MLM: PRACTICUM

Analyzing Life Outcome Trajectories Utilizing Categorical Functional Data Analysis within a Survival Analysis Framework

Howell Lu

New York University, New York, United States

ARTICLE HISTORY
Compiled December 16, 2025

ABSTRACT
We extend the usage of Categorical Functional Data Analysis (CFDA) onto the biofam dataset utilizing a Discrete-Time Hazards Model onto past sequence history data at differing periods along an individual's lifespan to predict a terminal outcome within the next time state. To enable CFDA onto different sequences lengths within a hazard dataset, we normalize sequence lengths using trajectory's relative duration. In addition, CFDA is compared against traditional methods such as Distinct Successive States (DSS) with Partition around Medoids (PAM) and a regression of the outcome variable against all distinct state sequences at given times.

KEYWORDS
Categorical Functional Data Analysis, Functional Data Analysis, Sequence Analysis, Longitudinal Data, Survival Analysis

1. Introduction

Sequence History Analysis (SHA) is a commonly used method for synthesizing life-course history, it measures longitudinal data by applying a distance measure on differing sequences; one commonly used such method is Longest Common Subsequence [3], which calculates the transitions needed, using only insert and delete to convert one sequence to another as a metric of dissimilarity between two sequences. This is compounded by utilizing a clustering technique such as PAM on a dissimilarity matrix to cluster observations together.

Event History Analysis (EHA) predicts risks at time t and utilizes the Person-Period dataset format on observations in a discrete time survival format to predict whether a hazard event occurs. This method is commonly paired with Sequence History Analysis by utilizing the clusters built by Sequence History Analysis as a predictor.

Longest Common Subsequence paired with clustering has limitations as not every state has the same distance. The distance between (Single - Relationship) is the exact same as (Relationship - Married), although the subjective human experience defines a large gap between the former compared to the latter. Additionally, the act of clustering yields a discretization error as observations are forced to be associated with a specific cluster rather than a mixture of many. In spite of these issues: uniform

distance and discretization, Longest Common Subsequences is frequently used as the combinatorial space is too large for traditional processes to fully capture all variation. Our specific dataset has a total of $(7^{15}) + (7^{14}) + \ldots + (7^{1})$ or more than 5 trillion possible sequences. However, not all sequences can be used, as one would deal with both overfitting and sparsity if all sequences were used. Additionally, recent techniques such as encoders are not a silver bullet as encoders take a large degree of fine-tuning to properly apply, given that the space is still sparse for encoder-based methods.

One possibility which circumvents the limitations of these traditional methods is Categorical Functional Data Analysis (CFDA) [4], which offers the possibility of representing sequences as continuous Functional Principal Components (FPC). Commonly used sequence normalization methods [1] followed by utilization of CFDA on said sequences allows for the usage of such FPCs as potential model features.

2. CFDA (Categorical Functional Data Analysis):

Representing the variation within the data between distinct categorical states as continuous variables is difficult. However, it solves the issues of overfitting and sparsity by storing representations of the combinatorial space in lower dimensions. CFDA specifically allows for the utilization of greater data in this manner:

(1) Represent sequence histories as a stochastic process.

Time	1	2	3	4	5	6
Living With Parents	1	1	0	0	0	0
Left Parents Abode	0	0	1	1	1	1

(2) Fit a series of curves that would approximately be the

$$\#\text{ of states} \times \#\text{ of people}$$

Each individual curve is joint on the other curves for the same individual so the probabilities sums to 1. This has a simplex constraint.

(3) Implement a Centered Log-Ratio which would log-transform the probabilities against the geometric mean so that the states are mapped into Euclidean space.

(4) Perform eigen-analysis on the joint covariance matrix and the inner product matrix to generate the eigenvectors.

(5) Multiply the eigenvectors by the basis function so that we have the principal functional components mapped to the basis function.

These principal functional components generated in step 5 generate a series of continuous vectors which capture the variance within the sequence.

3. Method

We utilize a mixture of both EHA and SHA to build our model which is based on past life histories of 2000 individuals born in Switzerland between the years of 1909 and 1957 [2]. These individuals have their habilitation status recorded annually from

the ages of 15 to 30 comprising of: P = Living with Parents, L = Left Parent's Home, M = Married but still living with Parents, LM = Left Parent's and Married, C = Has a Child but still living with Parents, LC = Left Parent's abode and has a Child, D = Divorced, LMC = Left, Married and has children. The outcome event is whether the individual arrives at "LMC" or Left + Married + Children in the next time state. All observations which include the terminal stage are censored and the outcome variable is defined as TRUE for the period immediately preceding the outcome event, all other periods are as FALSE. This structure utilizes the baseline formulated by Scott et al [5].

LCS (Longest Common Subsequences) creates a distance matrix between the unique sequence states using only insertion and deletion. For example, PPP-MMM and PPP-LL would both be reduced to P-M and P-L and these sequences would have a distance of 2 between each other (One Deletion and One Insertion). PAM (partitioning around medoids) is applied onto this distance matrix to cluster all observations into distinct clusters. The model then regresses the outcome variable (odds of LMC next time) against sex, time, and the assigned cluster.

One further test utilized is modeling all distinct prior states which have occurred prior to that moment for a person-period pair (including the current event) as categorical variables and utilizing that variable as a predictor within a regression rather than clusters.

The final method would fully utilize CFDA onto a normalized sequence, as CFDA cannot be utilized on datasets with different time lengths, as the matrix algebra would simply not compute. Sequences would be represented proportionally, irrespective of the total length. In this example, we would normalize past sequences onto 100 length sequences.

Original Dataset

Time	**State**
1	P
2	M
3	LC

Converted Dataset

Time	**State**
1	P
2	P
...	...
33	P
34	M
35	M
...	...
67	M
68	LC
...	...
100	LC

These CFDA FPCs are calculated on the converted dataset and appended back to the original DataFrame and subsequently used as predictors.

4. Results and Analysis

Method	AIC	Residual Deviance
Regression on Past Sequence History	6488	6444
Regression on 14 Clusters (PAM on LCS on DSS)	6480	6448
Regression on CFDA With 33 Functional Principal Components.	6325	6253

This CFDA methodology yields greater predictive accuracy than discrete states alone (Residuals of 6253 vs 6444). In addition, the functional principal components (FPCs) are informationally dense, as 25 functional principal components explain 0.969% of the variation. (Appendix A).

One unique observation is the fact that the harmonics are representative of specific classes and trends. Harmonic 2 has a strong positive association between $a_x(t)$ and the state of M (Married) and (C) (Has a child). Instances with the highest value of the second FPC incidentally spend large sections of their time in the "M" and "C" state. Conversely, we notice that for Harmonic 3, the value of the $a_x(t)$ is positively associated with the LC state and we also see that the instances with the lowest values for the third FPC also spend a vast amount of their time within the state of LC (Appendix B).

One further emergent property of CFDA, is that some of these functional principal components begin to encapsulate trends which directly relate to our outcome variable. Functional Principal Component 19 has the highest significance and the lowest p-value and it captures what appears to be the departure of individuals from their family home at a specific time. The specific trajectories noted is that an increased prevalence of M (Married) and LM (Left-Married) is directly associated with the outcome occurring, which coincides with the dataset and rational assessment that children generally follow marriage, especially given the LMC dependency in the dataset. (Appendix C).

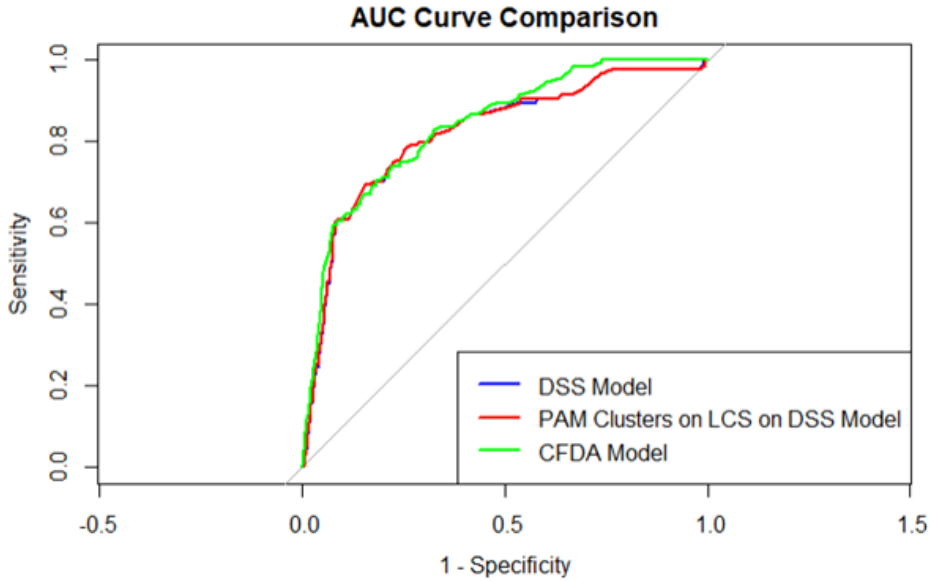
It does appear that CFDA is capable of extracting more fidelity from sequences than traditional methods whilst also maintaining generally decipherable components.

4.1. Cross-Validation

To verify that the functional principal components are not merely overfitting to the original dataset, the Biofam dataset was split into an 80:20 train-test split. Three models were trained on the dataset: 1.) Regression on DSS. 2.) Regression on PAM on LCS on DSS and 3.) Regression on CFDA FPC's.

These models were then applied to the held-off test set and AUC was used as a metric:

Model	Training AIC	Testing AUC
Regression on DSS	5180	0.8202
Regression on PAM on LCS on DSS	5170	0.8209
Regression on CFDA FPCs	5026	0.8375



4.2. Correlation between Distance matrices

Correlation between Distance matrices: A distance matrix was computed for on the CFDA FPC Coefficients and on a second distance matrix, which was the LCS on the DSS matrix. A mantel test was run against both matrices and the correlation between matrices is 0.625. However, what should be noted is that the mantel distance on both matrices subset on different lengths yield different results. As sequence lengths get progressively longer, the correlation between the LCS on DSS distance matrix and the CFDA distance matrix gets progressively weaker.

5. Discussion: Extensions, Limitations and Future Directions

5.1. Representation of different data lengths

One limitation CFDA cannot solve even post - normalization is the inability to account for differences between the homogenous sequence length. A sequence such as P and P-P-P-P-P-P-P-P-P would have the exact same FPC's despite different context. One naïve solution would be to simply add an interaction term between each of FPC and time, but this better results at the cost of adding many more predictors and does not uniquely benefit the CFDA FPCS relative to the other models.

5.2. Representation of different data lengths

Model Results with An Interaction Term Between the Predictor Variable and Time

Model	Residual Deviance on Whole Dataset	AIC (Full Dataset)	AIC (Training Dataset)	AUC (Test Split)
Model 1 (DSS)	6267	6348	5065	0.8351
Model 2 (PAM)	6266	6326	5034	0.8367
Model 3 (CFDA)	6106	6244	4978	0.8406

5.3. Examining the Drift Between different successive states

One perplexity observation is the growing distance between the CFDA distance matrices and the PAM on LCS on DSS distance matrix as the sequence lengthens. One potential explanation is this: Life pathways are not homogenous; there are specific trends and periods which are more correlated with transitions. However, LCS on DSS always enforces near constant distance while CFDA FPCs matrices maintain variability; here are some explanations. (Appendix D)

(1) Growing number of cumulative transitions: At early periods times, transitions are rare. However, with more transitions occur with every subsequent period. Each transition yields a different cost for CFDA compared to LCS on DSS which increases the cumulative errors.

(2) Non-uniform heterogeneity at different periods. Movements to different states occur non-uniformly, leading to differences in measurement. As an example, presume that there are very few changes in relationship status before the age of 25 and then the majority of the dataset marries at the age of 26. At the period of age of 26, CFDA represents marriage as a small proportion of the total data to be mapped which is represented as a minimal change, but the LCS on DSS would have marriage as a large-scale transition.

(3) Transitions to unique states may yield different costs. It is likely transitions between different states yield different costs and cannot be compared to a uniform transition space. Transitions to different states also occur more frequently during later timespans within the Biofam data.

(4) Growing heterogeneity within the CFDA distance matrix when subset within a distinct state, while LCS on DSS remains homogenous. Observe the subset of P-L at time = 2. This unique transition can only be represented with one permutation for both CFDA and DSS, thereby there is no distance between the observations within the P-L substate in the CFDA distance matrix and no distance between the LCS on the DSS Matrix. However, at time = 3, there is distance between the distinct successive states within the CFDA matrix P-L-L and P-P-L are different and will have distance from one another and also different distances between all other states. However, on the LCS on DSS, everything within that subset has the same distance with itself and other states. At each successive state the number of unique states increases exponentially increasing the variance within the CFDA distance matrix with time, thereby pulling the mantel correlation lower.

A worthwhile idea is to examine the changes within the CFDA distance matrix between different observations within the same LCS-DSS set. A further possibility

would be to exclude outliers and to create synthetic datasets to analyze how quickly the correlation these two methods drifts under real world circumstances.

5.4. Predictive Value of CFDA FPCs

Although the embeddings represent what appears to be pathways and the functional principal components are statistically significant with respect to the outcome variable. A further challenge is associating specific principal components with a specific Distinct Successive State. Each embedding alone does not seem to predict a specific state with a high degree of confidence but does to generally point towards specific trends and pathways which are rather difficult to quantitatively define.

5.5. Conclusion and Summary

CFDA is a powerful tool that can be applied to survival analysis and has been shown to outperform traditional methods such as PAM on LCS on DSS on our dataset for prediction without incurring issues regarding overfitting and sparsity. However, the FPC's of CFDA are not fully interpretable in the same manner that PAM on LCS on DSS is but this method generally retain a level of basic interpretability.

6. References

References

[1] Gesche Brandt and Susanne de Vogel. Normalising sequence lengths using the relative duration of episodes: an application to doctoral trajectories in germany. *Longitudinal and Life Course Studies*, 15(4):492 – 505, 2024. .

[2] Alexis Gabadinho, Gilbert Ritschard, Nicolas S. Müller, and Matthias Studer. Analyzing and visualizing state sequences in r with traminer. *Journal of Statistical Software*, 40(4): 1–37, 2011. .

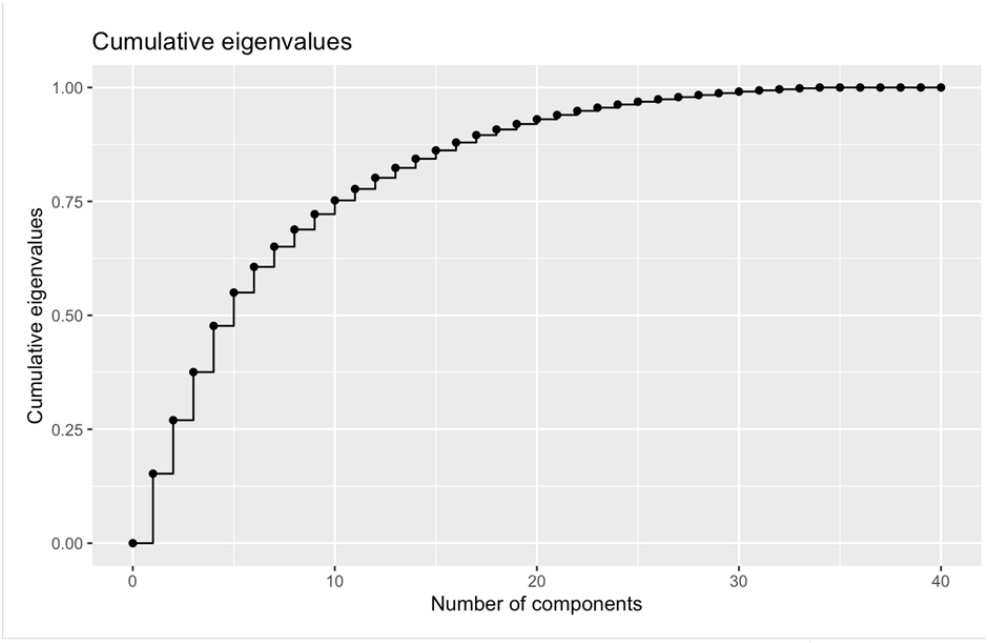
[3] S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970. ISSN 0022-2836. .

[4] Cristian Preda, Quentin Grimonprez, and Vincent Vandewalle. Categorical functional data analysis. the cfda r package. *Mathematics*, 9(23), 2021. ISSN 2227-7390. . URL `https://www.mdpi.com/2227-7390/9/23/3074`.

[5] Marc Scott, Jean-Marie Le Goff, and Gauthier Jacques-Antoine. History matters: the statistical modelling of the life course. *Quality Quantity*, 58:1–25, 03 2023. .

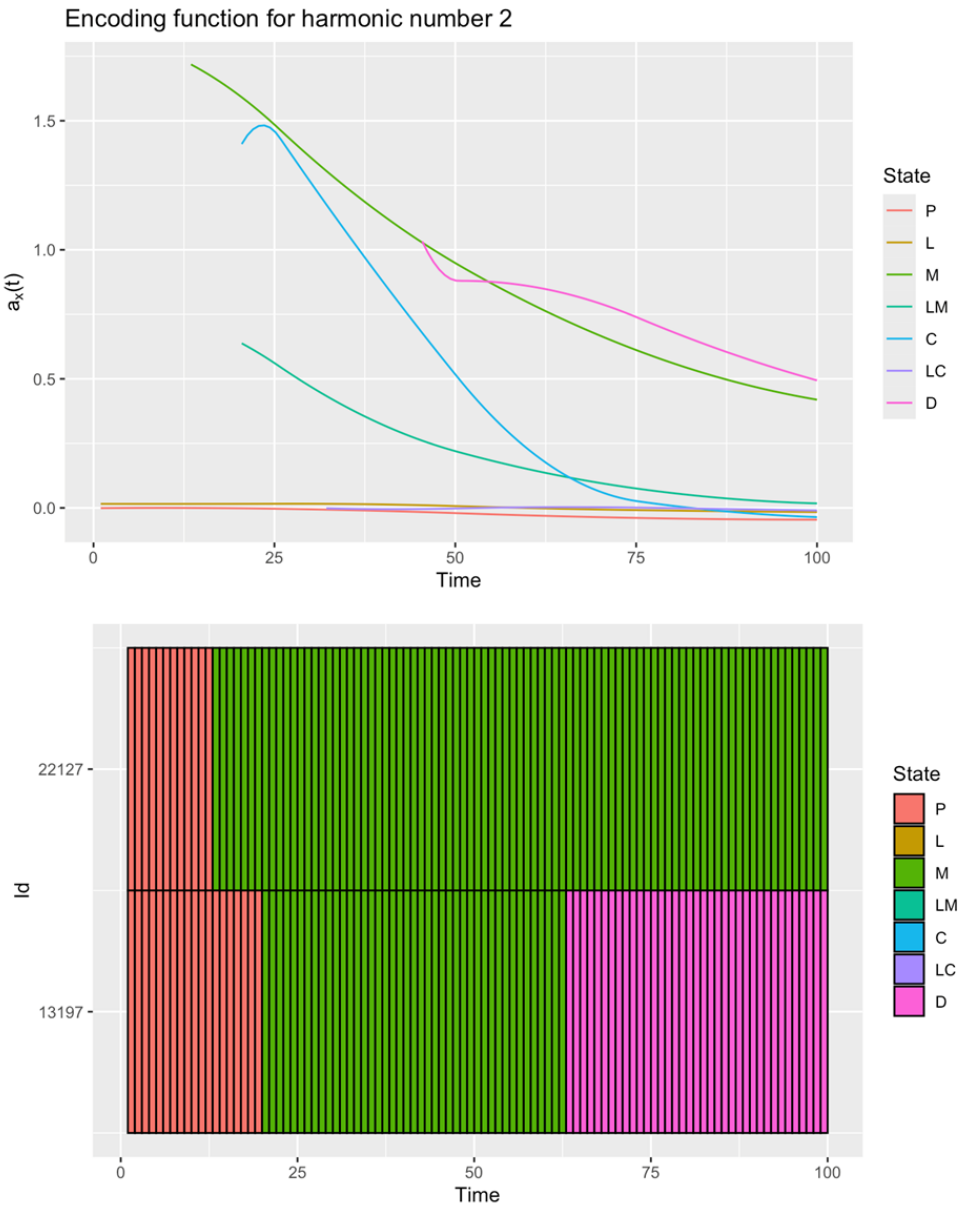
7. Appendices

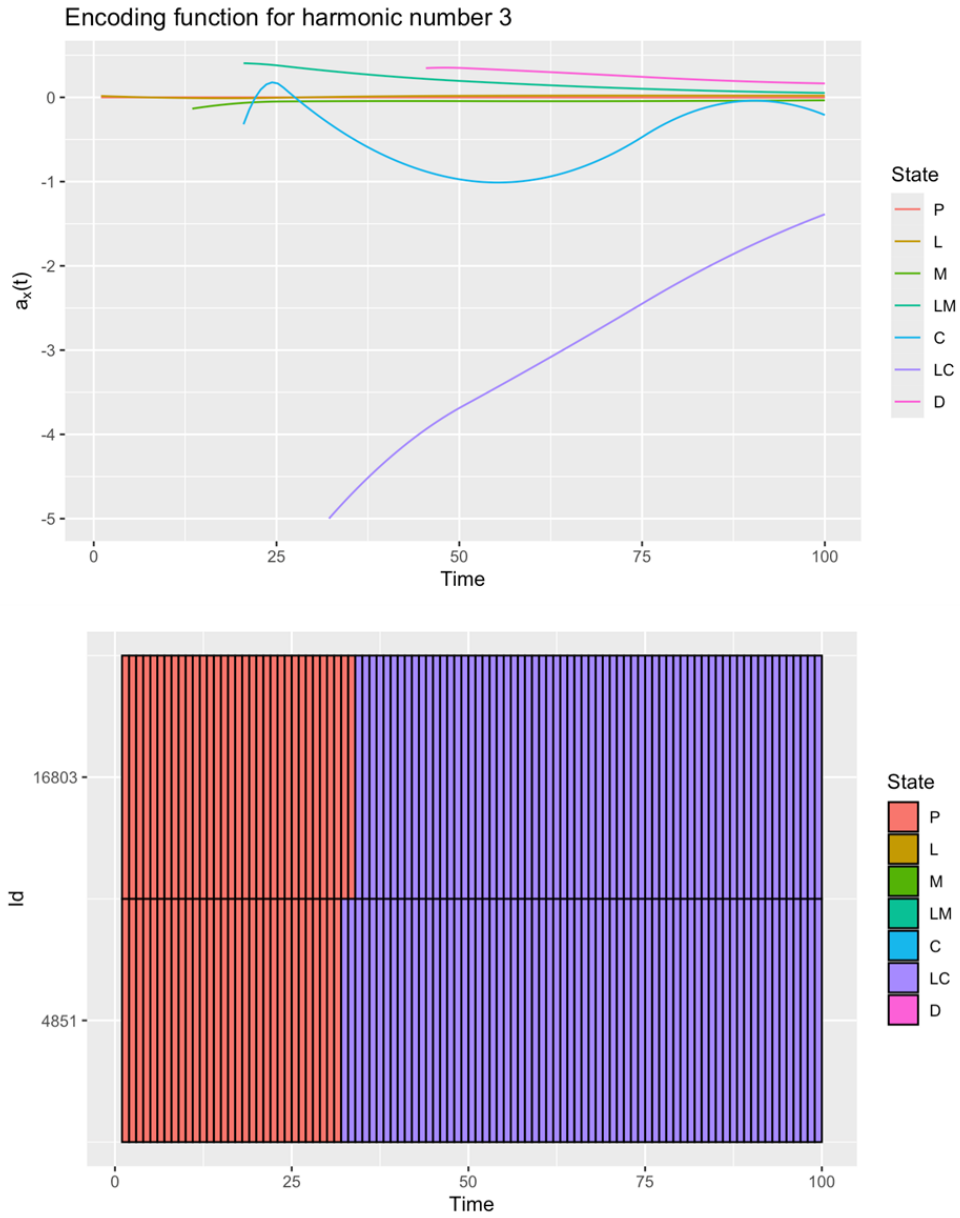
Appendix A. Variance Explained by Cumulative FPCS



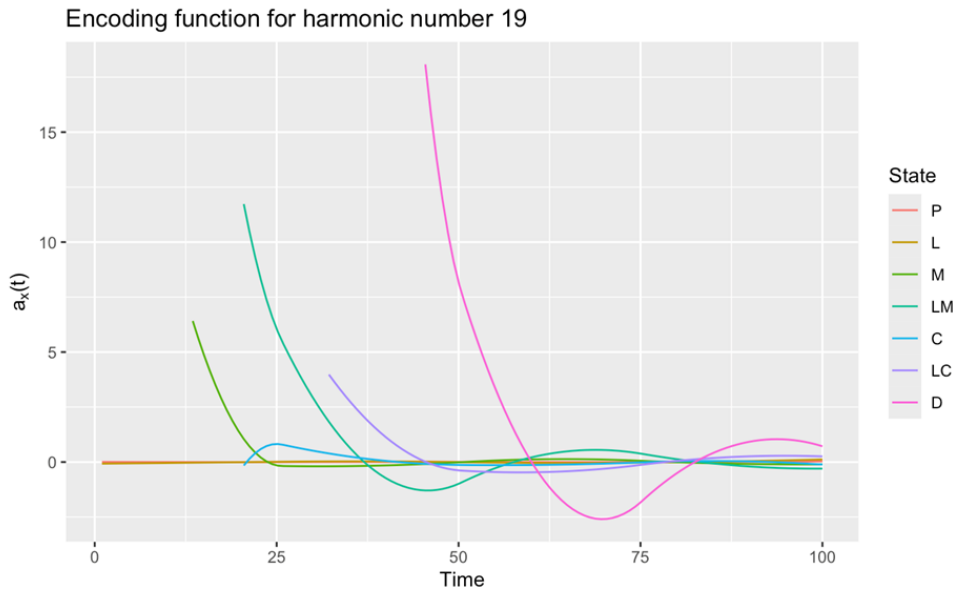
Appendix B. FPC Covariates and Harmonics Graphed

```
Call:
bayesglm(formula = as.formula(formula_str), family = binomial,
    data = sha3)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.419694   0.155801 -34.786  < 2e-16 ***
time         0.057619   0.011394   5.057 4.26e-07 ***
sexwoman     0.184025   0.074236   2.479  0.01318 *
V1           0.009457   0.011537   0.820  0.41237
V2          -0.681177   0.111996  -6.082 1.19e-09 ***
V3          -0.134357   0.026368  -5.095 3.48e-07 ***
V4           0.842295   0.089556   9.405  < 2e-16 ***
V5           1.785040   0.194242   9.190  < 2e-16 ***
V6          -0.367894   0.176181  -2.088  0.03678 *
V7          -0.111068   0.025803  -4.304 1.67e-05 ***
V8          -0.193831   0.221551  -0.875  0.38164
V9          -1.690994   0.223540  -7.565 3.89e-14 ***
V10         -0.042671   0.162547  -0.263  0.79292
V11         -3.440491   0.463137  -7.429 1.10e-13 ***
V12         -0.397640   0.109180  -3.642  0.00027 ***
V13         -0.042835   0.208889  -0.205  0.83753
V14         -1.545096   0.146242 -10.565  < 2e-16 ***
V15          0.666705   0.298370   2.234  0.02545 *
V16          0.367666   0.114599   3.208  0.00134 **
V17         -0.322601   0.358334  -0.900  0.36797
V18          0.274695   0.041393   6.636 3.22e-11 ***
V19         -2.398744   0.185417 -12.937  < 2e-16 ***
V20          0.305821   0.266519   1.147  0.25119
V21         -1.563605   0.193092  -8.098 5.60e-16 ***
V22          1.177941   0.150618   7.821 5.25e-15 ***
V23         -0.119513   0.185985  -0.643  0.52049
V24         -0.252035   0.146114  -1.725  0.08454 .
V25         -0.327988   0.214591  -1.528  0.12641
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7943.9  on 27072  degrees of freedom
Residual deviance: 6267.3  on 27045  degrees of freedom
AIC: 6323.3
```

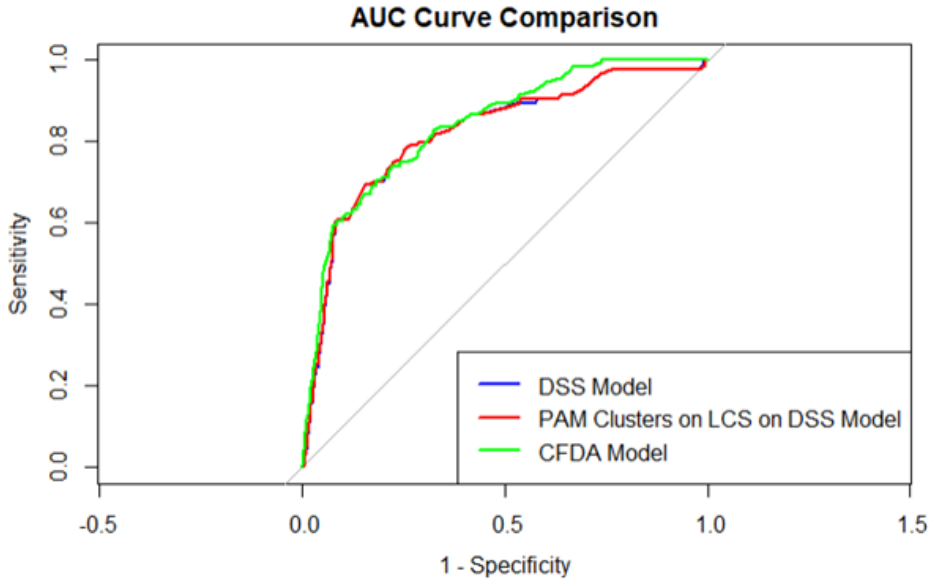

Encoding function for harmonic number 2
1.5
1.0
0.5
0.0
$a_x(t)$
0
25
50
75
100
Time
State
P
L
M
LM
C
LC
D
22127
13197
Id
0
25
50
75
100
Time
State
P
L
M
LM
C
LC
D


Encoding function for harmonic number 3
0
-1
-2
-3
-4
-5
$a_k(t)$
0
25
50
75
100
Time
State
P
L
M
LM
C
LC
D
16803
4851
Id
0
25
50
75
100
Time
State
P
L
M
LM
C
LC
D

Appendix C.


Encoding function for harmonic number 19
15
10
5
0
$a_x(t)$
0
25
50
75
100
Time
State
P
L
M
LM
C
LC
D

Appendix D.


AUC Curve Comparison
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
-0.5
0.0
0.5
1.0
1.5
1 - Specificity
DSS Model
PAM Clusters on LCS on DSS Model
CFDA Model