LMLFM: Longitudinal Multi-Level Factorization Machine

Junjie Liang,[1,2] **Dongkuan Xu,**[2] **Yiwei Sun,**[1,3] **Vasant Honavar**[1,2,3,4]
[1] Artificial Intelligence Research Laboratory, Pennsylvania State University
[2]College of Information Sciences and Technology, Pennsylvania State University
[3] Department of Computer Science and Engineering, Pennsylvania State University
[4] Institute of Computational and Data Sciences, Pennsylvania State University
{jul672,dux19,vhonavar}@ist.psu.edu, yus162@psu.edu

Abstract

We consider the problem of learning predictive models from longitudinal data, consisting of irregularly repeated, sparse observations from a set of individuals over time. Such data often exhibit *longitudinal correlation* (LC) (correlations among observations for each individual over time), *cluster correlation* (CC) (correlations among individuals that have similar characteristics), or both. These correlations are often accounted for using *mixed effects models* that include *fixed effects* and *random effects*, where the fixed effects capture the regression parameters that are shared by all individuals, whereas random effects capture those parameters that vary across individuals. However, the current state-of-the-art methods are unable to select the most predictive fixed effects and random effects from a large number of variables, while accounting for complex correlation structure in the data and non-linear interactions among the variables. We propose Longitudinal Multi-Level Factorization Machine (LMLFM), to the best of our knowledge, the first model to address these challenges in learning predictive models from longitudinal data. We establish the convergence properties, and analyze the computational complexity, of LMLFM. We present results of experiments with both simulated and real-world longitudinal data which show that LMLFM outperforms the state-of-the-art methods in terms of predictive accuracy, variable selection ability, and scalability to data with large number of variables. The code and supplemental material is available at https://github.com/junjieliang672/LMLFM.

Introduction

Longitudinal data consist of repeated observations from a set of individuals over time. Such data are common in many areas, including health sciences, social sciences and economics. Consider for example, the scenario shown in Fig. 1. To predict an individual $\boldsymbol{x}_i$'s health status at the age of 38, we should take into account both $\boldsymbol{x}_i$'s history of physical examinations, as well as those of other individuals who are similar to $\boldsymbol{x}_i$ in age and other characteristics. Clearly, such longitudinal data often exhibit *longitudinal correlation* (LC) (correlations among observations for each individual over time), *cluster correlation* (CC) (correlations among individuals that have similar characteristics), or both (multi-level correlation) (Finch, Bolin, and Kelley 2016), and hence are not independent and identically distributed. Analysis that does not account for such correlations can lead to misleading statistical inferences (Gibbons and Hedeker 1997). To account for data correlations, state-of-the-art longitudinal data analysis (Gibbons and Hedeker 1997; Lozano and Swirszcz 2012; Groll and Tutz 2014) often relies on *mixed effects models* that include *fixed effects* and *random effects*, where the fixed effects capture the regression parameters that are shared by all individuals, whereas random effects capture those parameters that vary across individuals. In practice, the design of mixed effects models relies on expert input to decide which variables are subject to random effects as opposed to fixed effects, or a process of trial and error. However, existing mixed effects models are very computationally intensive, with the computational cost scaling with $O(q^3)$, where q is the number of variables that are subject to random effects which limits their applicability to relatively low-dimensional data. While with the advent of big data, variants of dimensionality reduction methods such as LASSO have been explored in the longitudinal setting (Schelldorfer, Bühlmann, and van de Geer 2011; Lozano and Swirszcz 2012; Groll and Tutz 2014; Xu, Sun, and Bi 2015; Ratkovic and Tingley 2017; Marino, Buxton, and Li 2017; Lu et al. 2017; Finch and others 2018), most such methods are limited to selecting only fixed effects. There is limited work on jointly selecting fixed and random effects, e.g., penalized likelihood methods (Ibrahim et al. 2011; Bondell, Krishna, and Ghosh 2010; Hui, Müller, and Welsh 2017b) and Bayesian models (Chen and Dunson 2003; Yang, Wang, and Dong 2019). However, their applicability is limited by their high computational cost and reliance on linear models.

Contributions. This paper aims to address the urgent need for effective models that can handle high-dimensional longitudinal data where the number of variables is very large compared to the size of the population, the interactions among variables can be nonlinear, the fixed and random effects are a priori unspecified, and the data exhibit correlation structure (LC, CC, or both). Specifically, we introduce Longitudinal Multi-Level Factorization Machine (LMLFM), a novel, efficient, provably convergent extension of *Factoriza-*

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

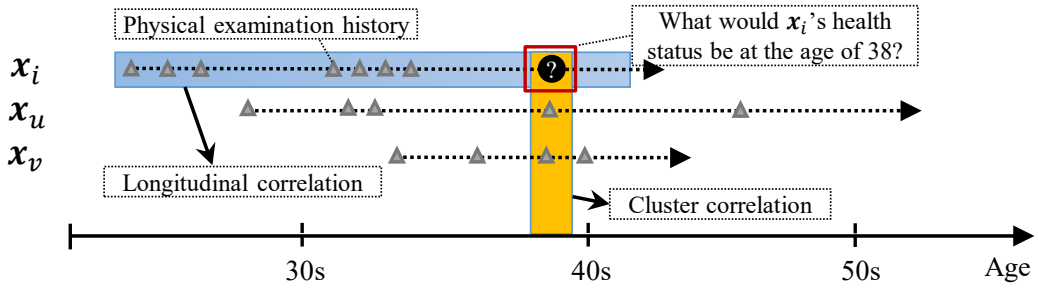


Figure 1: An example of longitudinal data. Note that the type of correlation behind the data is unknown. Hence, a predictive model should not only learn to accurately predict the outcome, but also exploit and identify the type of data correlation.

tion Machines (FM) (Rendle 2012) for predictive modeling of high-dimensional longitudinal data. LMLFM inherit the advantages of FM, e.g., the ability to reliably estimate the model parameters from high-dimensional data and model non-linear interactions. Further, LMLFM can automatically select fixed and random effects even in the presence of multi-level correlation, and greatly reduce the need for hyper-parameter tuning using a novel hierarchical Bayesian formulation. Specifically, LMLFM adopts two layers of Laplace prior, one for sparsifying the latent representation and one for identifying fixed effects and random effects. We solve the LMLFM using the iterated conditional modes (ICM) algorithm (Besag 1986) which offers efficient optimization with strong convergence guarantees. Experimental results with simulated data show that LMLFM can readily handle longitudinal data with over 5000 variables whereas the existing mixed effects models fail when the number of variables exceeds 100. Experiments with two real-world data sets show that LMLFM (i) compares favorably with the state-of-the-art baselines in terms of predictive accuracy; (ii) yields sparse and easy-to-interpret predictive models; and (iii) effectively selects the relevant variables, which are consistent with the published findings (Bromberger and Kravitz 2011; Dolan, Peasgood, and White 2008).

Related Work

Popular longitudinal data analysis methods include Generalized Estimating Equations (GEE) (Liang and Zeger 1986) and Generalized Mixed Effects Models (GMEM) (Fitzmaurice, Laird, and Ware 2012). GEE are marginal models which only estimate the average outcome (or fixed effects) over the population (Liang and Zeger 1986). In contrast, GMEM are conditional models that provide the expectation of the conditional distribution of the outcome given the random effects.

There is much interest in the problem of variable selection in longitudinal data (Schelldorfer, Bühlmann, and van de Geer 2011; Groll and Tutz 2014). Existing techniques focus on selecting only the fixed effects, under the assumption that the type of correlation is correctly specified and the random and fixed effects are correctly identified. Their high computational cost limits their applicability to data with small numbers of variables (Chen, Xu, and Bi 2018). There is limited work on the more challenging problem of selecting both fixed effects and random effects. Existing methods typically rely on adding a sparsity inducing penalty, e.g., LASSO or its variants, to the GMEM objective function (Bondell, Krishna, and Ghosh 2010; Ibrahim et al. 2011; Müller et al. 2013; Hui, Müller, and Welsh 2017a; 2017b). While Bayesian methods, *e.g.,* (Chen and Dunson 2003; Yang, Wang, and Dong 2019), offer a conceptually attractive alternative to penalized likelihood methods for variable selection, they are currently applicable only to 2-level data which exhibit only LC or only CC but not both. Furthermore, most assume a linear mixed model, and hence cannot accommodate non-linear interactions among variables. Because they rely on matrix decomposition and matrix inversion for parameter estimation, their computational complexity is $O(q^3)$, making them unsuitable for high-dimensional longitudinal data.

While there have been a few attempts at applying factorization techniques (Zhou et al. 2014; Stamile et al. 2017; Kidzinski and Hastie 2018), and deep representation learning techniques (Xu et al. 2019a; 2019b), their primary focus is to improve the predictive accuracy. These techniques do not explicitly account for the complex correlation structure in the data or distinguish between random effects and fixed effects. In contrast, LMLFM efficiently accounts for complex correlation structure in the data and selects the most predictive fixed and random effects.

Preliminaries

Notation. Scalars are denoted by lowercase letters and vectors by bold lowercase letters. All vectors are column vectors. $|\boldsymbol{\theta}|$ refers to the length of $\boldsymbol{\theta}$ and $\|\boldsymbol{\theta}\|_p$ is the ℓ_p norm of $\boldsymbol{\theta}$. Matrices are denoted by uppercase letters, *e.g.,* Θ and a set of objects by a bold uppercase letter, *e.g.,* $\boldsymbol{\Theta} = \{a, \boldsymbol{\theta}\}$. The calligraphic letters $\mathcal{I}$ and $\mathcal{O}$ denote information related to the individuals and the observations respectively. For example, $\Theta^{\mathcal{I}}$ refers to the sub-matrix of Θ associated to the individuals. We use the letters i, o to denote an arbitrary individual and observation respectively. We use $\mathrm{diag}(A)$ to denote the vector of diagonal components of a square matrix A. Because observations occur at discrete time points, we use observation and time point interchangeably.

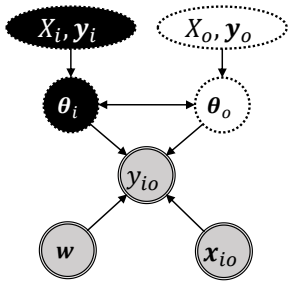


Figure 2: Model structure. Sub-graph with grey nodes encodes the structure of standard 1-level regression models; Sub-graph with grey and black nodes encodes the structure of 2-level LC models; Sub-graph with grey and white nodes denotes the structure of 2-level CC models; and the entire graph encodes the structure of multi-level models.

Factorization Machines (FM). Given a real-valued design feature vector $\boldsymbol{x}_{io} \in \mathbb{R}^p$ corresponding to an individual i and observation o, factorization machines (FM) (Rendle 2012) model all nested interactions up to order d. For example, the prediction of FM of order $d = 2$ is given by:

$$\hat{y}_{io} = \boldsymbol{x}_{io}^\intercal \boldsymbol{w} + \frac{1}{2}\boldsymbol{x}_{io}^\intercal M \boldsymbol{x}_{io} \tag{1}$$

where M is a squared matrix with zeros on the diagonal. The off-diagonal component $m_{qt} \in M$ is parameterized as a dot product of two low dimensional embeddings $\boldsymbol{\theta}_q, \boldsymbol{\theta}_t$. FM can be readily solved by coordinate descent (Rendle 2012). The time and space complexity are $O(k\,|S|)$ and $O(|\boldsymbol{y}|)$ respectively where $|\boldsymbol{y}|$ and $|S|$ denote the total number of observations across all individuals, and the data size (*i.e.,* $p\,|\boldsymbol{y}|$), respectively.

Linear Mixed Model (LMM). We introduce LMM to motivate the design of LMLFM. Let $y_{io} \in \mathbb{R}$ denote the scalar outcome of individual $i = 1, \cdots, n$ measured at observation $o = 1, \cdots, m$. Let $\boldsymbol{x}_{io} \in \mathbb{R}^p$, $\boldsymbol{z}_{io} \in \mathbb{R}^q \subseteq \boldsymbol{x}_{io}$ be variables associated with fixed effects (denoted by $\boldsymbol{\beta}$) and random effects (denoted as $\boldsymbol{\gamma}_i$) respectively. LMM assumes that the outcome is predicted by[1]:

$$\hat{y}_{io} = \boldsymbol{x}_{io}^\intercal \boldsymbol{\beta} + \boldsymbol{z}_{io}^\intercal \boldsymbol{\gamma}_i \tag{2}$$

The random effects matrix $(\gamma_1, ..., \gamma_n)^\intercal$ captures the time-invariant patterns for each individual. For all $i \in \mathcal{I}$, $\boldsymbol{\gamma}_i \sim N(\boldsymbol{0}, Q)$, where $Q \in \mathbb{R}^{q \times q}$ is the covariance matrix. The random effects $\boldsymbol{\gamma}_i$ serve two purposes: (i) regularizing the effects (similar to the ℓ_2 norm); and (ii) inducing correlation between the longitudinal observations, *i.e.,* $cov(y_{io}, y_{ij}) = \boldsymbol{z}_{io}^\intercal Q \boldsymbol{z}_{ij}$.

Longitudinal Multi-Level Factorization Machine (LMLFM)

The structures of multi-level models are shown in Fig. 2. Standard regression models assume that given the variables and regression parameters, the outcomes are i.i.d. and hence yield biased estimates of parameters in the presence of LC or CC. 2-level models account for the LC or CC by either directly or indirectly specifying a correlation matrix that models the corresponding correlations. Mixed effects models introduce individual (or observation) specific random effects as proxies for the relevant information (see Fig. 2). A natural approach to extend this design is to incorporate both individual factors $\boldsymbol{\theta}_i$ and observation factors $\boldsymbol{\theta}_o$, as proxies for the individual and observation specific information.[2] The pairwise interactions between such factors in the latent space (see Eq. (1)) are shown by arrows in Fig. 2. However, in its current form, the model does not provide a way to relate latent factors to the random effects or to explicitly accommodate complex correlation structures. **Given the observed design matrix X and outcomes $\boldsymbol{y}$, our goals are to: (i) predict the unknown outcomes $\hat{\boldsymbol{y}}$, (ii) jointly select both fixed and random effects and (iii) recover the correlation structure from the data.**

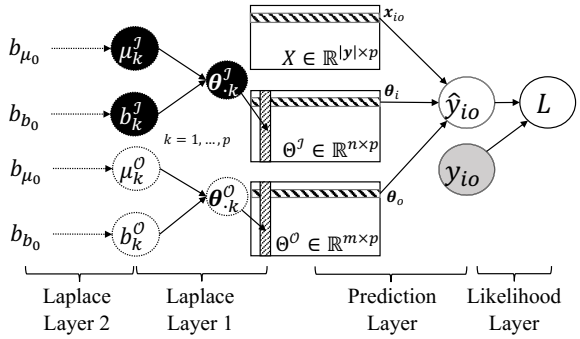


Figure 3: The hierarchical Bayesian structure of LMLFM. L stands for the objective function as presented in Eq. (4).

Prediction

The prediction layer of LMLFM (see Fig. 3) is inspired by both LMM and FM. With a Bayesian framework, all variables are first assumed random. Hence, the LMM prediction in Eq. (2) reduces to $\hat{y}_{io} = \boldsymbol{x}_{io}^\intercal \boldsymbol{\gamma}_{io}$ with $\boldsymbol{\gamma}_{io} \sim N(\boldsymbol{\beta}, Q)$. Further, to accommodate multi-level correlation, we decompose the random effects as the summation of two subsets of latent factors $\boldsymbol{\gamma}_{io} = \boldsymbol{\theta}_i + \boldsymbol{\theta}_o$, where $\boldsymbol{\theta}_i, \boldsymbol{\theta}_o$ denote the individual factors and observation factors respectively. Considering the interaction between individual and observation factors, we introduce the following prediction function:

$$\hat{y}_{io} = \boldsymbol{x}_{io}^\intercal (\boldsymbol{\theta}_i + \boldsymbol{\theta}_o) + \boldsymbol{\theta}_i^\intercal \boldsymbol{\theta}_o \tag{3}$$

Recall that in FM, the design feature vector $\boldsymbol{x}_{io}$ includes individual one-hot encoding, observation one-hot encoding and observed features. The original design of FM as in Eq. (1) embeds each component of the design feature vector into a latent vector. In contrast, LMLFM (Eq. (3)) embeds only the individuals and observations into latent vectors, and considers

[1] We omit the error term since it can be readily incorporated into the random effects.

[2] It is worth noting that individual/observation factors are distinct from individual/observation random effects, in this work, we use the former to estimate the latter.

only the interactions among individuals, observations, and the design feature vector, thus simplifying the model. Instead of setting the latent dimension empirically as in the case of a FM, we initially let the latent dimension k to be as large as the feature dimension p, and then apply variable selection to identify the relevant subset of latent factors (see below). This makes the proposed model almost as interpretable as a simple linear model while making use of factorization to accommodate nonlinear interactions among variables.

Hierarchical Bayesian Model

The hierarchical Bayesian structure of LMLFM is shown in Fig. 3. We decompose the task of joint selection of fixed and random effects into two sub-tasks: (i) identifying fixed and random effects by shrinking the variance of some variables towards 0; and (ii) selecting the relevant latent factors by shrinking some components towards 0. We handle the first sub-task by imposing a Laplace prior on μ_k and b_k, respectively (see Laplace layer 2 in Fig. 3). For the second sub-task, we enforce sparsity in $\boldsymbol{\theta}_{\cdot k}$ by imposing Laplace prior on $\boldsymbol{\theta}_{\cdot k}$ (see Laplace layer 1 in Fig. 3). We denote the model parameters and hyper-priors of LMLFM by $\boldsymbol{\Theta} = \{\alpha, \Theta, \boldsymbol{\mu}, \boldsymbol{b}\}$ and $\boldsymbol{\Theta}_0 = \{\alpha_0, \beta_0, b_{\mu_0}, b_{b_0}\}$ respectively, yielding the following generative model:

$$
\begin{array}{ll}
(y_{io}|\boldsymbol{x}_{io}, \boldsymbol{\theta}_i, \boldsymbol{\theta}_o, \alpha) \sim N(y_{io}|\hat{y}_{io}, \alpha^{-1}) & (\alpha|\alpha_0, \beta_0) \sim Gamma(\alpha|\alpha_0, \beta_0) \\
(\theta_{ik}|\mu_k^{\mathcal{I}}, b_k^{\mathcal{I}}) \sim Laplace(\theta_{ik}|\mu_k^{\mathcal{I}}, b_k^{\mathcal{I}}) & (\theta_{ok}|\mu_k^{\mathcal{O}}, b_k^{\mathcal{O}}) \sim Laplace(\theta_{ok}|\mu_k^{\mathcal{O}}, b_k^{\mathcal{O}}) \\
(\mu_k^{\mathcal{I}}|b_{\mu_0}) \sim Laplace(\mu_k^{\mathcal{I}}|0, b_{\mu_0}) & (\mu_k^{\mathcal{O}}|b_{\mu_0}) \sim Laplace(\mu_k^{\mathcal{O}}|0, b_{\mu_0}) \\
(b_k^{\mathcal{I}}|b_{b_0}) \sim Laplace(b_k^{\mathcal{I}}|0, b_{b_0}) & (b_k^{\mathcal{O}}|b_{b_0}) \sim Laplace(b_k^{\mathcal{O}}|0, b_{b_0})
\end{array}
$$

Unlike Bayesian FM (Rendle 2012), to accommodate different degrees of sparsity in relation to the numbers of individuals and observations, we allow different hierarchical priors for different latent factors for individuals ($\Theta^{\mathcal{I}}$) and for observations ($\Theta^{\mathcal{O}}$) (See the grey nodes for $\Theta^{\mathcal{I}}$ and black nodes for $\Theta^{\mathcal{O}}$ in Fig. 3). We use $\boldsymbol{\mu}^\intercal = (\mu_k^{\mathcal{I}}, \mu_k^{\mathcal{O}})_{k=1}^p, \boldsymbol{b}^\intercal = (b_k^{\mathcal{I}}, b_k^{\mathcal{O}})_{k=1}^p$ to denote the mean and scale of the Laplace distribution respectively. The choice of the distribution of y_{io} can be application-dependent. For the ease of exposition, we assume that the outcome variable follows a Gaussian distribution.

We adopt the iterated conditional modes (ICM) algorithm (Besag 1986) to estimate the parameters of LMLFM. ICM updates blocks of parameters with the modes of their conditional posterior while keeping the remaining parameters fixed. Our choice of priors permits the analytical closed-form derivation of the modes of the conditional posterior density, yielding substantial speedup. Specifically, we consider the Maximum A Posteriori (MAP) formulation:

$$\arg\max_{\boldsymbol{\Theta}} L = \pi\left(\boldsymbol{\Theta}|\boldsymbol{y}, X, \boldsymbol{\Theta}_0\right) \tag{4}$$

Due to space constraints, we include only the update equations for $\boldsymbol{\Theta}^{\mathcal{I}} = \left\{\alpha, \boldsymbol{\Theta}^{\mathcal{I}}, \boldsymbol{\mu}^{\mathcal{I}}, \boldsymbol{b}^{\mathcal{I}}\right\}$ here, omitting the superscript $\mathcal{I}$ to minimize notational clutter.

Update of α. The posterior of α is a Gamma distribution, whose mode is given by $\left(\beta_0 + \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2/2\right)^{-1}(\alpha_0 + |\boldsymbol{y}|/2 - 1)$.

Update of Θ. For each model parameter $\boldsymbol{\theta}_i \in \Theta$, the prediction is a linear combination of two functions $g(i)$ and $h(i)$ that are independent of the value of $\boldsymbol{\theta}_i$:

$$\hat{\boldsymbol{y}}_i = g(i) + h(i)\boldsymbol{\theta}_i \tag{5}$$

with $g(i) = \mathrm{diag}(X_i \cdot \Theta_i^{\mathcal{O}\intercal})$ and $h(i) = X_i + \Theta_i^{\mathcal{O}}$. Here $\Theta_i^{\mathcal{O}}$ is the matrix of latent factors constructed by the observations associated with i. Hence, we have:

$$\theta_{ik} = \mu_k + \left(\boldsymbol{h}_{ik}^\intercal \boldsymbol{h}_{ik}\right)^{-1} \mathrm{sgn}\left(\boldsymbol{r}_{ik}\right)\left(|\boldsymbol{r}_{ik}| - 1/\alpha b_k^{\mathcal{I}}\right)_+ \tag{6}$$

where $(\cdot)_+$ is the ReLU function; $\boldsymbol{r}_{ik} = \boldsymbol{h}_{ik}^\intercal(\boldsymbol{y}_i - g(i) - \sum_{q\in\{1:p\}\setminus k}\theta_{iq}\boldsymbol{h}_{iq} - \boldsymbol{h}_{ik}\mu_k)$, with $\{1:p\}\setminus k$ denoting the set of integers ranging from 1 to p excluding k and $\boldsymbol{h}_{ik}$ is the k-th column of $h(i)$. Clearly, sparsity is achieved if $|\boldsymbol{r}_{ik}| \leq 1/\alpha b_k^{\mathcal{I}}$ and $\mu_k = 0$.

Update of $\boldsymbol{\mu}$. The problem of finding the optimal μ_k $(k = 1, \cdots, p)$ reduces to finding the weighted median of the vector $\boldsymbol{\theta}_{\cdot k} \cup \{0\}$ with the weights $\{b_k\}_{i=1}^n \cup \{b_{\mu_0}\}$, yielding a linear-time algorithm (Gurwitz 1990).

Update of $\boldsymbol{b}$. The optimal b_k $(k = 1, \cdots, p)$ is updated by $2b_{b_0}\left(\sqrt{n^2 + \frac{4}{b_{b_0}}\|\boldsymbol{\theta}_{\cdot k} - \mu_k\|_1} - n\right)$.

We note that the computational complexity of LMLFM for one complete iteration is $O(|S|)$, which is linear in the size of the training data. Our approach is more efficient than FM (Rendle 2012), whose computational complexity is $O(k\,|S|)$ (k is the latent dimension). The space complexity of LMLFM is the same as that of FM, i.e., $O(|\boldsymbol{y}|)$.

Effects and Outcome Estimation

We proceed to describe how to estimate random effects, fixed effects and outcomes for seen and unseen individual/observation from the model.

Temporal Individual-Specific Random Effects (TISE). As shown in Eq. (3), we can rewrite the prediction function of LMLFM in a form resembling ordinary linear regression where $\hat{y}_{io} = \boldsymbol{x}_{io}^\intercal \boldsymbol{\gamma}_{io} + \epsilon_{io}$ with coefficients $\boldsymbol{\gamma}_{io} = \boldsymbol{\theta}_i + \boldsymbol{\theta}_o$ and error $\epsilon_{io} = \boldsymbol{\theta}_i^\intercal \boldsymbol{\theta}_o$. Hence, we let $\boldsymbol{\gamma}_{io}$ be the estimator of TISE.

Averaged Individual-Specific Random Effects (AISE). AISE is computed by integrating out the observation effects:

$$\boldsymbol{\gamma}_i = \mathbb{E}_{\pi(\boldsymbol{\theta}_o|\boldsymbol{y},X,\boldsymbol{\Theta}_0)}[\boldsymbol{\gamma}_{io}] = \boldsymbol{\theta}_i + \mathbb{E}_{\pi(\boldsymbol{\theta}_o|\boldsymbol{y},X,\boldsymbol{\Theta}_0)}\left[\boldsymbol{\theta}_o\right]$$

Solving $\mathbb{E}_{\pi(\boldsymbol{\theta}_o|\boldsymbol{y},X,\boldsymbol{\Theta}_0)}\left[\boldsymbol{\theta}_o\right]$ is non-trivial. Hence, we approximate it using the estimated observation factors, where $\mathbb{E}_{\pi(\boldsymbol{\theta}_o|\boldsymbol{y},X,\boldsymbol{\Theta}_0)}\left[\boldsymbol{\theta}_o\right] \approx \frac{1}{m}\sum_{o\in\mathcal{O}}\boldsymbol{\theta}_o$.

Temporal Population Averaged Random Effects (TPAE). Similar to solving AISE, TPAE is solved by integrating out the individual factors: $\boldsymbol{\gamma}_o = \mathbb{E}_{\pi(\boldsymbol{\theta}_i|\boldsymbol{y},X,\boldsymbol{\Theta}_0)}[\boldsymbol{\gamma}_{io}] \approx \boldsymbol{\theta}_o + \frac{1}{n}\sum_{i\in\mathcal{I}}\boldsymbol{\theta}_i$.

Fixed Effects. We say that a variable k has fixed effect if and only if $b_k^{\mathcal{I}} = b_k^{\mathcal{O}} = 0$. The fixed effect for variable k is computed by $\beta_k = \mu_k^{\mathcal{I}} + \mu_k^{\mathcal{O}}$.

Outcomes. The outcome y_{io} for seen individual i and observation o is computed using Eq. (3). In the case of unseen individuals, we replace the posterior of the latent factors with

their prior. Thus, for an unseen individual i, the outcome y_{io} is given by:

$$\mathbb{E}[y_{io}|\boldsymbol{x}_{io}, \boldsymbol{\theta}_i, \boldsymbol{\theta}_o, \alpha] = \boldsymbol{x}_{io}^{\intercal}(\boldsymbol{\mu}^{\mathcal{I}} + \boldsymbol{\theta}_o) + \boldsymbol{\theta}_o^{\intercal}\boldsymbol{\mu}^{\mathcal{I}}$$

Convergence Analysis

We establish two important properties in LMLFM: *Ascent property* and *Convergence*.[3] Let $\mathbb{Z}^+$ denotes the set of all positive integers and $\theta^{(t)}$ denotes the value of θ at the t-th iteration. We denote $\pi\left(\theta|\right)$ as the full conditional posterior of θ.

Proposition 1 ***Ascent property.*** $\pi\left(\boldsymbol{\Theta}^{(t+1)}|\right) \geq \pi\left(\boldsymbol{\Theta}^{(t)}|\right)$ *holds for all iterations* $t \in \mathbb{Z}^+$.

The proof of Proposition 1 follows from the observation that the joint posterior density of LMLFM is non-decreasing with the update of each component of $\boldsymbol{\Theta}$.

Proposition 2 ***Convergence.*** *If* $\pi\left(\Theta^{(t)}|\right)$ *is bounded above, there exists an iteration* $t \in \mathbb{Z}^+$*, such that* $\forall i \in \mathbb{Z}^+, \left|\pi\left(\Theta^{(t+i)}|\right) - \pi\left(\Theta^{(t)}|\right)\right| < \epsilon$ *holds for* $\epsilon > 0$.

We prove Proposition 2 by contradiction, *i.e.,* the negation of the proposition implies that $\lim_{t\to\infty} \pi\left(\Theta^{(t)}|\right) \to \infty$, yielding a contradiction.

Experimental Evaluation

Experiments with Simulated Data

We report results of experiments with simulated data to answer the following questions: (RC1) Can LMLFM handle high-dimensional data? (RC2) Can LMLFM accurately select the relevant variables? (RC3) How does LMLFM perform in the presence of LC and CC?

Simulated data. Following (Hui, Müller, and Welsh 2017b), we construct simulated longitudinal data sets with 40 individuals and 40 observations per individual. We consider several choices of p from $\{50, 100, 500, 1000, 5000\}$. We consider three types of correlation, *i.e.,* pure LC, pure CC and both (See supplemental material for details.).

Methods compared. We compare LMLFM with several baseline methods: (i) State-of-the-art multi-level linear mixed model (M-LMM) (Bates et al. 2015); (ii) State-of-the-art 2-level models: LMMLASSO, a linear mixed model with adaptive LASSO penalty on the fixed effects (Schelldorfer, Bühlmann, and van de Geer 2011); GLMMLASSO, a generalized linear mixed model with standard LASSO penalty on the fixed effects (Groll and Tutz 2014) and rPQL (Hui, Müller, and Welsh 2017b), a joint selection mixed model with adaptive LASSO penalty on the fixed effects and group LASSO penalty on the random effects; and (iii) Factorization-based multi-level Lasso (MLLASSO),[4] which factorizes the fixed effects as a product of global effects and individual effects, both regularized by ℓ_1 norm (Lozano and Swirszcz 2012); and (iv) the standard LASSO regression (LASSO) (Tibshirani 1996). We report performance statistics obtained from 100 independent runs. Hyper-parameters are selected using cross validation on the training data. Evaluation scores are computed on the held-out data set. We report execution failure if an algorithm fails to converge within 48 hours or generates an execution error.

Table 1: Performance comparison on simulated data in the presence of multi-level correlation. We use '-' to denote execution failure.

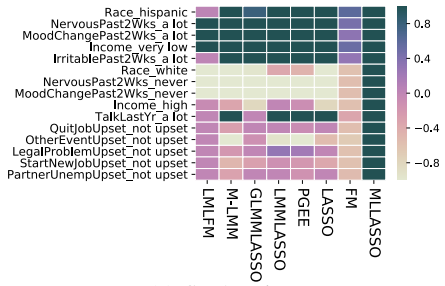
Method	$p = 100$			$p = 5000$		
	R^2 (%)	f.p.	f.n.	R^2 (%)	f.p.	f.n.
LMLFM	**92±1**	**0.2**	0.8	**88±2**	2.2	4.2
rPQL	88±2	20.6	**0**	-	-	-
M-LMM	90±1	92	**0**	-	-	-
GLMMLASSO	83±4	91	**0**	-	-	-
LMMLASSO	88±2	92	**0**	-	-	-
LASSO	88±1	42.4	**0**	84±4	415.8	0.4
MLLASSO	40±8	23.8	0.8	1±1	**0**	6.2

Evaluation Measures. We evaluate the performance of all methods in terms of both predictive accuracy and the ability to select random and fixed effects. We measure the predictive accuracy using the r-squared (R^2) score. To assess a method's ability to select the relevant variables, we consider a variable to be selected if the corresponding coefficient is non-zero. We use false positive (f.p.), the number of variables that are incorrectly selected, and false negative (f.n.), the number of variables that are incorrectly discarded, to assess the models.

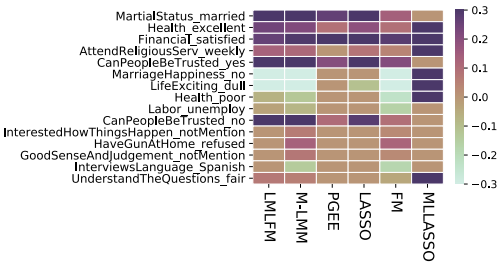
Results. A subset of our results summarized in Table 1 answer RC1-RC3. RC1: the performance of the state-of-the-art mixed effects models are highly sensitive to the number of random effects, *i.e.,* M-LMM, LMMLASSO and GLMMLASSO fail when p exceeds 100 due to execution error; Only LMLFM, LASSO and MLLASSO ran to completion. RC2: Selecting random effects is more challenging compared to selecting fixed effects. Among all models, only LMLFM and rPQL are designed to select random effects. To enforce model sparsity, say on variable k, LMLFM and rPQL shrink the vector $\boldsymbol{\theta}_{\cdot k}$ to 0 whereas other methods shrink a scalar β_k to 0. Our results show that LMLFM achieves the best terms of f.p. score. Although LASSO has attractive R^2 when $p = 5000$, LASSO is unable to select variables with random effects, leading to very high f.p. in the presence of LC, CC and both. RC3: Our results (omitted due to space constraints) show that 2-level models work poorly when a CC model is used on data with LC or vice versa. In contrast, multi-level models (M-LMM and LMLFM) achieve better fit on the data that exhibit LC, CC, or both. LMLFM consistently outperforms M-LMM in terms of accuracy, variable selection ability and computational efficiency. We conclude that LMLFM is the only method among those compared in this study, that can effectively model high-dimensional longitudinal data, and select the relevant variables, regardless of whether they are associated with random effects or fixed effects, in the presence of LC, CC, or both.

[3] See supplemental material for detailed proofs.

[4] Despite its name, MLLASSO works only as a 2-level model and does not provide a simple way to associate the latent factors with random effects.



(a) SWAN data



(b) GSS data

Figure 4: Comparison of population averaged effects on the selected variables for SWAN and GSS data. The Top 5, middle 5 and bottom 5 variables have positive, negative and neutral effects respectively.

Experiments with Real-World Data

We compare LMLFM with the state-of-the-art baselines on two real-world longitudinal data sets: (i) Study of Women's Health Across the Nation (SWAN) [5] (Sutton-Tyrrell et al. 2005) (on predicting depression); and (ii) General social survey (GSS)[6] (Smith et al. 2017) (on predicting general happiness). We chose these two data sets because they have attracted much interest in the field of social sciences. We use the same settings of hyper-parameters for LMLFM as in our experiments with simulated data. We exclude rPQL because it fails on all of the experiments due to memory issue. In addition to the aforementioned baselines, we include some popular 1-level models in our comparison: Random Forest (RF), FM and Penalized GEE (PGEE) (Inan and Wang 2017).

We seek answers to the following question: (RC4) How does LMLFM compare with the state-of-the-art baselines with respect to its ability to correctly identify the fixed and random effects and predictive accuracy? To answer RC4, for each data set, we choose as "ground truth", 5 "positive", 5 "negative" variables identified in the existing literature (see below for specifics) and add 5 additional variables that are believed to be relatively uninformative.

Evaluation on SWAN Data. In the case of SWAN data, we consider the task of predicting the CESD score (Dugan et al. 2015), which is used for screening for depression. The variables of interest include aspects of physical and mental health, and demographic factors, such as race and income. The data set includes 3,300 individuals, with 1-22 observations per individual, and 137 variables. The outcome we aim to predict is defined by $y_{io} = \text{CESD}_{io} - 15$ (i is individual and o is the age of the individual) since CESD ≥ 16 has been observed to be highly indicative of depression. Existing research (Dugan et al. 2015; Prairie et al. 2015) suggests that hispanic ethnicity, depressed or fluctuating mood and low household income are highly positively correlated with depression, whereas Caucasian/white ethnicity, stable mood and high income are negatively correlated with depression. The variables used to answer RC4 and the experimental results are summarized in Fig. 4(a) and Table 2 respectively. We note that LMLFM outperforms all other methods in R^2 score and correctly recovers the relevant variables. Performance of the factorization baselines (FM, MLLASSO) is unsatisfactory. This is because of the lack of intuitive way to relate estimated latent factors to the corresponding effects. Note that the variables related to depressed mood are generally selected by our baselines (not shown), which is consistent with the findings in (Prairie et al. 2015). FM renders menopausal status as strong factors to depression, a finding supported by existing literature (Bromberger and Kravitz 2011; Prairie et al. 2015). However, we argue that depressed and fluctuated mood is more likely to be direct causes to depressive symptoms because menopausal status usually causes abnormal hormone level, which could further affect the mood of the patients. Results of LMLFM show that $\boldsymbol{\mu}^{\mathcal{I}} = \mathbf{0}$ and the sparsity rate of $\Theta^{\mathcal{I}}$ (*i.e.,* the number of zero components in $\Theta^{\mathcal{I}}$ divided by $\left|\Theta^{\mathcal{I}}\right|$) is 98.3%, which further implies that the random effects related to the individuals are less predictive. Nonetheless, the analysis on $\boldsymbol{\mu}^{\mathcal{O}}$ reveals a different story: 59 out of 136 $\boldsymbol{\mu}^{\mathcal{O}}$ are non-zero and the sparsity rate of $\Theta^{\mathcal{O}}$ is 56.4%. This suggests that the depressive symptoms for multiple individuals with similar age are correlated (CC) as are the depressive symptoms for a single individual across time (LC), with CC dominating LC.

Evaluation on GSS Data. In our experiments with the GSS data, we consider the problem of predicting the self-reported happiness. We define $y_{io} = 1$ by individual i reports happy at year o and $y_{io} = -1$ as the opposite. The GSS data consists of 4,510 individuals, 1-30 observations per individual and 1,553 variables. Existing research (Dolan, Peasgood, and White 2008; Oishi, Kesebir, and Diener 2011) indicates that, being married, good physical and psychological health, satisfactory with financial situation, having strong religious beliefs and being trusted are positively correlated with happiness, whereas the absence of these characteristics and unemployment are negatively correlated with happiness. The variables used to answer RC4 and the results of our experiments are shown in Fig. 4(b) and Table 2 respectively. Though we see that PGEE and LASSO have the lowest f.p., their R^2 is relatively low. They tend to shrink the negative effects to zero, and perform poorly even on the training set, which strongly suggests that they underfit the data. LMLFM is competitive with the best performing methods in recovering the relevant variables, while significantly outperforming them in predictive perfor-

[5] https://www.swanstudy.org/

[6] http://gss.norc.org/

Table 2: Comparison of predictive accuracy on two real-life data sets. We use subscript '$-$' and '$+$' to denote the score on data with the 15 selected variables and all variables, respectively. We use '-' to denote execution failure.

Method	SWAN			GSS				
	$R^2_-(\%)$	f.p.$_-$	f.n.$_-$	$R^2_+(\%)$	$R^2_-(\%)$	f.p.$_-$	f.n.$_-$	$R^2_+(\%)$
LMLFM	**30±4**	**0**	**0**	**49±2**	**17±2**	2	**0**	**55±2**
M-LMM	29±2	1	5	-	16±1	2	1	-
GLMMLASSO	19±3	**0**	2	-	-	-	-	-
LMMLASSO	26±3	2	2	-	-	-	-	-
PGEE	25±4	2	5	-	3±1	**1**	**0**	-
LASSO	21±4	1	2	47±1	0±1	**1**	**0**	47±1
FM	29±3	**0**	5	45±2	12±4	4	2	31±2
RF	23±5	10	**0**	47±1	4±1	10	**0**	**55±2**
MLLASSO	0±2	10	**0**	20±2	-42±10	4	**0**	-2±1

mance. We note that variable selection with the GSS data is far more challenging than with the SWAN data because of the substantially larger number of variables and collinearity of the variables. The low R^2 of FM and MLLASSO indicate that they significantly underfit the data. Though RF has a high R^2 score, variables selected by RF are harder to explain compared to that of the other baselines (not shown). In contrast, LASSO achieves lower R^2, but selects variables that are consistent with those selected by LMLFM. We further find that all of the variables selected by LMLFM are consistent with the findings reported in (Dolan, Peasgood, and White 2008). We find that $\Theta^{\mathcal{I}} = 0$, thus ruling out LC. This is perhaps explained by the huge gap between consecutive observations (the survey is taken once per one to three years) within which many unobserved factors could potentially affect subjective happiness. We find that the sparsity rate of $\Theta^{\mathcal{O}}$ is 84.9%. Our analysis of the fixed effects shows that 126 out of 1199 effects are non-zero and among which, only 6 features have absolute effects greater than 0.1, thus vast majority of variables are uninformative.

Summary of Experimental Results. We conclude that LMLFM outperforms all the baselines and is the only multi-level mixed effects model that can reliably select variables associated with fixed as well as random effects from high-dimensional longitudinal data.

Conclusions

We have introduced LMLFM, for predictive modeling from longitudinal data when the number of variables is large compared to the population size, the fixed and random effects are a priori unspecified, the interactions among variables are nonlinear, and the data exhibit complex correlation (LC, CC, or both). LMLFM, a natural generalization of FM to longitudinal data setting, adopts a novel hierarchical Bayesian model with two layers of Laplace prior, where the first layer induces a sparse latent representation and the second layer identifies fixed effects and random effects. We train LMLFM using iterated conditional modes algorithm which offers both computational efficiency and strong convergence guarantee. Compared to the state-of-the-art alternatives, LMLFM yields more compact, easy-to-interpret, rapidly trainable, and hence scalable, models with minimal need for hyper-parameter tuning. Our experiments with simulated data with thousands of variables, and two widely studied real-world longitudinal data sets have shown that LMLFM outperforms the 1-level baselines and state-of-the-art 2-level and multi-level longitudinal models in terms of predictive accuracy, variable selection ability, and scalability to data with large number of variables.

Acknowledgments

This work was funded in part by the NIH NCATS through the grant UL1 TR002014 and by the NSF through the grants 1518732, 1640834, and 1636795, the Edward Frymoyer Endowed Professorship at Pennsylvania State and the Sudha Murty Distinguished Visiting Chair in Neurocomputing and Data Science funded by the Pratiksha Trust at the Indian Institute of Science (both held by Vasant Honavar). The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

References

Bates, D.; Mchler, M.; Bolker, B.; and Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles* 67(1):1–48.

Besag, J. 1986. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)* 259–302.

Bondell, H. D.; Krishna, A.; and Ghosh, S. K. 2010. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* 66(4):1069–1077.

Bromberger, J. T., and Kravitz, H. M. 2011. Mood and menopause: findings from the study of women's health across the nation (swan) over 10 years. *Obstetrics and Gynecology Clinics* 38(3):609–625.

Chen, Z., and Dunson, D. B. 2003. Random effects selection in linear mixed models. *Biometrics* 59(4):762–769.

Chen, K.-S.; Xu, T.; and Bi, J. 2018. Latent sparse modeling of longitudinal multi-dimensional data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Dolan, P.; Peasgood, T.; and White, M. 2008. Do we really know what makes us happy? a review of the economic literature on the factors associated with subjective well-being. *Journal of economic psychology* 29(1):94–122.

Dugan, S. A.; Bromberger, J. T.; Segawa, E.; Avery, E.; and Sternfeld, B. 2015. Association between physical activity and depressive symptoms: midlife women in swan. *Medicine and science in sports and exercise* 47(2):335.

Finch, W. H., et al. 2018. Modeling high dimensional multilevel data using the lasso estimator: A simulation study. *Journal of Statistical and Econometric Methods* 7(1):51–75.

Finch, W. H.; Bolin, J. E.; and Kelley, K. 2016. *Multilevel modeling using R*. Crc Press.

Fitzmaurice, G. M.; Laird, N. M.; and Ware, J. H. 2012. *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

Gibbons, R. D., and Hedeker, D. 1997. Random effects probit and logistic regression models for three-level data. *Biometrics* 1527–1537.

Groll, A., and Tutz, G. 2014. Variable selection for generalized linear mixed models by l 1-penalized estimation. *Statistics and Computing* 24(2):137–154.

Gurwitz, C. 1990. Weighted median algorithms for l 1 approximation. *BIT* 30(2):301–310.

Hui, F. K.; Müller, S.; and Welsh, A. 2017a. Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statistica Sinica* 501–518.

Hui, F. K.; Müller, S.; and Welsh, A. 2017b. Joint selection in mixed models using regularized pql. *Journal of the American Statistical Association* 112(519):1323–1333.

Ibrahim, J. G.; Zhu, H.; Garcia, R. I.; and Guo, R. 2011. Fixed and random effects selection in mixed effects models. *Biometrics* 67(2):495–503.

Inan, G., and Wang, L. 2017. Pgee: An r package for analysis of longitudinal data with high-dimensional covariates. *R Journal* 9(1):393–402.

Kidzinski, L., and Hastie, T. J. 2018. Longitudinal data analysis using matrix completion. *CoRR* abs/1809.08771.

Liang, K.-Y., and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73(1):13–22.

Lozano, A. C., and Swirszcz, G. 2012. Multi-level lasso for sparse multi-task regression. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 595–602.

Lu, P.; Colliot, O.; Initiative, A. D. N.; et al. 2017. Multilevel modeling with structured penalties for classification from imaging genetics data. In *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*. Springer. 230–240.

Marino, M.; Buxton, O. M.; and Li, Y. 2017. Covariate selection for multilevel models with missing data. *Stat* 6(1):31–46.

Müller, S.; Scealy, J. L.; Welsh, A. H.; et al. 2013. Model selection in linear mixed models. *Statistical Science* 28(2):135–167.

Oishi, S.; Kesebir, S.; and Diener, E. 2011. Income inequality and happiness. *Psychological science* 22(9):1095–1100.

Prairie, B. A.; Wisniewski, S. R.; Luther, J.; Hess, R.; Thurston, R. C.; Wisner, K. L.; and Bromberger, J. T. 2015. Symptoms of depressed mood, disturbed sleep, and sexual problems in midlife women: cross-sectional data from the study of women's health across the nation. *Journal of Women's Health* 24(2):119–126.

Ratkovic, M., and Tingley, D. 2017. Sparse estimation and uncertainty with application to subgroup analysis. *Political Analysis* 25(1):1–40.

Rendle, S. 2012. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3(3):57.

Schelldorfer, J.; Bühlmann, P.; and van de Geer, S. 2011. Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics* 38(2):197–214.

Smith, T.; Marsden, P.; Hout, M.; and Kim, J. 2017. General social surveys, 1972–2014 [machine-readable data file]/principal investigator. *Sponsored by national science foundation. Chicago: National Opinion Research Center at the University of Chicago [producer and distributor]*.

Stamile, C.; Kocevar, G.; Cotton, F.; Maes, F.; Sappey-Marinier, D.; and Van Huffel, S. 2017. Multiparametric non-negative matrix factorization for longitudinal variations detection in white-matter fiber bundles. *IEEE journal of biomedical and health informatics* 21(5):1393–1402.

Sutton-Tyrrell, K.; Wildman, R. P.; Matthews, K. A.; Chae, C.; Lasley, B. L.; Brockwell, S.; Pasternak, R. C.; Lloyd-Jones, D.; Sowers, M. F.; Torréns, J. I.; et al. 2005. Sex hormone–binding globulin and the free androgen index are related to cardiovascular risk factors in multiethnic premenopausal and perimenopausal women enrolled in the study of women across the nation (swan). *Circulation* 111(10):1242–1249.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Xu, D.; Cheng, W.; Luo, D.; Gu, Y.; Liu, X.; Ni, J.; Zong, B.; Chen, H.; and Zhang, X. 2019a. Adaptive neural network for node classification in dynamic networks. In *ICDM*. IEEE.

Xu, D.; Cheng, W.; Luo, D.; Liu, X.; and Zhang, X. 2019b. Spatio-temporal attentive rnn for node classification in temporal attributed graphs. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 3947–3953. AAAI Press.

Xu, T.; Sun, J.; and Bi, J. 2015. Longitudinal lasso: Jointly learning features and temporal contingency for outcome prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1345–1354. ACM.

Yang, M.; Wang, M.; and Dong, G. 2019. Bayesian variable selection for mixed effects model with shrinkage prior. *Computational Statistics* 1–17.

Zhou, J.; Wang, F.; Hu, J.; and Ye, J. 2014. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 135–144. ACM.