Bayesian Analysis (2023) **TBA**, Number TBA, pp. 1–31

Bayesian Semiparametric Hidden Markov Tensor Models for Time Varying Random Partitions with Local Variable Selection*

Giorgio Paulon[†], Peter Müller[‡], and Abhra Sarkar[§,¶]

Abstract. We present a flexible Bayesian semiparametric mixed model for longitudinal data analysis in the presence of potentially high-dimensional categorical covariates. Building on a novel hidden Markov tensor decomposition technique, our proposed method allows the fixed effects components to vary between dependent random partitions of the covariate space at different time points. The mechanism not only allows different sets of covariates to be included in the model at different time points but also allows the selected predictors' influences to vary flexibly over time. Smooth time-varying additive random effects are used to capture subject specific heterogeneity. We establish posterior convergence guarantees for both function estimation and variable selection. We design a Markov chain Monte Carlo algorithm for posterior computation. We evaluate the method's empirical performances through synthetic experiments and demonstrate its practical utility through real world applications.

Keywords: B-splines, factorial hidden Markov models (fHMM), higher order singular value decomposition (HOSVD), local variable selection, longitudinal data, partition models.

1 Introduction

We propose a novel statistical framework for modeling longitudinally varying continuous response trajectories in the presence of categorical covariates. The main innovations include a parsimonious dynamic random partition model for the covariate space using a two-layer construction that builds on a novel hidden Markov tensor decomposition mechanism. This allows for time-varying variable selection while also parsimoniously representing higher order interactions between the selected predictors. Adapting the model further in a novel way to mix with B-spline bases then allows us to flexibly characterize the longitudinal evolution of the associated parameter trajectories, including their data-adaptive forking and merging over time.

*This work was supported in part by National Science Foundation grants DMS 1952679 to Mueller and DMS 1953712 to Sarkar.

[†]Berry Consultants, LLC, 3345 Bee Caves Road, Suite 201, Austin, TX 78746-6692, USA, giorgio@berryconsultants.net

[‡]Department of Mathematics, University of Texas at Austin, 2515 Speedway C1200, Austin, TX 78712-1202, USA, pmueller@math.utexas.edu

[§]Department of Statistics and Data Sciences, University of Texas at Austin, 2317 Speedway D9800, Austin, TX 78712-1823, USA, abhra.sarkar@utexas.edu

[¶]Corresponding Author.

 https://doi.org/10.1214/23-BA1383

The settings analyzed here may be viewed as longitudinal adaptations of static analysis of variance (ANOVA) designs and hence are very generic and almost ubiquitously encountered in modern scientific research in many diverse fields, examples from recent statistics literature including pharmacodynamics (De Iorio et al., 2004), mass spectroscopy (Morris and Carroll, 2006), early pregnancy loss studies (MacLehose and Dunson, 2009), etc. In such scenarios, assessing the local variations in the response profiles, including especially how the associated predictors might influence the response differently in different stages of the longitudinal process, can provide valuable insights into the underlying data generating mechanisms. Figure 1 shows a synthetic illustrative example where the mean profiles of a continuous response y vary smoothly over time. Two associated covariates, x_1 and x_3, from a set of ten total available $\{x_1, \ldots, x_{10}\}$ are important and they jointly influence the response means differently in different longitudinal stages. The goal of this article is to understand such complex dynamics from data on response and covariate values.

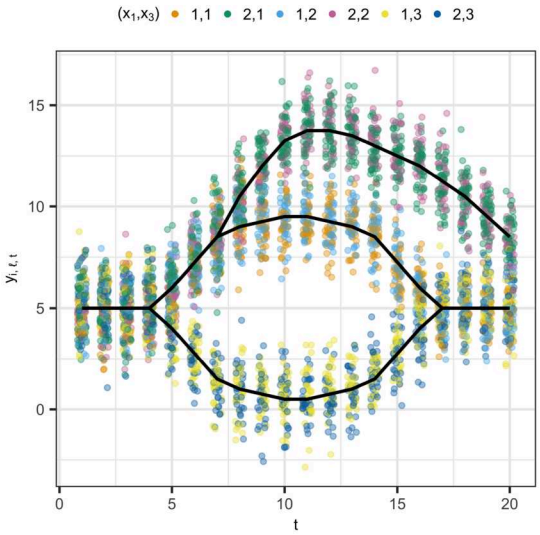

Figure 1: A synthetic scenario with ten covariates $(x_1, \ldots, x_{10})$. No covariate is globally important but (x_1, x_3) are locally important: They have no influence on y for $t \in [1, 4]$ but have a complex pattern of joint influence on y for $t \in (4, 20]$. The response values $y_{i,\ell,t}$ are represented here as slightly jittered points for all combination of the levels of the significant predictors (x_1, x_3). The true underlying mean functions are superimposed (black lines).

Existing Methods The literature on longitudinal data analysis is really vast (see, e.g., books by Diggle et al., 2002; Singer et al., 2003; Fitzmaurice et al., 2008, and the references therein). Bayesian methods for longitudinal data have also been extensively developed (Daniels and Pourahmadi, 2002; Chib and Hamilton, 2002; Li et al., 2010; Müller et al., 2013; Quintana et al., 2016, etc.). However, the problem of characterizing

dynamically varying variable importance in such settings has not received much attention. This article presents a novel Bayesian semiparametric method that addresses such needs.

Our work in this direction was inspired by the existing sparse literature on local clustering in functional data (Duan et al., 2007; Petrone et al., 2009; Nguyen, 2010; Nguyen and Gelfand, 2011). These Bayesian nonparametric approaches assume that the mean functions can be represented by a smaller set of canonical curves that are in turn modeled, for instance, as independently and identically distributed (i.i.d.) realizations from a stationary Gaussian process. Gelfand et al. (2005) specify an infinite mixture of these global functional atoms in which each observation is a noisy realization around a draw from a set of canonical curves. Such approaches however only allow for curves that are either completely different or completely identical across the entire functional domain, capturing only their global difference patterns. Many applications, however, involve data exhibiting local heterogeneity. Local clustering in such cases could greatly improve estimation and prediction, borrowing information across locally homogenous regions, as well as interpretability and inference, providing potentially interesting insights into the underlying causes of local heterogeneity. Toward this goal, Duan et al. (2007) proposed a solution by defining a stick-breaking construction at each location, which allows for local selection of curves. Petrone et al. (2009) assumed that the individual curves can be obtained as hybrid species defined as recombinations of different portions of the canonical curves. Both these approaches define the local allocation rules using a single hidden labeling process that indicates which canonical curve is chosen at each time stamp. Additional challenges are represented by the choice of functional dependence in the labeling process, whose theoretical properties have been studied by Nguyen and Gelfand (2011). Suarez and Ghosal (2016) proposed an alternative approach, using independent priors at different time points to cluster wavelet basis coefficients first, but then using these local features to find global functional clusters as the final inference goal.

The approaches mentioned above have limitations that deserve attention. First, defining the mean functions as recombinations of canonical curves implies that these curves are discontinuous, which can be an impractical assumption in most applications. Continuous curves may be desired, for instance, in dose-response relationships (De Iorio et al., 2004). Second, the inclusion of covariates in these models has only been accomplished via an additive term in the mean function. Even when flexible random effects are used, the linearity assumption of the covariate effects can be quite restrictive in practice. Furthermore, the problem of dynamically characterizing variable importance in these settings has not been addressed.

Alternative approaches to model time-varying predictor effects and interactions in longitudinal data include tree based methods. Bayesian additive regression trees (BART) (Chipman et al., 2010) perform well when the regression function consists of low order nonlinear interactions. With time as an additional covariate, these models can be adapted to capture longitudinally varying influences of the predictors (Sparapani et al., 2016). Separate ideas involving a single tree have also been proposed (Taddy et al., 2011; Gramacy et al., 2013) where the tree structure evolves when new data streams become available. Linero and Yang (2018) and Starling et al. (2020) proposed

smoothing the covariate effects which yields more appropriate results when the outcome is expected to vary smoothly over time. These models, albeit flexible, do not directly assess the local influence of each individual predictor but measure variable importance by calculating their contributions to reducing the in-sample mean squared error. With such heavy emphasis on prediction, they often include many weakly informative or even spurious predictors in the ensembles and lack parsimony and interpretability as a result (Breiman, 2001; Efron, 2020).

Yet another related strategy comprises varying coefficients (VC) models where the regression coefficients are allowed to smoothly vary over a set of chosen modifiers (Hastie and Tibshirani, 1993). VC models have been adapted to longitudinal data by considering time as the only modifier (Hoover et al., 1998). More recently BART priors (Deshpande et al., 2020) and variable selection techniques (Koslovsky et al., 2020) have also been adapted to VC settings. While VC models allow for an easy assessment of the predictors' importance, they are restricted in their ability to accommodate interactions between predictors. For example, for p categorical predictors $x_j \in \{1, \ldots, x_{j,\max}\}, j = 1, \ldots, p$, it is necessary to include $\sum_{j=1}^{p}(x_{j,\max} - 1)$ dummy variables for the main effects, $\sum_{j_1 \neq j_2}(x_{j_1,\max} - 1)(x_{j_2,\max} - 1)$ for the first order interactions, and so on.

Our Proposed Approach We propose a longitudinal functional mixed effects model that combines predictive power and interpretability by addressing the limitations of the local clustering approaches cited above. Most existing methods imply a tension between the main goals of statistical analysis (Breiman, 2001), namely estimation, attribution and prediction (Efron, 2020). Our proposed approach tries to strike a balance – it is highly flexible, being able to accommodate higher order interactions between the predictors, but also favors parsimony, modeling these complex effects implicitly and compactly, while also allowing some ease of interpretation, including explicitly encoding each predictor's varying overall significance at different time points. Our method also comes with theoretical guarantees for both function estimation and variable selection.

The construction of our proposed model proceeds by characterizing the longitudinal evolution of both the predictor dependent fixed effects and the subject specific random effects as flexible functions of time (Ramsay and Silverman, 2007; Morris, 2015; Wang et al., 2016) modeled by mixtures of locally supported spline bases (de Boor, 1978; Eilers and Marx, 1996). The fixed effects model spline coefficients are allowed to vary with the associated predictors' level combinations, thereby accommodating all order interactions between them. Structuring these coefficients as multi-way tensors and applying a novel higher order singular value (HOSVD) type decomposition (Tucker, 1966; De Lathauwer et al., 2000; Kolda and Bader, 2009), we reduce the high-dimensional problem of modeling the complex joint influence of many different predictors to that of estimating much smaller-dimensional core coefficients. In effect, this induces a local partitioning of the joint covariate space such that the different predictor level combinations belonging to the same partition set will have a similar effect on the response variable. The local partitions constructed this way can in fact be indexed by combinations of separate latent allocation indicators, one for each level of the associated categorical predictors, facilitating separate assessment of the influences of each individual covariate

(Sarkar and Dunson, 2016). To induce dependence between adjacent local partitions, we allow the latent allocation indicators evolve according to a factorial hidden Markov model (fHMM) (Ghahramani and Jordan, 1997). In constructing the model this way, we break free from the assumption of separate canonical curves of the previously existing Bayesian nonparametric literature cited above but allow the dependencies across adjacent temporal locations be further informed by the local partition configurations through a novel conditionally Markov prior on the core spline coefficients, conditional on the partition structure, improving model interpretability and estimation efficiency. The proposed functional approach also has the important advantage of avoiding to have to impute missing data when they are missing under simple mechanisms (Little and Rubin, 2019). We establish theoretical results on posterior consistency of the proposed method for both function estimation and variable selection. We evaluate its numerical performance in simulation experiments where it significantly outperformed its competitors not just on average but also uniformly in all simulation instances. Finally, we illustrate the method's practical performance in real data applications from diverse domains.

The methodology presented here is highly generic and broadly adaptable to diverse problems. For instance, the special case with a single categorical predictor x with a small number of levels was previously adapted to a drift-diffusion model in Paulon et al. (2021). The focus of this article, however, is on developing a general methodology with an emphasis on the multivariate case $(x_1, \ldots, x_p)$ which presents significant and unique additional modeling and computational challenges. Note here that redefining the $\prod_{j=1}^{p} x_{j,\max}$ level combinations of $(x_1, \ldots, x_p)$ as the levels of a new single predictor x, while conceptually straightforward, does not provide a practically effective solution as it does not allow separate characterization of the local importances of the different predictors and, with $\prod_{j=1}^{p} x_{j,\max}$ increasing exponentially fast with p, quickly becomes computationally inefficient even in small to moderate dimensional problems. The strategy is practically useless, for instance, in a real data applications we discuss in Section 6, where $\prod_{j=1}^{p} x_{j,\max} = 580{,}608$. Our proposed formulation based on tensors and their dynamic data adaptive HOSVD, in contrast, not only provides a flexible and highly efficient tool for dimension reduction and simultaneous variable selection but also does this locally at each time point while borrowing information across a number of levels.

Our proposed approach does not partition the response values directly, which has been considered by many in the static setting (Hartigan, 1990; Denison et al., 2002; Quintana and Iglesias, 2003) and some in the dynamic setting (Barry and Hartigan, 1992; Page et al., 2020). Instead, we partition the covariate space according to their influences on the response. Separately, the literature on HMMs and fHMMs is also vast (Rabiner, 1989; Scott, 2002; Frühwirth-Schnatter, 2006; Zucchini et al., 2017). To our knowledge, however, they have never been adapted in the novel ways proposed in this article to dynamic variable selection problems. There is also a growing body of literature on regression methods for tensor valued predictors with tensor factorization techniques used as a dimension reduction tool. These methods, however, apply tensor factorizations with all continuous components, where the general Tucker decomposition runs into identifiability and interpretability problems. To avoid these issues, the literature has focused on parallel factor (PARAFAC) type decomposition (see, e.g., Guhaniyogi et al.

2017; Papadogeorgou et al. 2019, etc.), a much simpler but also much restrictive special case of the Tucker. Aside from the development of sophisticated dependence models for the tensor components in a longitudinal setting, our proposal is also novel in that we employ a compact HOSVD, a flexible but interpretable version of the Tucker decomposition, where the core tensors take continuous values but the mode matrices comprise specially structured binary elements, resulting in interpretable partition structures that allow dynamic variable selection.

Outline of the Article The rest of this article is organized as follows. Section 2 develops the generic longitudinal mixed model framework. Section 3 develops Markov chain Monte Carlo (MCMC) algorithms for posterior computation. Section 4 establishes posterior convergence guarantees for the proposed model, for both function estimation and variable selection. Section 5 presents the results of simulation experiments. Section 6 presents real data applications. Section 7 contains concluding remarks. Substantive additional details are presented in the supplementary materials (Paulon et al., 2023a,b).

2 Longitudinal Functional Mixed Model

In this section, we develop a novel generic statistical framework for longitudinal functional mixed model (LFMM), where a response y is generated under the influence of p categorical predictors $x_j \in \{1, \ldots, x_{j,\max}\} = \mathcal{X}_j, j = 1, \ldots, p$ longitudinally over time. To be precise, data y_{i,ℓ,t_i}, available for individuals $i \in \{1, \ldots, n\}$ and trials $\ell \in \{1, \ldots, L_{i,t_i}\}$ at time points $t_i \in \{t_{i,1}, \ldots, t_{i,T}\}$, are generated under the influence of the predictors $x_j, j = 1, \ldots, p$. Importantly, we are not only interested in assessing the overall global influences of the predictors but also how they affect the responses locally at various times of the longitudinal studies.

We consider the following generic class of LFMMs

$$\{y_{i,\ell,t} \mid x_{j,i,\ell,t} = x_j, j = 1, \ldots, p\} = f_{x_1,\ldots,x_p}(t) + u_i(t) + \varepsilon_{i,\ell,t}, \quad \varepsilon_{i,\ell,t} \sim f_\varepsilon, \quad (2.1)$$

where $f_{x_1,\ldots,x_p}(t)$ denotes time-varying fixed effects due to associated predictors $\mathbf{x} = (x_1, \ldots, x_p) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_p = \mathcal{X}$, $u_i(t)$ are time-varying subject specific random effects, and $\varepsilon_{i,\ell,t}$ are random errors, i.i.d. from f_ε, satisfying $\mathbb{E}_{f_\varepsilon}(\varepsilon_{i,\ell,t}) = 0$. We assume that $f_{x_1,\ldots,x_p}(t)$ and $u_i(t)$ evolve continuously with time. In this article, we focus on normally distributed errors with an inverse-Gamma prior on the error variance as

$$f_\varepsilon = \text{Normal}(0, \sigma_\varepsilon^2), \quad \sigma_\varepsilon^2 \sim \text{Inv-Ga}(a_\sigma, b_\sigma).$$

For ease of exposition, we assume in (2.1) and henceforth that the data points are measured at a common set of equidistant time points $\{t_1, \ldots, t_T\}$, denoted simply as $\{1, \ldots, T\}$. With some abuse of notation, generic values taken by the response y, the predictors x_j are also denoted by y, x_j, etc. Without loss of generality, we also assume henceforth the same number of replicates $L_{i,t} = L$ for all i, t. To further simplify notation, generic data recording time stamps in $\{1, \ldots, T\}$ as well as other generic time points in $[1, T]$ will both be denoted by t.

For longitudinal data observed on a regular time grid, continuous functional parameter trajectories may still be more appealing and interpretable to a practitioner. A functional approach to modeling longitudinal data also does not require to impute missing data when they are missing at random (Little and Rubin, 2019).

The focus of this article is on continuous responses with categorical predictors. In many applications, the covariates are exogenous, that is, for each i, the $x_{j,i,\ell,t}$'s equal some fixed level x_j for all ℓ, t. When they are time-varying, we assume that all levels of x_j are present in the sample at each t for each j. An easy, highly robust and practically useful approach to include continuous and ordinal predictors in model (2.1) would be to categorize them by binning their values into intervals (e.g., using their quantiles) or by ignoring their order. Non-continuous responses of various types can likewise be conveniently analyzed via latent continuous variable augmentations (Albert and Chib, 1993; Dunson, 2000; Polson et al., 2013).

2.1 Fixed Effects Model

We propose a novel approach to model the latent functions $f_{x_1,\ldots,x_p}(t)$ using basis decomposition methods that allow them to flexibly vary with time t while also locally depending on the predictor combinations $(x_1, \ldots, x_p)$. Specifically, we let

$$f_{x_1,\ldots,x_p}(t) = \sum_{k=1}^{K} \beta_{k,x_1,\ldots,x_p} b_k(t), \tag{2.2}$$

where $\mathbf{b}(t) = \{b_1(t), \ldots, b_K(t)\}^{\mathrm{T}}$ are a set of known locally supported basis functions and $\boldsymbol{\beta}_{x_1,\ldots,x_p} = \{\beta_{1,x_1,\ldots,x_p}, \ldots, \beta_{K,x_1,\ldots,x_p} : (x_1, \ldots, x_p) \in \mathcal{X}\}$ are unknown coefficients to be estimated from the data. We use B-spline bases (de Boor, 1978) which are non-negative, continuous and have desirable local support properties (Figure 2). Allowing the $\beta_{k,x_1,\ldots,x_p}$'s to vary with all predictor combinations $(x_1, \ldots, x_p)$, the model also accommodates all order interactions among the predictors.

While other higher order B-splines can also be used, in this work we use linear B-splines whose local support properties result in locally linear approximations of the fixed effects function (Figure 2). In the following, we use knots at the observed locations, hence $K = T$. This allows local clustering at the set of all observable time points. In the case of irregularly spaced data, a suitable fine grid can be chosen where such inference is desired. As shown in Ruppert (2002), when smoothing is controlled by data adaptive penalty parameters, the number of knots K is not a crucial parameter as long as it is larger than a minimum threshold.

For most practical applications, the size $K \prod_{j=1}^{p} x_{j,\max}$ of the unstructured model (2.2) may be too big to allow efficient model estimation. It is also difficult to assess local influences of the predictors using such unstructured models. A potentially efficient solution that can greatly reduce dimensions while also facilitating the assessment of predictors' importance is to cluster the parameters by allowing them to have common shared values across different predictor combinations. If, e.g., $\boldsymbol{\beta}_{x_1,\ldots,x_{j,1},\ldots,x_p} = \boldsymbol{\beta}_{x_1,\ldots,x_{j,2},\ldots,x_p}$ for all combinations of $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$, then not only have we reduced the number of parameters to be modeled but have also established that the two levels $x_{j,1}$ and $x_{j,2}$ of x_j have no differential effect on the data generating mechanism.

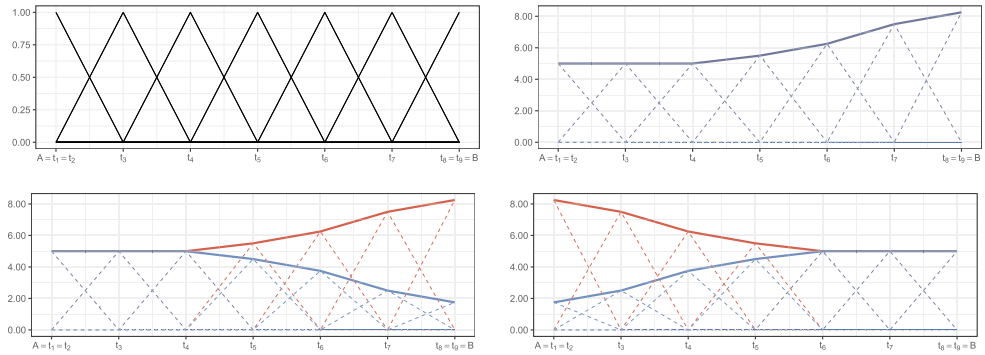

Figure 2: Top left panel: Plot of 7 linear B-splines on an interval $[A, B]$ defined by 9 equidistant knot points that divide $[A, B]$ into $K = 6$ equal subintervals. Top right panel: Example of global clustering of two curves with shared spline coefficients $\boldsymbol{\beta}_{x_1,\ldots,x_p} = \boldsymbol{\beta}_{x'_1,\ldots,x'_p} = (5, 5, 5, 5.5, 6.25, 7.5, 8.25)^{\mathrm{T}}$; the solid lines represent the estimated functions, the dashed lines represent the weighted B-spline bases. Bottom left panel: Example of splitting of two curves with partially shared spline coefficients $\boldsymbol{\beta}_{x_1,\ldots,x_p} = (5, 5, 5, 5.5, 6.25, 7.5, 8.25)^{\mathrm{T}}$, $\boldsymbol{\beta}_{x'_1,\ldots,x'_p} = (5, 5, 5, 4.5, 3.75, 2.5, 1.75)^{\mathrm{T}}$. Bottom right panel: Example of merging of two curves with partially shared spline coefficients $\boldsymbol{\beta}_{x_1,\ldots,x_p} = (8.25, 7.5, 6.25, 5.5, 5, 5, 5)^{\mathrm{T}}$, $\boldsymbol{\beta}_{x'_1,\ldots,x'_p} = (1.75, 2.5, 3.75, 4.5, 5, 5, 5)^{\mathrm{T}}$.

Such global clustering of all elements of $\boldsymbol{\beta}_{x_1,\ldots,x_p}$ together will still be highly restrictive in most practical applications. More realistically, the elements of $\boldsymbol{\beta}_{x_1,\ldots,x_p}$ should be allowed to cluster locally. In the following, we exploit the local support properties of B-splines in a novel way to achieve this desirable property. In principle, other basis decomposition methods whose bases have compact local support can also be used in a similar way.

Dimension Reduction and Local Clustering via HOSVD To achieve simultaneous dimensionality reduction and local clustering, we structure the parameters for different predictor combinations at each location k as a $x_{1,\max} \times \cdots \times x_{p,\max}$ dimensional tensor $\boldsymbol{\beta}_k = \{\beta_{k,x_1,\ldots,x_p} : (x_1, \ldots, x_p) \in \mathcal{X}\}$ and then apply an HOSVD-type (Tucker, 1966; De Lathauwer et al., 2000) factorization, arriving at

$$\{\beta_{k,x_1,\ldots,x_p} \mid z_{j,k}^{(x_j)} = z_{j,k}, j = 1, \ldots, p\} = \beta^{\star}_{k,z_{1,k},\ldots,z_{p,k}}, \tag{2.3}$$

where $z_{j,k}^{(x_j)}$'s are cluster indicator variables associated with each covariate j for its specific value x_j at the knot-location k, and the $\beta^{\star}_{k,z_{1,k},\ldots,z_{p,k}}$'s are the associated unique cluster specific spline coefficients. Our construction using locally supported linear B-splines then implies

$$\{f_{k,x_1,\ldots,x_p} \mid z_{j,k}^{(x_j)} = z_{j,k}, j = 1, \ldots, p\} = \beta^{\star}_{k,z_{1,k},\ldots,z_{p,k}},$$

allowing simple interpretations for the allocation variables $z_{j,k}^{(x_j)}$'s and also easier theoretical treatment and posterior computation.

Let the $z_{j,k}^{(x_j)}$'s take values in $\mathcal{Z}_{j,k} = \{1, \ldots, \ell_{j,k}\}$ for different possible values $x_j \in \mathcal{X}_j = \{1, \ldots, x_{j,\max}\}$.[1] Separately, $z_{j,k}^{(x_j)} \in \mathcal{Z}_{j,k}$ therefore forms $\ell_{j,k} \leq x_{j,\max}$ marginal clusters of the predictor levels $x_j \in \mathcal{X}_j$, while jointly, $(z_{1,k}^{(x_1)}, \ldots, z_{p,k}^{(x_p)}) \in \mathcal{Z}_k = \mathcal{Z}_{1,k} \times \cdots \times \mathcal{Z}_{p,k}$ forms $\prod_{j=1}^{p} \ell_{j,k} \leq \prod_{j=1}^{p} x_{j,\max}$ joint clusters of the predictor level combinations $(x_1, \ldots, x_p) \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$.

To see the HOSVD formulation behind this, note that (2.3) can be rewritten as

$$\{\beta_{k,x_1,\ldots,x_p} \mid z_{j,k}^{(x_j)}, j = 1, \ldots, p\} = \sum_{z_{1,k}} \cdots \sum_{z_{p,k}} \beta^{\star}_{k,z_{1,k},\ldots,z_{p,k}} \prod_{j=1}^{p} 1\{z_{j,k}^{(x_j)} = z_{j,k}\}, \quad (2.4)$$

where $\boldsymbol{\beta}_k^{\star} = \{\beta^{\star}_{k,z_{1,k},\ldots,z_{p,k}} : (z_{1,k}, \ldots, z_{p,k}) \in \mathcal{Z}_k\}$ is a $\ell_{1,k} \times \cdots \times \ell_{p,k}$ dimensional core tensor, $\mathbf{z}_{j,k} = \{1\{z_{j,k}^{(x_j)} = z_{j,k}\} : x_j \in \mathcal{X}_j, z_{j,k} \in \mathcal{Z}_{j,k}\}$ are $x_{j,\max} \times \ell_{j,k}$ dimensional mode matrices (Figure 3).

The allocation variables are assigned probability models supported on $\mathcal{X}_j$, so that the number of distinct values taken on by the $z_{j,k}^{(x_j)}$'s, namely $\ell_{j,k}$, lies between 1 and $|\mathcal{X}_j| = x_{j,\max}$. If $\ell_{j,k} = x_{j,\max}$, the $z_{j,k}^{(x_j)}$'s take on different values for different levels of x_j, implying that the spline coefficients are all different for different levels of x_j at location k. In this case, all levels of x_j differently influence the response generating mechanism at location k. If $\ell_{j,k} \leq x_{j,\max}$, local clustering of the predictor's effects is performed and the problem of modeling the original parameter tensors $\boldsymbol{\beta}_k$ is effectively reduced to that of modeling the smaller-dimensional core tensors $\boldsymbol{\beta}_k^{\star}$. For instance, when $z_{j,k}^{(x_{j,1})} = z_{j,k}^{(x_{j,2})} = z_{j,k}$ for two different levels $x_{j,1}$ and $x_{j,2}$ of the j^{th} predictor x_j, the spline coefficients at location k do not differ between $x_{j,1}$ and $x_{j,2}$, i.e., $\beta_{k,x_1,\ldots,x_{j,1},\ldots,x_p} = \beta_{k,x_1,\ldots,x_{j,2},\ldots,x_p} = \beta^{\star}_{k,z_{1,k},\ldots,z_{j,k},\ldots,z_{p,k}}$. There is thus no significant difference between how the two levels $x_{j,1}$ and $x_{j,2}$ influence the response y at location k. Importantly, when

[1] For notational simplicity, here we assumed that the values taken by the cluster allocation variables $z_{j,k}^{(x_j)}$'s for different values of $x_j \in \mathcal{X}_j = \{1, \ldots, x_{j,\max}\}$ are sequentially ordered without gaps, i.e., $\mathcal{Z}_{j,k} = \{1, \ldots, \ell_{j,k}\}$. In what follows, we allow other general configurations of $z_{j,k}^{(x_j)}$'s that induce the same equivalent partition of $\mathcal{X}_j$.

For example, consider some $x_j \in \mathcal{X}_j = \{1, 2, 3\}$ with $x_{j,\max} = 3$ partitioned into $\{\{1, 3\}, \{2\}\}$ at location k. Here our notation allows the configurations $(1, 2, 1)$ or $(2, 1, 2)$ of the corresponding cluster allocation variables $(z_{j,k}^{(1)}, z_{j,k}^{(2)}, z_{j,k}^{(3)})$ with $\mathcal{Z}_{j,k} = \{1, \ell_{j,k}\} = \{1, 2\}$. Going forward, we also allow other general configurations $(1, 3, 1)$ or $(3, 1, 3)$ with $\mathcal{Z}_{j,k} = \{1, 3\}$, or $(2, 3, 2)$ or $(3, 2, 3)$ with $\mathcal{Z}_{j,k} = \{2, 3\}$ which induce the same partition $\{\{1, 3\}, \{2\}\}$ of $\mathcal{X}_j$ with $\ell_{j,k} = |\mathcal{Z}_{j,k}| = 2$. In this example, the effects of $x_j = 1$ and $x_j = 3$ on the response curve are the same and hence these levels are clustered together, but these effects are different from the effect of $x_j = 2$ which therefore forms its own cluster. The predictor x_j therefore is important at location k.

Consider also the example when the levels $\mathcal{X}_j = \{1, \ldots, 3\}$ are partitioned into a single cluster at location k. Here our notation only allows the configuration $(1, 1, 1)$ of $(z_{j,k}^{(1)}, z_{j,k}^{(2)}, z_{j,k}^{(3)})$ with $\mathcal{Z}_{j,k} = \{\ell_{j,k}\} = \{1\}$. Going forward, we also allow the configurations $(2, 2, 2)$ or $(3, 3, 3)$ with $\mathcal{Z}_{j,k} = \{2\}$ and $\mathcal{Z}_{j,k} = \{3\}$ respectively which induce the same partition $\{\{1, 2, 3\}\}$ of $\mathcal{X}_j$ with $\ell_{j,k} = |\mathcal{Z}_{j,k}| = 1$. In this example, the effects of all three levels are the same on the response curve. The predictor x_j therefore is unimportant at location k.

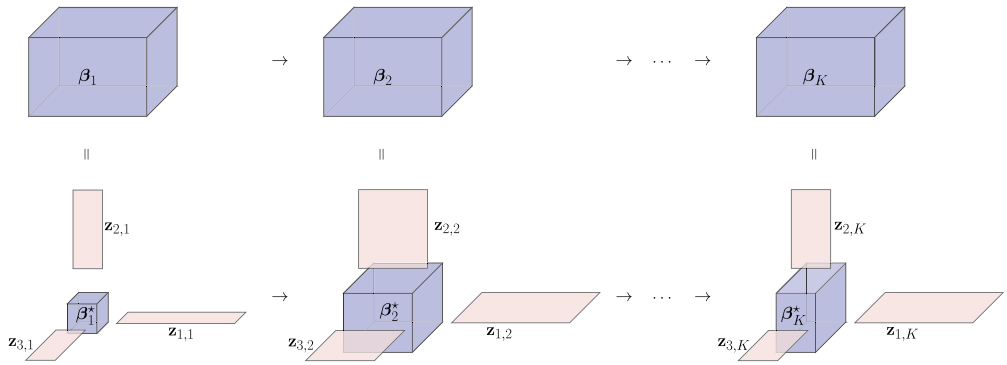

Figure 3: Model (2.4) with three covariates viewed as a dynamic HOSVD. Three-way parameter tensors $\boldsymbol{\beta}_k$ associated with the covariates (x_1, x_2, x_3) at each knot location k are first decomposed via HOSVD into smaller core tensors $\boldsymbol{\beta}_k^\star$ and associated binary mode matrices $\mathbf{z}_{j,k}$'s. Markov dependence structures are later imposed on these components to characterize their temporal dynamics across $k = 1, \ldots, K$.

$\ell_{j,k} = 1$, the $z_{j,k}^{(x_j)}$'s all take on the same value for all different levels of x_j, characterizing the scenario when x_j has no influence on y at location k and local variable selection is achieved. The set of important predictors at location k is thus $\{j : \ell_{j,k} > 1\}$. Significant reduction in model size is achieved at the location k when $\prod_{j=1}^p \ell_{j,k} \ll \prod_{j=1}^p x_{j,\max}$, i.e., when the size of the core tensors is much smaller than the original coefficient tensor (Figure 3). The varying side lengths $\ell_{j,k}$ of the core tensors $\boldsymbol{\beta}_k^\star$ at different locations k (Figure 3) also crucially allow the model to identify different sets of important predictors at different locations k.

In effect, our HOSVD formulation of the continuous coefficient tensors $\boldsymbol{\beta}_k$ in (2.4) into a continuous core but binary mode matrices thus induces local random partitions of the joint covariate space $\mathcal{X}$ into $\ell_k = \prod_{j=1}^p \ell_{j,k}$ sets at each knot location k. This is different from traditional PARAFAC decompositions of continuous tensors into all continuous components as in Guhaniyogi et al. (2017); Papadogeorgou et al. (2019), etc. Section S.3 in the supplementary materials provides some additional discussions on the novelty and advantages our formulation over these other existing approaches.

Second-Layer Clustering We note, however, that the partitions of the joint covariate space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$ induced by the HOSVD in (2.4) may still lead to some overparametrization as they are constructed as the product of p marginal partitions of $\mathcal{X}_j$ into $\ell_{j,k}$ sets (Figure 4, left panel). To obtain an unrestricted partition, say $\rho_k = \{S_{k,1}, \ldots, S_{k,m_k}\}$, we thus further cluster the elements of the core tensors using a second layer of latent variables $z_k^{(z_{1,k},\ldots,z_{p,k})} \in \{1, \ldots, \ell_k\}$ such that

$$\beta^\star_{k,z_{1,k},\ldots,z_{p,k}} = \textstyle\sum_{z_k=1}^{\ell_k} \beta^{\star\star}_{k,z_k} 1\{z_k^{(z_{1,k},\ldots,z_{p,k})} = z_k\}.$$

Such clustering further refines the model (Figure 4, right panel), making the final partition structure of the covariate space fully flexible.[2]

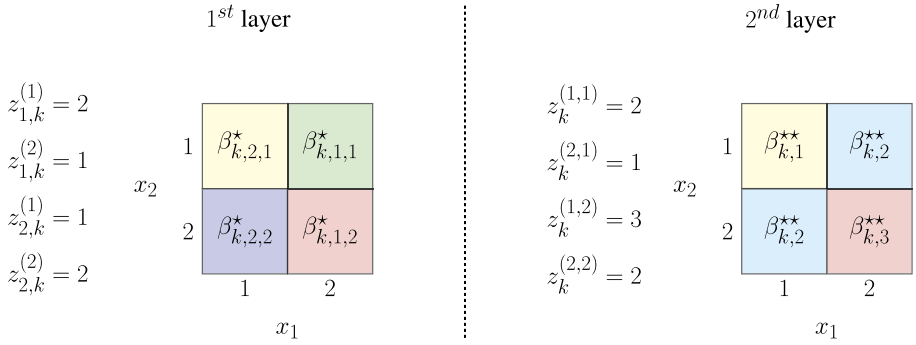

Figure 4: Illustration of the two layers of latent variables that induce the partition of the covariate space at a fixed location k in the case with two categorical predictors x_1, x_2 with $x_{1,\max} = x_{2,\max} = 2$ levels each. In this example, $S_{k,1} = \{(1,1)\}$, $S_{k,2} = \{(1,2),(2,1)\}$, $S_{k,3} = \{(2,2)\}$, $\ell_k = 4$, $m_k = 2$.

Our two clustering layers thus play different roles in achieving parsimonious partitions of the covariate space in high dimensions – efficiently exploring and partitioning $\mathcal{X}$ in high dimensions is an extremely challenging task – we break this into two parts – the HOSVD first reduces the dimensions by creating a product of marginal partitions while also efficiently removing the unimportant covariates – the second layer then refines this smaller space to arrive at a fully flexible model.

In summary, we achieve the challenging task of creating and representing arbitrary clusters of p-dimensional categorical covariates by a 2-step construction; first marginal partitions for each covariate (represented by the $z_{j,k}^{(x_j)}$'s), creating the cross product of these p partitions; and then merging these cross-product clusters to finally form arbitrary p-dimensional clusters (represented by the $z_k^{(z_1,\ldots,z_p)}$'s). The first and second level cluster-membership indicators, namely $z_{j,k}^{(x_j)}$ and $z_k^{(z_1,\ldots,z_p)}$, are only used for constructing the partitions, but are not an inference target themselves, and hence do not give rise to concerns about label-switching.

[2]Consider, e.g., two drugs A and B, each with two dosage levels $\mathcal{X}_A = \mathcal{X}_B = \{1,2\}$. Further let the two levels of each drug have different marginal effects, i.e., for both of them the effect of dosage level 1 is significantly different from that of dosage level 2. The basic idea of our approach to cluster the levels of the predictors (here drugs) according to their effects therefore would produce a clustering of $\{\{1\},\{2\}\}$ for both drugs. The tensor product of these sets then gets us to the joint clustering

$$\{\{1\},\{2\}\} \times \{\{1\},\{2\}\} = \{\{(1,1)\},\{(1,2)\},\{(2,1)\},\{(2,2)\}\} \quad \text{(four clusters)}.$$

It may be possible, however, that there are significant interactions between the levels of A and B, so that the effects of the dose combinations $(1,2)$ and $(2,1)$ are in fact the same. The correct final joint cluster configuration therefore should be

$$\{\{(1,1)\},\{(1,2),(2,1)\},\{(2,2)\}\} \quad \text{(three clusters)}.$$

The second layer, which further clusters $(1,2)$ and $(2,1)$ together, allows us to perform such inference.

So far, we have developed the HOSVD model separately for each knot location k. Next, we focus on introducing time-varying dependency structures between these building blocks appropriate for longitudinal settings. We do this by assuming fHMM dynamics on the allocation variables $\mathbf{z}_{j,k}$ that introduce dependencies between the local partitions at adjacent knot locations, and then assigning novel Markovian priors on the coefficients $\boldsymbol{\beta}_k^{\star\star}$ that make these coefficients vary smoothly over time.

Dynamically Evolving Partition Structures We first consider the problem of specifying probability models for the allocation variables $z_{j,k}^{(x_j)}$ that allow them to be temporally dependent across k. We model the temporal evolution of the $z_{j,k}^{(x_j)}$'s using hidden Markov models (HMMs). For each predictor combination $(x_1, \ldots, x_p)$, the collection $\mathbf{z}^{(x_1,\ldots,x_p)} = \{z_{j,k}^{(x_j)}, k = 1, \ldots, K, j = 1, \ldots, p\}$ then defines a factorial HMM (Ghahramani and Jordan, 1997) (Figure 6). We characterize the dynamics of the fHMM component chains as

$$(z_{j,k}^{(x_j)} \mid z_{j,k-1}^{(x_j)} = z_{k-1}) \sim \text{Mult}(\pi_{z_{k-1},1}^{(j)}, \ldots, \pi_{z_{k-1},z_{j,\max}}^{(j)}) \quad \text{for } j = 1, \ldots, p.$$

We assign Dirichlet priors on the transition probabilities

$$\boldsymbol{\pi}_z^{(j)} = (\pi_{z,1}^{(j)}, \ldots, \pi_{z,z_{j,\max}}^{(j)})^{\mathrm{T}} \sim \text{Dir}(\alpha^{(j)}/z_{j,\max}, \ldots, \alpha^{(j)}/z_{j,\max}),\ \alpha^{(j)} \sim \text{Ga}(a_\alpha, b_\alpha).$$

In general, the maximum number of distinct values of the $z_{j,k}^{(x_j)}$'s is $x_{j,\max}$. However, in most applications, $|\mathcal{Z}_{j,k}|$ will be much smaller than $x_{j,\max}$ uniformly for all k and the restricted support $z_{j,k}^{(x_j)} \in \{1, \ldots, z_{j,\max}\}$, $z_{j,\max} < x_{j,\max}$ will suffice. We impose parsimony by assigning exponentially decaying priors with finite support on the partition sizes $|\mathcal{Z}_{j,k}| = \ell_{j,k}$, favoring smaller partitions as

$$\ell_{j,k} \propto \exp(-\varphi_j \ell_{j,k}), \quad \varphi_j \sim \text{Ga}(a_{\varphi,j}, b_{\varphi,j}), \quad j = 1, \ldots, p, \quad k = 1, \ldots, K.$$

Larger values of φ_j here induce faster decay and hence smaller model sizes. Gamma hyper-priors on the φ_j's further make these shrinkage strengths data adaptive. Being shared across k, the φ_j's also allow to share information on partition sizes across k for each predictor j separately. This is desirable since it is expected that for most predictors, especially the unimportant ones, the influence will be similar across all locations k.

The second layer latent allocation variables $z_k^{(z_{1,k},\ldots,z_{p,k})}$ are assigned multinomial distributions with Dirichlet priors on the probability parameters as

$$(z_k^{(z_{1,k},\ldots,z_{p,k})} \mid \boldsymbol{\pi}_k^\star) \sim \text{Mult}(\pi_{k,1}^\star, \ldots, \pi_{k,\ell_k}^\star),$$
$$\boldsymbol{\pi}_k^\star = (\pi_{k,1}^\star, \ldots, \pi_{k,\ell_k}^\star)^{\mathrm{T}} \sim \text{Dir}(\alpha^\star/\ell_k, \ldots, \alpha^\star/\ell_k) \quad \text{with} \quad \alpha^\star \sim \text{Ga}(a_{\alpha^\star}, b_{\alpha^\star}).$$

When the $z_{j,k}^{(x_j)}$'s corresponding to two different categories of x_j are equal in a temporal region, the local support properties of B-splines then cause the underlying curves to be the same in that region. Conversely, if the $z_{j,k}^{(x_j)}$'s corresponding to two different values of x_j are different, the underlying curves will be distinct unless the second layer of latent variables maps them to the same joint partition element.

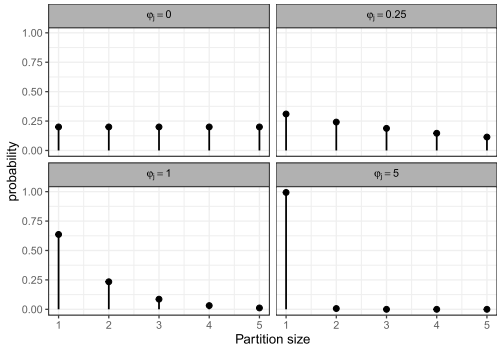

Figure 5: Illustration of the prior distribution on the partition sizes for the categorical predictor x_j with $x_{j,\max} = 5$ levels. Each panel corresponds to different *fixed* values of the parameter φ_j. As φ_j increases, the prior goes from discrete uniform ($\varphi_j = 0$) to a point mass at 1 ($\varphi_j \to +\infty$).

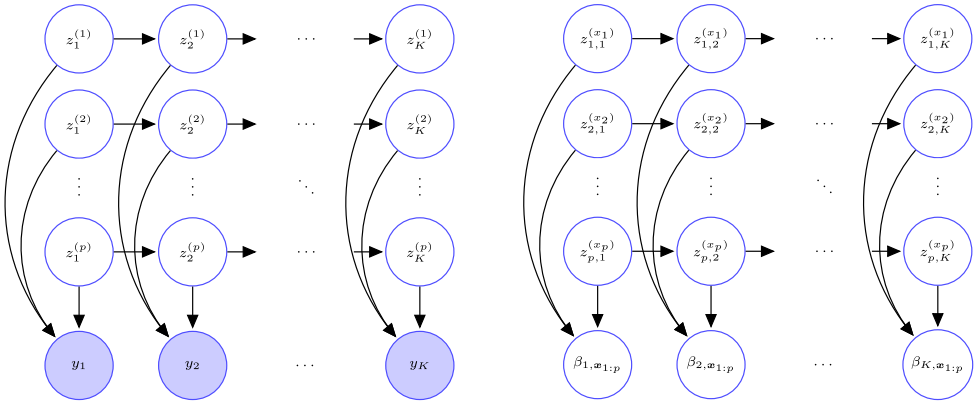

Figure 6: Left panel: The directed acyclic graph (DAG) of a conventional fHMM with p latent layers. Right panel: DAG of our proposed fixed effects model (2.4) with p categorical predictors $\mathbf{x}_{1:p} = (x_1, \ldots, x_p)$.

Conditionally Markov Core Coefficients We next consider priors for the unique core tensors $\beta^{\star\star}_{k,z_k}$. Conditional on the $z^{(x_j)}_{j,k}$'s, $z^{(z_{1,k},\ldots,z_{p,k})}_{k}$'s, and the coefficients at the previous locations, for $k = 2, \ldots, K$, we construct the priors sequentially as

$$\beta^{\star\star}_{k,z_k} \sim \prod_{h \in \mathcal{Z}^{-}_{k,z_k}} \text{Normal}\left(\beta^{\star\star}_{k-1,h}, \sigma^2_\beta\right), \tag{2.5}$$

where $\mathcal{Z}^{-}_{k,z_k} = \left\{ z_{k-1} : z_{k-1} = z^{(z_{1,k-1},\ldots,z_{p,k-1})}_{k-1}; (z_{1,k-1}, \ldots, z_{p,k-1}) = (z^{(x_1)}_{1,k-1}, \ldots, z^{(x_p)}_{p,k-1}); (x_1, \ldots, x_p) \in S_{k,z_k} \right\}$ and S_{k,z_k} is the partition element comprising the covariates levels $(x_1, \ldots, x_p)$ that, at location k, are assigned the label z_k. Simply put, we center

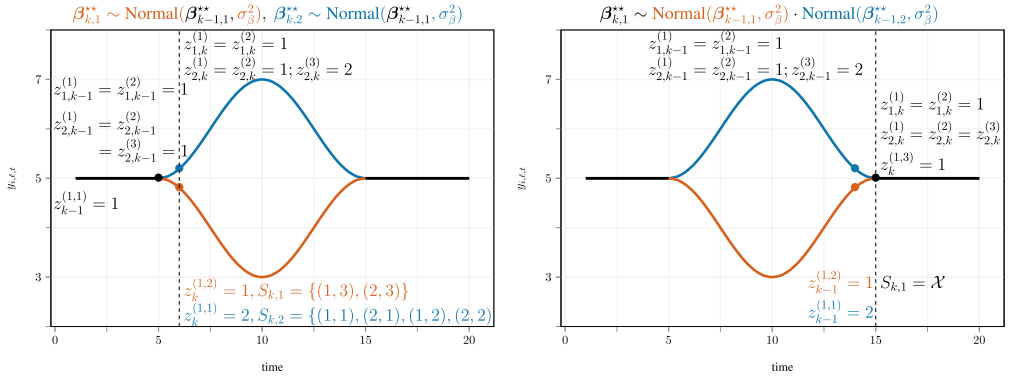

Figure 7: An illustration of the prior on the spline core coefficients $\beta^{\star\star}_{k,z_k}$ at location k (the dashed vertical lines) in the fixed effects model developed in Section 2.1 for a scenario with two categorical covariates $x_1 \in \{1,2\}$ and $x_2 \in \{1,2,3\}$, where the curves corresponding to all levels of (x_1, x_2) are initially equal, the curves for $x_2 = 1, 2$ (in blue) and $x_2 = 3$ (in red) then diverge at $t = 5$, merging back again at $t = 15$.

the core coefficients around the ones that are 'expressed' at the previous location (Figure 7), thus effectively penalizing their differences. The initial coefficients are assigned non-informative flat priors as $\beta^{\star\star}_{1,z_1} \sim 1$. The smoothness of the curves is thus controlled by the parameter σ^2_β and is assigned a prior, allowing it to be informed by the data. We let

$$\sigma_\beta \sim \text{C}^+\text{IG}(0, s_\sigma, c_\sigma, d_\sigma, M_\sigma),$$

where $\text{C}^+\text{IG}(a, b, c, d, M)$ denotes a prior constructed by combining a half-Cauchy distribution with location and scale parameters a and b, respectively, truncated to the interval $[0, M]$ and an inverse-Gamma distribution with shape parameters c and d restricted to (M, ∞), thereby mimicking the behavior of the half-Cauchy near zero and that of the inverse gamma in the tails. The half-Cauchy prior, having a higher probability mass near zero, is well equipped to induce strong smoothing on the β coefficients and is also very appropriate for random effects variance parameters (Gelman, 2006; Polson and Scott, 2012) discussed later in Section 2.2. The inverse gamma component, on the other hand, stops the variances from becoming arbitrarily large (Piironen and Vehtari, 2017) while also satisfying an exponentially decaying tail condition useful for establishing posterior consistency of our model in Section 4. We expect, however, the response error variance σ^2_ϵ to neither be very close to zero nor be very large relative to the overall response variance. A simple inverse-gamma prior thus suffices for it.

Characterization of Main and Interaction Effects As may be noted from our model description above, the HOSVD characterizes each x_j's overall significance explicitly and their joint influences implicitly but very compactly, efficiently eliminating the redundant variables and achieving significant reduction in dimensions, but avoids explicitly describing their main and lower-dimensional interaction effects which are often very useful to

practitioners for their easy interpretation. These effects may, however, be meaningfully *defined* (and easily estimated from the posterior samples) directly as

$$\text{overall mean: } f_0(t) = \frac{\sum_{\mathbf{x}} f_{x_1,\ldots,x_p}(t)}{|\mathcal{X}|}, \quad \text{main effects: } f_{x_j}(t) = \frac{\sum_{\mathbf{x}_{-j}} f_{x_1,\ldots,x_p}(t)}{|\mathcal{X}_{-j}|} - f_0(t),$$
$$\text{interactions: } f_{x_{j_1},x_{j_2}}(t) = \frac{\sum_{\mathbf{x}_{-j_1,-j_2}} f_{x_1,\ldots,x_p}(t)}{|\mathcal{X}_{-j_1,-j_2}|} - f_{x_{j_1}}(t) - f_{x_{j_2}}(t) - f_0(t), \text{ etc.}, \tag{2.6}$$

where $\mathbf{x}_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)^{\mathrm{T}} \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_{j-1} \times \mathcal{X}_{j+1} \times \cdots \times \mathcal{X}_p = \mathcal{X}_{-j}$, and so on. Section S.4 in the supplementary materials provides additional details and plots, a general recipe for testing these effects, etc.

Single Predictor Special Cases The HOSVD approach is relevant particularly for the challenging multivariate predictor problem $(x_1, \ldots, x_p)$ but not for a single predictor x, e.g., the adaptation considered in Paulon et al. (2021), in which case no tensor structure arises, the fHMM (Figure 6, right panel) simplifies to an HMM with a single layer $z_k^{(x)}$, the second layer clustering of the spline coefficients $\beta^{\star}_{k,z_k^{(x)}}$ (Figure 4) is not needed, the smoothness inducing Markovian prior (2.5) greatly simplifies, and the interaction effects discussed just above do not arise. As discussed in the Introduction, the focus of the article is on the general multivariate case. Our implementation, however, is automated to adjust to both scenarios.

2.2 Random Effects Model

We model the random effects components $u_i(t)$ as

$$\begin{aligned} &u_i(t) = \textstyle\sum_{k=1}^{K} \beta_{k,i}^{(u)} b_k(t), \\ &\boldsymbol{\beta}_i^{(u)} \sim \mathrm{MVN}_K\{\mathbf{0}, (\sigma_{u,a}^{-2}\mathbf{I}_K + \sigma_{u,s}^{-2}\mathbf{P}_u)^{-1}\}, \\ &\sigma_{u,s} \sim \mathrm{C}^{+}\mathrm{IG}(0, s_\sigma, c_\sigma, d_\sigma, M), \quad \sigma_{u,a} \sim \mathrm{C}^{+}\mathrm{IG}(0, s_\sigma, c_\sigma, d_\sigma, M), \end{aligned} \tag{2.7}$$

where $\boldsymbol{\beta}_i^{(u)} = (\beta_{1,i}^{(u)}, \ldots, \beta_{K,i}^{(u)})^{\mathrm{T}}$ are subject specific spline coefficients, $\mathrm{MVN}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a K dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The zero mean of the random effects distribution ensures that the random effects are separately nonparametrically identifiable (Guo, 2002; Morris and Carroll, 2006). We choose $\mathbf{P}_u = \mathbf{D}_u^{\mathrm{T}}\mathbf{D}_u$, where the $(K-1)\times K$ matrix $\mathbf{D}_u$ is such that $\mathbf{D}_u\boldsymbol{\beta}_i^{(u)}$ computes the first order differences in $\boldsymbol{\beta}_i^{(u)}$. The model thus penalizes $\sum_{k=1}^{K}(\nabla\beta_{k,i}^{(u)})^2 = \boldsymbol{\beta}_i^{(u)\mathrm{T}}\mathbf{P}_u\boldsymbol{\beta}_i^{(u)}$, the sum of squares of first order differences in $\boldsymbol{\beta}_i^{(u)}$ (Eilers and Marx, 1996). This induces a first order Markov dynamics for the spline coefficients, evident from the tridiagonal structure of the precision matrix in (2.7) that encodes their conditional dependence relationships. The random effects variance parameter $\sigma_{u,s}^2$ models the smoothness of the random effects curves, smaller $\sigma_{u,s}^2$ inducing smoother $u_i(t)$'s. Additional variations

from the constant zero curve are explained by $\sigma^2_{u,a}$. The absence of random effects is signified by the limiting case $\sigma^2_{u,s} = \sigma^2_{u,a} = 0$.

A similar model for functional random effects but with additional assumptions on the covariance matrix has previously been developed in Guo (2002). If we ignore the sharing of information through model hierarchies, the data for estimating an individual level effect come from that individual alone whereas the data for estimating the fixed effects come from many individuals with shared predictor levels. In the literature on mixed models, the models for random effects are thus often kept much simpler compared to fixed effects models. In similar vein, we have focused here on time-varying random intercept type models. When categorical covariates, say $x'_1, \ldots, x'_{p'}$, are desired to be included in the random effects model, $u_i(t)$ can be modified as $u_{x'_1,\ldots,x'_{p'},i}(t)$ and the strategies for modeling fixed effects, as described in Section 2.1, can potentially be used.

A summary of the proposed model components, all at one place, is provided in Section S.10 in the supplementary materials.

3 Posterior Inference

Inference for the proposed LFMM is based on samples drawn from the posterior using an MCMC algorithm. In our model, the values of $\ell_{j,k}$'s are crucial in controlling the model size since they act as local covariate importance indicators. Varying values of $\ell_{j,k}$'s, however, result in varying dimensional models, posing daunting computational challenges. Dynamic message passing algorithms, such as the forward-backward sampler, are popular strategies for inference in HMMs and fHMMs (Rabiner, 1989; Scott, 2002). However, it is not clear how message passing strategies can be adapted to include inferences about the $\ell_{j,k}$'s.

We address these challenges by designing an efficient trans-dimensional transition step which updates the partition structure and the corresponding local curves at every location. First, for every location k, an update in the partition structure ρ_k is proposed. Second, conditional on ρ_k, samples of the spline coefficients $\boldsymbol{\beta}^{\star\star}_k = \{\beta^{\star\star}_{k,h}\}_{h=1}^{m_k}$ are drawn from their Gaussian full conditional distributions.

Specifically, the first step involves updating, for every predictor j at each location k, the first layer of latent variables $\mathbf{z}_{j,k} = (z^{(1)}_{j,k}, \ldots, z^{(x_{j,\max})}_{j,k})$, the implied partition sizes $(\ell_{1,k}, \ldots, \ell_{p,k})$, and the corresponding second layer of latent variables $\mathbf{z}_k = z_k^{(z_{1,k},\ldots,z_{p,k})}$. Designing an efficient such proposal is made challenging by the discrete and potentially high-dimensional support of the latent variables $\mathbf{z}_{j,k}$ and $\mathbf{z}_k$. However, the proposal distribution can be defined sequentially as

$$q(\ell'_{j,k}, \mathbf{z}'_{j,k}, \mathbf{z}'_k \mid \ell_{j,k}, \mathbf{z}_{j,k}, \mathbf{z}_k) = q_1(\ell'_{j,k}, \mathbf{z}'_{j,k} \mid \ell_{j,k}, \mathbf{z}_{j,k}) q_2(\mathbf{z}'_k \mid \ell'_{j,k}, \mathbf{z}'_{j,k}).$$

First, we perturb the current state $\mathbf{z}_{j,k}$ to a new configuration $\mathbf{z}'_{j,k}$ by sampling uniformly in a Hamming ball of radius r around $\mathbf{z}_{j,k}$ (Titsias and Yau, 2014), resulting in an efficient first layer proposal that shares many of the old components as

$$q_1(\ell'_{j,k}, \mathbf{z}'_{j,k} \mid \ell_{j,k}, \mathbf{z}_{j,k}) = \text{Unif}\{\mathbf{z}'_{j,k} \mid \mathcal{H}_m(\mathbf{z}_{j,k})\} 1\{\ell'_{j,k} = |\mathcal{Z}'_{j,k}|\}.$$

Conditioning on the first layer of latent variables, we update the second layer as

$$q_2(\mathbf{z}'_k \mid \ell'_{j,k}, \mathbf{z}'_{j,k}) = \text{Mult}\left(1/\ell_k, \ldots, 1/\ell_k\right).$$

In terms of the implied marginal partition structure $\mathcal{Z}_{j,k}$, when $r = 1$, this corresponds to (A) selecting a covariate level and either (Ba) merging it to one of the other existing partition elements or (Bb) creating a singleton by separating it from its partition element. Since the first layer proposal distribution is symmetric, the resulting acceptance rate of the Metropolis-Hastings (M-H) step is

$$r_{acc} = \frac{p(\mathbf{y}_k \mid \rho'_k, \sigma^2_\varepsilon, \sigma^2_\beta, \boldsymbol{\zeta})}{p(\mathbf{y}_k \mid \rho_k, \sigma^2_\varepsilon, \sigma^2_\beta, \boldsymbol{\zeta})} \cdot \frac{p(\mathbf{z}'_{j,k}) p(\ell'_{j,k}) p(\mathbf{z}'_k)}{p(\mathbf{z}_{j,k}) p(\ell_{j,k}) p(\mathbf{z}_k)} \cdot \frac{q_2(\mathbf{z}_k \mid \ell'_{j,k}, \mathbf{z}'_{j,k})}{q_2(\mathbf{z}'_k \mid \ell'_{j,k}, \mathbf{z}'_{j,k})}, \tag{3.1}$$

where $\mathbf{y}_k = \{y_{i,\ell,k}\}_{i,\ell}$, and $\boldsymbol{\zeta}$ denotes a generic variable that collects all other variables not explicitly mentioned here, including the data points. Importantly, the spline coefficient parameters $\boldsymbol{\beta}^{\star\star}_k$ at each location k can be analytically integrated out of the posterior of the corresponding partition structure. This allows for an efficient scheme for sampling the random partition structures based on their marginal likelihood

$$p(\mathbf{y}_k \mid \rho_k, \sigma^2_\varepsilon, \sigma^2_\beta, \boldsymbol{\zeta}) = \prod_{h=1}^{m_k} \int p(\mathbf{y}_k \mid \beta^{\star\star}_{k,h}, S_{k,h}, \sigma^2_\varepsilon) p(\beta^{\star\star}_{k,h} \mid \sigma^2_\beta, \boldsymbol{\zeta}) d\beta^{\star\star}_{k,h}.$$

The second term in the integral above is the conditional smoothing prior for the spline coefficients

$$\begin{aligned} p(\beta^{\star\star}_{k,h} \mid \sigma^2_\beta, \boldsymbol{\zeta}) &\propto \prod_{h' \in \mathcal{Z}^-_{k,h}} \text{Normal}\left(\beta^{\star\star}_{k,h} \mid \beta^{\star\star}_{k-1,h'}, \sigma^2_\beta\right) \prod_{h'' \in \mathcal{Z}^+_{k,h}} \text{Normal}\left(\beta^{\star\star}_{k+1,h''} \mid \beta^{\star\star}_{k,h}, \sigma^2_\beta\right) \\ &= \text{Normal}(\mu_{k,h}, \sigma^2_{k,h}), \end{aligned}$$

where $\mathcal{Z}^-_{k,h} = \left\{z_{k-1} : z_{k-1} = z^{(z_{1,k-1},\ldots,z_{p,k-1})}_{k-1}; (z_{1,k-1}, \ldots, z_{p,k-1}) = (z^{(x_1)}_{1,k-1}, \ldots, z^{(x_p)}_{p,k-1}); (x_1, \ldots, x_p) \in S_{k,h}\right\}$ and $\mathcal{Z}^+_{k,h} = \left\{z_{k+1} : z_{k+1} = z^{(z_{1,k+1},\ldots,z_{p,k+1})}_{k+1}; (z_{1,k+1}, \ldots, z_{p,k+1}) = (z^{(x_1)}_{1,k+1}, \ldots, z^{(x_p)}_{p,k+1}); (x_1, \ldots, x_p) \in S_{k,h}\right\}$ are the indexes of the coefficients expressed at the previous and following locations, respectively, $n^-_{k,h} = |\mathcal{Z}^-_{k,h}|$ and $n^+_{k,h} = |\mathcal{Z}^+_{k,h}|$ are the corresponding cardinalities, $\sigma^2_{k,h} = \sigma^2_\beta (n^-_{k,h} + n^+_{k,h})^{-1}$ and $\mu_{k,h} = \frac{\sum_{h'} \beta^{\star\star}_{k-1,h'} + \sum_{h''} \beta^{\star\star}_{k+1,h''}}{n^-_{k,h} + n^+_{k,h}}$ are the resulting smoothing prior variance and mean parameters. First order Markov priors we designed in (2.5), as opposed to second order differences considered in Eilers and Marx (1996) and elsewhere, make these calculations much more tractable here. Using this, we get

$$\begin{aligned} &p(\mathbf{y}_k \mid \rho_k, \sigma^2_\varepsilon, \sigma^2_\beta, \boldsymbol{\zeta}) \\ &= \prod_{h=1}^{m_k} \int \prod_{\substack{(i,\ell) \text{ s.t.} \\ \mathbf{x}_{i,\ell,t} \in S_{k,h}}} \left\{(2\pi\sigma^2_\varepsilon)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2_\varepsilon}(r^{(m)}_{i,\ell,k} - \beta^{\star\star}_{k,h})^2}\right\} (2\pi\sigma^2_{k,h})^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2_{k,h}}\left(\beta^{\star\star}_{k,h} - \mu_{k,h}\right)^2} d\beta^{\star\star}_{k,h} \\ &= \prod_{h=1}^{m_k} (2\pi\sigma^2_\varepsilon)^{-\frac{n_{k,h}}{2}} (\sigma^2_{k,h})^{-\frac{1}{2}} (\sigma^{\star 2}_{k,h})^{\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{\sum_{i,\ell} r^{(m)2}_{i,\ell,k}}{\sigma^2_\varepsilon} + \frac{\mu^2_{k,h}}{\sigma^2_{k,h}} - \frac{\mu^{\star 2}_{k,h}}{\sigma^{\star 2}_{k,h}}\right)}, \end{aligned}$$

where $\mathbf{r}^{(m)} = \{y_{i,\ell,t} - u_i(t)\}_{i,\ell,t}$ are the main effects residuals, $n_{k,h} = |\{(i,l) \text{ s.t. } \mathbf{x}_{i,\ell,t} \in S_{k,h}\}|$ is the number of observations allocated to the spline coefficient $\beta^{\star\star}_{k,h}$, $\sigma^{\star 2}_{k,h} = \left(\sigma_\varepsilon^{-2} n_{k,h} + \sigma_{k,h}^{-2}\right)^{-1}$ and $\mu^{\star}_{k,h} = \sigma^{\star 2}_{k,h}\left(\sigma_\varepsilon^{-2}\sum_{i,\ell} r^{(m)}_{i,\ell,k} + \sigma_{k,h}^{-2}\mu_{k,h}\right)$.

Conditional on the partition structure ρ_k, the group specific curves are sampled from their Gaussian full conditional distribution

$$\begin{aligned} p(\beta^{\star\star}_{k,h} \mid \mathbf{y}_k, S_{k,h}, \sigma^2_\varepsilon, \sigma^2_\beta, \boldsymbol{\zeta}) &\propto p(\mathbf{y}_k \mid \beta^{\star\star}_{k,h}, S_{k,h}, \sigma^2_\varepsilon) p(\beta^{\star\star}_{k,h} \mid \sigma^2_\beta, \boldsymbol{\zeta}) \\ &= \text{Normal}\left\{\mu^{\star}_{k,h}, \sigma^{\star 2}_{k,h}\right\}. \end{aligned} \tag{3.2}$$

The full conditionals for σ_β, $\sigma^2_{u,s}$ and $\sigma^2_{u,a}$ under their C$^+$IG priors do not have tractable close forms. M-H steps are therefore used for updating these parameters.

The full MCMC sampler comprises the steps reported in Algorithm 1 in Section S.7.1 in the supplementary materials. Our software implementation in `R` and `C++`, available as part of the online supplementary materials, is highly automated, requiring only the available data points and the values of a few prior hyper-parameters as inputs. These hyper-parameters appear deep inside the model hierarchy and inference is highly robust to their choices. Additional details on the default choices of the hyper-parameters, the runtime of the algorithm, etc. are provided in Sections S.6 and S.7.2 in the supplementary materials.

4 Posterior Consistency

This section presents some convergence results for our proposed longitudinal functional mixed model. We focus on the case where $n \to \infty$ but L, the number of replicates per individual, and T, the number of data recording time points, are kept fixed, which constitutes an appropriate asymptotic regime for the applications discussed later. Under this framework, we focus mainly on the recovery of the fixed effects components. When $L \to \infty$, similar results can also be established for the individual specific effects. We restrict ourselves to consistency at the knot points which coincide with the set of unique data observing time points in the setting of this article. The functional domain remaining fixed to a finite interval, say $[A, B]$, when the number of data recording time points inside the domain $T \to \infty$ and some additional mild smoothness assumptions are made on the true underlying functions, the results can also be extended to the entire domain.

Our proofs rely on some results and ideas from Ghosal et al. (1999) and Suarez and Ghosal (2016) and are presented in Section S.5 in the supplementary materials. We first show consistency for the functional fixed effects. Using this result, we then show that our proposed model can also recover the underlying true local partitions of the covariate space and hence perform consistent variable selection.

We let $\Pi(\cdot)$ denote the prior distribution induced by our model on the space of fixed effects functions $f_{\mathbf{x}}(t)$ and $\Pi(\cdot|\text{data})$ denote the corresponding posterior. We let $g(\mathbf{x})$ denote the probability distribution of $\mathbf{x}$. We consider the g-weighted local L_2-norm of the function $f_{\mathbf{x}}(t)$, defined as $||f||^2_{2,g,loc} = \sum_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}) \sum_{k=1}^{K} f^2_{\mathbf{x}}(k)$. For the linear B-spline mixtures used in this article, $f_{\mathbf{x}}(k) = \beta_{k,\mathbf{x}}$.

Integrating out the random effects distribution (2.7) from model (2.1), we obtain

$$\{y_{i,\ell,t} \mid x_{j,i,\ell,t} = x_j, j = 1, \ldots, p\} \sim \text{Normal}\{f_{\mathbf{x}}(t), \sigma_\varepsilon^2 + \sigma_u^2(t)\}, \tag{4.1}$$

where $\sigma_u^2(t) = \{(\sigma_{u,a}^{-2}\mathbf{I}_K + \sigma_{u,s}^{-2}\mathbf{P}_u)^{-1}\}_{t,t}$. In our proof, we deviate slightly from our stated model in assuming exponentially decaying tails for the priors on the variance parameters $\sigma_{u,a}^2$ and $\sigma_{u,s}^2$ instead of the more non-informative half-Cauchy priors we used in our implementation.

Assuming a fixed dimension p of $\mathbf{x}$ over time t, let $\mathcal{F}$ denote the space of all longitudinal mean functions $\{f \equiv f_{\mathbf{x}}(t),\ \mathbf{x} \in \mathcal{X}\}$ represented by the selected basis functions. We assume the true longitudinal mean functions $f_0 \equiv f_{0,\mathbf{x}}(t) \in \mathcal{F}$.

Theorem 1 (function estimation)**.** For any $\epsilon > 0$, $\Pi(||f - f_0||_{2,g,loc} < \epsilon \mid \text{data}) \to 1$.

Details of the proof appear in Section S.5 in the supplementary material. The key steps are to (i) show that f_0 belongs to the Kullback-Leibler support of our prior, (ii) exhibit a sequence of increasing subsets $\mathcal{F}_n \subseteq \mathcal{F}$, with exponentially decreasing prior probability of $\mathcal{F}_n^c$; (iii) establish an exponential bound for the δ-covering number for $\mathcal{F}_n < e^{n\alpha}$ with a specific α; (iv) then use a theorem from Ghosal et al. (1999) to conclude convergence in L_1 norm, which implies the desired result. Specifically, step (ii) above makes use of the exponentially decaying tails of the priors on the β coefficients and the variance parameters.

Without any loss of generality, we assume that $g(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}$. If not, we can simply restrict ourselves to the set on which $g(\mathbf{x}) > 0$. We then have $n_{\mathbf{x}} \to \infty$ as $n \to \infty$ for all $\mathbf{x} \in \mathcal{X}$. The asymptotic regime can then be understood as averaging over $n_{\mathbf{x}}$ replications for each $\mathbf{x}$, thus replacing $\sigma_\varepsilon^2 + \sigma_u^2(t)$ by $\sigma_{n,\mathbf{x}}^2 = n_{\mathbf{x}}^{-1}\{\sigma_\varepsilon^2 + \sigma_u^2(t)\}$. Theorem 1 then implies that, for any $\mathbf{x} \in \mathcal{X}$ and $\epsilon > 0$, $\Pi(||f_{\mathbf{x}} - f_{\mathbf{x},0}||_{2,loc} < \epsilon \mid \text{data}) \to 1$, where $\mathbf{f}_{\mathbf{x}} = (f_{1,\mathbf{x}}, \ldots, f_{K,\mathbf{x}})^{\mathrm{T}}$ with $f_{k,\mathbf{x}} = f_{\mathbf{x}}(k)$ and $||f_{\mathbf{x}}||_{2,loc}^2 = \sum_{k=1}^K f_{\mathbf{x}}^2(k)$.

For a given location k, let $\rho_k = \{S_{k,1}, \ldots, S_{k,m_k}\}$ be a random partition of $\mathcal{X}$, the space of vectors of length p whose individual entries have values in $\mathcal{X}_j$, respectively. The partition ρ_k is defined in the following way:

$$\beta_{k,\mathbf{x}}^{\star\star} = \beta_{k,\mathbf{x}'}^{\star\star} \iff \mathbf{x}, \mathbf{x}' \in S_{k,h} \text{ for some } h \in \{1, \ldots, m_k\}.$$

Our hierarchical prior for the random partitions assigns a positive probability to each possible configuration. Let $\rho_{k,0}$ be the partition generated by the true values of the parameters at location k. Then the following theorem holds.

Theorem 2 (variable selection)**.** $\Pi(\rho_k = \rho_{k,0} \mid \text{data}) \to 1$.

Details of the proof appear in Section S.5 in the supplementary material. The key steps are to (i) argue that partitions that are not refinements of ρ_0 must have vanishing limiting posterior probability, lest it would violate Theorem 1; (ii) any partition ρ that is a refinements of ρ_0 represent the true $f_{0,\mathbf{x}}$ with additional clusters and cluster-specific parameters, but then, by Occam's razor, the Bayes factor for any such less parsimonious ρ relative to ρ_0 goes to zero, implying again vanishing limiting posterior probability.

The construction of our model in Section 2.1 is such that the influences of the predictors are encoded precisely by the model induced partition structures – the predictor x_j is important at location k if its levels belong to at least two different sets in the partition ρ_k. Consistency in recovering the local partitions thus immediately implies consistency in local variable selection.

5 Simulation Studies

In synthetic experiments, the proposed longitudinal framework achieved excellent empirical performance in recovering the true fixed and random effect curves and associated local cluster configurations from noisy subject level data. Figure 8 illustrates the scenario used in the simulation studies. We considered $T = 20$ time points $\{1, \ldots, T\}$. We generated $p = 10$ predictors, $x_1, x_2 \in \{1, 2\}$ and $x_3, \ldots, x_{10} \in \{1, 2, 3\}$. The total number of possible level combinations of $(x_1, \ldots, x_{10})$ across all time points to consider in a fully flexible but completely unstructured model would thus be $T \prod_{j=1}^{10} x_{j,\max} = 20 \times 2^2 \times 3^8 = 20 \times 26{,}244 = 524{,}880$. The true data generating mechanism is such that x_1 and x_3 are locally important whereas all other covariates are redundant at all time points. The fixed effects curves corresponding to the levels $\{1, 2\}$ and $\{3\}$ of x_3 are initially equal, then diverge at $t = 5$ and finally merge back at $t = 17$, conditional on $x_1 = 1$. The fixed effects curves corresponding to the levels $\{1\}$ and $\{2\}$ of x_1 are initially equal and then diverge at $t = 8$. The true unique spline coefficients are

$$\begin{aligned}
\boldsymbol{\beta}_1^{\star\star} &= (5, 5, 5, 5, 6, 7.25, 8.5, 9, 9.25, 9.5, 9.5, 9.25, 9, 8.5, 7.25, 6, 5, 5, 5, 5)^{\mathrm{T}}, \\
\boldsymbol{\beta}_2^{\star\star} &= (5, 5, 5, 5, 4, 2.75, 1.5, 1, 0.75, 0.5, 0.5, 0.75, 1, 1.5, 2.75, 4, 5, 5, 5, 5)^{\mathrm{T}} \text{ and} \\
\boldsymbol{\beta}_3^{\star\star} &= (5, 5, 5, 5, 6, 7.25, 8.5, 10.5, 12, 13.25, 13.75, 13.75, 13.5, 13, \\
&\qquad 12.5, 12, 11.25, 10.5, 9.5, 8.5)^{\mathrm{T}}.
\end{aligned}$$

We generated $n = 25, 50, 100$ individual specific curves with $L_{i,t} = 5$ repeated measurements at each time point. The residual variance was set at $\sigma_\varepsilon^2 = 1$, whereas the variance and the smoothness of the random effects were $\sigma_{u,s}^2 = 0.1$ and $\sigma_{u,a}^2 = 2$, respectively.

Our method correctly recovers x_1 and x_3 as the only significant predictors. In fact, the estimated number of groups $\ell_{j,k}$ (not shown here for brevity) associated with the other predictors consistently equals to one. The posterior probabilities also correctly estimate two groups for x_1 starting from $t = 8$ and two groups for x_3 starting from $t = 5$. For x_1, we find that the posterior probability $p(\ell_{1,k} = 1 \mid \mathbf{y}) > 0.8$ for $k < 7$, and that $p(\ell_{1,k} = 2 \mid \mathbf{y}) \approx 1$ for $k \geq 8$. For x_3, we find that $p(\ell_{3,k} = 2 \mid \mathbf{y})$ increases from 0.1 to 0.55 for $k = 1$ through $k = 5$, leveling off at $p(\ell_{3,k} = 2 \mid \mathbf{y}) > 0.8$ for $k = 6$ through $k = 20$. Estimates of the fixed effects curves and a few individual level curves obtained by our method are shown in Figure 8. Our model estimates the fixed (left panel) as well as the individual specific (right panel) effects very precisely by borrowing information whenever predictors are redundant or covariate levels are in the same cluster.

We compare the out-of-sample predictive performance of our proposed LFMM with state-of-the-art parametric and nonparametric regression alternatives. We focus particularly on BART models by fitting both the original BART (Chipman et al., 2010) and

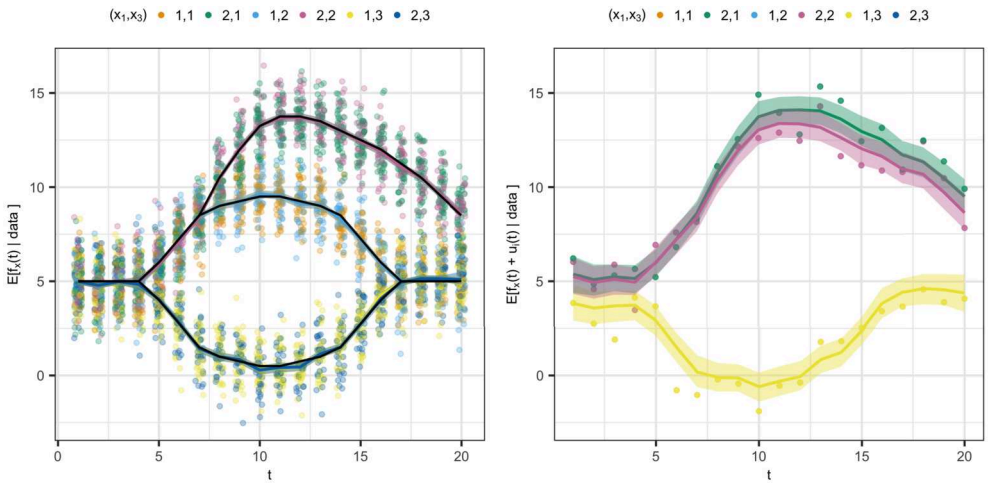

Figure 8: Results for synthetic data: Scenario with ten covariates $(x_1, \ldots, x_{10})$ where only (x_1, x_3) are locally important, as described in Section 5. Left panel: Estimated posterior means (colored lines) and 95% point wise credible intervals for the fixed effects, superimposed on slightly jittered response values $y_{i,\ell,t}$ for all combination of the levels of the significant predictors (x_1, x_3). The true fixed effects are superimposed (black lines). Right panel: Estimated posterior means (colored lines) and 95% point wise credible intervals for three individual specific curves, superimposed on the associated observed individual response values $y_{i,\ell,t}$. The figure here corresponds to the synthetic data set that produced the median root mean squared error.

the smooth BART (Linero and Yang, 2018) to the synthetic data sets. In addition, we apply a LASSO regression model independently at each time point, implemented using the function `glmnet` in `R`. Since these other methods do not accommodate random effects, we are compelled to restrict ourselves to comparing only the fixed effects estimates obtained by the different methods. Figure 9 compares the out-of-sample predictive performance (left panel) and the lengths of the associated 95% prediction intervals (right panel) for the different methods for 500 simulated data sets with 75%–25% training-test splits. All methods produced prediction intervals with coverages probabilities close to the nominal rate (not shown). Remarkably however, despite being very parsimonious, at least compared to BART and soft-BART, and despite having to estimate many more parameters constituting the random effects components, our proposed LFMM not only had substantially smaller out-of-sample root mean squared errors (RMSEs), it actually performed uniformly better than all other approaches in all simulated data sets. Furthermore, our method actually achieved this with uniformly smaller interval widths.

We present the results of some additional simulation experiments in Section S.8 of the supplementary materials to assess the performance of our proposed model in the special but unrealistic case when no individual specific information is available which makes it even more favorable for our competitors that do not accommodate random effects. Our findings are however similar.

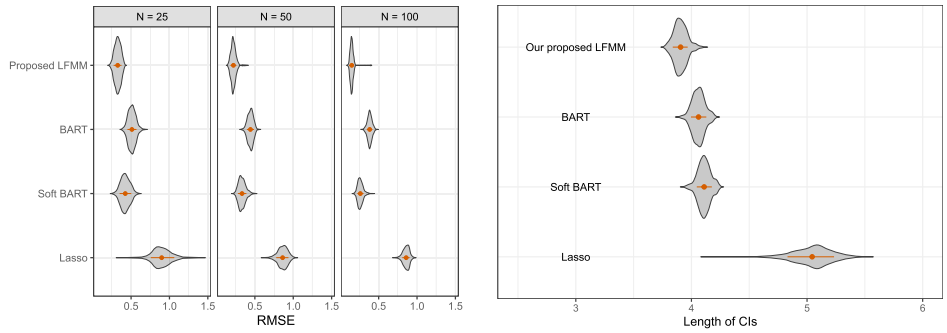

Figure 9: Results for synthetic data: The left panel shows the out-of-sample root mean squared error for $n = 25, 50$ and 100. The right panel shows the widths of the prediction intervals for $n = 50$. All measures reported are obtained over 500 75%-25% training-test splits. The red points represent the averages across simulations, whereas the red intervals represent the interquartile ranges across simulations.

6 Applications

In this section, we discuss the results of our method applied to two data sets. Three more examples, including one with time-varying predictors, are presented in Section S.9 of the supplementary materials.

6.1 Progesterone Data

We describe here an application of our proposed approach to modeling progesterone data (Brumback and Rice, 1998; Nguyen and Gelfand, 2011) that record the logarithm of the progesterone levels of women during the course of their menstrual cycles, measured by urinary hormone assay. Measurements of 51 female subjects occur during a monthly cycle ranging from -8 to 15 (8 days pre-ovulation to 15 days post-ovulation). There are a total of 91 cycles: the first 70 cycles belong to the non-conceptive group, the remaining 21 cycles belong to the conceptive group. The type of cycle is the single categorical predictor used in the analysis.

Figure 10 (left) shows the estimated posterior means and associated 95% point wise credible intervals for the group specific curves. The population level curves for conceptive and non-conceptive cycles are clustered together in the early part of the cycle but become different in the late post ovulation period. In particular, the late conceptive cycles are associated with higher levels of progesterone. Global clustering methods would not allow clustering of the groups in the pre-ovulation period and would simply separate the two groups across all time points. Figure 10 (right) shows the estimated posterior means and associated 95% point wise credible intervals for the individual specific curves. These estimates show how our model can flexibly recover the individual level variations.

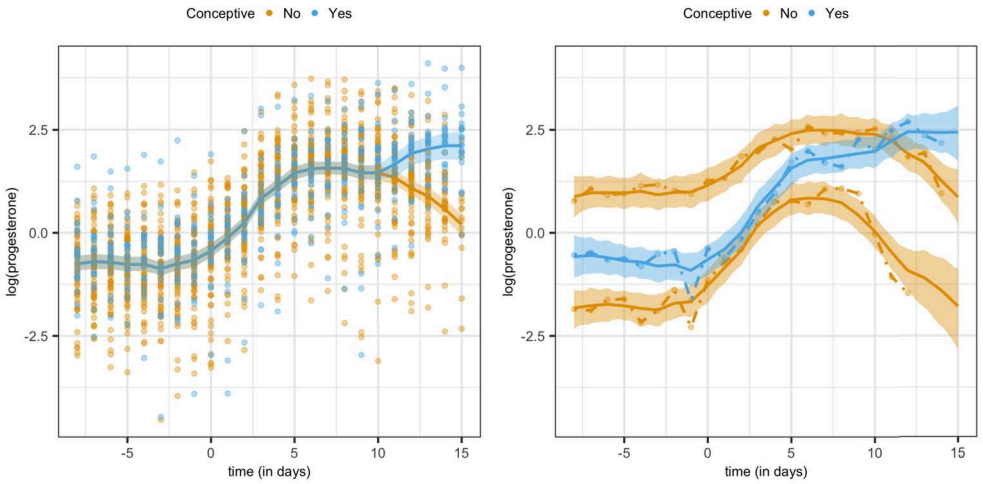

Figure 10: Results for the progesterone data: The left panel shows the estimated posterior means and 95% point wise credible intervals for the fixed effects curves, superimposed on slightly jittered response values $y_{i,\ell,t}$ for different levels of x. Here x is found to be only locally important: Its two levels have no effect on y for $t \in [-8, 10]$ but influence y significantly differently for $t \in (10, 15]$. The right panel shows three examples of individual specific curves, their estimated posterior means (solid lines) and 95% point wise credible intervals, superimposed on the associated observed individual response values (dashed lines) $y_{i,\ell,t}$.

6.2 Health and Retirement Study Data

We analyze publicly available data from a longitudinal survey of US adults, the Health and Retirement Study (HRS). The HRS was established to assess the health implications of aging at both individual and population levels and has been fielded biennially years since 1992. Three levels of data – public, sensitive and restricted – can be accessed on the `HRS website` or the `RAND HRS longitudinal file`. The HRS is sponsored by the National Institute on Aging and the University of Michigan and has previously been analyzed in Sonnega et al. (2014) and most recently in Deshpande et al. (2020).

The goal of the study is to understand how life course processes influence the trajectories of cognitive health. Therefore, we focus on predicting each subject's later-life cognitive function over time using life course socio-economic position (SEP) indicators. The $p = 13$ covariates include measures of SEP in childhood (SEP index), early adulthood (educational attainment), and later-life (household wealth) as well as measures of later-life mental and physical health (binary indicators of physical activity, diabetes, heart problems, high blood pressure, loneliness and stroke as well as body mass index BMI, and depression index) and socio-demographic factors (race, gender). The size of the unstructured model $T \prod_{j=1}^{p} x_{j,\max} = 32 \times 580{,}608 = 18{,}579{,}456$ makes it impossible to estimate the parameters without adopting a dimensionality reduction approach. The outcome is cognitive function as measured by a series of listening and memory tests

that the HRS used to construct a score ranging from 0 to 35. We restricted our analysis to subjects aged between 65 and 96 years with at least two cognitive scores recorded between 2000 and 2016. This resulted in a sample of $n = 4{,}167$ subjects who were administered a total of $N = 27{,}820$ surveys, each individual being recorded either at even or at odd numbered ages but missing the intermediate values.

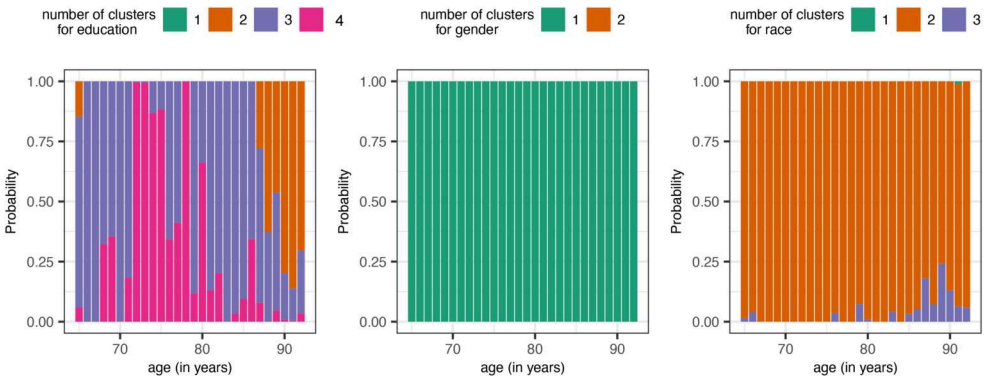

Figure 11: Results for the HRS: The estimated posterior probabilities for the number of clusters of the predictors' levels over time for $x_1 = education$, $x_2 = gender$, and $x_3 = race$. The predictors x_1 and x_3 were locally important. The remaining predictors, including x_2 shown here, were never included in the model since the number of clusters of their levels was always 1.

Figure 11 shows the posterior probabilities for the number of groups $\ell_{j,k}$ associated to three of the predictors (education, gender, race). The other predictors' levels were grouped together at each location k and therefore they did not affect the outcome. Figure 12 shows the effect of education and race, i.e., the two predictors that were selected by the model. These results highlight the importance of educational attainment due to its association with cognition. It appears that higher levels educational attainment are associated with higher cognitive function across adulthood. This confirms that socioeconomic position in early adulthood as measured by education can have later life effects on cognition. Conversely, it appears that the other SEP measures have no predictive effect on later-life cognition. In middle aged individuals, three groups of educational attainment seem to differently affect the outcome: 1–8, 9–12, 13+. In old aged individuals, instead, only two groups of educational attainment are significant: 1–8, and 9+. As far as race is concerned, it appears that after controlling for the other covariates in this study, white and non-white individuals have significant differences in cognitive scores during later-life. This finding also confirms the results in Deshpande et al. (2020), who estimated that white people's intercept parameter is larger than the one for other races, and is consistent with previous literature (Wilson et al., 2015; Díaz-Venegas et al., 2016). This result indicates that other factors that are unaccounted for (i.e., quality of education or literacy) are affecting the estimated cognitive scores for each race/ethnic group. Crucially, our model is able not only to flexibly estimate the cognitive score functions, but also to pool information across different covariate subgroups. Borrowing

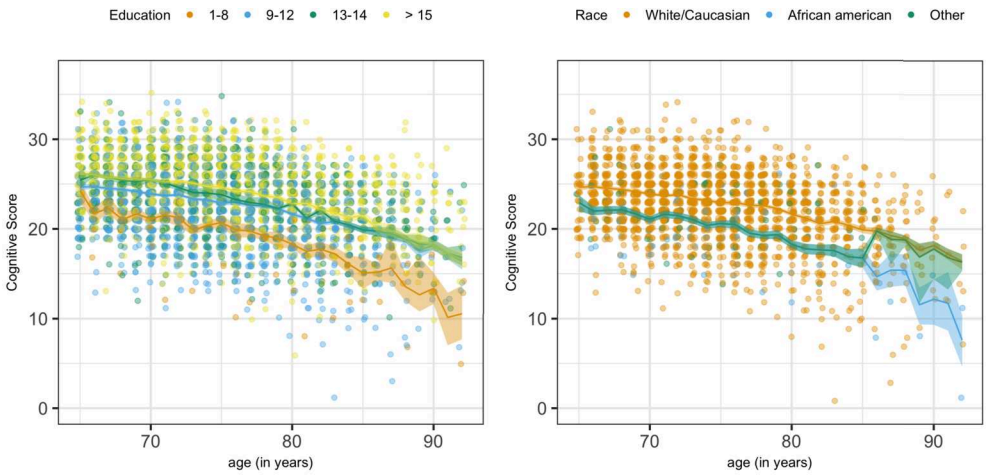

Figure 12: Results for the HRS: The left panel shows the estimated posterior means and 95% point wise credible intervals for the fixed effects curves corresponding to different education levels, superimposed on slightly jittered response values $y_{i,\ell,t}$. The right panel shows the estimated posterior means and 95% point wise credible intervals for the fixed effects curves corresponding to different races, superimposed on slightly jittered response values $y_{i,\ell,t}$.

information across curves becomes especially important to estimate the cognitive score of older aged individuals due to the decrease in sample size.

7 Discussion

In this article, we developed a flexible Bayesian semiparametric approach to longitudinal functional mixed models in the presence of categorical covariates. Building on novel fHMM infused mixtures of locally supported B-splines, our proposed method allows the fixed effects components to vary flexibly with the associated covariates, allowing potentially different sets of important covariates to be included in the model at different time points. The mechanism allows different sets of covariates to be included in the model at different time points while also allowing the selected predictors' influences to vary flexibly over time. Flexible time-varying additive random effects, modeled also by Markovian mixtures of B-splines, are used to capture subject specific heterogeneity. We established theoretical results on posterior consistency of the proposed method for both function estimation and variable selection. In simulation experiments, the method significantly outperformed the competitors. We illustrated the method's practical utility in real data applications.

The methodology presented here is broadly adaptable to diverse other problems. While the focus of this article has been on dynamically varying longitudinal data models, the methodology could also be useful in static multiway mixed ANOVA designs.

Methodological extensions we are pursuing as topics of separate research include dynamic partition models for observational units; models for spatial and spatiotemporal settings; models for multivariate responses; principled approaches to accommodate categorical and ordinal responses and continuous, ordinal and mixed type covariates; etc.

Our approach involves dynamically varying partitions of potentially high-dimensional covariate spaces. Achieving full posterior inference here is non-trivial. The critical computational demand comes from the size of the two-level random partition and the sample size. Our current implementation in `R` takes about 10 minutes for the progesterone data set and about 8 hours for the large HRS data set discussed in Section 6. A comparison of our computation times relative to the competitors considered in Section 5 is presented in Section S.7.2 in the supplementary material. The development of more efficient user-friendly software is in our future plans.

Supplementary Material

Supplementary Material. Bayesian Semiparametric Hidden Markov Tensor Partition Models for Longitudinal Data with Local Variable Selection
(DOI: 10.1214/23-BA1383SUPPA; .pdf). The supplementary materials present brief reviews of B-splines, fHMMs, and tensor factorization methods for easy reference. The supplementary materials also include additional discussions on the characterization of overall, main and interaction effects and associated tests; proofs of the theoretical results; choice of the prior hyper-parameters; additional details of the MCMC algorithm used to sample from the posterior; MCMC diagnostics; results of some additional simulation experiments; additional real data applications; etc.

Supplementary Material. `R` programs
(DOI: 10.1214/23-BA1383SUPPB; .zip). `R` programs implementing the methods developed in this article and an accompanying 'readme' file are also included as separate files.

References

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88: 669–679. MR1224394. 7

Barry, D. and Hartigan, J. A. (1992). "Product partition models for change point problems." *Annals of Statistics*, 20: 260–279. MR1150343. doi: https://doi.org/10.1214/aos/1176348521. 5

Breiman, L. (2001). "Statistical modeling: The two cultures." *Statistical Science*, 16: 199–231. MR1874152. doi: https://doi.org/10.1214/ss/1009213726. 4

Brumback, B. A. and Rice, J. A. (1998). "Smoothing spline models for the analysis of nested and crossed samples of curves." *Journal of the American Statistical Association*, 93: 961–976. MR1649194. doi: https://doi.org/10.2307/2669837. 22

Chib, S. and Hamilton, B. H. (2002). "Semiparametric Bayes analysis of longitudinal data treatment models." *Journal of Econometrics*, 110: 67–89. MR1920962. doi: https://doi.org/10.1016/S0304-4076(02)00122-7. 2

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). "BART: Bayesian additive regression trees." *Annals of Applied Statistics*, 4: 266–298. MR2758172. doi: https://doi.org/10.1214/09-AOAS285. 3, 20

Daniels, M. J. and Pourahmadi, M. (2002). "Bayesian analysis of covariance matrices and dynamic models for longitudinal data." *Biometrika*, 89: 553–566. MR1929162. doi: https://doi.org/10.1093/biomet/89.3.553. 2

de Boor, C. (1978). *A practical guide to splines*. Springer-Verlag. MR0507062. 4, 7

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99: 205–215. MR2054299. doi: https://doi.org/10.1198/016214504000000205. 2, 3

De Lathauwer, L., De Moore, B., and Vandewalle, J. (2000). "A multilinear singular value decomposition." *SIAM Journal on Matrix Analysis and Applications*, 21: 1253–1278. MR1780272. doi: https://doi.org/10.1137/S0895479896305696. 4, 8

Denison, D., Adams, N., Holmes, C., and Hand, D. (2002). "Bayesian partition modelling." *Computational Statistics & Data Analysis*, 38: 475–485. MR1884877. doi: https://doi.org/10.1016/S0167-9473(01)00073-1. 5

Deshpande, S. K., Bai, R., Balocchi, C., and Starling, J. E. (2020). "VC-BART: Bayesian trees for varying coefficients." *arXiv preprint* arXiv:2003.06416. 4, 23, 24

Díaz-Venegas, C., Downer, B., Langa, K. M., and Wong, R. (2016). "Racial and ethnic differences in cognitive function among older adults in the USA." *International Journal of Geriatric Psychiatry*, 31: 1004–1012. 24

Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K.-Y., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford University Press. MR2049007. 2

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). "Generalized spatial Dirichlet process models." *Biometrika*, 94: 809–825. MR2416794. doi: https://doi.org/10.1093/biomet/asm071. 3

Dunson, D. B. (2000). "Bayesian latent variable models for clustered mixed outcomes." *Journal of the Royal Statistical Society: Series B*, 62: 355–366. MR1749544. doi: https://doi.org/10.1111/1467-9868.00236. 7

Efron, B. (2020). "Prediction, estimation, and attribution." *Journal of the American Statistical Association*, 115: 636–655. MR4107663. doi: https://doi.org/10.1080/01621459.2020.1762613. 4

Eilers, P. H. and Marx, B. D. (1996). "Flexible smoothing with B-splines and penalties." *Statistical Science*, 11: 89–102. MR1435485. doi: https://doi.org/10.1214/ss/1038425655. 4, 15, 17

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press. MR1500110. 2

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media. MR2265601. 5

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). "Bayesian nonparametric spatial modeling with Dirichlet process mixing." *Journal of the American Statistical Association*, 100: 1021–1035. MR2201028. doi: https://doi.org/10.1198/016214504000002078. 3

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1: 515–534. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 14

Ghahramani, Z. and Jordan, M. I. (1997). "Factorial hidden Markov models." *Machine Learning*, 29: 245–273. 5, 12

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999). "Posterior consistency of Dirichlet mixtures in density estimation." *Annals of Statistics*, 27: 143–158. MR1701105. doi: https://doi.org/10.1214/aos/1018031105. 18, 19

Gramacy, R. B., Taddy, M., and Wild, S. M. (2013). "Variable selection and sensitivity analysis using dynamic trees, with an application to computer code performance tuning." *Annals of Applied Statistics*, 7: 51–80. MR3086410. doi: https://doi.org/10.1214/12-AOAS590. 3

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2017). "Bayesian tensor regression." *The Journal of Machine Learning Research*, 18: 2733–2763. MR3714242. 5, 10

Guo, W. (2002). "Functional mixed effects models." *Biometrics*, 58: 121–128. MR1891050. doi: https://doi.org/10.1111/j.0006-341X.2002.00121.x. 15, 16

Hartigan, J. A. (1990). "Partition models." *Communications in Statistics – Theory and Methods*, 19: 2745–2756. MR1088047. doi: https://doi.org/10.1080/03610929008830345. 5

Hastie, T. and Tibshirani, R. (1993). "Varying-coefficient models." *Journal of the Royal Statistical Society: Series B*, 55: 757–796. MR1229881. 4

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data." *Biometrika*, 85: 809–822. MR1666699. doi: https://doi.org/10.1093/biomet/85.4.809. 4

Kolda, T. G. and Bader, B. W. (2009). "Tensor decompositions and applications." *SIAM review*, 51: 455–500. MR2535056. doi: https://doi.org/10.1137/07070111X. 4

Koslovsky, M. D., Hébert, E. T., Businelle, M. S., Vannucci, M., et al. (2020). "A Bayesian time-varying effect model for behavioral mHealth data." *Annals of Applied Statistics*, 14: 1878–1902. MR4194252. doi: https://doi.org/10.1214/20-AOAS1402. 4

Li, Y., Lin, X., and Müller, P. (2010). "Bayesian inference in semiparametric mixed

models for longitudinal data." *Biometrics*, 66: 70–78. MR2756692. doi: https://doi.org/10.1111/j.1541-0420.2009.01227.x. 2

Linero, A. R. and Yang, Y. (2018). "Bayesian regression tree ensembles that adapt to smoothness and sparsity." *Journal of the Royal Statistical Society: Series B*, 80: 1087–1110. MR3874311. doi: https://doi.org/10.1111/rssb.12293. 3, 21

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons. MR1925014. doi: https://doi.org/10.1002/9781119013563. 5, 7

MacLehose, R. F. and Dunson, D. B. (2009). "Nonparametric Bayes kernel-based priors for functional data analysis." *Statistica Sinica*, 19: 611–629. MR2514178. 2

Morris, J. S. (2015). "Functional regression." *Annual Review of Statistics and Its Application*, 2: 321–359. 4

Morris, J. S. and Carroll, R. J. (2006). "Wavelet-based functional mixed models." *Journal of the Royal Statistical Society: Series B*, 68: 179–199. MR2188981. doi: https://doi.org/10.1111/j.1467-9868.2006.00539.x. 2, 15

Müller, P., Quintana, F. A., Rosner, G. L., and Maitland, M. L. (2013). "Bayesian inference for longitudinal data with non-parametric treatment effects." *Biostatistics*, 15: 341–352. 2

Nguyen, X. (2010). "Inference of global clusters from locally distributed data." *Bayesian Analysis*, 5: 817–845. MR2740158. doi: https://doi.org/10.1214/10-BA529. 3

Nguyen, X. and Gelfand, A. E. (2011). "The Dirichlet labeling process for clustering functional data." *Statistica Sinica*, 21: 1249–1289. MR2827523. doi: https://doi.org/10.5705/ss.2008.285. 3, 22

Page, G. L., Quintana, F. A., and Dahl, D. B. (2020). "Dependent random partition models." *arXiv preprint* arXiv:1912.11542. 5

Papadogeorgou, G., Zhang, Z., and Dunson, D. B. (2019). "Soft tensor regression." *arXiv preprint* arXiv:1910.09699. MR4329798. 6, 10

Paulon, G., Llanos, F., Chandrasekaran, B., and Sarkar, A. (2021). "Bayesian semiparametric longitudinal drift-diffusion mixed models for tone learning in adults." *Journal of the American Statistical Association*, 116: 1114–1127. MR4309254. doi: https://doi.org/10.1080/01621459.2020.1801448. 5, 15

Paulon, G., Müller, P., and Sarkar, A. (2023). "Supplementary Material. Bayesian Semiparametric Hidden Markov Tensor Partition Models for Longitudinal Data with Local Variable Selection." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1383SUPPA. 6

Paulon, G., Müller, P., and Sarkar, A. (2023). "Supplementary Material. R programs." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1383SUPPB. 6

Petrone, S., Guindani, M., and Gelfand, A. E. (2009). "Hybrid Dirichlet mixture models for functional data." *Journal of the Royal Statistical Society: Series B*, 71: 755–782. MR2750094. doi: https://doi.org/10.1111/j.1467-9868.2009.00708.x. 3

Piironen, J. and Vehtari, A. (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics*, 11: 5018–5051. MR3738204. doi: https://doi.org/10.1214/17-EJS1337SI. 14

Polson, N. G. and Scott, J. G. (2012). "On the half-Cauchy prior for a global scale parameter." *Bayesian Analysis*, 7: 887–902. MR3000018. doi: https://doi.org/10.1214/12-BA730. 14

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American Statistical Association*, 108: 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 7

Quintana, F. A. and Iglesias, P. L. (2003). "Bayesian clustering and product partition models." *Journal of the Royal Statistical Society: Series B*, 65: 557–574. MR1983764. doi: https://doi.org/10.1111/1467-9868.00402. 5

Quintana, F. A., Johnson, W. O., Waetjen, L. E., and B. Gold, E. (2016). "Bayesian nonparametric longitudinal data analysis." *Journal of the American Statistical Association*, 111: 1168–1181. MR3561940. doi: https://doi.org/10.1080/01621459.2015.1076725. 2

Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." *IEEE*, 77: 257–286. 5, 16

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: Methods and case studies*. Springer. MR2168993. 4

Ruppert, D. (2002). "Selecting the number of knots for penalized splines." *Journal of Computational and Graphical Statistics*, 11: 735–757. MR1944261. doi: https://doi.org/10.1198/106186002321018768. 7

Sarkar, A. and Dunson, D. B. (2016). "Bayesian nonparametric modeling of higher order Markov chains." *Journal of the American Statistical Association*, 111: 1791–1803. MR3601736. doi: https://doi.org/10.1080/01621459.2015.1115763. 5

Scott, S. L. (2002). "Bayesian methods for hidden Markov models recursive computing in the 21st century." *Journal of the American Statistical Association*, 97: 337–351. MR1963393. doi: https://doi.org/10.1198/016214502753479464. 5, 16

Singer, J. D., Willett, J. B., Willett, J. B., et al. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press. 2

Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W., and Weir, D. R. (2014). "Cohort profile: The health and retirement study (HRS)." *International Journal of Epidemiology*, 43: 576–585. 23

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). "Nonparametric survival analysis using Bayesian additive regression trees (BART)." *Statistics in Medicine*, 35: 2741–2753. MR3513715. doi: https://doi.org/10.1002/sim.6893. 3

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K., and Scott, J. G. (2020). "BART with targeted smoothing: An analysis of patient-specific stillbirth risk." *Annals of Applied Statistics*, 14: 28–50. MR4085082. doi: https://doi.org/10.1214/19-AOAS1268. 3

Suarez, A. J. and Ghosal, S. (2016). "Bayesian clustering of functional data using local features." *Bayesian Analysis*, 11: 71–98. MR3447092. doi: https://doi.org/10.1214/14-BA925. 3, 18

Taddy, M. A., Gramacy, R. B., and Polson, N. G. (2011). "Dynamic trees for learning and design." *Journal of the American Statistical Association*, 106: 109–123. MR2816706. doi: https://doi.org/10.1198/jasa.2011.ap09769. 3

Titsias, M. K. and Yau, C. (2014). "Hamming ball auxiliary sampling for factorial hidden Markov models." In *Advances in Neural Information Processing Systems*, 2960–2968. 16

Tucker, L. (1966). "Some mathematical notes on three-mode factor analysis." *Psychometrica*, 31: 273–282. MR0205395. doi: https://doi.org/10.1007/BF02289464. 4, 8

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). "Functional data analysis." *Annual Review of Statistics and Its Application*, 3: 257–295. 4

Wilson, R. S., Capuano, A. W., Sytsma, J., Bennett, D. A., and Barnes, L. L. (2015). "Cognitive aging in older Black and White persons." *Psychology and Aging*, 30: 279–285. 24

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov models for time series: An introduction using R*. CRC press. MR3618333. 5