Amyotrophic lateral sclerosis functional data analysis

Minjun Park[1], Yusha Liu[1], Meng Li[1]

***Abstract*— Amyotrophic lateral sclerosis (ALS) is a motor neuron disease that gradually weakens muscle functionality. Unfortunately, there is no cure for the disease; it is imperative to treat the patients accordingly to the disease stage. The goal of this research to find a generalizable model that summarizes functional rating scores reported by patients over 12 months period. We registered the data by aligning the longitudinal curves across the length of symptom duration. After curve registration, we grouped patients into subgroups based on each symptom duration. We found that patients with longer symptom duration report lower functional rating scores, and within subgroups, we applied clustering method and identified the trends for the disease progression.**

I. Background and Motivation

ALS is an incurable, progressive nervous system disease that reduces functionality in the muscles. However, the cause for this loss of muscle control remains unexplained. This longitudinal study aims to infer critical stages for the current and new patients. State-of-the-art medicine can only relieve muscle cramps and delay the progression, so it is extremely important to infer how ALS will most likely to behave in the next few months.

II. Methods

A. Dataset

Biotechnology company Biogen requested an analysis on dataset – the collection of demographics, such as patient identification, gender, location, and functional scores rated by each patient from the scale of 0 to 4 on different muscle functionalities.

B. Challenges and Limitations

The challenge of the study is that patients enter the trial at different stages of the disease, making it difficult to find the generalizable model. In addition, visualization of 982 patients' longitudinal functional scores is very noisy without alignment or clustering.

C. Solutions

To resolve this issue, we used curve registration and clustering of patient scores by symptom duration. Curve registration aligns all curves to the developmental stage of the disease. However, other challenges arose from this alignment; most patients have unique symptom duration, and aligned curves become left-censored that require imputation and interpolation for the missing domain. To resolve this issue, we applied R clustering package kml to cluster longitudinal data, which is an imputation method for trajectories and chooses the best cluster. With homogeneous patient samples and identically and independently distributed symptom duration, we clustered the dataset with dropouts.

[1]Rice University Department of Statistics, Houston TX

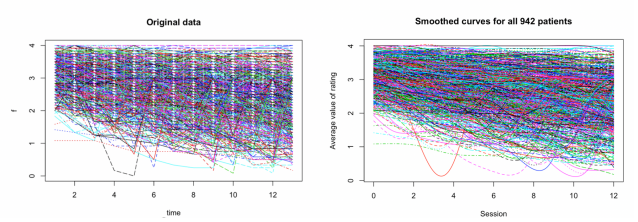

Fig. 1: Before and after smoothing 942 patients using functional data analysis package

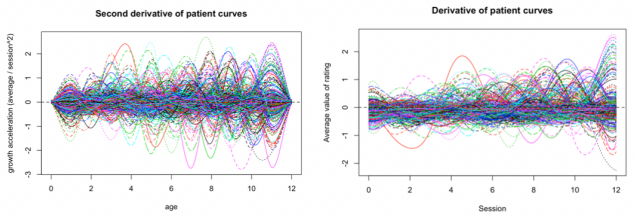

Fig. 2: First and Second derivatives of smoothing curve

III. Proposed solution and major contributions

A. Smoothing curve

We applied *smooth.basisPar* function in R package *fda* to map discrete observations into continuous space. The comparison between before smoothing and after smoothing is shown in fig. 1. As shown in fig. 2, Because all curves are continuous, we can apply first and second derivatives, and the result shows that many curves are stretching horizontally.

B. Curve registration

After curve smoothing, we applied curve registration to find the curve that best explains all curves. From the original data, curve registration method adopts warping functions (fig. 3, and we visualized the mean of all curves with the interval of one standard deviation with respect to each time point(fig. 4).

Registration of all curves is shown in fig. 5. Because we are aligning 942 curves all at same time, registration method forces all curves to into one cluster, leading to very noisy and yielding low interpretability. In registration, we also divided into various domains (fig. 6), but selections are more counter-intuitive

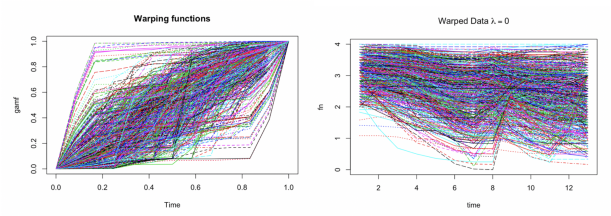

Fig. 3: Warping functions used for curve registration

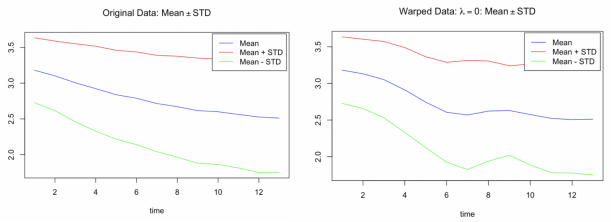

Fig. 4: mean and standard deviation graphs for all curves

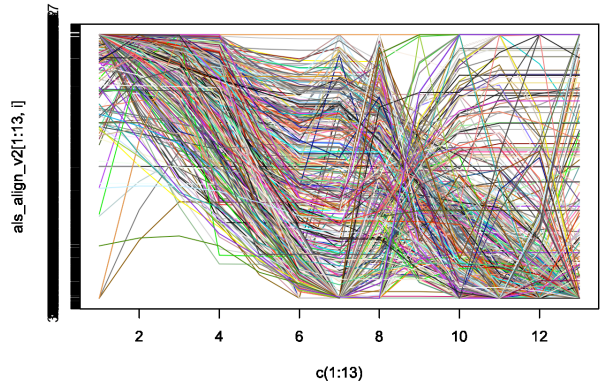

Fig. 5: Visualization of curve alignment

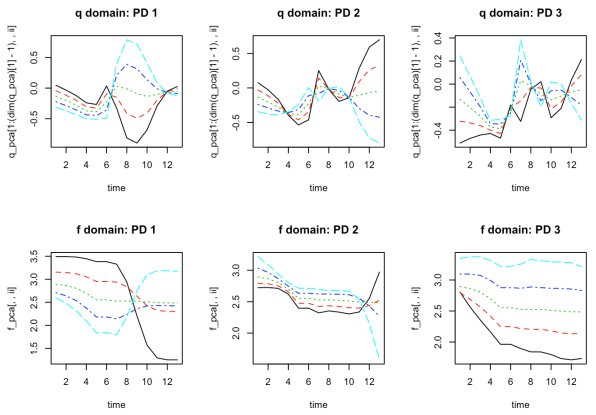

Fig. 6: visualization of registration into 6 different domains

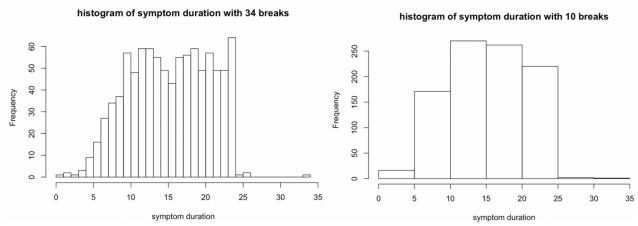

Fig. 7: visualization of registration into 6 different domains

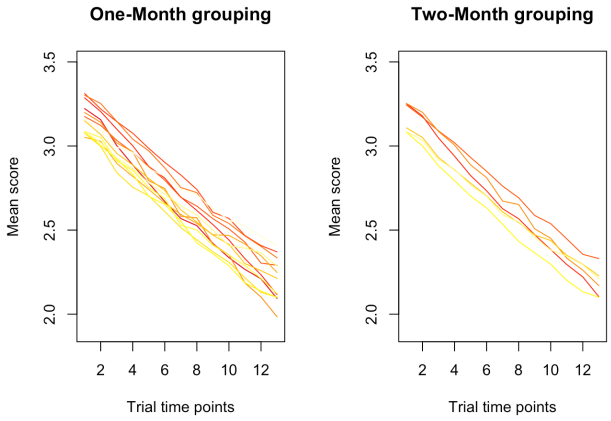

Fig. 8: Heatmap of average scores on each subgroups

C. *Clustering within each subgroups*

1) Introducing new variable: Symptom Duration: The biggest limitation of curve registration and smoothing is the low interpretability. Because all 942 patients enter the trail at different stage of the disease, finding the best explaining model remains difficult. Fortunately, the dataset has the variable called SYMPDUR that records the symptom duration(in months) before the patient has entered the trial. As shown in fig. 7, patients' symptom durations are uniformly distributed, and we used this variable to partition and register the data.

The heatmap in fig. 8 shows disparity in average ALSFRS score among groups with different symptom duration. For example, red and yellow lines are separable in the early time point, and two groups start to blend as the independent variable increases. This result encouraged intuitive search into clustering literature to identify clusters within each subgroups.

In addition, patients report randomly distributed average scores across all 13 visits with respect to symptom duration (fig.10). Also, clustering method within subgroups is more appropriate as the data is not skewed.

2) Clustering method: After dividing the patients into four groups as shown in fig. 9, I applied clustering package *kml* to find the best clusters. This method attempts to find the best cluster by adjusting clustered curves. These steps are shown in fig. 11. For each subgroups by symptom duration, I have the best clustering found, which are described in fig. 12.

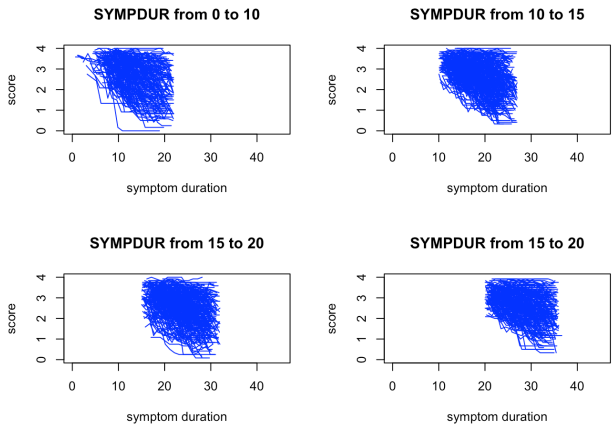

Fig. 9: Plot of registered data for each subgroups with respect to symptom duration

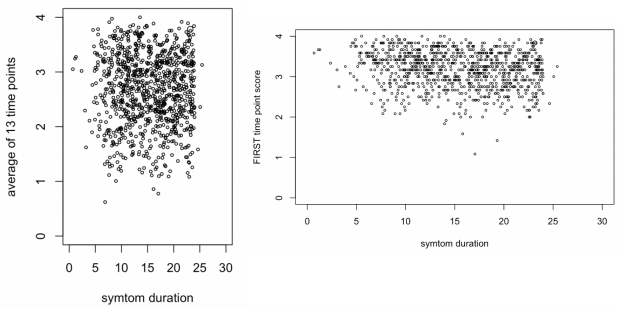

Fig. 10: Plot for averaged score across all visits and first visit with respect to symptom duration

IV. Envisioned Future Work and Potential Impact

After clustering the patients by symptom duration, envisioned future work is to identify features, such as gender, age, and the use of medication (riluzole), for the clustering method, and run multiple linear regression to see which predictors are statistically significant. Potential methods for feature selection include Lasso regularization, elastic net, and random forest.

By grouping patients according to symptom duration and demographics, current and new patients' disease progression can be diagnosed, and patients can be effectively treated with anticipation.

References

[1] Marron, J. S. et al. "Functional Data Analysis of Amplitude and Phase Variation." Statistical Science 30.4 (2015): 468–484. Crossref. Web.

[2] Aurore Delaigle & Peter Hall (2013) Classification Using Censored Functional Data, Journal of the American Statistical Association, 108:504, 1269-1283, DOI: 10.1080/01621459.2013.824893

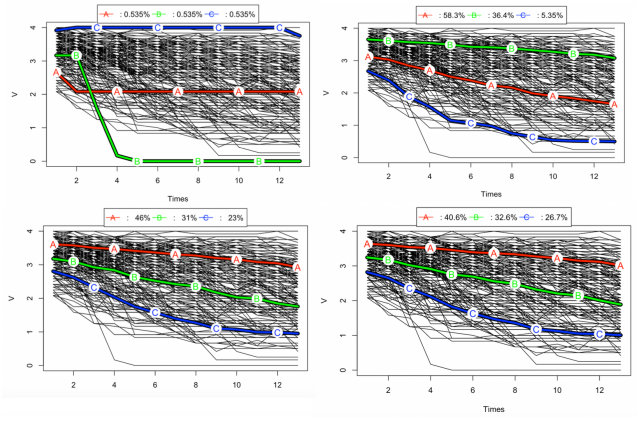

Fig. 11: Steps for learning clustering method (top left to botton right)

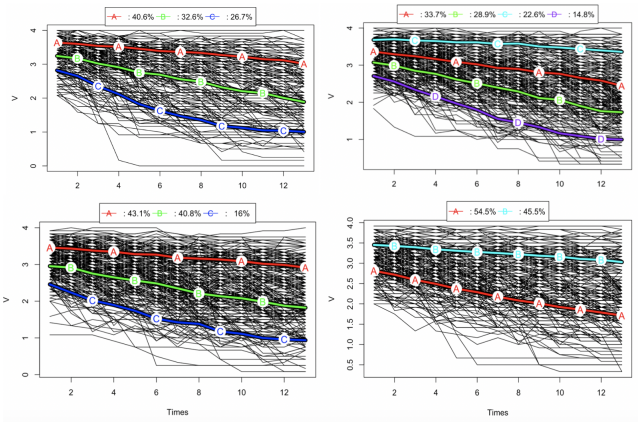

Fig. 12: Best clusters for four groups with respect to symptom duration