Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

Deep Mixture of Linear Mixed Models for Complex Longitudinal Data

Lucas Kock[1] | Nadja Klein[2] | David J. Nott[1]

[1]Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore | [2]Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany

Correspondence: Nadja Klein (nadja.klein@kit.edu)

Received: 2 October 2024 | **Revised:** 17 September 2025 | **Accepted:** 19 September 2025

Funding: This work was supported by Deutsche Forschungsgemeinschaft (KL 3037/1-1), Volkswagen Foundation (96932), Ministry of Education—Singapore (MOE-T2EP20123-0009).

Keywords: deep mixture of factor analyzer | irregularly sampled data | random effects | temporal trends | variational inference

ABSTRACT

Mixtures of linear mixed models are widely used for modeling longitudinal data for which observation times differ between subjects. In typical applications, temporal trends are described using a basis expansion, with basis coefficients treated as random effects varying by subject. Additional random effects can describe variation between mixture components or other known sources of variation in complex designs. A key advantage of these models is that they provide a natural mechanism for clustering. Current versions of mixtures of linear mixed models are not specifically designed for the case where there are many observations per subject and complex temporal trends, which require a large number of basis functions to capture. In this case, the subject-specific basis coefficients are a high-dimensional random effects vector, for which the covariance matrix is hard to specify and estimate, especially if it varies between mixture components. To address this issue, we consider the use of deep mixture of factor analyzers models as a prior for the random effects. The resulting deep mixture of linear mixed models is well suited for high-dimensional settings, and we describe an efficient variational inference approach to posterior computation. The efficacy of the method is demonstrated in biomedical applications and on simulated data.

1 | Introduction

Longitudinal data play an important role in many biomedical applications [1, 2]. In their practical use, mixtures of linear mixed models (MLMMs) [3] are widely used for the analysis of longitudinal data for which observation times differ by subject, and in cases where there is a need to "borrow strength" between subjects in a flexible way. A common approach to modeling temporal trends in MLMMs is to use flexible basis expansions, with basis coefficients treated as a random effect varying across individuals. The mixture structure for the distribution of the random effects provides flexibility when the random effects are non-Gaussian, and also provides a natural mechanism for clustering which enhances interpretability. In settings where there are a large number of observations per subject and the temporal trends are complex, many basis functions may be required, which results in high-dimensional random effects. The main contribution of this paper is to address the issue of high-dimensionality in MLMMs by using a deep mixture of factor analyzers (DMFA) model as the prior for the random effects distribution. The result is a new deep mixture of linear mixed model (DMLMM) specification. We discuss efficient variational methods for computation and demonstrate the good performance of our approach in simulations and a number of real biomedical examples involving

within-subject prediction for unbalanced longitudinal biomarker data, likelihood-free inference (LFI) for modeling the temporal dynamics of malaria transmission, and missing data imputation for gene expression data.

A common application of MLMMs has been in clustering of time course gene expression data. Several authors have considered linear mixed models (LMMs) with basis expansions for modeling temporal trends, and extensions to mixtures for clustering [4–6]. A similar approach in the functional data analysis literature is described by James and Sugar [7]. Celeux et al. [8] consider MLMMs for clustering of gene expression datasets with replication, where gene level random effects are shared between replicates. Ng et al. [9] extend this model with a random effect for different tissues, and Tan and Nott [10] consider a similar model with two random effects, one for subjects and one for the mixture component, and allow for covariate-dependent mixing weights. They consider Bayesian inference in their model, with computations carried out using variational approximation methods. Scharl et al. [11] consider initialization of EM algorithms for mixtures of regression models, including MLMMs, for clustering time series gene expression data. Pfeifer [12] clusters longitudinal data using LMMs, where the random effects distribution is either a finite mixture of normals or some arbitrary distribution approximated discretely. Coke and Tsao [13] consider clustering of electrical load series. MLMMs also arise in the literature on model-based functional clustering, where approximations to continuous time processes can lead to processes defined from finite basis expansions and a LMM formulation. Examples include Chiou and Li [14], who consider a nonparametric random effects model and a truncated Karhunen-Loève expansion, and Jacques and Preda [15] in which the authors cluster multivariate functional data assuming that multivariate functional principal components are normally distributed. McDowell et al. [16] perform functional clustering of gene expression data using a Dirichlet process Gaussian process mixture model. Shi and Wang [17] develop a mixture of Gaussian process functional regressions model where the mixing weights can be covariate-dependent.

There are a variety of generalizations or closely related models to finite MLMMs. These include partition models [18, 19] mixtures of generalized LMMs (GLMMs) [20, 21] and mixtures of nonlinear hierarchical models [22, 23]. Bai et al. [24] robustify mixtures of linear mixed models by assuming a multivariate-t distribution for the responses and random effects jointly within each mixture component. LMMs with nonparametric priors, include infinite mixtures of LMMs or more general hierarchical models have been considered in the literature on Bayesian nonparametrics [25–28]. Sigrist [29] combines boosting and latent Gaussian processes to specify a random effects models for longitudinal data.

There have been several recent works integrating mixed effects models with deep learning methods. Kilian et al. [30] introduce techniques to introduce random effects post hoc into arbitrary supervised regression models. Tran et al. [31] represent fixed and random effects of GLMMs through deep networks and use variational methods for inference in the resulting complex model. Similarly, Mandel et al. [32] replace the linear effects of a mixed effects model with neural networks. The resulting model is especially suited to handle densely sampled longitudinal data. A recent overview on machine learning techniques for longitudinal biomedical data can be found in Cascarano et al. [33]. In these approaches, deep models are typically used to increase flexibility in the fixed effects. This is in contrast to our approach, which considers a flexible random effects distribution.

In our model, the DFMA introduced by Viroli and McLachlan [34] serves as a prior for the random effects in MLMMs. It is based on a mixture of factor analyzers model [35, 36] but instead of assuming factors to be Gaussian, allows the factors to themselves be modeled as a mixture of factor analyzers recursively for multiple layers. The model of Viroli and McLachlan [34] builds on an earlier formulation described in Tang et al. [37], where components are split recursively and the fitting is done layerwise. However, Viroli and McLachlan [34] use a similar architecture to that in van den Oord and Schrauwen [38], where the authors allow parameter sharing between mixture components, although they do not consider factor structures for the mixture component covariance matrices. Other related mixture models are considered in Yang et al. [39], Li [40] and Malsiner-Walli et al. [41]. We build on the Bayesian formulation of DMFAs proposed by Kock et al. [42] and implement efficient variational methods for computation.

The DMFA prior allows for complex high-dimensional random effects distributions. Conditional distributions derived from our DMLMM approach are analytically tractable, thus predictive distributions for unobserved time points can be derived in a computationally attractive manner. One scenario where this is useful is predictive LFI. Simulators with intractable likelihoods are commonly used in biomedical applications [43, 44] and inference is often based on a large sample from the simulator. If each sample is a high-dimensional time series, a large number of basis functions is needed to estimate the temporal trend. Mixture models are a well-established tool in LFI, where the goal is parameter inference [45–47], but predictive LFI has not been explored within the MLMM literature before.

Throughout this paper, we demonstrate the adaptability of our DMLMM approach across a range of biomedical applications, each presenting distinct challenges in modern biostatistics. Firstly, we consider within-subject prediction for an unbalanced longitudinal study. Secondly, we consider the task of predicting the number of malaria cases in Afghanistan based on an intractable simulator. Lastly, an application to missing data imputation for gene expression data is given within the Supporting Information. Mixture modeling allows adaptive local sharing of information, which improves imputation. Across all these applications, the Gaussian mixture model (GMM) representation of the DMLMM is helpful for interpretation and the derivation of additional insights. Python code for the DMLMM is publicly available at github.com/kocklucx/DMLMM.

The structure of the paper is as follows. In the next section, we introduce the DMLMM for longitudinal data based on a Bayesian version of the DMFA model considered in Viroli and McLachlan [34] and outline efficient variational inference methods for posterior estimation in Section 3. Sections 4 and 5 investigate the properties of our approach empirically. Section 6 gives some concluding discussion.

2 | Deep Mixture of LMMS

This section introduces the DMLMM. Section 2.1 describes the overall model, whereas Section 2.2 discusses the DMFA prior for the random effects in more detail. A formal mathematical description can be found in the Supporting Information.

2.1 | The DMLMM—Notation and Model Specification

Consider a longitudinal study where data $y_i = (y_{i1}, \ldots, y_{in_i})^\top$ is observed for subject i, $i = 1, \ldots, n$, with y_{ij} an observation at time t_{ij}, $j = 1, \ldots, n_i$. Writing $t_i = (t_{i1}, \ldots, t_{in_i})^\top$, it is assumed that

$$y_i = B(t_i)\beta_i + \varepsilon_i, \tag{1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i})$, $B(t_i) = (B(t_{i1}), \ldots, B(t_{in_i}))^\top$ is a known $n_i \times d$ design matrix where $B(t_{ij})$ is a d-dimensional column vector of basis functions evaluated at t_{ij}, and $\beta_i \in \mathbb{R}^d$ is a subject-specific random coefficient vector. Note, that we do not enforce an explicit relationship between n_i and d. In particular, we explicitly allow $n_i < d$ for some individuals i. Our DMLMM approach assumes the additive errors ε_i to be uncorrelated. Longitudinal studies, where the random errors exhibit within-subject correlation, are for example considered in de Alencar et al. [48] and Lin and Wang [49]. We consider Bayesian inference, and use a half-Cauchy prior $\sigma \sim \mathcal{HC}(A)$ for the standard deviation of the error terms ε_i, which we express hierarchically as

$$\sigma^2 \mid \psi \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{\psi}\right), \qquad \psi \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{A^2}\right).$$

We choose this thick-tailed prior for the error variance as it robustifies the model against conflicts with the data for example through outliers. Section 2.2 introduces a DMFA model which we use as a flexible prior distribution for the random effects β_i. Write $\beta = (\beta_1^\top, \ldots, \beta_n^\top)^\top$, and $\theta = (\eta^\top, \beta^\top)$ where θ are the unknown parameters, so that η contains the unknowns except for β. The DMFA prior density for β_i is a GMM with density of the form

$$p(\beta_i|\eta) = \sum_{k=1}^{K} w_k \phi(\beta_i, \mu_k; \Sigma_k),$$

where $\sum w_k = 1$, and $\phi(\cdot; \mu, \Sigma)$ denotes the multivariate normal density function with mean μ and covariance matrix Σ. In the DMFA the parameters w_k, μ_k and Σ_k are parametrized parsimoniously and this is described in detail later. Integrating out β in (1) using $p(\beta_i|\eta)$ gives the marginal likelihood

$$p(y_i|\eta) = \sum_{k=1}^{K} w_k \phi\left(y_i, B(t_i)\mu_k; B(t_i)\Sigma_k B(t_i)^\top + \sigma^2 I_{n_i}\right). \tag{2}$$

The random effects β_i can be interpreted as projections of the unequal length observations y_i into a joint d-dimensional latent space. Our later applications demonstrate that the flexible DMFA prior allows complex trends to be modeled well when the number of basis functions is large, whereas borrowing strength between similar observations to stabilize estimation for subjects having little available data.

A key task that we address in these applications is within-subject prediction. Suppose that for subject i we need predictive inferences about unobserved data $\tilde{y}$ at time points $\tilde{t} = t_1, \ldots, t_T$. Integrating out β, the joint density of $(y_i, \tilde{y})$ given η is a high-dimensional GMM,

$$p(y_i, \tilde{y}|\eta) = \sum_{k=1}^{K} w_k \phi\left(\begin{bmatrix} y_i \\ \tilde{y} \end{bmatrix}; \begin{bmatrix} B(t_i)\mu_k \\ B(\tilde{t})\mu_k \end{bmatrix}, \times \begin{bmatrix} B(t_i)\Sigma_k B(t_i)^\top + \sigma^2 I_{n_i} & B(t_i)\Sigma_k B(\tilde{t})^\top \\ B(\tilde{t})\Sigma_k B(t_i)^\top & B(\tilde{t})\Sigma_k B(\tilde{t})^\top + \sigma^2 I_T \end{bmatrix}\right)$$

leading to a conditional density for $\tilde{y}$ given y_i, η which is also a GMM:

$$p(\tilde{y}, y_i|\eta) = \sum_{k=1}^{K} \tilde{w}_k \phi(\tilde{y}, \tilde{\mu}_k; \tilde{\Sigma}_k), \tag{3}$$

where

$$\begin{aligned}
\tilde{w}_k &= \frac{w_k \phi\left(y_i, B(t_i)\mu_k; B(t_i)\Sigma_k B(t_i)^\top\right)}{\sum_{k=1}^{K} w_k \phi\left(y_i, B(t_i)\mu_k; B(t_i)\Sigma_k B(t_i)^\top\right)}, \\
\tilde{\mu}_k &= B(\tilde{t})\mu_k - B(\tilde{t})\Sigma_k B(t_i)^\top \left(B(t_i)\Sigma_k B(t_i)^\top + \sigma^2 I_{n_i}\right)^{-1} \\
&\quad \times (y_i - B(t_i)\mu_k), \\
\tilde{\Sigma}_k &= B(\tilde{t})\Sigma_k B(\tilde{t})^\top + \sigma^2 I_T - B(\tilde{t})\Sigma_k B(t_i)^\top \\
&\quad \times \left(B(t_i)\Sigma_k B(t_i)^\top + \sigma^2 I_{n_i}\right)^{-1} B(t_i)\Sigma_k B(\tilde{t})^\top.
\end{aligned}$$

Predictive inference can be obtained from (3) either in a plug-in fashion, using a point estimate of η, or by integrating out the parameters over the posterior distribution or some approximation to it. In Section 3, we will consider posterior approximations and point estimates obtained using variational inference. Figure 1 illustrates the full DMLMM including the DMFA prior and model training process which we will discuss further next.

2.2 | DMFA Prior for Subject Specific Random Effects

The DMFA model was motivated by Viroli and McLachlan [34] as a deep extension of the mixture of factor analyzers (MFA) model, which can be thought of as a DMFA model with only one layer. Although Viroli and McLachlan [34] and Kock et al. [42] consider DMFA models for multivariate data directly, here it will be used as a prior for random effects in a LMM.

The hierarchical DMFA prior is a generative model for the random effects β_i expressed in terms of latent variables arranged in a number of layers. Define $z_i^{(0)} := \beta_i$, and write $z_i^{(l)} \in \mathbb{R}^{D^{(l)}}$, $i = 1, \ldots, n$, for latent variables at layer $l \in \{1, \ldots, L\}$. We define the model for $z_i^{(l-1)}$ $l = 1, \ldots, L$, in terms of $z_i^{(l)}$ as a mixture model, with $K^{(l)}$ components. At level l, the mixing weights for the mixture are denoted $w_k^{(l)}$, $k = 1, \ldots, K^{(l)}$, $\sum_k w_k^{(l)} = 1$. The model for $z_i^{(l-1)}$ given $z_i^{(l)}$ is expressed generatively as follows: for $l = 1, \ldots, L$, with probability $w_k^{(l)}$, the latent variable $z_i^{(l-1)}$ is generated as

$$z_i^{(l-1)} = \mu_k^{(l)} + B_k^{(l)} z_i^{(l)} + \epsilon_{ik}^{(l)}, \tag{4}$$

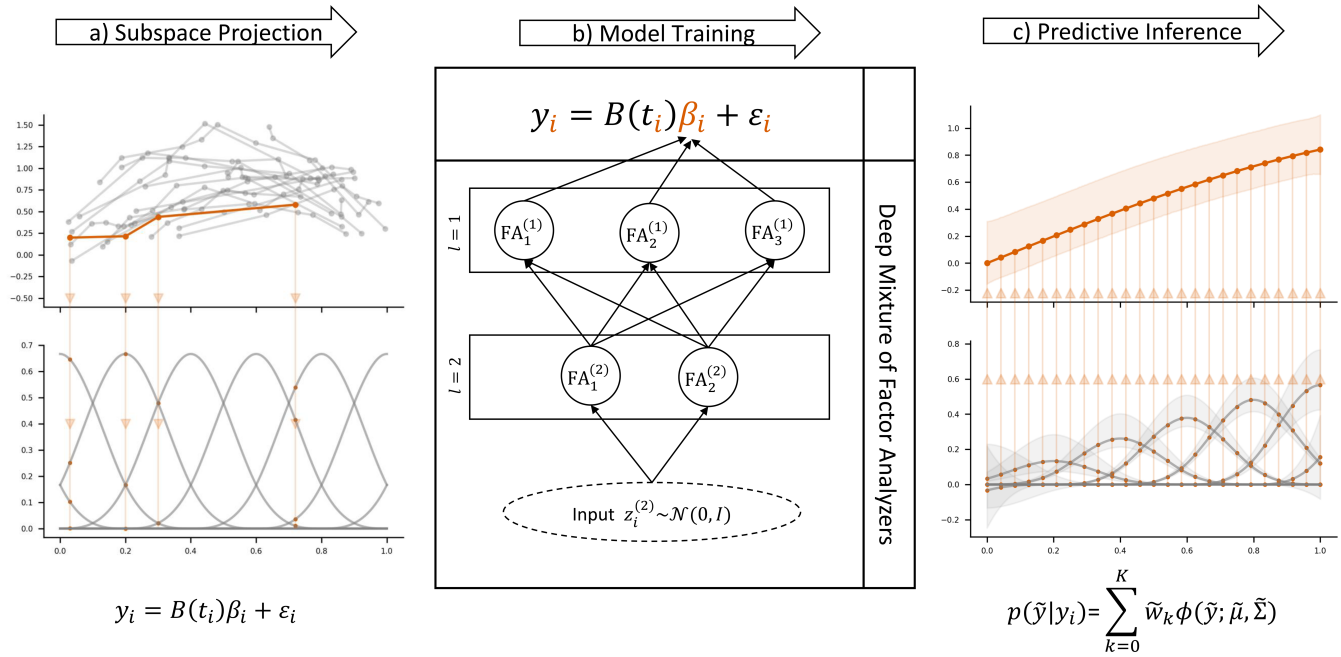

FIGURE 1 | Schematic description of the DMLMM. (a) represents the sub-space projection. The number of observations n_i as well as the time points t_i may vary between individuals i. By means of a basis approximation, all vectors y_i get projected into a lower dimensional sub-space of dimension d controlled through the random coefficients β_i. (b) represents the model training. The DMLMM consists of a regression layer of the form (1) and the DMFA prior for β_i. Here, we give an examplary DMFA architecture with $L = 2$ layers. The latent input variables $z_i^{(2)}$ are fed through a fully connected network with $L = 2$ layers, with $K^{(1)} = 3$ and $K^{(2)} = 2$ components, respectively. The components of the network correspond to a factor analyzer of form (4). Each of the $K^{(1)} \cdot K^{(2)} = 6$ possible paths through this model corresponds to a GMM component. (c) represents posterior predicted inference based on the fitted DMLMM for new unobserved data $\tilde{y}$ conditional on y_i exploiting (3).

where $\epsilon_{ik}^{(l)} \sim \mathcal{N}\left(0, \delta_k^{(l)}\right)$, $\mu_k^{(l)}$ is a $D^{(l-1)}$-vector, $B_k^{(l)}$ is a $D^{(l-1)} \times D^{(l)}$ lower triangular matrix, $\delta_k^{(l)} = \text{diag}\left(\delta_{k1}^{(l)}, \ldots, \delta_{kD^{(l-1)}}^{(l)}\right)$ is a $D^{(l-1)} \times D^{(l-1)}$ diagonal matrix with diagonal elements $\delta_{kj}^{(l)} > 0$. At the final layer $z_i^{(L)} \sim \mathcal{N}\left(0, I_{D^{(L)}}\right)$. In the specification of the DFMA prior, we restrict the dimensionality of the latent variables to satisfy the Anderson-Rubin condition [50] $D^{(l+1)} \leq \frac{D^{(l)}-1}{2}$ for $l = 0, \ldots, L$, as it is a necessary condition for ensuring model identifiability. (Figure 1b) gives an example for a DMFA prior architecture with $L = 2$ layers. Kock et al. [42] recommend architectures with few layers and a rapid decrease in dimension. Following this recommendation, we consider models with $L = 2$ layers throughout our experiments.

Following the discussion of Viroli and McLachlan [34], the DMFA prior can be regarded as a GMM with $K = \prod_{l=1}^{L} K^{(l)}$ components. The components correspond to "paths" through the factor mixture components at the different levels. Write $k_l \in \left\{1, \ldots, K^{(l)}\right\}$ for the index of a factor mixture component at level l and let $k = \left(k_1, \ldots, k_L\right)^\top$ index a path. Let $w_k = \prod_{l=1}^{L} w_{k_l}^{(l)}$,

$$\mu_k = \mu_{k_1}^{(1)} + \sum_{l=2}^{L}\left(\prod_{m=1}^{l-1} B_{k_m}^{(m)}\right)\mu_{k_l}^{(l)}, \quad \text{and}$$

$$\Sigma_k = \delta_{k_1}^{(1)} + \sum_{l=2}^{L}\left(\prod_{m=1}^{l-1} B_{k_m}^{(m)}\right)\delta_{k_l}^{(l)}\left(\prod_{m=1}^{l-1} B_{k_m}^{(m)}\right)^\top.$$

Then the DMFA prior corresponds to the Gaussian mixture density $\sum_{k=1}^{K} w_k \phi\left(y; \mu_k, \Sigma_k\right)$.

To get some intuition for the DMFA prior construction, it is helpful to consider the case of a single layer, $L = 1$. In this case, the DMFA prior is a mixture of factor analyzers (MFA) prior on the random effects. Abusing notation by writing simply $K = K^{(1)}$, $w_k = w_k^{(1)}$, $\mu_k = \mu_k^{(1)}$, $B_k = B_k^{(1)}$, $\delta_k = \delta_k^{(1)}$, $k = 1, \ldots, K$, and $z_i = z_i^{(1)}$, $i = 1, \ldots, n$, (4) specifies the prior for β_i through the following single generative layer: with probability w_k, generate β_i as

$$\beta_i = \mu_k + B_k z_i + \epsilon_{ik},$$

where $\epsilon_{ik} \sim \mathcal{N}\left(0, \delta_k\right)$. Integrating out the latent variables z_i, the corresponding density of β_i is

$$\sum_{k=1}^{K} w_k \phi\left(\beta_i, \mu_k; B_k B_k^\top + \delta_k\right).$$

The low-dimensional latent variables z_i allow a parsimonious description of the dependence between the possibly high-dimensional components in β_i; conditionally on z_i, components of β_i are independent. The latent variables z_i are called factors, and the matrices B_k are called factor loadings or factor loading matrices. The key idea of the DMFA prior is to replace the Gaussian assumption $z_i \sim \mathcal{N}(0, I)$ with the assumption that the z_i's themselves follow a MFA model.

In a Bayesian framework, Kock et al. [42] propose the following marginally independent priors for the parameters of a DMFA model, and we use similar priors for the hyperparameters on the DMFA prior for the random effects. Selecting marginally independent priors, which do not share information across layers and

components, is crucial for a closed form mean field approximation as derived in Section 3. They use thick-tailed Cauchy priors on the component mean parameters $\mu_k^{(l)}$ and half-Cauchy priors on the standard deviations $\delta_k^{(l)}$. Thus integration over the model parameters yields a prior centered on zero for the random effect distribution, which does not introduce an unwanted bias for β_i. In the DMLMM the same prior is used also for the standard deviation σ of the error terms ε_i. For the component factor loading matrices $B_k^{(l)}$, they use the sparsity-inducing horseshoe prior of Carvalho and Polson [51]. Kock et al. [42] show that this prior choice is helpful with regularizing the estimation. Additionally, in the DMLMM imposing sparsity on the factor loadings is motivated by the fact that the entries of the coefficient vector β_i control local information and therefore each of the latent factors should control only a subset of components, but not the full vector. Typically the basis functions are chosen such that $B(t)$ is sparse as well. Lastly, the marginal prior for $w^{(l)}$ is a Dirichlet distribution allowing to select the number of clusters in a computationally thrifty way, using overfitted mixtures [52]. A precise description of the priors is given in Web Appendix A in Supporting Information.

3 | Posterior Computation

Next we review basic ideas of variational inference (VI) [53] and explain how the scalable variational inference algorithm for the DMFA model in Kock et al. [42] can be extended to the new DMLMM with DMFA prior for the random effects. Further details can be found in the Supporting Information.

3.1 | Variational Inference

VI learns an approximation to the posterior density $p(\theta|y)$ in Bayesian inference using an approximating family of densities $\{q_\lambda(\theta), \lambda \in \Lambda\}$ where λ are variational parameters to be chosen. The optimal approximation is obtained by finding the value λ^* of λ minimizing some measure of dissimilarity between $p(\theta|y)$ and $q_\lambda(\theta)$. A common choice for the dissimilarity measure is the reverse Kullback–Leibler (KL) divergence,

$$\mathcal{D}_{\mathrm{KL}}\left[q_\lambda(\theta)\middle\|p(\theta|y)\right] = \mathbb{E}_{q_\lambda}\left\{\log(q_\lambda(\theta)/p(\theta|y))\right\},$$

where $\mathbb{E}_{q_\lambda}(\cdot)$ denotes expectation with respect to $q_\lambda(\theta)$. Minimizing the reverse KL divergence is equivalent to maximizing the Evidence Lower Bound (ELBO)

$$\mathcal{L}(\lambda) = \mathbb{E}q_\lambda\left\{\log(h(\theta)) - \log\left(q_\lambda(\theta)\right)\right\}, \tag{5}$$

where $h(\theta) = p(y|\theta)p(\theta)$. For the DMLMM with a DMFA prior for the random effects, we consider variational approximations leading to a closed form expression for the ELBO. We optimize the ELBO using a stochastic gradient ascent (SGA) method which uses mini-batch sampling to effectively deal with large datasets. We give a high level discussion of the approach next, a detailed discussion can be found in Kock et al. [42].

3.2 | VI for the DMFA

The SGA algorithm for the original DMFA model of Kock et al. [42] adapts stochastic VI [54] by partitioning the variational parameters into "global" parameters λ_G, which parametrize variational posterior terms for shared model parameters such as the factor loading matrices $B_k^{(l)}$ or the component mean shift vectors $\mu_k^{(l)}$, and "local" parameters λ_L, which parametrize variational posterior terms for observation specific latent variables, such as $z_i^{(l)}$. Write $\lambda = \left(\lambda_G^\top, \lambda_L^\top\right)^\top$, and denote the value of λ_L maximizing the ELBO for a given value of λ_G as $M\left(\lambda_G\right)$. We then consider the ELBO as a function of λ_G, with λ_L fixed at $M\left(\lambda_G\right)$:

$$\overline{\mathcal{L}}\left(\lambda_G\right) \coloneqq \mathcal{L}\left(\lambda_G, M\left(\lambda_G\right)\right).$$

The stochastic VI algorithm we use optimizes $\overline{\mathcal{L}}\left(\lambda_G\right)$ where at step $m = 1, \ldots, M$ of the SGA algorithm there are two nested steps. First, the optimal local parameters $\widehat{\lambda}_L$ for the current global parameter vector $\lambda_G^{(m-1)}$ are updated. Then, the global parameters are updated as

$$\lambda_G^{(m)} = \lambda_G^{(m-1)} + a_m \circ \widehat{\nabla_{\lambda_G}\overline{\mathcal{L}}}\left(\lambda_G^{(m-1)}\right), \tag{6}$$

where a_m is a vector-valued step size sequence, $\circ$ denotes elementwise multiplication, and $\widehat{\nabla_{\lambda_G}\overline{\mathcal{L}}}\left(\lambda_G^{(m-1)}\right)$ is an unbiased estimate of the natural gradient [55] of $\overline{\mathcal{L}}\left(\lambda_G^{(m-1)}\right)$ based on a random data mini-batch, where $\overline{\mathcal{L}}(\cdot)$ denotes the ELBO with local parameters fixed at $\widehat{\lambda}_L$. Optimization of local variational parameters is only required for the observations in the data mini-batch, which leads to an efficient algorithm for large data sets.

3.3 | VI for DMLMMs

The deep structure of the DMLMM corresponds to a DMFA model with an additional regression layer of the form (1) on top (see Figure 1b). The regression layer has a very similar structure to a single layer in the DMFA model, (4), where the factor loading matrix is fixed at $B\left(t_i\right)$ and the mean shift vector is zero. This perspective allows us to extend the efficient VI scheme for DMFA to DMLMM as follows.

Let θ_{DMFA} denote the vector of all unknown model parameters for the DMFA prior and $\theta_{\mathrm{Reg}} = \left(\beta^\top, \sigma^2, \psi\right)^\top$ be the vector of the remaining parameters. The full set of unknown model parameters for the DMLMM is then $\theta = \left(\theta_{\mathrm{DMFA}}^\top, \theta_{\mathrm{Reg}}^\top\right)^\top$. We assume a factorized variational approximation to the posterior density of the form

$$q_\lambda(\theta) = q_{\lambda_{\mathrm{DMFA}}}\left(\theta_{\mathrm{DMFA}}\right) q_{\lambda_{\mathrm{Reg}}}\left(\theta_{\mathrm{Reg}}\right), \tag{7}$$

where $q_{\lambda_{\mathrm{DMFA}}}\left(\theta_{\mathrm{DMFA}}\right)$ is the density for θ_{DMFA} with variational parameters λ_{DMFA} and

$$q_{\lambda_{\mathrm{Reg}}}\left(\theta_{\mathrm{Reg}}\right) = q\left(\sigma^2\right) q(\psi) \prod_{i=1}^{n} q\left(\beta_i\right),$$

where $q(\sigma^2)$ and $q(\psi)$ are inverse gamma densities and $q(\beta_i)$ is a multivariate Gaussian density with independent marginals. Then,

$$h(\theta) = p(\theta)\prod_{i=1}^{n} p(y_i|\theta)$$
$$= p(\theta_{\text{DMFA}})p(\sigma^2|\psi)p(\psi)\prod_{i=1}^{n} p(y_i|\beta_i,\sigma^2)p(\beta_i|\theta_{\text{DMFA}})$$

and (5) can be decomposed as

$$\mathcal{L}(\lambda) = \mathcal{L}^{\text{DMFA}}(\lambda) + \mathcal{L}^{\text{Reg}}(\lambda),$$

where

$$\mathcal{L}^{\text{DMFA}}(\lambda) = \mathbb{E}q_\lambda\left[\sum_{i=1}^{n}\log(p(\beta_i|\theta_{\text{DMFA}})) + \log(p(\theta_{\text{DMFA}})) - \log(q_{\lambda_{\text{DMFA}}}(\theta_{\text{DMFA}}))\right]$$

can be derived from the ELBO for the DFMA model and

$$\mathcal{L}^{\text{Reg}}(\lambda) = \mathbb{E}q_\lambda\left[\sum_{i=1}^{n}\log(p(y_i,\beta_i|\sigma^2)) + \log(p(\sigma^2|\psi)p(\psi)) - \log\left(q_{\lambda_{\text{Reg}}}(\theta_{\text{Reg}})\right)\right]$$

is available in closed form. More details on the calculation of $\mathcal{L}(\lambda)$ can be found in the Web Appendix B in Supporting Information.

$\mathcal{L}(\lambda)$ has a similar structure to the ELBO for the DMFA model, where β is an additional "local" parameter and σ, ψ are "global" parameters. As a result, it is straightforward to adapt the updating approach explained in Section 3.2 to the DMLMM.

In the DMFA model the use of overfitted mixtures and ELBO values of short runs allows to choose a suitable architecture in a computationally thrifty way and this idea directly translates to the DMLMM. The choice of the number of layers and factors in our DMLMM also follows the choices made in the DMFA model. Due to the parameter sharing, some components of the GMM representation (2) might be empty, even when there are data points assigned to every component in each layer [56]. Although this does not affect the clustering induced by the DMFA prior it can have negative impact on the resulting density estimation. Hence, we recommend that after the full model is fitted the weights for empty components of the GMM density are manually set to zero and remaining weights are rescaled. Predictive inference in the DMLMM is carried out using the variational posterior mean as a point estimate for η.

4 | Real Data Illustrations

In this section we showcase our DMLMM in diverse real data applications. First, we consider longitudinal CD4 counts, which are an established illustration in the longitudinal literature. Then, we consider a novel application on malaria transmission. Here, the deep structure of our approach is helpful in capturing the complex temporal structure of the data. An application on missing data imputation for gene expression data is presented in Appendix C in Supporting Information.

4.1 | Longitudinal CD4 Counts

4.1.1 | Data and Model Description

CD4 percentages are a popular prognostic marker of disease stage among human immunodeficiency virus (HIV)-infected individuals. Here, we consider data from the Multicenter AIDS Cohort Study [57] which has been analyzed by many previous authors, for example by [58–60]. The dataset contains repeated measurements for 283 MSM (men who have sex with men) who were tested HIV-positive between 1984 and 1991. Even though individuals were expected to get their measurements taken at regular 6-month intervals, the number of measurements and the measurement times differ per individual. The observed trajectories for all individuals are shown in Figure 2a.

The goal is to model the CD4 percentage trajectories in continuous time as well as to dynamically predict CD4 percentages at future time points. To this end, we denote by y_i the n_i-dimensional vector of observed CD4 measurements on the probit scale for individual i. The design matrices $B(t_i)$ are constructed from $d = 7$ Legendre polynomials. Since n_i varies greatly between individuals ranging from 1 to 14, $n_i < d$ for more than half of the individuals. Let $\tilde{y}$ denote the unobserved CD4

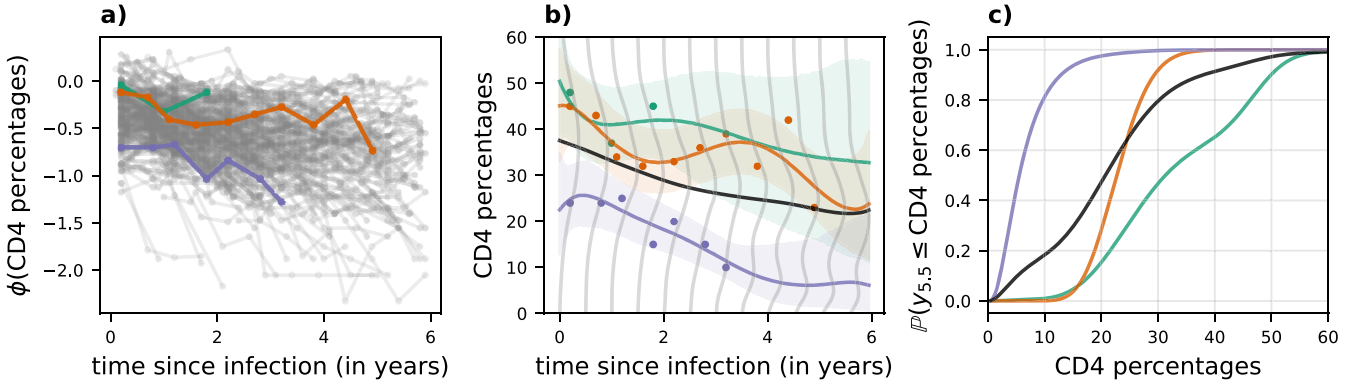

FIGURE 2 | CD4 data. (a) Spaghetti plot for all observed trajectories on the probit scale. Three randomly selected trajectories y_i are marked in color. (b) Predicted trajectories (bold) and 95% credible intervals for the three randomly selected individuals, with observed measurements given by dots. The gray lines correspond to the estimated marginal densities $p(\tilde{y}|\hat{\eta})$ at different times $\tilde{t}$ and the mean $\mathbf{E}(\tilde{y}|\hat{\eta})$ is given in black. (c) Plot of the predicted CDF $\mathbb{P}(\tilde{y}_j \leq \cdot|y_i,\hat{\eta})$ at $\tilde{t}_j = 5.5$ for the three individuals. The CDF of the marginal $\mathbb{P}(\tilde{y}_j \leq \cdot|\hat{\eta})$, $\tilde{t}_j = 5.5$ is given in black.

measurements on a fine equidistant grid $\tilde{t}$ over $[0, 6]$ with 120 grid points.

4.1.2 | Results

Figure 2b shows the estimated mean effects $\mathbb{E}(\tilde{y}|y_i, \hat{\eta})$ with 95% pointwise credible intervals for three randomly selected individuals based on the observed measurements. Even in cases with limited measurement data, the method reconstructs meaningful trajectories by combining information from both the specific individual and the entire dataset. As expected, credible intervals are wider in regions where no measurements are observed and near the end of the time interval, where fewer data points are observed. In a diagnostic context, within-subject forecasting is of particular interest. By (2), the cumulative distribution function (CDF) of a GMM can be expressed as a mixture of Gaussian CDFs. This allows for a simple calculation of the risk of the CD4 percentage of an individual falling below a threshold at a given time. Figure 2c shows the predicted CDFs $\mathbb{P}(\cdot|y_i, \hat{\eta})$ for three selected individuals at $\tilde{t} = 4.5$.

Further insight can be obtained through the predictive marginal density for an individual for which no data has been observed, $p(\tilde{y}|\hat{\eta})$, which is depicted in Figure 2b. Computation of this marginal density is simple as the mean effect, the variance function and the correlation function are available in closed form for a Gaussian mixture density. The mean effect shows an overall decreasing trend among individuals. The variance function is non-stationary and increases over time. The marginal distribution for time points near the end of the observation period becomes bimodal. As expected, time points close to each other are estimated to be highly correlated.

4.2 | Predicting Malaria Transmission in Afghanistan

The DMLMM approach is motivated by scenarios where both the number of observations and the dimension of the random effect are large. Such a scenario is commonly encountered when analyzing complex dynamical systems. Here, we reanalyze monthly data reporting malaria cases registered in Afghanistan from January 2005 to September 2015 [61]. In a recent analysis, Alahmadi et al. [62] considered the data in a classical Bayesian parameter inference setting. Here we are interested in forecasting future case counts based on the observed data. There is an extensive literature on models for infectious diseases [63] with Susceptible–infected–recovered (SIR) models being a popular choice [64, 65].

White et al. [66] and Alahmadi et al. [62] propose a nonlinear ordinary differential equation (ODE) model based on the SIR model to describe the temporal population dynamics associated with malaria transmission. The model uses four coupled ODEs modeling four population compartments (uninfected and non-immune, infected with no prior immunity, uninfected with immunity and infected with prior immunity). These ODEs are highly parameterized to describe the complex evolution of the population compartments over time. As none of the population compartments can be directly observed a fifth ODE describing the total number of treated cases is incorperated into the model. We observe $y_j \sim \mathcal{N}(\log(c_j), \sigma^2)$, where c_j denotes the number of new cases at time $t_j \in [0, T]$. A full description of the underlying latent ODE model can be found in Alahmadi et al. [62]. We write $y_{t_1:t_2}$ for the vector of observations at time points $t = (t_1, \ldots, t_2)^\top$ and θ for the vector of parameters of the ODE model.

Since the ODE model is not fully observed, $p(y_{t:T}|\theta, y_{1:t})$ is not available in closed form and $p(y_{1:T}|\theta)$ is costly to evaluate as it involves numerically approximating a solution to the ODE. However, simulating data from the marginalized likelihood $p(y_{1:T}) = \int p(y_{1:T}|\theta)d\theta$ is straight forward and the underlying model can be regarded as a black-box simulator.

Although simulator-based or LFI methods such as Approximate Bayesian Computation (ABC) [67] are commonly used for parameter estimation and model comparison with computationally expensive likelihoods, predictive inference, such as computing the posterior predictive distribution of future observations or missing data, remains challenging due to the complexity of the underlying dynamics and the high dimensionality of the observations. In contrast, the DMLMM enables closed form calculations of the predictive distribution without the need for direct inference on the model parameters, tedious calibration of hyperparameters, or selection of summary statistics based on expert knowledge. Instead, the flexibility of the DMLMM allows us to learn a low-dimensional representation of the high-dimensional observations that captures the relevant information for prediction.

4.2.1 | Experimental Design

As the uninformative prior $p(\theta)$ used in Alahmadi et al. [62] results in many unrealistic time series, we reject any simulated time series for which the number of simulated cases never exceeds 100. Since we regard $p(y) = p(y_{1:128})$ as a black-box simulator, we do not need to make this constraint explicit in the model formulation. We generate 7500 samples from $p(y)$, which we split into a training set with 5000 samples and a test set with 2500 samples. Our goal is to approximate $p(y_{81:128}|y_{1:80})$ using the joint samples from $p(y_{1:128})$. Järvenpää and Corander [68] discuss how ordinary ABC can be used in this setting and we use their approach, which we label ABC, as a benchmark. The design matrices $B(\cdot)$ used for DMLMM incorporate a 20-dimensional spline basis, with 6 splines modeling a yearly seasonality to account for the seasonal forcing associated with malaria transmission, and the remaining basis functions modeling an additive trend.

4.2.2 | Results

Figure 3a shows the predicted time series for the observed data. Both, ABC and the DMLMM recover the general behavior of the unobserved data points well. Studying the 95% credible intervals for both methods shows no large difference between the DMLMM approach and ABC, although our approach has slightly better coverage properties.

The DMLMM also performs slightly better in terms of the root mean square error (RMSE) $\sqrt{\frac{1}{T-t}\sum_{t'=t}^{T}(y_{t'} - \hat{y}_{t'})^2}$ with a mean

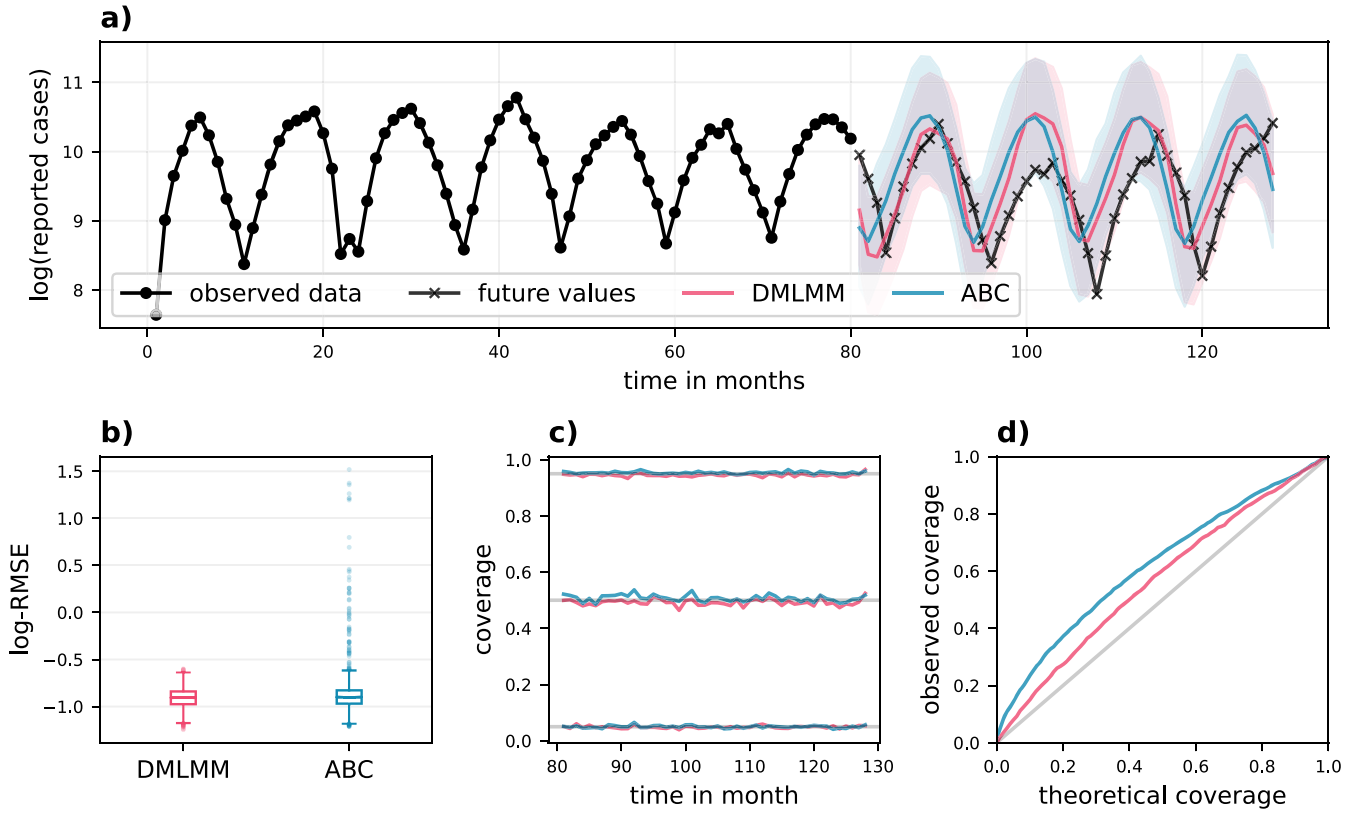

FIGURE 3 | Malaria data. (a) Prediction of the real, observed time series on registered malaria cases in Afghanistan. Shown is the predictive mean (bold) as well as a 95% credible interval for DMFA (red) and ABC (blue). (b) Boxplots for the logarithmic RMSE across 2500 independent realizations from the true model for DMFA (left) and ABC (right). (c) Observed coverage rates for pointwise 95%, 50% and 5% credible intervals for DMFA (red) and ABC (blue). (d) Observed coverage of elliptical credible sets from the 48-dimensional posterior predictive distribution $p(y_{t+1:T}|y_{1:t})$.

of 0.41 (ABC: 0.43), median of 0.4 (ABC: 0.41) and a standard deviation of 0.04 (ABC: 0.2) across all repetitions from the test data, as summarized in Figure 3b. Here, $\hat{y}_{t'}$ are the posterior predictive mean estimates for $y_{t'}$. While the pointwise credible intervals for both methods seem very well calibrated at levels 0.05,0.5,0.95 (Figure 3c), observed coverage rates of elliptical credible sets from the 48-dimensional predictive distribution $p(y_{81:128}|y_{1:80})$ are closer to the nominal levels for the DMLMM as shown in Figure 3d.

4.2.3 | Prior-Data Conflict

Recently, Nott et al. [69] proposed a method for detecting prior-data conflicts in Bayesian models based on comparing prior-to-posterior Rényi divergences of the observed data with the prior-to-posterior divergence under the prior predictive distribution for the data. Since the marginal distribution $p(y_{t+1:T})$ acts as a prior to the implicit likelihood $p(y_{1:t}|y_{t+1:T})$, these checks translate directly to the predictive model described above. A tail probability for a model check can be computed as $p = \mathbb{P}\left[G(y_{1:t}) \geq G\left(y_{1:t}^{(\text{obs})}\right)\right]$, where $G(y_{1:t}) = \mathcal{D}_{\text{KL}}\left[p(y_{t+1:T}|y_{1:t}) \,\middle\|\, p(y_{t+1:T})\right]$. A small p value indicates that the observed data is surprising under the assumed model and Chakraborty et al. [70] discuss how p can be estimated in likelihood-free models through GMM-approximations. It is straightforward to translate their approach to the DMLMM, as the DMFA prior allows for closed form GMM approximations of all quantities necessary to calculate p. The tail probability is estimated as $p = 0.0024$ indicating that the latent ODE model might need to be reexamined. This result is in line with the posterior predictive checks for the malaria data conducted in Alahmadi et al. [62].

5 | Simulation

To further illustrate in which scenarios the deep structure of the DMLMM is beneficial, we consider three distinct simulation set-ups motivated by real-world applications. We compare our DMLMM method to several established benchmarks, including a (non-deep) mixture of linear mixed models fitted by expectation maximization (MLMM), a random coefficient model (LMM), a mixture of linear models (MLM), functional principal component analysis (FPCA), and a latent Gaussian process with subject-specific auto-regressive innovations (GPAR) [29].

5.1 | Simulation Design

We consider the following three different data generating processes (DGPs). For each DGP, we draw 250 independent datasets.

DGP 1: We reanalyze the simulation study conducted in Wang et al. [71]. In particular, for $i = 1, \ldots, 600$, draw $0 \leq t_1, \ldots, <$

$t_{10} \leq 1$ uniformly on $[0,1]$, $g_i \sim \mathcal{U}\{-1,1\}$ and $\xi_{i1} \sim \mathcal{N}\left(0,0.1^2\right)$, $\xi_{i2} \sim \mathcal{N}\left(0,0.045^2\right)$, $\xi_{i3} \sim \mathcal{N}\left(0,0.01^2\right)$, $\xi_{i4} \sim \mathcal{N}\left(0,0.001^2\right)$. Then,

$$y_{ij} = g_i \sin\left(4\pi t_{ij}\right) + \sqrt{2}\sum_{k=1}^{4} \xi_{ik} \sin\left(k\pi t_{ij}\right) + \varepsilon_{ij},$$

where $\varepsilon_{ij} \sim \mathcal{N}\left(0,0.3^2\right)$, $j = 1, \ldots, 10$. This data set contains only two groups with means $\pm \sin(4\pi t)$ and observation specific functional errors $\sqrt{2}\sum_{k=1}^{4} \xi_{ik} \sin(k\pi t)$. Observation specific errors based on a truncated Karhunen-Loève expansion are often considered in the analysis of biomedical functional data [72].

DGP 2: We consider $i = 1, \ldots, 100$ observations of the form $y_{ij} = f_i\left(t_{ij}\right)$ for $n_i \sim \mathcal{U}\{15, 16, \ldots, 25\}$ random time points $t_{ij} \in [10, 20]$, where $f_i(t)$ is a solution to the following system of stochastic differential equations describing a Van der Pol oscillator

$$\frac{d}{dt} f(t) = g(t) + 0.5 \frac{d}{dt} W_{if}(t),$$
$$\frac{d}{dt} g(t) = \theta_i\left(1 - f(t)^2\right) g(t) - f(t) + 0.5 \frac{d}{dt} W_{ig}(t),$$

with $f(0) = 1$, $g(0) = 0.1$. W_{ig} and W_{if} are independent Brownian motions incorporating complex randomness into the observations. This induces a complex dependence structure between nearby time points, for which the DMLMM is misspecified. Additionally, the parameter $\log\left(\theta_i\right) \sim \mathcal{U}(1,5)$ has a continuous prior so that the observations cannot be easily separated into distinct groups. DGP 2 has a similar structure to the malaria model analyzed in Section 4.2.

DGP 3: This DGP is motivated by missing value imputation in time course gene expression studies and related to experiments conducted by Mao and Nott [73]. Let

$$y_{ij} = \beta_{i1} \cos\left(w_{i1}\pi \frac{j-1}{39}\right) + \beta_{i2} \sin\left(w_{i2}\pi \frac{j-1}{39}\right) + \varepsilon_{ij},$$
$$i = 1, \ldots, 120; j = 1, \ldots, 40,$$

where $\varepsilon_{ij} \sim \mathcal{N}\left(0,0.1^2\right)$ is iid noise, and $\beta_{i1}, \beta_{i2} \sim \mathcal{U}\{1,0.1\}$, $w_{i1} \sim \mathcal{U}\{1,2,3\}$, $w_{i2} \sim \mathcal{U}\{7,8,9\}$ are parameters controlling the temporal trend. In each row of the matrix $\left(y_{ij}\right)_{ij}$ 15 to 20 randomly selected data points are removed. This data set contains 36 clusters, some of which are difficult to distinguish, and a comparable small number of observations.

5.2 | Results

We consider $d = 10$ basis functions for each DGP. Here, 1000 iterations of the SGA algorithm for DMLMM take about 4 min on a standard laptop. Figure 4 shows simulations from the three DGPs. The GMM structure of the DMLMM approximation facilitates an implicit clustering of the subjects y_i and the clustering for one run is highlighted by color. Notably, DMLMM recovers the two clusters for DGP 1 well. For both DGP 2 and DGP 3 a large number of Gaussian components is utilized. There is no ground truth clustering for DGP 2 available, but DMLMM groups trajectories with similar shapes in a meaningful way. Performance is evaluated in terms of the RMSE for the predictive distributions $p\left(\tilde{y}|y_i\right)$ and the negative log-score $-\log\left(p\left(y_i\right)\right)$ evaluated on an additional hold-out test set. The summarized results in Table 1 indicate that DMLMM performs robustly across all datasets and is competitive when compared to benchmark methods. On DGP 1 DMLMM is slightly outperformed by the non-deep MLMM. DGP 1 has only two Gaussian components, so the deep structure of DMLMM might not be fully leveraged. Conversely, on the complexer data sets DGP 2 and DGP 3 DMLMM is the best performing method in terms of the negative log-score. Both, DGP 2 and DGP 3 exhibit intricate temporal trends necessitating a complex random effects distribution. This is precisely the scenario DMLMM is tailored for. DMLMM has superior performance in terms of density estimation for unobserved data, as measured by the log-score. This indicates that the DMLMM predictive distributions are useful for capturing predictive uncertainty, which is important for both prediction as well as other purposes such as the prior-data conflict checks considered in Section 4.2.

6 | Conclusion and Discussion

In this paper, we have introduced the DMLMM, which leverages the DMFA model as a prior for the random effects distribution. Our approach complements existing literature on models for complex longitudinal data, and it is particularly suited for high-dimensional settings. We demonstrate the effectiveness

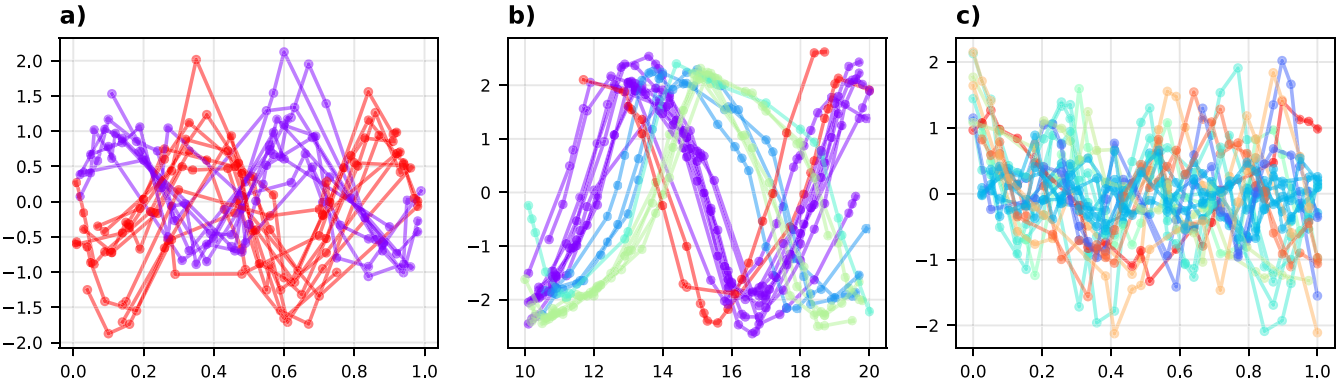

FIGURE 4 | Simulation. Twenty draws from DGP 1—DGP 3 (a–c). The colors correspond to the implicit clustering by DMLMM. Trajectories from the same cluster have the same color.

TABLE 1 | Simulation.

	DGP 1		DGP 2		DGP 3	
	Log-RMSE	**Neg. log-score**	**Log-RMSE**	**Neg. log-score**	**Log-RMSE**	**Neg. log-score**
DMLMM	−1.52 (0.33)	7.38 (0.44)	**−1.65 (0.46)**	**6.60 (1.54)**	−1.09 (0.19)	**14.94 (0.88)**
MLMM	**−1.62 (0.27)**	**6.31 (0.18)**	−1.57 (0.43)	9.87 (3.99)	−1.05 (0.20)	16.63 (1.13)
LMM	−1.18 (0.37)	8.85 (0.06)	−1.53 (0.49)	9.20 (0.54)	−0.94 (0.23)	20.43 (0.51)
MLM	−1.42 (0.50)	8.15 (0.21)	−0.75 (0.46)	17.89 (1.35)	−0.84 (0.32)	18.91 (0.89)
FPCA	−1.11 (0.23)	8.98 (0.09)	−0.95 (0.31)	16.64 (0.86)	−0.69 (0.37)	21.61 (0.59)
GPAR	−0.90 (0.28)	12.79 (0.09)	−1.37 (0.52)	31.43 (0.61)	**−1.19 (0.26)**	24.10 (0.56)

Note: Log-RMSE and negative log-score values for the three DGPs (columns) for the five benchmark methods considered. The mean and the standard deviation (in brackets) across the 250 repetitions are reported rounded to two digits. Bold values indicate the lowest mean value across each column.

of the approach in simulations and biomedical applications in various scenarios, including within-subject prediction for unbalanced longitudinal data, LFI, and missing data imputation. Our DMLMM outperforms existing methods in these applications. Although our focus has been on longitudinal data analysis, the DMLMM framework can be applied in other domains, including functional data analysis and Bayesian nonparametrics, and it is a flexible model for researchers across different fields. Although we have focused on temporal trends, many applications involve covariates that can influence the response. Extending the DMLMM to accommodate covariate-dependent effects is a further direction for future research.

Acknowledgments

Nadja Klein acknowledges support through the Emmy Noether grant KL 3037/1-1 of the German research foundation (DFG). The work of Lucas Kock and Nadja Klein was supported by the Volkswagenstiftung (grant: 96932). David Nott's research was supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 2 (MOE-T2EP20123-0009), and he is affiliated with the Institute of Operations Research and Analytics at the National University of Singapore. Open Access funding enabled and organized by Projekt DEAL.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All data used in this manuscript is openly available. The CD4 data and the gene expression data are available in the R-packages timereg at 10.32614/CRAN.package.timereg and GeneCycle at 10.32614/CRAN.package.GeneCycle respectively. The malaria transmission data is available as electronic Supporting Information to Alahmadi et al. at 10.1098/rsos.191315.

References

1. J. Ren, S. Tapert, C. C. Fan, and W. K. Thompson, "A Semi-Parametric Bayesian Model for Semi-Continuous Longitudinal Data," *Statistics in Medicine* 41, no. 13 (2022): 2354–2374, https://doi.org/10.1002/sim.9359.

2. Y. Cao, H. Allore, B. Vander Wyk, and R. Gutman, "Review and Evaluation of Imputation Methods for Multivariate Longitudinal Data With Mixed-Type Incomplete Variables," *Statistics in Medicine* 41, no. 30 (2022): 5844–5876, https://doi.org/10.1002/sim.9592.

3. G. Verbeke and E. Lesaffre, "A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population," *Journal of the American Statistical Association* 91, no. 433 (1996): 217–221.

4. Z. Bar-Joseph, G. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon, *A New Approach to Analyzing Gene Expression Time Series Data* (Association for Computing Machinery, 2002), 39–48.

5. Y. Luan and H. Li, "Clustering of Time-Course Gene Expression Data Using a Mixed-Effects Model With B-Splines," *Bioinformatics* 19, no. 4 (2003): 474–482.

6. L. X. Qin and S. G. Self, "The Clustering of Regression Models Method With Applications in Gene Expression Data," *Biometrics* 62, no. 2 (2006): 526–533.

7. G. M. James and C. A. Sugar, "Clustering for Sparsely Sampled Functional Data," *Journal of the American Statistical Association* 98, no. 462 (2003): 397–408.

8. G. Celeux, O. Martin, and C. Lavergne, "Mixture of Linear Mixed Models for Clustering Gene Expression Profiles From Repeated Microarray Experiments," *Statistical Modelling* 5, no. 3 (2005): 243–267.

9. S. K. Ng, G. J. McLachlan, K. Wang, L. Ben-Tovim Jones, and S. W. Ng, "A Mixture Model With Random-Effects Components for Clustering Correlated Gene-Expression Profiles," *Bioinformatics* 22, no. 14 (2006): 1745–1752.

10. S. L. Tan and D. J. Nott, "Variational Approximation for Mixtures of Linear Mixed Models," *Journal of Computational and Graphical Statistics* 23, no. 2 (2014): 564–585.

11. T. Scharl, B. Grün, and F. Leisch, "Mixtures of Regression Models for Time Course Gene Expression Data: Evaluation of Initialization and Random Effects," *Bioinformatics* 26, no. 3 (2010): 370–377.

12. C. Pfeifer, "Classification of Longitudinal Profiles Based on Semi-Parametric Regression With Mixed Effects," *Statistical Modelling* 4, no. 4 (2004): 314–323.

13. G. Coke and M. Tsao, "Random Effects Mixture Models for Clustering Electrical Load Series," *Journal of Time Series Analysis* 31, no. 6 (2010): 451–464.

14. J. M. Chiou and P. L. Li, "Functional Clustering and Identifying Substructures of Longitudinal Data," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 69, no. 4 (2007): 679–699.

15. J. Jacques and C. Preda, "Model-Based Clustering for Multivariate Functional Data," *Computational Statistics & Data Analysis* 71 (2014): 92–106.

16. I. C. McDowell, D. Manandhar, C. M. Vockley, A. K. Schmid, T. E. Reddy, and B. E. Engelhardt, "Clustering Gene Expression Time Series Data Using an Infinite Gaussian Process Mixture Model," *PLoS Computational Biology* 14, no. 1 (2018): e1005896.

17. J. Q. Shi and B. Wang, "Curve Prediction and Clustering With Mixtures of Gaussian Process Functional Regression Models," *Statistics and Computing* 18, no. 3 (2008): 267–283.

18. N. A. Heard, C. C. Holmes, and D. A. Stephens, "A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes," *Journal of the American Statistical Association* 101, no. 473 (2006): 18–29.

19. J. G. Booth, G. Casella, and J. P. Hobert, "Clustering Using Objective Functions and Stochastic Search," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 70, no. 1 (2008): 119–139.

20. P. Lenk and W. DeSarbo, "Bayesian Inference for Finite Mixtures of Generalized Linear Models With Random Effects," *Psychometrika* 65, no. 1 (2000): 93–119.

21. C. Proust and H. Jacqmin-Gadda, "Estimation of Linear Mixed Models With a Mixture of Distribution for the Random Effects," *Computer Methods and Programs in Biomedicine* 78, no. 2 (2005): 165–173, https://doi.org/10.1016/j.cmpb.2004.12.004.

22. D. K. Pauler and N. M. Laird, "A Mixture Model for Longitudinal Data With Application to Assessment of Noncompliance," *Biometrics* 56, no. 2 (2000): 464–472.

23. D. l R. Cruz-Mesa, F. A. Quintana, and G. Marshall, "Model-Based Clustering for Longitudinal Data," *Computational Statistics and Data Analysis* 52, no. 3 (2008): 1441–1457.

24. X. Bai, K. Chen, and W. Yao, "Mixture of Linear Mixed Models Using Multivariate t Distribution," *Journal of Statistical Computation and Simulation* 86, no. 4 (2016): 771–787, https://doi.org/10.1080/00949655.2015.1036431.

25. C. A. Bush and S. N. MacEachern, "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika* 83, no. 2 (1996): 275–285.

26. K. P. Kleinman and J. G. Ibrahim, "A Semiparametric Bayesian Approach to the Random Effects Model," *Biometrics* 54, no. 3 (1998): 921–938.

27. P. Müller and G. L. Rosner, "A Bayesian Population Model With Hierarchical Mixture Priors Applied to Blood Count Data," *Journal of the American Statistical Association* 92, no. 440 (1997): 1279–1292.

28. F. Heinzl and G. Tutz, "Clustering in Linear Mixed Models With Approximate Dirichlet Process Mixtures Using EM Algorithm," *Statistical Modelling* 13, no. 1 (2013): 41–67.

29. F. Sigrist, "Latent Gaussian Model Boosting," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 2 (2023): 1894–1905, https://doi.org/10.1109/TPAMI.2022.3168152.

30. P. Kilian, S. Ye, and A. Kelava, "Mixed Effects in Machine Learning–A Flexible mixedML Framework to Add Random Effects to Supervised Machine Learning Regression," *Transactions on Machine Learning Research* (2023), https://openreview.net/forum?id=MKZyHtmfwH.

31. M. N. Tran, N. Nguyen, D. Nott, and R. Kohn, "Bayesian Deep Net GLM and GLMM," *Journal of Computational and Graphical Statistics* 29, no. 1 (2020): 97–113, https://doi.org/10.1080/10618600.2019.1637747.

32. F. Mandel, R. P. Ghosh, and I. Barnett, "Neural Networks for Clustered and Longitudinal Data Using Mixed Effects Models," *Biometrics* 79, no. 2 (2021): 711–721, https://doi.org/10.1111/biom.13615.

33. A. Cascarano, J. Mur-Petit, J. Hernandez-Gonzalez, et al., "Machine and Deep Learning for Longitudinal Biomedical Data: A Review of Methods and Applications," *Artificial Intelligence Review* 56, no. Suppl 2 (2023): 1711–1771.

34. C. Viroli and G. J. McLachlan, "Deep Gaussian Mixture Models," *Statistics and Computing* 29, no. 1 (2019): 43–51.

35. Z. Ghahramani and M. Beal, "Variational Inference for Bayesian Mixtures of Factor Analysers," in *Advances in Neural Information Processing Systems*, vol. 12, ed. S. Solla, T. Leen, and K. Müller (MIT Press, 2000), 449–455.

36. G. J. McLachlan, D. Peel, and R. Bean, "Modelling High-Dimensional Data by Mixtures of Factor Analyzers," *Computational Statistics & Data Analysis* 41, no. 3–4 (2003): 379–388.

37. Y. Tang, R. Salakhutdinov, and G. Hinton, *Deep Mixtures of Factor Analysers* (ICML'12. Omnipress, 2012), 1123–1130.

38. v. d A. Oord and B. Schrauwen, "Factoring Variations in Natural Images With Deep Gaussian Mixture Models," in *Advances in Neural Information Processing Systems*, vol. 27, ed. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., 2014).

39. X. Yang, K. Huang, and R. Zhang, "Deep Mixtures of Factor Analyzers With Common Loadings: A Novel Deep Generative Approach to Clustering," in *Neural Information Processing*, ed. D. Liu, S. Xie, Y. Li, D. Zhao, and E. S. M. El-Alfy (Springer International Publishing, 2017), 709–719.

40. J. Li, "Clustering Based on a Multilayer Mixture Model," *Journal of Computational and Graphical Statistics* 14, no. 3 (2005): 547–568.

41. G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün, "Identifying Mixtures of Mixtures Using Bayesian Estimation," *Journal of Computational and Graphical Statistics* 26, no. 2 (2017): 285–295.

42. L. Kock, N. Klein, and D. J. Nott, "Variational Inference and Sparsity in High-Dimensional Deep Gaussian Mixture Models," *Statistics and Computing* 32, no. 5 (2022): 70.

43. A. Tancredi, "Approximate Bayesian Inference for Discretely Observed Continuous-Time Multi-State Models," *Biometrics* 75, no. 3 (2019): 966–977, https://doi.org/10.1111/biom.13019.

44. S. Clémençon, A. Cousien, M. D. Felipe, and V. C. Tran, "On Computer-Intensive Simulation and Estimation Methods for Rare-Event Analysis in Epidemic Models," *Statistics in Medicine* 34, no. 28 (2015): 3696–3713, https://doi.org/10.1002/sim.6596.

45. F. V. Bonassi, L. You, and M. West, "Bayesian Learning From Marginal Data in Bionetwork Models," *Statistical Applications in Genetics and Molecular Biology* 10, no. 1 (2011): 49.

46. F. V. Bonassi and M. West, "Sequential Monte Carlo With Adaptive Weights for Approximate Bayesian Computation," *Bayesian Analysis* 10, no. 1 (2015): 171–187.

47. F. Forbes, H. D. Nguyen, T. T. Nguyen, and J. Arbel, "Approximate Bayesian Computation With Surrogate Posteriors," (2021), Preprint hal-03139256.

48. d F H. Alencar, L. A. Matos, and V. H. Lachos, "Finite Mixture of Censored Linear Mixed Models for Irregularly Observed Longitudinal Data," *Journal of Classification* 39, no. 3 (2022): 463–486.

49. T. I. Lin and W. L. Wang, "Multivariate Contaminated Normal Linear Mixed Models Applied to Alzheimer's Disease Study With Censored and Missing Data," *Statistical Methods in Medical Research* 34, no. 3 (2025): 490–507.

50. S. Frühwirth-Schnatter, D. Hosszejni, and H. F. Lopes, "Sparse Bayesian Factor Analysis When the Number of Factors Is Unknown," *Bayesian Analysis* 1, no. 1 (2024): 1–31.

51. C. M. Carvalho and G. Polson, "The Horseshoe Estimator for Sparse Signals," *Biometrika* 97, no. 2 (2010): 465–480.

52. J. Rousseau and K. Mengersen, "Asymptotic Behaviour of the Posterior Distribution in Overfitted Mixture Models," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 73, no. 5 (2011): 689–710.

53. D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association* 112, no. 518 (2017): 859–877.

54. M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic Variational Inference," *Journal of Machine Learning Research* 14, no. 1 (2013): 1303–1347.

55. S. Amari, "Natural Gradient Works Efficiently in Learning," *Neural Computation* 10, no. 2 (1998): 251–276.

56. M. Selosse, C. Gormley, J. Jacques, and C. Biernacki, "A Bumpy Journey: Exploring Deep Gaussian Mixture Models," (2020).

57. R. A. Kaslow, D. G. Ostrow, R. Detels, J. P. Phair, B. F. Polk, and J. C. R. Rinaldo, "The Multicenter AIDS Cohort Study: Rationale, Organization, and Selected Characteristics of the Participants," *American Journal of Epidemiology* 126, no. 2 (1987): 310–318.

58. J. Fan and J. T. Zhang, "Two-Step Estimation of Functional Linear Models With Applications to Longitudinal Data," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 62, no. 2 (2000): 303–322, https://doi.org/10.1111/1467-9868.00233.

59. C. O. Wu and C. T. Chiang, "Kernel Smoothing on Varying Coefficient Models With Longitudinal Dependent Variable," *Statistica Sinica* 10, no. 2 (2000): 433–456.

60. F. Yao, H. G. Müller, and J. L. Wang, "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association* 100, no. 470 (2005): 577–590.

61. M. Y. Anwar, J. A. Lewnard, S. Parikh, and V. E. Pitzer, "Time Series Analysis of Malaria in Afghanistan: Using ARIMA Models to Predict Future Trends in Incidence," *Malaria Journal* 15, no. 1 (2016): 1–10.

62. A. A. Alahmadi, J. A. Flegg, D. G. Cochrane, C. C. Drovandi, and J. M. Keith, "A Comparison of Approximate Versus Exact Techniques for Bayesian Parameter Inference in Nonlinear Ordinary Differential Equation Models," *Royal Society Open Science* 7, no. 3 (2020): 191315.

63. H. W. Hethcote, "The Mathematics of Infectious Diseases," *SIAM Review* 42, no. 4 (2000): 599–653.

64. R. Xu and Z. Ma, "Global Stability of a SIR Epidemic Model With Nonlinear Incidence Rate and Time Delay," *Nonlinear Analysis: Real World Applications* 10, no. 5 (2009): 3175–3189, https://doi.org/10.1016/j.nonrwa.2008.10.013.

65. H. Wu, D. A. Stephens, and E. E. M. Moodie, "An SIR-Based Bayesian Framework for COVID-19 Infection Estimation," *Canadian Journal of Statistics* 52 (2024): e11817, https://doi.org/10.1002/cjs.11817.

66. L. J. White, R. J. Maude, W. Pongtavornpinyo, et al., "The Role of Simple Mathematical Models in Malaria Elimination Strategy Design," *Malaria Journal* 8 (2009): 1–10.

67. S. Sisson, Y. Fan, and M. Beaumont, *Handbook of Approximate Bayesian Computation*, 1st ed. (Chapman and Hall/CRC, 2018).

68. M. Järvenpää and J. Corander, "On Predictive Inference for Intractable Models via Approximate Bayesian Computation," *Statistics and Computing* 33, no. 2 (2023): 42.

69. D. J. Nott, X. Wang, M. Evans, and B. G. Englert, "Checking for Prior-Data Conflict Using Prior-To-Posterior Divergences," *Statistical Science* 35, no. 2 (2020): 234–253, https://doi.org/10.1214/19-STS731.

70. A. Chakraborty, D. J. Nott, and M. Evans, "Weakly Informative Priors and Prior-Data Conflict Checking for Likelihood-Free Inference," (2022), https://doi.org/10.48550/arXiv.2202.09993.

71. Q. Wang, A. Farahat, C. Gupta, and S. Zheng, "Deep Time Series Models for Scarce Data," *Neurocomputing* 456 (2021): 504–518.

72. N. Margaritella, V. Inácio, and R. King, "Parameter Clustering in Bayesian Functional Principal Component Analysis of Neuroscientific Data," *Statistics in Medicine* 40, no. 1 (2021): 167–184, https://doi.org/10.1002/sim.8768.

73. Y. Mao and D. J. Nott, "Bayesian Clustering Using Random Effects Models and Predictive Projections," (2021), https://doi.org/10.48550/arXiv.2106.15847.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Additional supporting information including the formal mathematical description of the DMLMM (A), details on the variational Bayes approach for posterior inference (B), and an application to missing data imputation for gene expression data not discussed within the main text (C) may be found in the online version of the article at the publisher's website. Python code for the DMLMM is publicly available github.com/kocklucx/DMLMM.