Machine Learning for Disease Progression

Yong Deng, Xuxin Huang and Guanyang Wang
{yongdeng, xxhuang, guanyang}@stanford.edu

I. Introduction

Analyzing the disease progression in individual patients is one of the fundamental questions in medical practice. In this project, we focus on the study of the progression of motor impairment in children with Cerebral Palsy by analyzing the Gait Deviation Index (GDI)[1], a quantitative characterization of gait impairments, collected over time for each patient.

Due to the sparsity and irregularity of the data in time, we have applied regression methods with rank-constraints relying on matrix completion to the dataset[2] and have successfully explained 40% of the error compared with the baseline.

II. Dataset Summary

The GDI dataset used in this project is provided by Dr. Łukasz Kidziński from the Department of Statistics at Stanford.

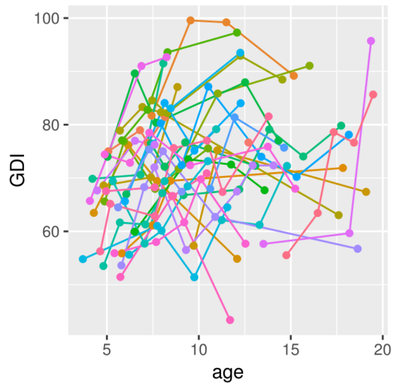

Fig. 1: Sample data from a subset of about 40 patients.

III. Model and Method

- **Direct approach:**

Y represents the data matrix. Each row is the GDI of a patient collected over time.
B is a time dependent spline basis.
P_Ω is a projection on the observed indices Ω.
The goal is to fit the coefficient matrix W.

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

We implemented Soft-Longitudinal-Impute (SLI), a singular value thresholding (SVT) based algorithm[2] to solve for W.
The regularization term is added to restrict the rank of W.

- **Regression:**

Extend to the case of multiple variables:

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

We formulate our trajectory prediction problem as a regression of Y on **X**:

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

This is solved by Sparse-Regression algorithm[2]

- **Dimensionality Reduction + Regression:**

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

U is the latent component retrieved from $W=USV^T$ obtained from dimensionality reduction of **X**.
A can be solved by Sparse-Longitudinal-Regression (SLR) algorithm[2]

IV. Results and Discussion

SLI and functional principal component analysis (fPCA) are applied to the original data. The results are compared with baseline (column mean of Y) in Table. 1. To improve the prediction, we modify the data using surgery information and compare results from different algorithms in Table. 2.

Table. 1: Original data.

	MSE	sd
SLI	81.54	10.12
fPCA	84.87	14.58
baseline	119.75	10.00

Table. 2: Data with surgery information.

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

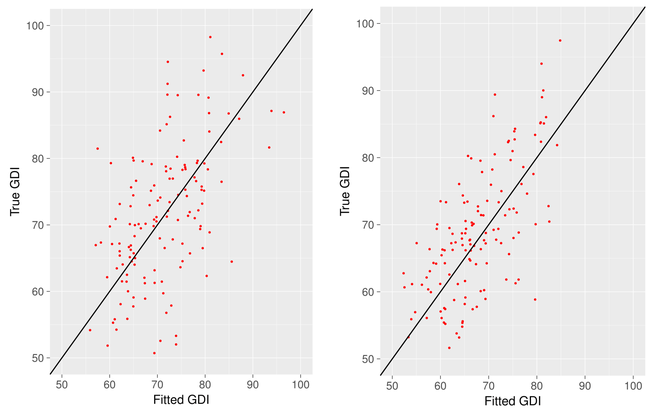

Fig. 2: Fitted GDI vs true GDI from SLI on original data (left) and data with surgery information (right)

Discussion: 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.

References & Acknowledgement

[1]Schwartz, Michael H., and Adam Rozumalski. "The Gait Deviation Index: a new comprehensive index of gait pathology." *Gait & posture* 28.3 (2008): 351-357.

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

[3]James, Gareth M., Trevor J. Hastie, and Catherine A. Sugar. "Principal component models for sparse functional data." *Biometrika* 87.3 (2000): 587-602.

This project is in corporation with postdoctoral researchers Dr. Łukasz Kidziński and Dr. Yumeng Zhang from the Department of Statistics at Stanford.