta}_l|\mathbf{1}\alpha_l + \mathbf{X}\boldsymbol{\beta}_l, \omega_l^2\mathbf{I}) \cdot \pi(\alpha_l); \tag{19}$$

Step 4. Sample $\boldsymbol{B}$ from its full conditional distribution

$$\pi(\boldsymbol{B}|\boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Omega}, \mathbf{y}_{1:N}) \propto \prod_{l=1}^{K} \mathcal{N}_N(\boldsymbol{\theta}_l|\mathbf{1}\alpha_l + \mathbf{X}\boldsymbol{\beta}_l, \omega_l^2\mathbf{I}) \cdot \pi(\boldsymbol{\beta}_l); \tag{20}$$

Step 5. Sample $\boldsymbol{\Omega}$ from its full conditional distribution

$$\pi(\boldsymbol{\Omega}|\boldsymbol{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \mathbf{y}_{1:N}) \propto \prod_{l=1}^{K} \mathcal{N}_N(\boldsymbol{\theta}_l|\mathbf{1}\alpha_l + \mathbf{X}\boldsymbol{\beta}_l, \omega_l^2\mathbf{I}) \cdot \pi(\omega_l^2). \tag{21}$$

Sampling the model matrix $\boldsymbol{\Theta}$ (5) at ***Step 1*** is independent of the choice of the priors, which we discuss shortly. On the other hand, sampling the parameters $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ at ***Steps 2***, ***3***, ***4***, and ***5*** depends on the prior choices for the parameters in Stage 3 (9); we discuss this topic in Section 7.

6.3. Parallel Computation for Model Matrix

One of the most computer-intensive steps to implement the Gibbs sampler in Section 6.2 is ***Step 1*** to sample the model matrix $\boldsymbol{\Theta} \in \mathbb{R}^{K \times N}$ (5), or equivalently its entries $\{\theta_{li}\}_{l=1,i=1}^{K,N}$, from the full conditional distribution $\pi(\boldsymbol{\Theta}|\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}, \mathbf{y}_{1:N})$ (17). Clearly, the nonlinear participation of the model parameters to the function f makes the conditional distribution intractable, hence, non-conjugate sampling is unavoidable, which may suffer from a slow convergence. At the same times, due to the Markovian nature of the Gibbs algorithm, it is difficult to parallelize the whole steps of the Gibbs sampler, which creates difficulties in slower languages like R [176]. Nevertheless, the increasing number of parallel cores that are available at a very low price drives more and more interest in 'parallel sampling algorithms' that can benefit from the available parallel processing units on computers [177,178].

We suggest a framework of parallel computations to efficiently update the model matrix $\boldsymbol{\Theta} \in \mathbb{R}^{K \times N}$. This framework can be particularly appreciated under the setting of Bayesian nonlinear mixed effects models when the number of subjects N is a lot larger than the number of model parameters K ($N \gg K$).

The first version of parallel sampling algorithms is based on scalar updating. For the derivation, we start with analyzing the full conditional posterior distribution of a single element θ_{li} $(l = 1, \cdots, K; i = 1, \cdots, N)$:

$$\begin{aligned}\pi(\theta_{li}|-) = \pi(\theta_{li}|\theta_{-li}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}, \mathbf{y}_{1:N}) &\propto \pi(\boldsymbol{\Theta}|\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}, \mathbf{y}_{1:N}) \\ &\propto \left\{\prod_{i=1}^{N} \mathcal{N}_{M_i}(\mathbf{y}_i|f_i(\mathbf{t}_i, \theta_{1i}, \cdots, \theta_{li}, \cdots, \theta_{Ki}), \sigma^2\mathbf{I})\right\} \cdot \left\{\prod_{l=1}^{K}\prod_{i=1}^{N} \mathcal{N}(\theta_{li}|\alpha_l + \mathbf{x}_i^\top \boldsymbol{\beta}_l, \omega_l^2)\right\} \\ &\propto \mathcal{N}_{M_i}(\mathbf{y}_i|f_i(\mathbf{t}_i, \theta_{1i}, \cdots, \theta_{li}, \cdots, \theta_{Ki}), \sigma^2\mathbf{I}) \cdot \mathcal{N}(\theta_{li}|\alpha_l + \mathbf{x}_i^\top \boldsymbol{\beta}_l, \omega_l^2), \end{aligned} \quad (22)$$

where the notation θ_{-li} represents the all the entries of $\boldsymbol{\Theta}$ except for θ_{li}, that is, $\theta_{-li} = \{\theta_{li}\}_{l=1,i=1}^{K,N} - \{\theta_{li}\}$. Here, we used a conventional notation in Bayesian computation: '$\pi(\theta_{li}|-)$' indicates the density $\pi(\theta_{li}|\theta_{-li}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}, \mathbf{y}_{1:N})$, where the notation '$-$' in $\pi(\theta_{li}|-)$ implies all the parameters except for the θ_{li} in the basic model (2)–(4) along with N observations.

Note that the proportional part of the full conditional $\pi(\theta_{li}|-)$ (22) only involves the i-th column vector of the model matrix $\boldsymbol{\Theta}$ (5), that is, $\boldsymbol{\theta}^i = (\theta_{1i}, \cdots, \theta_{Ki})^\top \in \mathbb{R}^K$ in its analytic expression. This implies that we can update the K entries of the vector $\boldsymbol{\theta}^i$ $(i = 1, \cdots, N)$ independently across subjects. Parallel sampling algorithm can be completed by assigning a single CPU process to each of the subjects i. Within the step to sample the K entries from the vector $\boldsymbol{\theta}^i$, it is required to use Gibbs iterative procedure to update the scalar components. Authors [30,31] applied this technique to update the model matrix for a Bayesian nonlinear mixed effects model to train the dataset explained in Sections 3.3 and 3.5. Figure 10 displays the schematic idea of the parallel sampling algorithm.

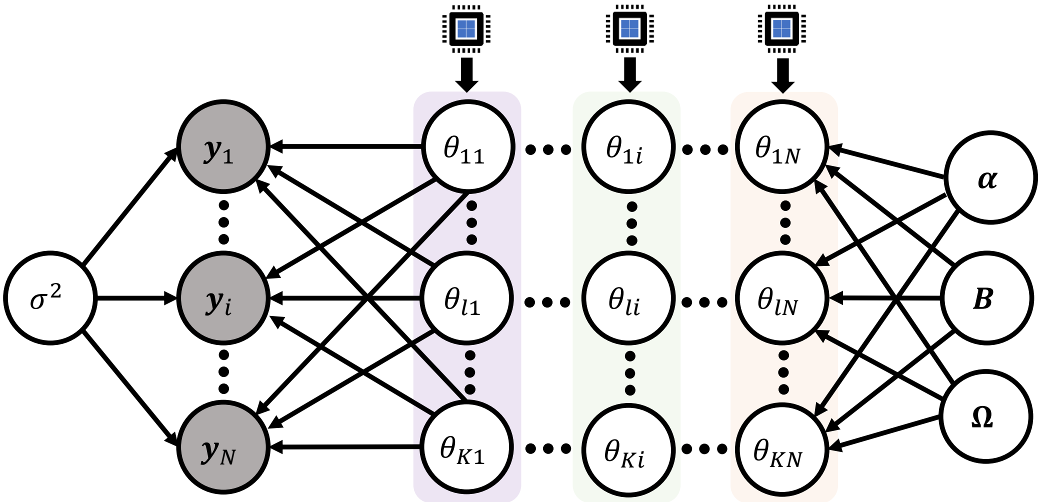

Figure 10. A pictorial description on the use of parallel computation to the basic model (2)–(4) to update the model matrix $\boldsymbol{\Theta} = [\boldsymbol{\theta}^1 \cdots \boldsymbol{\theta}^i \cdots \boldsymbol{\theta}^N] \in \mathbb{R}^{K \times N}$ (5). In the parallel computation, a single CPU is assigned to an individual subject $i = 1, \cdots, N$.

The second version of parallel sampling algorithms is based on vector updating. We analyze the full conditional posterior distribution of the vector $\boldsymbol{\theta}^i$ $(i = 1, \cdots, N)$:

$$\begin{aligned}\pi(\boldsymbol{\theta}^i|-) = \pi(\boldsymbol{\theta}^i|\boldsymbol{\theta}^{-i}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}, \mathbf{y}_{1:N}) &\propto \pi(\boldsymbol{\Theta}|\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}, \mathbf{y}_{1:N}) \\ &\propto \prod_{i=1}^{N} \mathcal{N}_{M_i}(\mathbf{y}_i|f_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2\mathbf{I}) \cdot \mathcal{N}_K(\boldsymbol{\theta}^i|\boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i, \boldsymbol{\Omega}) \\ &\propto \mathcal{N}_{M_i}(\mathbf{y}_i|f_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2\mathbf{I}) \cdot \mathcal{N}_K(\boldsymbol{\theta}^i|\boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i, \boldsymbol{\Omega}), \end{aligned} \quad (23)$$

where the notation $\boldsymbol{\theta}^{-i}$ represents the all the column vectors of $\boldsymbol{\Theta}$ except for $\boldsymbol{\theta}^i$. Similar to the first version, we can use the parallel computation to update the model matrix $\boldsymbol{\Theta}$ by simultaneously sampling from the full-conditional density $\pi(\boldsymbol{\theta}^i|-)$ (23) across subjects by assigning one CPU to each individual.

6.4. Elliptical Slice Sampler

Due to the issue of non-conjugacy to sample from the univariate density $\pi(\theta_{li}|-)$ (22) or K-dimensional density $\pi(\boldsymbol{\theta}^i|-)$ (23), the choice of a suitable MCMC method and further the choice of a proposal distribution is crucial for the fast convergence of the Markov chain simulated from the ***Step 1*** within the Gibbs sampler in Section 6.2. The Metropolis-Hastings (MH) algorithm [179,180] is the first solution to consider in such intractable situations: see the Algorithm 1 from [181]. In practice, the performances of the MH algorithm are highly dependent on the choice of the proposal density [182]. In the past decades, numerous MH-type algorithms to improve computational efficiency have been developed, and these fall broadly into two classes, according to whether the proposal density reflects a gradient information [41,43,44,183] or not [169,184]. In specific, the gradient information, here, refers to the first-order derivative of the minus of the log of the target density (i.e., $\nabla U(\theta_{li}) = -\nabla \log\ \pi(\theta_{li}|-) \in \mathbb{R}$ or $\nabla U(\boldsymbol{\theta}^i) = -\nabla \log\ \pi(\boldsymbol{\theta}^i|-) \in \mathbb{R}^K$, where the notation ∇ represents the gradient operator). Typically, gradient-based samplers are attractive in terms of rapid exploration of the state space, but the cost of the gradient computation can be prohibitive when the sample size N or model dimension K is extremely large [185,186]. Fortunately, this requirement can be made less burdensome by using automatic differentiation [187].

In the present subsection, we introduce an efficient gradient-free sampling technique, the elliptical slice sampler (ESS) proposed by [169], to simulate a Markov chain from the density $\pi(\theta_{li}|-)$ (22). The sampling logic can be directly applied to the situation to sample from the density $\pi(\boldsymbol{\theta}^i|-)$ (23) by simply replacing θ_{li} by $\boldsymbol{\theta}^i$ and changing the dimension of relevant distributions, stochastic processes, etc, from 1 to K. Conceptually, MH and ESS algorithms are similar in that both comprise two steps: a proposal step and a criterion step. A difference between the two algorithms arises in the criterion step. If a new candidate does not pass the criterion, then the MH algorithm takes the current state as the next state: whereas ESS re-proposes a new candidate until rejection does not take place, rendering the algorithm rejection-free. The utility of ESS can be appreciated when an analytic expression of a target density can be factored to have a Gaussian prior distribution. Unlike the traditional MH algorithm that requires the proposal variance or density, ESS is fully automated, and no tuning is required.

To adapt the ESS to our example, we re-write the density $\pi(\theta_{li}|-)$ (22) as the following form:

$$\pi(\theta_{li}|-) = \frac{1}{Z}\mathcal{L}(\theta_{li}) \cdot \mathcal{N}(\theta_{li}|\alpha_l + \mathbf{x}_i^\top \boldsymbol{\beta}_l, \omega_l^2), \tag{24}$$

where $\mathcal{L}(\theta_{li}) = \exp\{-\|\mathbf{y}_i - \boldsymbol{f}_i(\mathbf{t}_i, \theta_{1i}, \cdots, \theta_{li}, \cdots, \theta_{Ki})\|_2^2/(2\sigma^2)\}$, and Z is the normalization constant. By introducing the notation $\mathcal{L}(\theta_{li})$, it is our intention that we shall treat the function as a likelihood part temporarily only when sampling from the density (24). Alternatively, one can proceed with the choice $\mathcal{L}_2(\theta_{li}) = \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, \theta_{1i}, \cdots, \theta_{li}, \cdots, \theta_{Ki}), \sigma^2\mathbf{I})$ as a likelihood part to operate ESS, which then changes the normalization constant accordingly. We recommend to use the simplest functional form for the likelihood part, if possible, to reduce the computation cost.

Algorithm 1 summarizes the ESS in an algorithmic form, where the situation is at the $(s+1)$-th iteration of the Gibbs sampler. Therefore, the goal is to draw $\theta_{li}^{(s+1)}$ from the target density $\pi(\theta_{li}|-)$ ($l = 1, \cdots, K; i = 1, \cdots, N$) (24), where we already have $\theta_{li}^{(s)}$ as the current state for the target variable θ_{li} realized from the s-th iteration:

Algorithm 1: ESS to sample from $\pi(\theta_{li}|-)$ (22)

```
**Goal:** Sampling from the full conditional posterior distribution

    $\pi(\theta_{li}|-) \propto \mathcal{L}(\theta_{li}) \cdot \mathcal{N}(\theta_{li}|\mu_{li}, \omega_l^2),$

 where $\mathcal{L}(\theta_{li}) = \exp\{-\|\mathbf{y}_i - \boldsymbol{f}_i(\mathbf{t}_i, \theta_{1i}, \cdots, \theta_{li}, \cdots, \theta_{Ki})\|_2^2/(2\sigma^2)\}$ and
 $\mu_{li} = \alpha_l + \mathbf{x}_i^\top \boldsymbol{\beta}_l$.
**Input:** Current state $\theta_{li}^{(s)}$.
**Output:** A new state $\theta_{li}^{(s+1)}$.
a.  Choose an ellipse centered at $\mu_{li}$: $\nu \sim \mathcal{N}(\mu_{li}, \omega_l^2)$.
b.  Define a criterion function:

    $\alpha(\theta_{li}, \theta_{li}^{(s)}) = \min\{\mathcal{L}(\theta_{li})/\mathcal{L}(\theta_{li}^{(s)}), 1\} : \mathbb{R} \to [0,1].$

c.  Choose a threshold and fix: $u \sim \mathcal{U}nif[0,1]$.
d.  Draw an initial proposal $\theta_{li}^*$:

    $\phi \sim \mathcal{U}nif(-\pi, \pi];$
    $\theta_{li}^* = (\theta_{li}^{(s)} - \mu_{li}) \cos\phi + (\nu - \mu_{li}) \sin\ \phi + \mu_{li}.$

e.  **if (** $u < \alpha(\theta_{li}^*, \theta_{li}^{(s)})$ **)** { $\theta_{li}^{(s+1)} = \theta_{li}^*$ } **else** {
      Define a bracket : $(\phi_{\min}, \phi_{\max}] = (-\pi, \pi]$.
      **while (** $u \geq \alpha(\theta_{li}^*, \theta_{li}^{(s)})$ **)** {
        Shrink the bracket and try a new point :
        **if (** $\phi > 0$ **)** $\phi_{\max} = \phi$ **else** $\phi_{\min} = \phi$
        $\phi \sim \mathcal{U}nif(\phi_{\min}, \phi_{\max}]$
        $\theta_{li}^* = (\theta_{li}^{(s)} - \mu_{li}) \cos\phi + (\nu - \mu_{li}) \sin\ \phi + \mu_{li}.$
        }
        $\theta_{li}^{(s+1)} = \theta_{li}^*$
      }
```

6.5. Metropolis Adjusted Langevin Algorithm

We introduce the Metropolis adjusted Langevin algorithm (MALA) [43,44] which is popular for its use of problem-specific proposal distribution based on the gradient information of the target density. The main idea of MALA is to use Langevin dynamics to construct the Markov chain. To adapt the sampling technique to our example, we re-write the density $\pi(\theta_{li}|-)$ (22) as the following form:

$$\pi(\theta_{li}|-) = \frac{\exp(-U(\theta_{li}))}{\int_{-\infty}^{\infty} \exp(-U(\theta_{li}))d\theta_{li}} \quad \text{for all } \theta_{li} \in \mathbb{R}, \tag{25}$$

where $U(\theta_{li}) = (1/\{2\sigma^2\}) \cdot \|\mathbf{y}_i - \boldsymbol{f}_i(\mathbf{t}_i; \theta_{1i}, \cdots, \theta_{li}, \cdots, \theta_{Ki})\|_2^2 + (1/\{2\omega_l^2\}) \cdot (\theta_{li} - \alpha_l - \mathbf{x}_i^\top \boldsymbol{\beta}_l)^2$. Now, we consider a stochastic differential equation [188] that characterizes the evolution of the Langevin diffusion with the drift term set by the gradient of the log of the density (25):

$$d\theta_{li}(t) = \nabla \log \pi(\theta_{li}(t)|-)dt + \sqrt{2}dW(t) = -\nabla U(\theta_{li}(t))dt + \sqrt{2}dW(t), \tag{26}$$

where $\{W(t) \mid t \geq 0\}$ is a standard 1-dimensional Wiener process, or Brownian motion [189]. In (26), t indexes a fictitious continuous time. ∇ represents the gradient operator with respect to θ_{li}. Under fairly mild conditions on the function $U(\theta_{li})$, Equation (26) has a strong solution $\{\theta_{li}(t) \mid t \geq 0\}$ that is a Markov process [190]. Furthermore, it can be shown that the distribution of $\{\theta_{li}(t) \mid t \geq 0\}$ converges to the invariant distribution $\pi(\theta_{li}|-)$ (25) as $t \to \infty$.

Since solving the Equation (26) is very difficult, a first-order Euler-Maruyama discretization [191] is used to approximate the solution to the equation:

$$\theta_{li}^{[s+1]} \leftarrow \theta_{li}^{[s]} - \delta \cdot \nabla U(\theta_{li}^{[s]}) + \sqrt{2\delta}Z, \quad Z \sim \mathcal{N}(0,1), \tag{27}$$

where δ is the step size of discretization, and $[s]$ indexes the discrete time steps. This recursive update defines the Langevin Monte Carlo algorithm. Typically, to handle the discretization error and satisfy the detailed balance [192] to make Markov chain converge to the target distribution $\pi(\theta_{li}|-)$ (25), the MH correction is needed. Algorithm 2 details MALA to sample from the $\pi(\theta_{li}|-)$ $(l = 1, \cdots, K; i = 1, \cdots, N)$ (25):

```
Algorithm 2: MALA to sample from $\pi(\theta_{li}|-)$ (22)
    Goal: Sampling from the full conditional posterior distribution
        $\pi(\theta_{li}|-) \propto \exp\left(-U(\theta_{1i})\right),$
      where
      $U(\theta_{li}) = \|\mathbf{y}_i - \boldsymbol{f}_i(\mathbf{t}_i; \theta_{1i}, \cdots, \theta_{li}, \cdots, \theta_{Ki})\|_2^2/(2\sigma^2) + (\theta_{li} - \alpha_l - \mathbf{x}_i^\top \boldsymbol{\beta}_l)^2/(2\omega_l^2).$
    Input: Current state $\theta_{li}^{(s)}$ and step size $\delta$.
    Output: A new state $\theta_{li}^{(s+1)}$.
    a. Define a criterion function:
        $\alpha(\theta_{li}, \theta_{li}^{(s)}) = \min\left\{ \frac{\exp\left(-U(\theta_{li})\right)}{\exp\left(-U(\theta_{li}^{(s)})\right)} \cdot \frac{\mathcal{J}(\theta_{li}^{(s)}|\theta_{li})}{\mathcal{J}(\theta_{li}|\theta_{li}^{(s)})}, 1 \right\} : \mathbb{R} \to [0,1].$
    b. Choose a threshold $u$: $u \sim \mathcal{U}nif[0,1]$.
    c. Draw a proposal $\theta_{li}^*$:
        $\theta_{li}^* \sim \mathcal{J}(\theta_{li}|\theta_{li}^{(s)}) = \mathcal{N}(\theta_{li}|\theta_{li}^{(s)} - \delta \cdot \nabla U(\theta_{li}^{(s)}), 2\delta).$
    d. if ( $u < \alpha(\theta_{li}^*, \theta_{li}^{(s)})$ ) { $\theta_{li}^{(s+1)} = \theta_{li}^*$ } else { $\theta_{li}^{(s+1)} = \theta_{li}^{(s)}$ }
```

6.6. Hamiltonian Monte Carlo

We introduce the Hamiltonian Monte Carlo (HMC) algorithm that employs Hamiltonian dynamics to efficiently explore the parameter space [41,183]. Among many MH-type sampling algorithms, HMC has been recognized as one of the most effective algorithms due to its rapid mixing rate and small discretization error. By that reason, HMC has been deployed as the default sampler in many open packages such as STAN [193] and TENSORFLOW [194]. A key idea of HMC distinctive from ESS and MALA is the introduction of an auxiliary momentum variable, which is typically assumed to follow as a Gaussian distribution and independent of the target variable. By doing so, the HMC can produce distant proposals for the target variable, thereby avoiding the slow exploration of the state space that results from the diffusive behavior of simple random-walk proposals.

We adapt the HMC to our example. We shall first take a look at a joint density:

$$\begin{aligned} \pi(\theta_{li}, \phi_{li}|-) &= \pi(\theta_{li}|-) \cdot \pi(\phi_{li}) \\ &= \frac{\exp(-U(\theta_{li}))}{\int_{-\infty}^{\infty} \exp(-U(\theta_{li}))d\theta_{li}} \cdot \frac{1}{\sqrt{2\pi m_{li}}} \exp\left(-\frac{\phi_{li}^2}{2m_{li}}\right) \\ &\propto \exp\left(-U(\theta_{li}) - K(\phi_{li})\right) \quad \text{for all } (\theta_{li}, \phi_{li}) \in \mathbb{R} \times \mathbb{R}, \end{aligned} \tag{28}$$

where $U(\theta_{li}) = (1/\{2\sigma^2\}) \cdot \|\mathbf{y}_i - \boldsymbol{f}_i(\mathbf{t}_i; \theta_{1i}, \cdots, \theta_{li}, \cdots, \theta_{Ki})\|_2^2 + (1/\{2\omega_l^2\}) \cdot (\theta_{li} - \alpha_l - \mathbf{x}_i^\top \boldsymbol{\beta}_l)^2$ and $K(\phi_{li}) = \phi_{li}^2/(2m_{li})$. The auxiliary variable ϕ_{li} is distributed according to the univariate Gaussian distribution $\mathcal{N}(\phi_{li}|0, m_{li})$ with variance m_{li}. Note that it holds $\int \pi(\theta_{li}, \phi_{li}|-) d\phi_{li} = \pi(\theta_{li}|-)$ due to the independence between θ_{li} and ϕ_{li}. Therefore, our ultimate goal is to sample from the joint density $\pi(\theta_{li}, \phi_{li}|-)$ (28), and take only θ_{li} by marginalization.

Noting from (28), the negative of joint log-posterior is

$$H(\theta_{li}, \phi_{li}) = U(\theta_{li}) + K(\phi_{li}), \quad \text{for all } (\theta_{li}, \phi_{li}) \in \mathbb{R} \times \mathbb{R}. \tag{29}$$

The physical analogy of the bivariate function $H(\theta_{li}, \phi_{li}) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ (29) is a Hamiltonian [183,195], which describes the sum of a potential energy $U(\theta_{li})$, defined at the position θ_{li}, and a kinetic energy $K(\phi_{li}) = \phi_{li}^2/(2m_{li})$, where the auxiliary variable ϕ_{li} can be interpreted as a momentum variable and the variance m_{li} denotes a mass.

Now, we construct a Hamiltonian system by taking a derivative of H (29) with respect to θ_{li} and ϕ_{li}, and by introducing a continuous fictitious time t:

$$\frac{d\theta_{li}(t)}{dt} = \frac{\partial H(\theta_{li}, \phi_{li})}{\partial \phi_{li}} = \frac{\partial K(\phi_{li})}{\partial \phi_{li}} = \frac{\phi_{li}(t)}{m_{li}}, \tag{30}$$

$$\frac{d\phi_{li}(t)}{dt} = -\frac{\partial H(\theta_{li}, \phi_{li})}{\partial \theta_{li}} = -\frac{\partial U(\theta_{li})}{\partial \theta_{li}} = -\nabla U(\theta_{li}(t)), \tag{31}$$

where ∇ represents the gradient operator with respect to θ_{li}.

The Hamiltonian systems (30) and (31) have three nice properties. Assume that $(\theta_{li}(t), \phi_{li}(t)) : [a,b] \to \mathbb{R} \times \mathbb{R}$ is a solution curve of the system, where $a, b \in \mathbb{R} \cup \{\pm\infty\}$. Then, the following relationships hold:

(a) *Preservation of total energy* : $H(\theta_{li}(t), \phi_{li}(t)) = H(\theta_{li}(0), \phi_{li}(0))$ for all $t \in [a,b]$;

(b) *Preservation of volume*: $d\theta_{li}(t) d\phi_{li}(t) = d\theta_{li}(0) d\phi_{li}(0)$ for all $t \in [a,b]$;

(c) *Time reversibility*: The mapping T_s from state at t, $(\theta_{li}(t), \phi_{li}(t))$, to the state at time $t+s$, $(\theta_{li}(t+s), \phi_{li}(t+s))$, is one-to-one, and hence has an inverse T_{-s}.

Three properties are eventually related with the following nice properties of the HMC: (a) a high probability of acceptance of proposals; (b) a simple analytic form of acceptance ratio (no need to consider a hard-to-compute Jacobian factor); and (c) a detailed balance with respect to the target density $\pi(\theta_{li}|-)$. For a detailed description and extensive review, see [41].

For practical applications, the differential equation system (30) and (31) cannot be solved analytically and numerical methods are required. As the Hamiltonian H in the system is separable (or equivalently, the joint density $\pi(\theta_{li}, \phi_{li}|-)$ is factorizable), to traverse the state space more efficiently, the leapfrog integrator method is typically used, which involves a discretized step of the dynamics. As similar to the construction of MALA, the discretization errors arising from the leapfrog integration are addressed by MH correction step. Algorithm 3 details the HMC to sample from the target density $\pi(\theta_{li}|-) = \int \pi(\theta_{li}, \phi_{li}|-) d\phi_{li}$ (22). In the algorithm, (s) indexes the sampling iteration within the Gibbs sampler, while $[d]$ represents the index introduced due to the discretization.

One caveat in the HMC is that no matter whether we accept or reject the proposal, we draw a new momentum from the kinetic energy at every iteration. To check this, see the Step a in Algorithm 3, where $\phi_{li}^{[0]}$ is drawn from the kinetic density $K(\phi_{li}) \propto \mathcal{N}(0, m_{li})$. The momentum $\phi_{li}^{[0]}$ is only used to formulate the initial pair $(\theta_{li}^{[0]}, \phi_{li}^{[0]})$, where $\theta_{li}^{[0]}$ is the current target state $\theta_{li}^{(s)}$, that will be guided by Hamiltonian dynamics (30) and (31) via leapfrog integrator, eventually reaching the last pair $(\theta_{li}^{[L]} \phi_{li}^{[L]})$ which is used as the proposal $(\theta_{li}^*, \phi_{li}^*)$. The momentum $\phi_{li}^{[0]}$ is deleted and we will draw a new momentum in the next iteration. This independent drawing of the momentum is the engine that enables HMC to produce distant proposals, but nevertheless maintains a high probability of acceptance.

```
Algorithm 3: HMC to sample from $\pi(\theta_{li}|-) = \int \pi(\theta_{li}, \phi_{li}|-)d\phi_{li}$ (22)
    Goal: Sampling from the full conditional posterior distribution
        $$\pi(\theta_{li}|-) = \int \pi(\theta_{li}, \phi_{li}|-)d\phi_{li} \propto \exp\,(-U(\theta_{li})),$$
      where $\pi(\theta_{li}, \phi_{li}|-) \propto \exp\,(-H(\theta_{li}, \phi_{li})) = \exp\,(-U(\theta_{li}) - K(\phi_{li}))$ with
      $U(\theta_{li}) = \|\mathbf{y}_i - \boldsymbol{f}_i(\,\mathbf{t}_i; \theta_{1i}, \cdots, \theta_{li}, \cdots, \theta_{Ki})\|_2^2/(2\sigma^2) + (\theta_{li} - \alpha_l - \mathbf{x}_i^\top \boldsymbol{\beta}_l)^2/(2\omega_l^2)$
      and $K(\phi_{li}) = \phi_{li}^2/(2m_{li})$.
    Input: Current state $\theta_{li}^{(s)}$, step size $\delta$, number of steps $L$, and mass $m_{li}$.
    Output: A new state $\theta_{li}^{(s+1)}$.
    a.  Generate the initial momentum with mass $m_{li}$: $\phi_{li}^{[0]} \sim \mathcal{N}(0, m_{li})$.
    b.  Define a criterion function:
        $$\alpha((\theta_{li}, \phi_{li}), (\theta_{li}^{(s)}\phi_{li}^{[0]})) = \min\left\{\frac{\exp\,(-H(\theta_{li}, \phi_{li}))}{\exp\,(-H(\theta_{li}^{(s)}, \phi_{li}^{[0]}))}, 1\right\} : \mathbb{R} \times \mathbb{R} \to [0, 1].$$
    c.  Simulate discretization of Hamiltonian dynamics (30) and (31):
        i.   Set the initial pair of the solution curve: $(\theta_{li}^{[0]}, \phi_{li}^{[0]}) = (\theta_{li}^{(s)}\phi_{li}^{[0]})$.
        ii.  Make a half step for the momentum: $\phi_{li}^{[0]} \leftarrow \phi_{li}^{[0]} - (\delta/2)\nabla U(\theta_{li}^{[0]})$.
        iii. Alternate full steps for position and momentum:
             for ( $d = 1, \cdots, L$ ) {
             Update position: $\theta_{li}^{[d]} \leftarrow \theta_{li}^{[d-1]} + (\delta/m_{li})\phi_{li}^{[d-1]}$.
             Update momentum: if ( $d! = L$ ) { $\phi_{li}^{[d]} \leftarrow \phi_{li}^{[d-1]} - \delta\nabla U(\theta_{li}^{[d]})$. }
             }
        iv.  Make a half step for momentum: $\phi_{li}^{[L]} \leftarrow \phi_{li}^{[L-1]} - (\delta/2)\nabla U(\theta_{li}^{[L]})$.
        v.   Negate the last momentum: $\phi_{li}^{[L]} \leftarrow -\phi_{li}^{[L]}$.
        vi.  Set the last pair of the solution curve as the proposal: $(\theta_{li}^*, \phi_{li}^*) = (\theta_{li}^{[L]}\phi_{li}^{[L]})$.
    d.  Choose a threshold $u$: $u \sim \mathcal{U}nif[0, 1]$.
    e.  if ( $u < \alpha((\theta_{li}^*, \phi_{li}^*), (\theta_{li}^{(s)}\phi_{li}^{[0]}))$ ) { $\theta_{li}^{(s+1)} = \theta_{li}^*$ } else { $\theta_{li}^{(s+1)} = \theta_{li}^{(s)}$ }
```

The naive HMC (Algorithm 3) requires the users to specify at least three parameters: a step size δ, a number of steps L, and a mass m_{li}, for which to run a simulated Hamiltonian system. A poor choice of either of these parameters will result in a dramatic drop in the efficiency HMC. No-U-Turn Sampler (NUTS) developed by [42] is an extension of HMC, which is designed to automatically turn the parameters (δ, L) while fixing $m_{li} = 1$, making it possible to run NUTS with no hand-tuning at all. HMC and NUTS are general-purpose inference engines deployed in STAN.

We would like to highlight a difference between MALA (Algorithm 2) and HMC (Algorithm 3). Although both algorithms utilizes the gradient information (that is, $\nabla U(\theta_{li}) = -\nabla \log\ \pi(\theta_{li}|-)$), the former is based on stochastic differential Equation (26) and the latter is based on ordinary differential Equations (30) and (31). From the algorithmic perspective, MALA exhibits a single loop structure and is designed to directly employ the discretization of the underlying Langevin dynamics: see that the index $[s]$ resulting from the discretization in (27) is directly used as the sampling index (s) in Algorithm 2. On the other hand, HMC has a double loop structure: the inner loop (i.e., Step c in Algorithm 3) solves the Hamiltonian dynamics (30) and (31) to make a proposal, while the outer loop judges the proposals. The index $[d]$ in the inner loop, resulting from the leapfrog integrator, and the sampling index (s) of the outer loop in Algorithm 3 are not related [196]. Therefore, one can set the number of steps L for the leapfrog integrator by an arbitrary integer.

7. Prior Options

7.1. Priors for Variance

Provided the assumption of the basic form (2)–(4), the random error terms $\{\epsilon_{ij}\}_{i=1,j=1}^{N,M_i}$ in Stage 1 and $\{\eta_{li}\}_{l=1,i=1}^{K,N}$ in Stage 2 are the stochastic sources of (remaining) intra-individual and inter-individual variabilities, respectively [8]. Both terms are assumed to follow univariate Gaussian distributions in the basic model. This assumption can be generalized to multivariate Gaussian distribution, t-distribution, mixture of Gaussian distributions, etc., depending on the exhibition of the data or prior guess of perturbation associated with model matrix $\boldsymbol{\Theta}$ (5) [197].

Recall that the basic model assumes that data-level errors ϵ_{ij} are distributed according to $\mathcal{N}(0, \sigma^2)$ with variance σ^2, independently across times $j = 1, \cdots, M_i$ and subjects $i = 1, \cdots, N$. We discuss about the η-term in Section 7.3. Therefore, the standard deviation σ describes a vertical difference (i.e., measurement error) between the observation y_{ij} and theory $f(t_{ij}; \boldsymbol{\theta}^i)$ across time and individuals. We can generalize the basic setting by replacing σ^2 with (a) σ_i^2 $(i = 1, \cdots, N)$ or (b) σ_{ij}^2 $(i = 1, \cdots, N; j = 1, \cdots, M_i)$ to accommodate the heterogeneity the measurement error (a) across subjects and (b) across subjects and time, respectively, provided sufficiently large sampling times M_i [28].

For any prior $\pi(\sigma^2)$, the full conditional posterior distribution of σ^2 (18) is given as

$$\pi(\sigma^2|\boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}, \mathbf{y}_{1:N}) \propto (\sigma^2)^{-\sum_{i=1}^N M_i/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N} \|\mathbf{y}_i - \boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i)\|_2^2\right) \cdot \pi(\sigma^2), \quad (32)$$

where $\|\mathbf{a}\|_2^2$ represents the Euclidean norm of the vector $\mathbf{a}$.

Popularly used priors $\pi(\sigma^2)$ (or $\pi(\sigma)$) are (i) the Jeffreys prior $\pi(\sigma^2) \propto 1/\sigma^2$ [198]; (ii) inverse-gamma prior $\pi(\sigma^2) = \mathcal{IG}(a_{\sigma^2}, b_{\sigma^2})$ with shape $a_{\omega_l^2} > 0$ and scale $b_{\omega_l^2} > 0$; and (iii) half-Cauchy prior $\pi(\sigma) = \mathcal{C}^+(0, b_\sigma) = \{2/(\pi b_\sigma)\} \cdot 1/\{1 + (\sigma/b_\sigma)^2\}$ with scale $b_\sigma > 0$. Note that half-Cauchy distribution should be given to the standard deviation σ, not variance σ^2. The first two prior options lead to the conjugate update to sample from the density $\pi(\sigma^2|-)$ (32). Although the third one induces non-conjugate update to sample from the density $\pi(\sigma|-)$, computationally efficient sampling can be constructed by using parameter expansion technique [199] or slice sampler [45].

7.2. Priors for Intercept and Coefficient Vector

One of the central goals of using nonlinear mixed effects models is to identify significant covariates among the P covariates $\mathbf{x} = (x_1, \cdots, x_P)^\top$, explaining each of the model parameters θ_l $(l = 1, \cdots, K)$. This is because the function f in Stage 1 (2) is typically derived from a differential equation system. Such a differential equation has model parameters $\{\theta_l\}_{l=1}^K$ that control the dynamic of the solution of the system, and how the parameters are related with covariates is vital to understand causality. For example, in PK analysis, understanding whether and to what extent weight, renal status, disease status, etc., are associated with drug clearance may dictate how these factors can be considered in a dosing schedule.

We explain popularly used priors for the intercept and coefficient vector by taking the vector-form (a) (Stage 1—(6), Stage 2—(7), and Stage 3—(9)) because it directly embeds the framework of linear regression. For each model parameter index $l = 1, \cdots, K$, we re-write the Equation (7) for the purpose of illustration:

$$\begin{bmatrix} \theta_{l1} \\ \vdots \\ \theta_{li} \\ \vdots \\ \theta_{lN} \end{bmatrix} = \begin{bmatrix} \alpha_l \\ \vdots \\ \alpha_l \\ \vdots \\ \alpha_l \end{bmatrix} + \begin{bmatrix} x_{11} & \cdots & x_{1b} & \cdots & x_{1P} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ib} & \cdots & x_{iP} \\ \vdots & & \vdots & & \vdots \\ x_{N1} & \cdots & x_{Nb} & \cdots & x_{NP} \end{bmatrix} \begin{bmatrix} \beta_{l1} \\ \vdots \\ \beta_{lb} \\ \vdots \\ \beta_{lP} \end{bmatrix} + \begin{bmatrix} \eta_{l1} \\ \vdots \\ \eta_{li} \\ \vdots \\ \eta_{lN} \end{bmatrix} \tag{33}$$

where $\boldsymbol{\eta}_l = (\eta_{l1}, \cdots, \eta_{li}, \cdots, \eta_{lN})^\top \sim \mathcal{N}_N(\mathbf{0}, \omega_l^2 \mathbf{I})$. By the assumption (9), we have priors $\alpha_l \sim \pi(\alpha_l)$ and $\beta_{lb} \sim \pi(\beta_{lb})$ $(b = 1, \cdots, P)$. Note that the Equation (33) is a Bayesian multivariate linear regression (page 149 of [60]), and the only difference from the usual context is that the response vector in (33) is latent. Therefore, almost all Bayesian regression techniques [200,201] can be used to the latent regression (33) provided that the model matrix $\boldsymbol{\Theta}$ (5) is efficiently realized in ***Step 1*** within the Gibbs sampler.

The default choice for the prior $\pi(\alpha_l)$ is the flat prior $\pi(\alpha_l) \propto 1$, also called a uniform prior. Alternatively, a diffuse Gaussian prior $\pi(\alpha_l) = \mathcal{N}(a_l, b_l^2)$ is also often used by fixing b_l to be sufficiently large (saying $b_l = 10$ or 100) and $a_l = 0$. In either case, the full conditional density (19) enjoys the conjugate update, hence, ***Step 3*** in the Gibbs sampler in Section 6.2 seldom imposes computational burden. The idea behind the use of (nearly) non-informative priors for the intercept is that such priors induce almost minimal degree of Bayesian shrinkage, and hence allow the data to have (nearly) maximum effect on the posterior estimate for the intercept [199].

Now, we discuss the Bayesian analysis for the coefficients. There are numerous choices for the prior of the coefficient vector $\boldsymbol{\beta}_l = (\beta_{l1}, \cdots, \beta_{lb}, \cdots, \beta_{lP}) \sim \pi(\boldsymbol{\beta}_l)$, which is not surprising because the linear regression is arguably one of the most researched topics in statistics. Here, we suggest some popular priors whose main utility fall broadly into two settings, according to whether the design matrix $\mathbf{X} \in \mathbb{R}^{N \times P}$ in (33) is tall $(N \geq P)$ or fat $(N < P)$. In regression theory, the former setting is referred to as low-dimensional regression, and the latter one is called high-dimensional regression [202–204].

Under the *tall design* $(N \geq P)$, particularly when the number of subjects N is much larger than the number of covariates P $(N \gg P)$, one important theoretical consideration is that it is expected to see Bernstein–von Mises type results [205–207] on the posterior inference for coefficients $\boldsymbol{\beta}_l$. Roughly speaking, the theorem, sometimes called the "Bayesian Central Limit Theorem", states that the posterior distribution of $\boldsymbol{\beta}_l$ is approximately a normal distribution following a likelihood theory as sample size N goes to infinity for any prior choice $\pi(\boldsymbol{\beta}_l)$ under certain regularity conditions. In reality, it is possible that due to an ill-conditioned design matrix $\mathbf{X}$, a misspecification of the error distribution for η_{li}, a small sample size N, an inappropriate choice of prior $\pi(\boldsymbol{\beta}_l)$, etc., the Bernstein–von Mises Phenomenon (page 151 from [208]) may not be empirically observed. However, even in such cases, it is known that the influence of the prior distribution diminishes as N grows, which means that using different priors for $\boldsymbol{\beta}_l$ may not sensitively change the resulting Bayesian inferences about $\boldsymbol{\beta}_l$, and furthermore, the inference outcomes obtained by Bayesian and frequentist methods agree in most instances under the tall design. For example, see results of [209] and [28] for pharmacokinetic applications. Some possible options for prior $\pi(\boldsymbol{\beta}_l)$ are (i) flat prior $\pi(\boldsymbol{\beta}_l) \propto 1$; (ii) Gaussian diffuse prior $\pi(\boldsymbol{\beta}_l) = \mathcal{N}_P(\mathbf{0}, \sigma_{\beta_l}^2 \mathbf{I})$ with a large variance $\sigma_{\beta_l}^2$ [210], and (iii) g-prior $\pi(\boldsymbol{\beta}_l) = \mathcal{N}_P(\mathbf{0}, g \cdot \omega_l^2 [\mathbf{X}^\top \mathbf{X}]^{-1})$ for some positive value g [211]. The suggested priors yield the conjugate update for ***Step 4*** within the Gibbs sampler.

Now, we discuss some priors for $\pi(\boldsymbol{\beta}_l)$ in the linear regression (33) under the *fat design setting* $(N \ll P)$. This setting can be applied to pharmacogenetics where one of the main interests is to find important genes that may influence pharmacokinetics or pharmacodynamics [212–214], where the number of genes P is allowed to be a few thousand, while the number of patients N is confined to a few hundred. A fundamental assumption in this setting is *sparsity assumption* on the coefficients $\boldsymbol{\beta}_l$. This means that many of the coefficients of $\boldsymbol{\beta}_l = (\beta_{l1}, \cdots, \beta_{lP})^\top$ are (close to) zero. The true non-zero coefficients in the $\boldsymbol{\beta}_l$ are referred to as signal coefficients, while the remaining ones are called noise coefficients.

Statisticians have devised a number of penalized regression techniques for estimating $\boldsymbol{\beta}_l$ under the sparsity assumption [215]. From a Bayesian point of view, sparsity favoring mixture priors with separate control on the signal and noise coefficients have been proposed [216–219], which is called the ‘spike-and-slab priors’. Although these priors often lead to attractive theoretical properties [200,220], computational issues and considerations that many of the β_{lb}’s ($b = 1, \cdots, P$) may be small but not exactly zero has led to a wide variety of ‘continuous shrinkage priors’ [221–225], which can be unified through a global-local scale mixture representation [226]. The following hierarchies describe the sparse favoring priors:

- **Spike-and-slab priors**. Each component of the coefficients $\boldsymbol{\beta}_l$ is assumed to be drawn from
$$\beta_{lb}|\tau_l \sim (1-\tau_l)\cdot\delta_0(\beta_{lb}) + \tau_l \cdot f(\beta_{lb}), \quad (l = 1, \cdots, K; b = 1, \cdots, P),$$
where $\tau_l = \mathbf{Pr}[\beta_{lb} \neq 0]$. The function $\delta_0(\beta_{lb})$ is the Direc-delta function and $f(\beta_{lb})$ is a density supported on $\mathbb{R}$, called the spike and slab densities, respectively. The spike density shrinks noise coefficients to the exact zero, while the slab density captures signal coefficients by allowing a positive mass on the tail region [200,227,228];
- **Continuous shrinkage priors**. Each component of the coefficients $\boldsymbol{\beta}_l$ is assumed to be drawn from
$$\beta_{lb}|\lambda_{lb}, \tau_l, \omega_l^2 \sim \mathcal{N}(0, \lambda_{lb}^2\tau_l^2\omega_l^2), \quad (l = 1, \cdots, K; b = 1, \cdots, P),$$
$$\lambda_{lb} \sim f(\lambda_{lb}), \tau_l \sim g(\tau_l), \omega_l^2 \sim h(\omega_l^2), \quad (l = 1, \cdots, K; b = 1, \cdots, P),$$
where f, g, and h are priors for λ_{lb}, τ_l, and ω_l^2, respectively, supported on $(0, \infty)$. Here, λ_{lb} and τ_l are referred to as local-scale and global-scale parameters, respectively. The choices of f and g play a key role in controlling the effective sparsity and concentration of the prior and posterior distributions [226,229–233].

Roughly speaking, the roles of the τ_l in both prior frameworks are similar in the sense that they control the degree of the sparsity [226]. Slab density and local-scale prior density are expected to put a sufficient mass on the tail regions of the densities to detect signal coefficients and produce a robust Bayes estimator for $\boldsymbol{\beta}_l$ [201,224]. For that reason, heavy-tailed densities (e.g., double generalized Pareto distribution, Cauchy distributions [234,235]) are preferably used. Refer to [236,237] for comprehensive surveys on Bayesian variable selection.

7.3. Priors for Covariance Matrix

Consider a nonlinear function $f(t; \boldsymbol{\theta})$ indexed by a K-dimensional model parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_l, \cdots, \theta_K)^\top$ to describe an individual trajectory. A basic assumption is that all the components θ_l are unrelated across l ($l = 1, \cdots, K$), which is referred to as *uncorrelated design setting*. In many practical problems, this setting is reasonably accepted since one of the fundamental assumptions on f is that each component θ_l has its own role in modifying a functional shape of f. A central goal of a researcher when using nonlinear mixed models is to examine these roles mathematically, endowed with interpretations by domain experts in terms of physiology, epidemiology, or pharmacology, etc. The basic model (2)–(4) that we illustrated so far is designed under this assumption; recall that the covariance matrix $\boldsymbol{\Omega} \in \mathbb{R}^{K\times K}$ on Stage 2 was assumed to be diagonal, $\boldsymbol{\Omega} = \text{diag}(\omega_1^2, \cdots, \omega_l^2, \cdots, \omega_K^2)$. Under this uncorrelated design setting, the possible options for priors for the scale components ω_l^2 (or ω_l) are (i) Jeffreys prior $\pi(\omega_l^2) \propto 1/\omega_l^2$ [198]; (ii) inverse-gamma prior

$\pi(\omega_l^2) = \mathcal{IG}(a_{\omega_l^2}, b_{\omega_l^2})$ with shape $a_{\omega_l^2} > 0$ and scale $b_{\omega_l^2} > 0$; and (iii) half-Cauchy prior $\pi(\omega_l) = \mathcal{C}^+(0, b_{\omega_l}) = \{2/(\pi b_{\omega_l})\} \cdot 1/\{1 + (\omega_l / b_{\omega_l})^2\}$ with scale $b_{\omega_l} > 0$. See discussion by [238] for the prior options implemented on *8-schools example*.

We discuss Bayesian inference about a population covariance matrix $\mathbf{\Omega}$ under a *correlated design setting*, where the off-diagonal entries of $\mathbf{\Omega}$ (34) are allowed to be non-zeros:

$$\mathbf{\Omega} = \begin{bmatrix} \omega_{11} & \cdots & \omega_{1l} & \cdots & \omega_{1K} \\ \vdots & & \vdots & & \vdots \\ \omega_{l1} & \cdots & \omega_{ll} & \cdots & \omega_{Kl} \\ \vdots & & \vdots & & \vdots \\ \omega_{K1} & \cdots & \omega_{Kl} & \cdots & \omega_{KK} \end{bmatrix} \in \mathbb{R}^{K \times K}. \tag{34}$$

In (34), the (l_1, l_2)-th entry is denoted by $\mathbf{\Omega}[l_1, l_2] = \omega_{l_1 l_2} = \text{cov}(\eta_{l_1 i}, \eta_{l_2 i})$ $(l_1, l_2 = 1, \cdots, K; i = 1, \cdots, N)$. With $l_1 = l_2$, we have $\omega_{l_1 l_2} = \omega_{l_1}^2$ $(l_1 = 1, \cdots, K)$.

Researchers often wish to work with the correlated design setting to examine whether any pair of model parameters, θ_{l_1} and θ_{l_2}, are physiologically (or epidemiology, pharmacology, financially, etc.) associated or not. Taking the term structure modeling discussed in Section 3.4 as an example, it is a valid question whether the Nelson–Siegel parameters are correlated or not as they are all associated with the interest rate [87]. Statistically, having a well-designed covariance structure can also improve the model fitting and produce reliable estimators for the model parameters compared to uncorrelated designs. To make a fully Bayesian inference about the $\mathbf{\Omega}$ (34), we need to specify an appropriate prior $\pi(\mathbf{\Omega})$ which we will discuss shortly. After that, we operate the Gibbs sampler in Section 6.2, with some modifications, if necessary. For instance, to implement the parallel computation in ***Step 1***, we recommend to sample from the joint density $\pi(\boldsymbol{\theta}^i|-)$ across i, instead of sampling from the individual $\pi(\theta_{li}|-)$. Especially, among the five steps of the Gibbs sampler, implementation of ***Step 5*** needs special care in sampling from the full-conditional posterior density $\mathbf{\Omega}$. This step can be highly complicated depending on the chosen prior.

A challenge in choosing a workable prior $\pi(\mathbf{\Omega})$ is briefly mentioned. Research regarding this subject are vast, growing, and deep. Similar to the obstacles encountered in classical covariance estimation [239–243], there are three major aspects, among many others, in the consideration of a thoughtful prior $\pi(\mathbf{\Omega})$ to produce a reliable Bayes estimator of $\mathbf{\Omega}$: (1) sample size N; (2) the number of model parameters K; and (3) positive-definiteness of $\mathbf{\Omega}$ [244]. The first two aspects are related to theoretical constraints. In general, it is known that the estimation of the covariance $\mathbf{\Omega}$ can be distorted unless the ratio K/N is sufficiently small (see, e.g., [245–247]). The third one is germane to the modeling consideration and computation strategies to estimate $\mathbf{\Omega}$, typically resolved via principal component analysis, Cholesky decomposition, and Gaussian graphical models, etc. [248,249].

For any prior $\pi(\mathbf{\Omega})$, the full conditional density of $\mathbf{\Omega}$ is analytically expressed as follows (see [250] for a similar derivation):

$$\begin{aligned} \pi(\mathbf{\Omega}|\mathbf{\Theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \mathbf{y}_{1:N}) \propto \pi(\mathbf{\Theta}|\boldsymbol{\alpha}, \boldsymbol{B}, \mathbf{\Omega}) \cdot \pi(\mathbf{\Omega}) &\propto \left\{ \prod_{i=1}^{N} \mathcal{N}_K(\boldsymbol{\theta}^i | \boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i, \mathbf{\Omega}) \right\} \cdot \pi(\mathbf{\Omega}) \\ &\propto (\det\mathbf{\Omega})^{-N/2} \exp\left(-\frac{1}{2} \text{tr}\left[\mathbf{\Omega}^{-1} \left\{ (N-1)\mathbf{G} + N(\bar{\boldsymbol{\phi}} - \boldsymbol{\alpha})(\bar{\boldsymbol{\phi}} - \boldsymbol{\alpha})^\top \right\} \right] \right) \cdot \pi(\mathbf{\Omega}), \end{aligned} \tag{35}$$

where $\boldsymbol{\phi}^i = \boldsymbol{\theta}^i - \boldsymbol{B}\mathbf{x}_i$ $(i = 1, \cdots, N)$, $\bar{\boldsymbol{\phi}} \in \mathbb{R}^K$, and $\mathbf{G} \in \mathbb{R}^{K \times K}$ are defined by

$$\bar{\boldsymbol{\phi}} = N^{-1} \sum_{i=1}^{N} \boldsymbol{\phi}^i = N^{-1} \sum_{i=1}^{N} (\boldsymbol{\theta}^i - \boldsymbol{B}\mathbf{x}_i), \quad \mathbf{G} = (N-1)^{-1} \sum_{i=1}^{N} (\boldsymbol{\phi}^i - \bar{\boldsymbol{\phi}})(\boldsymbol{\phi}^i - \bar{\boldsymbol{\phi}})^\top;$$

$$\sum_{i=1}^{N} (\boldsymbol{\phi}^i - \boldsymbol{\alpha})(\boldsymbol{\phi}^i - \boldsymbol{\alpha})^\top = (N-1)\mathbf{G} + N(\bar{\boldsymbol{\phi}} - \boldsymbol{\alpha})(\bar{\boldsymbol{\phi}} - \boldsymbol{\alpha})^\top.$$

In (35), we used the vector-form (b) (i.e., i-indexing) to express the prior for $\mathbf{\Theta}$. Notations $\det(\mathbf{A})$ and $\mathrm{tr}(\mathbf{A})$ denote the determinant and trace of a square matrix $\mathbf{A}$, respectively. Matrix $\mathbf{G}$ is the 'latent' covariance matrix based on the model matrix $\mathbf{\Theta} \in \mathbb{R}^{K\times K}$ (5) and coefficient matrix $\boldsymbol{G} \in \mathbb{R}^{K\times P}$, whose form resembles sample covariance matrix assuming $\boldsymbol{\phi}^i$ are observed [247].

Traditionally used priors for the covariance matrix $\mathbf{\Omega}$ (34) are the Jefferys prior and the conjugate inverse Wishart prior (see [251,252] for the reviews of the earlier works):

- **Jeffreys prior**. The common non-informative prior has been the Jeffreys improper prior

$$\pi(\mathbf{\Omega}) \propto (\det\mathbf{\Omega})^{-(K+1)/2}.$$

  This prior was originally derived from an invariance argument by [253] for the case $K = 1, 2$; and it was considered for arbitrary K by [250,254,255] to develop Bayesian multivariate theory.
- **Inverse-Wishart prior**. The common informative prior is the inverse-Wishart prior [256]

$$\pi(\mathbf{\Omega}) = \mathcal{IW}(\mathbf{V}, d) = \frac{(\det\mathbf{V})^{d/2}}{2^{dK/2}\Gamma_K(d/2)}(\det\mathbf{\Omega})^{-(d+K+1)/2}\exp\left(-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Omega}^{-1}\mathbf{V}\right]\right),$$

  where $\mathbf{\Omega}$ and $\mathbf{V}$ are K-by-K positive definite matrices, and $\Gamma_K(\cdot)$ is the multivariate gamma function [257]. $\mathbf{V}$ is a scale matrix, and $d(> K-1)$ is the number of degrees of freedom. Conventionally, d is chosen to be as small as possible to reflect vague prior knowledge. A univariate specialization ($K = 1$) is the inverse-gamma distribution.

The success of Bayesian computation and MCMC in the late 1980s opened up the potential of using more flexible non-conjugate priors for covariance matrices [258–260]. Limitations of the traditional priors studied by many statisticians also motivated them to develop a new prior. For example, some of them argued that the Jeffreys prior may not be really non-informative, particularly in high dimensional setting [249,261], and inverse Wishart prior is very restrictive and lacks flexibility [262]. Among many new priors developed for particular applications [263], a combination of *separation strategy* developed by [249] and LKJ prior [264] has been successful, heavily used in a variety of industrial problems, and relevant software has been developed, including R package STAN [17,193].

We illustrate a central idea of using the separation strategy [249] to estimate the population covariance matrix $\mathbf{\Omega} \in \mathbb{R}^{K\times K}$. First, we decompose the matrix $\mathbf{\Omega}$ (34) into two components, K standard deviations $\omega_l = \sqrt{\omega_{ll}} = \sqrt{\mathbf{\Omega}[l,l]}$ $(l = 1, \cdots, K)$ and correlation matrix $\mathbf{R} \in \mathbb{R}^{K\times K}$:

$$\mathbf{\Omega} = \mathrm{diag}(\boldsymbol{\omega})\,\mathbf{R}\,\mathrm{diag}(\boldsymbol{\omega}) \in \mathbb{R}^{K\times K}, \tag{36}$$

where $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_K)^\top$ is the K-dimensional vector of standard deviations, $\mathrm{diag}(\boldsymbol{\omega})$ is the diagonal matrix with diagonal elements $\boldsymbol{\omega}$, and $\mathbf{R}$ is the K-by-K correlation matrix. Second, we specify priors independently for $\boldsymbol{\omega}$ and $\mathbf{R}$, denoted as $\pi(\boldsymbol{\omega})$ and $\pi(\mathbf{R})$, respectively, so that we have the joint prior for $(\boldsymbol{\omega}, \mathbf{R})$, $\pi(\boldsymbol{\omega}, \mathbf{R}) = \pi(\boldsymbol{\omega}) \cdot \pi(\mathbf{R})$. Following notation from [249], let $\mathcal{R}^K$ denote the correlation matrix space. Then, the priors $\pi(\boldsymbol{\omega})$ and $\pi(\mathbf{R})$ are supported on $(0, \infty)^K$ and $\mathcal{R}^K$, respectively. Finally, draw sample from each of the full conditional posterior distributions $\pi(\boldsymbol{\omega}|\mathbf{R}, -)$ and $\pi(\mathbf{R}|\boldsymbol{\omega}, -)$ at time in ***Step 5*** within the Gibbs sampler in Section 6.2. This means that, we are not directly sampling the covariance matrix $\mathbf{\Omega}$ from $\pi(\mathbf{\Omega}|-)$ (35) as we would do when Jeffreys or inverse-Wishart prior were used for the prior.

Normally used prior options for the scale vector $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_l, \cdots, \omega_K)^\top$ are (i) $\log(\boldsymbol{\omega}) \sim \mathcal{N}_K(\mathbf{a}_\omega, \mathbf{B}_\omega)$, where $\log(\boldsymbol{\omega}) = (\log\omega_1, \cdots, \log\omega_l, \cdots, \log\omega_K)^\top$, with hyperparameters, mean $\mathbf{a}_\omega \in \mathbb{R}^K$ and covariance $\mathbf{B}_\omega \in \mathbb{R}^{K\times K}$ which is often diagonal [249]; and

(ii) $\omega_l \sim \mathcal{C}^+(0, b_{\omega_l})$ with the scale hyperparameter $b_{\omega_l} > 0$ [17]. As for the prior distribution for the correlation $\mathbf{R} \in \mathbb{R}^{K \times K}$, the LKJ prior proposed by [264] is popularly used:

- **LKJ prior**. LKJ prior is supported over the correlation matrix space $\mathcal{R}^K$, or equivalently over the set of $K \times K$ Cholesky factors of real symmetric positive definite matrces

$$\pi(\mathbf{R}) = \left[2^{\sum_{q=1}^{Q-1}(2\gamma-2+Q-q)(Q-q)} \prod_{q=1}^{Q-1} \mathcal{B}\left(\gamma + \frac{Q-q-1}{2}, \gamma + \frac{Q-q-1}{2}\right)^{Q-q}\right] (\det \mathbf{R})^{\gamma-1}, \tag{37}$$

with the shape parameter $\gamma > 0$. The function $\mathcal{B}(\alpha, \beta)$ is the beta function. If $\gamma = 1$, the density is uniform over the space $\mathcal{R}^K$; for $\gamma > 1$, the density increasingly concentrates mass around the identity matrix $\mathbf{I} \in \mathbb{R}^{K \times K}$ (i.e., favoring less correlation); for $\gamma < 1$, the density increasingly concentrates mass in the other direction, and has a trough at the identity matrix (i.e., favoring more correlation).

Note that the normalizing constant of the LKJ prior (37) is constant with respect to γ, therefore, we have $\pi(\mathbf{R}) \propto (\det \mathbf{R})^{\gamma-1}$, with the shape hyperparameter $\gamma > 0$. The behavior of LKJ prior with $\gamma = 1$ (i.e., $\pi(\mathbf{R}) \propto 1$) was studied by [249], where the author found that as K increases, the marginal correlations tend to concentrate around zeros (see Figure 1 from [249]), hence, model matrix $\boldsymbol{\Theta}$ (5) are more likely to be treated as in the uncorrelated design setting.

As for the hyperparameter specification for the LKJ prior, Stan Development Team [17] recommends to use $\gamma \geq 1$. This suggestion is also well-aligned with the original intention of using the separation strategy to make a variance-correlation structure by [249] in that: (1) the authors intend to choose a diffuse prior for $\pi(\mathbf{R})$ to reflect weak knowledge about the correlation $\mathbf{R}$, while (2) prior knowledge, possibly informative, shall be put on the scale parameters by specifying $\pi(\boldsymbol{\omega})$, as most statisticians are normally trained to do. Computational algorithm and theory concerning the LKJ prior can be found in [264–266].

8. Model Selection

8.1. Setting

The recent development of MCMC methods has made it possible to fit enormously large classes of models with the aim of exploring real world complexities of data [267]. This ability naturally led us to wish to compare several candidate models that vary substantially in the model complexities and choose the best model out of them. For example, authors [31] tried to compare four different rate decline curves to fit the production data from the 360 wells in Figure 4. Indeed, upstream petroleum engineers endeavor to find a rate decline curve describing the production trajectories as accurately as possible so that EUR can be accurately estimated. Another application of model selection can be found in PK analysis. Taking the theophylline data in Figure 3 as an example, PK modelers may debate whether they need to use a two- or three-compartment model with a nonlinear clearance to describe the PK exposure, or if just one-compartment model with a linear clearance is sufficient.

In the current section, our primary focus is to illustrate a Bayesian approach to compare multiple Bayesian nonlinear mixed effects models explaining the data $\{(\mathbf{y}_i, \mathbf{t}_i, \mathbf{x}_i)\}_{i=1}^N$ introduced in Section 3.6. To that end, we want to lay out the set-up that underlies our model selection procedure. Consider Stages 1 and 2 of the basic model (i.e., the hierarchy (2)–(3)), endowed with a joint prior $\pi(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$. Very importantly, we do *not* assume the independent prior assumption as we did in Stage 3 (4): any prior assumption on the parameters $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$ works fine in our framework for the model comparison. Therefore, the basic model (2)–(4) is considered as a subclass of candidate models that we want to compare. In our framework, we can also consider the correlated design setting discussed in Section 7.3, where the covariance matrix $\boldsymbol{\Omega}$ (34) is allowed to be any positive-definite matrix (i.e., does not need to be a diagonal matrix), as one of candidates. Eventually, in our framework, modelers have the freedom to choose (i) the nonlinear function f to describe the temporal profile $\mathbf{y}_i$ and (ii) prior distribution $\pi(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega})$.

Assume that a researcher wants to consider H functions, denoted as $f(t; \boldsymbol{\theta}_{[\mathcal{M}_1]})$, $f(t; \boldsymbol{\theta}_{[\mathcal{M}_2]}), \cdots, f(t; \boldsymbol{\theta}_{[\mathcal{M}_H]})$, as a possible option for the use of f in Stage 1. Here, the subscript '$[\mathcal{M}_h]$' on the model parameter $\boldsymbol{\theta}$ is noted to indicate for the model index ($h = 1, \cdots, H$). Obviously, the candidate functions $f(t; \boldsymbol{\theta}_{[\mathcal{M}_h]})$ ($h = 1, \cdots, H$) can have different functional forms dictated by different dimensions for the model parameters $\boldsymbol{\theta}_{[\mathcal{M}_h]} \in \mathbb{R}^{K_{[\mathcal{M}_h]}}$. This will consecutively change the dimensions of the parameter blocks, $\boldsymbol{\alpha}_{[\mathcal{M}_h]} \in \mathbb{R}^{K_{[\mathcal{M}_h]}}$, $\boldsymbol{B}_{[\mathcal{M}_h]} \in \mathbb{R}^{K_{[\mathcal{M}_h]} \times P}$, and $\boldsymbol{\Omega}_{[\mathcal{M}_h]} \in \mathbb{R}^{K_{[\mathcal{M}_h]} \times K_{[\mathcal{M}_h]}}$, accordingly, yet the support of the σ^2 remains the same with $(0, \infty)$ because we still consider the additive error model. After that, she now has the freedom to choose a prior $\pi(\sigma^2, \boldsymbol{\alpha}_{[\mathcal{M}_h]}, \boldsymbol{B}_{[\mathcal{M}_h]}, \boldsymbol{\Omega}_{[\mathcal{M}_h]})$ ($h = 1, \cdots, H$). There are infinitely many choices for the prior, and one can use priors discussed in Section 7.

Now, the fundamental question naturally arising at this point is "what is the best model among the H candidate models?" To illustrate, we write the situation more technically. With the above specifications, Stage 1 of each of the H candidate models is given as

$$\begin{aligned}
\mathcal{M}_1: &\quad y_{ij} = f(t_{ij}; (\boldsymbol{\theta}_{[\mathcal{M}_1]})^i) + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad (i = 1, \cdots, N; j = 1, \cdots, M_i); \\
&\quad \vdots \\
\mathcal{M}_h: &\quad y_{ij} = f(t_{ij}; (\boldsymbol{\theta}_{[\mathcal{M}_h]})^i) + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad (i = 1, \cdots, N; j = 1, \cdots, M_i); \\
&\quad \vdots \\
\mathcal{M}_H: &\quad y_{ij} = f(t_{ij}; (\boldsymbol{\theta}_{[\mathcal{M}_H]})^i) + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad (i = 1, \cdots, N; j = 1, \cdots, M_i),
\end{aligned}$$

where each of the models has the corresponding model matrix

$$\begin{aligned}
\mathcal{M}_1: &\quad \boldsymbol{\Theta}_{[\mathcal{M}_1]} = \left[(\boldsymbol{\theta}_{[\mathcal{M}_1]})^1 \quad \cdots \quad (\boldsymbol{\theta}_{[\mathcal{M}_1]})^i \quad \cdots \quad (\boldsymbol{\theta}_{[\mathcal{M}_1]})^N \right] \in \mathbb{R}^{K_{[\mathcal{M}_1]} \times N}; \\
&\quad \vdots \\
\mathcal{M}_h: &\quad \boldsymbol{\Theta}_{[\mathcal{M}_h]} = \left[(\boldsymbol{\theta}_{[\mathcal{M}_h]})^1 \quad \cdots \quad (\boldsymbol{\theta}_{[\mathcal{M}_h]})^i \quad \cdots \quad (\boldsymbol{\theta}_{[\mathcal{M}_h]})^N \right] \in \mathbb{R}^{K_{[\mathcal{M}_h]} \times N}; \\
&\quad \vdots \\
\mathcal{M}_H: &\quad \boldsymbol{\Theta}_{[\mathcal{M}_H]} = \left[(\boldsymbol{\theta}_{[\mathcal{M}_H]})^1 \quad \cdots \quad (\boldsymbol{\theta}_{[\mathcal{M}_H]})^i \quad \cdots \quad (\boldsymbol{\theta}_{[\mathcal{M}_H]})^N \right] \in \mathbb{R}^{K_{[\mathcal{M}_H]} \times N},
\end{aligned}$$

obtained by stacking individual-based vector horizontally as we did to obtain $\boldsymbol{\Theta}$ (5). Again, the number of rows of the matrix $\boldsymbol{\Theta}_{[\mathcal{M}_h]}$, that is, $K_{[\mathcal{M}_h]}$, depends on the choice of the function f. Stage 2 of each of the H models will then be

$$\begin{aligned}
\mathcal{M}_1: &\quad (\boldsymbol{\theta}_{[\mathcal{M}_1]})^i = \boldsymbol{\alpha}_{[\mathcal{M}_1]} + \boldsymbol{B}_{[\mathcal{M}_1]} \mathbf{x}_i + (\boldsymbol{\eta}_{[\mathcal{M}_1]})^i, \quad (\boldsymbol{\eta}_{[\mathcal{M}_1]})^i \sim \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Omega}_{[\mathcal{M}_1]}), \quad (i = 1, \cdots, N); \\
&\quad \vdots \\
\mathcal{M}_h: &\quad (\boldsymbol{\theta}_{[\mathcal{M}_h]})^i = \boldsymbol{\alpha}_{[\mathcal{M}_h]} + \boldsymbol{B}_{[\mathcal{M}_h]} \mathbf{x}_i + (\boldsymbol{\eta}_{[\mathcal{M}_h]})^i, \quad (\boldsymbol{\eta}_{[\mathcal{M}_h]})^i \sim \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Omega}_{[\mathcal{M}_h]}), \quad (i = 1, \cdots, N); \\
&\quad \vdots \\
\mathcal{M}_H: &\quad (\boldsymbol{\theta}_{[\mathcal{M}_H]})^i = \boldsymbol{\alpha}_{[\mathcal{M}_H]} + \boldsymbol{B}_{[\mathcal{M}_H]} \mathbf{x}_i + (\boldsymbol{\eta}_{[\mathcal{M}_H]})^i, \quad (\boldsymbol{\eta}_{[\mathcal{M}_H]})^i \sim \mathcal{N}_K(\mathbf{0}, \boldsymbol{\Omega}_{[\mathcal{M}_H]}), \quad (i = 1, \cdots, N).
\end{aligned}$$

Finally, Stage 3 of each of the H models is comprised of the prior:

$$\begin{aligned}
\mathcal{M}_1: &\quad (\sigma^2, \boldsymbol{\alpha}_{[\mathcal{M}_1]}, \boldsymbol{B}_{[\mathcal{M}_1]}, \boldsymbol{\Omega}_{[\mathcal{M}_1]}) \sim \pi(\sigma^2, \boldsymbol{\alpha}_{[\mathcal{M}_1]}, \boldsymbol{B}_{[\mathcal{M}_1]}, \boldsymbol{\Omega}_{[\mathcal{M}_1]}); \\
&\quad \vdots \\
\mathcal{M}_h: &\quad (\sigma^2, \boldsymbol{\alpha}_{[\mathcal{M}_h]}, \boldsymbol{B}_{[\mathcal{M}_h]}, \boldsymbol{\Omega}_{[\mathcal{M}_h]}) \sim \pi(\sigma^2, \boldsymbol{\alpha}_{[\mathcal{M}_h]}, \boldsymbol{B}_{[\mathcal{M}_h]}, \boldsymbol{\Omega}_{[\mathcal{M}_h]}); \\
&\quad \vdots \\
\mathcal{M}_H: &\quad (\sigma^2, \boldsymbol{\alpha}_{[\mathcal{M}_H]}, \boldsymbol{B}_{[\mathcal{M}_H]}, \boldsymbol{\Omega}_{[\mathcal{M}_H]}) \sim \pi(\sigma^2, \boldsymbol{\alpha}_{[\mathcal{M}_H]}, \boldsymbol{B}_{[\mathcal{M}_H]}, \boldsymbol{\Omega}_{[\mathcal{M}_H]}).
\end{aligned}$$

We describe three model comparison criteria that are popularly used in the literature: deviance information criterion (DIC) [55,268], widely applicable information criterion (WAIC) [269], and posterior predictive loss criterion (PPLC) [270]. As in frequentist information criteria [271–273], formulation of the DIC, WAIC, and PPLC also takes the two terms into a consideration: *goodness-of-fit* and *penalty for model complexity*. Because increasing (or decreasing) model complexity is accompanied by the risk of over-fitting (or under-fitting), models should be compared by trading-off these two terms. Particularly, as we are currently discussing about a Bayesian hierarchical model, these criteria are obviously depending on what part of the model specification is considered to be part of the likelihood, and what is not. Ref. [268] refer to this as the *focus issue*. For example, in the general form of a Bayesian hierarchical model consisting of a top-level likelihood $p(\mathbf{y}|\boldsymbol{\Upsilon})$ for data $\mathbf{y}$, a prior model $\pi(\boldsymbol{\Upsilon}|\eta)$, and a hyperprior $\pi(\eta)$, one might choose as the likelihood either the conditional density $p(\mathbf{y}|\boldsymbol{\Upsilon})$, or the marginal density $p(\mathbf{y}|\eta) = \int p(\mathbf{y}|\boldsymbol{\Upsilon})\pi(\boldsymbol{\Upsilon}|\eta)d\boldsymbol{\Upsilon}$. Based on [268], the former situation is referred to as "focus on $\boldsymbol{\Upsilon}$", while the latter situation is referred to as "focus on η", respectively.

In our case, we shall "focus on parameters $(\boldsymbol{\Theta}_{[\mathcal{M}_h]}, \sigma^2)$" $(h = 1, \cdots, H)$ used in the conditional density in Stage 1. For notational simplicity, we denote $\boldsymbol{\Upsilon}_{[\mathcal{M}_h]} = (\boldsymbol{\Theta}_{[\mathcal{M}_h]}, \sigma^2)$. Then, likelihood of each of the H models based on N observations $\{(\mathbf{y}_i, \mathbf{t}_i, \mathbf{x}_i)\}_{i=1}^N$ is

$$\begin{aligned}
\mathcal{M}_1: &\quad \mathcal{L}(\boldsymbol{\Upsilon}_{[\mathcal{M}_1]}|\mathbf{y}_{1:N}) = \prod_{i=1}^{N} \mathcal{L}((\boldsymbol{\Upsilon}_{[\mathcal{M}_1]})^i|\mathbf{y}_i) = \prod_{i=1}^{N} \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, (\boldsymbol{\theta}_{[\mathcal{M}_1]})^i), \sigma^2\mathbf{I}); \\
&\quad \vdots \\
\mathcal{M}_h: &\quad \mathcal{L}(\boldsymbol{\Upsilon}_{[\mathcal{M}_h]}|\mathbf{y}_{1:N}) = \prod_{i=1}^{N} \mathcal{L}((\boldsymbol{\Upsilon}_{[\mathcal{M}_h]})^i|\mathbf{y}_i) = \prod_{i=1}^{N} \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, (\boldsymbol{\theta}_{[\mathcal{M}_h]})^i), \sigma^2\mathbf{I}); \\
&\quad \vdots \\
\mathcal{M}_H: &\quad \mathcal{L}(\boldsymbol{\Upsilon}_{[\mathcal{M}_H]}|\mathbf{y}_{1:N}) = \prod_{i=1}^{N} \mathcal{L}((\boldsymbol{\Upsilon}_{[\mathcal{M}_H]})^i|\mathbf{y}_i) = \prod_{i=1}^{N} \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i, (\boldsymbol{\theta}_{[\mathcal{M}_H]})^i), \sigma^2\mathbf{I}),
\end{aligned}$$

where $\mathcal{L}((\boldsymbol{\Upsilon}_{[\mathcal{M}_h]})^i|\mathbf{y}_i)$ is the likelihood (i.e., data distribution) based on an individual data with parameter $(\boldsymbol{\Upsilon}_{[\mathcal{M}_h]})^i = ((\boldsymbol{\theta}_{[\mathcal{M}_h]})^i, \sigma^2)$ $(i = 1, \cdots, N)$. One caveat of the practical calculation to obtain the three criteria is that, we mainly need the posterior samples of $\boldsymbol{\Upsilon}_{[\mathcal{M}_h]}$ drawn from the joint posterior density $\pi(\boldsymbol{\Upsilon}_{[\mathcal{M}_h]}, \boldsymbol{\alpha}_{[\mathcal{M}_h]}, \boldsymbol{B}_{[\mathcal{M}_h]}, \boldsymbol{\Omega}_{[\mathcal{M}_h]}|\mathbf{y}_{1:N})$ and the explicit form the likelihood function $\mathcal{L}(\boldsymbol{\Upsilon}_{[\mathcal{M}_h]}|\mathbf{y}_{1:N})$ from each of the models $\mathcal{M}_h$ $(h = 1, \cdots, H)$ due to our assumption of the focus.

In the next subsections, we provide some brief summaries of the criteria and then adapt them to our context. In what follows, to simplify the notation, we suppress the arguments $[\mathcal{M}_h]$ in the parameters. For a detailed explanation of the criteria, refer to [274,275].

8.2. Deviance Information Criterion

Ref. [55] suggested DIC, a generalized version of Akaike information criterion [271] for a Bayesian hierarchical model, given by

$$\text{DIC} = D(\overline{\mathbf{\Psi}}) + 2 \cdot p_{\text{D}}. \tag{38}$$

In (38), the function $D(\mathbf{\Psi}) = -2 \log \mathcal{L}(\mathbf{\Psi}|\mathbf{y}_{1:N})$ is referred to as deviance. Deviance is a goodness-of-fit statistics whose lower value indicates a better fitting [276]. Goodness-of-fit term of DIC (i.e., $D(\overline{\mathbf{\Psi}})$) is the value of deviance evaluated at the posterior mean of $\mathbf{\Psi}$, denoted by $\overline{\mathbf{\Psi}} = \mathbb{E}[\mathbf{\Psi}|\mathbf{y}_{1:N}] = \int \mathbf{\Psi}\pi(\mathbf{\Psi}|\mathbf{y}_{1:N})\mathbf{\Psi}$, where $\pi(\mathbf{\Psi}|\mathbf{y}_{1:N})$ represents the posterior distribution of $\mathbf{\Psi} = (\mathbf{\Theta}, \sigma^2)$. The effective number of parameters (i.e., penalty term for model complexity) of DIC in (38) is obtained by $p_{\text{D}} = \text{Var}[D(\mathbf{\Psi})|\mathbf{y}_{1:N}]/2 = 2 \cdot \text{Var}[\log \mathcal{L}(\mathbf{\Psi}|\mathbf{y}_{1:N})|\mathbf{y}_{1:N}]$. A model with a smaller value for DIC indicates a better predictive performance among considered models.

Some intuition behind having two competing additive terms in (38) is as follows. Typically, complex models get rewards in terms of the deviance than simple models: therefore, over-fitted models normally have a preference over under-fitted models when only the deviance is considered in model comparison, which is undesirable. By adding a penalty term for the model complexity to the deviance term, we hope that the resulting criterion produces a reasonable value based on fair comparison regardless of model complexity. Roughly speaking, this principle (i.e., a trade-off between the goodness-of-fit and penalty terms) is commonly manifested in the three criteria DIC, WAIC, and PPLC.

Going back to our examples, we can obtain the DIC corresponding to each of the H candidate models $\mathcal{M}_h$ $(h = 1, \cdots, H)$:

$$D(\mathbf{\Psi}) = -2 \sum_{i=1}^{N} \log \mathcal{N}_{M_i}(\mathbf{y}_i | \boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i), \sigma^2 \mathbf{I}),$$

where only mean function $\boldsymbol{f}_i(\mathbf{t}_i, \boldsymbol{\theta}^i)$ differs across H models $\mathcal{M}_h$, $h = 1, \cdots, H$. In practice, posterior mean $\overline{\mathbf{\Psi}}$ and effective number of parameters p_{D} are not expressed in closed-forms, hence, the DIC (38) is stochastically approximated though MCMC techniques [277].

8.3. Widely Applicable Information Criterion

Ref. [269] introduced WAIC, which is regarded as a fully Bayesian version of the DIC (38) in the sense that a goodness-of-fit term exploits the entire posterior distribution. Note that the goodness-of-fit term of the DIC (38) is obtained by plugging the posterior mean $\overline{\mathbf{\Psi}}$ into the deviance $D(\mathbf{\Psi})$, which lacks a fully Bayesian sense. It is known that WAIC is asymptotically equivalent to Bayesian cross-validation [278], and also applicable to singular models.

WAIC is defined by

$$\text{WAIC} = -2 \cdot \text{LPPD} + 2 \cdot p_{\text{W}}, \tag{39}$$

where the goodness-of-fit term is called the log posterior predictive density (LPPD), which is defined as $\text{LPPD} = \sum_{i=1}^{N} \log \mathbb{E}[\mathcal{L}(\mathbf{\Psi}^i|\mathbf{y}_i)|\mathbf{y}_{1:N}]$, and the effective number of parameter in the penalty term is defined by $p_{\text{W}} = \sum_{i=1}^{N} \text{Var}[\log \mathcal{L}(\mathbf{\Psi}^i|\mathbf{y}_i)|\mathbf{y}_{1:N}]$.

In practice, as similar to DIC (38), WAIC (39) is obtained by stochastic approximations. Given posterior samples $\{(\mathbf{\Psi})^{(s)}\}_{s=1}^{S} \sim \pi(\mathbf{\Psi}|\mathbf{y}_{1:N})$, the LPPD and p_W terms may be approximated by

$$\widehat{\text{LPPD}} = \sum_{i=1}^{N} \log \left(\frac{1}{S} \sum_{s=1}^{S} \mathcal{L}((\mathbf{\Psi}^i)^{(s)}|\mathbf{y}_i) \right), \tag{40}$$

$$\widehat{p_{\text{W}}} = \sum_{i=1}^{N} \left\{ \frac{1}{S-1} \sum_{s=1}^{S} \left(\log \mathcal{L}((\mathbf{\Psi}^i)^{(s)}|\mathbf{y}_i) - \frac{1}{S} \sum_{s=1}^{S} \log \mathcal{L}((\mathbf{\Psi}^i)^{(s)}|\mathbf{y}_i) \right)^2 \right\}. \tag{41}$$

Returning to our examples, we can approximate the value of WAIC corresponding to each of the H models as follows. First, replace $\mathcal{L}(\mathbf{\Psi}^i|\mathbf{y}_i)$ in (40) and (41) with the individual-based data distribution $p(\mathbf{y}_i|\mathbf{\Psi}^i) = \mathcal{N}_{M_i}(\mathbf{y}_i|\boldsymbol{f}_i(\mathbf{t}_i,\boldsymbol{\theta}^i),\sigma^2\mathbf{I})$ $(i = 1,\cdots,N)$, where only the mean function $\boldsymbol{f}_i(\mathbf{t}_i,\boldsymbol{\theta}^i)$ differs across the H candidate models, and second, approximate $\widehat{\text{LPPD}}$ and $\widehat{p_{\text{W}}}$ by using a MCMC method, and finally, obtain an approximation of WAIC (39) corresponding to each model.

8.4. Posterior Predictive Loss Criterion

Ref. [270] introduced PPLC as an alternative to DIC (38) or WAIC (39). A notable feature of PPLC different from DIC and WAIC is its use of replicated observations, denoted by $\mathbf{y}_i^{rep} = (y_{i1}^{rep}, y_{i2}^{rep}, \cdots, y_{iM_i}^{rep})^\top \in \mathbb{R}^{M_i}$, corresponding to the actual observations $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{iM_i})^\top \in \mathbb{R}^{M_i}$, for each $i = 1,\cdots,N$. Here, the replicate $\mathbf{y}_i^{rep}$ for the subject i is drawn from its posterior predictive density

$$f(\mathbf{y}_i^{rep}|\mathbf{y}_{1:N}) = \int p(\mathbf{y}_i^{rep}|\mathbf{\Psi}^i)\cdot\pi(\mathbf{\Psi}|\mathbf{y}_{1:N})d\mathbf{\Psi}, \quad (i = 1,\cdots,N), \tag{42}$$

where $p(\mathbf{y}_i^{rep}|\mathbf{\Psi}^i)$ is the data density for the i-th subject and $\pi(\mathbf{\Psi}|\mathbf{y}_{1:N})$ is posterior distribution. The idea of using replicates $\{\mathbf{y}_i^{rep}\}_{i=1}^N$ for a criticism of the model in light of the observed data $\{\mathbf{y}_i\}_{i=1}^N$ is also purported by [279].

A general rule of the PPLC is principled on a balanced loss function [280]. Given any loss function $l(\cdot)$ and a positive real number k, a balanced loss function is defined by

$$l(\mathbf{y}_i^{rep},\mathbf{a}_i;\mathbf{y}_{1:N}) = l(\mathbf{y}_i^{rep},\mathbf{a}_i) + k\cdot l(\mathbf{y}_i,\mathbf{a}_i), \quad k > 0,\, i = 1,\cdots,N, \tag{43}$$

where $\mathbf{a}_i$ is a non-stochastic action vector, k is a weight, and $\mathbf{y}_i^{rep}$ is a replicate for its observed counterpart $\mathbf{y}_i$. Conceptually, the role of action vector $\mathbf{a}_i$ is to accommodate both $\mathbf{y}_i$, and what we predict for $\mathbf{y}_i^{rep}$. Note that the loss function on the left-hand side of (43) penalizes actions $\mathbf{a}_i$ both for departure from the corresponding observed value (fit) as well as for departure from what we expect the replicate to be (smoothness) [274]. A generic version of PPLC is defined by $D_k = \sum_{i=1}^N \min_{\mathbf{a}_i} \mathbb{E}[l(\mathbf{y}_i^{rep},\mathbf{a}_i;\mathbf{y}_{1:N})|\mathbf{y}_{1:N}]$, where the expectation $\mathbb{E}[\cdot|\mathbf{y}_{1:N}]$ is taken with respect to the predictive density $f(\mathbf{y}_i^{rep}|\mathbf{y}_{1:N})$ (42) for some specified $k \geq 0$. Note that the resulting value D_k does not depend on the action vector $\mathbf{a}_i$ and replicates $\{\mathbf{y}_i^{rep}\}_{i=1}^N$ as they are marginalized out by the minimization and expectation, respectively, but is dependent on the constant $k > 0$.

By choosing the quadratic loss $l(\mathbf{y},\mathbf{a}) = \|\mathbf{y}-\mathbf{a}\|_2^2$ in (43), the generic PPLC D_k may be simplified as

$$D_k = \frac{k}{k+1}G + P, \quad k \geq 0, \tag{44}$$

where $G = \sum_{i=1}^N \|\boldsymbol{\nu}_i - \mathbf{y}_i\|_2^2$ and $P = \sum_{i=1}^N \varsigma_i^2$ represent the goodness-of-fit and penalty terms, respectively, with $\boldsymbol{\nu}_i = \mathbb{E}[\mathbf{y}_i^{rep}|\mathbf{y}_{1:N}]$ and $\varsigma_i^2 = \mathbb{E}[\|\mathbf{y}_i^{rep} - \boldsymbol{\nu}_i\|_2^2|\mathbf{y}_{1:N}]$, $i = 1,\cdots,N$. Eventually, a model with a smaller value for the D_k (44) is preferable. It is known that ordering of models is insensitive to the particular choice of k [274].

Finally, we adapt the PPLC (44) to our examples. Due to the definition of notation $\mathbf{\Psi}^i = (\boldsymbol{\theta}^i,\sigma^2)$ $(i = 1,\cdots,N)$, the posterior predictive distribution of $\mathbf{y}_i^{rep}$ (42) can be detailed as follows

$$f(\mathbf{y}_i^{rep}|\mathbf{y}_{1:N}) = \int \mathcal{N}_{M_i}(\mathbf{y}_i^{rep}|\boldsymbol{f}_i(\mathbf{t}_i,\boldsymbol{\theta}^i),\sigma^2\mathbf{I})\cdot\pi(\mathbf{\Psi}^i|\mathbf{y}_{1:N})d\mathbf{\Psi}^i, \quad (i = 1,\cdots,N).$$

To approximate D_k (44) for each model, first, choose a number k, saying $k = 1$, and second, approximate $\boldsymbol{\nu}_i$ and ς_i^2 through replicates $\mathbf{y}_i^{rep}$ drawn from the predictive density $f(\mathbf{y}_i^{rep}|\mathbf{y}_{1:N})$ (42) for each $i = 1,\cdots,N$, and finally, complete the G and P to get an approximation to the D_k (44).

9. Extensions and Recent Developments

9.1. Residual Error Models

In the basic version of the Bayesian nonlinear mixed effects model (2)–(4), we assume that residual errors in the individual-level model are additive to the mean function f across all subjects and times. Under this assumption, the (conditional) variance of the i-th subject's trajectory $\mathbb{V}[y_{ij}|\boldsymbol{\theta}^i]$ is constant with σ^2 over time t_{ij} $(j = 1, \cdots, M_i)$. The additive error model is the most standard assumption used in a variety of problems arising from many industrial and academic researches [28,30,31,281,282]. However, when there exist some systematic temporal trend in the volatility of individual trajectories, for instance, the variance $\mathbb{V}[y_{ij}|\boldsymbol{\theta}^i]$ seems to decrease over times t_{ij} as shown in Figures 3 and 4, the additive residual assumption may not be adequate to fully account for the reality of the data.

The list in Table 3 shows popularly used residual error models that can be used in Stage 1 (2). Some of them are deployed as options for the user to choose in industrial software such as MONOLIX [48] and NONMEM [47,101], and open source R package such as NLMIXR [18]. If we assume $\epsilon_{ij} = \varepsilon_{ij} = 0$, then all the error models leads to the same deterministic equation $y_{ij} = f(t_{ij}; \boldsymbol{\theta}^i)$. That being said, if the variances of the residuals $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \varsigma^2)$, that is, σ^2 and ς^2, are quite small, then the inference outcome based on each of the error models will be similar each other.

Table 3. List of residual error models that can be used in Stage 1 (2).

Residual Error Type	**Individual-Level Model**	**Mean $\mathbb{E}[y_{ij}\|\boldsymbol{\theta}^i]$**	**Variance $\mathbb{V}[y_{ij}\|\boldsymbol{\theta}^i]$**
Additive	$y_{ij} = f(t_{ij}; \boldsymbol{\theta}^i) + \epsilon_{ij}$	$f(t_{ij}; \boldsymbol{\theta}^i)$	σ^2
Proportional	$y_{ij} = f(t_{ij}; \boldsymbol{\theta}^i) \cdot (1 + \epsilon_{ij})$	$f(t_{ij}; \boldsymbol{\theta}^i)$	$\{f(t_{ij}; \boldsymbol{\theta}^i)\}^2 \cdot \sigma^2$
Exponential	$y_{ij} = f(t_{ij}; \boldsymbol{\theta}^i) \cdot \exp(\epsilon_{ij})$	$f(t_{ij}; \boldsymbol{\theta}^i) \cdot \exp(\sigma^2/2)$	$\{f(t_{ij}; \boldsymbol{\theta}^i)\}^2 \cdot (\exp(\sigma^2) - 1) \cdot \exp(\sigma^2)$
Additive and proportional	$y_{ij} = f(t_{ij}; \boldsymbol{\theta}^i) \cdot (1 + \epsilon_{ij}) + \varepsilon_{ij}$	$f(t_{ij}; \boldsymbol{\theta}^i)$	$\{f(t_{ij}; \boldsymbol{\theta}^i)\}^2 \cdot \sigma^2 + \varsigma^2$
Additive and exponential	$y_{ij} = f(t_{ij}; \boldsymbol{\theta}^i) \cdot \exp(\epsilon_{ij}) + \varepsilon_{ij}$	$f(t_{ij}; \boldsymbol{\theta}^i) \cdot \exp(\sigma^2/2)$	$\{f(t_{ij}; \boldsymbol{\theta}^i)\}^2 \cdot (\exp(\sigma^2) - 1) \cdot \exp(\sigma^2) + \varsigma^2$

Random errors are assumed to be distributed according to $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \varsigma^2)$ $(i = 1, \cdots, N; j = 1, \cdots, M_i)$, with independence between ϵ_{ij} and ε_{ij}.

The exponential error model (i.e., $y_{ij} = f(t_{ij}; \boldsymbol{\theta}^i) \cdot \exp(\epsilon_{ij})$) is a routine option, which is used when the ranges of the response y_{ij} and mean function f are positive real numbers, while we want to systematically describe the temporal volatility. In the practical implementation, we take the natural logarithm on both sides of the equation of the error model so that the model is converted to an additive error model in log-scale (i.e., $\log y_{ij} = \log(f(t_{ij}; \boldsymbol{\theta}^i)) + \epsilon_{ij}$). That way, relevant Bayesian computation becomes much straightforward. Ref. [31] analyzed the shale oil data shown in Figure 4 in this formulation.

9.2. Bayesian Nonparametric Methods

Recently, the use of the Bayesian nonparametric (BNP) statistical models has received increasing attention in the statistical literature because they allow modelers to gain model flexibility and robustness compared to its parametric counterpart [283,284]. BNP methods can be applied to the formulation of the basic model (2)–(4), when the parametric specification for the error distributions is too restrictive to achieve a certain purpose of the analysis, or inference results lead to poor performance due to the inappropriate parametric form. Typically, BNP methods are applied to the population-level model, by extending or relaxing the parametric assumption on the random errors η_{li}, while retaining the individual-level model as fully parametric [285]. A Gaussian process prior [286,287] or a Dirichlet process prior [288–290] is popularly used for such extension and relaxation. Mathematical concepts of the processes are explained in [287,288].

To illustrate some motivation behind the application of BNP methods, we take the shale oil production data in Figure 4 researched by [31] as an example. Their goal was to predict EUR at a new location before the actual drilling takes place. Because the geological location is not stochastically incorporated into the basic model (2)–(4), authors extended the linear regression in Stage 2 into a spatial linear regression as follows

$$\theta_{li} = \theta_l(\mathbf{s}_i) = \alpha_l + \mathbf{x}_i^\top \boldsymbol{\beta}_l + v_l(\mathbf{s}_i) + \eta_l(\mathbf{s}_i), \quad (i = 1, \cdots, N; l = 1, \cdots, K),$$

where $\eta_l(\cdot) \sim \mathcal{GP}(0, \omega_l^2 I(\cdot, \cdot))$ represents a Gaussian white noise process with the indicator function $I(\cdot, \cdot)$ with variance ω_l^2. The stochastic process $v_l(\cdot) \sim \mathcal{GP}(0, \mathcal{K}(\cdot, \cdot))$ is the newly introduced Gaussian process with a radial basis function kernel $\mathcal{K}(\mathbf{s}_{i_1}, \mathbf{s}_{i_2})) = \gamma_l^2 \exp[-\|\mathbf{s}_{i_1} - \mathbf{s}_{i_2}\|_2^2 / \{2\rho_l^2\}]$ with variance γ_l^2 and range parameter ρ_l^2, and $\mathbf{s}_i$ represents the (longitude, latitude) of the i-th shale oil well. The existence of $v_l(\cdot)$ enables spatial prediction of EUR at a new location, taking an advantage of the geological proximity information, which is called the latent kriging technique.

An another motivation on the use BNP methods is the situation where there exists multimodality in the distribution $\mathcal{P}$ of model parameter vector $\{\boldsymbol{\theta}^i\}_{i=1}^N \sim \mathcal{P}$. Note that, in the basic model (2)–(4), the distribution $\mathcal{P}$ is assumed to be a single multivariate normal distribution $\mathcal{N}_K(\boldsymbol{\alpha} + \boldsymbol{B}\mathbf{x}_i, \boldsymbol{\Omega})$. In the multimodality case, the population may consist of disparate subpopulations, and the single multivariate normal distribution of the basic model can produce a poor model performance due to the lack of flexibility. A natural generalization to accommodate such multimodality is an extension to a finite mixture of multivariate normal distributions [291], or furthermore, to a countably infinite number of mixtures of multivariate normal distributions [292]. Particularly, in the latter case, if a Dirichlet process prior [288] is placed on the mixture components, then the resulting infinite mixture models are generally called Dirichlet process mixture (DPM) model [293,294]. The DPM model is one of the most studied topics in BNP methods in recent years [284]. See [295] for a survey of the posterior computations of using DPM models.

A number of authors have studied DPM models under the basic model (2)–(4) or similar forms with their own specifications [24,282,285,296]. For example, ref. [285] placed a DPM model only for the model parameter vector $\boldsymbol{\theta}^i$ $(i = 1, \cdots, N)$, while the covariates $\mathbf{x}_i$ are incorporated into the base measure of Dirichlet process. In contrast, refs. [282,296] used a DPM model jointly for the model parameter vector and covariates, $(\boldsymbol{\theta}^i, \mathbf{x}_i)$ $(i = 1, \cdots, N)$, to induce a nonparametric regression function $\mathbb{E}[\boldsymbol{\theta}^i | \mathbf{x}_i]$. Ref. [24] used a Dirichlet process prior only for a certain component θ_{li} $(i = 1, \cdots, N)$ corresponding to a block indicator in an analysis-of-variance setup. Refer to the [297] for a review and references therein for more specifications.

9.3. Software Development

Recent years have seen the great success of Bayesian nonlinear mixed effects models, or more generally, Bayesian hierarchical models (BHM), in a variety of disciplines such as biology, medical research, physics, social, and educational sciences [30,31,141,298,299]. This was partly due to the widespread introduction of non-commercial software packages that enabled applied researchers to answer substantive research questions through applications of BHM [17,34,35,47,50,77]. Most of the Bayesian software such as JAGS [34], BUGS [35], and STAN [17] are designed to require a reasonable understanding of the MCMC sampling scheme. From the perspective of implementation, spirits of most Bayesian software are similar in that researchers only need to designate a DAG structure [300,301] of a BHM. Such a DAG structure can be abstractly represented as the collection $\{$data y, likelihood $p(y|\theta)$, prior $\pi(\theta|\eta)$, hyperprior $\pi(\eta)\}$ that should be programmed by textually or graphically, after which Bayesian software prints out simulated Markov chains from the posterior distribution $\pi(\theta, \eta | y)$. See [35,49] for an overall idea about how Bayesian software operates.

From the algorithmic perspective, the performance of Bayesian software may highly depend on two aspects: (i) whether the program has been designed to exploit a conditional independence structure arising from the hierarchy; and (ii) what sampling algorithms have been deployed to simulate Markov chains from a non-closed form distribution, possibly high-dimensional. As discussed in Section 6.3, conditional independence is inherent in the formulation of BHM, of which proper exploitation can greatly improve the computational efficiency [39,302]. This can be mostly done by constructing a Gibbs sampling algorithm with a blocking strategy into the consideration [303]. A general rule is that the convergence of the Gibbs sampler can be improved by grouping correlated latent variables as a single parameter block to sample from as a whole [304]. On the other hand, in the task of sampling from a non-closed form distribution, we know that a naive MH algorithm [180] requires the specification of proposal density, which can be problematic in developing software. Therefore, fully automated sampling algorithms such as ESS (Algorithm 1) [169], NUTS [42], and slice sampler [45] are appreciated as general-purpose inference engines in the development of Bayesian software when conjugate-update is infeasible.

Most of the Bayesian software packages, for instance, WinBUGS [49,305], OpenBUGS [306], and JAGS [34], use three family of MCMC algorithms: Gibbs [302], MH [180], and slice sampling [45]. In contrast, Stan [17] implements HMC [41,183] and its extension, NUTS [42]. Perhaps, Stan is one of the most extensively used Bayesian software packages in recent years due to the fast converge of the inference engines regardless of whether the priors are conjugate or not. By that reason, and its great modeling flexibility, Stan has been used as a basic platform for other high-level packages like brms [50] and Torsten [77].

9.4. Future Research Topics

We briefly mention two topics that have generated great recent interest in the Bayesian statistical community. The first topic we want to bring out is the use of Bayesian optimization techniques, namely variational inference [173,175] and expectation propagation [172,307], for the Bayesian nonlinear mixed effects models. These methods have received significant attention in the recent past because of their scalability to large-scale problems, enabling 'Big Bayesian Learning' [308,309]. Essentially, the main goal of these methods is to approximate the joint posterior density (16) via optimization, rather than via sampling such as MCMC sampling, which may cause scalability problems. The basic idea behind them is to first posit a family of densities and then to find a member of that family which is close to the target density, where the closeness is often measured by Kullback–Leibler divergence [310]. See Chapter 10 in [144] for a general idea for the methods. Although there were published research works for a new algorithmic development of the methods for the application to (generalized) linear mixed effect models [311–313], to our knowledge, there is no relevant published research for the application to nonlinear linear mixed effect models.

Another topic untapped in the literature is the development of the Bayesian version of mixed effects machine learning models [314–317]. This is a relatively new branch in statistics and machine learning community, where most research works were published in the past five years. The central idea of the models is to estimate the nonlinear function f in the individual-level model (2) nonparametrically by using a variety of machine learning methods, rather than specifying a parametric function, while maintaining the mixed effects modeling framework. Therefore, the modeling framework will be similar to a nonparametric regression of which the primary purpose is to estimate the unknown function f [287,318,319]. However, the main difference is that, in mixed effects machine learning models, (1) there exist some random variables to describe inter-subject variability, and (2) curve fitting mechanism (i.e., estimation of f) is mostly done by machine learning models, including deep learning [145], random forest [320], gradient boosted machine [321], etc.

10. Discussion

This review of Bayesian nonlinear mixed effects models is of necessity incomplete, as the literature is too vast to attempt even a moderate review. We have chosen to focus much of our attention on providing some of the most recent literature on the Bayesian analysis of the underlying basic model, with an emphasis on implementation of the model and introduction of recently developed prior distributions. We hope that this review can be read as a guideline to develop computational algorithms for complex and realistic Bayesian hierarchical nonlinear models. We wish that this review will offer readers familiar with frequentist analysis some pedagogical insight into the Bayesian approach, and provide those new to nonlinear mixed effects modeling a foundation of the implementation of Bayesian and frequentist computations for appreciating its idea and utility. We look forward to continuing methodological developments, software developments, and new applications of this rich class of models in industrial and academic research.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Sterba, S.K. Fitting nonlinear latent growth curve models with individually varying time points. *Struct. Equ. Model. Multidiscip. J.* **2014**, *21*, 630–647. [CrossRef]
2. McArdle, J.J. Latent variable growth within behavior genetic models. *Behav. Genet.* **1986**, *16*, 163–200. [CrossRef] [PubMed]
3. Cook, N.R.; Ware, J.H. Design and analysis methods for longitudinal research. *Annu. Rev. Public Health* **1983**, *4*, 1–23. [CrossRef] [PubMed]
4. Mehta, P.D.; West, S.G. Putting the individual back into individual growth curves. *Psychol. Methods* **2000**, *5*, 23. [CrossRef] [PubMed]
5. Zeger, S.L.; Liang, K.Y. An overview of methods for the analysis of longitudinal data. *Stat. Med.* **1992**, *11*, 1825–1839. [CrossRef] [PubMed]
6. Diggle, P.; Diggle, P.J.; Heagerty, P.; Liang, K.Y.; Zeger, S. *Analysis of Longitudinal Data*; Oxford University Press: Oxford, UK, 2002.
7. Demidenko, E. *Mixed Models: Theory and Applications with R*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
8. Snijders, T.A.; Bosker, R.J. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*; Sage: Los Angeles, CA, USA, 2011.
9. Goldstein, H. *Multilevel Statistical Models*; John Wiley & Sons: Hoboken, NJ, USA, 2011; Volume 922.
10. Raudenbush, S.W.; Bryk, A.S. *Hierarchical Linear Models: Applications and Data Analysis Methods*; Sage: Thousand Oaks, CA, USA, 2002; Volume 1.
11. Efron, B. The future of indirect evidence. *Stat. Sci. A Rev. J. Inst. Math. Stat.* **2010**, *25*, 145. [CrossRef] [PubMed]
12. Sheiner, L.B.; Rosenberg, B.; Melmon, K.L. Modelling of individual pharmacokinetics for computer-aided drug dosage. *Comput. Biomed. Res.* **1972**, *5*, 441–459. [CrossRef]
13. Lindstrom, M.J.; Bates, D.M. Nonlinear mixed effects models for repeated measures data. *Biometrics* **1990**, *46*, 673–687. [CrossRef]
14. Davidian, M.; Giltinan, D.M. Nonlinear models for repeated measurement data: An overview and update. *J. Agric. Biol. Environ. Stat.* **2003**, *8*, 387–419. [CrossRef]
15. Davidian, M.; Giltinan, D.M. *Nonlinear Models for Repeated Measurement Data*; Routledge: New York, NY, USA, 1995.
16. Beal, S. The NONMEM System. 1980. Available online: https://iconplc.com/innovation/nonmem/ (accessed on 20 February 2022).
17. Stan Development Team. *RStan: The R Interface to Stan*; R Package Version 2.21.3. 2021. Available online: https://mc-stan.org/rstan/ (accessed on 20 February 2022).
18. Fidler, M.; Wilkins, J.J.; Hooijmaijers, R.; Post, T.M.; Schoemaker, R.; Trame, M.N.; Xiong, Y.; Wang, W. Nonlinear mixed-effects model development and simulation using nlmixr and related R open-source packages. *CPT Pharmacometr. Syst. Pharmacol.* **2019**, *8*, 621–633. [CrossRef]
19. Wang, W.; Hallow, K.; James, D. A tutorial on RxODE: Simulating differential equation pharmacometric models in R. *CPT Pharmacometr. Syst. Pharmacol.* **2016**, *5*, 3–10. [CrossRef] [PubMed]
20. Stegmann, G.; Jacobucci, R.; Harring, J.R.; Grimm, K.J. Nonlinear mixed-effects modeling programs in R. *Struct. Equ. Model. Multidiscip. J.* **2018**, *25*, 160–165. [CrossRef]
21. Vonesh, E.; Chinchilli, V.M. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*; CRC Press: Boca Raton, FL, USA, 1996.
22. Lee, S.Y. *Structural Equation Modeling: A Bayesian Approach*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
23. Dellaportas, P.; Smith, A.F. Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *J. R. Stat. Soc. Ser. C* **1993**, *42*, 443–459. [CrossRef]
24. Bush, C.A.; MacEachern, S.N. A semiparametric Bayesian model for randomised block designs. *Biometrika* **1996**, *83*, 275–285. [CrossRef]
25. Zeger, S.L.; Karim, M.R. Generalized linear models with random effects; a Gibbs sampling approach. *J. Am. Stat. Assoc.* **1991**, *86*, 79–86. [CrossRef]

26. Brooks, S.P. Bayesian computation: A statistical revolution. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **2003**, *361*, 2681–2697. [CrossRef] [PubMed]
27. Bennett, J.; Wakefield, J. A comparison of a Bayesian population method with two methods as implemented in commercially available software. *J. Pharmacokinet. Biopharm.* **1996**, *24*, 403–432. [CrossRef]
28. Wakefield, J. The Bayesian analysis of population pharmacokinetic models. *J. Am. Stat. Assoc.* **1996**, *91*, 62–75. [CrossRef]
29. Gelman, A.; Bois, F.; Jiang, J. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Am. Stat. Assoc.* **1996**, *91*, 1400–1412. [CrossRef]
30. Lee, S.Y.; Lei, B.; Mallick, B. Estimation of COVID-19 spread curves integrating global data and borrowing information. *PLoS ONE* **2020**, *15*, e0236860. [CrossRef] [PubMed]
31. Lee, S.Y.; Mallick, B.K. Bayesian Hierarchical Modeling: Application Towards Production Results in the Eagle Ford Shale of South Texas. *Sankhya B* **2021**, 1–43. [CrossRef]
32. Hammersley, J. *Monte Carlo Methods*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
33. Green, P.J.; Łatuszyński, K.; Pereyra, M.; Robert, C.P. Bayesian computation: A summary of the current state, and samples backwards and forwards. *Stat. Comput.* **2015**, *25*, 835–862. [CrossRef]
34. Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria, 20–22 March 2003; Volume 124, pp. 1–10.
35. Lunn, D.; Spiegelhalter, D.; Thomas, A.; Best, N. The BUGS project: Evolution, critique and future directions. *Stat. Med.* **2009**, *28*, 3049–3067. [CrossRef] [PubMed]
36. Beal, S.L.; Sheiner, L.B. Estimating population kinetics. *Crit. Rev. Biomed. Eng.* **1982**, *8*, 195–222. [PubMed]
37. Wolfinger, R. Laplace's approximation for nonlinear mixed models. *Biometrika* **1993**, *80*, 791–795. [CrossRef]
38. Delyon, B.; Lavielle, M.; Moulines, E. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Stat.* **1999**, *27*, 94–128. [CrossRef]
39. Lee, S.Y. Gibbs sampler and coordinate ascent variational inference: A set-theoretical review. *Commun. Stat. Theory Methods* **2021** , 1–21. [CrossRef]
40. Robert, C.P.; Casella, G. The metropolis—Hastings algorithm. In *Monte Carlo Statistical Methods*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 231–283.
41. Neal, R.M. MCMC using Hamiltonian dynamics. *Handb. Markov Chain Monte Carlo* **2011**, *2*, 2.
42. Hoffman, M.D.; Gelman, A. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **2014**, *15*, 1593–1623.
43. Dwivedi, R.; Chen, Y.; Wainwright, M.J.; Yu, B. Log-concave sampling: Metropolis-Hastings algorithms are fast! In Proceedings of the Conference on Learning Theory, Stockholm, Sweden, 6–9 July 2018; pp. 793–797.
44. Ma, Y.A.; Chen, Y.; Jin, C.; Flammarion, N.; Jordan, M.I. Sampling can be faster than optimization. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 20881–20885. [CrossRef]
45. Neal, R.M. Slice sampling. *Ann. Stat.* **2003**, *31*, 705–767. [CrossRef]
46. SAS Institute. SAS OnlineDoc, Version 8. 1999. Available online: http://v8doc.sas.com/sashtml/main.htm (accessed on 20 February 2022).
47. Beal, S.L.; Sheiner, L.B.; Boeckmann, A.; Bauer, R.J. *NONMEM Users Guides*; NONMEM Project Group, University of California: San Francisco, CA, USA, 1992.
48. Lavielle, M. *Monolix User Guide Manual*. 2005. Available online: https://monolix.lixoft.com/ (accessed on 20 February 2022).
49. Lunn, D.J.; Thomas, A.; Best, N.; Spiegelhalter, D. WinBUGS-a Bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput.* **2000**, *10*, 325–337. [CrossRef]
50. Bürkner, P.C. brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **2017**, *80*, 1–28. [CrossRef]
51. Chernoff, H. Large-sample theory: Parametric case. *Ann. Math. Stat.* **1956**, *27*, 1–22. [CrossRef]
52. Wand, M. Fisher information for generalised linear mixed models. *J. Multivar. Anal.* **2007**, *98*, 1412–1416. [CrossRef]
53. Kang, D.; Bae, K.S.; Houk, B.E.; Savic, R.M.; Karlsson, M.O. Standard error of empirical bayes estimate in NONMEM® VI. *Korean J. Physiol. Pharmacol.* **2012**, *16*, 97–106. [CrossRef] [PubMed]
54. Breslow, N.E.; Clayton, D.G. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **1993**, *88*, 9–25.
55. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Chapman and Hall/CRC: London, UK, 2004.
56. Smid, S.C.; McNeish, D.; Miočević, M.; van de Schoot, R. Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Struct. Equ. Model. Multidiscip. J.* **2020**, *27*, 131–161. [CrossRef]
57. Rupp, A.A.; Dey, D.K.; Zumbo, B.D. To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Struct. Equ. Model.* **2004**, *11*, 424–451. [CrossRef]
58. Bonangelino, P.; Irony, T.; Liang, S.; Li, X.; Mukhi, V.; Ruan, S.; Xu, Y.; Yang, X.; Wang, C. Bayesian approaches in medical device clinical trials: A discussion with examples in the regulatory setting. *J. Biopharm. Stat.* **2011**, *21*, 938–953. [CrossRef] [PubMed]
59. Campbell, G. Bayesian methods in clinical trials with applications to medical devices. *Commun. Stat. Appl. Methods* **2017**, *24*, 561–581. [CrossRef]
60. Hoff, P.D. *A First Course in Bayesian Statistical Methods*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 580.
61. O'Hagan, A. Bayesian statistics: Principles and benefits. *Frontis* **2004**, *3*, 31–45.

62. van de Schoot, R.; Depaoli, S.; King, R.; Kramer, B.; Märtens, K.; Tadesse, M.G.; Vannucci, M.; Gelman, A.; Veen, D.; Willemsen, J.; et al. Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* **2021**, *1*, 1–26. [CrossRef]
63. Blaxter, L.; Hughes, C.; Tight, M. *How to Research*; McGraw-Hill Education: New York, NY, USA, 2010.
64. Neuman, W.L. *Understanding Research*; Pearson: New York, NY, USA, 2016.
65. Pinheiro, J.; Bates, D. *Mixed-Effects Models in S and S-PLUS*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
66. Gelman, A.; Simpson, D.; Betancourt, M. The prior can often only be understood in the context of the likelihood. *Entropy* **2017**, *19*, 555. [CrossRef]
67. Garthwaite, P.H.; Kadane, J.B.; O'Hagan, A. Statistical methods for eliciting probability distributions. *J. Am. Stat. Assoc.* **2005**, *100*, 680–701. [CrossRef]
68. O'Hagan, A.; Buck, C.E.; Daneshkhah, A.; Eiser, J.R.; Garthwaite, P.H.; Jenkinson, D.J.; Oakley, J.E.; Rakow, T. *Uncertain Judgements: Eliciting Experts' Probabilities*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2006.
69. Howard, G.S.; Maxwell, S.E.; Fleming, K.J. The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychol. Methods* **2000**, *5*, 315. [CrossRef]
70. Levy, R. Bayesian data-model fit assessment for structural equation modeling. *Struct. Equ. Model. Multidiscip. J.* **2011**, *18*, 663–685. [CrossRef]
71. Wang, L.; Cao, J.; Ramsay, J.O.; Burger, D.; Laporte, C.; Rockstroh, J.K. Estimating mixed-effects differential equation models. *Stat. Comput.* **2014**, *24*, 111–121. [CrossRef]
72. Botha, I.; Kohn, R.; Drovandi, C. Particle methods for stochastic differential equation mixed effects models. *Bayesian Anal.* **2021**, *16*, 575–609. [CrossRef]
73. Fucik, S.; Kufner, A. *Nonlinear Differential Equations*; Elsevier: Amsterdam, The Netherlands, 2014.
74. Verhulst, F. *Nonlinear Differential Equations and Dynamical Systems*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
75. Cohen, S.D.; Hindmarsh, A.C.; Dubois, P.F. CVODE, a stiff/nonstiff ODE solver in C. *Comput. Phys.* **1996**, *10*, 138–143. [CrossRef]
76. Dormand, J.R.; Prince, P.J. A family of embedded Runge-Kutta formulae. *J. Comput. Appl. Math.* **1980**, *6*, 19–26. [CrossRef]
77. Margossian, C.; Gillespie, B. Torsten: A Prototype Model Library for Bayesian PKPD Modeling in Stan User Manual: Version 0.81. Available online: https://metrumresearchgroup.github.io/Torsten/ (accessed on 20 February 2022).
78. Chipman, H.; George, E.I.; McCulloch, R.E.; Clyde, M.; Foster, D.P.; Stine, R.A. The practical implementation of Bayesian model selection. *Lect. Notes-Monogr. Ser.* **2001**, 38, 65–134.
79. Gibaldi, M.; Perrier, D. *Pharmacokinetics*; M. Dekker: New York, NY, USA, 1982; Volume 15.
80. Jambhekar, S.S.; Breen, P.J. *Basic Pharmacokinetics*; Pharmaceutical Press: London, UK, 2009; Volume 76.
81. Sheiner, L.; Ludden, T. Population pharmacokinetics/dynamics. *Annu. Rev. Pharmacol. Toxicol.* **1992**, *32*, 185–209. [CrossRef] [PubMed]
82. Ette, E.I.; Williams, P.J. Population pharmacokinetics I: Background, concepts, and models. *Ann. Pharmacother.* **2004**, *38*, 1702–1706. [CrossRef]
83. Lewis, J.; Beal, C.H. Some New Methods for Estimating the Future Production of Oil Wells. *Trans. AIME* **1918**, *59*, 492–525. [CrossRef]
84. Fetkovich, M.J. Decline curve analysis using type curves. *J. Pet. Technol.* **1980**, *32*, 1065–1077. [CrossRef]
85. Harris, S.; Lee, W.J. A Study of Decline Curve Analysis in the Elm Coulee Field. In *SPE Unconventional Resources Conference*; Society of Petroleum Engineers: The Woodlands, TX, USA, 2014.
86. Nelson, C.R.; Siegel, A.F. Parsimonious modeling of yield curves. *J. Bus.* **1987**, 60, 473–489. [CrossRef]
87. Diebold, F.X.; Li, C. Forecasting the term structure of government bond yields. *J. Econom.* **2006**, *130*, 337–364. [CrossRef]
88. Svensson, L.E. *Estimating and Interpreting forward Interest Rates: Sweden 1992–1994*; National Bureau of Economic Research Working Paper, no 4871. 1994. Available online: https://www.nber.org/papers/w4871 (accessed on 20 February 2022).
89. Dahlquist, M.; Svensson, L.E. Estimating the term structure of interest rates for monetary policy analysis. *Scand. J. Econ.* **1996**, *98*, 163–183. [CrossRef]
90. Wang, P.; Zheng, X.; Li, J.; Zhu, B. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos Solitons Fractals* **2020**, *139*, 110058. [CrossRef]
91. Wilke, C.O.; Bergstrom, C.T. Predicting an epidemic trajectory is difficult. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 28549–28551. [CrossRef]
92. Bonate, P.L. *Pharmacokinetic-Pharmacodynamic Modeling and Simulation*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 20.
93. Rowland, M.; Tozer, T.N. *Clinical Pharmacokinetics/Pharmacodynamics*; Lippincott Williams and Wilkins Philadelphia: New York, NY, USA, 2005.
94. Gabrielsson, J.; Weiner, D. *Pharmacokinetic and Pharmacodynamic Data Analysis: Concepts and Applications*; CRC Press: Boca Raton, FL, USA, 2001.
95. Dua, P.; Hawkins, E.; Van Der Graaf, P. A tutorial on target-mediated drug disposition (TMDD) models. *CPT Pharmacometr. Syst. Pharmacol.* **2015**, *4*, 324–337. [CrossRef]

96. Xu, X.S.; Yuan, M.; Zhu, H.; Yang, Y.; Wang, H.; Zhou, H.; Xu, J.; Zhang, L.; Pinheiro, J. Full covariate modelling approach in population pharmacokinetics: Understanding the underlying hypothesis tests and implications of multiplicity. *Br. J. Clin. Pharmacol.* **2018**, *84*, 1525–1534. [CrossRef]
97. Roses, A.D. Pharmacogenetics and the practice of medicine. *Nature* **2000**, *405*, 857–865. [CrossRef]
98. Food and Drug Administration. *Population Pharmacokinetics Guidance for Industry*; FDA Guidance Page. 1999. Available online: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/population-pharmacokinetics (accessed on 20 February 2022).
99. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*; SIAM: Philadelphia, PA, USA, 1982.
100. Broeker, A.; Wicha, S.G. Assessing parameter uncertainty in small-n pharmacometric analyses: Value of the log-likelihood profiling-based sampling importance resampling (LLP-SIR) technique. *J. Pharmacokinet. Pharmacodyn.* **2020**, *47*, 219–228. [CrossRef]
101. Bauer, R.J. NONMEM tutorial part I: Description of commands and options, with simple examples of population analysis. *CPT Pharmacometr. Syst. Pharmacol.* **2019**, *8*, 525–537. [CrossRef]
102. Giger, F.; Reiss, L.; Jourdan, A. The reservoir engineering aspects of horizontal drilling. In *SPE Annual Technical Conference and Exhibition*; OnePetro: Houston, TX, USA, 1984.
103. Al-Haddad, S.; Crafton, J. Productivity of horizontal wells. In *Low Permeability Reservoirs Symposium*; OnePetro: Denver, CO, USA, 1991.
104. Mukherjee, H.; Economides, M.J. A parametric comparison of horizontal and vertical well performance. *SPE Form. Eval.* **1991**, *6*, 209–216. [CrossRef]
105. Joshi, S. Cost/benefits of horizontal wells. In *SPE Western Regional/AAPG Pacific Section Joint Meeting*; OnePetro: Long Beach, CA, USA, 2003.
106. Valdes, A.; McVay, D.A.; Noynaert, S.F. Uncertainty quantification improves well construction cost estimation in unconventional reservoirs. In *SPE Unconventional Resources Conference Canada*; OnePetro: Calgary, AB, Canada, 2013.
107. Bellarby, J. *Well Completion Design*; Elsevier: Amsterdam, The Netherlands, 2009.
108. Currie, S.M.; Ilk, D.; Blasingame, T.A. Continuous estimation of ultimate recovery. In *SPE Unconventional Gas Conference*; OnePetro: Pittsburgh, PA, USA, 2010.
109. Arps, J.J. Analysis of decline curves. *Trans. AIME* **1945**, *160*, 228–247. [CrossRef]
110. Weibull, W. A statistical distribution function of wide applicability. *J. Appl. Mech.* **1951**, *18*, 293–297. [CrossRef]
111. Ilk, D.; Rushing, J.A.; Perego, A.D.; Blasingame, T.A. Exponential vs. hyperbolic decline in tight gas sands: Understanding the origin and implications for reserve estimates using Arps' decline curves. In *SPE Annual Technical Conference and Exhibition*; Society of Petroleum Engineers: Denver, CO, USA, 2008.
112. Valkó, P.P.; Lee, W.J. A better way to forecast production from unconventional gas wells. In *SPE Annual Technical Conference and Exhibition*; Society of Petroleum Engineers: Florence, Italy, 2010.
113. Clark, A.J. Decline Curve Analysis in Unconventional Resource Plays Using Logistic Growth Models. Ph.D. Thesis, The University of Texas Austion, Austin, TX, USA, 2011.
114. Duong, A.N. Rate-decline analysis for fracture-dominated shale reservoirs. *SPE Reserv. Eval. Eng.* **2011**, *14*, 377–387. [CrossRef]
115. Ali, T.A.; Sheng, J.J. Production Decline Models: A Comparison Study. In *SPE Eastern Regional Meeting*; Society of Petroleum Engineers: Morgantown, WV, USA, 2015.
116. Miao, Y.; Li, X.; Lee, J.; Zhao, C.; Zhou, Y.; Li, H.; Chang, Y.; Lin, W.; Xiao, Z.; Wu, N.; et al. Comparison of Various Rate-Decline Analysis Models for Horizontal Wells with Multiple Fractures in Shale gas Reservoirs. In *SPE Trinidad and Tobago Section Energy Resources Conference*; Society of Petroleum Engineers: Port of Spain, Trinidad and Tobago, 2018.
117. Duffee, G. Forecasting interest rates. In *Handbook of Economic Forecasting*; Elsevier: Amsterdam, The Netherlands, 2013; Volume 2, pp. 385–426.
118. Gürkaynak, R.S.; Sack, B.; Wright, J.H. The US Treasury yield curve: 1961 to the present. *J. Monet. Econ.* **2007**, *54*, 2291–2304. [CrossRef]
119. Zaloom, C. How to read the future: The yield curve, affect, and financial prediction. *Public Cult.* **2009**, *21*, 245–268. [CrossRef]
120. Hays, S.; Shen, H.; Huang, J.Z. Functional dynamic factor models with application to yield curve forecasting. *Ann. Appl. Stat.* **2012**, *6*, 870–894. [CrossRef]
121. Chen, Y.; Niu, L. Adaptive dynamic Nelson–Siegel term structure model with applications. *J. Econom.* **2014**, *180*, 98–115. [CrossRef]
122. Bank for International Settlements. *Zero-Coupon Yield Curves: Technical Documentation*; BIS Papers, no 2; Bank for International Settlements: Basel, Switzerland, 2005. Available online: https://www.bis.org/publ/bppdf/bispap25.htm (accessed on 20 February 2022).
123. Hautsch, N.; Yang, F. Bayesian inference in a stochastic volatility Nelson–Siegel model. *Comput. Stat. Data Anal.* **2012**, *56*, 3774–3792. [CrossRef]
124. Diebold, F.X.; Li, C.; Yue, V.Z. Global yield curve dynamics and interactions: A dynamic Nelson–Siegel approach. *J. Econom.* **2008**, *146*, 351–363. [CrossRef]
125. Cruz-Marcelo, A.; Ensor, K.B.; Rosner, G.L. Estimating the term structure with a semiparametric Bayesian hierarchical model: An application to corporate bonds. *J. Am. Stat. Assoc.* **2011**, *106*, 387–395. [CrossRef]

126. Richards, F. A flexible growth function for empirical use. *J. Exp. Bot.* **1959**, *10*, 290–301. [CrossRef]
127. Nelder, J.A. 182. note: An alternative form of a generalized logistic equation. *Biometrics* **1962**, *18*, 614–616. [CrossRef]
128. Seber, G.A.; Wild, C.J. *Nonlinear Regression*; John Wiley Sons: Hoboken, NJ, USA, 2003; Volume 62, p. 63.
129. Anton, H.; Herr, A. *Calculus with Analytic Geometry*; Wiley: New York, NY, USA, 1988.
130. Causton, D. A computer program for fitting the Richards function. *Biometrics* **1969**, *25*, 401–409. [CrossRef]
131. Birch, C.P. A new generalized logistic sigmoid growth equation compared with the Richards growth equation. *Ann. Bot.* **1999**, *83*, 713–723. [CrossRef]
132. Kahm, M.; Hasenbrink, G.; Lichtenberg-Fraté, H.; Ludwig, J.; Kschischo, M. grofit: Fitting biological growth curves with R. *J. Stat. Softw.* **2010**, *33*, 1–21. [CrossRef]
133. Cao, L.; Shi, P.J.; Li, L.; Chen, G. A New Flexible Sigmoidal Growth Model. *Symmetry* **2019**, *11*, 204. [CrossRef]
134. Wang, X.S.; Wu, J.; Yang, Y. Richards model revisited: Validation by and application to infection dynamics. *J. Theor. Biol.* **2012**, *313*, 12–19. [CrossRef] [PubMed]
135. Hsieh, Y.H.; Lee, J.Y.; Chang, H.L. SARS epidemiology modeling. *Emerg. Infect. Dis.* **2004**, *10*, 1165. [CrossRef] [PubMed]
136. Hsieh, Y.H. Richards model: A simple procedure for real-time prediction of outbreak severity. In *Modeling and Dynamics of Infectious Diseases*; World Scientific: London, UK, 2009; pp. 216–236.
137. Hsieh, Y.H.; Ma, S. Intervention measures, turning point, and reproduction number for dengue, Singapore, 2005. *Am. J. Trop. Med. Hyg.* **2009**, *80*, 66–71. [CrossRef] [PubMed]
138. Hsieh, Y.H.; Chen, C. Turning points, reproduction number, and impact of climatological events for multi-wave dengue outbreaks. *Trop. Med. Int. Health* **2009**, *14*, 628–638. [CrossRef]
139. Hsieh, Y.H. Pandemic influenza A (H1N1) during winter influenza season in the southern hemisphere. *Influenza Other Respir. Viruses* **2010**, *4*, 187–197. [CrossRef]
140. Wu, K.; Darcet, D.; Wang, Q.; Sornette, D. Generalized logistic growth modeling of the COVID-19 outbreak in 29 provinces in China and in the rest of the world. *arXiv* **2020**, arXiv:2003.05681.
141. Lee, S.Y.; Munafo, A.; Girard, P.; Goteti, K. Optimization of dose selection using multiple surrogates of toxicity as a continuous variable in phase I cancer trial. *Contemp. Clin. Trials* **2021**, *113*, 106657. [CrossRef]
142. Dugel, P.U.; Koh, A.; Ogura, Y.; Jaffe, G.J.; Schmidt-Erfurth, U.; Brown, D.M.; Gomes, A.V.; Warburton, J.; Weichselberger, A.; Holz, F.G.; et al. HAWK and HARRIER: Phase 3, multicenter, randomized, double-masked trials of brolucizumab for neovascular age-related macular degeneration. *Ophthalmology* **2020**, *127*, 72–84. [CrossRef]
143. Willyard, C. New human gene tally reignites debate. *Nature* **2018**, *558*, 354–356. [CrossRef] [PubMed]
144. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
145. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: London, UK, 2016.
146. Boyd, S.; Boyd, S.P.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
147. James, W.; Stein, C. Estimation with quadratic loss. In *Breakthroughs in Statistics*; Springer: Berlin/Heidelberg, Germany, 1992; pp. 443–460.
148. Dawid, A.P. Conditional independence in statistical theory. *J. R. Stat. Soc. Ser. B* **1979**, *41*, 1–15. [CrossRef]
149. Liu, Q.; Pierce, D.A. A note on Gauss—Hermite quadrature. *Biometrika* **1994**, *81*, 624–629.
150. Hedeker, D.; Gibbons, R.D. A random-effects ordinal regression model for multilevel analysis. *Biometrics* **1994**, *50*, 933–944. [CrossRef]
151. Vonesh, E.F.; Wang, H.; Nie, L.; Majumdar, D. Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *J. Am. Stat. Assoc.* **2002**, *97*, 271–283. [CrossRef]
152. Hinrichs, A.; Novak, E.; Ullrich, M.; Woźniakowski, H. The curse of dimensionality for numerical integration of smooth functions II. *J. Complex.* **2014**, *30*, 117–143. [CrossRef]
153. Vonesh, E.F.; Carter, R.L. Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics* **1992**, *48*, 1–17. [CrossRef]
154. Goldstein, H. Nonlinear multilevel models, with an application to discrete response data. *Biometrika* **1991**, *78*, 45–51. [CrossRef]
155. Vonesh, E.F. A note on the use of Laplaces approximation for nonlinear mixed-effects models. *Biometrika* **1996**, *83*, 447–452. [CrossRef]
156. Marsden, J.E.; Hoffman, M.J. *Elementary Classical Analysis*; Macmillan: New York, NY, USA, 1993.
157. Lindley, D.V.; Smith, A.F. Bayes estimates for the linear model. *J. R. Stat. Soc. Ser. B* **1972**, *34*, 1–18. [CrossRef]
158. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–22.
159. Meng, X.L.; Rubin, D.B. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Am. Stat. Assoc.* **1991**, *86*, 899–909. [CrossRef]
160. Walker, S. An EM algorithm for nonlinear random effects models. *Biometrics* **1996**, *52*, 934–944. [CrossRef]
161. Allassonnière, S.; Chevallier, J. A new class of stochastic EM algorithms. Escaping local maxima and handling intractable sampling. *Comput. Stat. Data Anal.* **2021**, *159*, 107159. [CrossRef]
162. Kuhn, E.; Lavielle, M. Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Stat. Data Anal.* **2005**, *49*, 1020–1038. [CrossRef]

163. Samson, A.; Lavielle, M.; Mentré, F. The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model. *Stat. Med.* **2007**, *26*, 4860–4875. [CrossRef]
164. Kuhn, E.; Lavielle, M. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM Probab. Stat.* **2004**, *8*, 115–131. [CrossRef]
165. Allassonnière, S.; Kuhn, E.; Trouvé, A. Construction of Bayesian deformable models via a stochastic approximation algorithm: A convergence study. *Bernoulli* **2010**, *16*, 641–678. [CrossRef]
166. Bernardo, J.M.; Smith, A.F. *Bayesian Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 405.
167. Lindley, D.V. *Bayesian Statistics, a Review*; SIAM: Philadelphia, PA, USA, 1972; Volume 2.
168. Casella, G.; George, E.I. Explaining the Gibbs sampler. *Am. Stat.* **1992**, *46*, 167–174.
169. Murray, I.; Prescott Adams, R.; MacKay, D.J. Elliptical Slice Sampling. In Proceedings of the Thirteenth International Conference on Artificial Intelligence And Statistics, Sardinia, Italy, 13–15 May 2010.
170. Ranganath, R.; Gerrish, S.; Blei, D. Black box variational inference. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, Reykjavik, Iceland, 22–25 April 2014.
171. Wang, C.; Blei, D.M. Variational inference in nonconjugate models. *J. Mach. Learn. Res.* **2013**, *14*, 1005–1031.
172. Minka, T.P. Expectation propagation for approximate Bayesian inference. *arXiv* **2013**, arXiv:1301.2294.
173. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
174. Andrieu, C.; De Freitas, N.; Doucet, A.; Jordan, M.I. An introduction to MCMC for machine learning. *Mach. Learn.* **2003**, *50*, 5–43. [CrossRef]
175. Zhang, C.; Bütepage, J.; Kjellström, H.; Mandt, S. Advances in variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2008–2026. [CrossRef] [PubMed]
176. Team, R.C. *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013. Available online: http://www.R-project.org (accessed on 20 February 2022).
177. Lee, A.; Yau, C.; Giles, M.B.; Doucet, A.; Holmes, C.C. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *J. Comput. Graph. Stat.* **2010**, *19*, 769–789. [CrossRef] [PubMed]
178. Suchard, M.A.; Wang, Q.; Chan, C.; Frelinger, J.; Cron, A.; West, M. Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. *J. Comput. Graph. Stat.* **2010**, *19*, 419–438. [CrossRef] [PubMed]
179. Hastings, W.K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57*, 97–109. [CrossRef]
180. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092. [CrossRef]
181. Robert, C.P. The Metropolis–Hastings Algorithm. In *Wiley StatsRef: Statistics Reference Online*; John Wiley and Sons, Ltd.: Hoboken, NJ, USA, 2015; pp. 1–15.
182. Chib, S.; Greenberg, E. Understanding the metropolis-hastings algorithm. *Am. Stat.* **1995**, *49*, 327–335.
183. Duane, S.; Kennedy, A.D.; Pendleton, B.J.; Roweth, D. Hybrid monte carlo. *Phys. Lett. B* **1987**, *195*, 216–222. [CrossRef]
184. Mengersen, K.L.; Tweedie, R.L. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Stat.* **1996**, *24*, 101–121. [CrossRef]
185. Chen, T.; Fox, E.; Guestrin, C. Stochastic gradient hamiltonian monte carlo. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1683–1691.
186. Aicher, C.; Ma, Y.A.; Foti, N.J.; Fox, E.B. Stochastic gradient mcmc for state space models. *SIAM J. Math. Data Sci.* **2019**, *1*, 555–587. [CrossRef]
187. Griewank, A.; Walther, A. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*; SIAM: Philadelphia, PA, USA, 2008.
188. Øksendal, B. Stochastic differential equations. In *Stochastic Differential Equations*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 65–84.
189. Uhlenbeck, G.E.; Ornstein, L.S. On the theory of the Brownian motion. *Phys. Rev.* **1930**, *36*, 823. [CrossRef]
190. Roberts, G.O.; Tweedie, R.L. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **1996**, *83*, 95–110. [CrossRef]
191. Asai, Y.; Kloeden, P.E. Numerical schemes for random ODEs via stochastic differential equations. *Commun. Appl. Anal.* **2013**, *17*, 521–528.
192. Casella, G.; Robert, C.P. *Monte Carlo Statistical Methods*; Springer: New York, NY, USA, 1999.
193. Carpenter, B.; Gelman, A.; Hoffman, M.D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; Riddell, A. Stan: A probabilistic programming language. *J. Stat. Softw.* **2017**, *76*, 1–32. [CrossRef]
194. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
195. Leimkuhler, B.; Reich, S. *Simulating Hamiltonian Dynamics*; Number 14; Cambridge University Press: Cambridge, UK, 2004.
196. Zou, D.; Gu, Q. On the convergence of Hamiltonian Monte Carlo with stochastic gradients. In Proceedings of the International Conference on Machine Learning (PMLR), Virtual, 18–24 July 2021; pp. 13012–13022.

197. Meza, C.; Osorio, F.; De la Cruz, R. Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Stat. Comput.* **2012**, *22*, 121–139. [CrossRef]
198. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A* **1946**, *186*, 453–461.
199. Makalic, E.; Schmidt, D.F. A simple sampler for the horseshoe estimator. *IEEE Signal Process. Lett.* **2016**, *23*, 179–182. [CrossRef]
200. Castillo, I.; Schmidt-Hieber, J.; Van der Vaart, A. Bayesian linear regression with sparse priors. *Ann. Stat.* **2015**, *43*, 1986–2018. [CrossRef]
201. Lee, S.Y.; Pati, D.; Mallick, B.K. Tail-adaptive Bayesian shrinkage. *arXiv* **2020**, arXiv:2007.02192.
202. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
203. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [CrossRef]
204. Fan, J.; Samworth, R.; Wu, Y. Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **2009**, *10*, 2013–2038.
205. Lu, Y.; Stuart, A.; Weber, H. Gaussian Approximations for Probability Measures on R^d. *SIAM/ASA J. Uncertain. Quantif.* **2017**, *5*, 1136–1165. [CrossRef]
206. Wang, Y.; Blei, D.M. Frequentist consistency of variational Bayes. *J. Am. Stat. Assoc.* **2019**, *114*, 1147–1161. [CrossRef]
207. Johnstone, I.M. High dimensional Bernstein-von Mises: Simple examples. *Inst. Math. Stat. Collect.* **2010**, *6*, 87.
208. Le Cam, L.; LeCam, L.M.; Yang, G.L. *Asymptotics in Statistics: Some Basic Concepts*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2000.
209. Davidian, M.; Gallant, A.R. Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine. *J. Pharmacokinet. Biopharm.* **1992**, *20*, 529–556. [CrossRef]
210. Wei, Y.; Higgins, J.P. Bayesian multivariate meta-analysis with multiple outcomes. *Stat. Med.* **2013**, *32*, 2911–2934. [CrossRef]
211. Zellner, A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*; Elsevier Science Publishers, Inc.: New York, NY, USA, 1986; pp. 233–243.
212. Pirmohamed, M. Pharmacogenetics and pharmacogenomics. *Br. J. Clin. Pharmacol.* **2001**, *52*, 345. [CrossRef] [PubMed]
213. Weinshilboum, R.M.; Wang, L. Pharmacogenetics and pharmacogenomics: Development, science, and translation. *Annu. Rev. Genom. Hum. Genet.* **2006**, *7*, 223–245. [CrossRef] [PubMed]
214. Arab-Alameddine, M.; Di Iulio, J.; Buclin, T.; Rotger, M.; Lubomirov, R.; Cavassini, M.; Fayet, A.; Décosterd, L.; Eap, C.B.; Biollaz, J.; et al. Pharmacogenetics-based population pharmacokinetic analysis of efavirenz in HIV-1-infected individuals. *Clin. Pharmacol. Ther.* **2009**, *85*, 485–494. [CrossRef] [PubMed]
215. Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*; Chapman and Hall/CRC: New York, NY, USA, 2015.
216. Mitchell, T.J.; Beauchamp, J.J. Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.* **1988**, *83*, 1023–1032. [CrossRef]
217. George, E.I.; McCulloch, R.E. Stochastic search variable selection. *Markov Chain Monte Carlo Pract.* **1995**, *68*, 203–214.
218. Johnson, V.E.; Rossell, D. On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B* **2010**, *72*, 143–170. [CrossRef]
219. Yang, Y.; Wainwright, M.J.; Jordan, M.I. On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Stat.* **2016**, *44*, 2497–2532. [CrossRef]
220. Castillo, I.; van der Vaart, A. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Stat.* **2012**, *40*, 2069–2101. [CrossRef]
221. Park, T.; Casella, G. The bayesian lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686. [CrossRef]
222. Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
223. Griffin, J.E.; Brown, P.J. Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **2010**, *5*, 171–188.
224. Carvalho, C.M.; Polson, N.G.; Scott, J.G. The horseshoe estimator for sparse signals. *Biometrika* **2010**, *97*, 465–480. [CrossRef]
225. Carvalho, C.M.; Polson, N.G.; Scott, J.G. Handling sparsity via the horseshoe. In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, Clearwater Beach, FL, USA, 16–18 April 2009.
226. Polson, N.G.; Scott, J.G. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Stat.* **2010**, *9*, 105.
227. George, E.I.; McCulloch, R.E. Approaches for Bayesian variable selection. *Stat. Sin.* **1997**, *7*, 339–373.
228. Johnstone, I.M.; Silverman, B.W. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Stat.* **2004**, *32*, 1594–1649. [CrossRef]
229. Pati, D.; Bhattacharya, A.; Pillai, N.S.; Dunson, D. Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Ann. Stat.* **2014**, *42*, 1102–1130. [CrossRef]
230. Song, Q.; Liang, F. Nearly optimal Bayesian shrinkage for high dimensional regression. *arXiv* **2017**, arXiv:1712.08964.
231. Martin, R.; Mess, R.; Walker, S.G. Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **2017**, *23*, 1822–1847. [CrossRef]
232. Bai, R.; Ghosh, M. High-dimensional multivariate posterior consistency under global–local shrinkage priors. *J. Multivar. Anal.* **2018**, *167*, 157–170. [CrossRef]
233. Zhang, R.; Ghosh, M. Ultra High-dimensional Multivariate Posterior Contraction Rate Under Shrinkage Priors. *arXiv* **2019**, arXiv:1904.04417.
234. Lee, S.; Kim, J.H. Exponentiated generalized Pareto distribution: Properties and applications towards extreme value theory. *Commun. Stat.-Theory Methods* **2019**, *48*, 2014–2038. [CrossRef]

235. Armagan, A.; Dunson, D.B.; Lee, J. Generalized double Pareto shrinkage. *Stat. Sin.* **2013**, *23*, 119. [CrossRef] [PubMed]
236. O'Hara, R.B.; Sillanpää, M.J. A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal.* **2009**, *4*, 85–117. [CrossRef]
237. Bhadra, A.; Datta, J.; Polson, N.G.; Willard, B. Lasso meets horseshoe: A survey. *Stat. Sci.* **2019**, *34*, 405–427. [CrossRef]
238. Gelman, A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **2006**, *1*, 515–534. [CrossRef]
239. Fan, J.; Liao, Y.; Liu, H. An overview of the estimation of large covariance and precision matrices. *Econom. J.* **2016**, *19*, C1–C32. [CrossRef]
240. Bickel, P.J.; Levina, E. Covariance regularization by thresholding. *Ann. Stat.* **2008**, *36*, 2577–2604. [CrossRef]
241. Lam, C.; Fan, J. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat.* **2009**, *37*, 4254. [CrossRef]
242. El Karoui, N. High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: Risk underestimation. *Ann. Stat.* **2010**, *38*, 3487–3566. [CrossRef]
243. Stein, C. Estimation of a covariance matrix, Rietz Lecture. In Proceedings of the 39th Annual Meeting IMS, Atlanta, GA, USA, 1975.
244. Pourahmadi, M. *High-Dimensional Covariance Estimation: With High-Dimensional Data*; John Wiley & Sons: Hoboken, NJ, USA, 2013; Volume 882.
245. Ledoit, O.; Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* **2004**, *88*, 365–411. [CrossRef]
246. Rajaratnam, B.; Massam, H.; Carvalho, C.M. Flexible covariance estimation in graphical Gaussian models. *Ann. Stat.* **2008**, *36*, 2818–2849. [CrossRef]
247. Won, J.H.; Lim, J.; Kim, S.J.; Rajaratnam, B. Condition-number-regularized covariance estimation. *J. R. Stat. Soc. Ser. B* **2013**, *75*, 427–450. [CrossRef]
248. Liu, C. Bartlett' s Decomposition of the Posterior Distribution of the Covariance for Normal Monotone Ignorable Missing Data. *J. Multivar. Anal.* **1993**, *46*, 198–206. [CrossRef]
249. Barnard, J.; McCulloch, R.; Meng, X.L. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Stat. Sin.* **2000**, *10*, 1281–1311.
250. Geisser, S. Bayesian estimation in multivariate analysis. *Ann. Math. Stat.* **1965**, *36*, 150–159. [CrossRef]
251. Lin, S.P.; Perlman, M.D. A Monte Carlo comparison of four estimators for a covariance matrix. In *Multivariate Analysis VI*; Krishnaiah, P.R., Ed.; North-Holland: Amsterdam, The Netherlands, 1985; pp. 411–429.
252. Brown, P.J.; Le, N.D.; Zidek, J.V. Inference for a Covariance Matrix. In *Aspects of Uncertainty*; Freeman, P.R., Smith, A.F.M., Eds.; John Wiley: Chichester, UK, 1994; pp. 77–90.
253. Jeffreys, H. *The Theory of Probability*; OUP Oxford: Oxford, UK, 1998.
254. Geisser, S.; Cornfield, J. Posterior distributions for multivariate normal parameters. *J. R. Stat. Soc. Ser. B* **1963**, *25*, 368–376. [CrossRef]
255. Villegas, C. On the a priori distribution of the covariance matrix. *Ann. Math. Stat.* **1969**, *40*, 1098–1099. [CrossRef]
256. Schervish, M.J. *Theory of Statistics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
257. James, A.T. Distributions of matrix variates and latent roots derived from normal samples. *Ann. Math. Stat.* **1964**, *35*, 475–501. [CrossRef]
258. Yang, R.; Berger, J.O. Estimation of a covariance matrix using the reference prior. *Ann. Stat.* **1994**, *22*, 1195–1211. [CrossRef]
259. Daniels, M.J.; Kass, R.E. Shrinkage estimators for covariance matrices. *Biometrics* **2001**, *57*, 1173–1184. [CrossRef]
260. Wong, F.; Carter, C.K.; Kohn, R. Efficient estimation of covariance selection models. *Biometrika* **2003**, *90*, 809–830. [CrossRef]
261. Sun, D.; Berger, J.O. Objective Bayesian analysis for the multivariate normal model. *Bayesian Stat.* **2007**, *8*, 525–562.
262. Daniels, M.J.; Pourahmadi, M. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* **2002**, *89*, 553–566. [CrossRef]
263. Smith, M.; Kohn, R. Parsimonious covariance matrix estimation for longitudinal data. *J. Am. Stat. Assoc.* **2002**, *97*, 1141–1153. [CrossRef]
264. Lewandowski, D.; Kurowicka, D.; Joe, H. Generating random correlation matrices based on vines and extended onion method. *J. Multivar. Anal.* **2009**, *100*, 1989–2001. [CrossRef]
265. Ghosh, S.; Henderson, S.G. Behavior of the NORTA method for correlated random vector generation as the dimension increases. *ACM Trans. Model. Comput. Simul.* **2003**, *13*, 276–294. [CrossRef]
266. Joe, H. Generating random correlation matrices based on partial correlations. *J. Multivar. Anal.* **2006**, *97*, 2177–2189. [CrossRef]
267. Gilks, W.R.; Richardson, S.; Spiegelhalter, D. *Markov Chain Monte Carlo in Practice*; CRC Press: Boca Raton, FL, USA, 1995.
268. Spiegelhalter, D.J.; Best, N.G.; Carlin, B.P.; Van Der Linde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* **2002**, *64*, 583–639. [CrossRef]
269. Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **2010**, *11*, 3571–3594.
270. Gelfand, A.E.; Ghosh, S.K. Model choice: A minimum posterior predictive loss approach. *Biometrika* **1998**, *85*, 1–11. [CrossRef]
271. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 199–213.
272. Efron, B. How biased is the apparent error rate of a prediction rule? *J. Am. Stat. Assoc.* **1986**, *81*, 461–470. [CrossRef]

273. Burnham, K.P.; Anderson, D.R. Practical use of the information-theoretic approach. In *Model Selection and Inference*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 75–117.
274. Banerjee, S.; Carlin, B.P.; Gelfand, A.E. *Hierarchical Modeling and Analysis for Spatial Data*; CRC Press: Boca Raton, FL, USA, 2014.
275. Gelman, A.; Hwang, J.; Vehtari, A. Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **2014**, *24*, 997–1016. [CrossRef]
276. Celeux, G.; Forbes, F.; Robert, C.P.; Titterington, D.M. Deviance information criteria for missing data models. *Bayesian Anal.* **2006**, *1*, 651–673. [CrossRef]
277. Robert, C.; Casella, G. *Monte Carlo Statistical Methods*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
278. Vehtari, A.; Gelman, A. *WAIC and Cross-Validation in Stan*; Aalto University: Helsinki, Finland, 2014.
279. Box, G.E. Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Stat. Soc. Ser. A* **1980**, *143*, 383–430. [CrossRef]
280. Zellner, A. Bayesian and non-Bayesian estimation using balanced loss functions. In *Statistical Decision Theory and Related Topics V*; Springer: Berlin/Heidelberg, Germany, 1994; pp. 377–390.
281. Vonesh, E.F. Non-linear models for the analysis of longitudinal data. *Stat. Med.* **1992**, *11*, 1929–1954. [CrossRef]
282. Müller, P.; Rosner, G.L. A Bayesian population model with hierarchical mixture priors applied to blood count data. *J. Am. Stat. Assoc.* **1997**, *92*, 1279–1292.
283. Müller, P.; Quintana, F.A. Nonparametric Bayesian data analysis. *Stat. Sci.* **2004**, *19*, 95–110. [CrossRef]
284. Hjort, N.L.; Holmes, C.; Müller, P.; Walker, S.G. *Bayesian Nonparametrics*; Cambridge University Press: Cambridge, UK, 2010; Volume 28.
285. Walker, S.; Wakefield, J. Population models with a nonparametric random coefficient distribution. *Sankhyā Indian J. Stat. Ser.* **1998**, *60*, 196–214.
286. MacKay, D.J. Introduction to Gaussian processes. *NATO ASI Ser. F Comput. Syst. Sci.* **1998**, *168*, 133–166.
287. Rasmussen, C.E. Gaussian processes in machine learning. In *Summer School on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 63–71.
288. Ferguson, T.S. Prior distributions on spaces of probability measures. *Ann. Stat.* **1974**, *2*, 615–629. [CrossRef]
289. Escobar, M.D. Estimating normal means with a Dirichlet process prior. *J. Am. Stat. Assoc.* **1994**, *89*, 268–277. [CrossRef]
290. Escobar, M.D.; West, M. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.* **1995**, *90*, 577–588. [CrossRef]
291. McLachlan, G.J.; Lee, S.X.; Rathnayake, S.I. Finite mixture models. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 355–378. [CrossRef]
292. Rasmussen, C.E. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 12*; 1999; Volume 12, pp. 554–560. Available online: https://papers.nips.cc/paper/1999/hash/97d98119037c5b8a9663cb21fb8ebf47-Abstract.html (accessed on 20 February 2022).
293. Antoniak, C.E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.* **1974**, *2*, 1152–1174. [CrossRef]
294. Teh, Y.W.; Jordan, M.I.; Beal, M.J.; Blei, D.M. Hierarchical dirichlet processes. *J. Am. Stat. Assoc.* **2006**, *101*, 1566–1581. [CrossRef]
295. Jara, A. Theory and computations for the Dirichlet process and related models: An overview. *Int. J. Approx. Reason.* **2017**, *81*, 128–146. [CrossRef]
296. Rosner, G.L.; Müller, P. Bayesian population pharmacokinetic and pharmacodynamic analyses using mixture models. *J. Pharmacokinet. Biopharm.* **1997**, *25*, 209–233. [CrossRef]
297. Müller, P.; Quintana, F.; Rosner, G. A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. Ser. B* **2004**, *66*, 735–749. [CrossRef]
298. Brown, H.; Prescott, R. *Applied Mixed Models in Medicine*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
299. Congdon, P.D. *Applied Bayesian Hierarchical Methods*; CRC Press: Boca Raton, FL, USA, 2010.
300. Jordan, M.I. Graphical models. *Stat. Sci.* **2004**, *19*, 140–155. [CrossRef]
301. Lauritzen, S.L.; Dawid, A.P.; Larsen, B.N.; Leimer, H.G. Independence properties of directed Markov fields. *Networks* **1990**, *20*, 491–505. [CrossRef]
302. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *PAMI-6*, 721–741. [CrossRef] [PubMed]
303. Liu, J.S. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc.* **1994**, *89*, 958–966. [CrossRef]
304. Park, T.; Lee, S. Improving the Gibbs sampler. *Wiley Interdiscip. Rev. Comput. Stat.* **2021**, e1546. [CrossRef]
305. Spiegelhalter, D.J.; Thomas, A.; Best, N.; Lunn, D. *WinBUGS Version 1.4 User Manual*; MRC Biostatistics Unit: Cambridge, UK, 2003. Available online: http://www.mrc-bsu.cam.ac.uk/bugs (accessed on 20 February 2022).
306. Spiegelhalter, D.; Thomas, A.; Best, N.; Lunn, D. OpenBUGS user manual. *Version* **2007**, *3*, 2007.
307. Barthelmé, S.; Chopin, N. Expectation propagation for likelihood-free inference. *J. Am. Stat. Assoc.* **2014**, *109*, 315–333. [CrossRef]
308. Zhu, J.; Chen, J.; Hu, W.; Zhang, B. Big learning with Bayesian methods. *Natl. Sci. Rev.* **2017**, *4*, 627–651. [CrossRef]
309. Jordan, M.I. Message from the president: The era of big data. *ISBA Bull.* **2011**, *18*, 1–3.
310. Johnson, D.; Sinanovic, S. Symmetrizing the kullback-leibler distance. *IEEE Trans. Inf. Theory* **2001**. Available online: https://scholarship.rice.edu/bitstream/handle/1911/19969/Joh2001Mar1Symmetrizi.PDF?sequence=1 (accessed on 20 February 2022).

311. Tan, L.S.; Nott, D.J. Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Stat. Sci.* **2013**, *28*, 168–188. [CrossRef]
312. Ormerod, J.T.; Wand, M.P. Gaussian variational approximate inference for generalized linear mixed models. *J. Comput. Graph. Stat.* **2012**, *21*, 2–17. [CrossRef]
313. Tan, L.S.; Nott, D.J. A stochastic variational framework for fitting and diagnosing generalized linear mixed models. *Bayesian Anal.* **2014**, *9*, 963–1004. [CrossRef]
314. Ngufor, C.; Van Houten, H.; Caffo, B.S.; Shah, N.D.; McCoy, R.G. Mixed Effect Machine Learning: A framework for predicting longitudinal change in hemoglobin A1c. *J. Biomed. Inform.* **2019**, *89*, 56–67. [CrossRef] [PubMed]
315. Capitaine, L.; Genuer, R.; Thiébaut, R. Random forests for high-dimensional longitudinal data. *Stat. Methods Med. Res.* **2021**, *30*, 166–184. [CrossRef] [PubMed]
316. Mandel, F.; Ghosh, R.P.; Barnett, I. Neural Networks for Clustered and Longitudinal Data Using Mixed Effects Models. *Biometrics* **2021**. [CrossRef] [PubMed]
317. Fu, W.; Simonoff, J.S. Unbiased regression trees for longitudinal and clustered data. *Comput. Stat. Data Anal.* **2015**, *88*, 53–74. [CrossRef]
318. Tsybakov, A.B. *Introduction to Nonparametric Estimation*; Springer: New York, NY, USA, 2009.
319. Schulz, E.; Speekenbrink, M.; Krause, A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *J. Math. Psychol.* **2018**, *85*, 1–16. [CrossRef]
320. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
321. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]