Digital Phenotyping for Mixed-Type mHealth Data

M^2FPCA — Multivariate functional principal component analysis for mixed-type functional data

Debangan Dey

Department of Statistics, Texas A&M University

IBC 2026, Seoul · Invited Session IS.08.03 · July 14, 2026

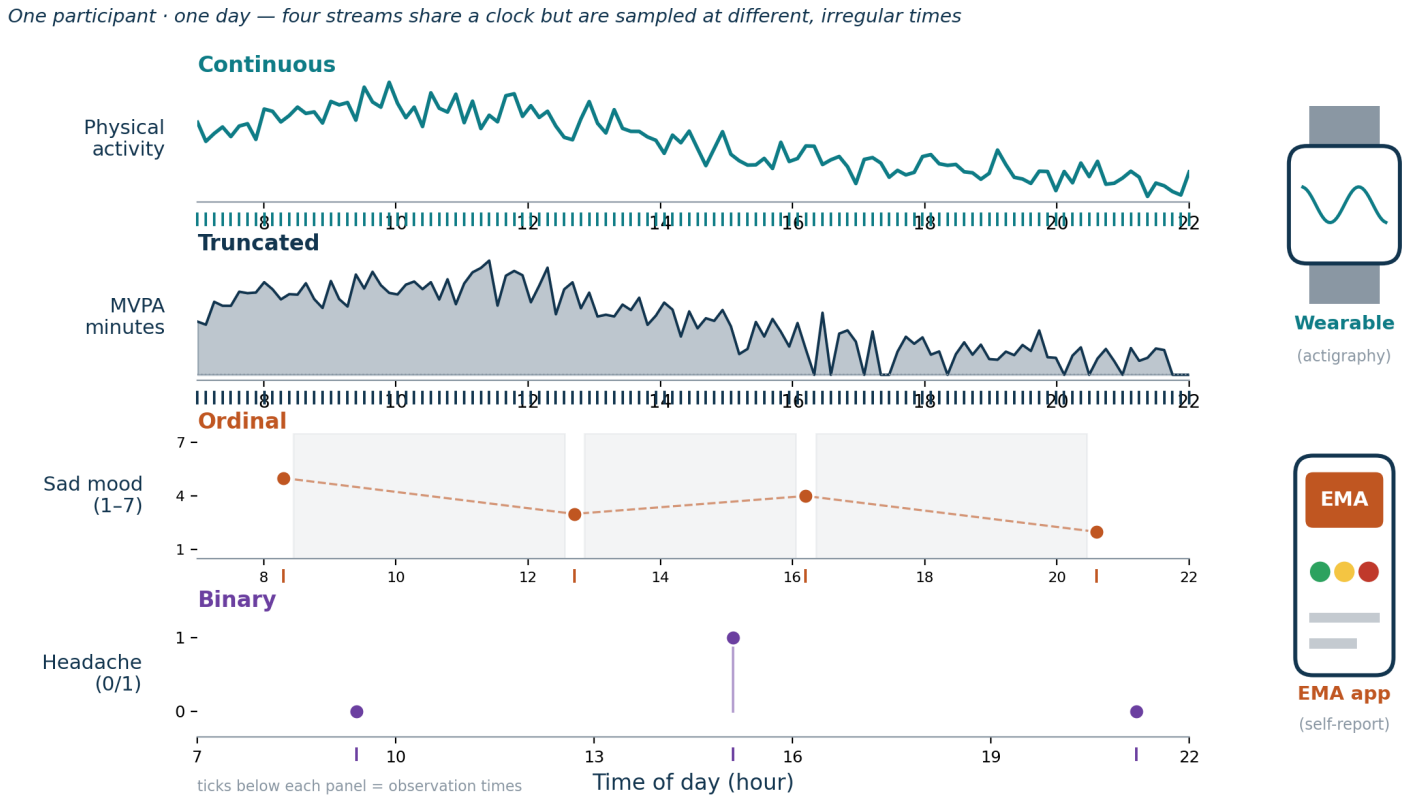
Disclosures

Conflict of interest. I have no current or past relationships with commercial entities.

Funding. Supported by grant ZIA MH002954, Motor Activity Research Consortium for Health (mMARCH), National Institutes of Health / National Institute of Mental Health Intramural Research Program. The views expressed are those of the author and do not represent the views of the U.S. Government.

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

Motivation: mHealth and digital phenotyping



EMA (Ecological Momentary Assessment): brief in-the-moment self-reports (mood, anxiousness, energy) collected on a smartphone a few times a day — **sparse and asynchronous.** Wrist actigraphy adds a dense, continuous activity stream. Together: mixed-type multivariate functional data.

Motivating questions

1. How do mood, anxiousness, energy, and activity co-vary across the day?
2. Can each person's diurnal profile be summarized into interpretable digital biomarkers?
3. Do within-day dynamics distinguish mood-disorder subtypes — bipolar I, bipolar II, MDD?

Goal: a unified framework for mixed-type multivariate functional data — enabling digital phenotyping and mood-disorder subtype stratification.

The statistical problem

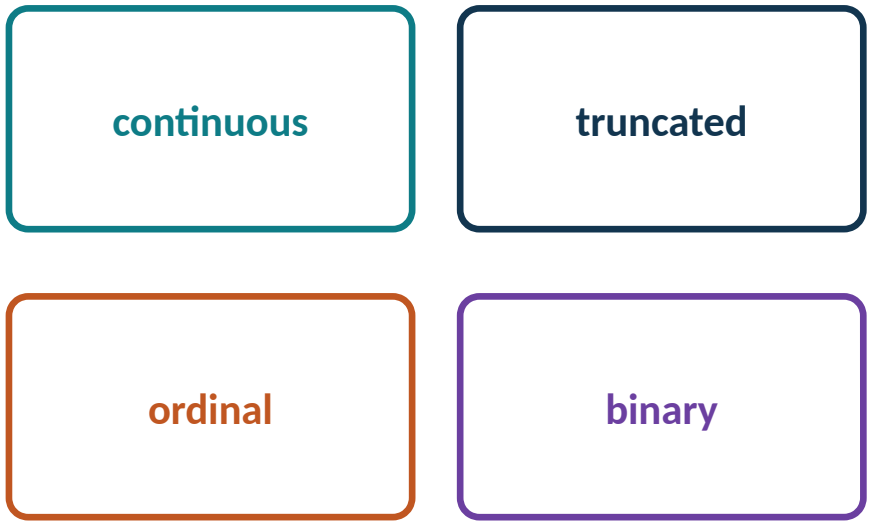
We observe a J-variate functional process

$$X(\cdot) = \{X_1(\cdot), \ldots, X_J(\cdot)\}, \quad X_j \in L^2(\mathcal{T})$$

- Each component may be continuous, truncated, ordinal, or binary.
- Sampled on irregular, component-specific grids.

One framework must:

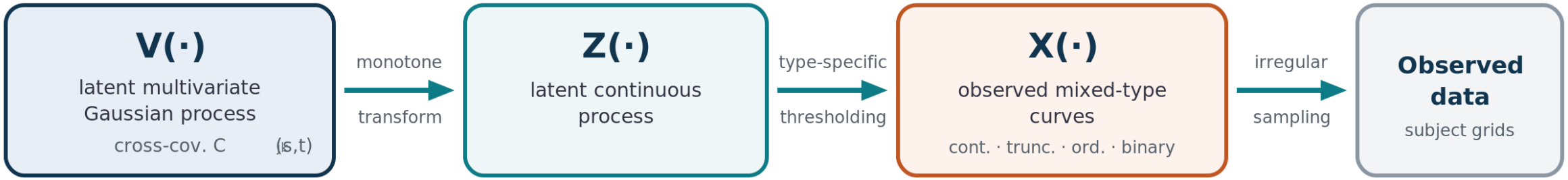
- jointly model all four measurement scales on a common domain
- estimate within- and cross-domain temporal dependence
- yield interpretable low-dimensional scores (biomarkers)
- remain estimable under sparse, asynchronous sampling



four mixed types, one clock

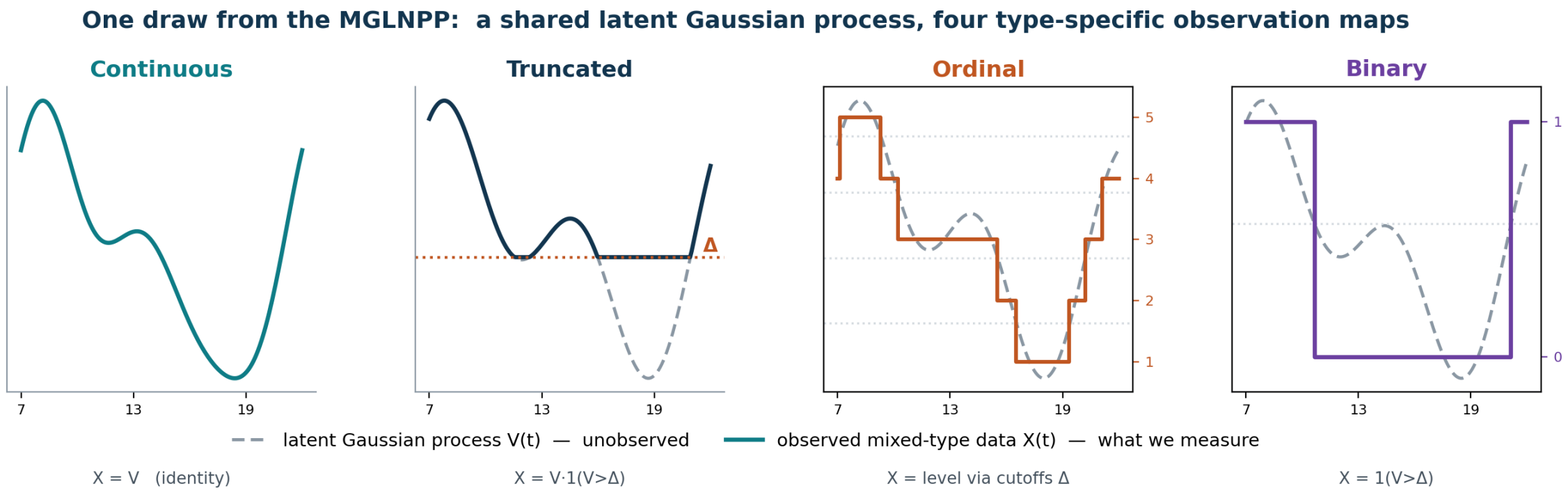
The modeling framework (MGLNPP)

Observed mixed-type curves are generated from a shared latent Gaussian process.



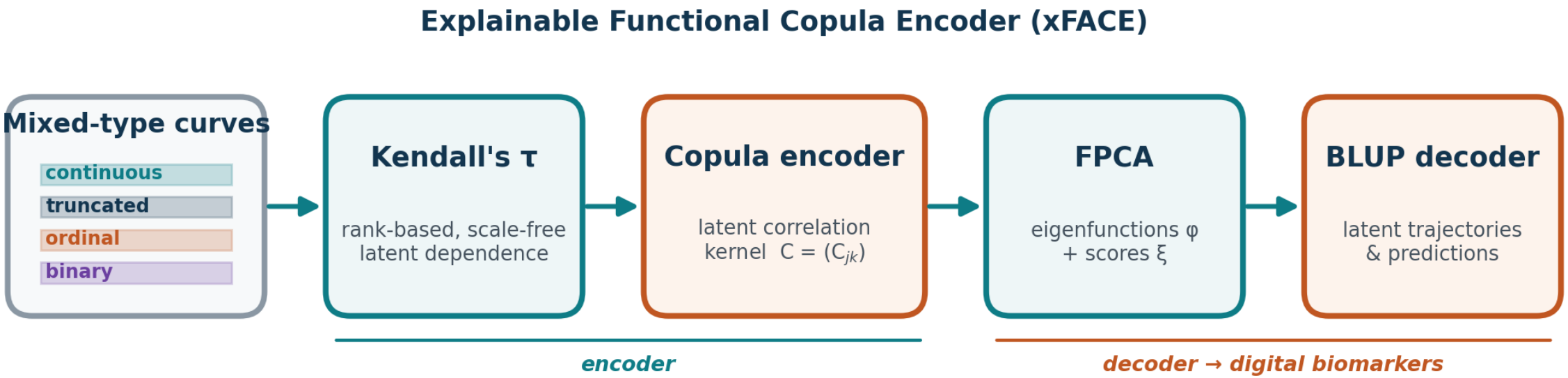
Cross-domain dependence is carried entirely by the latent Gaussian process V. Extends the univariate GLNPP of Dey et al. (Statistics in Medicine, 2024) to the multivariate, mixed-type setting.

From latent to observed: one draw from the MGLNPP



The latent Gaussian process V(t) is unobserved (grey, dashed). Type-specific observation maps produce what we actually measure **(colored)**: identity, truncation, ordinal cutoffs, or a binary threshold — the same latent dependence seen on four different scales.

The method at a glance



Mixed-type curves → rank-based Kendall's τ → copula encoder (latent correlation kernel) → functional PCs → BLUP decoder → latent trajectories and FPC scores (digital biomarkers).

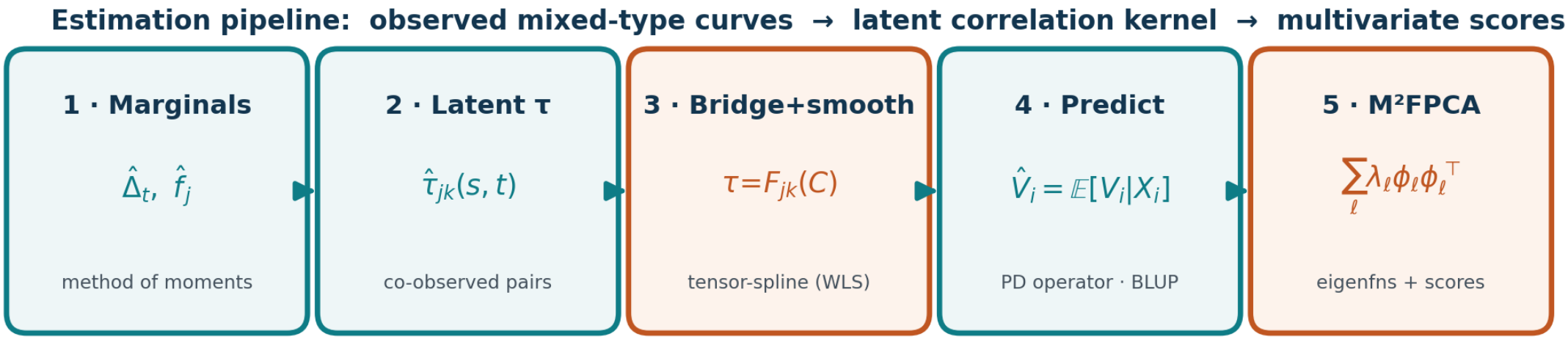
Estimation: the core idea (Kendall's τ)

Target: the latent correlation kernel $C = (C_{jk})$. Because X_j may be discrete or truncated, we use the invariance of Kendall's τ to monotone transformations:

$$\tau_{jk}(s,t) = F_{jk}\{ C_{jk}(s,t) \}$$

- F_{jk} — a type-specific bridging function for each pair of margins (we derive the missing ordinal × {ordinal, truncated, binary} bridges).
- $\hat{\tau}_{jk}(s,t)$ is computed only over concurrently observed subject pairs — the key to handling irregular, asynchronous sampling.

Estimation in five steps

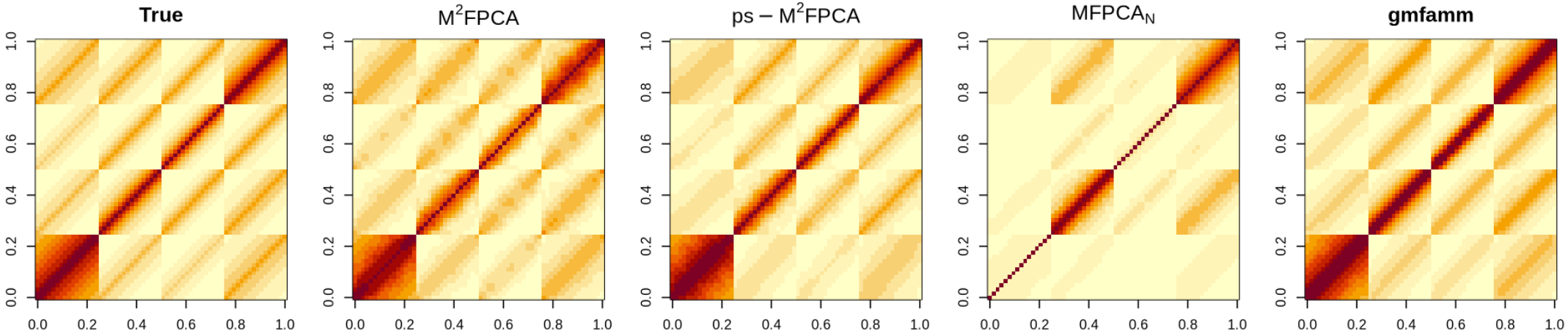


Two covariance estimators (steps 2–5)

M²FPCA — full blockwise cross-covariance $\{C_{jk}\}$; multivariate BLUP borrows strength across all domains.

ps-M²FPCA — shared temporal eigenbasis (partial separability); reuses univariate M¹FPCA, far cheaper.

Simulation: recovering the latent covariance



Correlation surfaces at n = 500 (four-domain MGLNPP; m = 16; 100 Monte-Carlo runs).

- Both proposed estimators recover the true surface; naive MFPCA collapses off-diagonal structure.
- Integrated squared error is 4–20× lower than naive MFPCA$_N$, and below the mixed-type gmfamm competitor.

Robustness to sparse, asynchronous sampling

Component curves are thinned by independent, component-specific masks — dense actigraphy vs. sparse EMA.

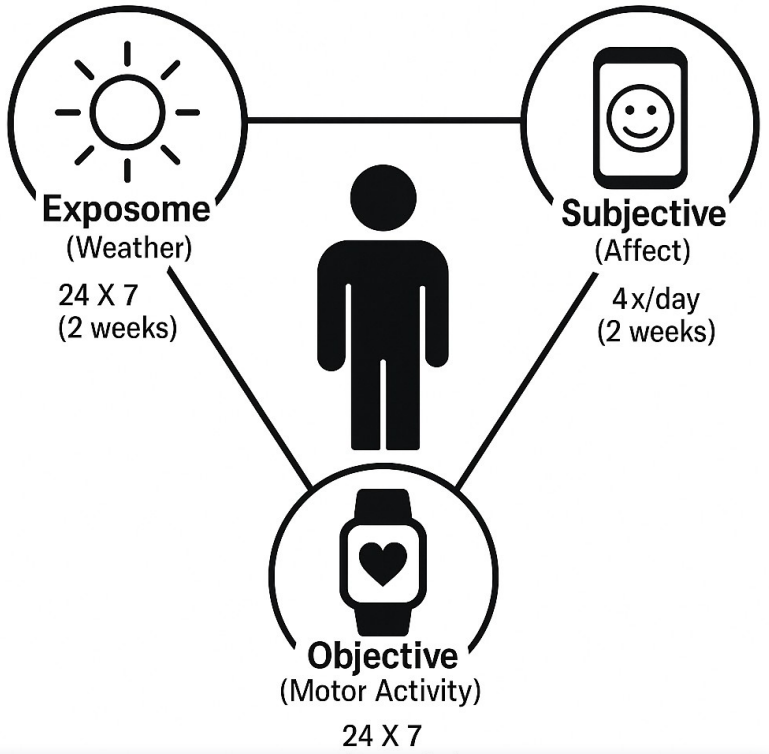
	Covariance ISE			Held-out prediction MSE		
Sampling regime	**M²FPCA**	**ps-M²FPCA**	**MFPCAN**	**M²FPCA**	**ps-M²FPCA**	**MFPCAN**
Bernoulli (0.3)	**0.0042**	0.0113	0.0328	0.579	**0.571**	0.676
Bernoulli (0.6)	**0.0021**	0.0051	0.0292	0.476	**0.458**	0.545
Poisson (activity vs. EMA)	**0.0050**	0.0107	0.0339	0.653	**0.637**	0.776
EMA-calibrated	**0.0068**	0.0115	0.0343	0.653	**0.626**	0.768

M²FPCA attains the smallest covariance ISE and ps-M²FPCA the smallest held-out prediction error in every regime. The gmfamm competitor is not estimable under sparse EMA.

n = 1000, non-stationary kernel; one-hour coarsening grid, c_0 = 30 co-observed subjects per cell.

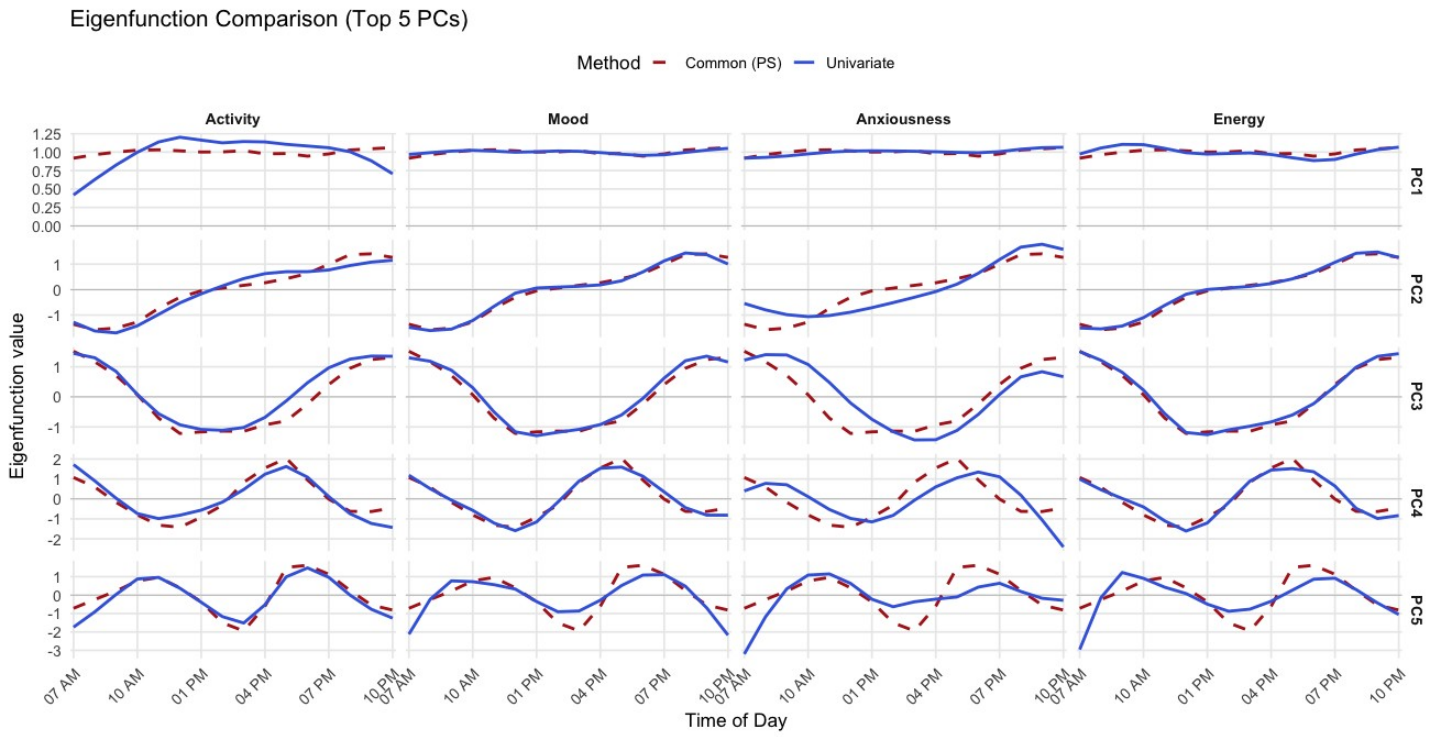
Application: NIMH Family Study of Mood Disorders

- N = 307 participants (ages 11–85): controls (130), MDD (106), bipolar I (41), bipolar II (30).
- 14 days of wrist actigraphy (minute-level, GGIR) + four EMA prompts/day (sad mood, anxiousness, energy).
- Prompts scheduled to each participant's wake–sleep cycle → information beyond four fixed points.
- Harmonized to a common 16-bin, 7 a.m.–10 p.m. grid → a four-dimensional daily function.
- 3,474 subject-days; 93% EMA completion.



Higher activity is inversely associated with sad mood and anxiousness, positively with energy; sad mood and anxiousness are strongly co-elevated.

Shared diurnal components across domains



- **Top three ps-M²FPCA components explain 82% of variability:**
- FPC1 (64%) — overall daily burden / activation.
- FPC2 (12%) — morning-to-evening contrast (homeostatic).
- FPC3 (6%) — midday inverted-U (circadian).

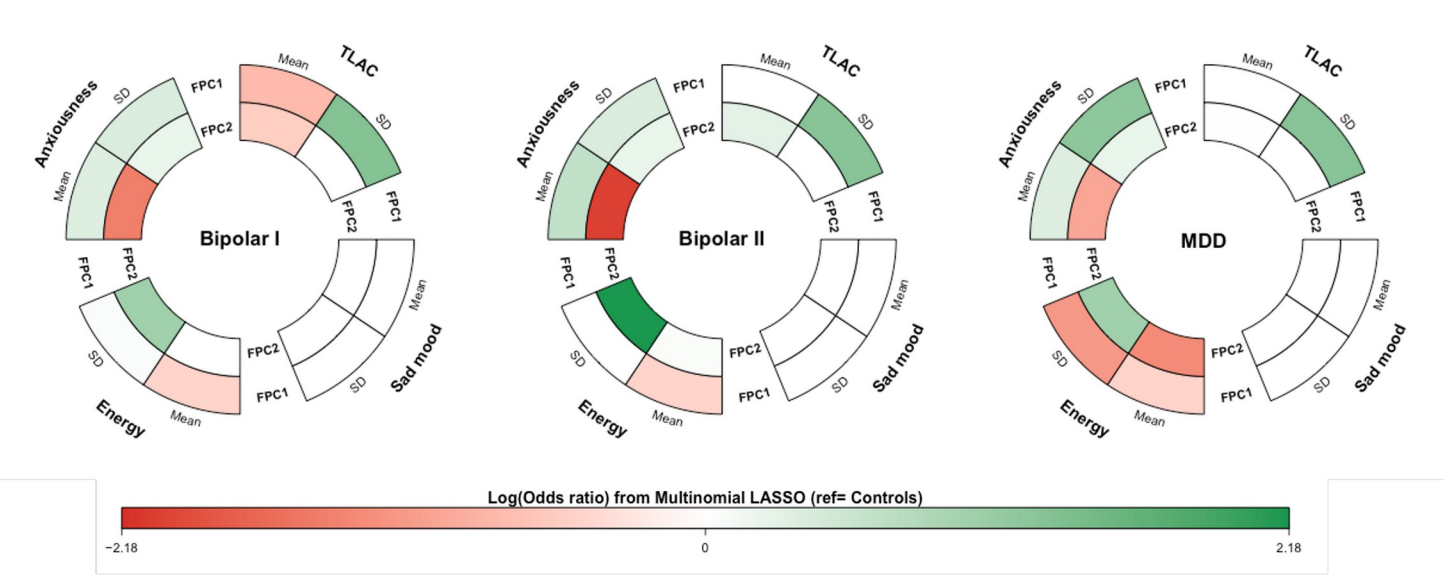
Univariate and shared eigenfunctions agree closely — the separability assumption holds.

Multivariate prediction borrows strength across domains

- Held-out EMA categories, interpolated from the multivariate covariance (orange) vs. each domain's univariate model (green).
- A curve recovers a category when it crosses the observed interval (bars).

Joint modeling reduces latent-scale prediction error over the univariate baseline by **16.2%** (sad mood), **11.0%** (anxiousness), **7.4%** (energy).

Digital biomarkers stratify mood disorders



16 participant-level FPC means and SDs → multinomial LASSO (log-odds vs. controls).

- Anxiousness informative across all disorders.
- Activity (TLAC) FPC1/FPC2 means negatively associated with bipolar I; bipolar II differs in sign.
- Energy — attenuated diurnal gradient and reduced variability uniquely flag MDD.
- Sad-mood severity alone is not a dominant discriminator.

Nested cross-validation + 400 bootstraps: core biomarkers selected in >90% of fits — reproducible.

Software: the M2FPCA R package

Install & fit

```
# install (SGCTools pulled automatically)
remotes::install_github("Ddey07/M2FPCA")

library(M2FPCA)
type <- c("cont", "bin", "ord")

# joint fit: marginal + cross-cov,
# shared eigenbasis + score covariances
fit <- mpfpca.dir(dat_list, type = type,
         argvals, df = 5, weights = TRUE)

# subject multivariate FPC scores
sc  <- mpfpca_scores(dat_list, type,
         fit, argvals, npc = 3)
```

Worked example — recovers a known truth

- Simulate p = 3 mixed-type variables — continuous, binary, ordinal — sharing a latent structure.
- Marginal + cross-covariance recovered: relative Frobenius error 0.18.
- Handles irregular / sparse sampling (pass NAs; min_no_pairs cutoff).

```
round(cor(sc$scores[[1]]), 2)  # true = 0.50
#>        V1_PC1 V2_PC1 V3_PC1
#> V1_PC1   1.00   0.43   0.46
#> V2_PC1   0.43   1.00   0.40
#> V3_PC1   0.46   0.40   1.00
```

Key functions: **mpfpca.dir** (joint fit) · **mpfpca_scores** (scores) · **pfpca_crosscov** (cross-cov) · **cov.from.mpfpca.dir** · **mpfpca** (ps-M²FPCA variant).
github.com/Ddey07/M2FPCA

Summary and outlook

- M^2FPCA synthesizes mixed-type mHealth streams into interpretable, participant-specific digital phenotypes — without discarding measurement scale.
- A latent-Gaussian (MGLNPP) model, a copula-based Kendall's τ covariance estimator, and multivariate FPCA — with a scalable ps-M^2FPCA variant.
- Robust to sparse, asynchronous sampling; joint modeling improves prediction and yields reproducible diagnostic signal.

Next: a multilevel M^3FPCA for the subject-day hierarchy — a foundation for this session's progression from digital phenotyping toward causal discovery and intervention effects.

Thank you

Questions?

email me: debangan@tamu.edu

Debangan Dey · Texas A&M University

R package: github.com/Ddey07/M2FPCA

Dey, Ghosal, Merikangas & Zipunnikov — Multivariate functional PCA for mixed-type mHealth data (M^2FPCA), (arXiv:2603.11385). Builds on Statistics in Medicine (2024).