

# Wearables: other functional approaches

JSM 2019

# Overview

- Multi-level functional methods
- Matrix-variate multi-level functional methods
- Multi-level methods for generalized (e.g binary) functional curves
- Registration of generalized (e.g. binary) functional curves

# Multi-level functional principal component analysis (MFPCA)

## NIMH Family Study of Affective Spectrum Disorders

- 350 participants; ages from 10 to 84
- 5 diagnosis groups: BPI, BPII, MDD, Other and Control

Data:

- 2 weeks of follow-up measurements
- minute-by-minute activity counts
- 4945 person-days; 7,120,800 data points

Goal:

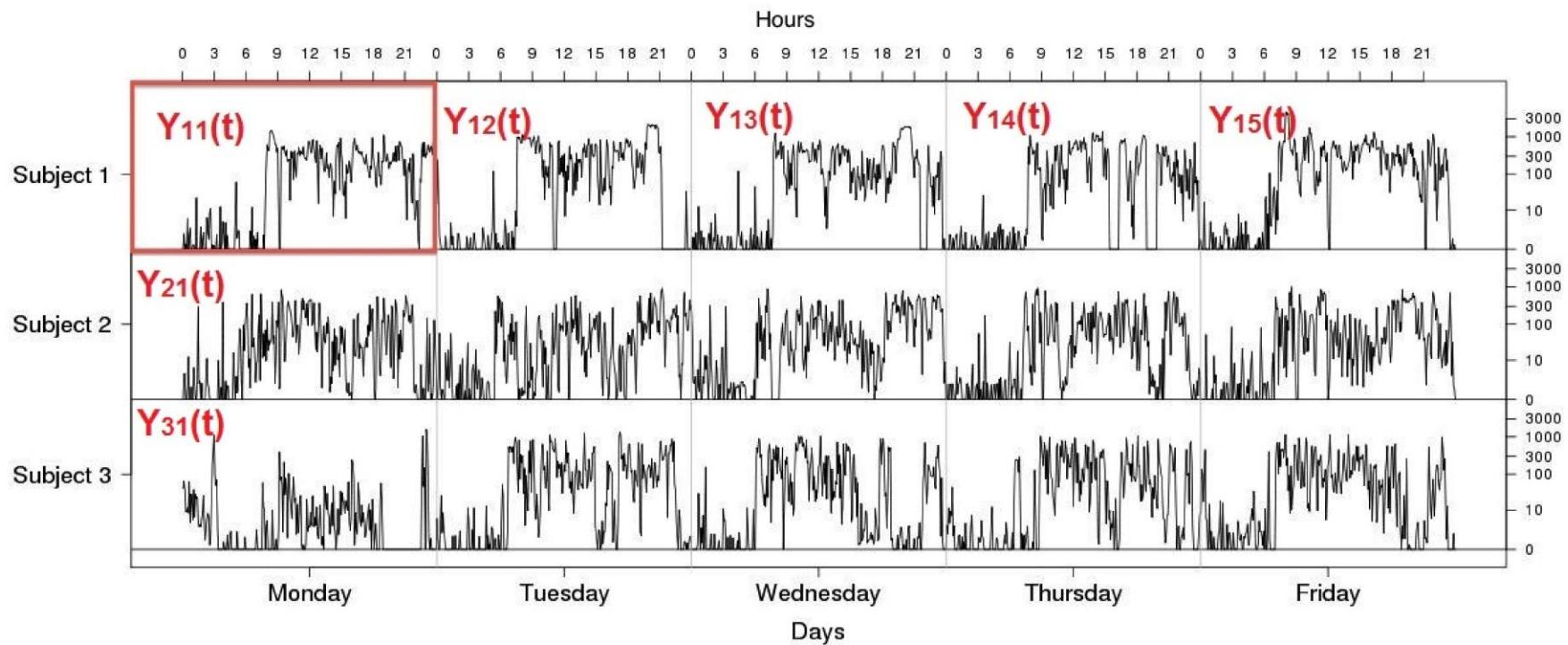
Extract representative patterns that comprise daily activity

Quantify multilevel variations after adjusting age and disease effects

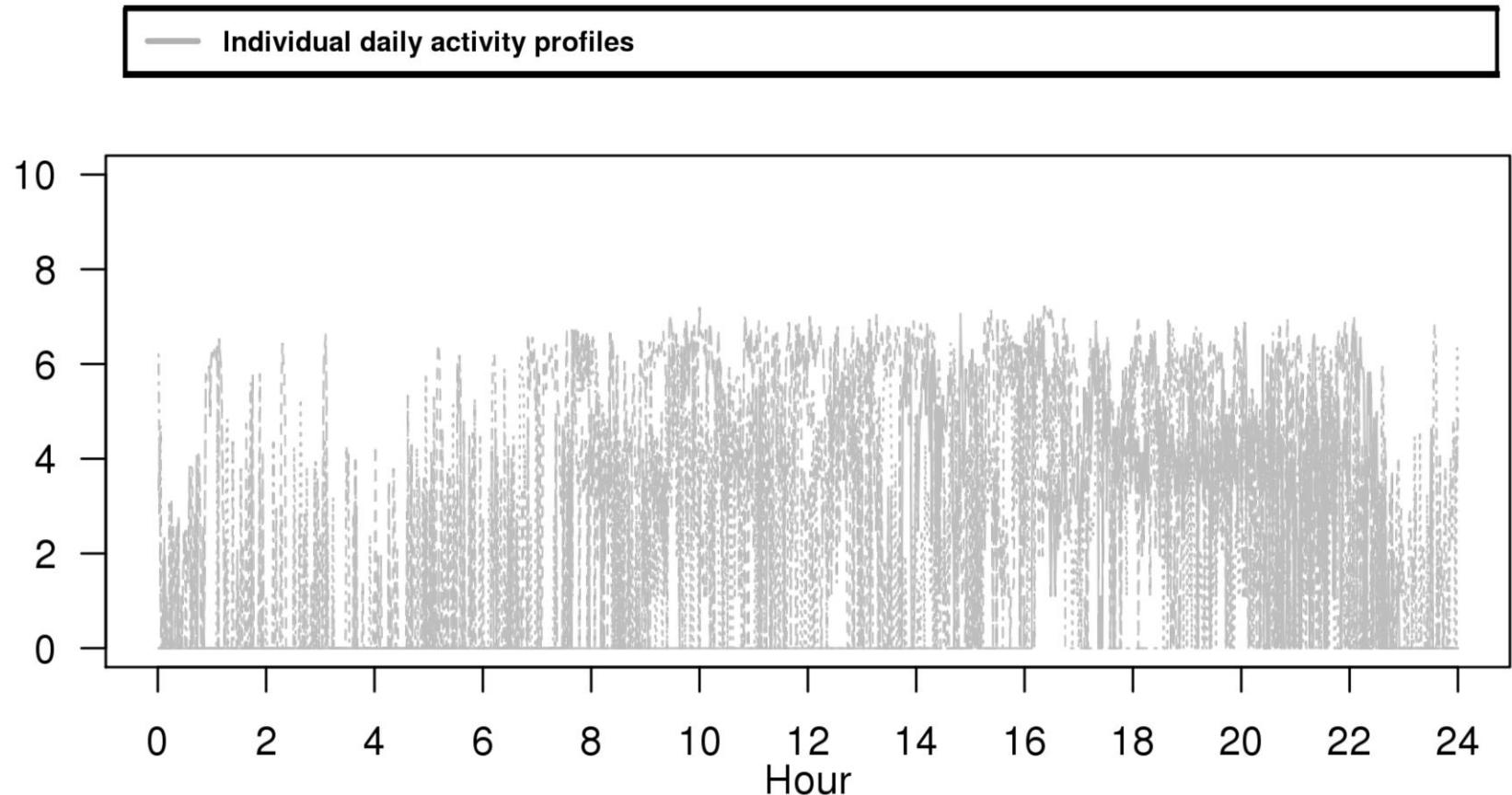
- Subject heterogeneity
- Day-to-day variation
- Age and disease effects in mean and variance components

# Daily Activity Profiles

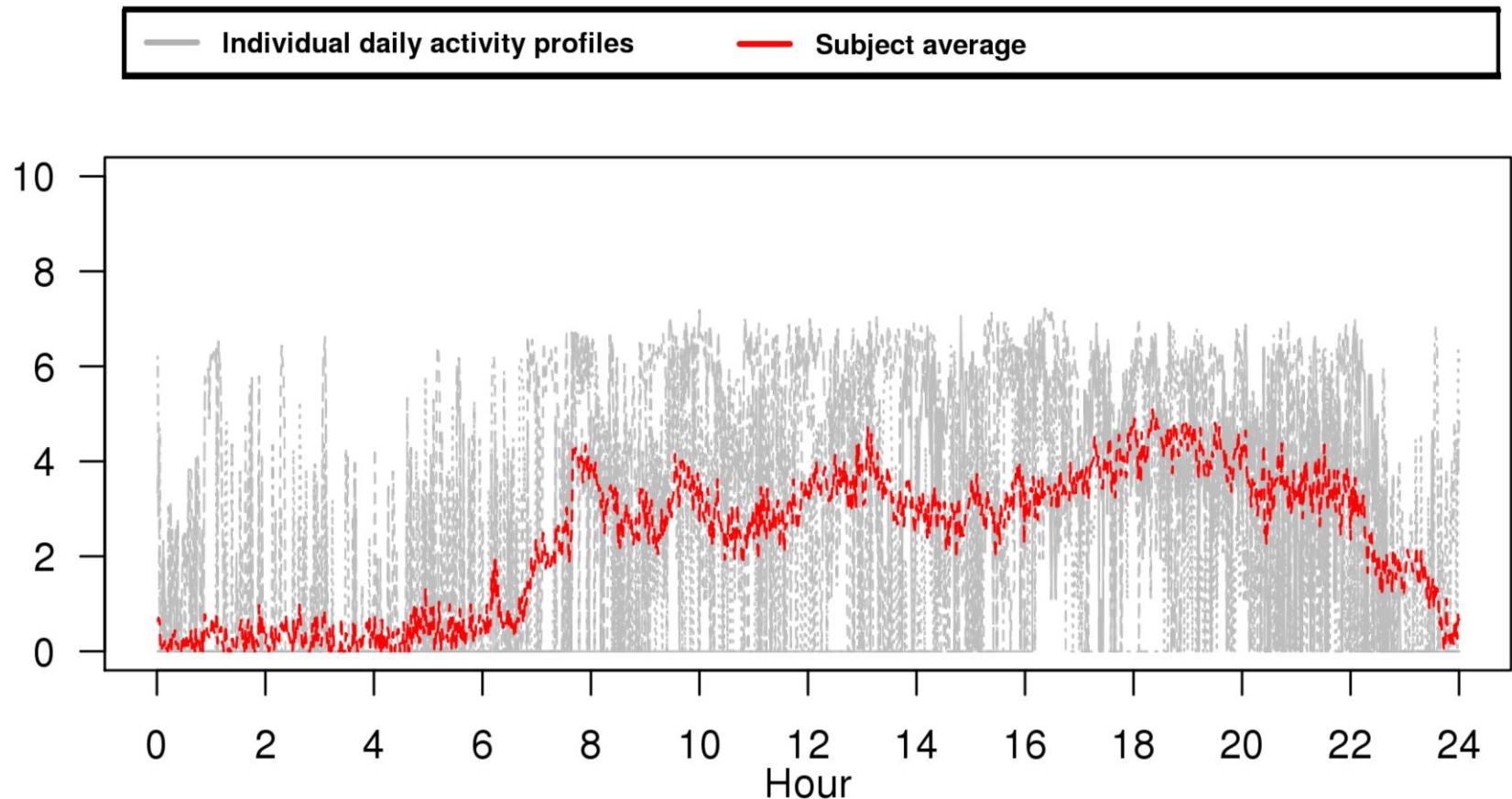
$$Y_{ij}(t), i = 1, 2, \dots, I; j = 1, 2, \dots, J_i$$



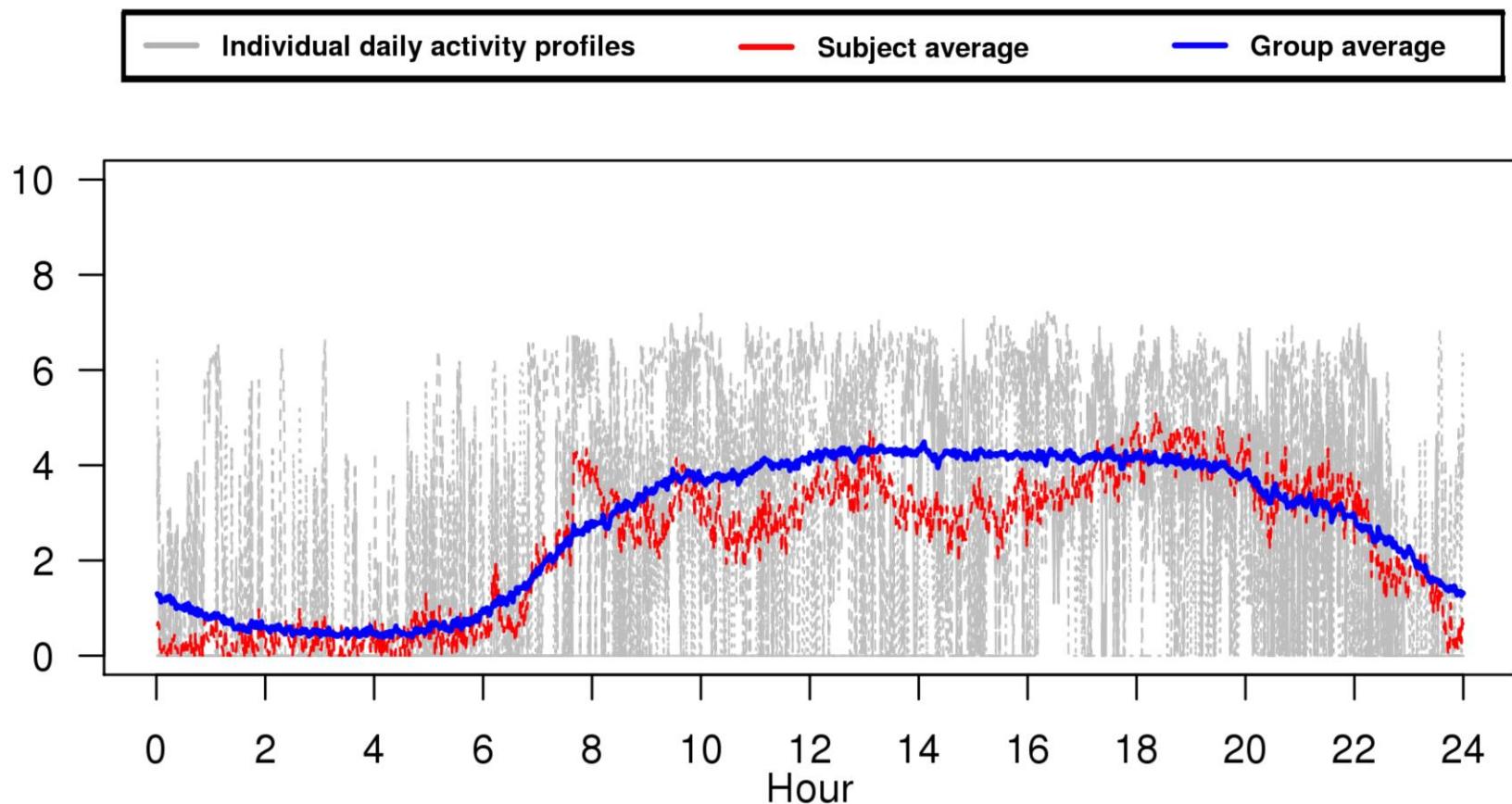
Control and age 67 ( $> 60$ )



Control and age 67 ( $> 60$ )



Control and age > 60



$$\begin{aligned} & \mu_{A_{(i)}, D_{(i)}}(t) \\ & \mu_{A_{(i)}, D_{(i)}}(t) + X_i(t) \\ & \mu_{A_{(i)}, D_{(i)}}(t) + X_i(t) + U_{ij}(t) \end{aligned}$$

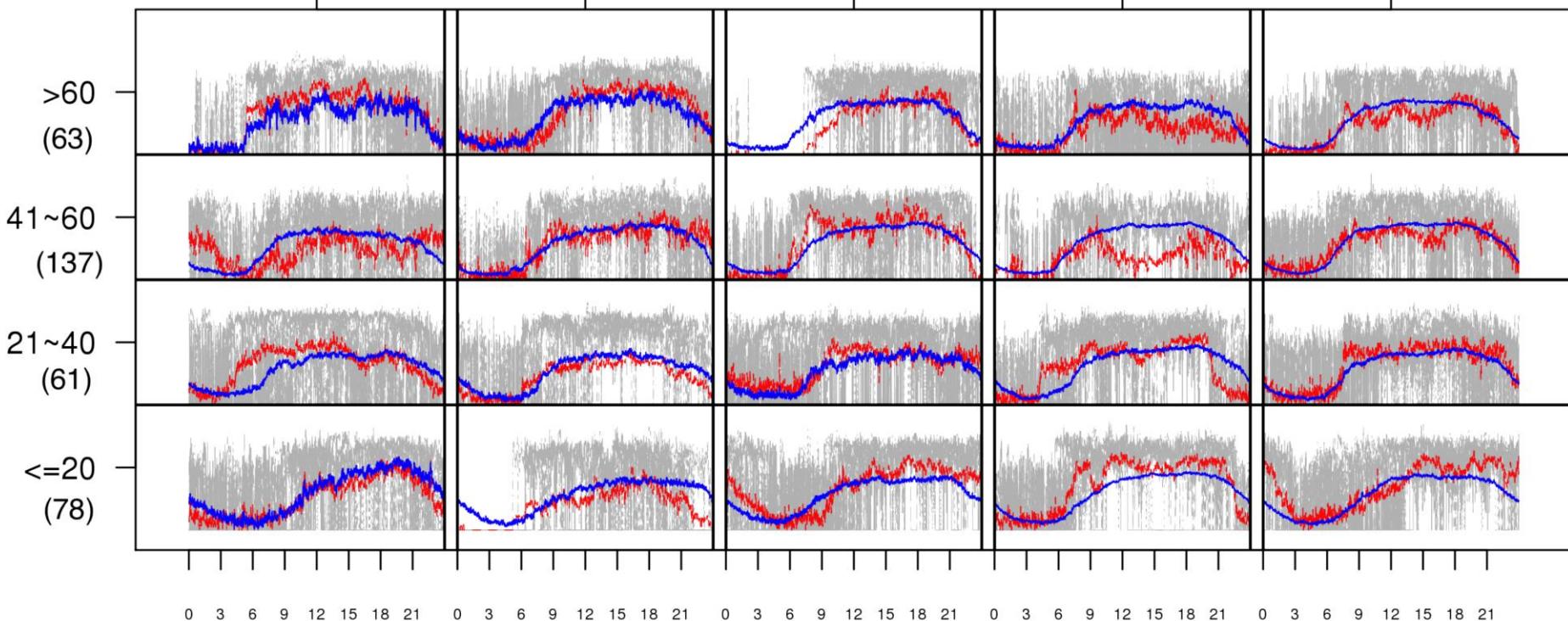
BPI (33)

BPII (31)

MDD (52)

OTHER (98)

CONTROL (125)



Without loss of generality, assume  $Y_{ij}(t) = Y'_{ij}(t) - \mu(t, v_i)$

- Decomposability and additivity

$$Y_{ij}(t) = X_i(t) + U_{ij}(t)$$

- Identifiability

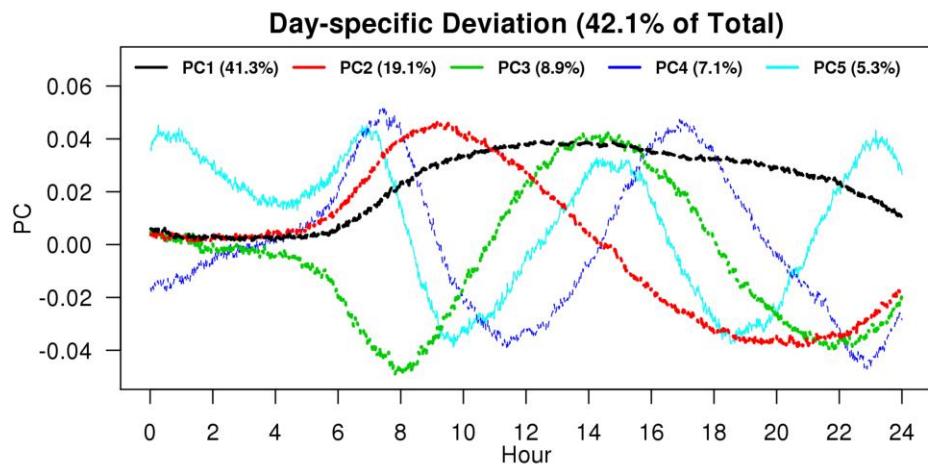
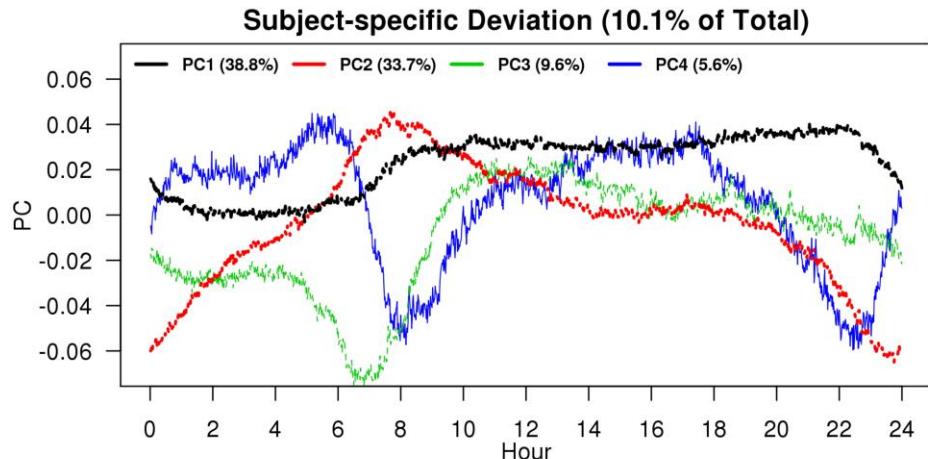
Latent processes are mean zero and mutually uncorrelated.

- Data correlation captured by covariance operators

$$E\{Y_{ij}(t) - Y_{kl}(t)\}\{(Y_{ij}(s) - Y_{kl}(s)\}^T = \begin{cases} 2K_U(t, s) & i = k, j \neq l \\ 2\{K_X(t, s) + K_U(t, s)\} & i \neq k \end{cases}$$

where  $K_X(t, s) := \text{Cov}\{X(t), X(s)\}$ , similar definitions for  $K_U$ .

- Subject heterogeneity accounts for 10.1% of total variability
- The first 4 principal components explain 87.8% of the subject heterogeneity
- Day-specific deviation and random noise along the curve together accounts for the remaining 89.9% of total variability
- The first 5 principal components explain 81.8% of day-to-day variation



# References

- Di, C.Z., Crainiceanu, C.M., Caffo, B.S. and Punjabi, N.M., 2009. Multilevel functional principal component analysis. *The annals of applied statistics*, 3(1), p.458.
- Zipunnikov V., Caffo B.S., Yousem D.M, Davatzikos C., Schwartz B.S., Crainiceanu C. (2011),Multilevel Functional Principal Component Analysis for High-Dimensional Data. *Journal of Computational and Graphical Statistics*, 20(4), pp. 852-873
- Shou, H., Zipunnikov, V., Crainiceanu C., Greven, S. (2015) Structured Functional Principal Component Analysis  
*Biometrics*, 71 (1), pp. 247-757

# Multi-level Matrix-Variate Analysis (MMVA)

# Background

- **Heart failure (HF)** is a leading chronic disease in the elderly
- Lifetime risk is 20% for those over age 40 in the US
- HF burden exceeds \$30 billion (> 50% on hospitalization costs)
- Identifying subjects with increased risk of hospitalization is important

# Background

- **Static risk models** include demographics, comorbidities (AFib, hypertension, diabetes mellitus), income, etc.
- **Dynamic risk models** may be more accurate by including real-time data from wearables
- A prospective longitudinal cohort study Advanced Cardiac Care Center of Columbia University Medical Center
- 59 individuals with clinical diagnosis of congestive heart failure (CHF)
  - 3-9 months of follow up
  - Actical (Respironics)
  - up to one month of minute-level activity counts

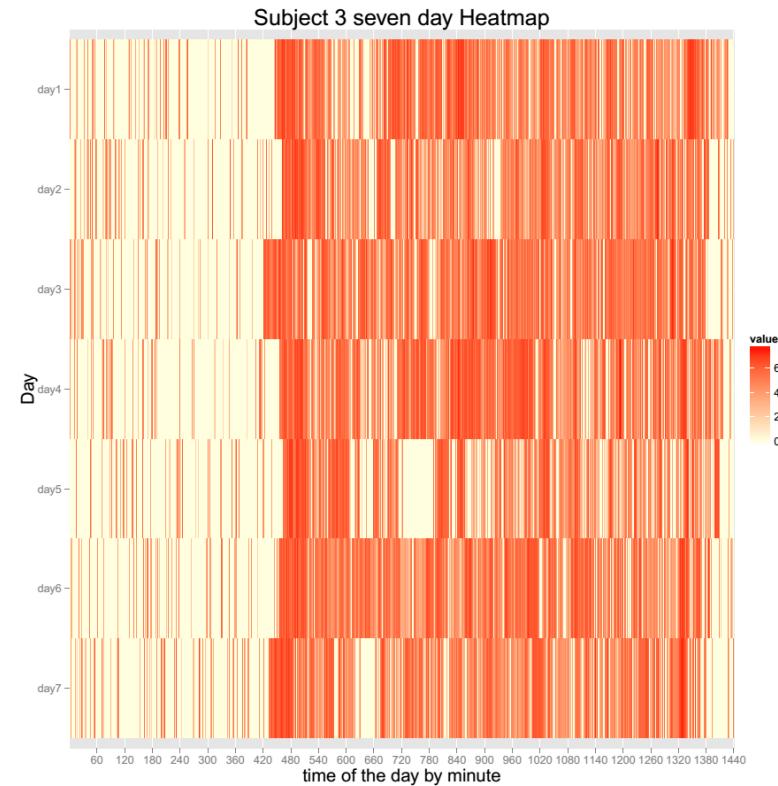
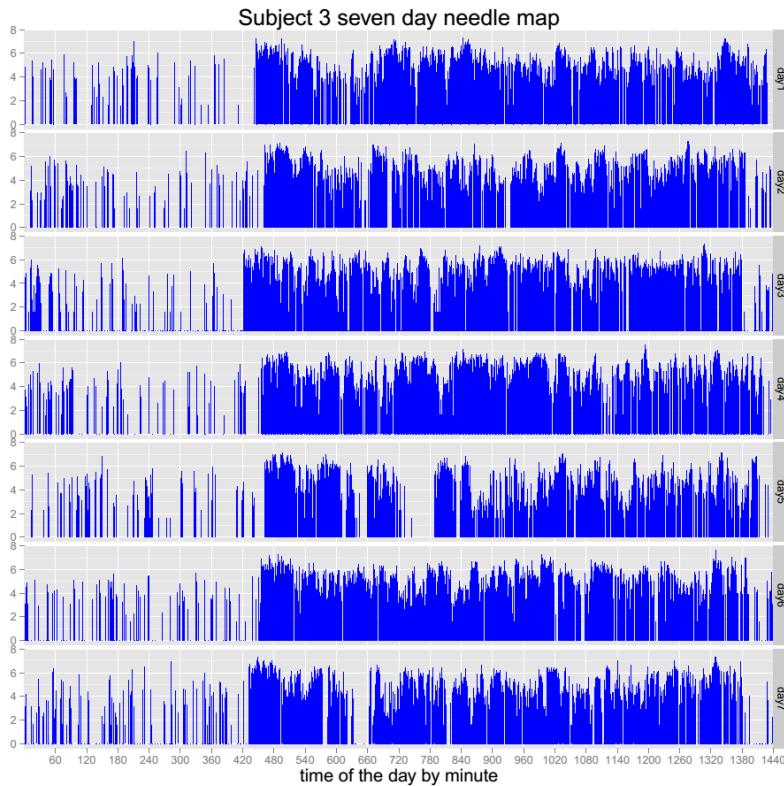


# Background

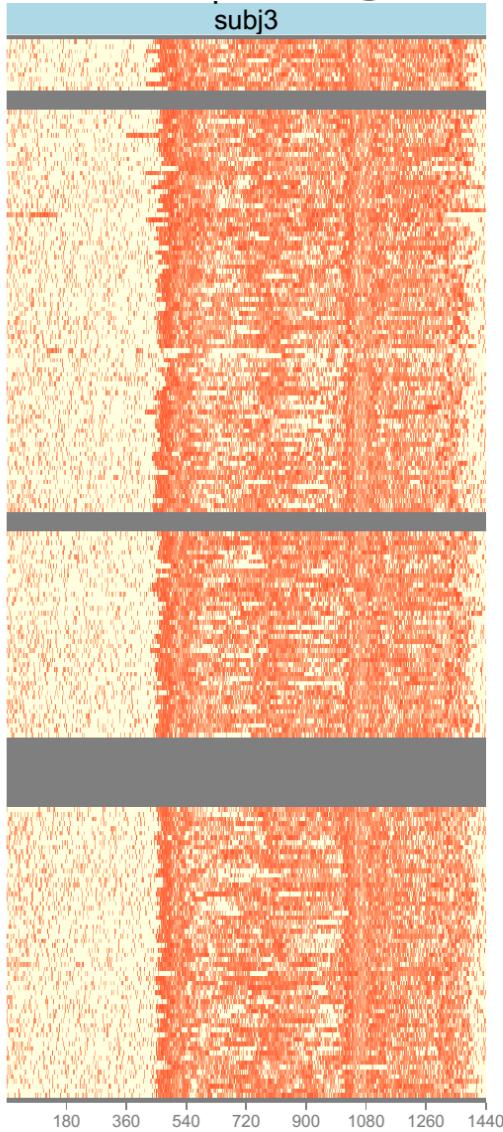
- 24 individuals had adverse clinical events
  - 14 hospitalizations
  - 10 emergency room visits
- **Goal:** model within-subject pre/post event change in patients status
- **Solution:** track week-to-week variability



# Minute-level activity profiles

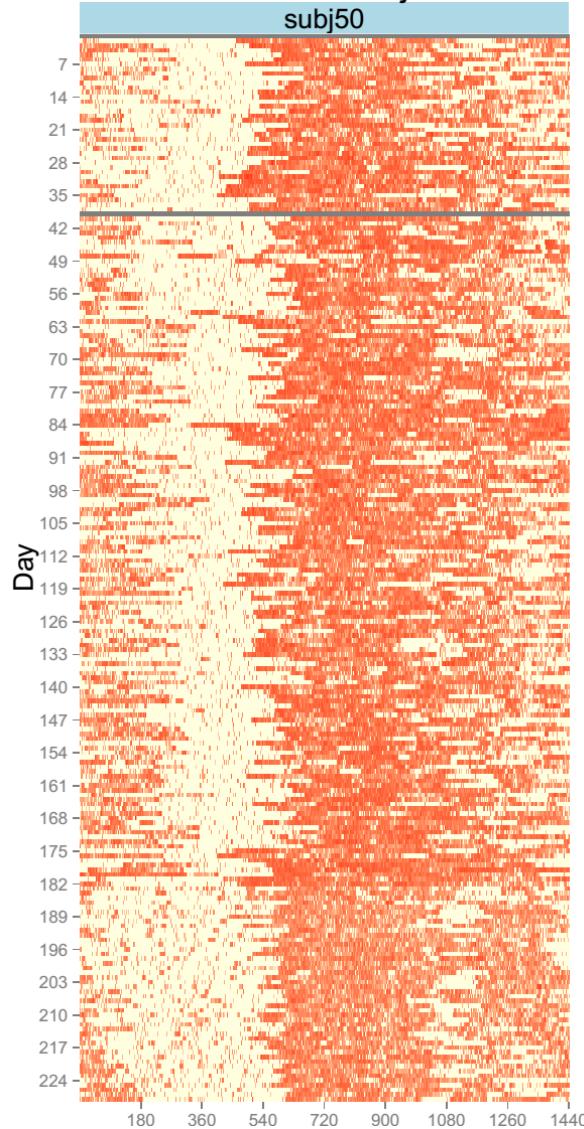


# No-event group subject



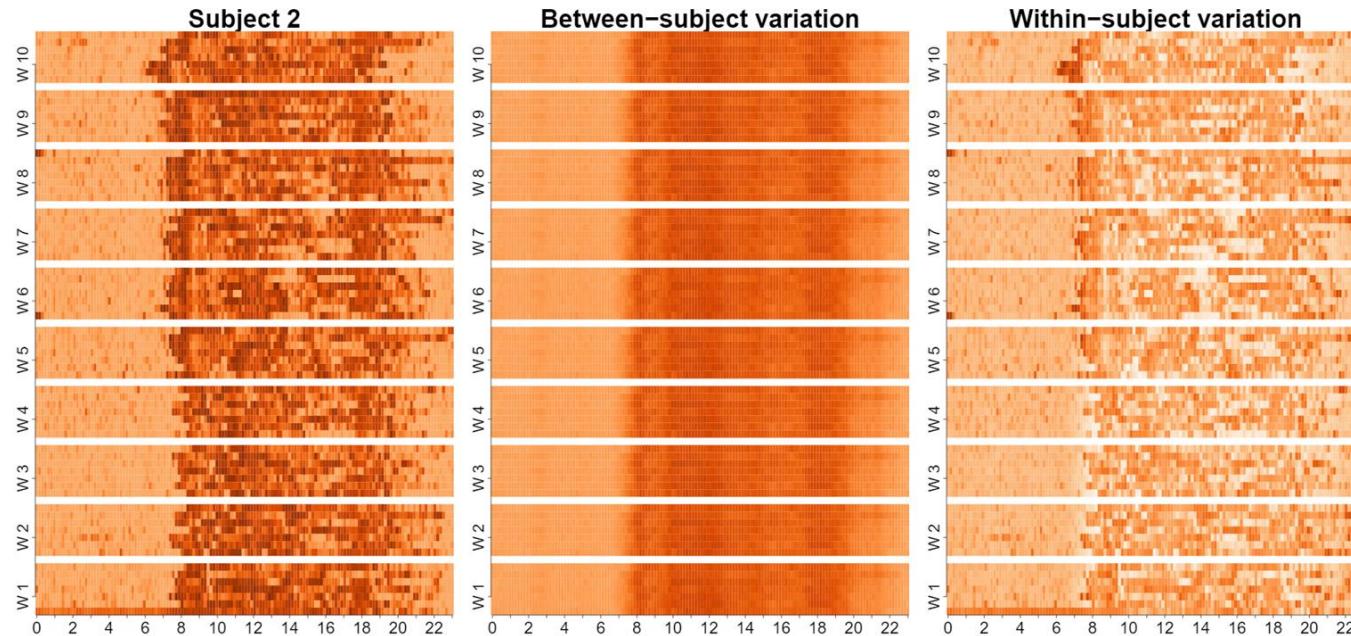
- 8 months of monitoring
- **Low week-to-week variability**
- **Had no hospitalizations**

# Event-group subject



- 8 months of monitoring
- **High week-to-week variability**
- **Had a hospitalization**

# Multilevel Matrix-Variate Analysis (MMVA)



- Keep within-week temporal structure across days & within-day
- Models subject-specific week-to-week variability
- uses a linear mixed effect model to account for the multilevel design
- 2D structure is handled via a matrix-variate distribution

# MMVA

Model:

$$\begin{cases} \mathbf{Y}_{ij} = \mathbf{M} + \mathbf{X}_i + \mathbf{W}_{ij}, i = 1, \dots, I, j = 1, \dots, n_i \\ \mathbf{X}_i \sim \text{MD}_{D,T}(\mathbf{0}, \mathbf{C}_X, \mathbf{R}_X), \\ \mathbf{W}_{ij} \sim \text{MD}_{D,T}(\mathbf{0}, \mathbf{C}_W, \mathbf{R}_W), \end{cases}$$

$\mathbf{Z}$  follows a matrix-variate distribution:  $\mathbf{Z} \sim \text{MD}_{D,T}(\mathbf{M}, \mathbf{C}, \mathbf{R})$   
if  $\text{vec}(\mathbf{Z}) \sim Q_{DT}(\text{vec}(\mathbf{M}), \mathbf{R} \otimes \mathbf{C})$

Normal matrix-variate distribution

$$p(\mathbf{Z} | \mathbf{M}, \mathbf{C}, \mathbf{R}) = \frac{\exp\left(-\frac{1}{2}\text{tr}\left[\mathbf{R}^{-1}(\mathbf{Z} - \mathbf{M})^T \mathbf{C}^{-1}(\mathbf{Z} - \mathbf{M})\right]\right)}{(2\pi)^{DT/2} \|\mathbf{R}\|^{D/2} \|\mathbf{C}\|^{T/2}}$$

$$\mathbf{R} = E[(\mathbf{Z} - \mathbf{M})^T (\mathbf{Z} - \mathbf{M})] / \text{tr}(\mathbf{C})$$

$$\mathbf{C} = E[(\mathbf{Z} - \mathbf{M})(\mathbf{Z} - \mathbf{M})^T] / \text{tr}(\mathbf{R})$$

# MMVA

1. Obtain the within- and between-subject covariance matrices

- method of moments: design-specific closed form
- calculations & storage:  $O(D^*D+T^*T)$

2. Estimate principal components by decomposing covariance matrices

$$K_X = R_X \otimes C_X = [\Phi_{R_X} \otimes \Phi_{C_X}] [D_{R_X} \otimes D_{C_X}] [\Phi_{R_X} \otimes \Phi_{C_X}]^T$$

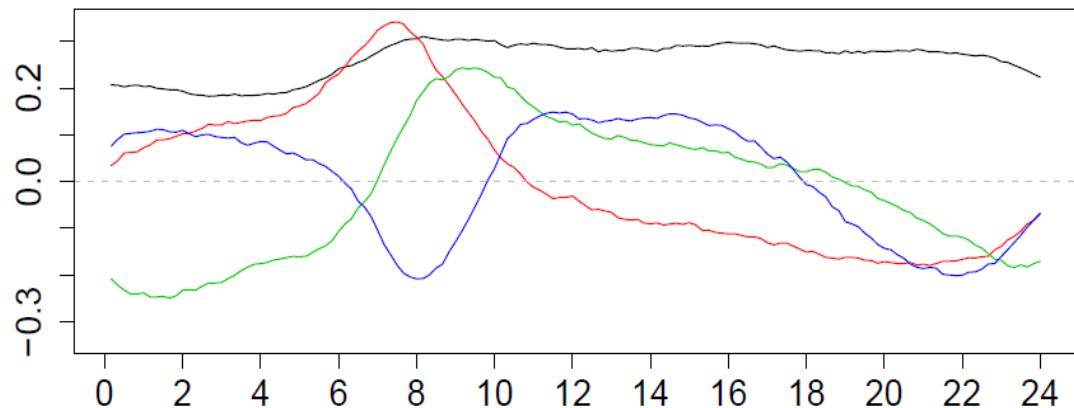
$$K_W = R_W \otimes C_W = [\Phi_{R_W} \otimes \Phi_{C_W}] [D_{R_W} \otimes D_{C_W}] [\Phi_{R_W} \otimes \Phi_{C_W}]^T$$

3. Use a projection method to estimate the principal scores

$$Y_{ij} = M + \Phi_{C_X} \Gamma_i^X \Phi_{R_X}^T + \Phi_{C_W} \Gamma_{ij}^W \Phi_{R_W}^T$$

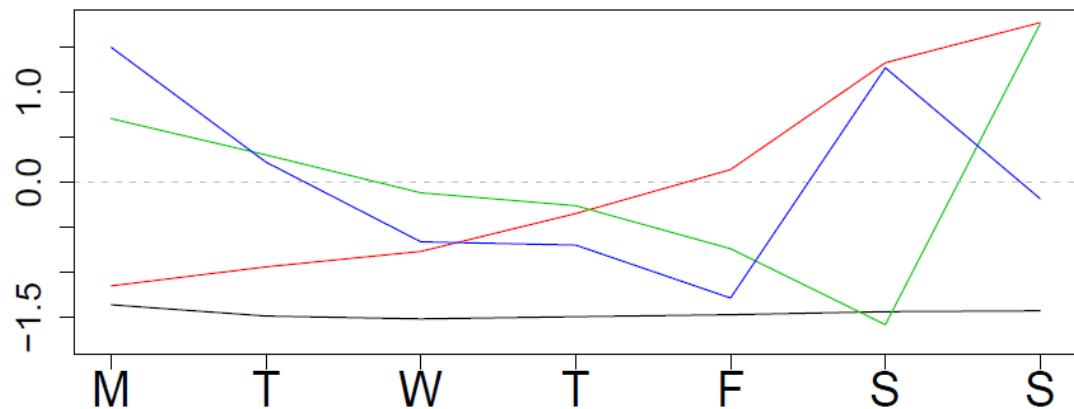
# MMVA

Time – Within  $R_W$



Black, red, green, blue  
are the 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, and  
4<sup>th</sup> principal  
components for rows  
(top) and columns  
(bottom)

Day – Within  $C_W$



# MMVA

- Between-subject matrix-variate distance

$$d(i, k) = \text{dist}(\mathbf{X}_i, \mathbf{X}_k) = \| \text{vec}(\boldsymbol{\Gamma}_i^X) - \text{vec}(\boldsymbol{\Gamma}_k^X) \|.$$

- Within-subject matrix-variate distance

$$d_i(j_1, j_2) = \text{dist}(\mathbf{W}_{ij_1}, \mathbf{W}_{ij_2}) = \| \text{vec}(\boldsymbol{\Gamma}_{ij_1}^W) - \text{vec}(\boldsymbol{\Gamma}_{ij_2}^W) \|.$$

# MMVA

- **Between-subject scores** can be used as a “static” biomarker to enrich and potentially improve accuracy of currently used “static” risk score models.

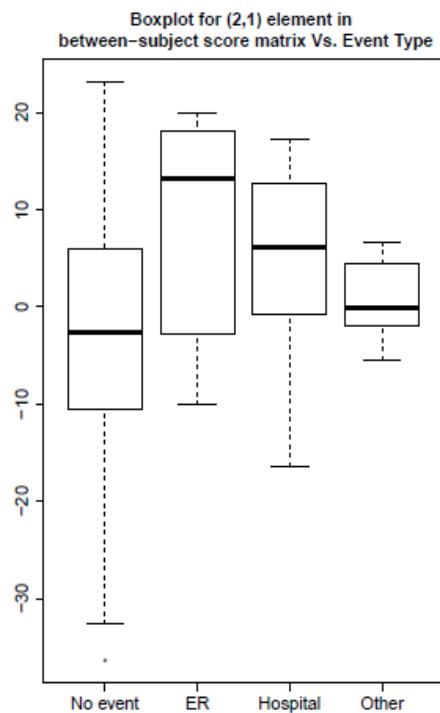
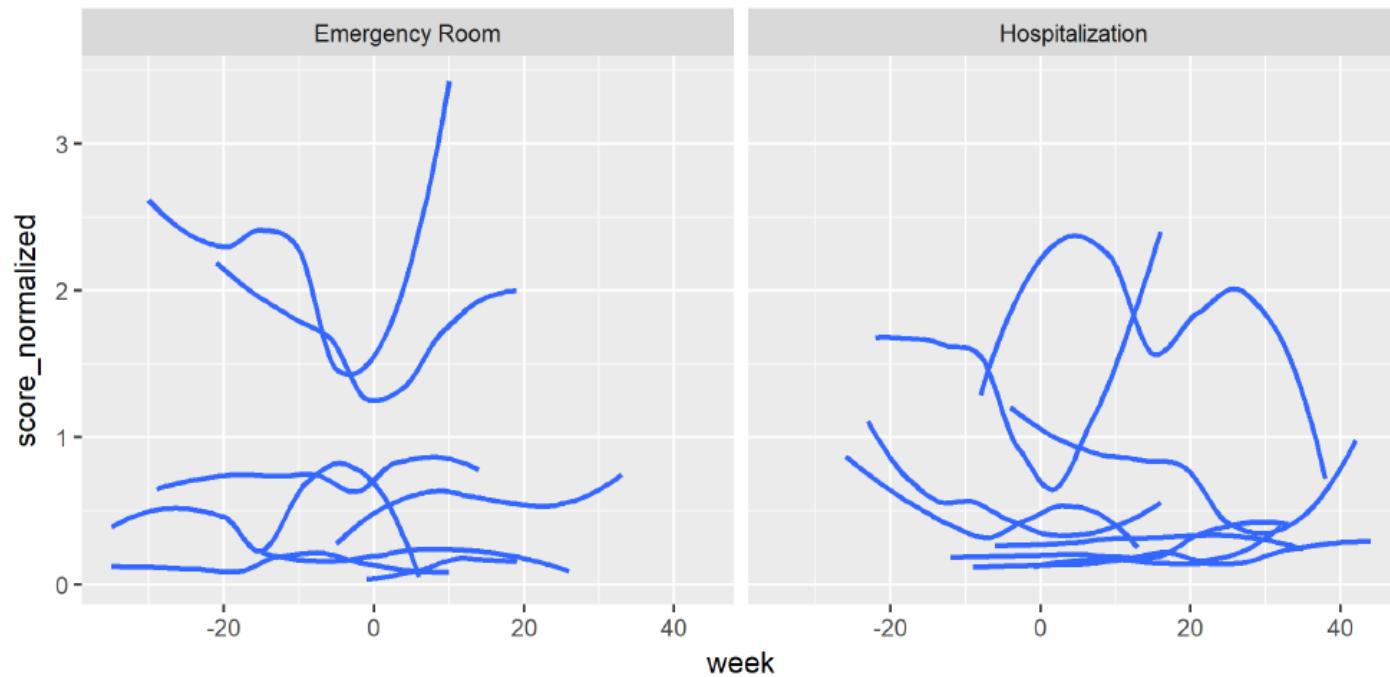


Table 3: Logistic models for between-subject scores; an asterisk indicates significance at level 0.05

	Model 1	Model 2	Model 3	Model 4
Score (2,1)	0.012(0.005)*	0.011(0.005)*	0.010(0.005)*	0.011(0.035)*
Score (2,3)	0.016(0.012)	0.015(0.013)	0.015(0.013)	0.013(0.013)
Score (4,2)	0.024(0.054)	0.014(0.056)	0.026(0.057)	0.029(0.057)
Sex		0.113(0.140)	0.099(0.141)	0.056(0.147)
Age			0.004(0.004)	0.005(0.005)
BMI				0.012(0.011)

# MMVA

- **Within-subject scores** can be used as “dynamic” biomarkers that inform about weekly changes in patient status.



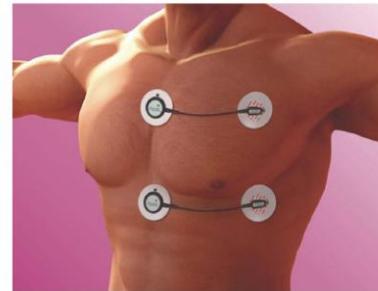
# Reference

Huang, L., Bai, J., Ivanescu, A., Harris, T., Maurer, M., Green, P. and Zipunnikov, V., 2019. Multilevel matrix-variate analysis and its application to accelerometry-measured physical activity in clinical populations. *Journal of the American Statistical Association*, 114(526), pp.553-564.

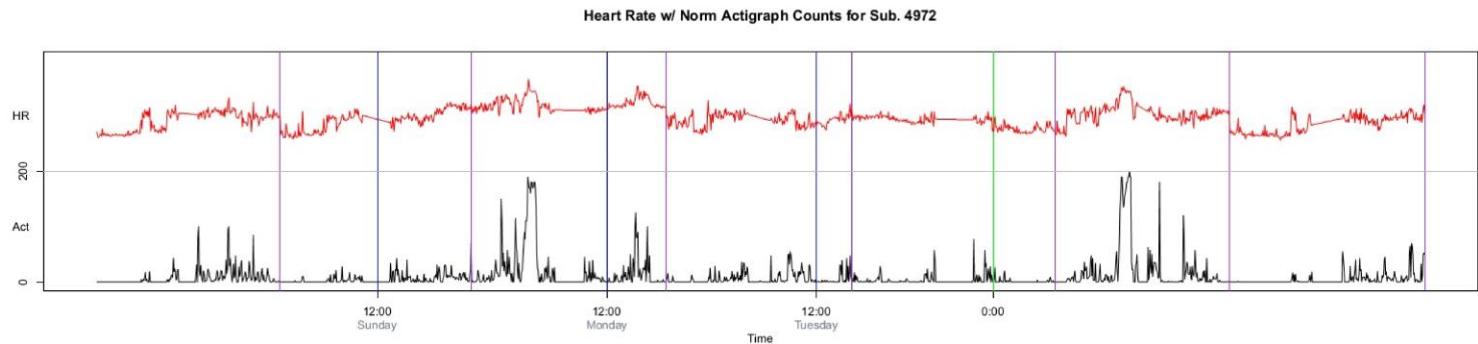
**Data is freely available with the submission**

# Multi-level Generalized Function-on-Scalar Regression and Principal Component Analysis

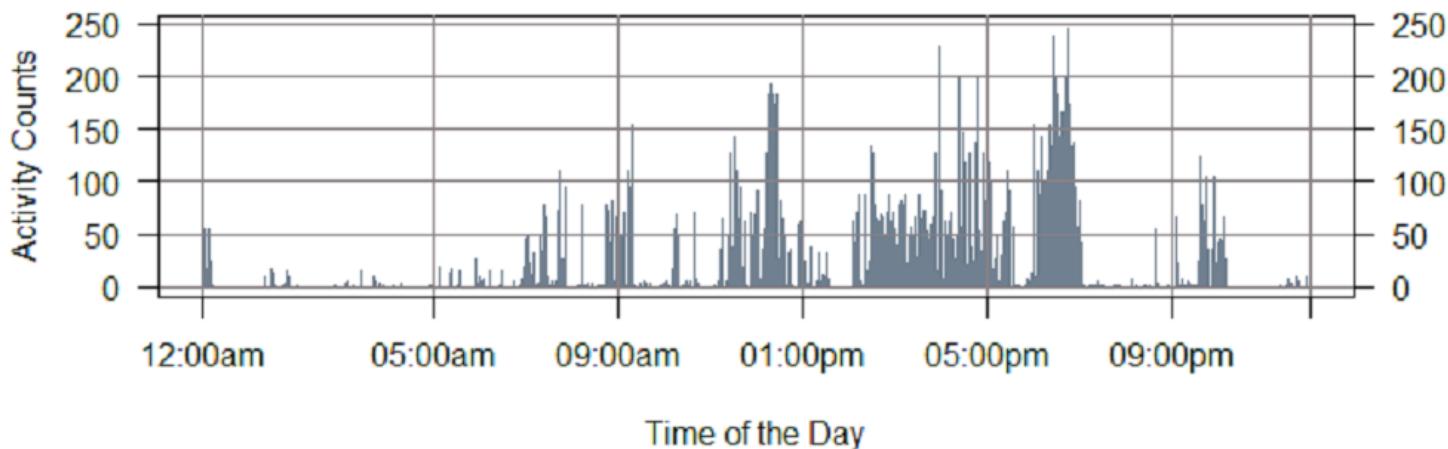
- normative human aging
- followed for life, visits every 1-4 years
- 631 subjects, wear for 7 days
- **Goal:** daily patterns of physical activity



- Every minute: activity counts and heart rate



- ▶ Binary time series  $Y_{ij}(t)$  : activity/inactivity
- ▶ Covariates  $x_{ij,k}$  : age, gender, BMI, etc



- Generalized Multilevel Functional-on-Scalar Regression

- ▶  $Y_{ij}(t)$  is a generalized response curve
- ▶  $Y_{ij}(t)$  comes from an exponential family

$$\begin{aligned}
 \text{E}[Y_{ij}(t)|b_i(t), v_{ij}(t)] &= \mu_{ij}(t) \\
 g(\mu_{ij}(t)) &= \beta_0(t) + \sum_{k=1}^p x_{ij,k} \beta_k(t) + b_i(t) + v_{ij}(t) \\
 &\approx \beta_0(t) + \sum_{k=1}^p x_{ij,k} \beta_k(t) + \sum_{k=1}^{K^{(1)}} c_{ik}^{(1)} \psi_k^{(1)}(t) + \sum_{k=1}^{K^{(2)}} c_{ijk}^{(2)} \psi_k^{(2)}(t).
 \end{aligned}$$

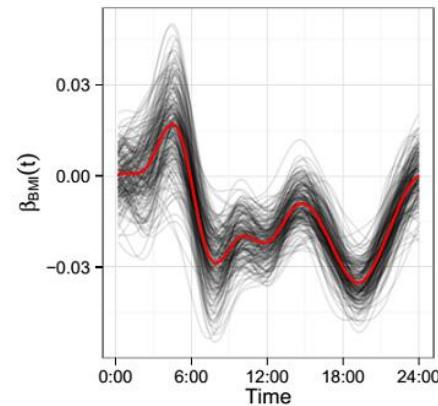
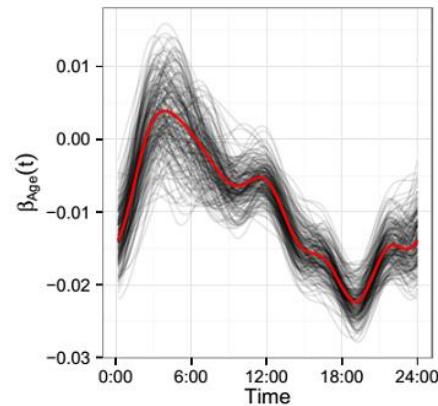
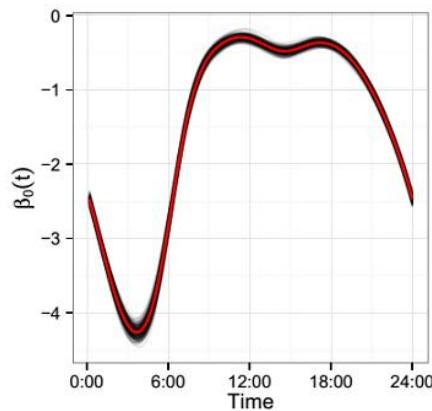
## • Generalized Multilevel Functional-on-Scalar Regression

- ▶  $\Theta$ : cubic B-spline basis functions
- ▶ computation using Stan (Hamiltonian MC)

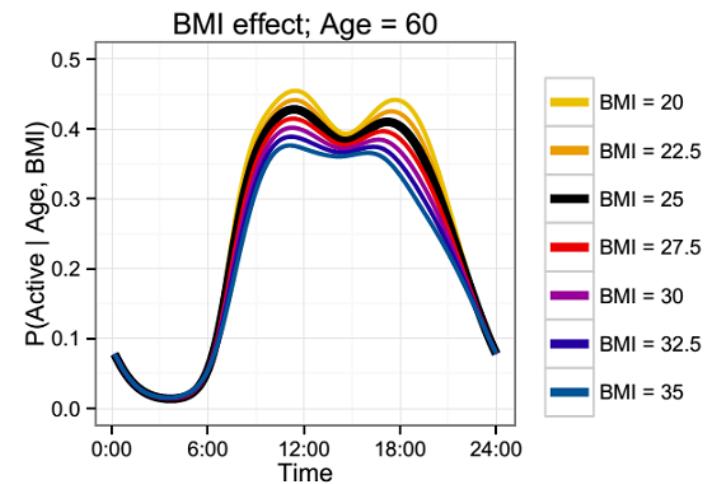
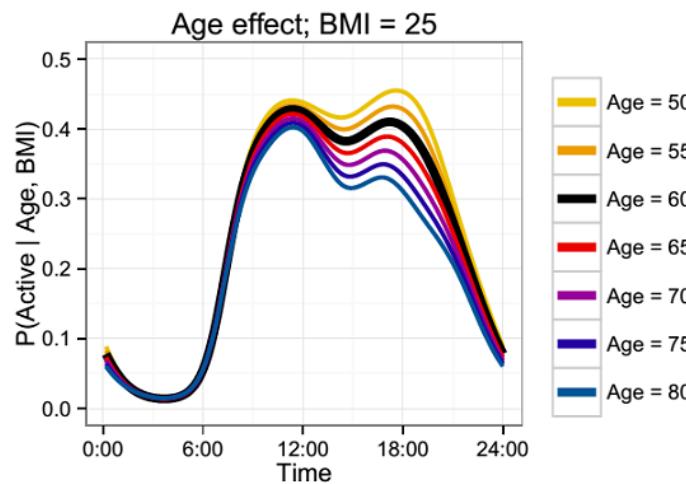
$$\text{E}[\mathbf{Y}|\mathbf{b}, \mathbf{v}] = \boldsymbol{\mu}$$

$$\begin{aligned} g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{v} \\ &= \mathbf{X}\mathbf{B}_X^T\boldsymbol{\Theta}^T + \mathbf{Z}\mathbf{C}^{(1)}\mathbf{B}_{\psi^{(1)}}^T\boldsymbol{\Theta}^T + \mathbf{C}^{(2)}\mathbf{B}_{\psi^{(2)}}^T\boldsymbol{\Theta}^T. \end{aligned}$$

- Generalized Multilevel Functional-on-Scalar Regression
  - ▶ Estimated functional mean, bmi, and age effects



- ▶ Estimated functional mean, bmi, and age effects



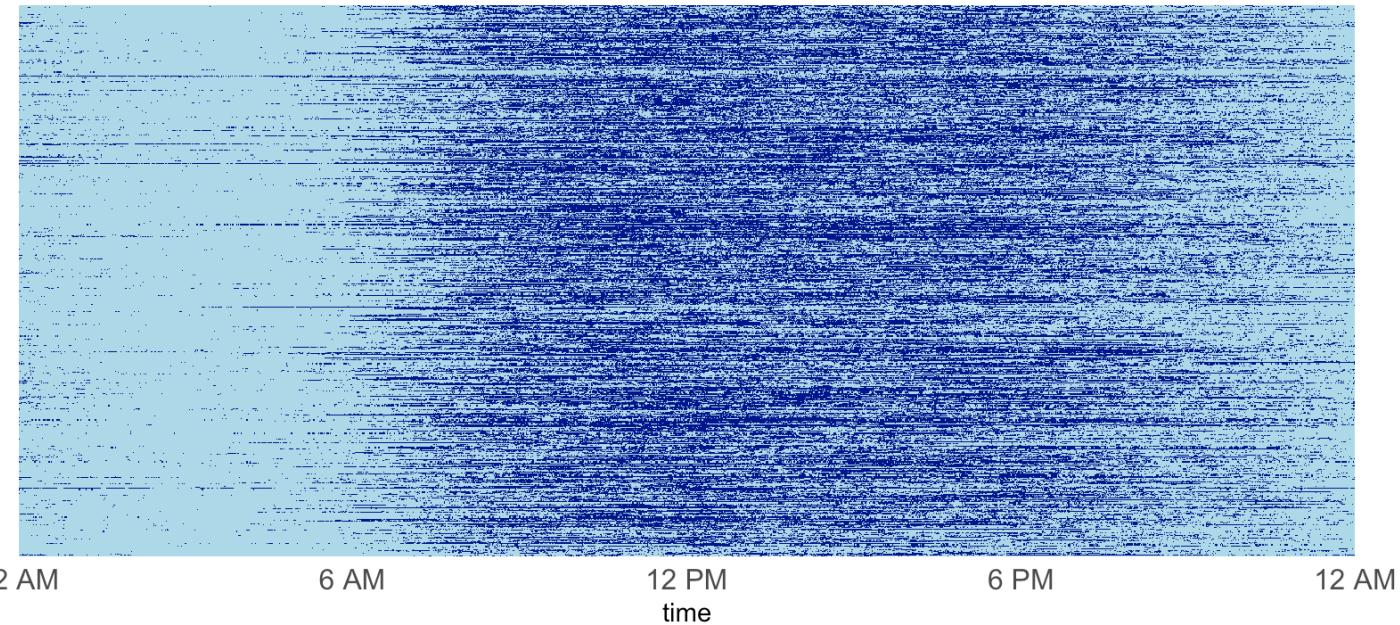
## References:

- Goldsmith, J., Zipunnikov, V, Schrack, J., Generalized Multilevel Function-on-Scalar Regression and Principal Component Analysis  
Biometrics, 71 (2), pp. 344-353

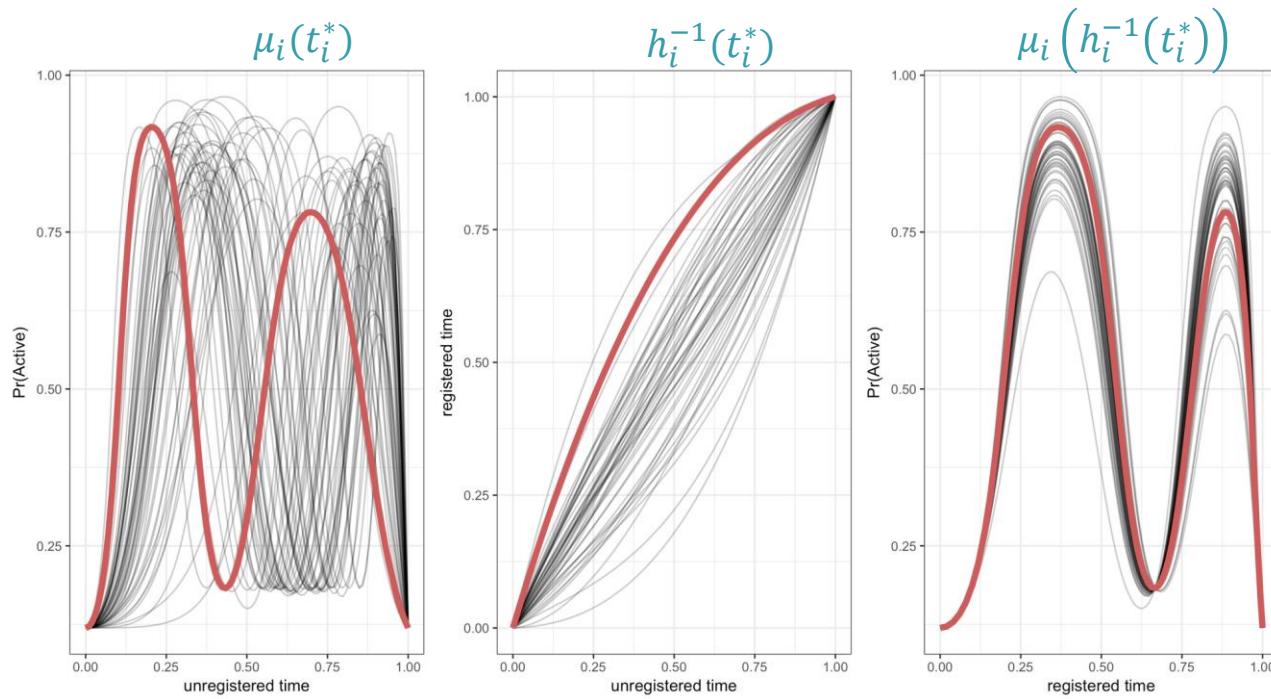
# Registration of binary (0/1) actigraphy profiles

# Binary “activity”

Subject



# Registration



# Exponential family registration

- Step 1
  - Underlying probability of being active is registration ‘template’
  - Estimate template,  $\mu_i(t)$ , conditional on current estimate of time  $\hat{t} = \hat{h}_i^{-1}(t_i^*)$
- Step 2
  - Inverse warping functions  $h_i^{-1}(t_i^*)$  map from observed to registered time
  - Estimate inverse warping function,  $h_i^{-1}(t_i^*) = t$
  - Warping step is MLE conditional on template estimate  $\hat{\mu}_i(t)$
- Iterate between step (1) and step (2) until alignment
- Focus on computational efficiency

# FPCA

- Small digression ...
- FPCA is a variation on PCA that involves smoothness in eigenfunctions
- Most often, the process is to
  - Compute a covariance from observed data vectors
  - Smooth the covariance, after removing off-diagonal elements
  - Decompose resulting covariance matrix to obtain eigenfunctions (vectors) and score variances
  - Scores are estimated in a mixed model

$$\mu_i(t) = \alpha(t) + \sum_{k=1}^K c_{ik} \psi_k(t) + \epsilon_i(t)$$

# Probabilistic (F)PCA

- Probabilistic / Bayesian methods provide an alternative to the usual covariance decomposition approach
- Model is of interest is still

$$\mu_i(t) = \alpha(t) + \sum_{k=1}^K c_{ik} \psi_k(t) + \epsilon_i(t)$$

- Add priors for the mean and eigenfunctions (vectors)
- MLE is equivalent, up to rotation, to the covariance decomposition method

# Back to registration – Step 1

- Estimate template  $\mu_i(t)$  by exponential family FPCA
- Model is

$$E[Y_i(t)|c_i] = \mu_i(t)$$

$$g[\mu_i(t)] = \alpha(t) + \sum_{k=1}^K c_{ik} \psi_k(t)$$

- Observations  $Y_i(t) \sim \text{Exponential Family}$
- Subject specific scores  $c_{ik} \sim (0, \lambda_k)$
- We implicitly condition on  $h_i^{-1}(\mathbf{t}_i^*)$  when we assume  $\mathbf{t}_i$  is known

# Exponential-family FPCA

- Expand functions using B-spline bases  $\Theta(\mathbf{t}_i)$

$$E[Y_i(\mathbf{t}_i)|\mathbf{c}_i] = \mu_i(\mathbf{t}_i)$$

$$g[\mu_i(\mathbf{t}_i)] = \Theta(\mathbf{t}_i)\boldsymbol{\alpha}_{\Theta} + \Theta(\mathbf{t}_i)\boldsymbol{\Psi}_{\Theta}\mathbf{c}_i$$

- Framework is general, but for binary data
  - $Y_i(\mathbf{t}_i) \sim \text{Bernoulli}[\mu_i(\mathbf{t}_i)]$
  - $\mathbf{c}_i \sim MVN(0, I_{K \times K})$
  - $g(\cdot) = \text{logit}(\cdot)$

## Registration Step 2: Estimate $\mathbf{h}_i^{-1}(\mathbf{t}_i^*)$

- Given a template, stretch time domain to match

$$E[Y_i \left( h_i^{-1}(\mathbf{t}_i^*) \right) | c_i] = \mu_i(\mathbf{t}_i)$$

$$g[\mu_i(\mathbf{t}_i)] = \Theta(\mathbf{t}_i)\alpha_\Theta + \Theta(\mathbf{t}_i)\Psi_\Theta c_i$$

- Constrained optimization problem

- $h_i^{-1}(0) = 0$
- $h_i^{-1}(1) = 1$
- Strictly increasing

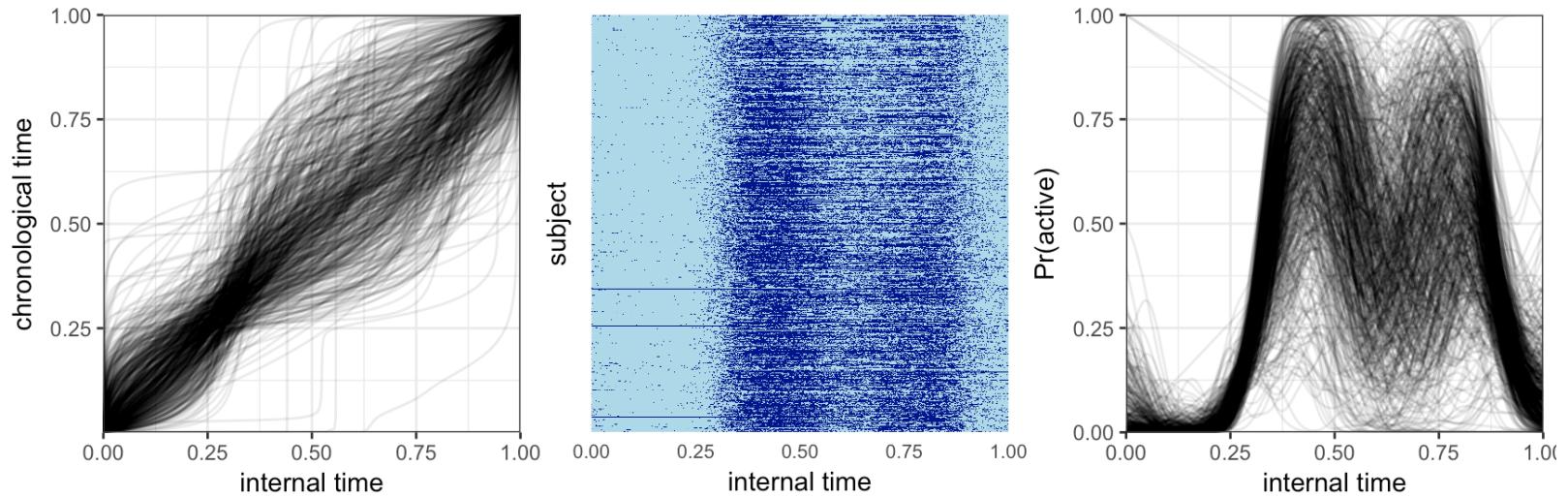
- Uses subject-specific means,  $\mu_i$ , from FPCA

# Implementation

- Fast Binary FPCA (C++ backend)
- Constrained optimization with known gradient in `constrOptim()`
- R package `registr`
- `refund.shiny` plotting
- `tidyfun` integration coming soon!



# BLSA – post-registration



## References:

- Wrobel, J., Zipunnikov, V., Schrack, J. and Goldsmith, J., 2019. Registration for exponential family functional data. *Biometrics*, 75(1), pp.48-57.