Supplementary Material for LMLFM

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

Derivation of LMLFM

In this section, we detail the process of solving LMLFM. Fig. outlines the hierarchical Bayesian model of LMLFM. The resulting generative model is as follows:

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

we consider solving the Maximum A Posteriori (MAP) problem. Thus, the objective function is expressed as:

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} \pi\left(\boldsymbol{\Theta}|\boldsymbol{y}, X, \boldsymbol{\Theta_0}\right) \tag{1}$$

where the model parameters and hyperpriors are represented by $\boldsymbol{\Theta} = \{\alpha, \Theta, \boldsymbol{b}, \boldsymbol{\mu}\}$ and $\boldsymbol{\Theta}_0 = \{\alpha_0, \beta_0, b_{b_0}, b_{\mu_0}\}$ respectively. In the following, we only provide the derivation w.r.t. the individual parameters, i.e., $\boldsymbol{\Theta}^{\mathcal{I}} = \left\{\alpha, \boldsymbol{\Theta}^{\mathcal{I}}, \boldsymbol{\mu}^{\mathcal{I}}, \boldsymbol{b}^{\mathcal{I}}\right\}$. To keep the notation light, we omit the superscript $\mathcal{I}$ till the end of this section.

Update of $\Theta^{\mathcal{I}}$: 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 $\boldsymbol{\theta}_i$:

$$\hat{\boldsymbol{y}}_i = g(i) + h(i)\boldsymbol{\theta}_i \tag{2}$$

with

$$g(i) = \mathrm{diag}(X_i \cdot \Theta_i^{\mathcal{O}\intercal}) \qquad h(i) = X_i + \Theta_i^{\mathcal{O}}$$

where $\Theta_i^{\mathcal{O}}$ is the matrix of latent factors constructed by the observations associated to i. Thus we have $\pi(\boldsymbol{\theta}_i) = \pi(\boldsymbol{y}_i|X_i, \boldsymbol{\theta}_i, \alpha) \cdot \pi(\boldsymbol{\theta}_i|\mu_k, b_k)$, which is given by:

$$\pi(\boldsymbol{\theta}_i|) \propto \exp\left\{-\frac{\alpha}{2}\left\|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\right\|_2^2\right\} \cdot \prod_{k=1}^{p} \exp\left\{-\frac{|\theta_{ik} - \mu_k|}{b_k}\right\}$$

The log of which is:

$$\ell(\boldsymbol{\theta}_i) = -\frac{\alpha}{2}\left\|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\right\|_2^2 - \sum_{k=1}^{p}\frac{|\theta_{ik} - \mu_k|}{b_k} + \text{const.} \tag{3}$$

Let's define the gradient of $|\theta_{ik} - \mu_k|$ by e. Then following the sub-gradient equations (see e.g., (Bertsekas 1999)), we have $e = \mathrm{sgn}\left(\theta_{ik} - \mu_k\right)$ if $\theta_{ik} \neq \mu_k$ and $e \in [-1, 1]$ otherwise. For cases where $|\theta_{ik} - \mu_k|$ is differentiable, the optimal θ_{ik}^* can be derived by simply setting the gradient of (3) to 0. Otherwise, $\theta_{ik} = \mu_k$. Consequently, we have the following results:

$$\theta_{ik}^* = \left(\boldsymbol{h}_{ik}^\intercal \boldsymbol{h}_{ik}\right)^{-1}\left(\boldsymbol{h}_{ik}^\intercal \boldsymbol{h}_{ik}\mu_k + \mathrm{sgn}\left(\boldsymbol{r}_{ik}\right)\left(|(\boldsymbol{r}_{ik}| - 1/\alpha b_k^{\mathcal{I}}\right)_+\right) \tag{4}$$

where $(\cdot)_+$ is the ReLU function; $\boldsymbol{r}_{ik} = \boldsymbol{h}_{ik}^\intercal(\boldsymbol{y}_i - g(i) - \sum_{q\in 1:p\backslash k}\theta_{iq}\boldsymbol{h}_{iq} - \boldsymbol{h}_{ik}\mu_k)$, with $1 : p\backslash k$ denoting the set of integers ranging from 1 to p excluding k and $\boldsymbol{h}_{ik}$ is the k-th column of $h(i)$.

Update of α: $\pi(\alpha|) = \pi(\boldsymbol{y}|X, \Theta, \alpha) \cdot \pi(\alpha|\alpha_0, \beta_0)$. The full conditional posterior of α can be obtained by:

$$
\begin{aligned}
\pi(\alpha|) &\propto \prod_{j=1}^{|\boldsymbol{y}|}\alpha^{1/2}\exp\left\{-\frac{\alpha}{2}\left(y_j - \hat{y}_j\right)^2\right\} \cdot \alpha^{\alpha_0 - 1}\exp\left\{-\alpha\beta_0\right\}\\
&\propto \alpha^{\alpha_0 + |\boldsymbol{y}|/2 - 1}\exp\left\{-\alpha\left(\beta_0 + \left\|\boldsymbol{y} - \hat{\boldsymbol{y}}\right\|_2^2/2\right)\right\}\\
&\propto Gamma\left(\alpha_0 + |\boldsymbol{y}|/2, \beta_0 + \left\|\boldsymbol{y} - \hat{\boldsymbol{y}}\right\|_2^2/2\right)
\end{aligned}
$$

the mode of $\pi(\alpha|)$ is then accessed by $\left(\beta_0 + \left\|\boldsymbol{y} - \hat{\boldsymbol{y}}\right\|_2^2/2\right)^{-1}(\alpha_0 + |\boldsymbol{y}|/2 - 1)$.

Update of $\boldsymbol{\mu}$: Since the components in $\boldsymbol{\mu}$ are i.i.d., we only consider computing the full conditional posterior of μ_k $(k = 1, \cdots, p)$. $\pi(\mu_k|) = \pi(\boldsymbol{\theta}_{\cdot k}|\mu_k, b_k) \cdot \pi(\mu_k|0, b_{\mu_0})$. The full conditional posterior of μ_k can be obtained by:

$$\pi(\mu_k|) \propto \prod_{i=1}^{n}\exp\left\{-\frac{|\theta_{ik} - \mu_k|}{b_k}\right\} \cdot \exp\left\{-\frac{|\mu_k|}{b_{\mu_0}}\right\}$$

The problem of finding the optimal μ_k can be converted to finding the weighted median of the vector $\boldsymbol{\theta}_{\cdot k} \cup \{0\}$ with the weights $\{b_k\}_{i=1}^{p} \cup \{b_{\mu_0}\}$, which, as discussed in (Gurwitz 1990), can be solved in linear time.

Update of $\boldsymbol{b}$: Since the components in $\boldsymbol{b}$ are i.i.d., we only consider computing the full conditional posterior of b_k $(k = 1, \cdots, p)$. $\pi(b_k|) = \pi(\boldsymbol{\theta}_{\cdot k}|\mu_k, b_k) \cdot \pi(b_k|0, b_{b_0})$. The

Algorithm 1 LMLFM

Input: Training set $S = \{X, \boldsymbol{y}\}$, hyperpriors $\boldsymbol{\Theta}_0$, convergence criteria T (See Sec.).
Output: $\boldsymbol{\Theta}$.
Initialize the parameters $\{\boldsymbol{b}, \alpha, \boldsymbol{\mu}, \Theta\}$
repeat
  Update the parameters in the following order: $\left\{\Theta, \alpha, \boldsymbol{\mu}^{\mathcal{I}}, \boldsymbol{\mu}^{\mathcal{O}}, \boldsymbol{b}^{\mathcal{I}}, \boldsymbol{b}^{\mathcal{O}}\right\}$
until convergence

full conditional posterior of b_k can be obtained by:

$$\pi(b_k|) \propto \prod_{i=1}^{n} b_k^{-1} \exp\left\{-\frac{|\theta_{ik} - \mu_k|}{b_k}\right\} \cdot \exp\left\{-\frac{|b_k|}{b_{b_0}}\right\}$$

Assume $b_k > 0$, by setting $d \log \pi(b_k|)/db_k = 0$, we arrive at:

$$\frac{b_k^2}{b_{b_0}} + nb_k - \|\boldsymbol{\theta}_{\cdot k} - \mu_k\|_1 = 0$$

This is a simple quadratic function. Combining the fact that $b_k > 0$, the root of is found by:

$$b_k^* = 2b_{b_0}\left(\sqrt{n^2 + \frac{4}{b_{b_0}}\|\boldsymbol{\theta}_{\cdot k} - \mu_k\|_1} - n\right) \tag{5}$$

We can easily verify that $b_k > 0$ only when $\|\boldsymbol{\theta}_{\cdot k} - \mu_k\|_1 > 0$. For the case where $\|\boldsymbol{\theta}_{\cdot k} - \mu_k\|_1 = 0$, we have $\theta_{ik} = \mu_k$ for all $i = 1, \cdots, n$. This is equivalent to rejecting the randomness of $\boldsymbol{\theta}_{\cdot k}$; thus, if $\|\boldsymbol{\theta}_{\cdot k} - \mu_k\|_1 = 0$, we can draw two conclusions: i) variable k is subject to fixed effect instead of random effects; ii) $\boldsymbol{\theta}_{\cdot k} \sim Laplace(\mu_k, 0)$, thus we get $b_k = 0$.

The pseudo-code of LMLFM is given in Algorithm 1.

Complexity Analysis.

The computation of Θ can be accelerated by pre-computing and caching $\hat{\boldsymbol{y}}$. Then, $\hat{\boldsymbol{y}}$ can be maintained up-to-date using:

$$g(i) + \sum_{q \in 1:p \setminus k} \theta_{iq} \boldsymbol{h}_{iq} = \hat{\boldsymbol{y}}_i - \theta_{ik}^{(t)} \boldsymbol{h}_{ik}; \quad \hat{\boldsymbol{y}}_i = \hat{\boldsymbol{y}}_i + \theta_{ik}^{(t+1)} \boldsymbol{h}_{ik}$$

where $\theta_{ik}^{(t)}$ is the value of θ_{ik} for the t-th iteration. Thus, updating Θ takes $O(|S|)$ time. Next, computing α requires $\hat{\boldsymbol{y}}$, which is already cached; hence, complexity for updating α is $O(|\boldsymbol{y}|)$. The time complexity of updating $\boldsymbol{\mu}, \boldsymbol{b}$ is $O(p(n+m))$. Therefore, the overall computational complexity for one complete iteration is $O(|S|)$, which is strictly linear in the size of the training data. It is worth noting that the time complexity of FM is $O(k\,|S|)$ (Rendle 2012), where k denotes the number of latent factors. The space complexity of LMLFM is $O(|\boldsymbol{y}|)$, identical to that of FM (Rendle 2012). We conclude that our algorithm is more efficient than FM.

Theoretical Analysis

In this section, we will prove two important properties in LMLFM: *Ascent property* and *Convergence*. Let $\mathbb{Z}^+$ denotes the set of all positive integers. To keep the notation light,

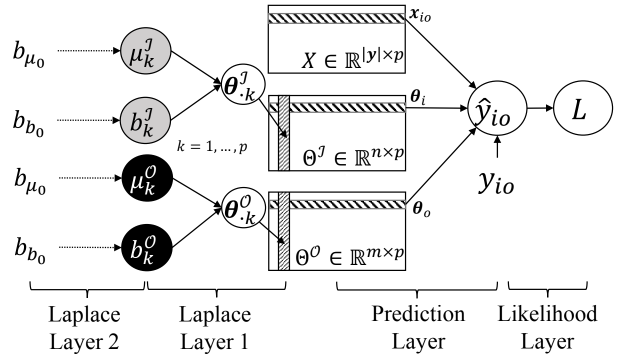

Figure 1: The hierarchical Bayesian structure of LMLFM. Laplace layer 1 is designed to select predictive latent factors; Laplace layer 2 is designed to identify random effects.

we overload the symbol θ_j to refer to the j-th component of the vectorized full parameter set $\boldsymbol{\Theta} = \{\theta_j\}_{j=1}^{d}$. The remaining set of parameters excluded θ_j is denoted by $\boldsymbol{\Theta}_{-j}$. We denote by $\pi(\theta_j|)$ the full conditional posterior of θ_j, i.e., $\pi(\theta_j|\boldsymbol{y}, X, \boldsymbol{\Theta}_0, \boldsymbol{\Theta}_{-j})$. Putting these together, the joint posterior density at the t-th iteration in (1) can be expended as:

$$\pi\left(\boldsymbol{\Theta}^{(t)}|\boldsymbol{y}, X, \boldsymbol{\Theta}_0\right) = \pi\left(\theta_j^{(t)}|\right) \cdot \pi\left(\boldsymbol{\Theta}_{-j}^{(t)}|\boldsymbol{y}, X, \boldsymbol{\Theta}_0\right) \tag{6}$$

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

Proof. Without loss of generality, we assume that $\boldsymbol{\Theta}$ is updated with the following order $\theta_1, \ldots, \theta_d$, where θ_j is updated by the mode of $\pi(\theta_j|), j = 1, \ldots, d$. Therefore, when $j = 1$, we must have $\pi\left(\theta_1^{(t+1)}|\right) \geq \pi\left(\theta_1^{(t)}|\right)$. Thus, followed by (6), we have $\pi\left(\theta_1^{(t+1)}, \theta_2^{(t)}, \ldots, \theta_d^{(t)}|\right) \geq \pi\left(\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_d^{(t)}|\right)$. Similarly, for $j = 2, \ldots, d$, $\theta_j^{(t+1)} = \arg\max_{\theta_j} \pi(\theta_j|)$ with other parameters fixed. Consequently, we have $\pi\left(\theta_1^{(t+1)}, \ldots, \theta_j^{(t+1)}, \theta_{j+1}^{(t)}, \ldots, \theta_d^{(t)}|\right) \geq \pi\left(\theta_1^{(t+1)}, \ldots, \theta_{j-1}^{(t+1)}, \theta_j^{(t)}, \theta_{j+1}^{(t)}, \ldots, \theta_d^{(t)}|\right)$. At $j = d$, we achieve $\pi\left(\boldsymbol{\Theta}^{(t+1)}|\right) \geq \pi\left(\boldsymbol{\Theta}^{(t)}|\right)$. $\square$

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 $\epsilon > 0$.*

Proof. According to Proposition 1, we only need to prove $\pi\left(\boldsymbol{\Theta}^{(t+i)}|\right) - \pi\left(\boldsymbol{\Theta}^{(t)}|\right) < \epsilon$. Let's assume that $\forall t \in \mathbb{Z}^+$, such that $\exists i \in \mathbb{Z}^+, \pi\left(\boldsymbol{\Theta}^{(t+i)}|\right) - \pi\left(\boldsymbol{\Theta}^{(t)}|\right) \geq \epsilon$. Then we can always find a monotonically increasing sequence $\{a_k\}_{k=0}^{\infty}$ with $a_0 = 0$, such that $\pi\left(\boldsymbol{\Theta}^{(t+a_k)}|\right) > \pi\left(\boldsymbol{\Theta}^{(t+a_{k-1})}|\right) > \cdots > \pi\left(\boldsymbol{\Theta}^{(t+a_0)}|\right)$. Let's further define ϵ_k as the lower bound of $\pi\left(\boldsymbol{\Theta}^{(t+a_k)}|\right) - \pi\left(\boldsymbol{\Theta}^{(t+a_{k-1})}|\right)$. Thus, $\sum_{j=1}^{k} \pi\left(\boldsymbol{\Theta}^{(t+a_j)}|\right) -$

$\pi\left(\boldsymbol{\Theta}^{(t+a_{j-1})}|\right) = \pi\left(\boldsymbol{\Theta}^{(t+a_k)}|\right) - \pi\left(\boldsymbol{\Theta}^{(t)}|\right) \geq \sum_{j=1}^{k}\epsilon_j$. Then we have $\pi\left(\boldsymbol{\Theta}^{(t+a_k)}|\right) \geq \pi\left(\boldsymbol{\Theta}^{(t)}|\right) + \sum_{j=1}^{k}\epsilon_j$. Notice that $\forall k, \epsilon_k > 0$, we then have $\lim_{k\to\infty}\pi\left(\boldsymbol{\Theta}^{(t+a_k)}|\right) \geq \pi\left(\boldsymbol{\Theta}^{(t)}|\right) + \sum_{j=1}^{\infty}\epsilon_j \geq \sum_{j=1}^{\infty}\epsilon_j \to \infty$. However, the density function $\pi\left(\boldsymbol{\Theta}|\right)$ is assumed bounded, thus leads to a contradiction. □

Proposition 1 establishes the convergence of the joint posterior density $\pi\left(\boldsymbol{\Theta}|\right)$. It also allows us to set the stopping criteria of our model: checking either the number of iterations exceeds a predefined threshold or the improvement of joint posterior density is below a preset threshold. However, because computing $\pi\left(\boldsymbol{\Theta}|\right)$ is intractable, we instead monitor the full joint density $\pi\left(\boldsymbol{y}, X, \boldsymbol{\Theta}, \boldsymbol{\Theta}_0\right)$.

Correlation Estimation

Longitudinal Correlation (LC) Let's denote o, j as two different observations. The LC is computed as follows:

$$\begin{aligned}
&cov\left(y_{io}, y_{ij}\right) \\
&= \mathbb{E}\left[\left(\boldsymbol{x}_{io}^{\intercal}\left(\boldsymbol{\theta}_i+\boldsymbol{\theta}_o\right)+\boldsymbol{\theta}_i^{\intercal}\boldsymbol{\theta}_o\right)\left(\boldsymbol{x}_{ij}^{\intercal}\left(\boldsymbol{\theta}_i+\boldsymbol{\theta}_j\right)+\boldsymbol{\theta}_i^{\intercal}\boldsymbol{\theta}_j\right)^{\intercal}\right] \\
&\;-\mathbb{E}\left[\left(\boldsymbol{x}_{io}^{\intercal}\left(\boldsymbol{\theta}_i+\boldsymbol{\theta}_o\right)+\boldsymbol{\theta}_i^{\intercal}\boldsymbol{\theta}_o\right)\right]\cdot\mathbb{E}\left[\left(\boldsymbol{x}_{ij}^{\intercal}\left(\boldsymbol{\theta}_i+\boldsymbol{\theta}_j\right)+\boldsymbol{\theta}_i^{\intercal}\boldsymbol{\theta}_j\right)\right] \\
&= 2\left(\boldsymbol{x}_{io}+\boldsymbol{\mu}^{\mathcal{O}}\right)^{\intercal}[\![\boldsymbol{b}^{\mathcal{I}}]\!]^2\left(\boldsymbol{x}_{ij}+\boldsymbol{\mu}^{\mathcal{O}}\right)
\end{aligned}$$

Cluster Correlation (CC) Let's denote i, v as two different individuals. The CC is computed as follows:

$$\begin{aligned}
&cov\left(y_{io}, y_{vo}\right) \\
&= \mathbb{E}\left[\left(\boldsymbol{x}_{io}^{\intercal}\left(\boldsymbol{\theta}_i+\boldsymbol{\theta}_o\right)+\boldsymbol{\theta}_i^{\intercal}\boldsymbol{\theta}_o\right)\left(\boldsymbol{x}_{vo}^{\intercal}\left(\boldsymbol{\theta}_v+\boldsymbol{\theta}_o\right)+\boldsymbol{\theta}_v^{\intercal}\boldsymbol{\theta}_o\right)^{\intercal}\right] \\
&\;-\mathbb{E}\left[\left(\boldsymbol{x}_{io}^{\intercal}\left(\boldsymbol{\theta}_i+\boldsymbol{\theta}_o\right)+\boldsymbol{\theta}_i^{\intercal}\boldsymbol{\theta}_o\right)\right]\cdot\mathbb{E}\left[\left(\boldsymbol{x}_{vo}^{\intercal}\left(\boldsymbol{\theta}_v+\boldsymbol{\theta}_o\right)+\boldsymbol{\theta}_v^{\intercal}\boldsymbol{\theta}_o\right)\right] \\
&= 2\left(\boldsymbol{x}_{io}+\boldsymbol{\mu}^{\mathcal{I}}\right)^{\intercal}[\![\boldsymbol{b}^{\mathcal{O}}]\!]^2\left(\boldsymbol{x}_{vo}+\boldsymbol{\mu}^{\mathcal{I}}\right)
\end{aligned}$$

Multi-level Correlation Finally, the multi-level correlation is computed as follows:

$$\begin{aligned}
&cov\left(y_{io}, y_{io}\right) \\
&= \mathbb{E}\left[\left(\boldsymbol{x}_{io}^{\intercal}\left(\boldsymbol{\theta}_i+\boldsymbol{\theta}_o\right)+\boldsymbol{\theta}_i^{\intercal}\boldsymbol{\theta}_o\right)\left(\boldsymbol{x}_{io}^{\intercal}\left(\boldsymbol{\theta}_i+\boldsymbol{\theta}_o\right)+\boldsymbol{\theta}_i^{\intercal}\boldsymbol{\theta}_o\right)^{\intercal}\right] \\
&\;-\mathbb{E}\left[\left(\boldsymbol{x}_{io}^{\intercal}\left(\boldsymbol{\theta}_i+\boldsymbol{\theta}_o\right)+\boldsymbol{\theta}_i^{\intercal}\boldsymbol{\theta}_o\right)\right]\cdot\mathbb{E}\left[\left(\boldsymbol{x}_{io}^{\intercal}\left(\boldsymbol{\theta}_i+\boldsymbol{\theta}_o\right)+\boldsymbol{\theta}_i^{\intercal}\boldsymbol{\theta}_o\right)\right] \\
&= 2\boldsymbol{x}_{io}^{\intercal}\left([\![\boldsymbol{b}^{\mathcal{I}}]\!]^2+[\![\boldsymbol{b}^{\mathcal{O}}]\!]^2\right)\boldsymbol{x}_{io}+4\cdot\mathrm{tr}\left([\![\boldsymbol{b}^{\mathcal{I}}]\!]^2\cdot[\![\boldsymbol{b}^{\mathcal{O}}]\!]^2\right)
\end{aligned}$$

where $\mathrm{tr}\left(A\right)$ is the trace of matrix A.

Parametrization

Initialization. Previous study has pointed out that ICM-based techniques are particularly sensitive to the choice of initializations (Szeliski et al. 2008). We find in our experiments that LMLFM is rather robust to the choice of parameter initialization. The set of parameters that need initialization are $\{\boldsymbol{b}, \alpha, \boldsymbol{\mu}, \Theta\}$. We simply let $\boldsymbol{b} = \mathbf{1}, \boldsymbol{\mu} = \mathbf{0}$ and $\Theta = \mathbf{0}$. The precision α tends to have stronger impact on the model performance. Our experiments show that setting α with the form $\tau\left[\mathrm{var}\left(\boldsymbol{y}\right)\right]^{-1}$ (where τ is a hyper-parameter controlling the value of precision) provides satisfactory results.[1]

Hyperprior settings. Due to the high number of explanatory variables, the impact of α_0, β_0 are negligible; therefore, we chose trivial values $\alpha_0 = \beta_0$. We also set $b_{\mu_0} = b_{b_0} = 1$. We claim convergence if the change of normalized log posterior density value of two consecutive iterations is less than 10^{-4} or the number of iterations exceed 10.

Experimental Protocol

The observations in the generated data are assigned to disjoint training and test sets while simultaneously ensuring that for any given individual, no observation that is included in the training data has a time stamp that is later than any observation that is included in the test data. This ensures that while making predictions no information from the future is used. Specifically, we use the following procedure: i) Randomly split the complete data set into two subsets, i.e., $S^{(a)}$ (70%) and $S^{(b)}$ (30%); ii) For each individual $i \in S^{(a)}$, we find out the observation with the latest time stamp and denote it as t_i; iii) For each individual $i \in S^{(b)}$, we split the observations associated to i into two half. Specifically, let $S_i^{(b_1)}$ denotes the subset of observations with time stamp less than t_i and $S_i^{(b_2)}$ denoting the rest; iv) The training and test data is given by $S^{(a)} \cup S^{(b_1)}$ and $S^{(b_2)}$ respectively.

Most of the implementations of our baselines are publicly available. we use the LASSO code from (Pedregosa et al. 2011). For M-LMM, LMMLASSO, GLMMLASSO and rPQL, we utilize the `lmer4`, `lmmlasso`, `glmmLasso` and `rpql` package, respectively from CRAN.[2] We implemented MLLASSO and LMLFM in Python. We use the implementation of Random forest in (Pedregosa et al. 2011)), FM in (Mikhail Trofimov 2016) and the Penalized GEE (PGEE) in the `PGEE` package in CRAN. All experiments are conducted on a desktop computer with i7-7700K CPU, 16G RAM and GeForce GTX 1060 graphics card. The hyperparameters of all methods are tuned to optimize their performance using 5-fold cross validation. We report performance statistics obtained from 100 independent runs. Evaluation scores are computed using only the test data.

Simulated data. The outcomes $\boldsymbol{y}$ are drawn independently from $N\left(y_{io}|\boldsymbol{x}_{io}^{\intercal}\left(\boldsymbol{\beta}+\boldsymbol{\theta}_i+\boldsymbol{\theta}_o\right)+\boldsymbol{\theta}_i^{\intercal}\boldsymbol{\theta}_o, 1\right)$, with $\boldsymbol{\beta} = \{1, 2, 3, -1, -2, -3, 7, 10, 0, \cdots, 0\}$. Each component $\theta_{zk} \in \Theta$ follows $Laplace(\theta_{zk}|0, b_k)$, where $b_k^{\mathcal{I}}$ for $k = 1, \cdots, 10$ and $b_k^{\mathcal{O}}$ for $k = 5, \cdots, 15$ are drawn from $U(0, 1)$. For other k, we set $b_k^{\mathcal{I}} = b_k^{\mathcal{O}} = 0$. Thus there are 15 and $p - 15$ relevant features with random effects and fixed effects respectively. Absence of LC or CC is simulated by manually setting $\left\{\Theta^{\mathcal{O}}, \boldsymbol{b}^{\mathcal{O}}\right\}$ or $\left\{\Theta^{\mathcal{I}}, \boldsymbol{b}^{\mathcal{I}}\right\}$ to zero, respectively.

References

Bertsekas, D. P. 1999. *Nonlinear programming*. Athena scientific Belmont.

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

[1] τ is the only tuning hyper-parameter in LMLFM.

[2] https://cran.r-project.org/

Mikhail Trofimov, A. N. 2016. tffm: Tensorflow implementation of an arbitrary order factorization machine. https://github.com/geffy/tffm.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research* 12(Oct):2825–2830.

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

Szeliski, R.; Zabih, R.; Scharstein, D.; Veksler, O.; Kolmogorov, V.; Agarwala, A.; Tappen, M.; and Rother, C. 2008. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE transactions on pattern analysis and machine intelligence* 30(6):1068–1080.

Depression Analysis on SWAN Data					
Rank	LMLFM	LASSO	FM	RF	MLLASSO
1	NervousPast2Wks_a lot (4.91)	HospitalStayPastYear_yes (7.33)	OtherEventUpset_not upset (-0.05)	NervousPast2Wks_a lot (0.13)	Race_white (10.27)
2	HospitalStayPastYear_yes (2.89)	HospitalStayPastYear_no (7.01)	STATUS_hysterectomy/both ovaries removed (-0.05)	MoodChangePast2Wks_never (0.08)	Race_black american (9.27)
3	HospitalStayPastYear_no (2.82)	Stroke_no (6.9)	NervousPast2Wks_a lot (0.05)	Stroke_no (0.04)	HadD& C_yes *(7.04)
4	MoodChangePast2Wks_never (-2.36)	Stroke_yes (5.66)	Race_hispanic (0.05)	Height (0.04)	ViolentEventUpset_no (6.71)
5	Income_very low (2.27)	NervousPast2Wks_a lot (3.88)	Income_very low (0.05)	DHAS*(0.03)	QuitJobUpset_no (6.7)
6	Race_hispanic (2.26)	MoodChangePast2Wks_a lot (3.2)	STATUS_unknown due to hysterectomy (-0.05)	Testosterone (0.03)	QuitJobUpset_not upset (6.68)
7	Stroke_no (1.77)	NervousPast2Wks_never (-2.72)	MoodChangePast2Wks_a lot (0.04)	SHBG*(0.03)	PartnerUnempUpset_no (6.63)
8	TalkLastYr_a lot (1.69)	Height (-2.29)	STATUS_pregnant/breastfeeding (0.04)	FSH*(0.03)	HipSize (6.51)
9	TalkLastYr_few (1.32)	Income_very low (2.26)	HotFlashesPast2Wks_a lot (-0.04)	Estradiol (0.03)	EndedRelationUpset_no (6.46)
10	WorsenRelationUpset_yes (1.25)	Migraines_yes (2.16)	STATUS_post-menopausal (-0.04)	HipBoneDensity (0.03)	HadD& C_no*(6.36)

General Happiness Analysis on GSS Data					
Rank	LMLFM	LASSO	FM	RF	MLLASSO
1	MarriageHappiness_no (-0.53)	Financial_satisfied (0.21)	FirstLanuage_Tagalog (-1.53)	WomansHealthEndangered_yes (0.08)	InterviewsLanguage_English (2.47)
2	LifeExciting_dull (-0.45)	LifeExciting_dull (-0.2)	LanguageAtHome_Vietnamese (-0.99)	HaveLifeAfterDeath_yes (0.07)	MemberIn.NationalityGp_no (2.38)
3	JobOrHousework_very dissatisfied (-0.21)	Financial_more or less satisfied (0.19)	FirstLanuage_KOREAN (0.97)	CourtsDealWithCriminals_not harse enough (0.06)	MemberIn.SchoolFraternity_no (2.37)
4	Health_poor (-0.15)	Oppo.RaceInNeighborhood_no (0.17)	2ndCountrySpouseOrigin_WestIndies (-0.85)	LifeExciting_dull (0.03)	MemberIn.PoliticalClub_no (2.37)
5	MarriageHappiness_very happy (0.11)	LifeExciting_exciting (0.17)	FirstLanuage_Hindu (-0.84)	Financial_not at all satisfied (0.03)	MemberIn.FarmOrg_no (2.37)
6	Unemployed (-0.1)	MarriageHappiness_very happy (0.17)	VoteIn1972Election_dont known (-0.83)	EverUnemplInLast10Yrs_no (0.03)	MemberIn.ArtGp_no (2.35)
7	MarriageHappiness_pretty happy (0.09)	Oppo.RaceInNeighborhood_yes (0.16)	3rdCountrySpouseOrigin_Philippines (0.83)	SeenX-RatedMovieLastYr_no (0.02)	MemberIn.ServiceGp_no (2.34)
8	JobOrHousework_dissatisfied (-0.08)	FinancialChange_better (0.16)	FirstLanuage_Norwegian (-0.79)	MarriageHappiness_very happy (0.02)	MemberIn.YouthGp_no (2.33)
9	Financial_not satisfied (-0.08)	Financial_stayed same (0.13)	3rdCountryYouOrigin_WestIndies (0.78)	Oppo.RaceInNeighborhood_yes (0.01)	MemberIn.FraternalGp_no (2.33)
10	SocialClass_lower class (-0.08)	PovertyStatus_unknown (-0.11)	2ndCountrySpouseOrigin_arabic (0.72)	PovertyStatus_unknown (0.01)	MemberIn.VeteranGp_no (2.33)

* D& C: Dilation and curettage; DHAS: Dehydroepiandrosterone sulfate; FSH: Follicle-stimulating hormone; SHBG: Sex hormone-binding globulin.

Table 1: Top 10 most relevant variable selected by different models on the real-life data sets. The feature importance (or PAE) is specified in the parenthesis.