Machine Learning for Disease Progression

Yong Deng
Department of Materials Science & Engineering
yongdeng@stanford.edu

Xuxin Huang
Department of Applied Physics
xxhuang@stanford.edu

Guanyang Wang
Department of Mathematics
guanyang@stanford.edu

1 Introduction

Disease progression in individual patients is one of the fundamental questions in medical practice. Since many medical tests are either harmful or inconvenient to perform frequently, it would be beneficial to develop a disease progression prediction method based on machine learning approaches. In this project, we focus on the study of the progression of motor impairment in children with Cerebral Palsy. In particular, Gait Deviation Index (GDI)[5] is collected over time for each patient and used to quantitatively characterize the development of gait impairments. Due to the sparsity and irregularity of the data in time, we would apply regression methods with rank-constraints relying on matrix completion to analyze the data set, as proposed in Ref. [3]. Specifically, our main input data is a matrix Y, each row of which is GDI of a patient measured over time. It is of our interest to find a coefficient matrix W so that Y can be described by WB, where B is a matrix of time dependent basis. Details can be found in section 2, section 3 and Ref. [3]. Our contributions in this project include:

- We have cleaned and preprocessed the 'Gillette Children's Specialty Healthcare dataset', which includes around 6000 observations of children visiting a gait clinic, and merged the dataset with the SEMLS dataset, which contains information of surgeries of each patient.
- We have implemented Soft-Longitudinal-Impute (SLI), Sparse-Longitudinal-Regression (SLR), functional principal components (fPCA) methods described by [3] on the dataset, which explains around 30% of the variance.
- We have studied the effect of a surgery for each patient. It turns out that after taking the surgery information into account, the model could perform significantly better and explains around 40% of the variance, which performs better than the current state-of-the-art approaches.

2 Dataset and Features

The original data contains records of 12078 exams on 2904 patients, mostly between age 3-18. Each exam record consists basic information (e.g. walking speed, cadence, bmi, height, weight, maximum knee flexion, O2 expenditure, and Gait Deviation Index (GDI) [5]) of a patient as well as about 300 clinical variables of a particular leg. In particular, GDI is a measurement of severity of pathology. It's a real number normally between 50 - 110, where normally developed children have GDI around 100. Accurate estimation of post-treatment GDI can lead to better clinical decisions. In this report, it is of our main interest to predict the GDI trajectories of the patients.

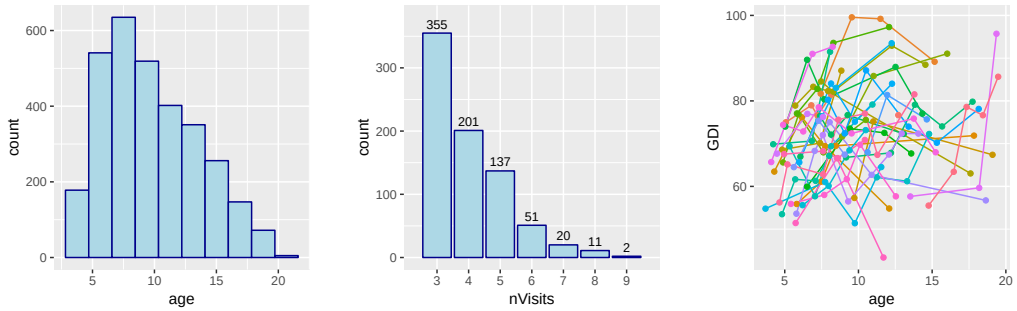
For robustness of the result, the following pre-processing procedures were performed:

1. To avoid co-linearity between the two legs of patients, we consider only exams on left legs.

2. We consider only records with valid GDI, age between 3 and 20, BMI (body mass index) between 10 and 30 and throw away the outliers.
3. To have enough data points to fit the individual progression curve, we consider only subset of patients with 3 or more remaining records after previous steps.

The remaining data set contains 3106 exams on 777 patients. Some summaries of the data set are shown below. As can be seen in Figure 1, the measurements are collected sparsely and irregularly in time.

Figure 1: Data set summaries



The three plots from left to right are: (a). histogram of age among all exams; (b). histogram of number of visits per patient; (c). Gait Deviation Index (GDI) from a subset of about 40 patients, individual patients are differing by color.

3 Models and Methods

3.1 Problem Formulation and Related Work

The general question is stated as following. Let n be the number of patients. For each patient $i \in \{1, ...n\}$, we have n_i observations $\{y_{i,1}, ..., y_{i,n_i}\}$ at time-points $\{t_{i,1}, t_{i,2}, ..., t_{i,n_i}\}$ where $0 < t_{\min} \leq t_{i,1} < t_{i,2} < ... < t_{i,n_i} \leq t_{\max}$. Let $\mathbf{b} \equiv \{b_i : i \in \{1, ..., K\}\}$ be a set of basis for $L_2([t_{\min}, t_{\max}])$ truncated to the first K elements. We would like to estimate a set of coefficients $\mathbf{w}_i \in \mathbb{R}^K$ so that $y_{i,j}$ can be approximated by $\mathbf{w}_i^T \mathbf{b}(t_{i,j})$. The state-of-art approaches to estimating $\mathbf{w}_i$ include direct approaches with functional principal component analysis [4, 7], linear mixed-effect models [6, 9] with low-rank approximations [1, 8].

Direct approach has two major drawbacks to modeling covariance. First, overfitting happens when the number of observations n_i for individual i is smaller or equal to the size of the basis k. Second, similarities between curves are ignored while they could improve the fit potentially. Linear mixed-effect models can solve this problem conveniently by estimating the covariance structure and the individual fit simultaneously. However, they are not applicable unless the number of observations per subject is relatively large, for we attempt to estimate K coefficients for every subject. Given the small number of observations, we could still fit a linear mixed-effect model in a smaller space spanned by functions with largest contribution to the random effect. Based on this, low-rank approximations are widely applied. However, due to their reliance on the distribution assumptions, these models need to be carefully fine-tuned for specific situations.

To avoid the potential bias caused by the assumption of an underlying probabilistic distribution in mixed-effect models, Ref.[3] approximates the optimization problem in the sparse matrix completion setting. Observed measurements are denoted as $\tilde{y}_{i,j}$. The time grid is discretized to T time-points $G = [\tau_1, ..., \tau_T]$. We assign $y_{i,g_i(j)} = \tilde{y}_{i,j}$ for $g_i(j) = \arg\min_{1 \leq k \leq T} |\tau_k - t_{i,j}|$. Then we can construct a $N \times T$ matrix Y of observed values for N patients. Notice that the matrix Y is a sparse matrix, as each patients only have a few measurements at different times. We considers all the unobserved entries in matrix Y as missing values, and our target is to impute all the unknown elements.

Denote the set of all observed elements by pairs of indices as Ω. Let $P_\Omega(Y)$ be the projection onto observed indices: $P_\Omega(Y) = M$, such that $M_{i,j} = Y_{i,j}$ for $(i, j) \in \Omega$ and $M_{i,j} = 0$ otherwise.

$P_\Omega^\perp(Y)$ is defined as the projection on the complement of Ω: $P_\Omega^\perp(Y) = Y - P_\Omega(Y)$. The basis now is a $T \times K$ matrix $B = [\tau_1, ..., \tau_T]^T$. The coefficients we would like to fit is denoted by a $N \times K$ matrix W. Therefore, the problem can be formulated as a matrix completion problem, heuristically speaking, the target is to find a matrix W, such that $WB^T \approx Y$ on our observed indices Ω, and thus we could impute the missing values of Y by WB^T. The details of the previous heuristics can be found in the next part of this section.

3.2 Models without adjusting for surgery

As our model can be described as:

$$Y \approx WB^T. \tag{1}$$

The direct way of finding such a W is to solve the following optimization problem:

$$\arg\min_W \|P_\Omega(Y - WB^T)\|_F^2 \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm, i.e. the square root of the sum of matrix elements. However, such approach consists two main drawbacks. First, if the number of basis functions K is larger than or equal to the number of observations n_i, which is the number of observations of individual i, then the error could be reduced to 0, which causes overfitting. Second, this methods ignores the similarities between the curves of different individuals, which could potentially improve the model performance.

One of the standard ways to remedy these issues it to assume that individual trajectories can be represented in a low-dimensional space by constraining the rank of W. Thus our optimization problem is now:

$$\arg\min_W \|P_\Omega(Y - WB^T)\|_F^2 + \lambda\|W\|_*, \tag{3}$$

where $\lambda > 0$ is a parameter, $\|\cdot\|_F$ is the Frobenius norm, and $\|\cdot\|_*$ is the nuclear norm, i.e. the sum of singular values. Ref.[2] shows that the optimization problem $\arg\min_W \frac{1}{2}\|Y - WB^T\|_F^2 + \lambda\|W\|_*$ has a unique solution $S_\lambda(YB)$, where $S_\lambda(X) = UD_\lambda V^T$ and $X = UDV^T$ is the singular value decomposition (SVD) of X. $D_\lambda = \mathrm{diag}((d_1 - \lambda)_+, (d_2 - \lambda)_+, ..., (d_p - \lambda)_+)$, where $(x)_+ = \max(x, 0)$, is soft-thresholding of a diagonal matrix $D = \mathrm{diag}(d_1, d_2, ..., d_p)$. We refer to $S_\lambda(X)$ as the singular value thresholding (SVT) of X. Inspired by this, algorithm 1 in Ref.[3] is proposed to solve the optimization problem Eq. 3 on a sparsely observed data set by iteratively imputing the missing elements in Y with SVT of $P_\Omega(Y)B$ obtained in the previous step.

The optimization problem above can be easily extended to multiple variables varying or constant in time that work together to characterize the progression of one disease:

$$\arg\min_W \|P_\Omega(\mathbf{X} - W\mathbf{B}^T)\|_F^2 + \lambda\|W\|_*. \tag{4}$$

X_i is some $N \times T$ matrices corresponding to the processes measured and $\mathbf{X} = (X_1 : X_2 : ... : X_p)$. $\mathbf{B} = I_p \otimes B$ is a $pT \times pK$ matrix with B stacked p times on the diagonal. W is a $N \times pK$ coefficient matrix that we want to fit. This optimization problem can also be solved with algorithm 1 in Ref.[3].

The above two optimization problems both aim to reduce the dimensionality of the sparse observations. In practice we would often want to predict the trajectory of one variable Y (GDI in our case) with the knowledge of other variables $\mathbf{X}$ related to the same disease. Then the problem can be formulated as a regression of Y on $\mathbf{X}$:

$$\arg\min_{\mathbf{A}} \|P_\Omega(Y - \mathbf{X}\mathbf{A}\mathbf{B}^T)\|_F^2 + \lambda\|\mathbf{A}\|_*. \tag{5}$$

Algorithm 3 in Ref.[3] is proposed to solve this sparse-regression problem.

Finally, combining the technique of dimensionality reduction and sparse regression we can predict the trajectory of one variable, given some other covariates that are varying or constant in time. We first solve Eq.(2) using algorithm 1 in Ref.[3] to reduce the dimension of covariates $\mathbf{X}$. The resulting coefficient matrix is given by W. Then we decompose W as $W = USV^T$ to retrieve the latent components U. Finally we regress Y on U solve the regression problem with algorithm 3 in Ref.[3]:

$$\arg\min_{\mathbf{A}} \|P_\Omega(Y - U\mathbf{A}\mathbf{B}^T)\|_F^2. + \lambda\|\mathbf{A}\|_*. \tag{6}$$

Some preliminary simulations and data studies are presented in Ref.[3].

Table 1: SLI and fPCA applied to the original data.

Method	MSE	sd
SLI	81.54	10.12
fPCA	84.87	14.58
baseline	119.75	10.00

Table 2: SLR, SLI and fPCA applied to the data with surgery information.

Method	MSE	sd
SLR	72.19	10.32
SLI	73.61	10.68
fPCA	70.28	10.27
baseline	119.75	10.00

3.3 Models after adjusting for surgery

Intuitively, having a surgery would impact the progression of disease, thus it is reasonable to build up a model which take surgery into account. For each patient i, our new model can be formulated as

$$y_i(t) = \sum_{j=1}^{k} w_{i,j} b_j(t) + \mu \cdot \mathbb{1}_{i,S}(t) + \epsilon_i, \tag{7}$$

where the indicator function $\mathbb{1}_S(t)$ equals 1 if patient i has received a surgery before time t and 0 otherwise, μ can be interpreted as the average effect of a surgery among all patients, and ϵ_i is a random effect which can be modeled as a normal distribution with mean 0 and variance σ^2.

In our case, we could first regress Y on the dummy feature $\mathbb{1}_S$ to get an estimation for the mean effect $\hat{\mu}$, after adjusting for the effect of surgery, we could get a new matrix $\tilde{Y}$ where each rows represents the 'adjusted GDI' for patient i. Then we could try to impute the missing values of $\tilde{Y}$ based on all the methodologies described in part 3.2.

3.4 Implementation

The code used can be found in `https://drive.google.com/file/d/1gkyf1IAwJICRcVAkKsn_FBI8qnfg1ed0/view?usp=sharing`.

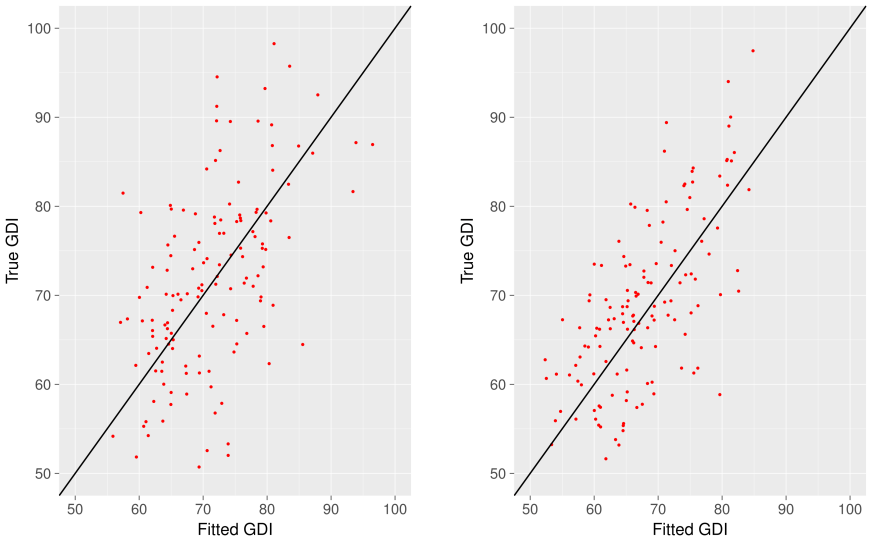
4 Results and Discussion

We first apply SLI to the GDI data. As a comparison, fPCA is also used to train the same dataset and generate predictions. The baseline is just a naive prediction calculated by averaging all GDI data we have. To estimate the accuracy and robustness, we have performed the experiment 20 times, at each time the data is randomly split into a training set (containing 90% of the data) and a test set (containing the rest 10% of the data), all models are trained on the training set and evaluated on the test set. The averaged mean square error (MSE) and the standard deviation (sd) of the MSE on the test set are summarized in Table. 1.

We can see that SLI is better than fPCA in this case and can explain 30% of variance from the baseline prediction, defined as the percentage of reduction of MSE from SLI with respect to MSE from baseline. To further improve the prediction we include the surgery information, which is expected to be highly correlated with GDI, into consideration. We then perform SLR, SLI and fPCA on the adjusted data $\tilde{Y}$, obtained as described in section 3.3, instead of Y and compare the results with the baseline, as shown in Table. 2

The inclusion of the surgery data significantly improves the performance of SLI, which can now explain 40% of variance from the baseline prediction. MSE from SLR is also smaller than MSE from SLI. This is not surprising since SLR makes use of more disease related information. However, both SLR and SLI could not outperform fPCA in this case. Plots of fitted GDI from SLI versus true GDI

Figure 2: Results



Fitted GDI vs true GDI from SLI on original data (left) and data with surgery information (right)

can be found in Fig. 2. It is convincing to conclude that, adjusting the effect of a surgery improves the performance of predictions significantly.

On the other hand, there is still a larger portion of variance remains unexplained, this may be due to the sparse and irregularity of the data, also it is important to note that in practice, it is often hard to predict the disease progression precisely based on the current features of an individual. But still, there is room for improvement, for example we could do variable selections to find the most relevant features or build more sophisticated models to capture the effect of a surgery, some of the directions of future works are discussed in Section 5.

5 Conclusion and Future Work

The SLI results using the original data can explain 30% of the error of the baseline. After we include the effect of surgery, the predictions of both SLI and SLR are improved and can explain up to 40% of the error of the baseline. However even after we consider the effect of surgery, the performances of SLI and SLR are still not as good as fPCA.

We can perform feature selection to further improve our matrix-completion based methods. This can be done by forward selection or by using the top components from the dimension reduction of covariates as new features.

Acknowledgments

This project is in corporation with postdoctoral researchers Dr. Łukasz Kidziński and Dr. Yumeng Zhang from the department of statistics in Stanford. The GDI dataset is provided by Dr. Łukasz Kidziński. The code we used is based on fcomplete (https://github.com/kidzik/fcomplete.git), a R package written by Dr. Kidziński. Dr. Zhang contributed helpful discussions and the code for data pre-processing.

References

[1] CS Berkey and RL Kent. Longitudinal principal components and non-linear regression models of early childhood growth. *Annals of human biology*, 10(6):523–536, 1983.

[2] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):2010, 1956.

[3] Łukasz Kidziński and Trevor Hastie. Longitudinal data analysis using matrix completion. *arXiv preprint arXiv:1809.08771*, 2018.

[4] DD Kosambi. Statistics in function space. In *DD Kosambi*, pages 115–123. Springer, 2016.

[5] Michael H Schwartz and Adam Rozumalski. The gait deviation index: a new comprehensive index of gait pathology. *Gait & posture*, 28(3):351–357, 2008.

[6] Geert Verbeke. Linear mixed models for longitudinal data. In *Linear mixed models in practice*, pages 63–153. Springer, 1997.

[7] Satosi Watanabe. Karhunen-loeve expansion and factor analysis: theoretical remarks and application. In *Trans. on 4th Prague Conf. Information Theory, Statistic Decision Functions, and Random Processes Prague*, pages 635–660, 1965.

[8] Fangrong Yan, Xiao Lin, Xuelin Huang, et al. Dynamic prediction of disease progression for leukemia patients by functional principal component analysis of longitudinal expression levels of an oncogene. *The Annals of Applied Statistics*, 11(3):1649–1670, 2017.

[9] Scott L Zeger, Kung-Yee Liang, and Paul S Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.