Parkinson's Progression Markers Initiative
Play a Part in Parkinson's Research

Modeling Progression of Parkinson's Disease

Liyang Sun, Suchan Vivatsethachai, Christina Ji
Supervisor: Prof. David Sontag

Abstract

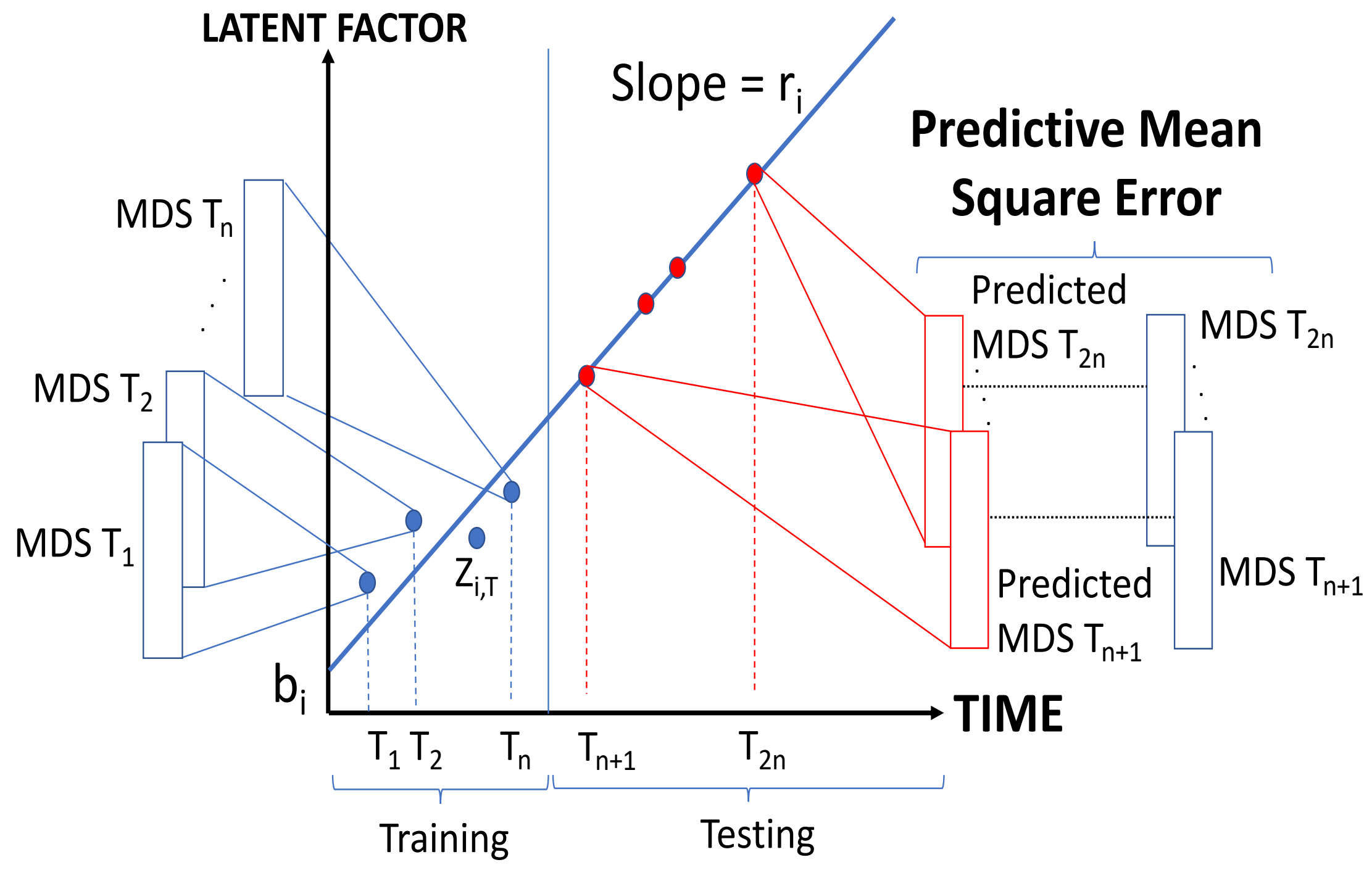
In this project, we learn a low-dimensional representation that reflects the progression of motor symptoms in Parkinson's disease patients. To accommodate categorical data, we develop auto-encoders that are capable of learning nonlinear relationships. We evaluate these auto-encoders based on clinical interpretability and ability to predict future timepoints.

1. Clinical problem

Parkinson's disease (PD) is the second most common neurodegenerative brain disorder. It is a complex and heterogenous disease characterized by decreased level of the neurotransmitter dopamine. Clinical decisions for treating Parkinson's disease are primarily based on motor symptoms, which are currently measured by high-dimensional clinical rating scales developed by Movement Disorder Society (MDS).

Workflow We represent each individual at a given timepoint t by a low-dimensional latent state $z_{i,t}$ using auto-encoders, as depicted in the following figure. With these $z_{i,t}$, we can

- estimate rates of PD progression (r_i), which can be used for subtyping patients;
- predict future PD symptoms.



2. Data

We used the dataset from the Parkinson's Progression Markers Initiative (PPMI).

- 423 patients in the PD cohort with motor assessed by MDS repeatedly (2221 patient×timepoint);
- Consists of 46 MDS questions on a 0-4 scale;
- To avoid confounding due to treatment, we only use untreated timepoints.

Feature	Mean (SD)	Feature	Perc.
# visits	5.3 (2.4)	Male	65.5%
Duration	0.45 (0.53)	White	94.8%
Age	61.6 (9.7)	History	24.3 %
MDS I	5.8 (4.2)	Right-dom.	42.3%
MDS II	5.7 (4.2)	Depressed	13.9%
MDS III	20.3 (8.9)	Sleepiness	15.6%

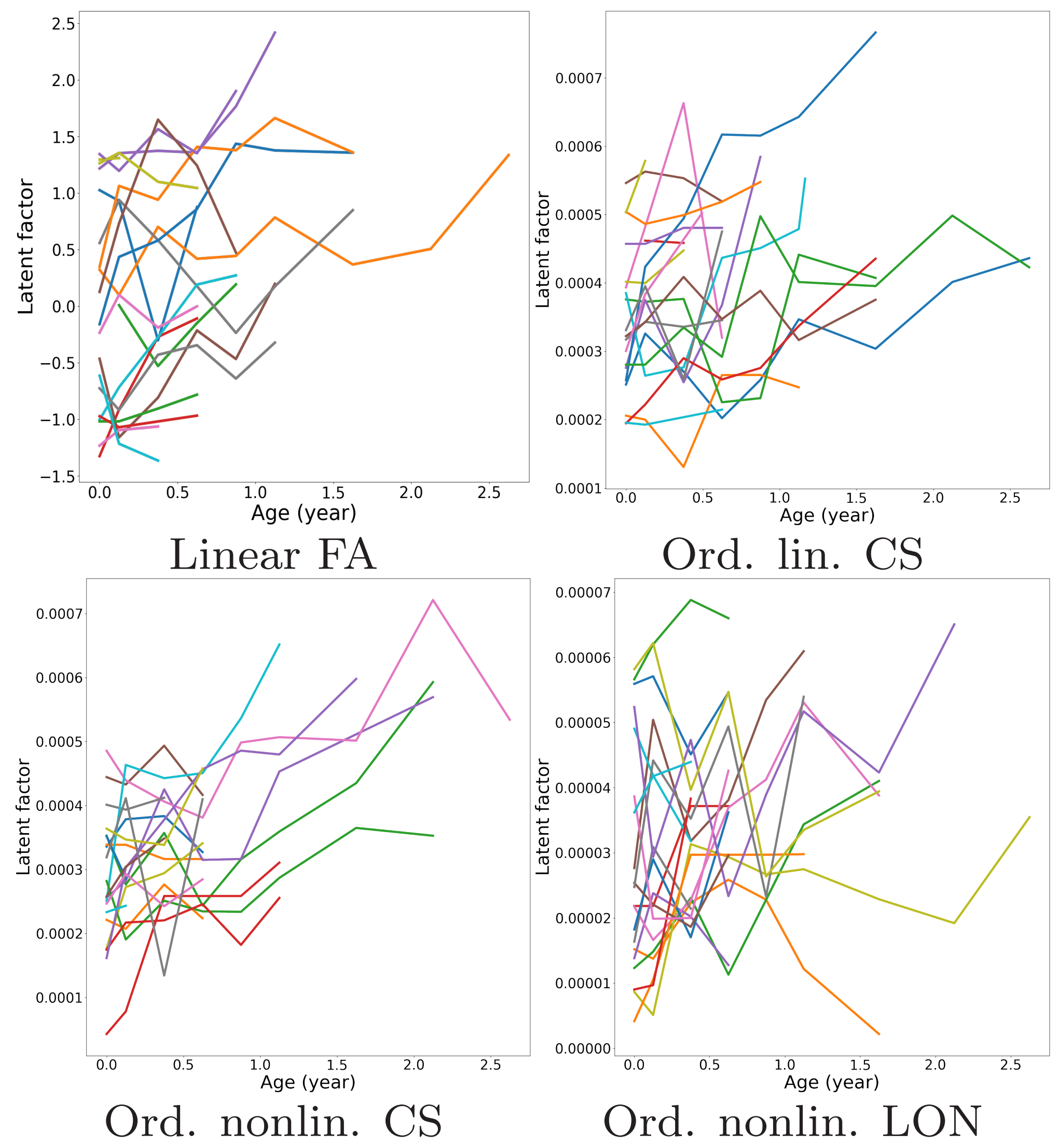
3. Methods

We denote by $y_{d,i,t}$ (as outcome) and $x_{d,i,t}$ (as feature) response to MDS question d for a patient-timepoint pair (i, t).

- Baseline: linear factor analysis (FA);
- Encoder $z_{i,t} = f(x_{d,i,t})$: linear or nonlinear, can be multi-dimensional;
- Decoder $\widehat{y}_{d,i,t} = g(z_{i,t})$: linear, monotonic polynomial by [2], ordinal regression by [3];
- Longitudinal constraint: $z_{i,t'} > z_{i,t}$ for $t' > t$ assuming PD does not improve over time [1]. We denote methods with this longitudinal constraint by LON. The remaining methods do not impose such a constraint and treat data as cross-sectional (CS).

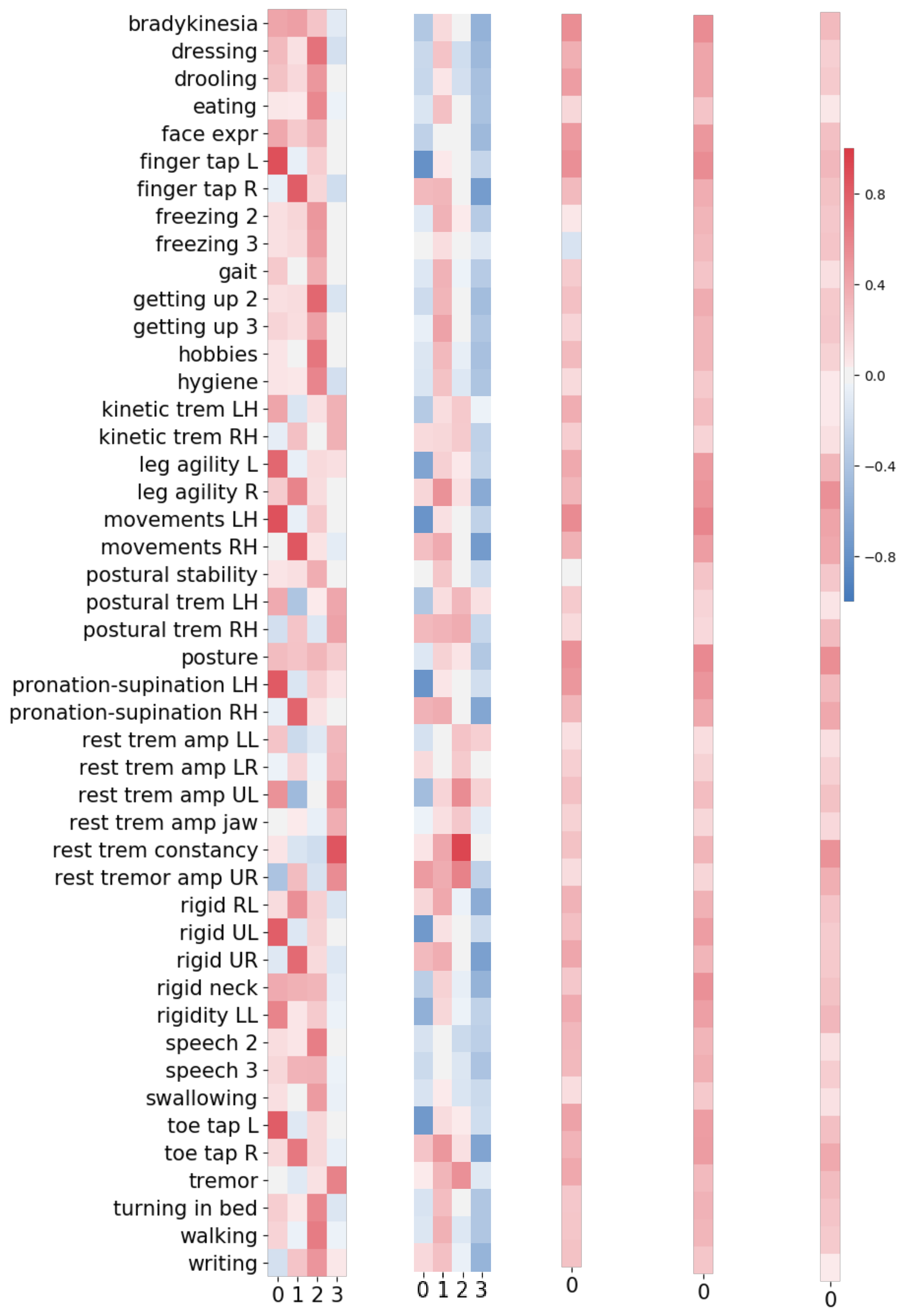
Method	Encoder	Decoder
Linear FA	Linear	Linear
VAE	Nonlinear	Nonlinear
Aging CS	Nonlinear	Mono. poly.
Aging LON	Nonlinear	Mono. poly.
Linear ordinal CS	Linear	Ordinal
Nonlinear ordinal CS	Nonlinear	Ordinal
Nonlinear ordinal LON	Nonlinear	Ordinal

4. Results



Linear FA
Ord. lin. CS
Ord. nonlin. CS
Ord. nonlin. LON

Latent states $z_{i,t}$ over time for 20 randomly selected patients. Age is time since enrollment in PPMI.



Linear FA
VAE
Ordinal
lin. CS
nonlin. CS
nonlin. LON

Correlation between latent states and input data. Each row is a MDS question. Each column is a dimension of the latent state.

5. Quantitative Evaluation

Using 5-fold train/validation/test = 70/10/20.

- CI : concordance index, as in the share of learned latent states $z_{i,t}$ that respect time ordering i.e. $z_{i,t'} > z_{i,t}$ for $t' > t$. For multidimensional $z_{i,t}$, we take the dimension with the highest CI;
- MSE1: mean squared error on reconstructing outcome $\sum(\widehat{y}_{d,i,t} - y_{d,i,t})^2$ on test sample for $t = T_1, \ldots, T_{2n}$;
- MSE2: mean squared error on predicting future outcome. We extrapolate $\widehat{z}_{i,t}$ for $t = T_{n+1}, \ldots, T_{2n}$ from the first half. Then we compute $\sum(g(\widehat{z}_{i,t}) - y_{d,i,t})^2$.

Method	CI	MSE1	MSE2
Linear FA	0.682	0.683	1.532
VAE	0.645	0.402	0.637
Aging CS	0.728	0.545	0.619
Aging LON	0.745	0.571	1.001
Linear ordinal CS	0.735	1.509	1.854
Nonlinear ordinal CS	0.738	1.302	1.581
Nonlinear ordinal LON	0.651	2.028	2.255

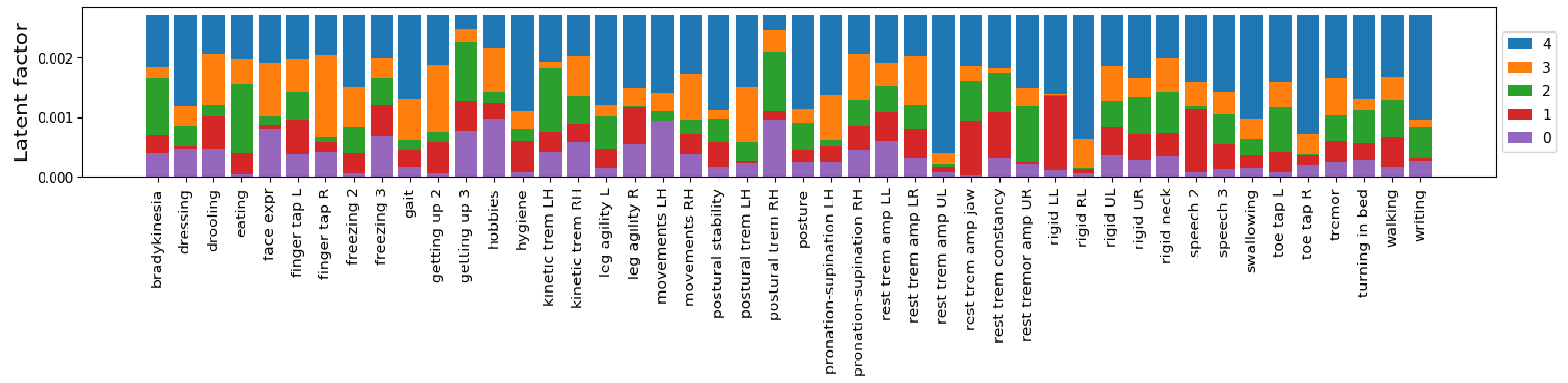
6. Subtyping

We separate patients into two clusters based on whether the estimated rate of progression r_i is above or below the median. Across models, the following patient characteristics are predictive of faster progression:

- Right-dominant, depression, older age
- Lower DaTscan ipsilateral putamen
- Cognitive: BJLO + HVLT retention
- Genetic PCA component 9

Method	AUROC	Prec.	Recall
Linear FA	0.632	0.574	0.577
Lin ord CS	0.581	0.545	0.545
Nonlin ord CS	0.557	0.531	0.531
Nonlin ord LON	0.515	0.507	0.507

Acknowlegments We thank Charles Venuto and Monica Javidnia for sharing their clinical expertise. Data used in the preparation of this article were obtained from the Parkinsons Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI-a public-private partnership-is funded by the Michael J. Fox Foundation for Parkinsons Research and funding partners, including abbvie, Allergan, Avid, Biogen, BioLegend, Bristol-Myers Squibb, Denali, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pfizer, Piramal, Prevail, Roche, Sanofi Genzyme, Servier, Takeda, Teva, UCB, and Golub Capital.



Estimated thresholds for predicting MDS responses using the nonlinear ordinal CS model.

7. Contribution

- We provide lower-dimensional representations of MDS responses that reflect PD motor progression.
- These representations allow us to identify subtypes of slow and fast motor progression.
- We plan to improve prediction beyond linear extrapolation.

References

[1] Bin Liu, Ying Li, Zhaonan Sun, Soumya Ghosh and Kenney Ng Early Prediction of Diabetes Complications from Electronic Health Records: A Multi-Task Survival Analysis Approach 2018.

[2] Emma Pierson, Pang Wei Koh, Tatsunori Hashimoto, Daphne Koller, Jure Leskovec, Nicholas Eriksson and Percy Liang Inferring Multidimensional Rates of Aging from Cross-Sectional Data 2019.

[3] Jason Rennie and Nathan Srebro Loss functions for preference levels: Regression with discrete ordered labels 2005.