Biostatistics (2020), **0**, 0, *pp.* 1–31
doi:10.1093/biostatistics/output

Penalized longitudinal mixed models with latent group structure, with an application in neurodegenerative diseases

FARHAD HATAMI[1,*], KONSTANTINOS PERRAKIS[2], JOHNATHAN COOPER-KNOCK[3],

SACH MUKHERJEE[4], FRANK DONDELINGER[1]

[1] *Centre for Health Informatics, Computation and Statistics, Lancaster Medical School, Lancaster University, UK*
[2] *Department of Mathematical Sciences, Durham University, UK*
[3] *Department of Neuroscience, The University of Sheffield*

[4] *German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany*

f.hatami@lancaster.ac.uk

Summary

Large-scale longitudinal data are often heterogeneous, spanning latent subgroups such as disease subtypes. In this paper, we present an approach called *longitudinal joint cluster regression* (LJCR) for penalized mixed modelling in the latent group setting. LJCR captures latent group structure via a mixture model that includes both the multivariate distribution of the covariates and a regression model for the response. The longitudinal dynamics of each individual are modeled using a random effect intercept and slope model. Inference is done via a profile likelihood approach that can handle high-dimensional covariates via ridge penalization. LJCR is motivated by questions in neurodegenerative disease research, where latent subgroups may reflect heterogeneity with respect to disease presentation, progression and diverse subject-specific factors. We

*To whom correspondence should be addressed.

study the performance of LJCR in the context of two longitudinal datasets: a simulation study and a study of amyotrophic lateral sclerosis (ALS). LJCR allows prediction of progression as well as identification of subgroups and subgroup-specific model parameters.

Key words: neurodegenerative disease, heterogeneity, longitudinal data, fixed-random effect models, clustering, latent group structure

1. Introduction

Longitudinal designs play a key role in biomedical research. In these studies, repeated measurements of the same quantities enable the study of temporal processes such as disease progression. Contemporary large-scale longitudinal datasets may include large numbers of observed variables, and are often heterogeneous, spanning multiple data subgroups such as disease subtypes. This leads to a subgroup structure that is often latent. In this situation, classical longitudinal models may be confounded by the latent group structure, or may simply be inapplicable due to the number of covariates.

In this paper, we propose an approach called *longitudinal joint cluster regression* (LJCR) for the heterogeneous data case that extends classical mixed modelling via a regularised mixture framework. In summary our approach posits models specific to latent subgroups indexed by k, i.e.:

$$\begin{aligned} y_{ijk} &= \alpha_k + \Lambda_{ik}(t_{ijk}) + \epsilon_{ijk}, \\ \Lambda_{ik}(t) &= \mathbf{x}_i(t)^T \boldsymbol{\beta}_k + \mathbf{z}_i(t)^T \mathbf{b}_{ik}, \end{aligned} \tag{1.1}$$

where y_{ijk} is the response for subject i at measurement j in subgroup k, α_k is the subgroup-specific intercept, ϵ_{ijk} are the (usually Gaussian-distributed) residuals. The term $\Lambda_{ik}(t)$ captures the time-dependent dynamics, with $\mathbf{x}_i(t)$ the vector of covariates, $\boldsymbol{\beta}_k$ the group-specific fixed effects, $\mathbf{z}_i(t)$ the time-dependent covariates and $\mathbf{b}_{ik}$ the subject-specific random effects.

The subgroup-specific model parameters are a key feature of our model, which allows both

the temporal dynamics, such as rate of progression, as well as the regression parameters $\boldsymbol{\beta}_k$ to differ between subgroups. LJCR estimates K models of the form (1.1). The subgroup labels are treated as latent, which allows us to cope with the situation where the subgroup structure is entirely unknown at the outset.

LJCR is thus a joint modelling approach aimed at capturing heterogeneous longitudinal dynamics of disease progression by combining clustering, regression and linear mixed modelling. As described in detail below, LJCR considers both the distribution of $\mathbf{Y}|\mathbf{X}$ and the distribution of $\mathbf{X}$. This is done within a mixture framework, extending recent work by Perrakis et al. [2019] to the mixed model setting. Like Perrakis et al. [2019], we employ a joint cluster regression approach with regularization, allowing for group/cluster-specific regression parameters via a latent variable model. To deal with longitudinal dynamics, we incorporate a linear mixed effects intercept and slope model, and we develop a combination of L1 and L2 penalization and an efficient inference method to deal with small n and moderate-to-large p scenarios. Our method treats both the outcome variable and the explanatory variables as random quantities whose covariance matrix can be estimated. To find the optimal number of latent clusters within a given range, we employ a heuristic based on the elbow technique [Joshi and Nalwade, 2013].

Our work is motivated by challenges in longitudinal data analysis in the study of neurodegenerative diseases (NDDs). These diseases have complex underlying aetiology and display considerable heterogeneity in presentation and progression. Furthermore, as for many complex diseases in neurology and psychiatry, disease subtyping remains an open area of investigation. Hence one cannot typically assume that all subjects in a given study follow the same distribution, nor that subgroups are known at the outset. In general, NDD patients are characterized by heterogeneous progression profiles, leading to very different disease trajectories and increases in impairment that progress at different time-scales for each individual. This effect is particularly striking in motor neurone disease, or amyotrophic lateral sclerosis (ALS), a disease targeting the voluntary motor

neurons. Most people with ALS succumb to the disease within 2-4 years, but around 10% of affected people survive more than 10 years [Swinnen and Robberecht, 2014]. The drivers behind these differences in progression remain incompletely understood. It is therefore imperative to develop computational methods that can infer underlying subgroup structure from observed data and leverage this to predict long- and short-term progression.

Modelling heterogeneous data is an active area for statistical research. For example, Dondelinger et al. [2020] used a joint penalized regression approach for estimation of high-dimensional fixed effects in heterogeneous data, but their approach is for cross-sectional rather than longitudinal data, and does not include random effects. Graphical models for heterogeneous data have also been considered in the literature [Danaher et al., 2011]. Various statistical models have been developed to infer both regression and group structures (latent variables); for example using regularized or unregularized mixture models [McLachlan], such as in [Khalili and Chen, 2007, Städler et al., 2010] where regularized mixtures of regressions were employed. Alternatively, [Xu et al., 2015] developed a multi-task approach using regularized LU-decomposition to map individual-specific models from k latent base models, where each individual-specific model is a linear combination of the base models. This approach, although flexible, is less convenient for identifying well-defined groups. Suresh et al. [2018] developed a deep learning multi-task model based on a LSTM (Long short-term memory) architecture. Patient groupings were first learned by clustering the embeddings of an autoencoder with LSTM structure using a standard Gaussian mixture model. Then a second neural net with a common LSTM layer and group-specific dense hidden layers was used to produce predictions for each patient. As this model is highly non-linear, interpreting the influence of specific variables becomes difficult.

An alternative approach is to extend conventional clustering approaches to the longitudinal setting. For instance, the k-means method for longitudinal data [Genolini and Falissard, 2010] is an implementation of k-means specifically designed to clustering longitudinal data. For a given

number of clusters (K), the algorithm determines a clustering of individuals, where the progression scores $\mathbf{y}_i = (y_i(t_{i,1}), ..., y_i(t_{i,n_i}))^T$ at time points $t_{i,1}$ to t_{i,n_i} are treated as observations for each individual, and the Euclidean distance with Gower correction is used to optimise the cluster assignments. Note that this method does not take any covariates into account, and will not work very well in cases where none of the observation times coincide across individuals. Another consideration is how to detect the optimal number of clusters (subgroups or subtypes) using such methods [Everitt et al., 2001]. Various efforts have been made, either using nonparametric [Ray and Turi, 1999, Davies and Bouldin, 1979] or parametric approaches [Hurvich and Tsai, 1989, Schwarz et al., 1978].

None of the models described above take the distribution of the features $\mathbf{X}$ into account. Mixture regression models propose a mixture approach for the conditional distribution of $\mathbf{Y}|\mathbf{X}$ and hence solely deal with the relation between the response variable $\mathbf{Y}$ and feature matrix $\mathbf{X}$, disregarding any signal that would arise from the distribution of $\mathbf{X}$ itself. This would make it difficult to predict the response value (for example progression of a disease) on new intakes (patients) with new design matrix $\mathbf{X}^*$, as we cannot assign these patients to a specific group, and would have to average across all possible groups. Motivated by this gap, Perrakis et al. [2019] extended the framework of mixture regressions to include the distribution of the features in the estimation of the latent group membership variable. Our work builds on this to incorporate longitudinal dynamics. Via a combination of the profile likelihood [Pinheiro and Bates, 2000] and some simple linear algebra, we show how an efficient Expectation-Maximisation algorithm can be developed, allowing for scaling to large p scenarios.

The remainder of the paper is organised as follows. We first present the methodological framework of our method, before describing the results of an in-depth simulation study to characterise its performance. We then apply our method to a real-world dataset of ALS patients, and analyse both the predictive performance of our model, as well as the properties of the inferred groups

and the main features associated with longitudinal progression

2. METHODOLOGY

Our model consists of a joint mixture regression model similar to the one described in Perrakis et al. [2019], but where the fixed effects regression model is replaced with a mixed effects model with a random intercept and slope term, as first explored in Laird and Ware [1982], to represent the longitudinal dynamics. In order to deal with small subgroup sizes, we apply l2 regularisation to the fixed effects, which requires the development of an efficient inference algorithm based on the profile likelihood [Pinheiro and Bates, 2006]. For ease of exposition, we first describe the linear mixed model and inference in Subsection 2.1 for the case where the subgroups are known. We then describe the mixture model in Subsection 2.2 and present the full expectation-maximization (EM) algorithm for inference in the combined model in Subsection 2.3.

We have M observational units (usually patients or study subjects). Suppose $\mathbf{y}_i = (y_i(t_{i,1}), ..., y_i(t_{i,n_i}))^T$ denotes an n_i-dimensional vector of responses at time points $\mathbf{t}_i$, and $\mathbf{X}_i$ denotes an $n_i \times p$ matrix of observed covariates for observational unit $i \in \{1, \ldots, M\}$. Let $k \in \{1, \ldots, K\}$ denote the group label and $z_i \in \{1, \ldots, K\}$ represent the true (latent) group label indicator for the sample $(\mathbf{y}_i, \mathbf{X}_i)$ with $p(z_i = k) = \tau_k$. Let $\mathbf{y}$ be the stacked response vector of length $N = \sum_i^M n_i$ collecting the $\mathbf{y}_i$, and let $\mathbf{X}$ be the stacked design matrix collecting the $\mathbf{X}_i$. Define $\boldsymbol{\theta}_k^X$ and $\boldsymbol{\theta}_k^Y$ to be the group-specific parameters; respectively parameterizing the marginal distribution of $\mathbf{X}$ and the regression model of $\mathbf{y}$ on $\mathbf{X}$.

2.1 *Linear mixed effects model for longitudinal dynamics*

We model the longitudinal dynamics of the outcome variable y_i with a mixed effects model:

$$p\left(\mathbf{y}_i|\boldsymbol{\theta}_k^Y, \mathbf{X}_i, z_i = k\right) \equiv p\left(\mathbf{y}_i|\alpha_k, \boldsymbol{\beta}_k, \mathbf{V}_i, \mathbf{X}_i, z_i = k\right) = \mathcal{N}\left(\mathbf{y}_i|\alpha_k + \mathbf{X}_i\boldsymbol{\beta}_k, \mathbf{V}_i\right) \tag{2.2}$$

In other words, we introduce a dependency between longitudinal observations of patient i via the covariance matrix $\mathbf{V}_i$. If $\mathbf{V}_i$ is diagonal with diagonal elements σ_k^2 then a fixed effects model is recovered.

We define V_i using a standard longitudinal mixed model approach with random effects for the intercept and slope (Laird and Ware [1982]). Conceptually, we model $\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\theta}_k^Y, z_i = k$ as

$$\mathbf{y}_i = \alpha_k + \mathbf{X}_i\boldsymbol{\beta}_k + b_{1,i} + b_{2,i}\mathbf{t}_i + \boldsymbol{\epsilon}_i, \tag{2.3}$$

where $\mathbf{b}_i = (b_{1,i}, b_{2,i})^T \sim \mathcal{N}(0, \mathbf{D}_k)$ and $\epsilon_i \sim \mathcal{N}(0, \mathbf{I}_{n_i}\sigma_k^2)$. Here $\mathbf{I}_{n_i}$ is the $n_i \times n_i$ identity matrix. Note that $\mathbf{b}_i$ implicitly depends on $z_i = k$; we chose not to make this explicit in the notation to avoid a redundant subscript. It follows that

$$\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i^T + \sigma_k^2\mathbf{I}_{n_i}, \tag{2.4}$$

where $\mathbf{Z}_i = (1, \mathbf{t}_i)$ is the $n_i \times 2$ design matrix of random effect covariates (in our case, an intercept and the observation time variable). The notation introduced in eqs. (2.2-2.4) differs slightly from the more standard notation for longitudinal models in eq. (1.1), but will simplify exposition in what follows.

Our method needs to be robust to low sample sizes and large numbers of covariates. We use L_2 penalization to regularize the model and allow for efficient estimation of the fixed effect parameters. In the following, we drop the latent group indicator $z_i = k$ and assume that the group labels are known.

The linear model for the response $\mathbf{y}_i$ can then be written as:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b}_i + \epsilon_i \text{ such that } \sum_{j=1}^{p} \beta_j^2 \leqslant \tau. \tag{2.5}$$

In general, our model will include an intercept term, as in eq. (2.3); to simplify notation we assume that this has been integrated into the design matrix as an additional column of ones. The corresponding objective function takes the form:

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b}_i)^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b}_i) + \lambda||\boldsymbol{\beta}||_2^2. \tag{2.6}$$

Straightforward optimization of this objective function would involve inverting M $p \times p$ matrices (where M is the number of patients). For large M and p this is not computationally feasible. Instead, we describe a more computationally efficient approach using the QR decomposition.

We follow Pinheiro and Bates [2000] in augmenting the vectors $\mathbf{y}_i$ and design matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ using a pseudo-data approach. First, note that the likelihood has the form

$$\begin{aligned} L\left(\boldsymbol{\beta}, \mathbf{D}, \sigma^2 | \mathbf{y}\right) &= p\left(\boldsymbol{\beta}|\tau^2\right) \prod_{i=1}^{M} p\left(\mathbf{y}_i | \boldsymbol{\beta}, \mathbf{D}, \sigma^2\right) \\ &= p\left(\boldsymbol{\beta}|\tau^2\right) \prod_{i=1}^{M} \int p\left(\mathbf{y}_i | \boldsymbol{b}_i, \boldsymbol{\beta}, \sigma^2\right) p\left(\boldsymbol{b}_i | \mathbf{D}, \sigma^2\right) d\boldsymbol{b}_i, \end{aligned} \tag{2.7}$$

where the $p\left(\boldsymbol{\beta}|\tau^2\right)$ term is a multivariate normal prior with variance τ^2 inducing a ridge penalization with parameter $\lambda = \sigma^2/\tau^2$. Let us parameterise the multivariate normal distribution for $\mathbf{b}_i$ as

$$p\left(\boldsymbol{b}_i | \boldsymbol{\theta}, \sigma^2\right) = \frac{\exp\left(-\boldsymbol{b}_i^T \mathbf{D}^{-1} \boldsymbol{b}_i\right)}{(2\pi)^{q/2}\sqrt{|\mathbf{D}|}} = \frac{\exp\left(-\left\|\boldsymbol{\Delta}\boldsymbol{b}_i\right\|^2 / 2\sigma^2\right)}{(2\pi\sigma^2)^{q/2}\,\mathrm{abs}|\boldsymbol{\Delta}|^{-1}}, \tag{2.8}$$

where $\sigma^2\mathbf{D}^{-1} = \boldsymbol{\Delta}^T\boldsymbol{\Delta}$, and $\boldsymbol{\theta}$ denotes the free parameters in $\mathbf{D}$ (or equivalently $\boldsymbol{\Delta}$). Note that in our case the number of random effect parameters $q = 2$. We can define augmented vectors and design matrices,

$$\tilde{\mathbf{y}}_i = \begin{bmatrix} \mathbf{y}_i \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\boldsymbol{X}}_i = \begin{bmatrix} \boldsymbol{X}_i \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\boldsymbol{Z}}_i = \begin{bmatrix} \boldsymbol{Z}_i \\ \boldsymbol{\Delta} \end{bmatrix}. \tag{2.9}$$

Pinheiro and Bates [2000] show that given an estimate of $\widehat{\boldsymbol{b}}_i$, we can express the likelihood (without the penalizing prior on $\boldsymbol{\beta}$) as

$$L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 | \mathbf{y}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(\frac{-\sum_{i=1}^{M}\left\|\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\widehat{\boldsymbol{b}}_i\right\|^2}{2\sigma^2}\right) \prod_{i=1}^{M} \frac{\mathrm{abs}(|\boldsymbol{\Delta}|)}{\sqrt{\left|\tilde{\boldsymbol{Z}}_i^T\tilde{\boldsymbol{Z}}_i\right|}}. \tag{2.10}$$

Now, we can show that with the prior on $\boldsymbol{\beta}$, we get:

$$\begin{aligned} L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 | \mathbf{y}\right) &\propto \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{\sum_{i=1}^{M}\left\|\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\widehat{\boldsymbol{b}}_i\right\|^2}{2\sigma^2} - \frac{\|\boldsymbol{\beta}\|^2}{2\tau^2}\right) \prod_{i=1}^{M} \frac{\operatorname{abs}(|\boldsymbol{\Delta}|)}{\sqrt{\left|\tilde{\boldsymbol{Z}}_i^T\tilde{\boldsymbol{Z}}_i\right|}} \\ &\propto \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{\sum_{i=1}^{M}\left\|\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\widehat{\boldsymbol{b}}_i\right\|^2 + \lambda\|\boldsymbol{\beta}\|^2}{2\sigma^2}\right) \prod_{i=1}^{M} \frac{\operatorname{abs}(|\boldsymbol{\Delta}|)}{\sqrt{\left|\tilde{\boldsymbol{Z}}_i^T\tilde{\boldsymbol{Z}}_i\right|}} \\ &\propto \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{\sum_{i=1}^{M}\left\|\tilde{\mathbf{y}^*}_i - \tilde{\boldsymbol{X}^*}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}^*}_i\widehat{\boldsymbol{b}}_i\right\|^2}{2\sigma^2}\right) \prod_{i=1}^{M} \frac{\operatorname{abs}(|\boldsymbol{\Delta}|)}{\sqrt{\left|\tilde{\boldsymbol{Z}}_i^T\tilde{\boldsymbol{Z}}_i\right|}}, \end{aligned} \tag{2.11}$$

where we have further augmented the vectors and matrices as

$$\tilde{\mathbf{y}}_i^* = \begin{bmatrix} \mathbf{y}_i \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\boldsymbol{X}}_i^* = \begin{bmatrix} \boldsymbol{X}_i \\ \mathbf{0} \\ \sqrt{\frac{\lambda}{M}}\mathbf{I_p} \end{bmatrix}, \quad \tilde{\boldsymbol{Z}}_i^* = \begin{bmatrix} \mathbf{Z}_i \\ \boldsymbol{\Delta} \\ \mathbf{0} \end{bmatrix}. \tag{2.12}$$

Following Pinheiro and Bates [2000], we can work out that the profiled likelihood is

$$L(\boldsymbol{\theta}) = L\left(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \widehat{\sigma}^2(\boldsymbol{\theta})\right) = \frac{\exp(-N/2)}{\left[2\pi\widehat{\sigma}^2(\boldsymbol{\theta})\right]^{N/2}} \prod_{i=1}^{M} \frac{\operatorname{abs}(|\boldsymbol{\Delta}|)}{\sqrt{\left|\tilde{\boldsymbol{Z}}_i^T\tilde{\boldsymbol{Z}}_i\right|}}, \tag{2.13}$$

with $\widehat{\sigma}^2(\boldsymbol{\theta})$ defined by the residual sum-of-squares. Instead of the naive approach of first estimating $\hat{\mathbf{b}}$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ in order to get $\widehat{\sigma}^2(\boldsymbol{\theta})$, we can more efficiently calculate the latter via the QR decomposition:

$$\tilde{Z}_i^* = \mathbf{Q}_{(i)} \begin{bmatrix} \boldsymbol{R}_{11(i)} \\ \mathbf{0} \end{bmatrix}, \tag{2.14}$$

where $\mathbf{Q}_{(i)}$ is $(n_i + q + p) \times (n_i + q + p)$ and $\boldsymbol{R}_{11(i)}$ is $q \times q$. In our case, $q = 2$, therefore

$$\begin{aligned} \left\|\tilde{\mathbf{y}}_i^* - \tilde{\boldsymbol{X}}_i^*\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i^*\boldsymbol{b}_i\right\|^2 &= \left\|\boldsymbol{Q}_{(i)}^T\left(\tilde{\mathbf{y}^*}_i - \tilde{\boldsymbol{X}}_i^*\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i^*b_i\right)\right\|^2 \\ &= \left\|\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\boldsymbol{\beta} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i\right\|^2 + \left\|c_{0(i)} - \boldsymbol{R}_{00(i)}\boldsymbol{\beta}\right\|^2, \end{aligned} \tag{2.15}$$

where

$$\begin{bmatrix} \boldsymbol{R}_{10(i)} \\ \boldsymbol{R}_{00(i)} \end{bmatrix} = \boldsymbol{Q}_{(i)}^T\tilde{\boldsymbol{X}}_i^* \quad \text{and} \quad \begin{bmatrix} \boldsymbol{c}_{1(i)} \\ c_{0(i)} \end{bmatrix} = \boldsymbol{Q}_{(i)}^T\tilde{\mathbf{y}}_i^*. \tag{2.16}$$

By integrating out the $\mathbf{b}_i$, eq. (2.11) can then be shown to become

$$\begin{aligned} L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 | \mathbf{y}\right) &= \prod_{i=1}^{M} \frac{\exp\left[-\left\|\boldsymbol{c}_{0(i)} - \boldsymbol{R}_{00(i)}\boldsymbol{\beta}\right\|^2 / 2\sigma^2\right]}{\left(2\pi\sigma^2\right)^{n_i/2}} \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{\left|\boldsymbol{R}_{11(i)}\right|}\right) \\ &= \frac{\exp\left(-\sum_{i=1}^{M}\left\|c_{0(i)} - R_{00(i)}\boldsymbol{\beta}\right\|^2 / 2\sigma^2\right)}{\left(2\pi\sigma^2\right)^{-N/2}} \prod_{i=1}^{M} \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{\left|\boldsymbol{R}_{11(i)}\right|}\right). \end{aligned} \tag{2.17}$$

The term in the exponent is a residual sum-of-squares across patients i, which can be calculated using an additional QR decomposition as follows:

$$\begin{bmatrix} \boldsymbol{R}_{00(1)} & \boldsymbol{c}_{0(1)} \\ \vdots & \vdots \\ \boldsymbol{R}_{00(M)} & \boldsymbol{c}_{0(M)} \end{bmatrix} = \boldsymbol{Q}_0 \begin{bmatrix} \boldsymbol{R}_{00} & \boldsymbol{c}_0 \\ \mathbf{0} & \boldsymbol{c}_{-1} \end{bmatrix}, \tag{2.18}$$

which leads to:

$$L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 | \mathbf{y}\right) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left(\frac{\|c_{-1}\|^2 + \|c_0 - R_{00}\beta\|^2}{-2\sigma^2}\right) \prod_{i=1}^{M} \operatorname{abs}\left(\frac{|\Delta|}{\left|R_{11(i)}\right|}\right). \tag{2.19}$$

Using the maximum likelihood estimates for $\boldsymbol{\beta}$ and σ^2:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \boldsymbol{R}_{00}^{-1}\boldsymbol{c}_0 \quad \text{and} \quad \widehat{\sigma}^2(\boldsymbol{\theta}) = \frac{\|\boldsymbol{c}_{-1}\|^2}{N}, \tag{2.20}$$

we get the final expression for the profile likelihood:

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{y}) &= L\left(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \widehat{\sigma}^2(\boldsymbol{\theta}) | \mathbf{y}\right) \\ &= \left(\frac{N}{2\pi\left\|\boldsymbol{c}_{-1}\right\|^2}\right)^{N/2} \exp\left(-\frac{N}{2}\right) \prod_{i=1}^{M} \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{\left|\boldsymbol{R}_{11(i)}\right|}\right). \end{aligned} \tag{2.21}$$

We are now able to optimize eq. (2.21) with respect to $\boldsymbol{\theta} = D$ and then use the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ in eq. (2.20) to get the estimates for $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\widehat{\sigma}^2(\boldsymbol{\theta})$. Note that this approach only involves a single matrix inversion of $\boldsymbol{R}_{00}$; however, since this is an upper triangular matrix, we can simply solve for $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ by forward substitution in the equation $\boldsymbol{R}_{00}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \boldsymbol{c}_0$.

Finally, we note that the QR decomposition in eq. (2.18) is rather inefficient for the high-dimensional case, because of the inflation of the starred matrices with p penalty terms, which leads to having to calculate the QR decomposition of a matrix of size $(M * (n_i + p)) \times (p + 1)$.

We can avoid the computational burden by first noting that the matrix on the left-hand side in eq. (2.18) is taller than it is wide. Let us define:

$$\mathbf{R_c} = \begin{bmatrix} \boldsymbol{R}_{00(1)} & \boldsymbol{c}_{0(1)} \\ \vdots & \vdots \\ \boldsymbol{R}_{00(M)} & \boldsymbol{c}_{0(M)} \end{bmatrix}, \tag{2.22}$$

and

$$\mathbf{R}_{\mathbf{c}(i)} = \begin{bmatrix} \boldsymbol{R}_{00(i)} & \boldsymbol{c}_{0(i)} \end{bmatrix}. \tag{2.23}$$

Then it can be shown that the R-matrix from the QR decomposition of $\mathbf{R_c}$ can be obtained by Cholesky decomposition of the cross-product $\mathbf{A} = {\mathbf{R_c}}^T\mathbf{R_c}$. But the cross-product of the $(M*(n_i+p))\times(p+1)$ matrix $\mathbf{R_c}$ is just the sum of M cross-products ${\mathbf{R}_{\mathbf{c}(i)}}^T\mathbf{R}_{\mathbf{c}(i)}$. We can further optimize the calculation by noting that each ${\mathbf{R}_{\mathbf{c}(i)}}^T\mathbf{R}_{\mathbf{c}(i)}$ is of the form:

$$\mathbf{R}_{\mathbf{c}(i)} = \begin{bmatrix} \boldsymbol{R}'^T_{00(i)}\boldsymbol{R}'_{00(i)} + \mathbf{P} & \boldsymbol{R}'^T_{00(i)}\boldsymbol{c}'_{0(i)} \\ \boldsymbol{R}'^T_{00(i)}\boldsymbol{c}'_{0(i)} & \boldsymbol{c}'^T_{0(i)}\boldsymbol{c}'_{0(i)} \end{bmatrix}, \tag{2.24}$$

where $\mathbf{P} = \frac{\lambda}{M}\mathbf{I_p}$, and $\boldsymbol{R}'^T_{00(i)}$ and $\boldsymbol{c}'_{0(i)}$ result from transforming the unpenalized vector and matrix $\tilde{\mathbf{y}}_i$ and $\tilde{\boldsymbol{X}}_i$ defined in eq (2.9) by the upper left $(n_i+q)\times(n_i+q)$ matrix of $\mathbf{Q}^T_{(i)}$, similarly to eq. (2.16). As a result, we can avoid calculating the cross-products of $(n_i+p)\times(p+1)$ matrices in favour of $n_i\times(p+1)$ matrices, followed by adding $\mathbf{P}$ to the upper left $p\times p$ matrix.

2.2 *Mixture model for latent group structure detection*

We are now ready to define the mixture model that combines a latent group membership variable with the longitudinal regression model defined in Subsection 2.1. Conditional on $z_i = k$, i.e. knowing the cluster memberships, the joint likelihood of $(\mathbf{y}_i, \mathbf{X}_i)$ can be decomposed into the mixed effects regression model for $\mathbf{y}_i|\mathbf{X}_i$, and a multivariate model for $\mathbf{X}_i$ as follows:

$$p\left(\mathbf{y}_i, \mathbf{X}_i|\boldsymbol{\theta}_k, z_i = k\right) \equiv p\left(\mathbf{y}_i|\boldsymbol{\theta}^Y_k, \mathbf{X}_i, z_i = k\right) p\left(\mathbf{X}_i|\boldsymbol{\theta}^X_k, z_i = k\right), \tag{2.25}$$

where $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_k^X, \boldsymbol{\theta}_k^Y)^T$. Marginalizing out the latent variables leads to a mixture regression of the form:

$$p\left(\mathbf{y}, \mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\tau}\right) = \prod_{i=1}^{M} \sum_{k=1}^{K} p\left(\mathbf{y}_i|\boldsymbol{\theta}_k^Y, \mathbf{X}_i, z_i = k\right) p\left(\mathbf{X}_i|\boldsymbol{\theta}_k^X, z_i = k\right) \boldsymbol{\tau}_k \tag{2.26}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)^T$ and $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_K)^T$. The model formulation presented in eq. 2.26 has been studied by Ingrassia et al. [2012] under the non-longitudinal setting and in the context of ML estimation.

Gaussian graphical model for $p\left(\mathbf{X}_i|\boldsymbol{\theta}_k^X, z_i = k\right)$. Throughout this paper we assume that covariates can be modelled via p-dimensional multivariate Gaussian distributions such that $\boldsymbol{\theta}_k^X = (\boldsymbol{\mu}_k, \mathbf{vec}(\boldsymbol{\Sigma}_k))^T$, where $\boldsymbol{\mu}_k$ is the mean and $\boldsymbol{\Sigma}_k$ is the $p \times p$ covariance matrix. To mitigate computational costs during inference, we do not attempt to model the time-dependencies for time-varying covariates, but instead asssume that the overall mean and covariance are sufficiently representative of the underlying phenotype that we want to capture via the mixture components. To deal with the potentially large number of parameters in $\boldsymbol{\Sigma}_k$, we use regularization via the graphical lasso introduced in Friedman et al. [2008] (package glasso in **R**, Friedman et al. [2015]). The graphical lasso induces sparsity in the inverse covariance matrix, denoted by $\boldsymbol{\Omega}_k = \boldsymbol{\Sigma}_k^{-1}$ for group k, where we set the graphical lasso penalty to be $-\xi\|\boldsymbol{\Omega}_k\|_1$, in such a way that $\xi > 0$ controls the strength of regularization and $\|.\|_1$ is the L_1 norm. Then for known group labels the graphical lasso estimate is given by solving the following maximization problem

$$\underset{\boldsymbol{\Omega}_k \in M^+}{\arg\max} \left\{ \mathbf{log}\,|\boldsymbol{\Omega}_k| - \mathrm{tr}(\boldsymbol{\Omega}_k \hat{\boldsymbol{S}}_k) - \xi\|\boldsymbol{\Omega}_k\|_1 \right\}, \tag{2.27}$$

where M^+ is the space of positive definite matrices and $\hat{\boldsymbol{S}}_k$ is the ML covariance estimate of X_k. In practice, this will be weighted by the responsibilities $m_{ik} = p(z_i = k|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\theta}_k)$ of each patient, using $\boldsymbol{\mathcal{M}}_k$, the $N \times N$ diagonal matrix with entries $\boldsymbol{\mathcal{M}}_k(r, r) = m_{ik}$ if row x_r corresponds to covariates for patient i. The empirical estimate for the covariance matrix becomes $\hat{\boldsymbol{S}}_k = \boldsymbol{X}^T \boldsymbol{\mathcal{M}}_k \boldsymbol{X}$. If the covariates are not time-varying, then for the purpose of eq. (2.27) the matrix

$\boldsymbol{X}$ can be simplified to an $M \times p$ matrix without loss of generality.

Linear mixed model for $p(\mathbf{y}_i|\boldsymbol{\theta}_k^Y, \mathbf{X}_i, z_i = k)$. The regression term $p(\mathbf{y}_i|\boldsymbol{\theta}_k^Y, \mathbf{X}_i, z_i = k)$ corresponds to the mixed effects model in eq. (2.2). However, in the case where the latent variable z_i is unobserved, we need to additionally account for the responsibilities m_{ik} of each patient. In other words, the likelihood function for group k is of the form

$$L\left(\boldsymbol{\beta}_k, \mathbf{D}_k, \sigma_k^2|\mathbf{y}\right) = p\left(\boldsymbol{\beta}_k|\tau^2\right) \prod_{i=1}^{M} p\left(\mathbf{y}_i|\boldsymbol{\beta}_k, \mathbf{D}_k, \sigma_k^2\right)^{m_{ik}}, \tag{2.28}$$

which means that eq. (2.17) can be shown to become

$$\begin{aligned} L\left(\boldsymbol{\beta}_k, \boldsymbol{\theta}, \sigma_k^2|\mathbf{y}\right) &= \frac{\exp\left(-\sum_{i=1}^{M} m_{ik} \left\|c_{0(i)} - R_{00(i)}\boldsymbol{\beta}_k\right\|^2 / 2\sigma_k^2\right)}{\left(2\pi\sigma_k^2\right)^{-N/2}} \prod_{i=1}^{M} \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{\left|\boldsymbol{R}_{11(i)}\right|}\right)^{m_{ik}} \\ &= \frac{\exp\left(-\sum_{i=1}^{M} \left\|\sqrt{m_{ik}}c_{0(i)} - \sqrt{m_{ik}}R_{00(i)}\boldsymbol{\beta}_k\right\|^2 / 2\sigma_k^2\right)}{\left(2\pi\sigma_k^2\right)^{-N/2}} \prod_{i=1}^{M} \operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{\left|\boldsymbol{R}_{11(i)}\right|}\right)^{m_{ik}}, \end{aligned} \tag{2.29}$$

and consequently the matrix that needs to undergo QR decomposition becomes:

$$\mathbf{R}_{\mathbf{c}}^k = \begin{bmatrix} \sqrt{m_{1k}}\boldsymbol{R}_{00(1)} & \sqrt{m_{1k}}\boldsymbol{c}_{0(1)} \\ \vdots & \vdots \\ \sqrt{m_{Mk}}\boldsymbol{R}_{00(M)} & \sqrt{m_{Mk}}\boldsymbol{c}_{0(M)} \end{bmatrix}. \tag{2.30}$$

Note that $L\left(\widehat{\boldsymbol{\beta}}_k(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k, \widehat{\sigma}_k^2(\boldsymbol{\theta}_k)|\mathbf{y}\right)$ can be optimized individually for each k, leading to potential efficiency gains using parallel computation.

2.3 *Expectation-Maximization Algorithm*

Inference of $\boldsymbol{\beta}_k$, $\mathbf{D}_k$ and σ_k for each group $k \in K$ is complicated by the fact that the group indicator variables z_i are unobserved. We employ an expectation-maximization (EM) algorithm, similar to Perrakis et al. [2019], to perform this inference. We describe the initialisation, expectation (E-step) and maximisation (M-step) below; the algorithm is also summarized in Algorithm 1.

Initialisation. We initialise $\boldsymbol{\beta}_k$ values across all groups, by running a simple penalised linear model using the `glmnet` package in R (Hastie [2020]). We initialise $\mathbf{D}_k$ and σ_k by choosing randomly generated positive definite matrices and scalar values, respectively, across all groups.

E-step. We estimate the responsibilities following Perrakis et al. [2019], using the the following formula:

$$\begin{aligned} m_{ik} &\equiv p\left(z_i = k | \mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\theta}_k\right) \\ &= \frac{p\left(\mathbf{y}_i | \boldsymbol{\theta}_k^Y, \mathbf{X}_i, z_i = k\right) p\left(\mathbf{X}_i | \boldsymbol{\theta}_k^X, z_i = k\right) \tau_k}{\sum_{k'} p\left(\mathbf{y}_i | \boldsymbol{\theta}_{k'}^Y, \mathbf{X}_i, z_i = k'\right) p\left(\mathbf{X}_i | \boldsymbol{\theta}_{k'}^X, z_i = k'\right) \tau_{k'}}, \end{aligned} \tag{2.31}$$

with the modification that in our work, i refers to patients rather than data points, and $p\left(\mathbf{y}_i | \boldsymbol{\theta}_k^Y, \mathbf{X}_i, z_i = k\right)$ is defined as in eq. (2.2).

In the next step, which is now the M-step, we need to optimise the objective function stated in eq. (2.28) (which comes in the form a profile log-likelihood function) and update $\boldsymbol{\theta}_k^Y$ and $\boldsymbol{\theta}_k^X$ using the responsibilities m_{ik} from the E-step. For $\boldsymbol{\theta}_k^X$, let $\mathbf{W}_k = \boldsymbol{\mathcal{M}}_k \mathbf{X}$ be the weighted covariate matrix, then $\mu_{kp} = \frac{\sum_r \mathbf{W}_k(r,p)}{N_k}$ is the update for $\boldsymbol{\mu}_k$, where $N_k = \sum_i m_{ik}$. $\boldsymbol{\Sigma}_k$ can be updated using the graphical lasso update in eq. (2.27). For $\boldsymbol{\theta}_k^Y$, we update $\boldsymbol{\beta}_k$, $\mathbf{D}_k$ and σ_k using the profile likelihood in eq. (2.21).

Algorithm 1 shows the pseudo-code of the EM procedure deployed in the LJCR algorithm. Here *objective_function* refers to the eq. (2.28) and $\boldsymbol{\mathcal{M}}_k$ is the $N \times N$ matrix containing the responsibilities on the diagonal and zero elsewhere. The maximum number of iterations N_{it} is set to 100, although the loop may terminate early if we reach convergence, or if the size of one of the groups (N_k) gets very small. This latter condition is necessary to avoid splitting individuals into very small groups where estimation of $\boldsymbol{\beta}_k$, $\mathbf{D}_k$ and σ_k would not be reliable.

Algorithm 1: Pseudo-code of the utilized EM algorithm

Result: Optimal values for $\boldsymbol{\beta}_k$, $\mathbf{D}_k$, σ_k and responsibilities m_{ik} for each $i \in \{1, \ldots, M\}$ and $k \in \{1, \ldots, K\}$

Initialize with some values for $\boldsymbol{\beta}_k$ $\mathbf{D}_k$ *and* σ_k;

$\rho = \dfrac{\sqrt{2M \log(p)}}{2}$;

for *iteration* $\leftarrow 1$ to $Niteration$ **do**

E-step

$p_{y_{ik}} = p(y_i | \mathbf{X}_i \boldsymbol{\beta}_k, \mathbf{Z}_i \mathbf{D}_k \mathbf{Z}_i^T + \sigma_k \mathbf{I}_{n_i})$;

$p_{x_{ik}} = p(\mathbf{X}_i | \mu'_k, \boldsymbol{\Omega}_k)$;

$m_{ik} = \dfrac{p_{y_{ik}} p_{x_{ik}}}{\sum_k p_{y_{ik}} p_{x_{ik}} m_k}$;

$N_k = \sum_i I(m_{ik} > \frac{1}{K})$;

M-step

$\mu'_k = \dfrac{\sum_i m_{ik} \times E_t[\mathbf{X}_i(t)]}{\sum_i m_{ik}}$;

$\hat{\boldsymbol{S}}_k = \boldsymbol{X}^T \boldsymbol{\mathcal{M}}_k \boldsymbol{X}$;

$\boldsymbol{\Omega}_k$ =**glasso**$(\hat{\boldsymbol{S}}_k, \rho,$ **penalize.diagonal** $= FALSE)\$w)$;

Optimise the objective function defined in the eq. (2.28) and set

$L_{\text{iteration}} = L\left(\boldsymbol{\beta}_k, \mathbf{D}, \sigma_k^2 | \mathbf{y}\right)$;

Update $\boldsymbol{\beta}_k$, $\mathbf{D}_k$ *and* σ_k;

if $(N_k \leqslant \frac{M}{10 \times K})$ ***or*** $(L_{iteration} - L_{iteration\text{-}1} < 1)$ **then**

Stop the for loop and output $\boldsymbol{\beta}_k$ $\mathbf{D}_k$, σ_k and m_{ik} for each $i \in \{1, \ldots, M\}$ and $k \in \{1, \ldots, K\}$

else

Continue the for loop

end

end

3. RESULTS

We apply the LJCR model to two different longitudinal datasets; a simulated dataset, which allows us to evaluate the ability of the model to recover known subgroups and regression coefficients; and the PRO-ACT dataset (Atassi et al. [2014]), the largest database of clinical trials of patients with ALS (Amyotrophic Lateral Sclerosis). The latter allows us to both evaluate the predictive performance of the model on a real-world dataset, and to gain novel insights into factors that may be associated with different clinical phenotypes and progression trajectories in this disease.

3.1 *Simulation Study*

We perform a simulation study to test the performance of the LJCR method under a range of scenarios. More precisely, we study how well the model can predict the three parameters $\boldsymbol{\beta}_k$ (coefficients for the fixed effects), σ_k (variance), and $\mathbf{D}_k$ (covariance matrix of the random effects) under scenarios with different number of individuals M and number of covariates (features) p. We then compare the performance of the LJCR algorithm with three different methods as follows:

- kml: k-means method for longitudinal data [Genolini and Falissard, 2010] (package kml in R).
- Clustering+LMM: In this method we first use the standard k-means algorithm to cluster the data (based on the response values y) and then apply a Linear Mixed Effects Model (LMM) (Schafer [1998]) (package LMM in R by Zhao [2020]) on each cluster.
- Baseline method: Here we cluster the data as above and then apply a simple linear model with a random slope only, without including the fixed effects for the design matrix $\mathbf{X}$. This provides a convenient baseline that is not affected by the dimensionality of $\mathbf{X}$.

Here we generate 10 sets of independent datasets where in each of those sets we generate 4 different datasets (in total 40) with the dimension of $(M \times n_i) \times p$ where $M = 400$ (number

of individuals), $n_i = 10$ (number of observations/data points for each individual i), and $p \in \{100, 500, 1000, 7000\}$ (number of features/covariates). We use these set of generated data to show the performance of the LJCR model when varying the number of features p (see figure 1). We then generate another 10 sets of independent datasets where in each of those sets we generate 4 different datasets (in total 40) with the dimension of $(M \times n_i) \times p$ where $M \in \{500, 1000, 4000\}$, $n_i = 10$, and $p = 7000$. Again, we use these set of generated data to show the performance of the LJCR model when varying the sample size M (see figure 1). The number of groups across all these simulation scenarios is assumed to be $K = 3$.

In both of the aforementioned sets of simulated data, we use normal distribution (using rnorm in R) with mean 0 and variance 1 to generate the data. We then sample the initial values for $\boldsymbol{\beta}_k$, $\mathbf{D}_k$ and σ_k from a normal distribution with mean 0 and variance 1, randomly generated positive definite matrices (using genPositiveDefMat() in clusterGeneration R package by Weiliang Qiu [2015]), and uniformly generated scalar values, respectively, across all groups.

Figure 1 shows a comparison between the LJCR, the k-means for longitudinal data (Genolini and Falissard [2010]), clustering and then running LMM on each cluster (Clustering+LMM), and baseline methods. Performance of the all models have been tested on the generated datasets we explained before, varying the number of covariates and samples. Notice that in the kml method, we compute the mean trajectories of each group k, using the function calculTrajMeanC in the kml package, and consider them as the values of $\boldsymbol{\beta}_k$. We then directly use the obtained $\boldsymbol{\beta}_k$ values to calculate responses y_i. We see that in the both scenarios where we increase the number of features p or the sample size M, the LJCR outperforms all the models in terms of prediction of both parameters $\boldsymbol{\beta}_k$ and response values y_i. We could not apply the Clustering+LMM method beyond $p = 500$, as multivariate nature and high-dimensional setting makes it infeasible. There is no covariate matrix $\mathbf{X}$ used in the baseline method, hence there would be no $\boldsymbol{\beta}$ obtained in this method.

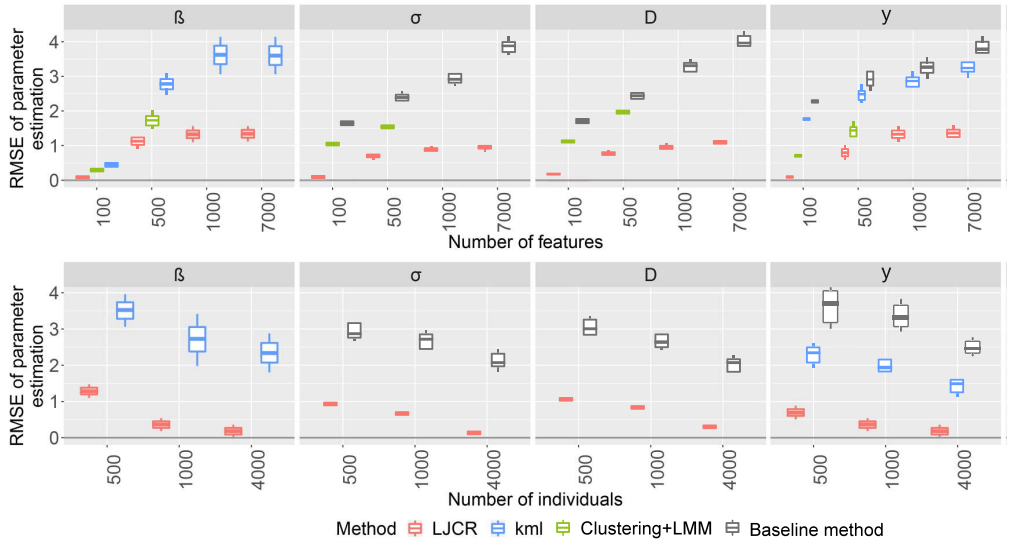
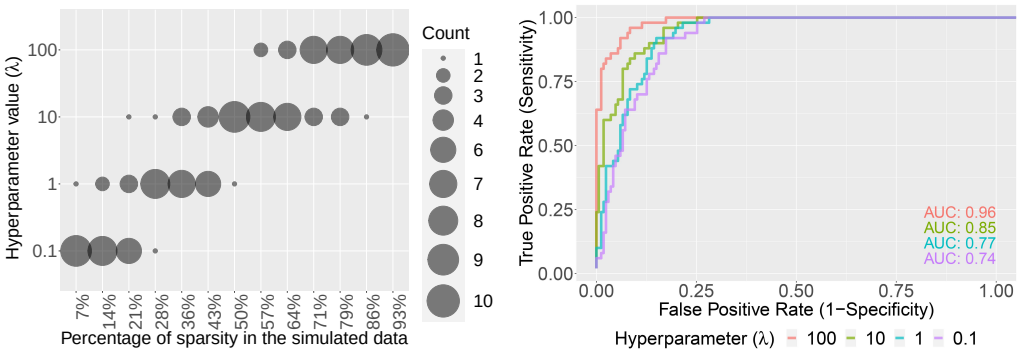


Fig. 1: Comparison of the performance of the LJCR, k-means for longitudinal data, clustering and then running LMM on each cluster (Clustering+LMM), and baseline methods on simulated data when varying the number of features (top row), and sample size (bottom row). The boxplots represent the estimates of RMSE obtained from 10 independent simulated datasets. Missing boxplots for Clustering+LMM indicate situations where the dimensionality was too high to apply this method. Also note that the baseline method will not return an estimate for the $\boldsymbol{\beta}_k$, and kml will not return an estimate for σ_k or $\boldsymbol{D}_k$.



(a) (b)

Fig. 2: (a) The number of times the hyperparameter (λ) is selected by a 10-fold cross validation. The x-axis represents the proportion of sparse betas (everything below the 0.01 threshold) in the initial vector $\boldsymbol{\beta}$ that is used to simulate the data. (b) ROC curves of the estimated $\hat{\boldsymbol{\beta}}$ and associated AUC values for different hyperparameter values.

Next we perform a sensitivity analysis of the LJCR model to look at the effect of different values of hyperparameters λ in eq. (2.6). We define a set of hyperparameters $\lambda \in \{0.1, 1, 10, 100\}$ and allow the LJCR model to pick up the best value via 10-fold cross-validation. Figure 2a shows which hyperparameter values are chosen for different degrees of sparsity of the initial vector $\boldsymbol{\beta}$. Here we have $p = 7000$, $M = 400$ and $n_i = 10$ for all i. The test is performed for 10 simulated datasets. The size of each circle represents the number of times each value of λ is selected in a simulation run. As expected, when having highly sparse $\boldsymbol{\beta}$, then a large hyperparameter is chosen by the model. Figure 2b shows the relative ROC curve for different values of λ under the highly sparse scenario where 6500 out of 7000 values are near zero (93% sparsity). While high values of the hyperparameter naturally result in larger areas under the ROC curve (AUC) values, we note that even for misspecified λ values, the drop in AUC is modest.

3.2 *ALS PRO-ACT Data*

PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials) is a publicly available database containing industry and academic clinical trials of patients with ALS disease (Amyotrophic Lateral Sclerosis) (Atassi et al. [2014]). It is a longitudinal dataset including records of each individual across repeated visits to clinic. PRO-ACT is the largest ALS clinical trials database ever created, with more than 8500 patient records, including demographic and laboratory data, medical histories and functional scores.

In our study, we have used a subset of PRO-ACT collected longitudinally at different observation time-points. The response variable (y_i) is the ALSFRS (ALS Function Rating Scale) score. This score captures the overall state of the disease and can be considered as a progression score for people living with ALS. The ALSFRS scale is a list of 10 different assessments of motor function (such as the ability to move an object, the ability to eat with cutlery, the ability to handwrite, etc.), with each measure ranging from 0 to 4, with 4 being the highest (normal function) and 0

being no function. The score for the individual questions are then summed together to generate the total ALSFRS score, which ranges between $0 - 40$.

First of all, we have discarded all patients with single observation (one visit to the clinic). Then to deal with the missing values we interpolate using inter-subject sectional linear interpolation; i.e. we look at each individual i and replace every missing value with the average value between its previous and next observations/datapoints. We are then left with a cohort of $M = 4821$ patients and $p = 55$ observed features (covariates).

Supplementary table S1 shows all the $p = 55$ features used in the PRO-ACT dataset (Atassi et al. [2014]).

We have applied the LJCR algorithm to find the underlying latent subtypes.

Figure 3 shows that using the elbow technique (Joshi and Nalwade [2013]), $K = 9$ latent subtypes (groups or clusters) represents the optimal number of groups when applying the LJCR to the PRO-ACT dataset. The idea behind the elbow technique is to choose a number of clusters so that adding another cluster does not result in better model to fit to the data. More precisely, we look at the relative change of the values in each pair of consecutive clusters (gradient slope) and then compare the differences.

Figure 4a shows that there are 4 groups labels which contain most of the population size ($k = \{1, 5, 7, 9\}$). Figure 4b demonstrates that in the pre-mentioned group labels men are susceptible to be diagnosed with ALS disease earlier in age than women (about $2 - 5$ years).

Figure 5 shows the estimated effect size ($\boldsymbol{\beta}_k$) for each feature in group $k \in \{1, 5, 7, 9\}$. As an interesting result, we observe that the effect of Mean Corpuscular Hemoglobin Concentration (MCHC) has a positive effect size in group label 7, unlike in the other groups. Similarly, Absolute Basophil Count has a negative effect size in group label 7 which is in contrary to all other group labels.

Many of these factors have been previously associated with the rate of ALSFRS decline

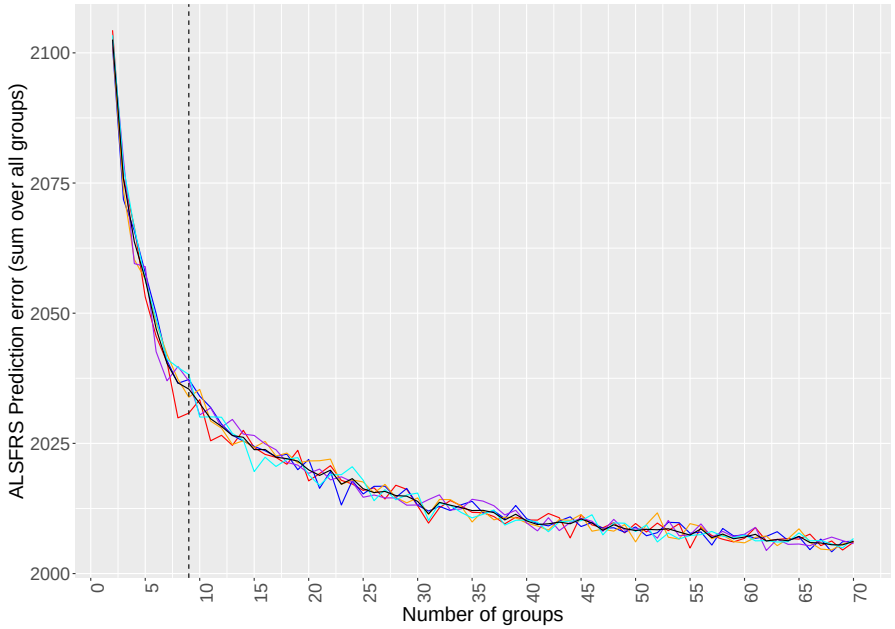
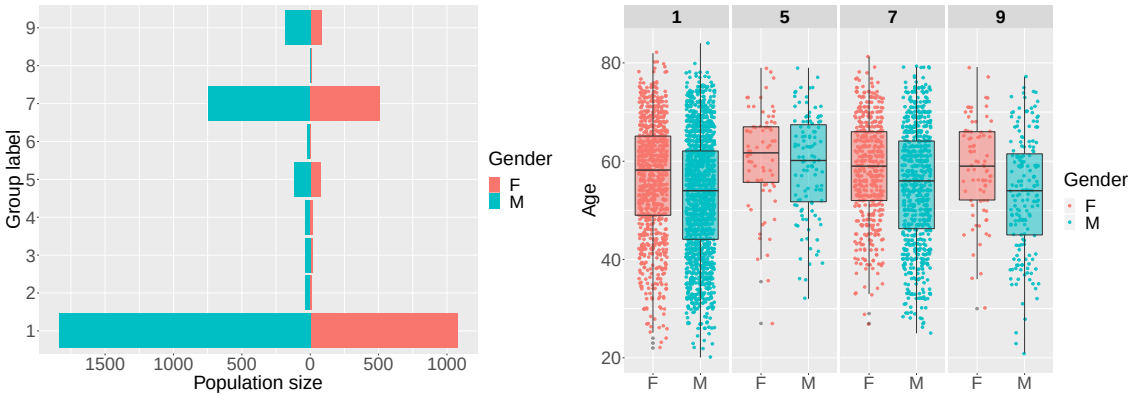


Fig. 3: Performance of the LJCR model on the PRO-ACT ALS data when varying the number of latent subtypes (groups). The LJCR model is executed five times (shown in different colors) each time with different initial values for the parameters $\boldsymbol{\beta}_k$, σ_k, and $\mathbf{D}_k$. The number of latent groups varies with $K \in \{2, \ldots, 70\}$). For each K (x-axis), the y-axis represents the mean prediction error for the ALSFRS scores across the groups. We calculate the error between the predicted response values (ALSFRS) (after running the LJCR and assigning the group membership) and the true ALSFRS values on the whole ALS PRO-ACT samples (training set). The black curve represents the mean of the colored curves. The elbow method (Joshi and Nalwade [2013]) is then used to identify $K = 9$ as the optimal number of subtypes for this dataset.

including weight, FVC and age (Mandrioli et al. [2015]). Plasma creatinine has been previously associated with outcome in ALS and may act as a marker of muscle reserve (Mitsumoto et al. [2020]). The observation of different relative effect sizes between groups for cholesterol and weight suggests that cholesterol is not acting only as a proxy for weight. Indeed there is evidence that serum cholesterol may be an important marker of the extent to which a dysfunctioning motor system is energy deficient (Dupuis et al. [2008]). The observation that serum cholesterol may have a different relative effect in different patient groups is important because clearly hyperlipidaemia can be harmful in certain contexts and lead to, for example, cardiovascular disease; therefore it would be important to recommend dietary changes to boost cholesterol only when clinically



(a) (b)

Fig. 4: Application of the LJCR model to the PRO-ACT ALS study. (a) Pyramid plot showing the mixture component sizes. (b) Box-plots showing the age at baseline distribution for the groups with the largest population size $k = \{1, 5, 7, 9\}$, with the blue color standing for male and red for female.

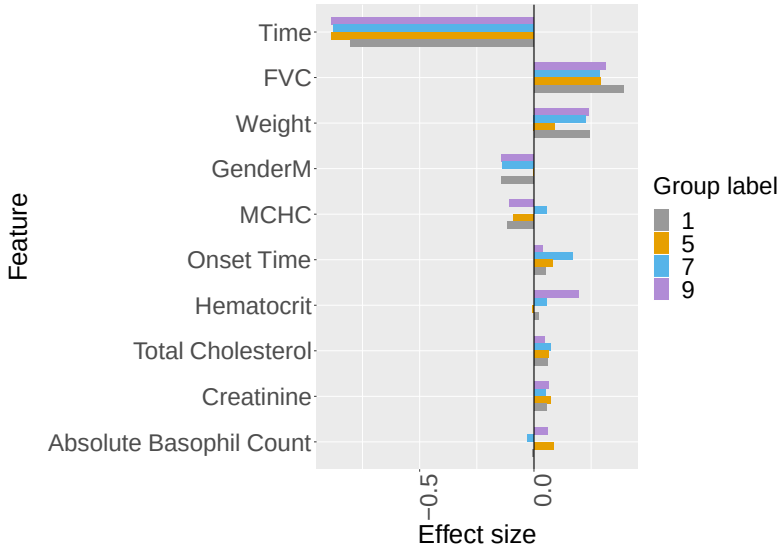


Fig. 5: Application of the LJCR model to the PRO-ACT ALS dataset. The bar plots show the estimated effect size (estimated $\boldsymbol{\beta}_k$ parameters) for the group labels with largest population size ($k = \{1, 5, 7, 9\}$). Note that MCHC stands for Mean Corpuscular Hemoglobin Concentration. Time refers to time on the study (in days, but scaled here to make the effect sizes comparable).

appropriate. The observation that MCHC and hematocrit are associated with outcome is a novel finding. This could be interpreted in the context of resistance to respiratory failure however this does not explain the different direction of effect in groups 9 and 1. Equally the finding that absolute basophil count is a predictor of outcome is novel although peripheral immune cells have been linked to CNS inflammation and disease progression (Butovsky et al. [2012]). Discovering the differences in CNS inflammation between groups 9 and 5, where basophils have a positive correlation, and groups 1 and 7 where there is no correlation or a negative correlation, could guide personalised immunotherapy for ALS.

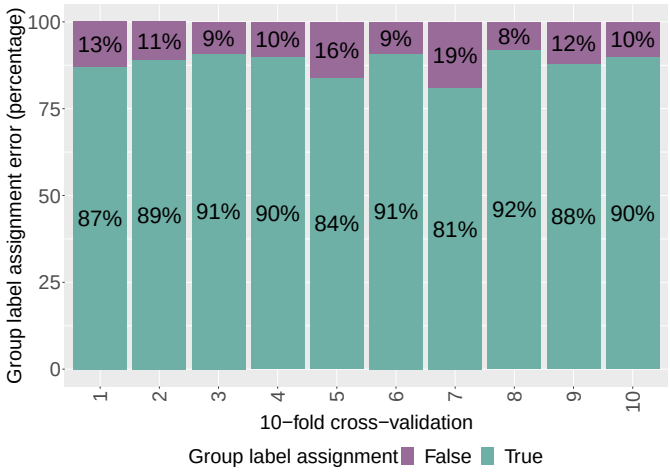
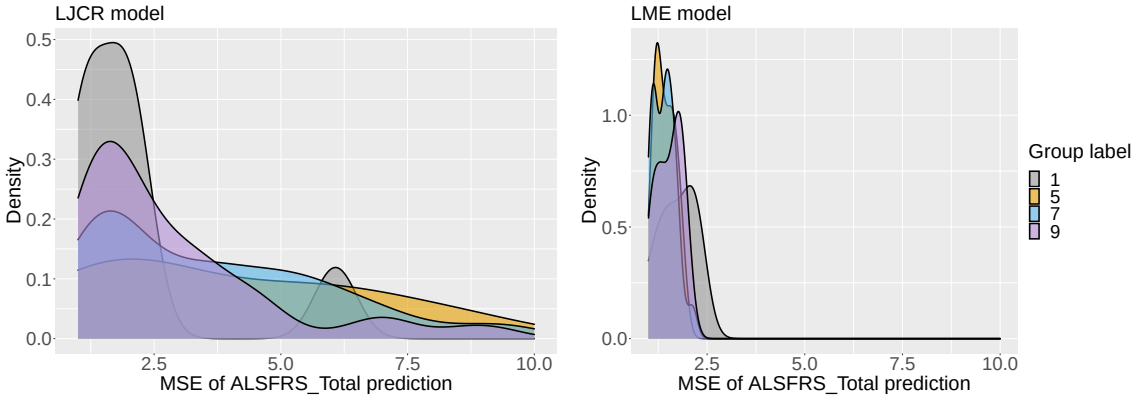


Fig. 6: Performance of the LJCR model in detecting how well patients could be classified based on their covariates alone (no ALSFRS), compared to the group label that they were assigned when including both covariates and responses (ALSFRS) in the training of the mixture model. We train the model 10 times with different randomly selected test sets containing 50 individuals. The bar plots represent the percentage of group assignment error for each fold.

One potential issue with prediction of progression for new patients is that we only have access to the covariates $\mathbf{X}_i$ for assigning the new subject to an existing mixture component. In order to test whether this is sufficient, we perform an experiment where we first train the model on the whole dataset (4821 individuals) to assign each individual a group label. Then we choose a random set of 50 people as our test set and re-run the model on the remaining training set (4771

people). Finally, for each individual in the test, we find the estimated assigned group label by solving this problem: Find the group label with the highest probability of test individual i falling into that group based on the mean and variance of distribution of the observed feature training set ($\mathbf{x}_{ik}$) for each group label k. Here we applied the graphical lasso [Friedman et al., 2008] to calculate mean and covariance matrix components for each subgroups ($\mathbf{X}_k$). We have repeat this procedure 10 times. Figure 6 shows that the LJCR models performs reasonably well in detecting the same group labels that would have been assigned when training on both the response and the covariates.

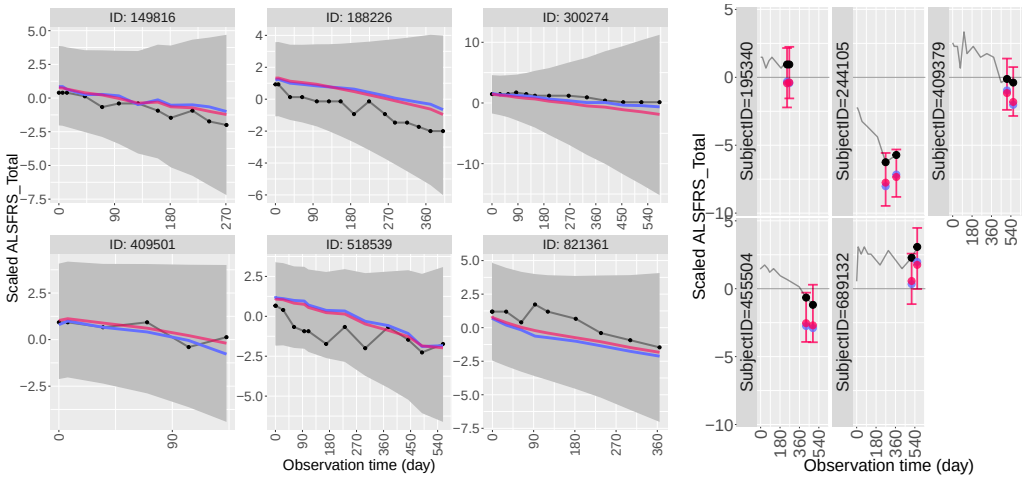


(a) (b)

Fig. 7: Distribution of ALSFRS prediction for each individual i and observation time t in different group labels ({1,5,7,9}).

Figure 7a shows the density plots for the ALSFRS prediction error of the different mixture components. Note that group 1 is the largest, hence the MSE prediction error for the ALSFRS total score should be lower, since more data is available to estimate the parameters. Figures 7b shows a density plot where we have used the group labels inferred from the LJCR model, but then refit for each group using the LME model (linear mixed effects model). We see that this re-fitting after inference of the group memberships by the LJCR algorithm improves the prediction error.

Figure 8 shows the performance of the LJCR model in ALSFRS prediction under two scenar-



(a) (b)

Fig. 8: Performance of the LJCR model on ALS PRO-ACT data in predicting value of ALSFRS Total, when number of optimal groups is fixed and equal to $K_{\text{optimal}} = 9$. In each case, we compare performance of the LJCR model without taking the random effect parameters into account (in red), with the linear mixed effect model prediction (in blue). (a) Prediction for unseen individuals over a test set of individuals of size $n = 50$. The black line represents the true values of ALSFRS Total and the gray ribbon area shows the 95% confidence interval of the LJCR model when taking the random effect parameters into account. (b) Prediction for new datapoints of seen individuals over a test set of observation time-points of size 2. The test is executed on 5 randomly selected individuals. The black line represents the true values of ALSFRS Total and error bars show the 95% confidence interval of the LJCR model when taking the random effect parameters into account.

ios; in the first scenario we test prediction for unseen individuals (Figure 8a) where we predict ALSFRS value of a test set of 50 new individuals. In the second scenario we test prediction for unseen time-points on individuals where the previous time points were included in the training set (Figure 8b). Here we predict ALSFRS scores at two time-points for 5 randomly selected individuals. Both figures 8a and 8b show that the prediction performance using the random and fixed effects are almost identical, indicating that the random effects are negligible for the prediction task.

4. Discussion

The aim of this article was to introduce longitudinal joint cluster regression (LJCR) to detect latent group (cluster) structures within longitudinal data and predict personalised disease outcomes informed by these latent structures. Latent group structure plays a key role in modern data-intensive applications as it can strongly confound estimates and lead to practical difficulties if ignored.

Latent group structures are modelled using a class of Gaussian mixture models that couple together the multivariate distribution of the covariates and response. This is different from classical mixture regression approaches, which focus on the distribution of the dependent variable only. Our approach could be further extended to the non-parametric realm using e.g. a Dirichlet process formulation [Hannah et al., 2011, Liverani et al., 2015]. This would also remove the need for determining the optimal number of clusters. To avoid excessive computational costs, we have not pursued this approach here.

We model the longitudinal dynamics of each individual using a random effect intercept and slope model. The inference is done via a profile likelihood approach that can handle high-dimensional covariates by incorporating sparsity assumptions via ridge penalization. While l1 penalisation is possible in the mixed model paradigm [Schelldorfer et al., 2011], this comes with computational disadvantages, and the benefit of additional sparsity obtained by setting some parameters to zero is not clear; in previous work [Dondelinger et al., 2020], l2 penalisation led to improved predictions in some settings.

We have compared the performance of the LJCR model with an alternative method based on k-means [Genolini and Falissard, 2010] under a scenario where we vary the sample size and the number of covariates. It was shown that the LJCR outperforms this method, both in prediction error for the response variable (benefitting from modeling longitudinal dynamics via the random effect parameters), and prediction error for the fixed effect parameters in the high-dimensional

case (benefitting from incorporating ridge penalization).

An alternative method is the one described in Bruckers et al. [2016], which uses a latent growth model for the longitudinal data. It is worth mentioning that Bruckers et al. [2016], like most conventional mixture model approaches, only relies on conditional distribution of responses $Y|X$, disregarding any signal arising from the distribution of feature matrix X itself. This is one of the key differences between the LJCR method and the other standard models as we also incorporate estimation of the distribution of X via a graphical lasso approach.

We applied LJCR to a cohort of patients with ALS disease to find the latent subtypes (groups) within the study. Our approach detected 9 group labels in total, with 4 groups hosting the largest population sizes. Note that we are not claiming this as a ground truth for the homogeneous groups within the dataset, but rather an estimate based on our linear mixed model approach for the dynamics within each mixture component. An interesting extension for our work would be to consider non-linear dynamics for the longitudinal model.

We evaluated the prediction performance on our real-world dataset for each of the larger groups, and found that post-inference refitting of a standard linear mixed model improves prediction error. As we do not have a gold standard for group membership, we investigate the group label assignments derived by the LJCR algorithm informally by looking at the group characteristics and interpreting the clinical and biochemical variables identified as important via the group-specific fixed effects. Further investigations should focus on confirmatory studies to establish whether these variables have a causal effect on disease progression in subsets of patients.

References

Nazem Atassi, James Berry, Amy Shui, Neta Zach, Alexander Sherman, Ervin Sinani, Jason Walker, Igor Katsovskiy, David Schoenfeld, Merit Cudkowicz, et al. The pro-act database: design, initial analyses, and predictive features. *Neurology*, 83(19):1719–1725, 2014.

Liesbeth Bruckers, Geert Molenberghs, Pim Drinkenburg, and Helena Geys. A clustering algorithm for multivariate longitudinal data. *Journal of Biopharmaceutical Statistics*, 26(4): 725–741, 2016.

Oleg Butovsky, Shafiuddin Siddiqui, Galina Gabriely, Amanda J Lanser, Ben Dake, Gopal Murugaiyan, Camille E Doykan, Pauline M Wu, Reddy R Gali, Lakshmanan K Iyer, et al. Modulating inflammatory monocytes with a unique microrna gene signature ameliorates murine als. *The Journal of clinical investigation*, 122(9):3063–3087, 2012.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *arXiv:1111.0324*, November 2011.

David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

Frank Dondelinger, Sach Mukherjee, and Alzheimer's Disease Neuroimaging Initiative. The joint lasso: high-dimensional regression for group structured data. *Biostatistics*, 21(2):219–235, 2020.

L Dupuis, P Corcia, A Fergani, J-L Gonzalez De Aguilar, D Bonnefont-Rousselot, R Bittar, D Seilhean, J-J Hauw, L Lacomblez, J-P Loeffler, et al. Dyslipidemia is a protective factor in amyotrophic lateral sclerosis. *Neurology*, 70(13):1004–1009, 2008.

Brian S Everitt, S Landau, and M Leese. Cluster analysis arnold. *A member of the Hodder Headline Group, London*, pages 429–438, 2001.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Jerome Friedman, Trevor Hastie, Rob Tibshirani, and Maintainer Rob Tibshirani. Package 'glasso', 2015.

Christophe Genolini and Bruno Falissard. Kml: k-means for longitudinal data. *Computational Statistics*, 25(2):317–328, 2010.

Lauren A Hannah, David M Blei, and Warren B Powell. Dirichlet process mixtures of generalized linear models. *J. Mach. Learn. Res.*, 12(6), 2011.

Trevor Hastie. glmnet v4. 0-2. 2020.

Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

Salvatore Ingrassia, Simona C Minotti, and Giorgio Vittadini. Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of classification*, 29(3):363–401, 2012.

Kalpana D Joshi and PS Nalwade. Modified k-means for better initial cluster centres. *International Journal of Computer Science and Mobile Computing, IJCSMC*, 2(7):219–223, 2013.

Abbas Khalili and Jiahua Chen. Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479):1025–1038, 2007.

N M Laird and J H Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4): 963–974, December 1982.

Silvia Liverani, David I Hastie, Lamiae Azizi, Michail Papathomas, and Sylvia Richardson. PReMiuM: An R package for profile regression mixture models using dirichlet processes. *J. Stat. Softw.*, 64(7):1–30, March 2015.

Jessica Mandrioli, Sara Biguzzi, Carlo Guidi, Elisabetta Sette, Emilio Terlizzi, Alessandro Ravasio, Mario Casmiro, Fabrizio Salvi, Rocco Liguori, Romana Rizzi, et al. Heterogeneity in alsfrs-r decline and survival: a population-based study in italy. *Neurological Sciences*, 36(12): 2243–2252, 2015.

G McLachlan. Peel., d.(2000). finite mixture models.

Hiroshi Mitsumoto, Diana C Garofalo, Regina M Santella, Eric J Sorenson, Björn Oskarsson, J americo M Fernandes Jr, Howard Andrews, Jonathan Hupf, Madison Gilmore, Daragh Heitzman, et al. Plasma creatinine and oxidative stress biomarkers in amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 21(3-4):263–272, 2020.

Konstantinos Perrakis, Frank Dondelinger, and Sach Mukherjee. Latent group structure and regularized regression. *arXiv preprint arXiv:1908.07869*, 2019.

José Pinheiro and Douglas Bates. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2006.

José C Pinheiro and Douglas M Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, NY, 2000.

Siddheswar Ray and Rose H Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143. Calcutta, India, 1999.

Joseph L Schafer. Some improved procedures for linear mixed models. *Submitted to Journal of*, 1998.

Jürg Schelldorfer, Peter Bühlmann, and Sara Van de Geer. Estimation for High-Dimensional linear Mixed-Effects models using l_1-penalization. *Scand. Stat. Theory Appl.*, 38(2):197–214, 2011.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978.

Nicolas Städler, Peter Bühlmann, and Sara Van De Geer. l1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

Harini Suresh, Jen J Gong, and John V Guttag. Learning tasks for multitask learning: Heterogenous patient populations in the ICU. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 802–810, New York, NY, USA, July 2018. Association for Computing Machinery.

Bart Swinnen and Wim Robberecht. The phenotypic variability of amyotrophic lateral sclerosis. *Nature Reviews Neurology*, 10(11):661, 2014.

Harry Joe Weiliang Qiu. Package 'clustergeneration', 2015.

Jianpeng Xu, Jiayu Zhou, and Pang-Ning Tan. FORMULA: FactORized MUlti-task LeArning for task discovery in personalized medical models. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, Proceedings, pages 496–504. Society for Industrial and Applied Mathematics, June 2015.

Jing Zhao. Package 'linear mixed models (lmm) package v1.3', 2020.