In studies of cognitive aging, it is crucial to distinguish subtypes of longitudinal cognition change while accounting for the effects of given covariates. The longitudinal cognition trajectories and the covariate effects can both be nonlinear with heterogeneous shapes that do not follow a simple parametric form, where flexible functional methods are preferred. However, most functional clustering methods for longitudinal data do not allow controlling for the possible functional effects of covariates. Although traditional mixture-of-experts methods can include covariates and be extended to the functional setting using nonlinear basis functions, satisfactory parsimonious functional methods required for robust functional coefficient estimation and clustering are still lacking. In this paper, we propose a novel latent class functional mixed-effects model, in which we assume the covariates have fixed functional effects, and the random curves follow a mixture of Gaussian processes that facilitates a model-based conditional clustering. A transformed penalized B-spline approach is employed for parsimonious modeling and robust model estimation. We propose a new iterative-REML method to choose the penalty parameters in heterogeneous data. The new method is applied to the latest data from the Religious Orders Study and Rush Memory and Aging Project; and four novel subtypes of cognitive changes are identified.

1. Introduction. Characterizing distinct longitudinal cognitive profiles and identifying their determinants are essential to prevent cognitive impairment and dementia in old age. In particular, the patterns of cognitive decline vary greatly across individuals and include both linear and nonlinear characteristics that do not follow a simple parametric form, see e.g. Hall et al. (2000), Proust et al. (2006), Dodge et al. (2011), and Jack et al. (2013) for different assumptions of the nonlinear patterns. This heterogeneity can partially be explained by individual differences in common age-related neuropathologies (see e.g. Markesbery, 2010), but such heterogeneity usually persists after accounting for the neuropathologies due to various unmeasured risk factors and cognitive resilience (see e.g. Stern et al., 2019). The risk factors can also have nonlinear effects on the cognitive decline that further complicates characterization of the longitudinal cognitive profiles (Proust-Lima, Dartigues and Jacqumin-Gadda, 2011). In this paper, we aim to propose a new conditional functional clustering method that flexibly accommodates various linear and nonlinear patterns under a unified model framework, and allows controlling for the nonlinear effects of covariates in clustering.

Our study was motivated by the Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) (referred to as ROSMAP), which are cohort studies of risk factors for cognitive impairment and dementia, and other health outcomes in aging (Bennett et al., 2018). ROS started in 1994 and enrolls older Catholic nuns, priests, and monks from orders across the United States. MAP started in 1997 and enrolls older lay persons from across northeastern Illinois. The two studies together had enrolled 3559 older participants at the
time of the analysis, and are still ongoing. A battery of 21 cognitive performance tests, 19 in common, was annually administered for each participant. Two tests (Mini-Mental State Examination and Complex Ideational Material) were used only for clinical diagnostic decisions. Performances on 17 tests were used to create a composite measure of global cognitive function. To make the composite measure of global cognition, we first converted raw scores on each component test to $z$ scores, using the baseline mean and SD from the combined parent studies, and then averaged the $z$ scores. Psychometric properties of the composite scores were described previously (Wilson et al., 2002). Figure 1A shows a spaghetti plot of the observed longitudinal global cognition trajectories for a random sample of 50 participants. The heterogeneity across individuals is evident with some having flat trajectories, some linear declines, some accelerating declines with increasing steepness, and some sigmoidal patterns. One of the primary research interests was to examine how this heterogeneity of cognitive decline is attributed to common age-related neuropathologies. Boyle et al. (2017) showed that the effects of several pathologies varied over time with nonlinear patterns by a time-varying effects model ignoring the remaining subject-specific variations. Nonetheless, these pathologic indices all together explain less than half of the variation in person-specific rates of cognitive change (Boyle et al., 2013, 2021). It is of great interest to pursue a more accurate estimation of the effects of neuropathologies with the remaining subject-specific variation explicitly considered, and more importantly, explore the possible latent clusters in the residual trajectories after adjusting for the effects of neuropathologies. Details of our proposed analyses are presented in Section 6.

![Spaghetti plots of longitudinal global cognition trajectories.](image)

**Fig 1.** Spaghetti plots of longitudinal global cognition trajectories: A) the observed trajectories of 50 randomly selected individuals; B) the fitted residual trajectories after regressing out the functional effects of demographics and neuropathologies. The thick trajectories highlight the heterogeneity of longitudinal cognitive decline patterns.

Given the substantial heterogeneity of the person-specific cognitive trajectories, constructing a parametric nonlinear function that includes all possible patterns is non-trivial. In this case, flexible functional approaches are advantageous (see Ramsay and Silverman, 2005, for an introduction). The functional approaches are traditionally proposed to analyze curve data
that are observed densely in a common time interval, where the basic unit of analysis is the entire curve. Although longitudinal data are generally somewhat different with a few scattered measurements at time points that may differ among subjects, we can still apply the functional approaches to longitudinal data by treating the longitudinal trajectories as sparse and noisy observations of a latent functional curve. To account for the covariates, Shi, Weiss and Taylor (1996) and Rice and Wu (2001) proposed to model the longitudinal trajectories by B-spline expansions with both fixed and random coefficients. Notably, Guo (2002) proposed a general functional mixed-effects (FME) model, where both the fixed and random effects were modeled as smoothing splines. Morris and Carroll (2006) proposed a novel wavelet-based FME model to handle spatially heterogeneous and spiky signals. Chen and Wang (2011) further extended the FME model to allow time-varying coefficients and heteroscedastic error variance functions. They proposed penalized splines for estimation and described the asymptotic properties of the estimated curves. The FME model is a powerful tool for investigating the possible nonlinear effects of any given risk factors. However, the classical FME model assumes that the random curves are generated from a single stochastic process, which is not applicable if we want to further investigate the heterogeneity of the random curves. Figure 1B shows the residual trajectories of the same participants in Figure 1A, obtained by a single-class FME model with the effects of neuropathologies and demographics regressed out as fixed functions. It shows that considerable heterogeneity persists in the residual cognitive trajectories, although this heterogeneity appears to be smaller than in the original trajectories.

Generally, the heterogeneity of the random curves can be modeled by classifying individuals into clusters with similar patterns, called latent classes. The mixture model-based methods (Titterington, Smith and Makov, 1985; McLachlan and Peel, 2000; Fraley and Raftery, 2002) are among the most popular approaches to this problem, see e.g. Chamroukhi and Nguyen (2019) for a nice review. In particular, James and Sugar (2003) proposed a general functional clustering method for sparsely sampled longitudinal data using spline basis, and employed a reduced-rank model for robust estimation. Different spline models have been used to estimate the functional curves, including regression B-splines (Luan and Li, 2003; Bar-Joseph et al., 2002; Qin and Self, 2006), smoothing splines (Ma et al., 2006; Ma and Zhong, 2008), and penalized splines (Grün, Scharl and Leisch, 2011; Coffey, Hinde and Holian, 2014). Similarly, Jacques and Preda (2013, 2014) used a truncated functional principal expansion as the basis functions. Alternatively, functional clustering methods are also proposed under other frameworks, such as Bayesian hierarchical models (Ramoni, Sebastiani and Kohane, 2002; Wakefield, Zhou and Self, 2003; Heard, Holmes and Stephens, 2006; Rodriguez and Dunson, 2014), k-means or k-centers clustering methods (Abraham et al., 2003; Serban and Wasserman, 2005; Genolini et al., 2016; Chou and Li, 2007; Peng and Müller, 2008; Delaigle, Hall and Pham, 2019), and pairwise-grouping penalization on B-spline coefficients (Zhu and Qu, 2018). We will focus on model-based methods in this paper.

Despite the extensive research on functional clustering, very few functional methods considered controlling for covariates in clustering. Among the few exceptions, Qin and Self (2006) included the covariates into their model by means of longitudinal regression, with a fixed coefficient for each covariate. Ma and Zhong (2008) proposed to incorporate the covariates into the random part with subject-specific random coefficients to complement the fixed mean functions. None of the aforementioned methods allow the effects of covariates to be nonlinear or functional which is preferred in modeling the effects of neuropathologies on cognitive changes (Boyle et al., 2017). Notably, the mixture-of-expects methods are designed to incorporate covariates into clustering, see e.g. Murphy and Murphy (2020). They are traditionally proposed for cross-sectional multivariate data but can potentially be extended to longitudinal data using nonlinear basis expansions to define the design matrices. For
example, the R package \texttt{lcmix} (Proust-Lima, Philipps and Liquet, 2017) could be used to fit a functional clustering model with functional covariate effects based on nonlinear basis expansions. However, it requires the covariance matrices of the random effects across different latent classes to be proportional to each other, which is usually too restrictive to fit the cognition data well. In general, we do need to use parsimonious covariance matrices for robust model estimation. But those covariance matrices should take into account the intrinsic functional property of the longitudinal data. We also need to put restrictions, e.g. smoothness penalties, on the functional effects to avoid boundary effects as the longitudinal data are typically sparse at either or both ends of follow-up times. Nonetheless, satisfactory methods to design the covariance matrices and impose smoothness restrictions in heterogeneous longitudinal data are still lacking in the literature.

In this paper, we propose a new latent class functional mixed effects model where the effects of covariates are modeled as flexible fixed functions, and the person-specific random curves are assumed to be generated from a mixture of Gaussian processes that facilitates model-based conditional clustering. Each component in the mixture corresponds to a cluster of individuals with similar patterns after accounting for the functional effects of the covariates. We allow the clusters to have different mean functions and covariance kernel functions, as well as different within-subject variances in terms of measurement errors. This flexibility offers a more accurate specification of the distribution of the random curves. As a by-product, it enables a more accurate estimation of the fixed functional effects in heterogeneous data than the single-class FME methods. A similar phenomenon was also reported in fixed effects estimation of the linear mixed-effects model with heterogeneous random effects (Verbeke and Lesaffre, 1996).

To fit the model, all functional components are estimated through penalized B-splines (Eilers and Marx, 1996). We propose to transform the B-spline expansion into a linear part and a nonlinear part, in the same spirit as the smoothing splines (Wahba, 1990; Guo, 2002), which offers a parsimonious parametrization for robust model estimation and rigorous statistical inference. We propose a penalized likelihood approach for model estimation using a penalized EM algorithm (Dempster, Laird and Rubin, 1977; Green, 1990), see also Ma and Zhong (2008) for a similar algorithm using smoothing splines in the context of mixture models. Ma and Zhong (2008) proposed to select the smoothing parameters by generalized cross-validation (GCV) applied to each cluster separately using single-class methods (Gu and Ma, 2005). Our algorithm distinguishes from Ma and Zhong (2008) in that we have fixed functional effects shared by all clusters. Consequently, the smoothing parameters can no longer be selected separately in each cluster by classical methods proposed for homogeneous data. Direct extension of the GCV and GML/REML methods (Wahba, 1985; Wang, 1998) to mixture data involves heavy computational burden. We develop a new iterative-REML method to select the smoothing parameters by generalized cross-validation (GCV) applied to each cluster separately using single-class methods (Gu and Ma, 2005). Our algorithm distinguishes from Ma and Zhong (2008) in that we have fixed functional effects shared by all clusters. Consequently, the smoothing parameters can no longer be selected separately in each cluster by classical methods proposed for homogeneous data. Direct extension of the GCV and GML/REML methods (Wahba, 1985; Wang, 1998) to mixture data involves heavy computational burden. We develop a new iterative-REML method to select the smoothing parameters, which is compatible with the EM algorithm by improving the likelihood in each iteration. For model selection, we propose to use a modified BIC derived from the penalized likelihood (Konishi, Ando and Imoto, 2004). Similar to smoothing splines, we can construct Bayesian confidence intervals (Wahba, 1983) for the penalized spline estimates. In addition, taking advantage of the spline basis representations, we can use simple Chi-square tests to check the significance and nonlinearity of the functional estimates (James and Silverman, 2005).

The rest of this paper is organized as follows. In Section 2, we first provide a brief review of the model-based functional clustering, and then introduce the latent class functional mixed effects model for conditional functional clustering. Details of the estimation methods are presented in Section 3. Section 4 discusses the statistical inferences. Simulation studies and application to the ROSMAP data are presented in Sections 5 and 6, respectively. We conclude with discussion in Section 7.
2. Methodology.

2.1. Model-based functional clustering for longitudinal data. In this section, we first provide a brief review of the model-based functional clustering for longitudinal data. Let \( y_{ij} \) denote the response for the \( i \)th individual, measured at time \( t_{ij} \). Suppose there is a latent function for each individual, denoted by \( f_i \), \( i = 1, \ldots, n \), and \( y_{ij} \) is a noisy observation of \( f_i \) at time \( t_{ij} \), i.e.,

\[
y_{ij} = f_i(t_{ij}) + e_{ij}, \quad j = 1, \ldots, m_i,
\]

where \( e_{ij} \) denotes the measurement error. For example, in cognitive aging studies, \( y_{ij} \) corresponds to the cognitive test score, and \( f_i \) denotes the latent trajectory of cognitive change; and clusters in \( f_i \), \( i = 1, \ldots, n \), shed light on the subtypes of cognitive changes.

Define vector \( \mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})^\top \) to be the longitudinal trajectory for the \( i \)th individual, and \( \mathbf{t}_i = (t_{i1}, \ldots, t_{im_i})^\top \) as the vector of measurement times. In many studies of cognitive aging, \( \mathbf{t}_i \) is unbalanced with possible missing observations, which causes difficulties in statistical clustering. First of all, classical multivariate clustering (see e.g. Fraley and Raftery, 2002) cannot be applied, because \( \mathbf{y}_i \)'s are not comparable if \( \mathbf{t}_i \)'s are different across individuals. Although it is possible to project \( \mathbf{y}_i \) onto an orthogonal basis expansion to obtain comparable basis coefficients as a data pre-processing step, and then perform multivariate clustering (e.g. k-means) as a second step, this method raises a non-trivial question of how to measure the quality of a clustering (see e.g. Tarpey, 2007). In this case, model-based functional clustering is preferred.

Model-based functional clustering assumes that the functions \( f_1, \ldots, f_n \) are generated by a mixture of random processes with \( G \) components (see e.g. James and Sugar, 2003). Each component corresponds to a cluster of functions. We assume that the components are all Gaussian processes. Define the mean function and covariance kernel of the \( g \)th component as \( \mu_g(\cdot) \) and \( R_g(\cdot, \cdot) \), i.e., if the \( i \)th individual is a member of the \( g \)th cluster, then

\[
E[f_i(t)] = \mu_g(t), \quad \text{Cov}(f_i(t), f_i(s)) = R_g(t, s),
\]

where \( \mu_g \) and \( R_g \) are continuously differentiable functions. The mixture of \( G \) Gaussian processes can be written as \( f_i \sim \sum_{g=1}^{G} \pi_g \text{GP}(\mu_g, R_g) \), where ‘GP’ stands for Gaussian process, and \( \pi_g \) is the mixing weight for the \( g \)th cluster satisfying \( \sum_{g=1}^{G} \pi_g = 1 \).

Here and in the sequel, we assume the cluster membership of \( i \)th individual is unknown which will be modelled as a latent variable. Then the distribution of \( \mathbf{y}_i \) has a mixture form as

\[
\mathbf{y}_i \sim \sum_{g=1}^{G} \pi_g N(\mathbf{u}_{i,g}, \Sigma_{i,g} + \sigma^2_g I_{m_i}), \tag{1}
\]

where \( \mathbf{u}_{i,g} = (\mu_g(t_{i1}), \ldots, \mu_g(t_{im_i})) \),

\[
\Sigma_{i,g} = \begin{pmatrix}
R_g(t_{i1}, t_{i1}) & R_g(t_{i1}, t_{i2}) & \cdots & R_g(t_{i1}, t_{im_i}) \\
R_g(t_{i2}, t_{i1}) & R_g(t_{i2}, t_{i2}) & \cdots & R_g(t_{i2}, t_{im_i}) \\
\vdots & \vdots & \ddots & \vdots \\
R_g(t_{im_i}, t_{i1}) & R_g(t_{im_i}, t_{i2}) & \cdots & R_g(t_{im_i}, t_{im_i})
\end{pmatrix},
\]

and \( \sigma^2_g \) is the variance of the measurement errors. Given the model parameters, the cluster classification of the \( i \)th individual is based on the model-derived posterior probabilities.

When covariates are involved and denoted by a vector \( X_i \), it is fairly straightforward to incorporate \( X_i \) into \( \pi_g \) using a multinomial probit model as

\[
\pi_g(X_i) = \frac{\pi_0 g \exp(\xi_g^\top X_i)}{\sum_{g'=1}^{G} \pi_0 g' \exp(\xi_{g'}^\top X_i)},
\]
where \( \pi_{0g} \geq 0 \) for \( g = 1, \ldots, G \). This extended model can be used to examine how the latent cluster membership depends on \( X_i \). But it does not provide information regarding how \( X_i \) contributes to the heterogeneity of the longitudinal trajectories. When the interest is in exploring the heterogeneity after controlling for the effects of the risk factors, such model is not appropriate. In the next section, we introduce a new approach that performs model-based functional clustering with covariates controlled.

### 2.2. Conditional functional clustering

To control for the effects of covariates in clustering, we assume \( y_{ij} \) follows a latent class functional mixed effects model:

\[
y_{ij} = \beta(t_{ij})^\top X_i + \alpha_i(t_{ij}) + e_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i,
\]

where \( X_i \) is a vector of covariates, \( \beta(t) = (\beta_1(t), \ldots, \beta_p(t))^\top \) is a \( p \times 1 \) vector of fixed functions, \( \alpha_i(t) \) is a random function assumed to be generated by a mixture of Gaussian processes, i.e., \( \alpha_i \sim \sum_{g=1}^{G} \pi_g \text{GP}(\mu_g, R_g) \), and \( e_{ij} \) is measurement error that follows a normal distribution with mean 0 and cluster-specific variance \( \sigma_y^2 \). Based on model (2), the conditional distribution of \( y_i \) given \( X_i \) is

\[
y_i \mid X_i \sim \sum_{g=1}^{G} \pi_g N \left( \beta(t_i)^\top X_i + \mathbf{u}_{i,g}, \Sigma_{i,g} + \sigma_y^2 I_{m_i} \right),
\]

where \( \beta(t_i)^\top = (\beta(t_{i1}), \ldots, \beta(t_{i,m_i}))^\top \), and \( \mathbf{u}_{i,g} \) and \( \Sigma_{i,g} \) are as defined in (1). Similar to the linear mixed effects model, \( \beta(\cdot) \) can be interpreted as the functional effects of \( X_i \) on the longitudinal profiles. Thus, model (2) facilitates a conditional functional clustering with the functional effects of \( X_i \) controlled. In a special case when \( \beta(t) \) and \( \alpha_i(t) \) are modelled as linear functions, model (2) reduces to be a linear mixed effects model with heterogeneous random effects (Verbeke and Lesaffre, 1996).

Generally, \( \beta(t) \) and \( \alpha_i(t) \) are modelled as flexible functions. Two popular choices are B-spline expansions (e.g. Rice and Wu, 2001) and smoothing splines (e.g. Guo, 2002). We chose B-spline because it is easier to implement under the latent class structure. We assume \( \beta(t) \) and \( \alpha_i(t) \) can be approximated as

\[
\beta_i(t) = B(t)^\top \mathbf{b}_i \quad \text{and} \quad \alpha_i(t) = B(t)^\top \mathbf{a}_i
\]

where \( B(t) \) is a \( K \times 1 \) vector of cubic B-spline basis functions with \( K - 4 \) interior knots defined in a fixed interval \( \mathcal{T} \) that contains all \( t_i \)'s, and \( \mathbf{b}_i \) and \( \mathbf{a}_i \) are respectively the fixed and random vectors of basis coefficients. The random vector \( \mathbf{a}_i \) is assumed to follow a mixture of \( G \) multivariate Gaussian distribution with different means \( \mathbf{u}_g \) and covariance matrices \( \Sigma_g, \quad g = 1, \ldots, G \), i.e., \( \mathbf{a}_i \sim \sum_{g=1}^{G} \pi_g N(\mathbf{u}_g, \Sigma_g) \). Therefore, the corresponding mean function and covariance kernel of the \( g \)th Gaussian process are approximated by

\[
\mu_g(t) = B(t)^\top \mathbf{u}_g \quad \text{and} \quad R_g(t,s) = B(t)^\top \Sigma_g B(s).
\]

### 2.3. A transformed penalized spline approach

Often we need to use a fairly large number of basis functions so as to capture all possible nonlinear patterns of the longitudinal trajectories. Naturally there is a concern of overfitting for both the fixed functions \( \beta_i \) and the mean functions \( \mu_g \). Moreover, if \( \Sigma_g \)'s are all unstructured, the number of unknown parameters in the covariance matrices is \( K(K + 1)/2 \times G \), which is usually too large to be estimated well. To alleviate these problems, we adopt a penalized spline approach (Eilers and Marx, 1996; Chen and Wang, 2011), and transform the B-spline basis functions into a linear part and a nonlinear part, in the same spirit as the smoothing spline approach (Guo, 2002). A similar decomposition was previously considered in Grün, Scharl and Leisch (2011) for the cluster-specific mean functions only.
We employ a penalty on the second order derivative of a functional curve to control the smoothness. For example, for the fixed functions, the penalty terms are

\[ P_{\beta,l} = \lambda_l \int_T \beta''_l(t)^2 dt, \quad l = 1, ..., p. \]

Based on the B-spline expansion, the penalty can also be written as

\[ P_{\beta,l} = \lambda_l \int_T \left\{ B''_l(t) b_l \right\}^2 dt = \lambda_l b_l^T A_2 b_l, \quad l = 1, ..., p. \]

where \( A_2 = \int_T B''(t) B''(t)^T dt \). It is easy to see that the rank of \( A_2 \) is \( K - 2 \), i.e., the eigen-decomposition of \( A_2 \) has a form as

\[ A_2 = [V_1, V_2] \times \begin{pmatrix} 0_{2 \times 2} & \Lambda \\ \Lambda & 0_{2 \times 2} \end{pmatrix} \times [V_1, V_2]^T, \]

where \( V_1 = (v_{10}, v_{11}) \) is a \( K \times 2 \) matrix, \( V_2 \) is a \( K \times (K - 2) \) matrix orthogonal to \( V_1 \), and \( \Lambda \) is a diagonal matrix of \( K - 2 \) strictly positive numbers. The solution for \( V_1 \) is not unique. Any two functions \( B(t)^T v_{10} \) and \( B(t)^T v_{11} \) that span the space of linear functions on \( T \) can be orthogonalized to provide a valid \( V_1 \). Thus, the functional space spanned by \( B(t)^T v_{10} \) and \( B(t)^T v_{11} \) is equivalent to that of a constant function 1 and \( t \).

Define the transformed B-spline basis functions as

\[ \tilde{B}(t) = \begin{pmatrix} 1, t, \tilde{B}_2(t)^T \end{pmatrix}^T, \]

where \( \tilde{B}_2(t)^T = B(t)^T V_2 \Lambda^{-\frac{1}{2}} \), and the corresponding transformed basis coefficients as

\[ \tilde{b}_l = (b_{l,0}, b_{l,1}, b_{l,2}), \]

where \( b_{l,2} = \Lambda^{\frac{1}{2}} V_2^T b_l, \quad l = 1, ..., p. \) By construction, the functional space spanned by \( \tilde{B}(t) \) is equivalent to that of \( B(t) \), and thus inherits all the nice mathematical properties of the classical B-splines. The fixed functions can then be written as

\[ \beta_l(t) = b_{l,0} + b_{l,1} t + \tilde{B}_2(t)^T b_{l,2}. \]

It divides \( \beta_l(t) \) into a linear component \( b_{l,0} + b_{l,1} t \) and a nonlinear component \( \tilde{B}_2(t)^T b_{l,2} \). The smoothness penalty term is simplified to be \( P_{\beta,l} = \lambda_l b_{l,2}^T b_{l,2} \), which only acts on the nonlinear component. Following the idea of Bayesian interpretation of the spline models (Wahba, 1978; Silverman, 1985), a penalized spline estimator of \( \beta_l \) can be regarded as a posterior estimate with diffuse priors on \( b_{l,0} \) and \( b_{l,1} \) and Gaussian prior \( b_{l,2} \sim N(0, \lambda_l^{-1} I_{K-2}) \), where \( I_{K-2} \) is the \( (K-2) \times (K-2) \) identity matrix. Therefore, we can adopt a similar approach to Wahba (1983) to construct the confidence intervals. Details are given in Section 4.

Similarly, the mean function of the \( g \)th cluster can be written as

\[ \mu_g(t) = u_{g,0} + u_{g,1} t + \tilde{B}_2(t)^T u_{g,2}. \]

The associated smoothness penalty term is \( P_{\mu,g} = \gamma_g u_{g,2}^T u_{g,2} \). Treating \( \mu_g \) as another fixed parameter, its inference is similar to \( \beta_l \). The random functions can then be written as

\[ \alpha_i(t) = a_{i,0} + a_{i,1} t + \tilde{B}_2(t)^T a_{i,2}, \]

where \( a_{i,0} \) and \( a_{i,1} \) are similar to the random intercept and random slope in a linear mixed effects model, and the last term is the nonlinear component of the random function. It implies
a structural assumption on covariance matrix of the random coefficients such that, for the $g$th cluster,

$$
\tilde{\Sigma}_g = \text{var}\{(a_{i,0}, a_{i,1}, a_{i,2})\} = \left(\tilde{\Sigma}_{g,1}, \kappa_g I_{K-2}\right),
$$

where $\tilde{\Sigma}_{g,1}$ is a $2 \times 2$ symmetric covariance matrix and $\kappa_g$ is a scalar parameter that controls the smoothness of the nonlinear components in the $g$th cluster. Each $\tilde{\Sigma}_g$ contains only 4 unknown parameters.

### 3. Estimation.

#### 3.1. The penalized likelihood function and estimation.

Let $\tilde{u}_g = (u_{g,0}, u_{g,1}, u_{g,2})$ and $\tilde{a}_i = (a_{i,0}, a_{i,1}, a_{i,2})$. Our working model is built on the transformed basis functions $B(t) = (1, t, \tilde{B}_2(t)^\top)$. Namely,

$$
\beta(t) = B(t)^\top \tilde{b}_l, \quad \mu_g(t) = B(t)^\top \tilde{u}_g, \quad \text{and} \quad \alpha_i(t) = B(t)^\top \tilde{a}_i,
$$

with $\tilde{a}_i \sim \sum_{g=1}^{G} \pi_g N(\tilde{u}_g, \tilde{\Sigma}_g)$. Let $\theta = (\tilde{b}, \tilde{u}, \tilde{\Sigma}, \pi)$ denote the vector of unknown parameters, where $\tilde{b} = (\tilde{b}_1^\top, ..., \tilde{b}_p^\top)^\top$, $\tilde{u} = (\tilde{u}_1^\top, ..., \tilde{u}_G^\top)^\top$, $\tilde{\Sigma} = (\text{vech}(\tilde{\Sigma}_{1,1}), ..., \text{vech}(\tilde{\Sigma}_{G,1}), \kappa_1, ..., \kappa_G, \sigma_1, ..., \sigma_G)$, and $\pi = (\pi_1, ..., \pi_G)$. The conditional density function of $y_i$ is approximated by

$$
f(y_i|X_i, t_i; \theta) = \sum_{g=1}^{G} \pi_g \phi\left(y_i \big| X_i^\top \otimes \tilde{B}_i \right)^\top \tilde{b}_l + \tilde{B}_i^\top \tilde{u}_g, \tilde{B}_i^\top \tilde{\Sigma}_g \tilde{B}_i + \sigma_g^2 I_{m_i},
$$

where $\otimes$ stands for the Kronecker product, $\tilde{B}_i = (\tilde{B}(t_{i1})^\top, ..., \tilde{B}(t_{ip})^\top)^\top$, and $\phi(y_i|u, \Sigma)$ is the Gaussian density of $y_i$ with mean $u$ and variance $\Sigma$.

Let $O_i = \{y_i, t_i, X_i\}$ be the observed data for the $i$th individual, and $O = \{O_i\}_{i=1}^n$. Define $\lambda = (\lambda_1, ..., \lambda_p, \gamma_1, ..., \gamma_G)$ which contains all the smoothing penalty parameters for $\beta_l$, $l = 1, ..., p$, and $\mu_g$, $g = 1, ..., G$. Suppose the measurement times $t_i$ are non-informative, i.e., their distributions do not depend on $\theta$. The penalized log-likelihood function is defined as,

$$
\ell_P(\theta|O, \lambda) = \sum_{i=1}^{n} \log f(y_i|X_i, t_i; \theta) - \frac{1}{2} \sum_{l=1}^{p} \lambda_l b_{l,2} b_{l,2}^\top - \frac{1}{2} \sum_{g=1}^{G} \gamma_g u_{g,2} u_{g,2}^\top.
$$

This likelihood function is invariant under permutations of cluster labels, known as the labeling non-identifiability (Redner and Walker, 1984). Nevertheless, we can assume the parameters are asymptotic identifiable, i.e., although there may be multiple parameter values in the parameter space that give the same sample distributions, we can always find an open neighborhood of each optimal parameter such that every other parameter in the neighborhood gives a unique and less optimal distribution (Rothenberg, 1971; Redner and Walker, 1984). Therefore, if the initial value lies in one of these neighborhoods, we can find the optimal value within this neighborhood using the classical unimodal methods. Titterington, Smith and Makov (1985, section 3.1) discussed in detail the concept of identifiability for mixtures, who pointed out that most finite mixtures of continuous densities are identifiable except for the mixture of uniform densities.

Because of the latent class structure, direct maximization of $\ell_P(\theta|O, \lambda)$ is a difficult non-convex problem. We employ a penalized EM algorithm (Dempster, Laird and Rubin, 1977;
Green, 1990; Ma and Zhong, 2008) to estimate the parameters. The latent variable to be considered as missing data is the cluster membership of an individual, denoted by $Q_i = (q_{i1}, ..., q_{iG})$, $i = 1, ..., n$. If the $i$th individual belongs to the $g$th cluster, then $q_{ig} = 1$ and $q_{ik} = 0$ for $k \neq g$. The joint density of $y_i$ and $Q_i$ given $X_i$ and $t_i$ can be written as

$$f(y_i, Q_i | X_i, t_i; \theta) = \prod_{g=1}^{G} \left[ \pi_g \phi \left( y_i \left| X_i^\top \otimes \tilde{B}_i \right. \right)^\top b + \tilde{B}_i^\top \tilde{u}_g, \tilde{B}_i^\top \tilde{S}_g \tilde{B}_i + \sigma_g^2 I_m, \right]^{q_{ig}}.$$ 

The complete-data penalized log-likelihood function is defined as

$$\ell^*_P(\theta | O, Q, \lambda) = \sum_{i=1}^{n} \log f(y_i, Q_i | X_i, t_i; \theta) - \frac{1}{2} \sum_{l=1}^{p} \lambda_l b_{l,2}^\top b_{l,2} - \frac{1}{2} \sum_{g=1}^{G} \gamma_g u_{g,2}^\top u_{g,2}$$

(5)

$$= \sum_{i=1}^{n} \sum_{g=1}^{G} \log \pi_g \phi_{ig} \left( y_i | X_i, t_i; \theta \right) - \frac{1}{2} \sum_{l=1}^{p} \lambda_l b_{l,2}^\top b_{l,2} - \frac{1}{2} \sum_{g=1}^{G} \gamma_g u_{g,2}^\top u_{g,2},$$

where $Q = \{ Q_i \}_{i=1}^{n}$, and $\phi_{ig} \left( y_i | X_i, t_i; \theta \right) = \phi \left( y_i \left| X_i^\top \otimes \tilde{B}_i \right. \right)^\top b + \tilde{B}_i^\top \tilde{u}_g, \tilde{B}_i^\top \tilde{S}_g \tilde{B}_i + \sigma_g^2 I_m, \right)$. 

At the iteration step $k + 1$ of the penalized EM algorithm, the E-step computes

$$\Omega_P(\theta | \theta^{(k)}; \lambda) = E \left[ \ell^*_P(\theta | O, Q, \lambda) | O, \theta^{(k)} \right]$$

$$= \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{\pi}_{ig} \left( \theta^{(k)} \right) \log \pi_g \phi_{ig} \left( y_i | X_i, t_i; \theta \right) - \frac{1}{2} \sum_{l=1}^{p} \lambda_l b_{l,2}^\top b_{l,2} - \frac{1}{2} \sum_{g=1}^{G} \gamma_g u_{g,2}^\top u_{g,2},$$

where $\hat{\pi}_{ig} \left( \theta^{(k)} \right) = E \left[ q_{ig} | O, \theta^{(k)} \right]$ is the posterior probability for the $i$th subject to belong to the $g$th cluster given parameter $\theta^{(k)}$, i.e.,

$$\hat{\pi}_{ig} \left( \theta^{(k)} \right) = E \left[ q_{ig} | O, \theta^{(k)} \right] = \frac{\pi_{ig}^{(k)} \phi_{ig} \left( y_i | X_i, t_i; \theta^{(k)} \right)}{\sum_{g'=1}^{G} \pi_{ig'}^{(k)} \phi_{ig'} \left( y_i | X_i, t_i; \theta^{(k)} \right)}.$$ 

The M-step updates the parameters by $\theta^{(k+1)} = \arg\max_{\theta} \Omega_P(\theta | \theta^{(k)}; \lambda)$.

The likelihood function of mixture models may have several local maxima. Choosing an appropriate initial value is crucial for finding the global maximum. In our implementations, the initial value is obtained by three steps. First, we fit a single-class functional mixed effects model based on the transformed penalized B-splines and the structural covariance assumption. The parameters and penalty terms can be estimated simultaneously by REML, using existing software such as the function `lme` in the R package `nlme` (Pinheiro and Bates, 2000), with two layers of random effects. Second, we estimate the individual spline coefficients by BLUP (Henderson, 1975), and cluster the estimated coefficients into $G$ groups. We use agglomerative hierarchical clustering for its nested structure, which is more consistent with the model selection method discussed in section 3.3. Third, we estimate the cluster-specific parameters using the individuals in each cluster respectively. From this initial value, denoted by $\theta^{(0)}$, the E-step and M-step are repeated until convergence.

### 3.2. Choice of the smoothing parameters

The smoothing parameters are $\lambda = (\lambda_1, ..., \lambda_p, \gamma_1, ..., \gamma_G)$, which control the nonlinearity of the fixed functions $\beta_l(t) = \tilde{B}(t)^\top \tilde{b}_l$, $l = 1, ..., p$ and the cluster-specific mean functions $\mu_g = \tilde{B}(t)^\top \tilde{u}_g$, $g = 1, ..., G$, respectively.

Since the model parameters $\tilde{b} = (\tilde{b}_1^\top, ..., \tilde{b}_p^\top)^\top$ and $\tilde{u} = (\tilde{u}_1^\top, ..., \tilde{u}_G^\top)^\top$ are estimated based
on a likelihood function, it is natural to consider using the REML method, which Wahba (1985) called ‘generalized maximum likelihood (GML)’, to estimate the optimal smoothing parameters. It regards the penalized likelihood estimates \( \hat{\beta} \) and \( \hat{\theta} \) as the posterior modes of the conditional distribution \((\hat{\beta}, \hat{\theta})|O\), under improper ‘prior’ densities, 

\[
    f_p(\hat{\beta}, \hat{\theta}; \lambda) = \frac{|\Lambda|^{1/2}}{(2\pi)^{(K-2)(p+G)/2}} \exp \left( -\frac{1}{2} \sum_{i=1}^{p} \lambda_i \beta_{i,2}^\top \beta_{i,2} - \frac{1}{2} \sum_{g=1}^{G} \gamma_g \theta_{g,2}^\top \theta_{g,2} \right),
\]

where \( \Lambda = \text{diag}(\lambda_1 I_{K-2}, ..., \lambda_p I_{K-2}, \gamma_1 I_{K-2}, ..., \gamma_G I_{K-2}) \). To obtain the objective function for REML method, we need to integrate \( \hat{\beta} \) and \( \hat{\theta} \) out of

\[
    L(\beta, \theta|O; \lambda, \theta_R) = f_p(\beta, \theta; \lambda) f(O|\beta, \theta_R),
\]

where

\[
    f(O|\beta, \theta_R) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{G} \pi_g \phi_{i,g}(y_i|\xi_i, \eta_i; \beta, \theta) \right\},
\]

\( \theta_R = (\Sigma, \pi) \) and \( \theta = (\beta, \theta, \Sigma, \pi) \). If \( G > 1 \), this integration does not have a closed form where approximation methods must be used. First of all, it involves other unknown parameters \( \theta_R \) in the integral. We need to decide how to handle these unknown parameters. Two feasible methods are: 1) for each given \( \lambda \), iterating the EM algorithm to convergence and then evaluating and minimizing \( \int L(\beta, \theta|O; \lambda, \theta_R) d\{\beta, \theta\} \) to update \( \lambda \), or 2) updating \( \lambda \) at each iteration based on the current estimates of the unknown parameters. Because the first method optimizes a properly defined function of \( \lambda \), convergence is usually guaranteed. However, the computational burden of the first method is very heavy. In contrast, although in the second method the penalty term changes at each iteration for which convergence may not be guaranteed, it is much more efficient in computations. Therefore, the second method is actually very popular in the literature, see e.g. Wood (2004) for likelihood-based methods and also Gu (1992) for prediction-error-based methods. Secondly, even if \( \theta_R \) is given, integrating \( L(\beta, \theta|O; \lambda, \theta_R) \) with regard to \((\beta, \theta)\) is still non-trivial due to the latent class structure. Hence, we propose an iterative REML approach for estimating \( \lambda \), where the smoothing parameters are updated based on an easy-to-compute REML criterion at each EM iteration.

Specifically, the objective function \( \Omega_p(\theta; \theta^{(k)}; \lambda) \) at EM iteration step \( k + 1 \) can be regarded as the logarithm of a ‘joint density’ function defined by

\[
    f_k(\beta, \theta|O; \theta^{(k)}, \lambda, \theta_R) \propto \exp \left\{ \sum_{i=1}^{n} \sum_{g=1}^{G} \pi_{ig} (\theta^{(k)}) \log \phi_{ig}(y_i|\xi_i, \eta_i; \theta) \right\} \times f_p(\beta, \theta; \lambda),
\]

i.e., the estimates \( \hat{\beta}^{(k+1)} \) and \( \hat{\theta}^{(k+1)} \) can be treated as ‘posterior modes’ of this density function. Therefore, we can estimate the smoothing parameter at iteration step \( k + 1 \) by optimizing a REML-type criterion \( \int f_k(\beta, \theta|O; \theta^{(k)}, \lambda, \theta_R) d\{\beta, \theta\} \). Because \( f_k \) is a quadratic function of \( \beta \) and \( \theta \), this integral has an explicit form as

\[
    \omega(\lambda, \Sigma|\theta^{(k)}) = \int f_k(\beta, \theta|O; \theta^{(k)}, \lambda, \theta_R) d\{\beta, \theta\} \\
    \propto |\Lambda|^{\frac{1}{2}} |\Sigma_L^{(k)}|^{-\frac{1}{2}} \exp \left\{ \frac{1}{2} \eta^{(k)} \Sigma_L^{(k)} \eta^{(k)} \right\},
\]
where
\[ \tilde{\Sigma}_\Lambda^{(k)} = \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{\pi}_{i,g}(\theta^{(k)}) W_i \Sigma_{i,g}^{-1} W_i^\top + \Lambda, \]
and \( \eta^{(k)} = \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{\pi}_{i,g}(\theta^{(k)}) W_i \Sigma_{i,g}^{-1} y_i, \)
with \( W_i = \left( (X_i^\top \otimes \hat{B}_i)^\top, \hat{B}_i^\top \right)^\top \) and \( \Sigma_{i,g} = \hat{B}_i^\top \Sigma_g \hat{B}_i + \sigma_g^2 I_{m_g}. \) The function \( \omega(\xi, \hat{\Sigma}|\theta^{(k)}) \) depends on \( \pi \) only through the concurrent parameter estimate \( \theta^{(k)} \), which is only involved in the posterior probability \( \hat{\pi}_{i,g}(\theta^{(k)}) \). Such probability often changes stably across the iterations. More importantly, the variance parameters \( \tilde{\Sigma} \) estimated in the M-step as a weighted mixed-effects model are equivalent to the REML estimates from \( \omega(\lambda, \hat{\Sigma}|\theta^{(k)}) \) (Wang, 1998), i.e., the estimation of \( \lambda \) is compatible with the EM algorithm in the sense of improving the likelihood function at each iteration. Thus, it together with the EM algorithm generally converges. In particular, this method reduces to be the classical REML/GML criterion when \( G = 1 \).

**Remark 3.1.** Although selecting smoothing parameters is a more important issue in penalized splines, we still need to specify the number and location of knots in practice. In general, the results are robust to the number and location of knots, as long as the number exceeds a minimum necessary value and spreads reasonably, typically at fixed quantiles (Ruppert, 2002). A rule of thumb is to select enough knots to fit the data well but not so many as to add unnecessary computational burden. In our experience with longitudinal cognitive data, choosing the number of knots to be close to the average number of observations per subject provides a good balance between approximation accuracy and computational burden. Equal-quantiles of the pooled longitudinal data are suggested for the location of knots.

### 3.3. Selection of the number of latent classes.

The selection of the number of latent classes is a challenging problem in modeling finite mixtures (McLachlan and Peel, 2000). The Bayesian information criterion (BIC; Schwarz, 1978) is by far the most popular choice (Fraley and Raftery, 2002). In this paper, we employ a modified BIC for the evaluation of models estimated by the penalized likelihood (Konishi, Ando and Imoto, 2004). It is based on the Bayesian model interpretation discussed in the previous section. Denote the estimated model parameters at convergence as \( \hat{\theta} = (\hat{\xi}, \hat{\mu}, \hat{\theta}_R) \) with \( \hat{\theta}_R = (\hat{\Sigma}, \hat{\pi}) \), and the estimated smoothing parameters as \( \hat{\lambda} \). We define the BIC as
\[
\text{BIC}(G) = -2 \log L_R(\hat{\theta}_R, \hat{\lambda}; O) + q \log(N),
\]
where
\[
L_R(\hat{\theta}_R, \hat{\lambda}; O) = \int L(\bar{b}, \bar{u}; O; \hat{\lambda}, \hat{\theta}_R) \cdot d\bar{b} \cdot d\bar{u},
\]
\( q \) is the total number of parameters of \( \theta_R \) and \( \hat{\lambda} \), and \( N \) is the total number of longitudinal observations. Note that \( (\bar{b}, \bar{u}) \) is the maximizer of the penalized likelihood \( L(\bar{b}, \bar{u}; O; \hat{\lambda}, \hat{\theta}_R). \)

We can approximate \( L_R(\hat{\theta}_R, \hat{\lambda}; O) \) by Laplace approximation as follows,
\[
L_R(\hat{\theta}_R, \hat{\lambda}; O) \approx L(\bar{b}, \bar{u}; O; \hat{\lambda}, \hat{\theta}_R)(2\pi)^{K/2}(p+G)/2 |\Sigma_N|^{1/2} N^{-K(p+G)/2} 
\approx \exp \left\{ \ell_p(\hat{\theta}|O, \hat{\lambda}) \right\} |\hat{\lambda}|^{1/2} |\Sigma_N|^{1/2} N^{-K(p+G)/2},
\]
where \( \ell_p(\cdot) \) is defined in equation (4), \( \Sigma_N = -N \hat{\Sigma}_{\bar{b}, \bar{u}} \) and \( \hat{\Sigma}_{\bar{b}, \bar{u}} = (D^2 \log L(\bar{b}, \bar{u}; O; \hat{\lambda}, \hat{\theta}_R))^{-1} \) is the inverse of the Hessian matrix evaluated at \( (\bar{b}, \bar{u}) \).

The modified BIC is then defined as
\[
\text{BIC}_M(G) = -2\ell_p(\hat{\theta}|O, \hat{\lambda}) - \log |\Lambda| - \log |\Sigma_N| + (q + K(p+G)) \log(N).
\]
We choose the number of clusters \( G \) that minimizes \( \text{BIC}_M(G) \).
3.4. Classification. Denote by \( c_i \) the latent cluster membership of the \( i \)th individual. The cluster classification of the \( i \)th individual is based on the conditional posterior probability defined as
\[
\hat{\pi}_{ig} = P(c_i = g | y_i, t_i, X_i; \hat{\theta}) = \frac{\hat{\pi}_g \phi_{ig} (y_i | X_i, t_i; \hat{\theta})}{\sum_{g'=1}^G \hat{\pi}_{g'} \phi_{ig'} (y_i | X_i, t_i; \hat{\theta})}.
\]
It quantifies how likely the \( i \)th individual is to be in cluster \( g \), given the longitudinal trajectory \( y_i \) observed at \( t_i \) and conditional on the covariates \( X_i \). The posterior probabilities offer a soft clustering (McLachlan and Peel, 2000) for the longitudinal trajectories. For hard clustering, the \( i \)th individual is assigned to the cluster \( g_i \) if \( \hat{\pi}_{ig_i} = \max_g \{ \hat{\pi}_{ig} \} \). Ideally, the maximal posterior probability \( \hat{\pi}_{ig_i} \) should be close to 1, with the remaining probabilities close to 0. In practice, we can check the goodness of separation by visualizing the soft clustering in barycentric plots, see Figure 5 in Section 6 for an example.

4. Inference. The confidence intervals can be constructed by the Bayesian model interpretation of the penalized likelihood, following the idea of Wahba (1983) and Du and Wang (2014). Specifically, the confidence intervals of the fixed functions \( \hat{\beta}_l \) and the cluster-specific mean functions \( \hat{\mu}_g \) are based on those of \( \hat{\beta} \) and \( \hat{\mu} \). As discussed in section 3.2, we can treat \( (\hat{\beta}, \hat{\mu}) \) as the posterior mode of the conditional density \( (b, u) | O \) denoted by \( L(\hat{\beta}, \hat{\mu}; O; \hat{\lambda}, \hat{\theta}_R) \). We can approximate this distribution through a second order Taylor expansion of \( \log L(\hat{\beta}, \hat{\mu}; O; \hat{\lambda}, \hat{\theta}_R) \) at this posterior mode. It gives a Gaussian distribution with mean \( (\hat{\beta}, \hat{\mu}) \) and covariance \( \hat{\Sigma}_{b, u} \) defined in section 3.3. To construct the confidence intervals of, say, \( \hat{\beta}_l(t) = \hat{B}(t)^\top \hat{b}_l \), we first extract the corresponding variance matrix of \( \hat{b}_l \) from \( \hat{\Sigma}_{b, u} \), denoted \( \hat{\Sigma}_{b_l} \). Then, we have \( \text{var}(\hat{\beta}_l(t)) = \hat{B}(t)^\top \hat{\Sigma}_{b_l} \hat{B}(t) \). The 100(1 - \( \tau \))% confidence interval of \( \hat{\beta}_l(t) \) can be constructed as
\[
\hat{\beta}_l(t) \pm z_{\tau/2} \sqrt{\hat{B}(t)^\top \hat{\Sigma}_{b_l} \hat{B}(t)},
\]
where \( z_{\tau/2} \) is the z-score of the standard Gaussian distribution at level \( \tau \).

The Bayesian confidence intervals are useful tools to visualize the significance of the estimates. But they do not offer formal hypothesis testing such as \( H_0: \beta_l(t) \equiv 0 \) or more specifically \( H'_0: \beta''_l(t) \equiv 0 \) which checks whether \( \beta_l(t) \) is significantly nonlinear. Taking advantage of the basis expansion, we can employ \( \chi^2 \)-test for these hypothesis testing problems (see e.g. Corollary 2, James and Silverman, 2005). For example, to test the hypothesis \( H_0: \beta_l(t) \equiv 0 \), we can use the statistic \( x_0 = \hat{b}_l^\top \hat{\Sigma}_{b_l}^{-1} \hat{b}_l \) whose distribution is approximated by \( \chi^2(K) \); and to test the hypothesis \( H'_0: \beta''_l(t) \equiv 0 \), we can use the statistic \( x_1 = (0, 0, \hat{b}_{l,2}^\top \hat{\Sigma}_{b_l}^{-1} (0, 0, \hat{b}_{l,2}^\top)^\top \) whose distribution is approximated by \( \chi^2(K - 2) \). These statistics are easily available once the model estimates are computed.

5. Simulation. A concern of using the latent class model is whether the number of clusters would be over-estimated, i.e., whether the estimated latent classes are merely an artifact of splitting a homogeneous data into several pieces. Another concern is the possible over/under-smoothing of the functional estimates by the proposed penalized spline approach. We employed a simulation study to check 1) the effectiveness of the proposed BIC_M in preventing over-estimation of latent classes, 2) the estimation accuracy of the fixed functional effects, and 3) the power of the \( \chi^2 \)-square tests for significance and nonlinearity of the functional components.
To simulate data similar to the longitudinal cognitive trajectories seen in the ROSMAP data, the longitudinal response $y_{ij}$ was generated by model

$$y_{ij} = \beta(t_{ij})^T X_i + \alpha_i(t_{ij}) + e_{ij}, \ i = 1, \ldots, n, \ j = 1, \ldots, m_i,$$

where $X_i = (X_{i,1}, \ldots, X_{i,12})$ was a $12 \times 1$ vector of covariates, and $\beta(t)^T = (\beta_1(t), \ldots, \beta_{12}(t))$. Specifically, $X_{i,1}, X_{i,2}$ and $X_{i,3}$ were i.i.d. normal distribution with mean $= 0.75$ and SD $= 0.6$; $X_{i,4}, X_{i,5}$ and $X_{i,6}$ were i.i.d. exponential distribution with mean $= 1.25$; $X_{i,7}, X_{i,8},$ and $X_{i,9}$ were i.i.d. binomial with probability $0.5$; $X_{i,10}, X_{i,11}$, and $X_{i,12}$ were i.i.d. binomial with probability $0.15$. For each of the triplets $X_{i,k}, X_{i,k+1}$ and $X_{i,k+2}$, $k = 1, 4, 7, 10$, the fixed functions were respectively defined as $\beta_k(t) = -0.15 - 0.005t - \exp((t + 5)/5)/5$ which is a nonlinear function similar to the fixed functional effects of the global AD pathology, $\beta_{k+1}(t) = -0.5 - 0.025t$, and $\beta_{k+2}(t) \equiv 0$.

Then, two simulation scenarios were considered for the random curves $\alpha_i(t_{ij})$ and measurement errors $e_{ij}$. In the first scenario, the random curves were generated by a mixture of Gaussian processes $\alpha_i \sim \sum_{g=1}^{4} \pi_g \text{GP}(\mu_g, R_g)$ where $\pi_1 = 0.25$, $\pi_2 = 0.35$, $\pi_1 = 0.1$, and $\mu_1(t) = 0.8 + 0.02t$, $\mu_2(t) = -\exp((t + 5)/5)/2 + 0.1 - 0.01t$, $\mu_3(t) = 0.2 - \exp((t + 5)/5)/2 - \exp((t + 2)/2) - 0.01t$, and $\mu_4(t) = -\exp((t + 2)/2)/10 + 0.2$, all of which were chosen to mimic the estimates in Section 6. The covariance kernels $R_g$ were parameters of less interest. We set $R_g(t, s) = P(t)^T \Sigma_g P(s)$ where $P(t)^T = (1, t, t^2, t^3)$ was the vector of cubic polynomial basis functions, and $\Sigma_g$ was obtained by projecting the covariances estimated in Section 6 with B-splines onto the space of $P(t)$. Thus, the random curves $\alpha_i(t)$ were generated by $P(t)^T \alpha_i$ where $\alpha_i \sim \mathcal{N}(0, \Sigma_g)$. We chose this setting mainly to show the robustness of our method to cases that did not exactly follow a penalized B-spline formulation. The measurement errors $e_{ij}$ were assumed to follow $\mathcal{N}(0, \sigma_e^2)$ with cluster-specific standard deviations that $\sigma_e = 0.17, 0.17, 0.25$ and $0.48$ respectively for $g = 1, 2, 3, 4$. This scenario was designed to check the accuracy of estimation and hypothesis testing on the functional components.

In the second scenario, the random curves were generated by a single Gaussian process $\alpha_i \sim \text{GP}\left(\sum_{g=1}^{4} \pi_g \mu_g, \sum_{g=1}^{4} \pi_g R_g\right)$ where $\pi_g$, $\mu_g$ and $R_g$ were the same as the first scenario; and the measurement errors $e_{ij}$ were i.i.d. $\mathcal{N}(0, \tilde{\sigma}_e^2)$ where $\tilde{\sigma}_e = \sqrt{\pi_1 \times 0.17^2 + \pi_2 \times 0.17^2 + \pi_3 \times 0.25^2 + \pi_4 \times 0.48^2} \approx 0.24$. This scenario was designed to check if the proposed BIC$_M$ would over-estimate the number of latent classes when the true data were homogenous.

We simulated 100 datasets for each model. Each dataset contained 1000 subjects. For the longitudinal observations of the $i$th subject, we first generated a temporary sequence of time points $(t_{i,-24}, \ldots, t_{i,-1}, t_{i,0})$, where $t_{i,k}$ followed a truncated normal distribution in the interval $[k - 0.5, k + 0.5]$ with mean $= k$ and SD $= 0.2$, $k = -24, \ldots, -1$, and $t_{i,0}$ followed a truncated normal distribution in $[-0.5, 0]$ with mean $= 0$ and SD $= 0.2$. Then, we generated the length of follow-up time $T_i$ by a Beta($-2, 3.5$) distribution multiplied by 25. Only the time points greater than $-T_i$ were retained in the dataset. This created an unbalanced longitudinal dataset similar to the ROSMAP data.

The frequency of selected number of clusters is reported in Table 1. In the first scenario, the true 4 clusters were selected the most in the replications, with some selected 3 because of the small designed probability of the 4th cluster. Very few selected 5, indicating that the proposed BIC$_M$ protected against model over-estimation very well. This was further confirmed in the second scenario, where the true 1 cluster was selected in almost all replications.

As an illustration, Figure 2 and 3 respectively show the estimated clusters and functional effects of one replication in the first scenario. Notably, both the nonlinear and the linear functional effects are estimated very well, which shows that our proposed model can handle both nonlinear and linear patterns in a unified framework.
TABLE 1
Frequency of the number of clusters selected by $BIC_M$ in simulation study.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$G_0$</th>
<th>$G = 1$</th>
<th>$G = 2$</th>
<th>$G = 3$</th>
<th>$G = 4$</th>
<th>$G = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1: $G_0 = 4$</td>
<td>0</td>
<td>0</td>
<td>38</td>
<td>60</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Scenario 2: $G_0 = 1$</td>
<td>98</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Cluster 1

Cluster 2

Cluster 3

Cluster 4

FIG 2. Estimated clusters of an example in the simulation study. The thick solid curves in each panel are the estimated mean functions of each cluster; the dotted lines are the 95% point-wise confidence intervals; the dashed lines are the corresponding true mean functions.

The accuracy of the estimated functional effects were assessed by the empirical mean squared error (MSE) defined by $\frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (\beta_k(t_{ij}) - \hat{\beta}_k(t_{ij}))^2$, where $m_i$ is the number of longitudinal observations for the $i$th subject, and $N = \sum_{i=1}^{n} m_i$. Table 2 summarizes the MSE and hypothesis testing results. In scenario 1, the model with 4 latent clusters clearly outperforms the single class model in the accuracy of both the fixed functional effects estimation and the hypothesis testing. In scenario 2, the performance of the model with 4 latent clusters was similar to the single class model in fixed functional effects estimation, but worse in hypothesis testing because of model misspecification.
Fig 3. Estimated functional effects of an example in the simulation study. The solid lines are the estimated functions; shaded areas are the 95% point-wise confidence intervals; dashed lines are the corresponding true functional effects; and dotted lines are the horizontal lines at zero.

6. Application to the ROSMAP Data. We applied the latent class functional mixed-effects model (2) to the longitudinal global cognition data described briefly in the Introduction. The main objective of our analyses was to characterize trajectories of norma-


tive cognitive aging. Cognitive change in late life is widely believed to reflect normative (i.e., affecting nearly all individuals) age-related processes and non-normative (i.e., affecting some individuals but not others) pathology-related and mortality-related processes, see e.g. Steinerman et al. (2010). Normative cognitive aging is thought to involve decline in function but the onset and rate of this normative cognitive decline are uncertain, owing to several factors. First, there is substantial heterogeneity in late life cognitive trajectories with little change in some persons and complex nonlinear change in others, making it difficult to model change uniformly. Second, there are multiple neurodegenerative and cerebrovascular conditions in old age that are associated with cognitive impairment. These conditions, which are not part of the normative age process, need to be controlled for in the analyses. Third, even after adjustment for neuropathologic burden, decline in cognitive function proximate to death still varies substantially across subjects, implying potential subtypes of cognition aging. The proposed model (2) addresses all three of these problems.

For robust model estimation, persons included in the analysis had to meet 4 criteria: 1) absence of dementia at the baseline clinical evaluation; 2) valid composite global cognition score at baseline and at least one follow-up evaluation; 3) death with a brain autopsy; and 4) a completed neuropathologic examination. In total we had 1010 individuals included in the analysis with 9331 longitudinal observations. Since the measured neuropathologies mostly reflect the brain condition at a time shortly after death, for ease of interpretation of the results, we aligned all longitudinal trajectories by the date of death, i.e, the time scale for longitudinal analysis was the time in years before death. In this sample, participants underwent annual cognitive testing for a mean of 9.9 years before death (SD = 5.0; range: 0.8 to 24.0). At the baseline clinical evaluation, 307 (30.4%) persons met criteria for mild cognitive impairment (MCI). On follow-up, 258 (25.5%) persons developed incident MCI, 249 (24.7%) developed incident dementia, and 196 (19.4%) remained free of cognitive impairment throughout the follow-ups.

### Table 2

Estimation and hypothesis testing results for the fixed functional effects in simulation study. Columns $H_0$ and $H_0^{[2]}$ respectively represent the frequency of $p < 0.05$ for the tests $f(t) \equiv 0$ and $f^\prime(t) \equiv 0$ in the 100 replications.

<table>
<thead>
<tr>
<th>Scenario 1: $G_0 = 4$</th>
<th>Scenario 2: $G_0 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model $G = 4$</strong></td>
<td><strong>Model $G = 1$</strong></td>
</tr>
<tr>
<td><em><em>MSE</em> $H_0$</em>*</td>
<td><em><em>MSE</em> $H_0^{[2]}$</em>*</td>
</tr>
<tr>
<td>$X_1$ 0.173 100 97 0.326 100 96</td>
<td>$X_{12}$ 0.319 0 0 0.601 1 0 0.283 2 1 0.267 4 1</td>
</tr>
<tr>
<td>$X_2$ 0.144 100 6 0.259 100 10</td>
<td>$X_3$ 0.306 100 0 0.510 90 1 0.286 100 0 0.280 100 0</td>
</tr>
<tr>
<td>$X_3$ 0.130 13 3 0.268 16 6</td>
<td>$X_4$ 0.359 100 95 0.601 97 49 0.350 100 69 0.322 100 63</td>
</tr>
<tr>
<td>$X_5$ 0.071 100 1 0.136 100 2</td>
<td>$X_6$ 0.510 90 1 0.286 100 0 0.280 100 0</td>
</tr>
<tr>
<td>$X_6$ 0.064 2 0 0.112 1 1</td>
<td>$X_7$ 0.154 100 99 0.317 100 96 0.177 100 100 0.152 100 98</td>
</tr>
<tr>
<td>$X_7$ 0.149 100 1 0.259 100 3</td>
<td>$X_8$ 0.149 100 1 0.259 100 3</td>
</tr>
<tr>
<td>$X_8$ 0.174 3 0 0.322 4 0</td>
<td>$X_9$ 0.174 3 0 0.322 4 0</td>
</tr>
</tbody>
</table>

*All MSE numbers were multiplied by 100 for clearer illustration.*
To estimate associations of neuropathologies with cognition, we included 9 common post-mortem neuropathologic markers in the model: composite AD pathology, TDP-43 pathology, cerebral amyloid angiopathy, neocortical Lewy bodies, hippocampal sclerosis, chronic gross infarct, chronic microinfarct, atherosclerosis, and arteriolosclerosis. Specifically, the composite AD pathology, TDP-43 pathology, and cerebral amyloid angiopathy were treated as continuous variables. The remaining 6 pathologies were either binary by measurement or transformed into binary variables based on the presence of moderate to severe levels of neuropathology. In this sample, the composite measure of AD pathology ranged from 0 to 3.07 (mean = 0.75, SD = 0.60) and the composite measure of TDP-43 pathology ranged from 0 to 4.67 (mean = 0.64, SD = 0.96); 13.5% had neocortical Lewy bodies and 8.9% had hippocampal sclerosis. The composite measure of cerebral amyloid angiopathy ranged from 0 to 4.00 (mean = 1.14, SD = 1.06) and 35.7% had at least one gross infarct, 28.5% had at least one microinfarct, 30.3% had moderate to severe atherosclerosis, and 29.5% had moderate to severe arteriolosclerosis.

We used cubic B-splines with 10 equal-quantile interior knots to generate the transformed spline basis functions. The optimal number of clusters that minimizes $BIC_M$ was 4 (Table 3). Each individual was classified into the cluster that had the highest posterior probability. Figure 4 shows the residual cognitive profiles of the 4 clusters, where the order of the clusters was chosen based on the overall trend of decline of the cluster-specific means, from high to low. The 4 clusters had very different within-subject variation, with $\sigma_1 = 0.174$, $\sigma_2 = 0.167$, $\sigma_3 = 0.251$ and $\sigma_4 = 0.480$. The mean function of cluster 1 was generally linear, while the other 3 clusters all had significantly nonlinear mean functions (Table 4). Cluster 1 ($n = 237, 23.5\%$) had slightly increasing residual cognitive trajectories, and cluster 2 ($n = 360, 35.6\%$) had relatively flat trajectories, with cluster 1 functioning at a slightly higher level than cluster 2. In contrast, cluster 3 ($n = 324, 32.1\%$) exhibited moderate decline and cluster 4 ($n = 89, 8.8\%$) exhibited precipitous residual cognitive decline. These data suggest that non-pathologic late life cognitive decline is confined to subset of older persons and mostly occurs proximate to death.

<table>
<thead>
<tr>
<th># of Clusters</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># of parameters ($n_p$)</td>
<td>200</td>
<td>220</td>
<td>240</td>
<td>260</td>
<td>280</td>
</tr>
<tr>
<td>$-2\ell_P$</td>
<td>7025.83</td>
<td>5164.02</td>
<td>5038.87</td>
<td>4877.83</td>
<td>4825.26</td>
</tr>
<tr>
<td>$-\log</td>
<td>\Lambda</td>
<td>$</td>
<td>-1907.81</td>
<td>-2787.46</td>
<td>-3421.65</td>
</tr>
<tr>
<td>$-\log</td>
<td>\Sigma_N</td>
<td>$</td>
<td>555.57</td>
<td>1334.69</td>
<td>1856.89</td>
</tr>
<tr>
<td>$n_p \log N$</td>
<td>1828.22</td>
<td>2011.04</td>
<td>2193.86</td>
<td>2376.69</td>
<td>2559.51</td>
</tr>
<tr>
<td>$BIC_M$</td>
<td>7501.80</td>
<td>5722.30</td>
<td>5667.98</td>
<td>5593.30</td>
<td>5624.50</td>
</tr>
</tbody>
</table>

To check the goodness of separation of the posterior probabilities, we transformed the estimated probabilities of $i$th individual, denoted by $\hat{\pi}_i = (\hat{\pi}_{i1}, \ldots, \hat{\pi}_{i4})$, into a 3D barycentric coordinate system. Each $\hat{\pi}_i$ is represented as a point in the triangular pyramid. Each of the four corners of the triangular pyramid stands for a cluster. For a particular cluster $g$, the point for the $i$th individual would be right on the $g$th corner if $\hat{\pi}_{ig} = 1$. Generally, the point would lie in the pyramid where the distance to the $g$th corner was positively correlated with $1 - \hat{\pi}_{ig}$. The 3D barycentric coordinates were then projected into a 2D plot for illustration. The algorithm was implemented by the function `quadplot` in R package klaR (Garczarek, 2002). Figure 5 shows the details of the 2D plot. We observed that overall the four clusters...
were reasonably well separated. Most points were located near a corner. We checked the number of individuals for whom the differences of the top two largest posterior probabilities were smaller than 0.1, where it was uncertain to which group the individual can be classified. Only 97 (less than 1%) out of 1010 individuals were in this category. Most of them were between the adjacent clusters: 53 (54.6%) were between clusters 1 and 2, 29 (29.9%) were between clusters 2 and 3, 6 (6.2%) were between clusters 3 and 4; the remaining 9 (9.3%) were between cluster 1 and 3. In particular, if we focus on the non-decliners (cluster 1 and 2) vs. decliners (cluster 3 and 4), only 38 individuals had uncertain classifications.

Figure 6 shows the estimated fixed functions of each neuropathologic marker on global cognitive trajectories. For comparison, we also plotted the linear component of each fixed function. The estimated functional effects of AD, Lewy bodies and TDP-43 were clearly nonlinear. Specifically, the difference in rates of cognitive decline between individuals with different AD pathologies became evident about a decade before death and progressed as death approached, whereas important associations of Lewy bodies and TDP-43 with cogni-
Table 4: P-values of the $\chi^2$-tests for non-zero and non-linear functional effects in the ROSMAP data.

<table>
<thead>
<tr>
<th>Cluster-specific mean functions:</th>
<th>$H_0: x(t) \equiv 0$ (Null: zero effect)</th>
<th>$H_0: x''(t) \equiv 0$ (Null: linear effect)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>&lt; 0.001</td>
<td>&gt; 0.999</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Functional effects of pathologies:</th>
<th>$H_0: x(t) \equiv 0$ (Null: zero effect)</th>
<th>$H_0: x''(t) \equiv 0$ (Null: linear effect)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD pathologies</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Lewy bodies</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>TDP-43</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Hippocampal sclerosis</td>
<td>&lt; 0.001</td>
<td>&gt; 0.999</td>
</tr>
<tr>
<td>Atherosclerosis</td>
<td>0.032</td>
<td>0.474</td>
</tr>
<tr>
<td>Arteriosclerosis</td>
<td>&gt; 0.999</td>
<td>&gt; 0.999</td>
</tr>
<tr>
<td>Gross infarcts</td>
<td>0.974</td>
<td>&gt; 0.999</td>
</tr>
<tr>
<td>Microinfarcts</td>
<td>&gt; 0.999</td>
<td>&gt; 0.999</td>
</tr>
<tr>
<td>Cerebral amyloid angiopathy</td>
<td>0.004</td>
<td>0.011</td>
</tr>
</tbody>
</table>

The nonlinear effect of cerebral amyloid angiopathy was statistically significant, the magnitude of this effect is relatively small. The nonlinear effects of the remaining 5 pathologies were not significant. Among them, two pathologies showed statistically significant effects: the effect of hippocampal sclerosis is virtually linear; and the effect of atherosclerosis is largely constant over time. Table 4 shows the details of the hypothesis tests.

7. Discussion. We have proposed a new conditional functional clustering approach for heterogeneous longitudinal trajectories with nonlinear patterns. This approach reveals the latent clusters of longitudinal trajectories after adjusting the functional effects of given covariates. The proposed model (2) as a mixture model could be regarded as a non-trivial functional extension of the classical mixture-of-experts method, which is typically designed for balanced cross-sectional multivariate data (Murphy and Murphy, 2020), to handle unbalanced longitudinal data with heterogeneous nonlinear patterns. The “experts” in the model (2) are the $G$ class-specific conditional distributions of $y_i$ given $X_i$, i.e., $N(\beta(t_i)^\top X_i + u_{i,g}, \Sigma_{i,g} + \sigma^2_g I_{m_i})$, for $g = 1, G$. Notably, the functional coefficients $\beta(t_i)$, the mean parameters $u_{i,g}$, and the covariance matrix $\Sigma_{i,g} + \sigma^2_g I_{m_i}$ all depend on the longitudinal measurement times $t_i$, and thus are person-specific instead of fixed for each cluster in a typical mixture-of-experts setting. In addition, we constrain the functional coefficients $\beta(t_i)$ to be the same across clusters for conditional clustering, in which it is different from the finite mixture of regressions (see, e.g. Yao, Fu and Lee, 2010).

Both the functional coefficients $\beta(t)$ and the flexible random curves $\alpha_i(t)$ require parsimonious modeling for robust parameter estimation and more interpretable clustering. To this end, we propose a transformed B-spline expansion to model these functional components, which enables a simple penalized estimation of the functional coefficients $\beta(t)$ and a novel parsimonious assumption of the covariance matrix for the random coefficients of $\alpha_i(t)$. Because the functional coefficients $\beta(t)$ are shared by all clusters in a heterogeneous dataset, for which a satisfactory method to select the smoothness penalty is still lacking, we propose a novel iterative-REML method that extends the classical REML/GML method (Wahba, 1985;
Barycentric plot of the estimated posterior probabilities for cluster classification in the ROSMAP data. This figure is obtained by projecting the 3D triangular pyramid onto a 2D plot. The 4 corners of the pyramid are located respectively at the three outer corners and the knot at left upper part of the triangle of the 2D plot. The star in the center represents the center of the pyramid. Each dashed line is the fragment line from the star to the middle point of each edge of the pyramid, which is the separation line of two adjacent clusters.

Wang, 1998) to this new setting. Finally, we use a modified BIC to select the number of clusters, which leverages the Bayesian interpretation of the proposed method.

In our application to the ROSMAP data, we identified four subtypes of longitudinal cognitive decline in late-life aging process. With the same functional coefficients for all participants, the residual global cognition trajectories used for clustering are comparable to each other, offering a more interpretable clustering result. Namely, after adjustment for the deleterious effects of non-normative pathologic processes, residual global cognitive decline was seen in less than half of the participants (cluster 3 and 4), which was related to mortality more than age as the decline was more precipitous approximate to death. The results suggest that most late-life cognitive loss is attributable to non-normative processes related to brain pathology or mortality and provide little support for the idea that a meaningful proportion of late life cognitive loss reflects normative age-related processes. It is worthy noting that the atypical residual trajectories with extremely high and low values are almost all included in the cluster 4. This is because in model (2), we allow both the covariance kernel $R_g$ and error variance $\sigma_g^2$ to be different across clusters. When there exist atypical trajectories that cannot be clustered with others, this model tends to use a separate cluster with a relaxed $R_g$ and a large $\sigma_g^2$ to accommodate these trajectories. In this way, these atypical trajectories would not affect the inference of other main clusters. Also, because the functional coefficients $\beta(t)$ are estimated using all subjects in all clusters, these atypical trajectories tend to be rare and thus have very limited impact on the inference of $\beta(t)$.

The proposed method was mainly motivated by the analysis of unbalanced longitudinal cognitive data observed in cohort studies, with the ability to handle nonlinear longitudinal
Estimated fixed functions of the neuropathologic markers. The solid lines are the estimated functions; shaded areas are the 95% point-wise confidence intervals; dashed lines are the linear components of each function; and dotted lines are the horizontal lines at zero.

Trajectories observed at sparse and irregular times. But the unbalanced data are merely additional complexity rather than requirements to apply the proposed method. Extension to other applications with dense and/or balanced measurements, such as time-course gene expression data, is fairly straightforward.

Acknowledgement. This research was funded by the NIH National Institute on Aging (R01AG17917, P30AG10161, R01AG15819, R01AG34374) and the Illinois Department of Public Health. We thank the study participants of the ROS and MAP cohorts, and the staffs of Rush Alzheimer’s Disease Center.
REFERENCES


Murphy, K. and Murphy, T. B. (2020). Gaussian parsimonious clustering models with covariates and a noise component. Advances in Data Analysis and Classification 14 293–325.


