Covariance Estimation and Functional Principal Component Analysis for Mixed-Type Functional Data

Application to mHealth

Debangan Dey
Postdoc Fellow
National Institute of Mental Health
Presenting at JSM ‘23

Joint work with
Rahul Ghosal,
Kathleen Merikangas, and
Vadim Zipunnikov

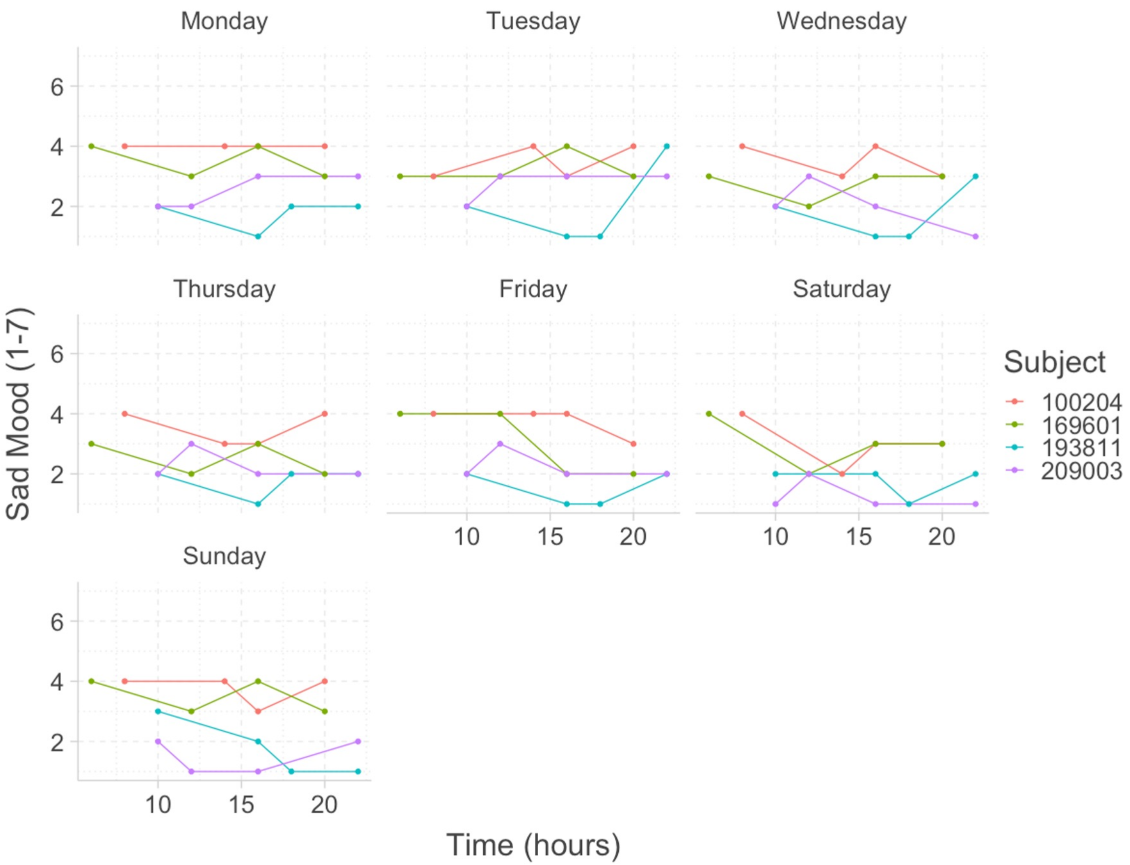
Disclaimer:

This work was supported by the National Institute of Mental Health Intramural Research Program.

The views and opinions expressed in this article are those of the authors and should not be construed to represent the views of any of the U.S. Government.

Motivation: mHealth Questionnaire data

- Real-time self-reports of mood, energy, stress, anxiety (1-7), headache (0-1) recorded through smartphones.
- Objectively recorded physical activity and sleep through smartwatches.
- Intensive mixed-type longitudinal data.
- Different measurement scales (binary, ordinal, truncated, continuous).
- Differences in subjective interpretation of scales.



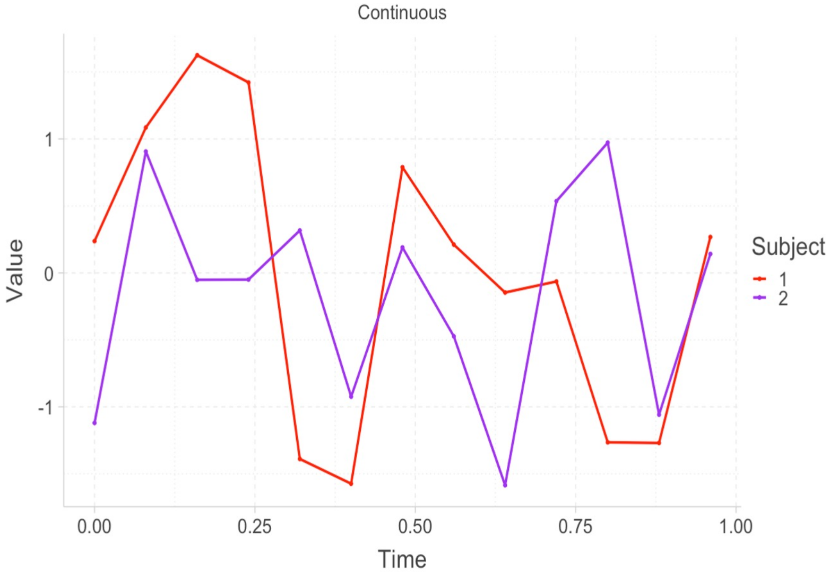
Example: Self-reported Mood (1-7)

Challenges

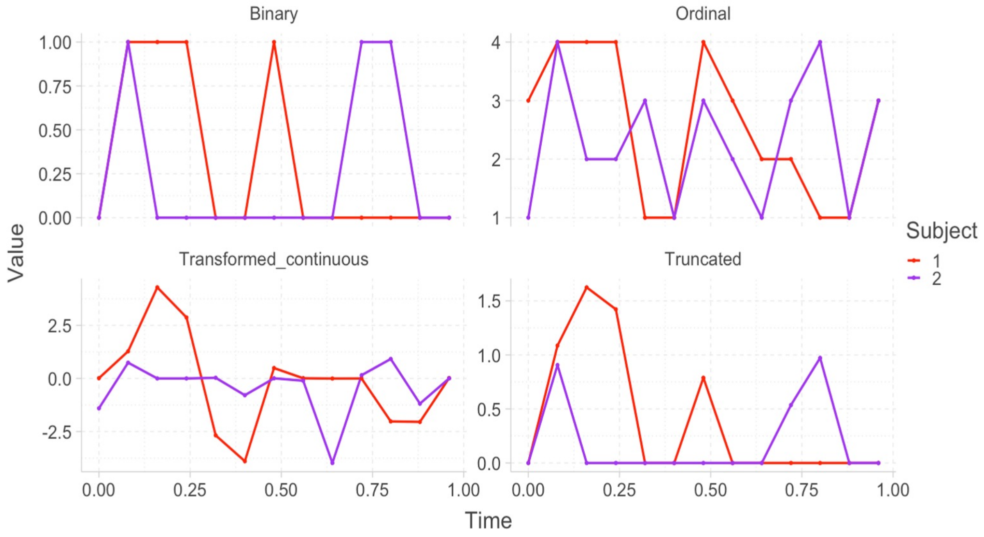
- Self-reported mood variables need to be treated as ordinal variables rather than continuous.
- Every person has subject-specific different scales.
- Can we build a general modeling framework for any of binary, ordinal, continuous and truncated type functional data?
- Can we build Functional Principal Component Analysis for such approaches?

Generalized Latent Non-paranormal Process

Latent Gaussian Process



Observed data



Generalized Latent Non-paranormal Process (X(t))

$$X(t) = f_t^{-1}(V(t)), \text{(Continuous)}$$

$$X(t) = f_t^{-1}(V(t)) I(V(t) > \Delta_t), \text{(Truncated)}$$

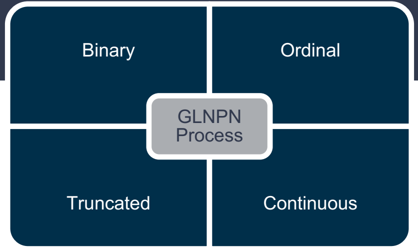
$$X(t) = \sum_{k=0}^{l_{j_t}-1} k I(\Delta_{jk_t} \leqslant V(t) < \Delta_{j(k+1)_t}), \Delta_{oj0_t} = -\infty, \Delta_{ojl_t} = \infty, \text{(Ordinal)}$$

$$X(t) = I(V(t) > \Delta_t), \text{(Binary)}$$

$$S = \{t_1, t_2, \ldots, t_m\}, ((V(t_1)), \cdots, (V(t_m))) \sim N(0, C(S, S)), f = (f_{t_1}, \cdots, f_{t_m})$$

Unknown monotone transformation functions

Estimation



Method of moments

Kendall's Tau

Estimated Time-point specific Cutoffs $(\widehat{\Delta}_t)$

$$\delta_{jj'}^{ii'} = sgn\{(X_i(t_j) - X_i'(t_j))(X_i(t_{j'}) - X_i'(t_{j'}))\}$$

$(\hat{\tau}_{jj'})_{m \times m}$ for pair of time-points

$(a-b)$ has the same sign as $(f(a)-f(b))$ for any increasing transformation f

Kendall's Tau is invariant under monotone increasing transformation

Estimation

Binary
Ordinal
GLNPN Process
Truncated
Continuous

Method of moments

Kendall's Tau

Estimated Time-point specific Cutoffs $(\widehat{\Delta}_t)$

$(\hat{\tau}_{jj'})_{m \times m}$ for pair of time-points

Covariance function (?)

Estimation

Binary

Ordinal

GLNPN Process

Truncated

Continuous

Method of moments

Kendall’s Tau

Estimated Time-point specific Cutoffs $(\widehat{\Delta}_t)$

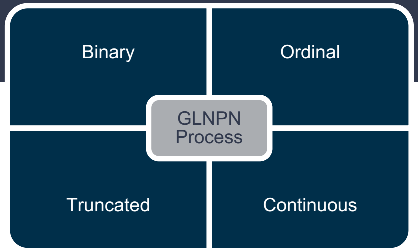
$(\hat{\tau}_{jj'})_{m \times m}$ for pair of time-points

$$C(s,t) = g\left(\sum_{k=1}^{d}\sum_{l=1}^{d} u_{kl} B_k(s) B_l(t)\right), s \neq t,$$

Covariance function (?)

Model the scaled covariance function as a tensor product spline.

Estimation



Method of moments

Kendall's Tau

Estimated Time-point specific Cutoffs $(\widehat{\Delta}_t)$

$(\hat{\tau}_{jj'})_{m \times m}$ for pair of time-points

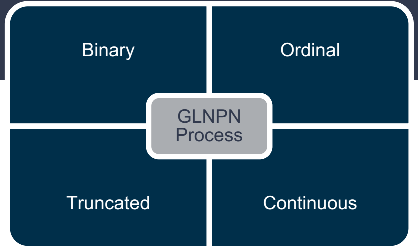
$$C(s,t) = g\left(\sum_{k=1}^{d}\sum_{l=1}^{d} u_{kl} B_k(s) B_l(t)\right), s \neq t,$$

$$g = \frac{e^x - 1}{e^x + 1}$$

Known scaling function to ensure the correlation values are within [-1,1]

Covariance function

Estimation



Method of moments

Kendall’s Tau

Estimated Time-point specific Cutoffs $(\widehat{\Delta}_t)$

$(\hat{\tau}_{jj'})_{m \times m}$ for pair of time-points

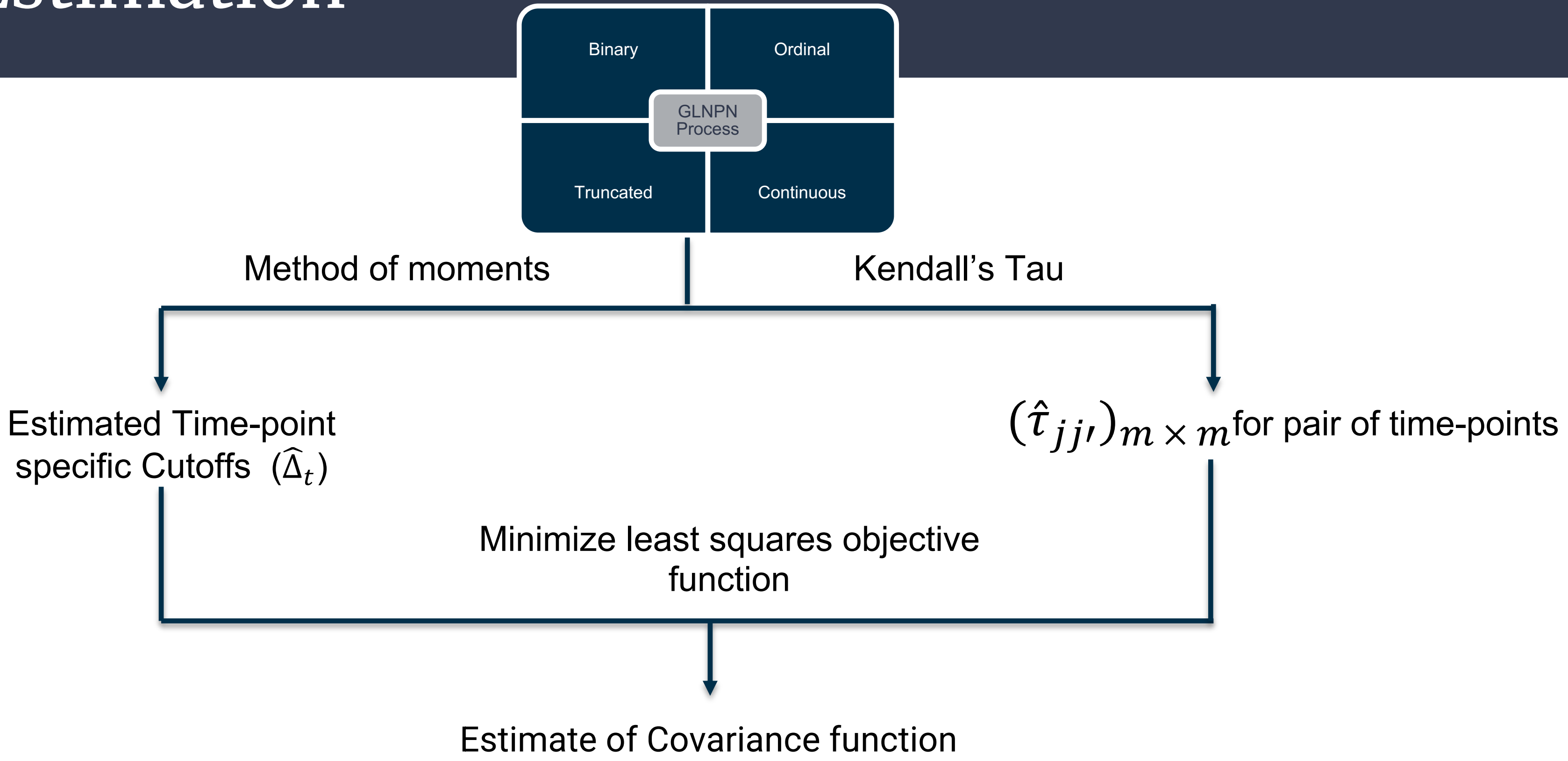
$$\hat{\boldsymbol{U}} = \underset{\boldsymbol{U}}{argmin} \sum_{1 \leqslant j < j' \leqslant m} \left(\hat{\tau}_{jj'} - F\left(g\left(\sum_{k=1}^{d} \sum_{l=1}^{d} u_{kl} B_k(t_j) B_l(t_{j'})\right)\right)\right)^2$$

Known Bridging functions*

Covariance function

* Dey, Debangan, and Vadim Zipunnikov. "Semiparametric Gaussian Copula Regression modeling for Mixed Data Types (SGCRM)." *arXiv preprint arXiv:2205.06868* (2022).

Estimation



Binary
Ordinal
GLNPN Process
Truncated
Continuous
Method of moments
Kendall's Tau
Estimated Time-point specific Cutoffs $(\widehat{\Delta}_t)$
$(\hat{\tau}_{jj'})_{m \times m}$ for pair of time-points
Minimize least squares objective function
Estimate of Covariance function

Advantages

- Can work with sparse data as we only need pairwise complete information for any pairs of time-points to calculate Kendall’s Tau.

- We can use conditional expectation to predict latent trajectory of a subject at missing time-points.

- We develop approaches to calculate latent Functional Principal Component Scores which can be used in further predictive modeling.

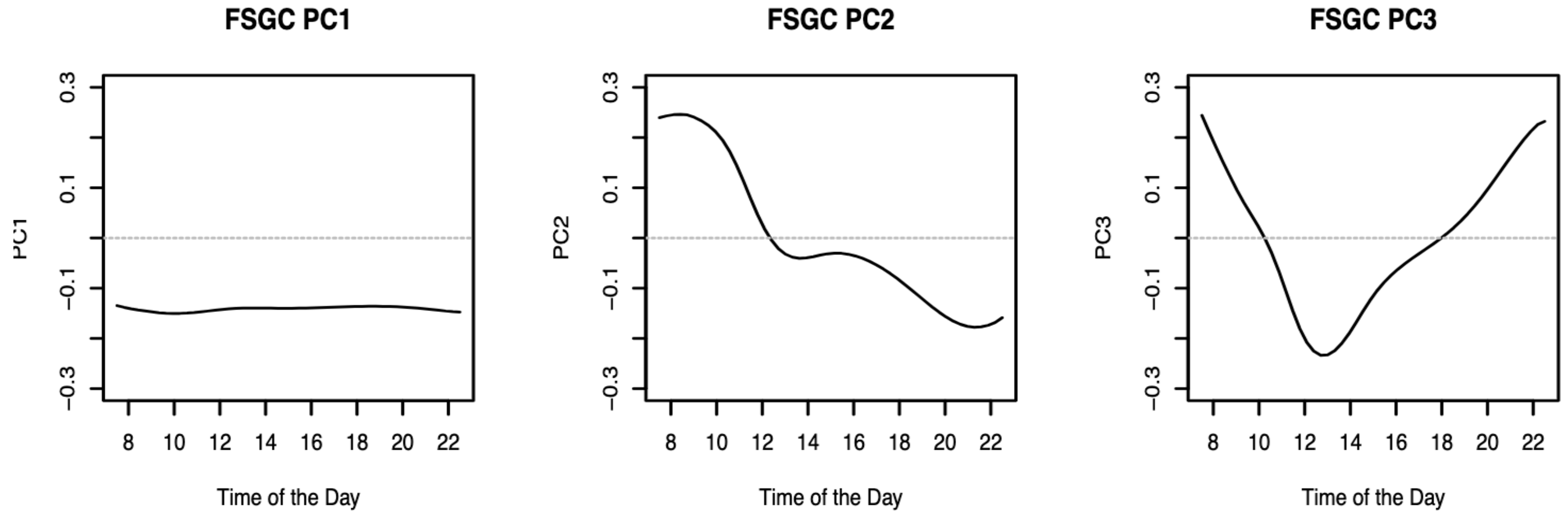
Data analysis: NIMH Family Study of Affective Disorders

Table 2

Descriptive statistics for the complete, male and female samples in the real data analysis. For continuous variable the mean and standard deviation is reported, for categorical variable the frequency in each group is mentioned. The P-values are from two-sample t-test and Chi-Square test of association with gender.

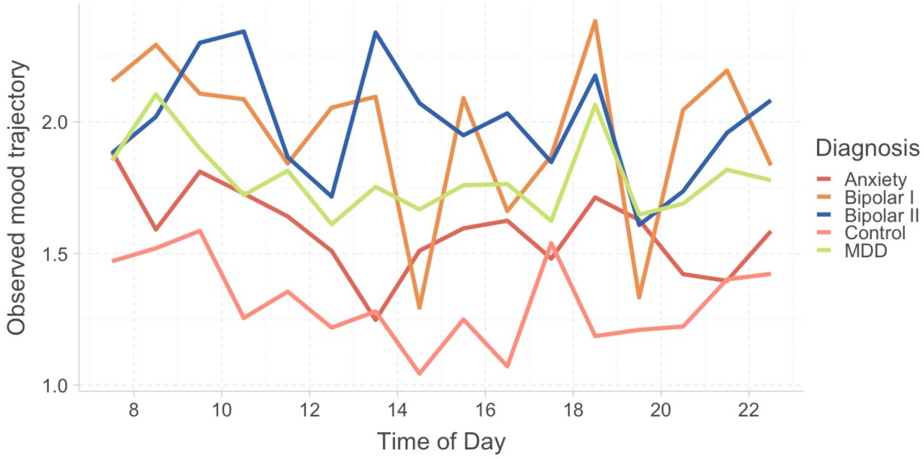
Characteristic	Complete (n=497)	Male (n=195)	Female (n=302)	P value
	Mean(sd)	Mean(sd)	Mean(sd)	
Age	41.8 (19.5)	41.2 (21.7)	42.2(17.9)	0.56
Diagnosis: control (N)	134	74	60	0.0001
Diagnosis: Anxiety (N)	97	35	62	
Diagnosis: bipolar I (N)	56	20	36	
Diagnosis: bipolar II (N)	54	22	32	
Diagnosis: MDD (N)	156	44	112	

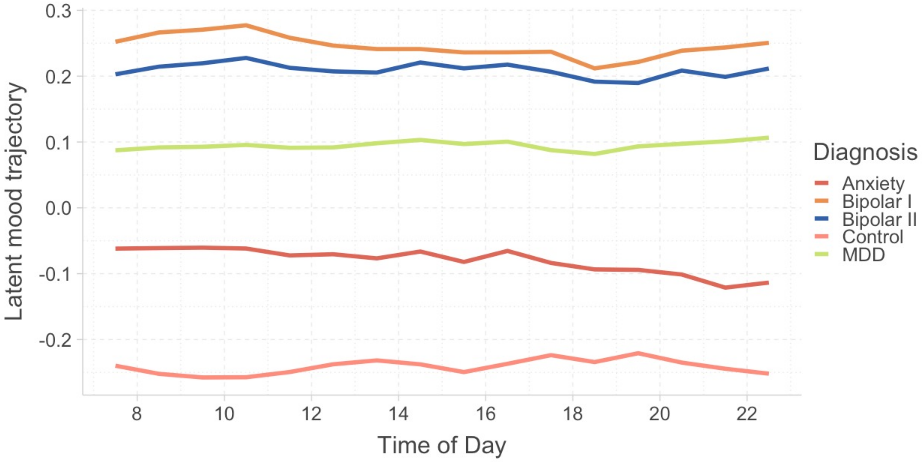
Global Functional Principal Components



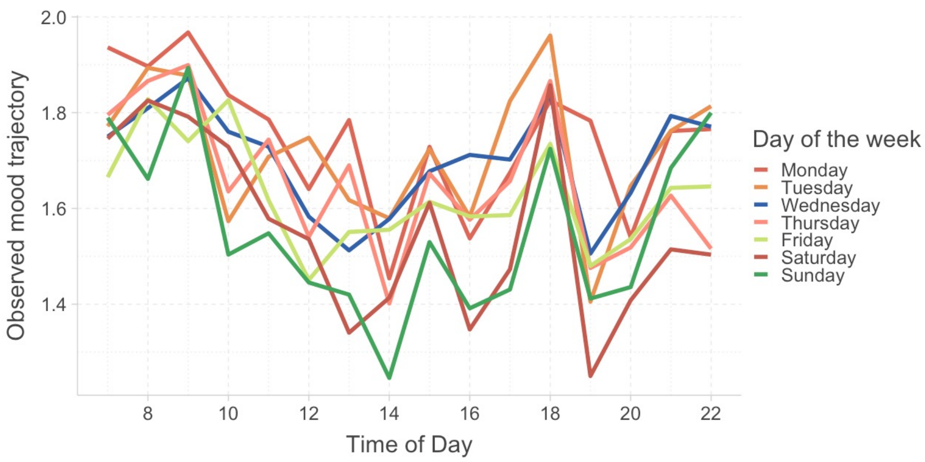
Global mean of happiness/unhappiness (66%)	A daily happiness/unhappiness angle (8%)	The degree of evening return to the morning happiness/unhappiness level (4%)

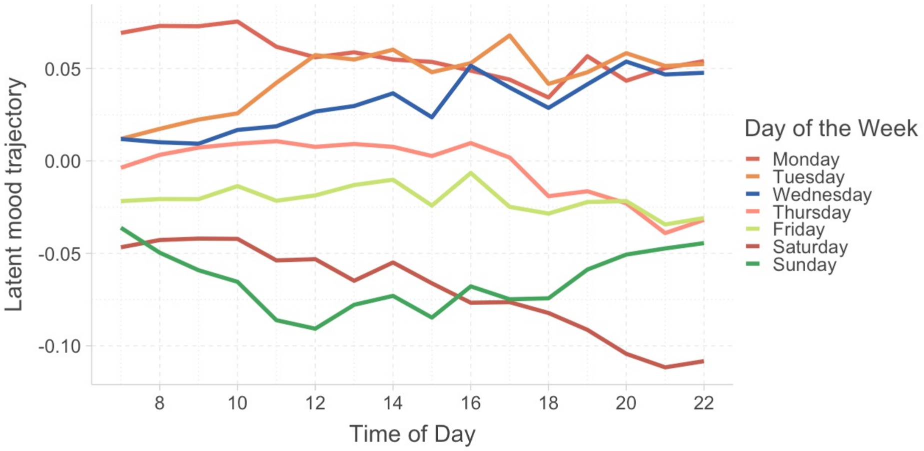
Predicted latent trajectory separates mood disorders





Predicted latent trajectory vs days of the week





Discussion

- General approach to perform Functional Principal Components analysis to any mixed-type functional data.

- Can be extended to model cross-covariances between mixed-type processes, which will extend to Generalized Function-on-regression and scalar-on-function regression for mixed-data types.

- Extensions involve adding day-specific random-effects and build a multilevel functional model for such intensive multilevel mixed-type longitudinal data in digital health.