Multivariate Principal Component Analysis for Mixed-Type Functional Data with application to mHealth

Debangan Dey
Postdoc Fellow
National Institute of Mental Health

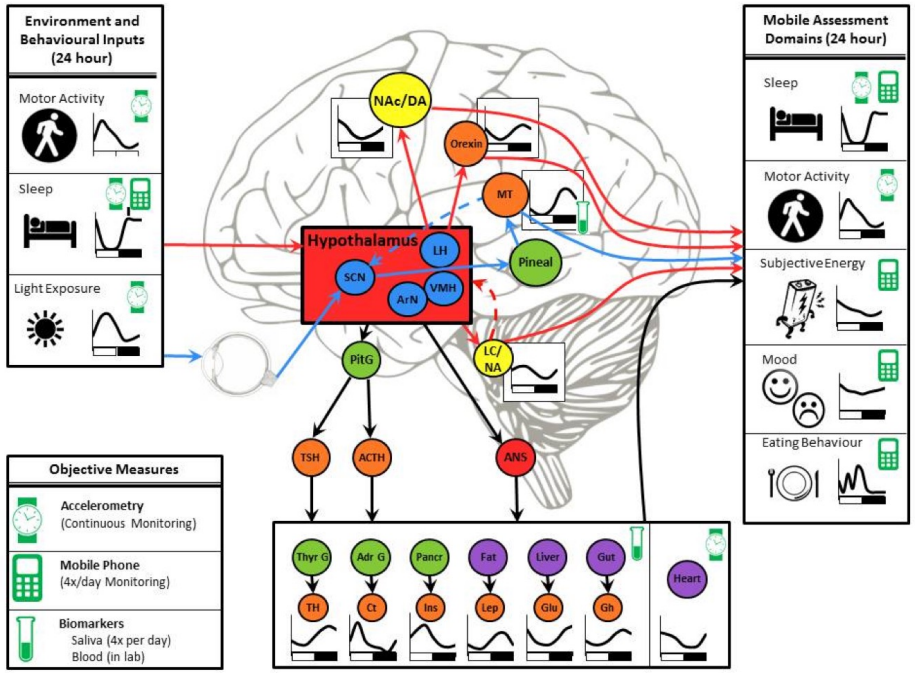
Joint work with
Rahul Ghosal,
Kathleen Merikangas, and
Vadim Zipunnikov

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.

Biological processes associated with regulation of homeostatic domains assessed by mHealth



Sleep  

Sleep  

Light 

Sleep 

Light 

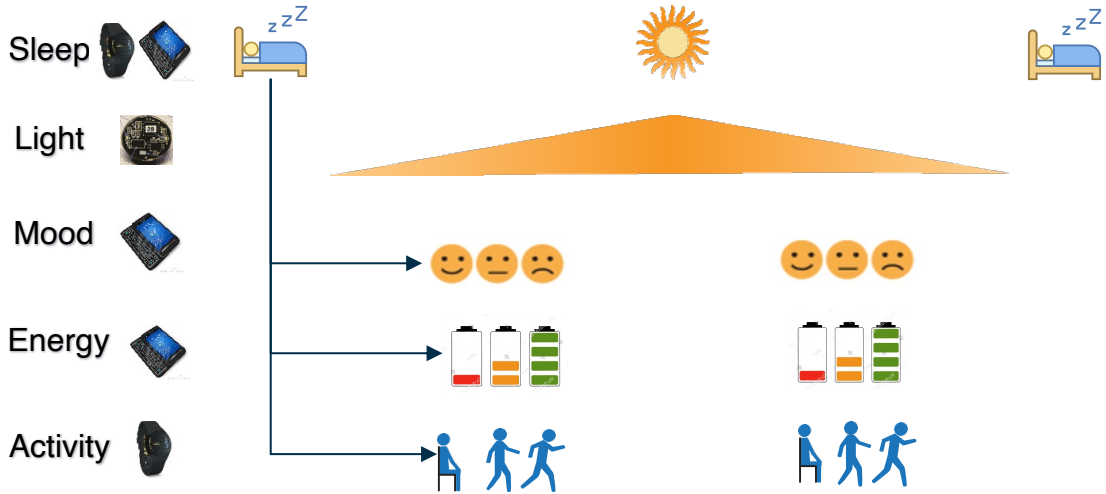
Mood 

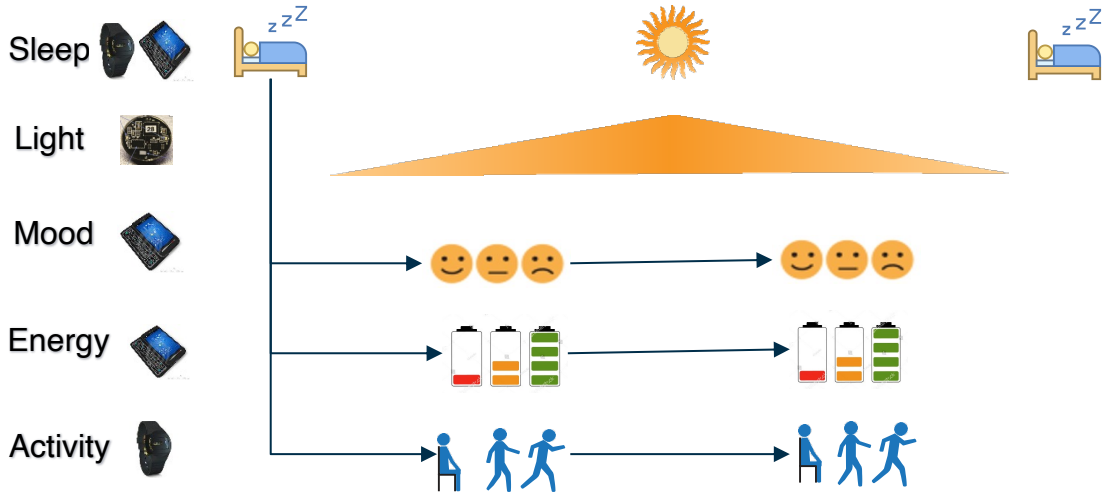
Sleep 

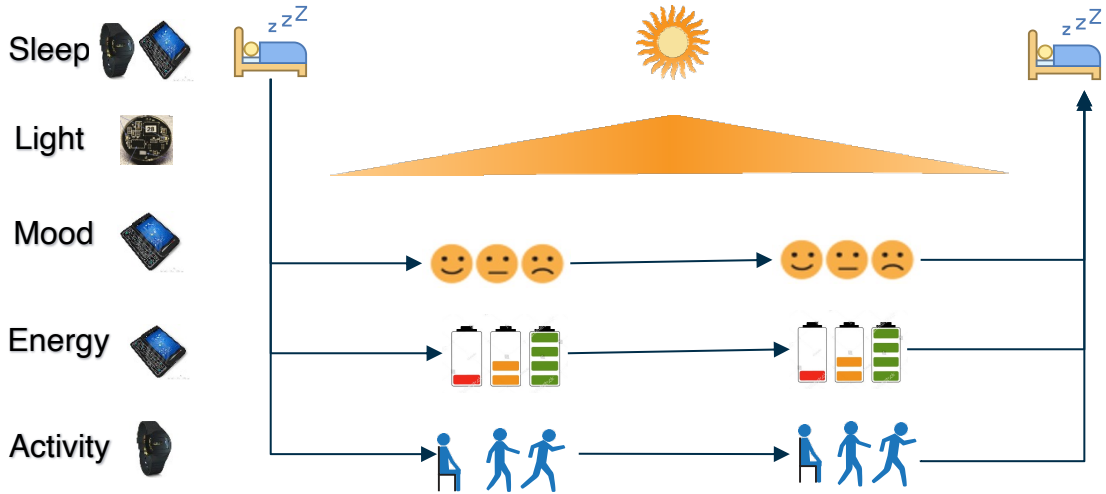
Light 

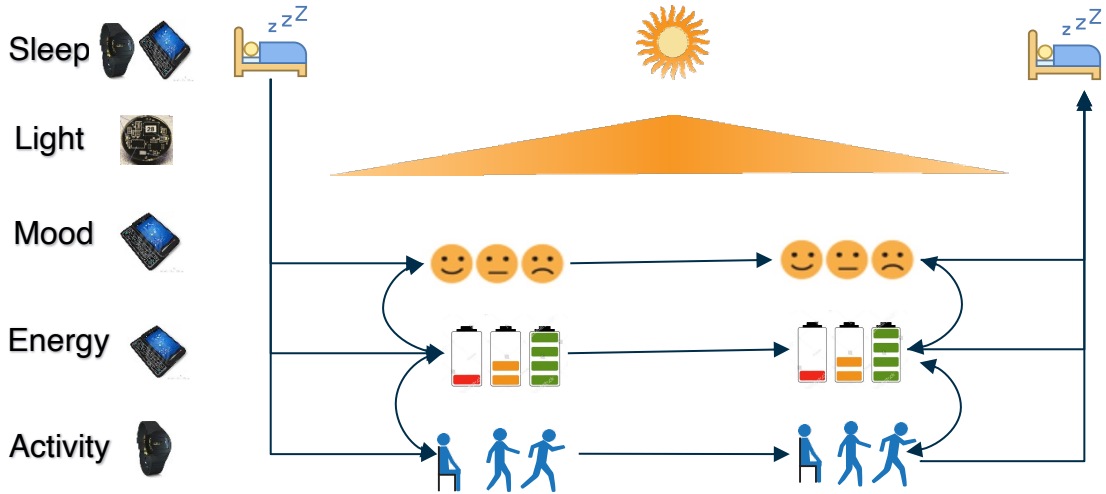
Mood 

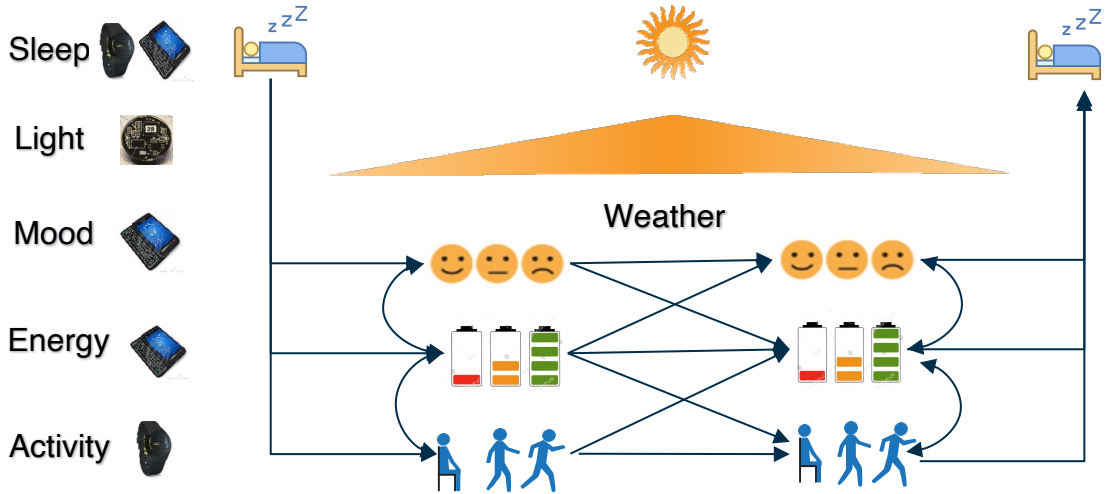
Energy 

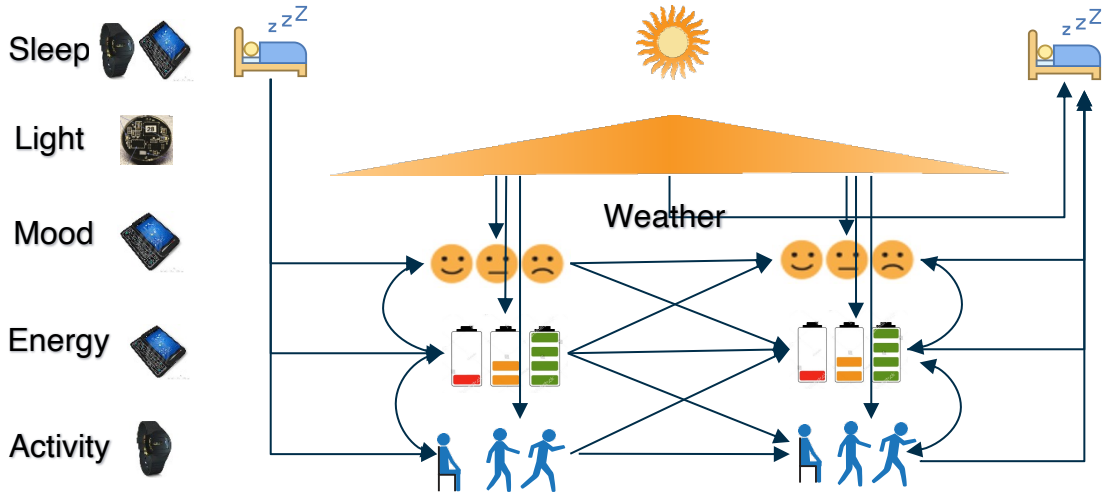

Sleep
Light
Mood
Energy
Activity

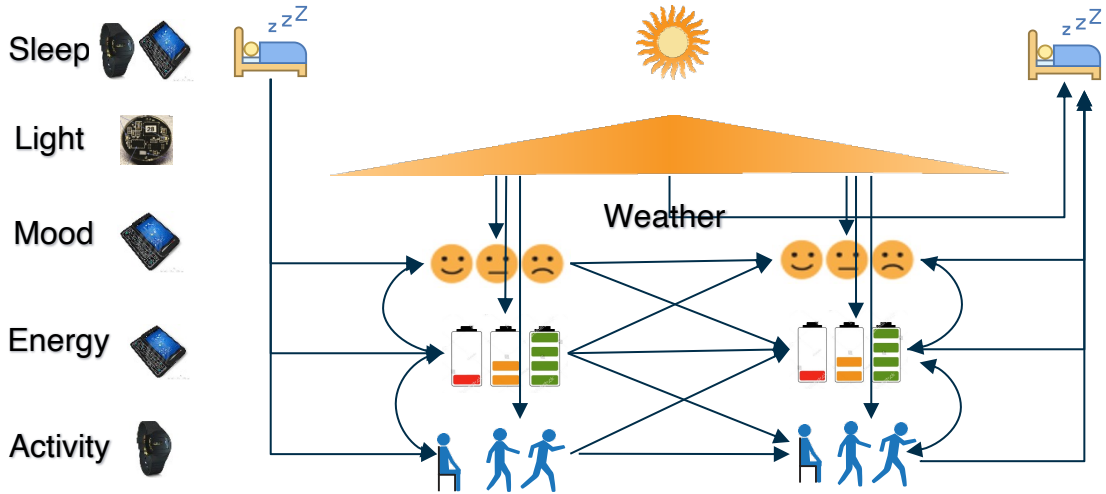

Sleep
Light
Mood
Energy
Activity


Sleep
Light
Mood
Energy
Activity


Sleep
Light
Mood
Energy
Activity


Sleep
Light
Mood
Energy
Activity
Weather

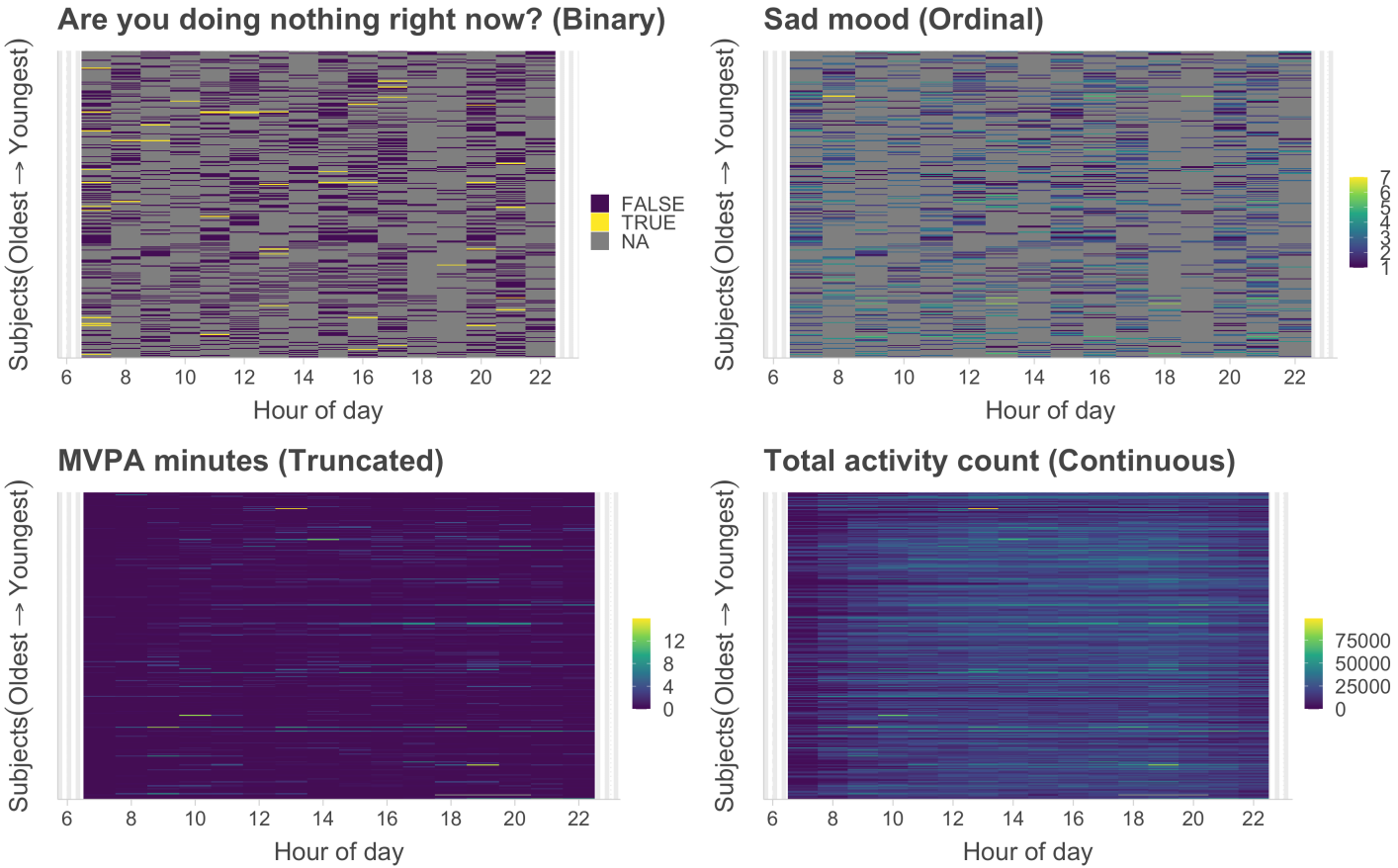

Sleep
Light
Mood
Energy
Activity
Weather



Multivariate Principal Component Analysis for Mixed-type functional data to understand cross- and inter-correlations between processes.

Challenges: mHealth

- 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.
- Mixed-type sparse data observed in misaligned time scales.
- Different measurement scales (binary, ordinal, truncated, continuous).
- Differences in subjective interpretation of scales.


Are you doing nothing right now? (Binary)
Subjects(Oldest → Youngest)
Hour of day
FALSE
TRUE
NA
Sad mood (Ordinal)
MVPA minutes (Truncated)
Total activity count (Continuous)
75000
50000
25000
0

What we need

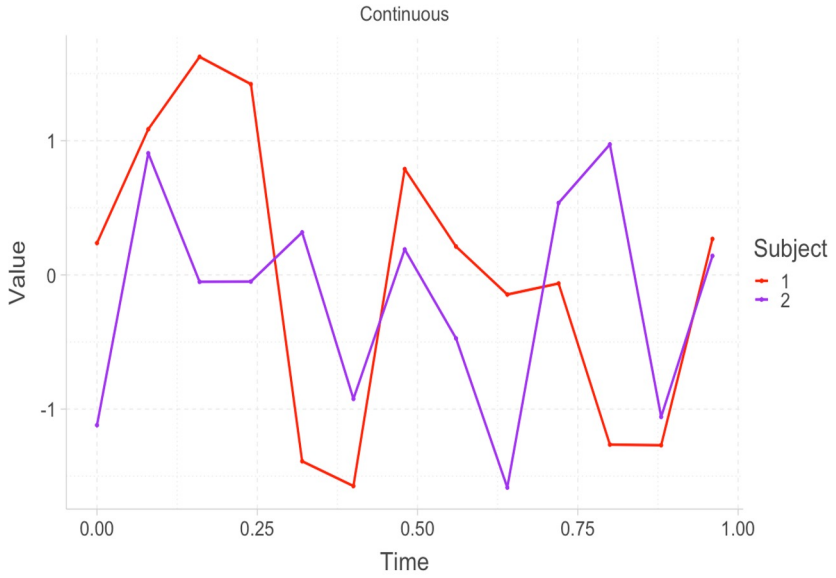
- Can we build a general modeling framework for joint modeling of binary, ordinal, continuous and truncated type functional data?

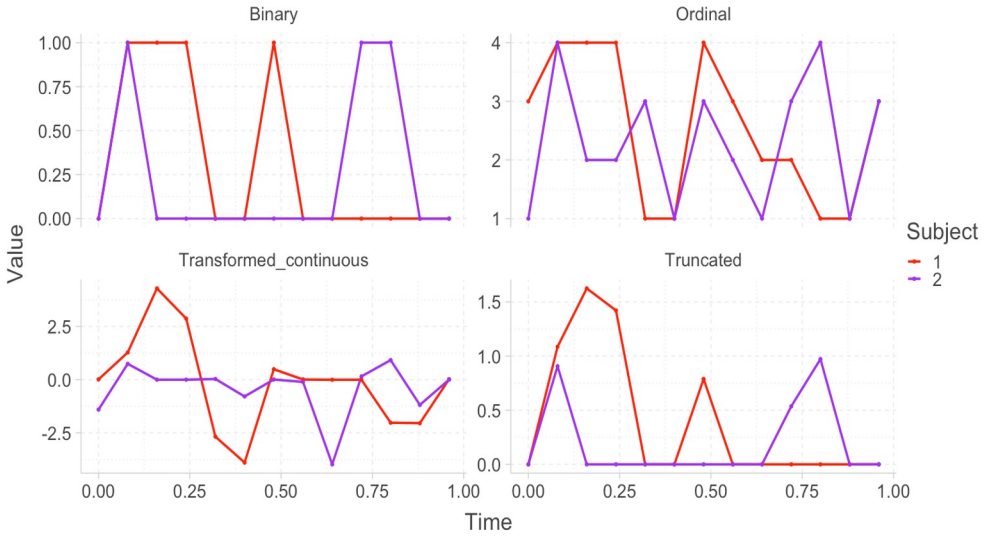
- Can we build Multivariate Functional Principal Component Analysis for such approaches?

Generalized Latent Non-paranormal Process

Latent Gaussian Process

Observed data





Latent Non-paranormal Distribution

Definition 1. (Nonparanormal distribution) A random vector $W = (W_1, \dots, W_p)'$ follows a non-paranormal distribution denoted as $W \sim NPN_p(0, \Sigma, f)$, if there exists monotone transformation functions $f = (f_1, ..., f_p)$ such that $L = f(W) = (f_1(W_1), \dots, f_p(W_p)) \sim N(0, \Sigma)$, with $\Sigma_{jj} = 1$ for $1 \leq j \leq p$. The constraints on diagonal elements of Σ are made to ensure the identifiability of the distribution as demonstrated in Liu et al.[22] and Fan et al.[26].

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

Let $X_{ij}(t); t \in \mathcal{T}$ be a function for jth variable measured over a continuous variable t within subject i for $i = 1, 2, \cdots, I$ and $j = 1, 2 \cdots, J$.

$$X_{ij}(t) = Z_{ij}(t) (\text{continuous scale}), \text{ or}$$

$$X_{ij}(t) = Z_{ij}(t) I(Z_{ij}(t) > \Delta(t)), (\text{truncated scale}), \text{ or}$$

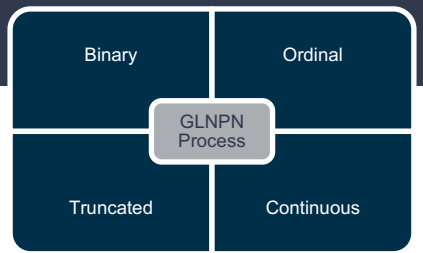
$$X_{ij}(t) = \sum_{k=0}^{l-1} k I(\Delta_{ijk}(t) \leqslant Z_{ij}(t) < \Delta_{ij(k+1)}(t)),$$

$$-\infty = \Delta_0(t) \leqslant \Delta_1(t) \leqslant \cdots \leqslant \Delta_l(t) = \infty, (\text{ordinal scale}), \text{ or}$$

$$X(t) = I(Z_{ij}(t) > \Delta(t)), (\text{binary scale}).$$

For multiple dependent outcomes $Z(\mathcal{T}) = (Z_1(\mathcal{T}), \ldots, Z_q(\mathcal{T}))^T$, specification is $(Z(\mathcal{T})) \sim NPN(0, C(\cdot, \cdot, f))$ where $C = (C_{ij})$ is the $J \times J$ multivariate covariance function.

Univariate Marginal Covariance Estimation*



Method of moments

Kendall's Tau

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

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

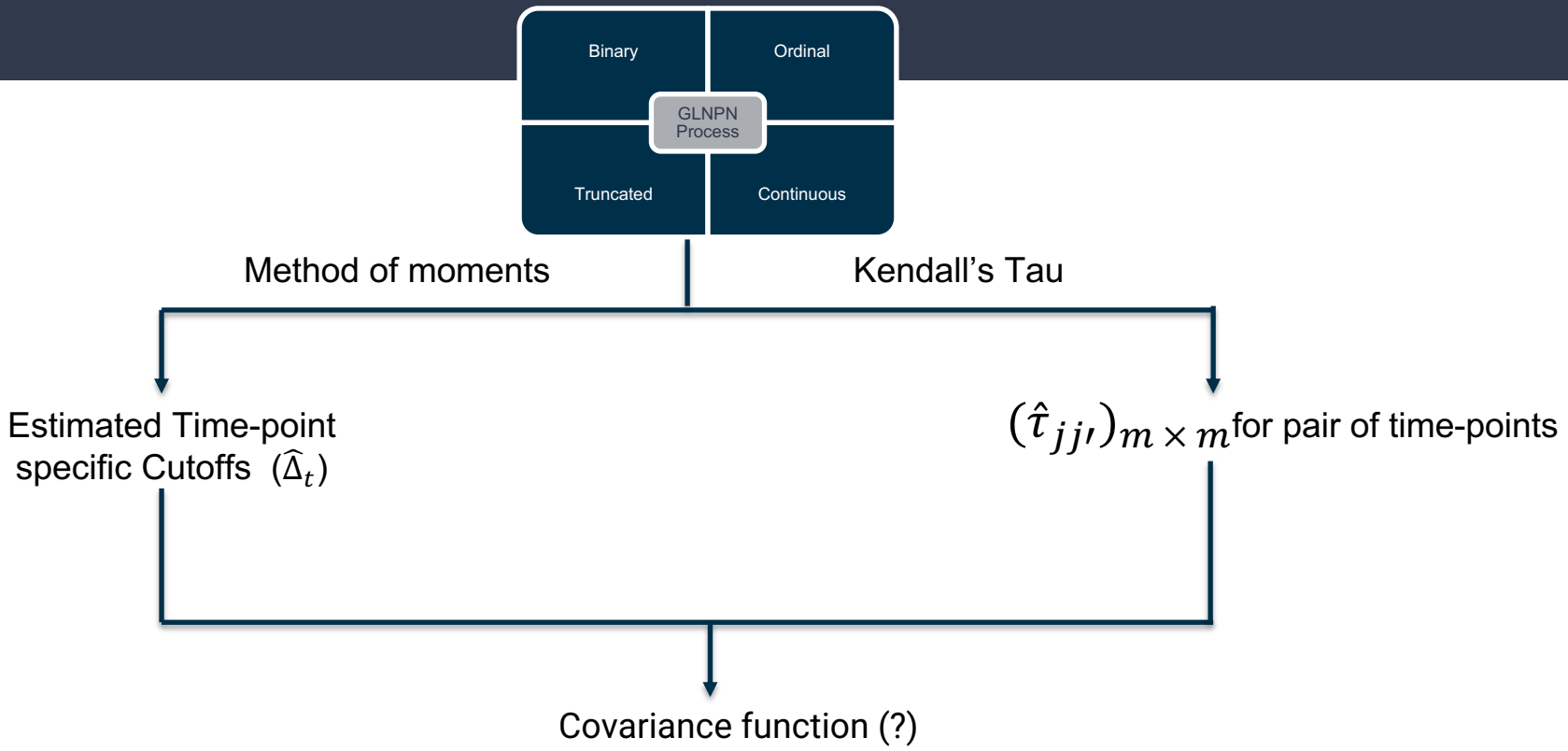
$(\hat{\tau}_{jj'})_{m \times m}$ for pair of time-points

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

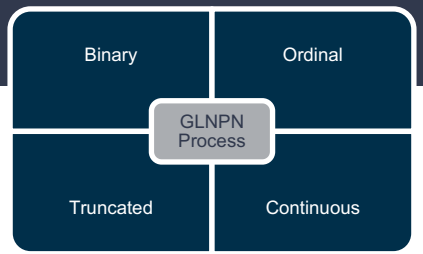
Kendall's Tau is invariant under monotone increasing transformation

*Functional Principal Component Analysis for Continuous non-Gaussian, Truncated, and Discrete Functional Data,
Dey, Ghosal, Merikangas, Zipunnikov, 2024

Univariate Marginal Covariance Estimation



Univariate Marginal Covariance Estimation



Method of moments

Kendall's Tau

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

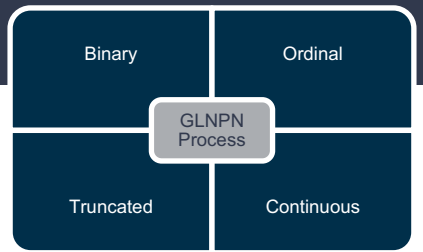
$(\hat{\tau}_{jj'})_{m \times m}$ for pair of time-points

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

Covariance function (?)

Model the scaled covariance function as a tensor product spline.

Univariate Marginal Covariance Estimation



Method of moments

Kendall's Tau

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

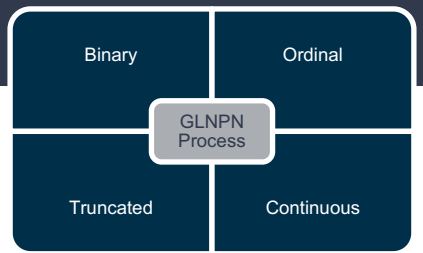
$(\hat{\tau}_{jj'})_{m \times m}$ for pair of time-points

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

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

Covariance function

Univariate Marginal Covariance Estimation



Method of moments

Kendall's Tau

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

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

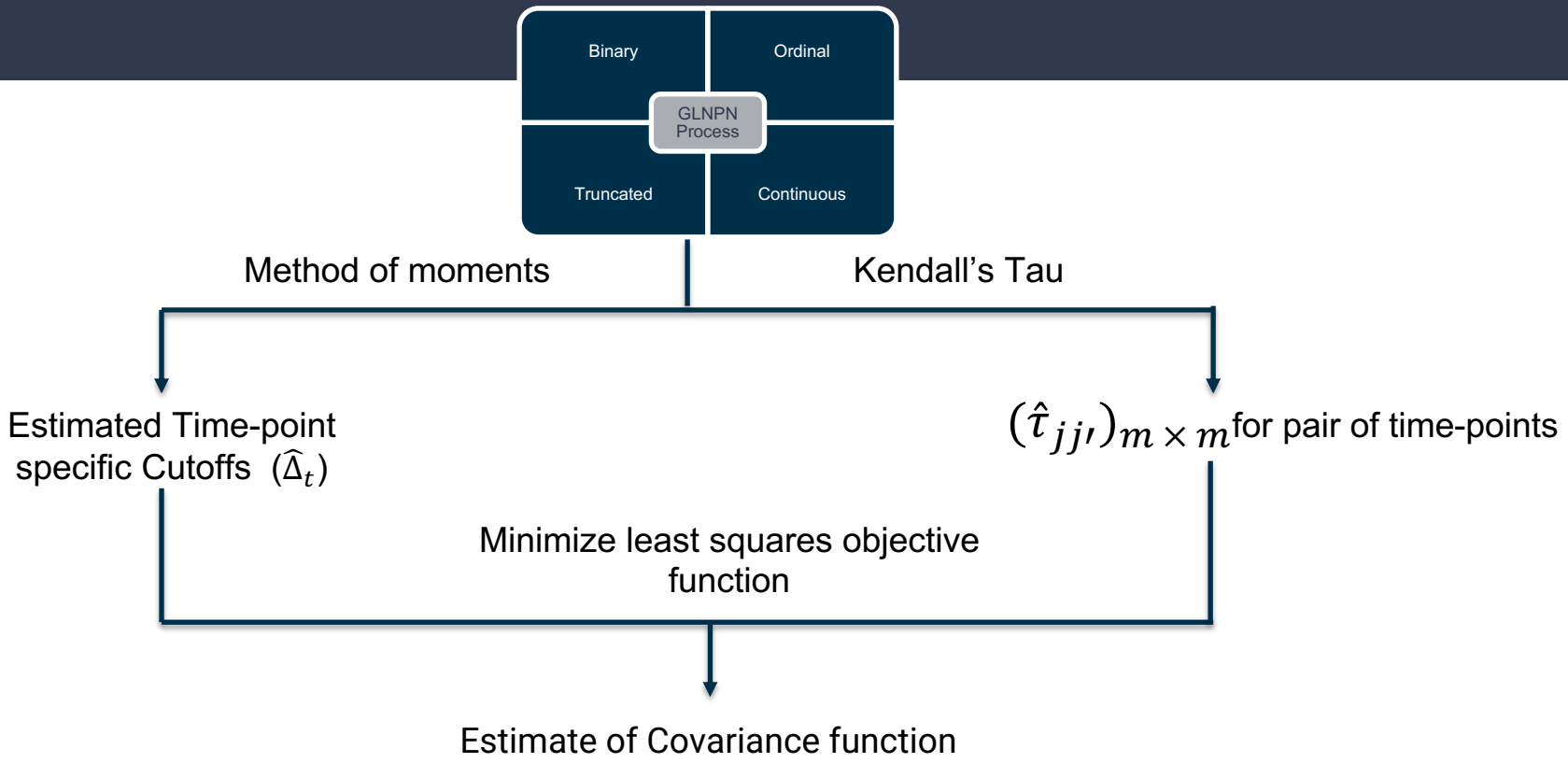
$$\hat{U} = \underset{U}{argmin} \sum_{1 \leqslant j < j' \leqslant m} (\hat{\tau}_{jj'} - F(g(\sum_{k=1}^{d}\sum_{l=1}^{d} u_{kl} B_k(t_j) B_l(t_{j'}))))^2$$

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).

Univariate Marginal Covariance Estimation



Full Covariance Estimation

- Estimate marginal covariance matrices (Σ_{jj} blocks).
- Following similar steps as marginal estimation, estimate cross-covariance (Σ_{jk} blocks) (use cross-variable bridging function)
- Use nearPD* function in R package Matrix to find the nearest correlation matrix made of the above blocks.
- Needs to estimate $O(J^2)$ covariances. Could be slow.

*Higham, Nick (2002) Computing the nearest correlation matrix - a problem from finance; IMA Journal of Numerical Analysis 22, 329--343.

Full Covariance Estimation (Faster)

- Estimate marginal covariance matrices ($\Sigma\_\{jj\}$ blocks).
- We can use conditional expectation to predict latent continuous variables for each marginal process.
- Treat these predictions as pseudo-observations (continuous) to perform partially separable* or full multivariate principal component analyses.
- Partial separable assumption can give us J vector principal component scores for desired number of levels and can be used to define graphical model between variables.

*Partial separability and functional graphical models for multivariate Gaussian processes By J. Zapata, S. Y. Oh and A. Petersen, Biometrika, 2022

Resulting methods

1. Function-on-function regression

$$E(V_1(s)|V_2(T)) \quad = \Sigma_{12}(s,T)\Sigma_{22}(T,T)^{-1}(V_2(T))$$

In our case, we can regress mood ratings on concurrent and past physical activity from the joint covariance structure.

2. Multivariate functional principal component scores

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.
- We provide asymptotical confidence intervals for regression coefficients and covariance values.

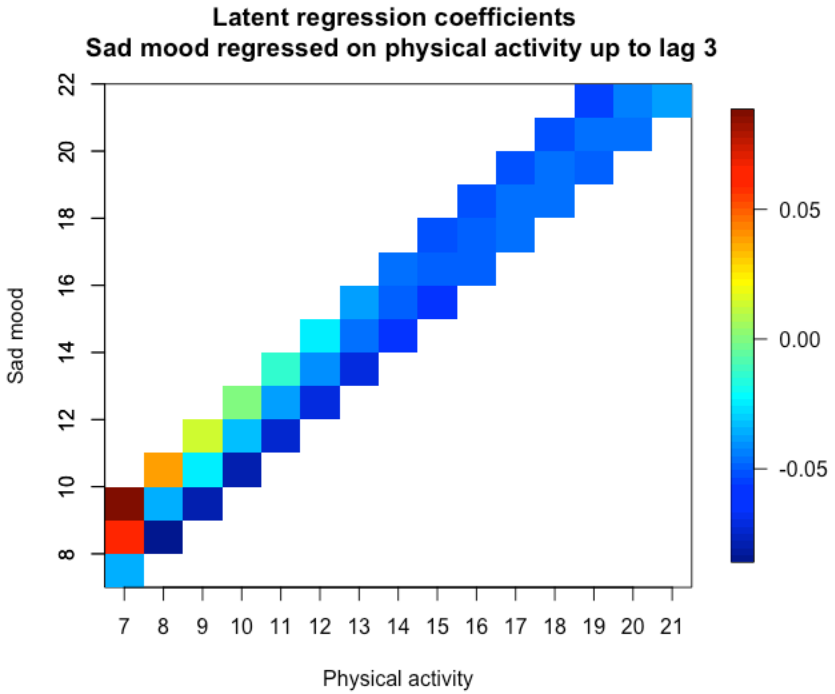
Data analysis: NIMH Family Study of Affective Disorders

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	

Mood on current and past physical activity



Discussion

- Our approach provides a unified, general solution for modeling multivariate mixed-type functional data.
- Since we use rank-based estimators, our approach is robust.
- Next, we want to extend to handle multilevel structure of the data and take care of within-subject correlation.

Thank you.
Questions?