Special Topics: Biostatistical Methods for Wearable Computing

Analyzing “macro" scale accelerometry data: NHANES

Andrew Leroux

January 13, 2020

Roadmap

- Last class
  - Introduced accelerometry broadly
    - Described what is being measured
    - Presented some ways of visualizating the data
  - Provided an example of raw (sub-second level) accelerometry data
- Today
  - Recap wearables, focusing on accelerometry
  - Describe data and motivate dimensionality reduction (raw data to aggregated data)
  - Describe the NHANES study
  - Walk through some example analyses using NHANES accelerometry data

Wearable and Implantable Technology

- Wearable and implantable devices are smart electronic devices that can be worn on the body as implants or accessories
- Emerging technology (increasing variety of sensors and signals measured)
- Growing popularity in health research and consumer tech

Wearable and Implantable Technology: Statistical Methods

- High dimensional time series data
  - Signal processing
  - Functional data analysis
  - Feature extraction
- Methodological challenges specific to wearables
  - Study protocol (device location, battery life, convenience, comfort)
  - Wear vs non-wear (complex missing data patterns)
- Computational challenges
  - Data storage
  - Data analysis

NHANES accelerometry: Reproducing these Analyses

- All analyses presented here can be replicated using the "wearables_shortcourse_week1.R script located at `https://github.com/andrew-leroux/wearables_special_topics_course`
- Steps:
  - Download or clone
  - Open R project ("data_science_accelerometry.Rproj")
  - Open R script "wearables_shortcourse_week1.R" in the "code" directory
  - Run code

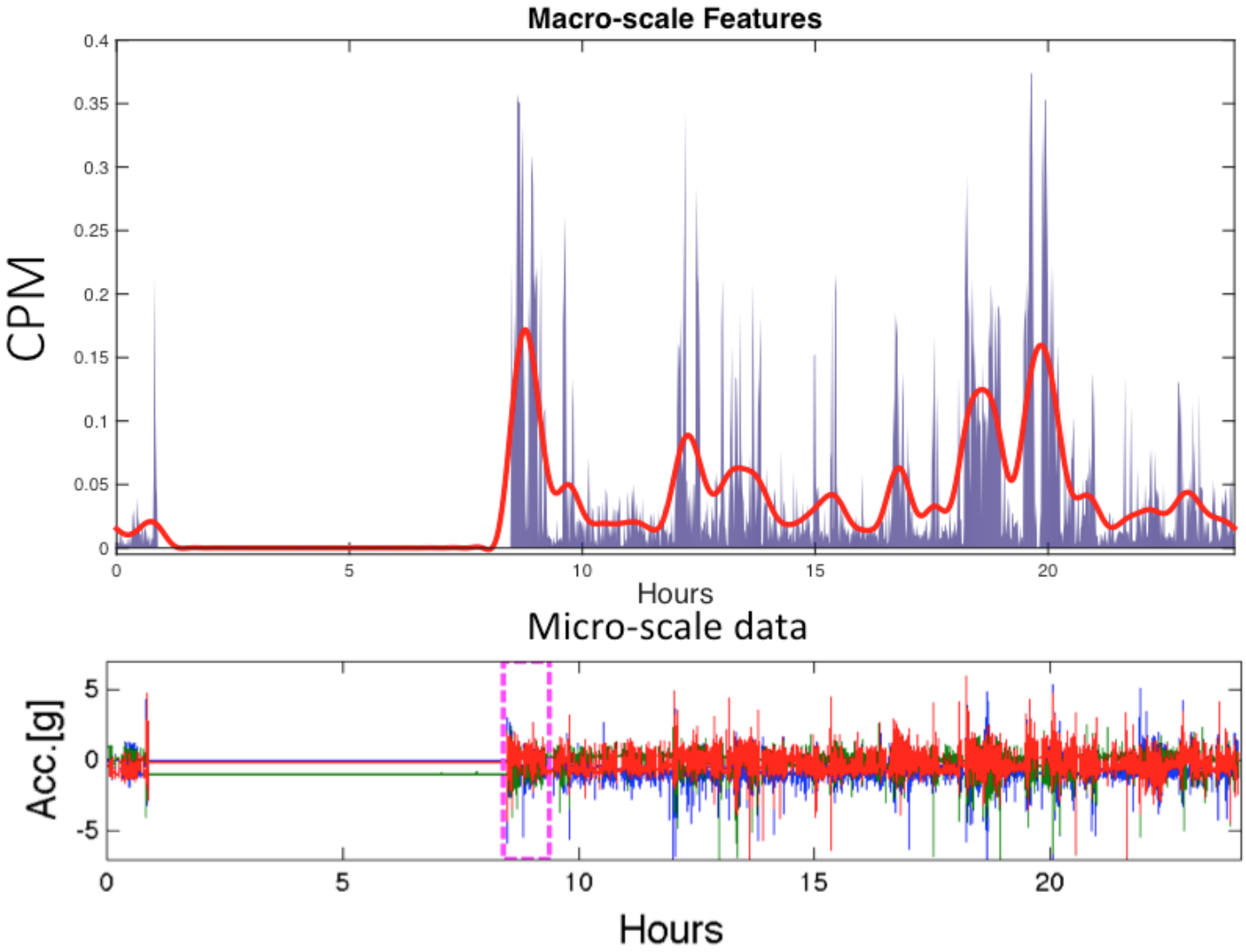
Why Accelerometry?

- Physical activity is **the** modifiable risk factor
  - Mortality
  - Morbidity
  - Pain
  - "Aging"
- Objective vs. self-reported PA
  - Recall bias
  - Subjective assessment of intensity
  - Questionairre format
  - "MET" equivalents

Accelerometry: Data Reduction

- Raw accelerometry data are enormous
- A tri-axial accelerometer recording at 50hz for 7 days generates $90,720,000$ data points per subject
- Can we reduce the data?
  1. Collapse the three channels (x,y,z) into a single time series
  2. Summarize within an epoch (1s, 5s, 1min, ...)
     - "Activity count"
     - Activity index
     - Vector magnitude
     - Mean amplitude deviation
- Choices for data redution depend on study goals

Micro- vs Macro-Scale Accelerometry Data



National Health and Nutrition Examination Survey (NHANES)

- Complex, multistage probability sample of the non-institutionalized US population
- Ongoing cross-sectional study conducted in 2-year "waves"
- 2003-2004 and 2005-2006 waves collected accelerometry data
- Oversamples certain groups
  - African Americans, Mexican Americans
  - Low income White Americans
  - Adolescents (12-19 years old)
  - Older adults (60+)
- All participants assigned a "weight" indicating the number of people in the US population they "represent"

NHANES Sampling Procedure

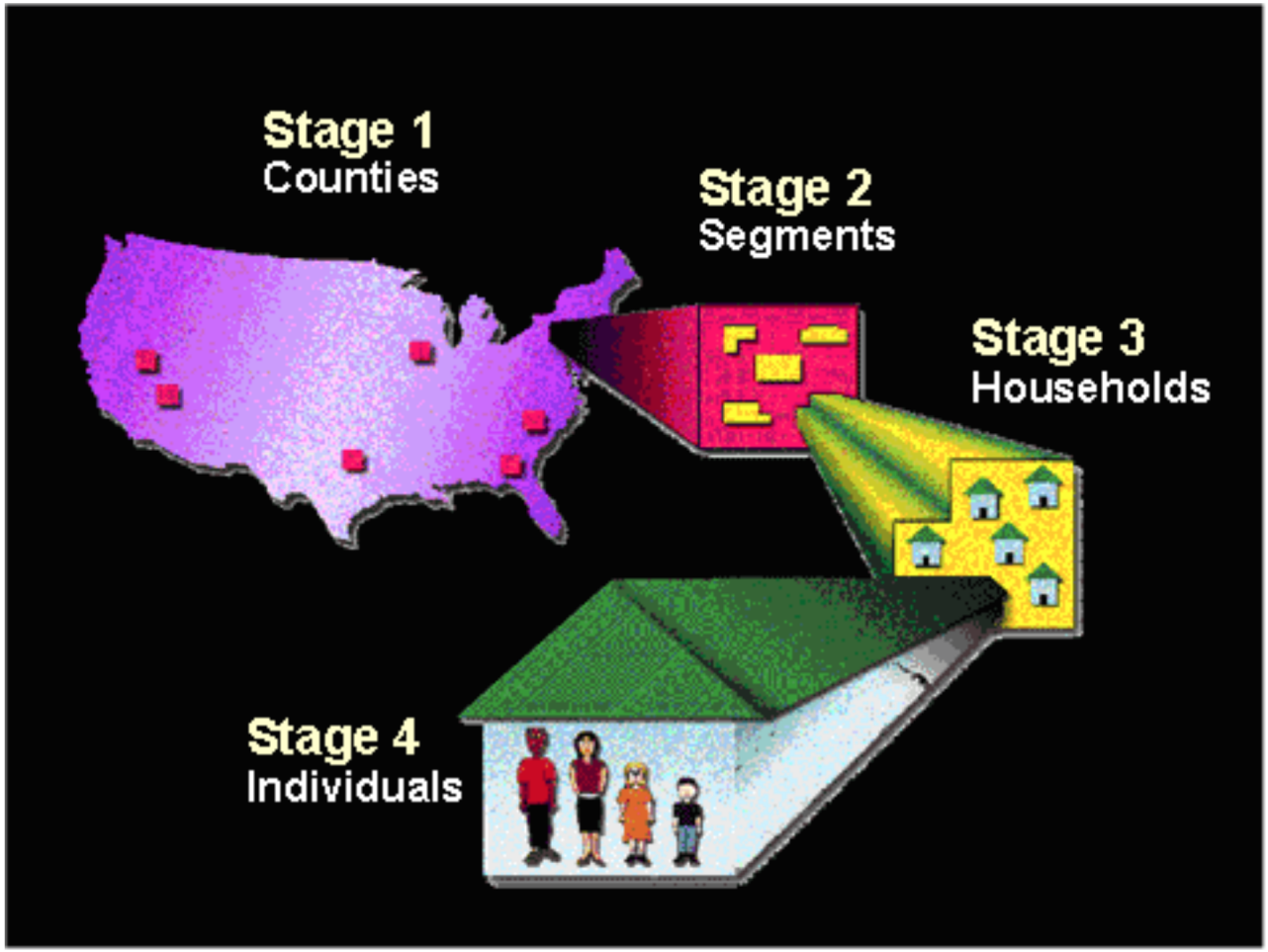


Image credit: https://www.cdc.gov/nchs/tutorials/NHANES/SurveyDesign/SampleDesign/Info1.htm

NHANES 2003-2006 Data

- Accelerometry data
  - Acceleration summarized into minute-level "activity counts"
  - Up to 7 days of data for each participant
  - Study protocol: remove the device at bedtime
- NHANES collects a lot of data on participants
  - `https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2003`
  - `https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2005`

NHANES accelerometry: data structure

- Accelerometry data downloadble from NHANES is in long format
- Very large file sizes ($\approx$ 2.5 GB)

SEQN	PAXSTAT	PAXCAL	PAXDAY	PAXN	PAXHOUR	PAXMINUT	PAXINTEN	PAXSTEP
31128	1	1	1	1	0	0	166	4
31128	1	1	1	2	0	1	27	0
31128	1	1	1	3	0	2	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

NHANES accelerometry: proposed data strucutre

- Wide format instead of long format[1] ($\approx$ 60 MB)
- 7 rows per participant, descending cronologoical order

	Unique Identifier		Quality Flags		NHANES wave	Activity Counts				
	SEQN	PAXDAY	PAXCAL	PAXSTAT	SDDSRVYR	MIN1	MIN2	MIN3	$\cdots$	MIN1440
(a)	31128	1	1	1	4	166	27	0	$\cdots$	0
	31128	2	1	1	4	0	0	0	$\cdots$	0
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	31128	7	1	1	4	0	0	0	$\cdots$	0
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(b)	31193	2	2	1	4	0	0	0	$\cdots$	1921
	31193	3	2	1	4	335	2598	2185	$\cdots$	46
	31193	4	2	1	4	0	0	0	$\cdots$	0
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(c)	31880	2	2	2	4	32767	32767	32767	$\cdots$	32767
	31880	3	2	2	4	32767	32767	32767	$\cdots$	32767
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(d)	32008	5	1	2	4	0	0	0	$\cdots$	0
	32008	6	1	2	4	NA	NA	NA	$\cdots$	NA

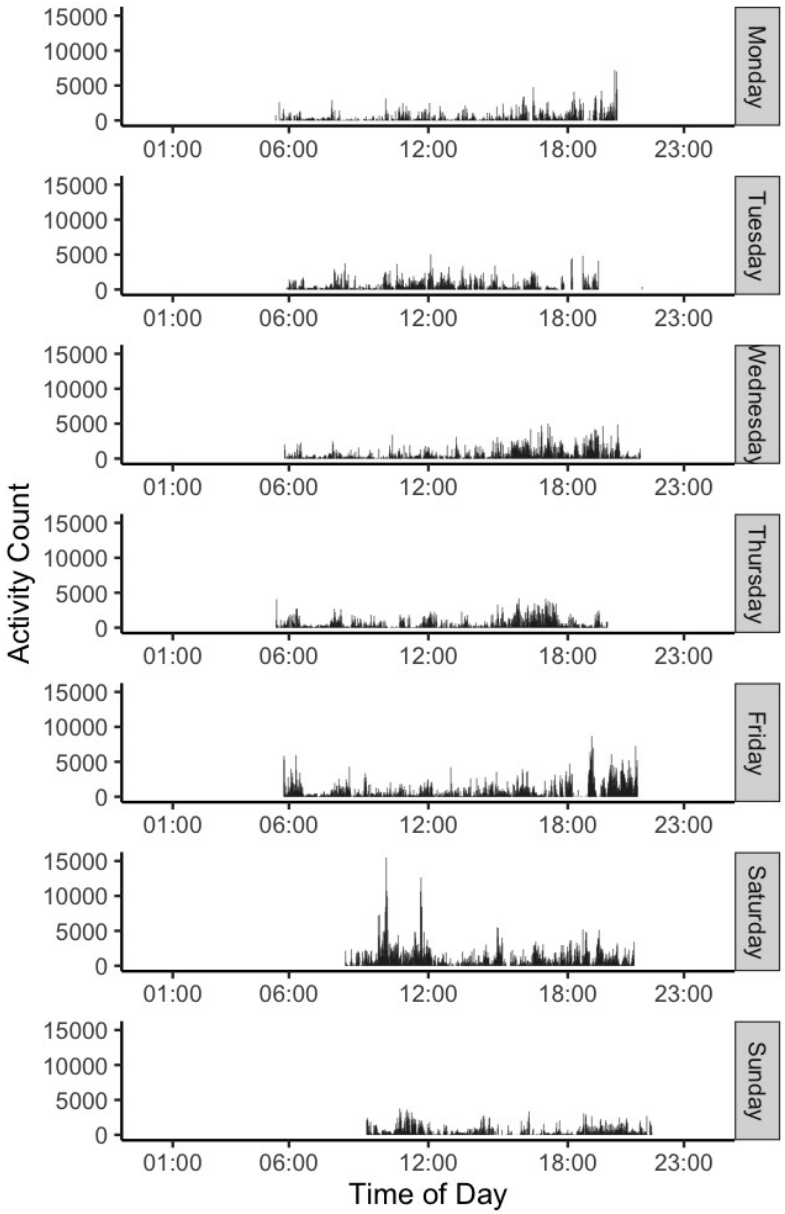
[1] Leroux A, Di J, Smirnova E, et al. Organizing and Analyzing the Activity Data in NHANES. Statistics in Biosciences. 2019. 10.1007/s12561-018-09229-9.

NHANES accelerometry: *rnhanesdata* package

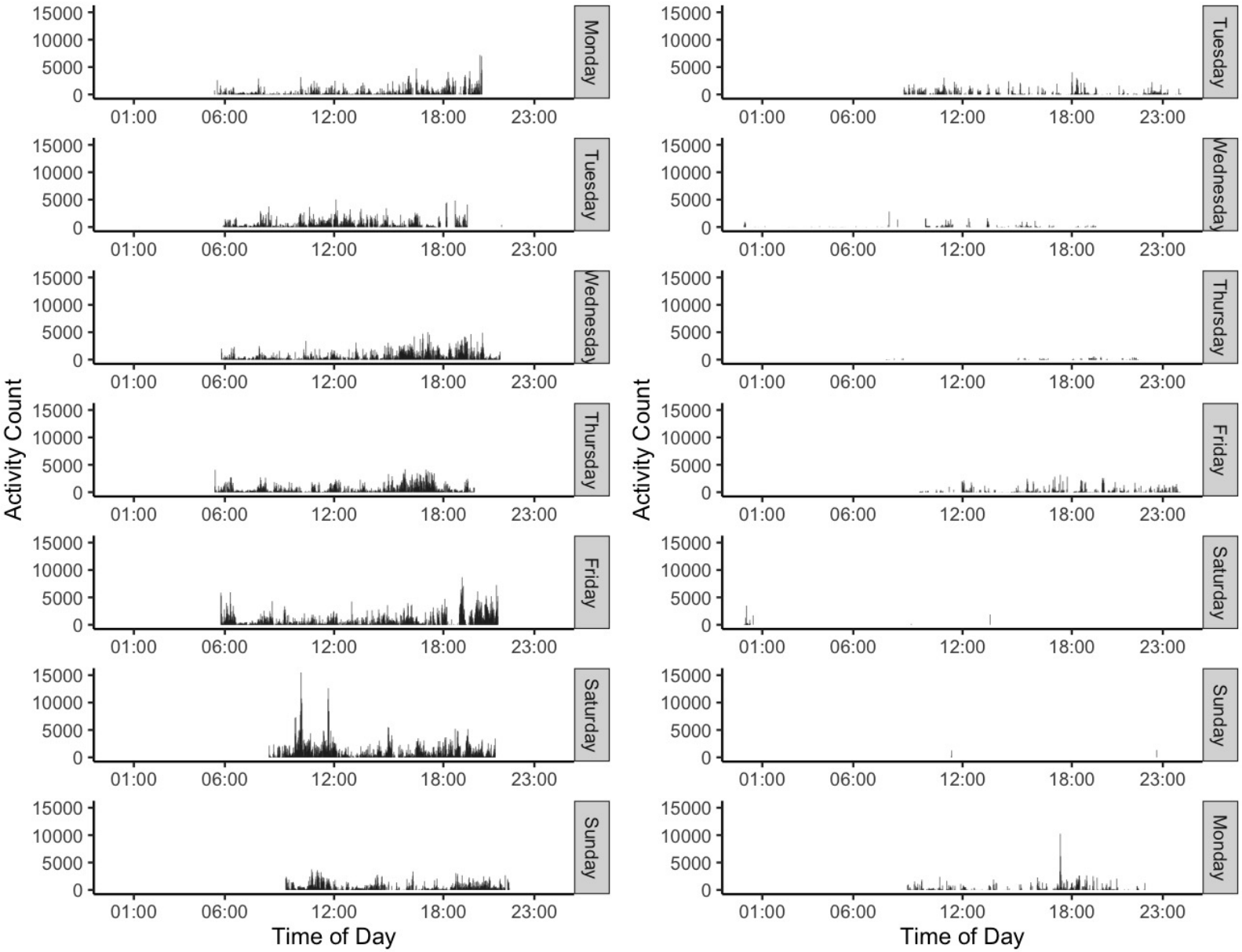
(1) Processed data	
processed physical activity data	"PAXINTEN_C.rda" and "PAXINTEN_D.rda"
wear/non-wear flags data	"Flags_C.rda" and "Flags_D.rda"
covariates data	"Covariate_C.rda" and "Covariate_D.rda"
mortality data	"Mortality_2011_C.rda" and "Mortality_2011_D.rda"
(2) Data processing functions	
NHANES activity processing code	"process_accel()"
NHANES wear/non-wear flag code	"process_flags()"
NHANES mortality	"process_mort()"
NHANES data merging	"process_covar()"
(3) Helper functions	
Calculate survey weights on subsets	"reweight_accel()"
Identify "good" days of accelerometry data	"exclude_accel()"
(4) Raw data	
NHANES covariate data	"ALQ_C.XPT","ALQ_D.XPT", "BMX_C.XPT","BMX_D.XPT", ...
NHANES linked mortality data	"NHANES_2005_2006_MORT_2011_PUBLIC.dat" "NHANES_2003_2004_MORT_2011_PUBLIC.dat"

[1] Leroux A, Di J, Smirnova E, et al. Organizing and Analyzing the Activity Data in NHANES. Statistics in Biosciences. 2019. 10.1007/s12561-018-09229-9.

Macro Scale Accelerometry Data: Compliant Participant



Macro Scale Accelerometry Data: Compliant vs Non-Compliant Participant



NHANES accelerometry: Analysis Procedure

- Load and merge any relevant data by unique identifier (SEQN)
- Apply exclusion criteria
  - Data quality: 1) device calibration (PAXCAL); and 2) NHANES supplied flag (PAXSTAT)
  - Adherence to wear-time protocol. Most studies use ≥ 10 hours.
  - Sufficient number of days of data. Most studies use ≥ 3 days of data with ≥ 10 hours of wear.
  - Other criteria: missing data, etc.
- Calculate features of interest
- Incorporate survey design? Survey weights?
- Regresison, machine learning, etc.

Macro Scale Accelerometry Data: Features and Dimensionality Reduction

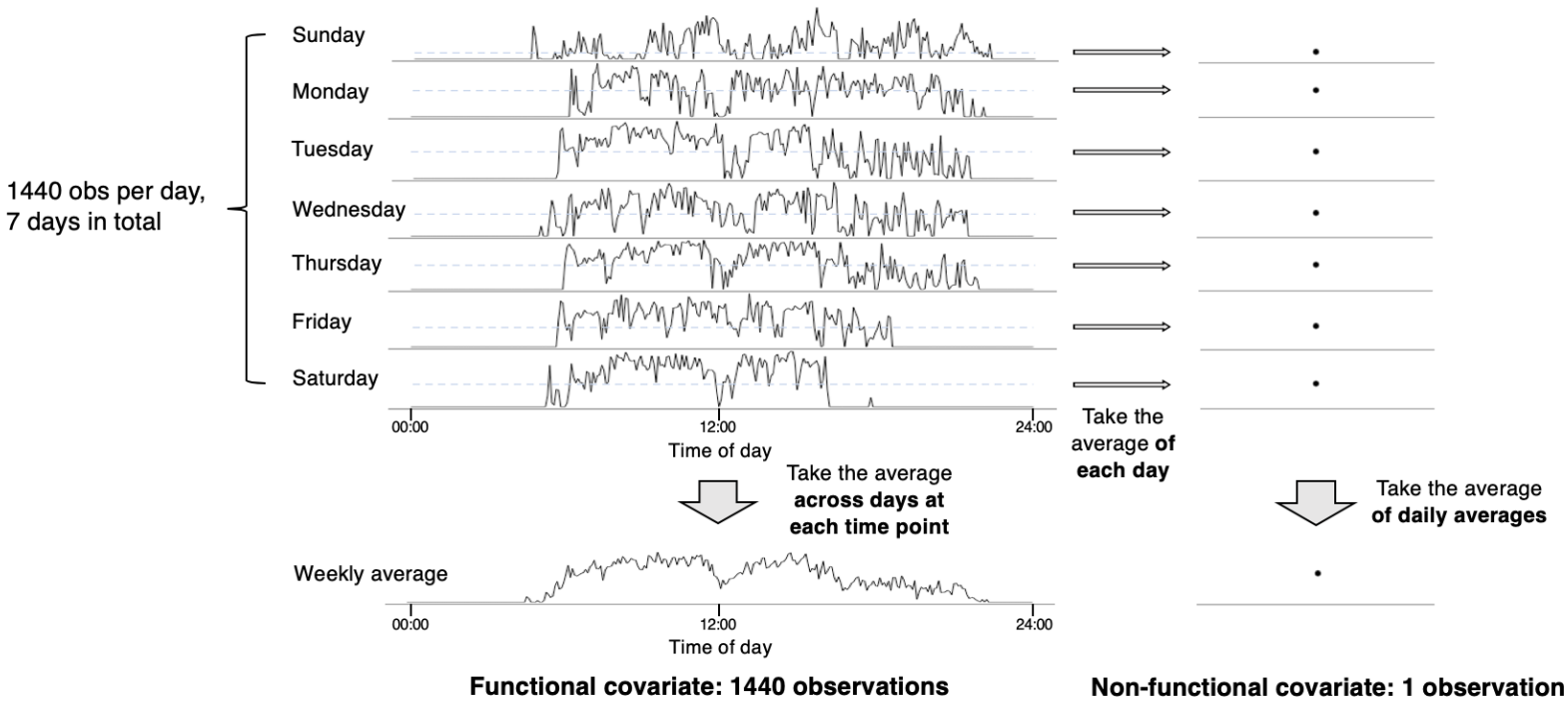


Image credit: Erjia Cui

NHANES accelerometry: Features

- Analyzing activity profiles falls under "Functional Data Analysis"
- Current standard: calculate single summaries of the data
  - Volume of activity[2]
    - Time spent in sedentary/light/moderate/vigorous behaviours. Require population-specific studies to determine thresholds.
    - Average daily total activity count (TAC). A proxy for total volume of moderate/vigorous activity
    - Average daily total log activity count (TLAC). A proxy for total volume of low/light activity
  - Patterns of activity
    - Fragmentation measures[3]
    - Timing of physical activity (activity profiles)

---

[3]Varma VR, Dey D, Leroux A, et al. Total volume of physical activity: TAC, TLAC or TAC(λ). Prev Med. 2017;106:233-235.

[3]Di, J., Leroux, A., Urbanek, J., et al. Patterns of sedentary and active time accumulation are associated with mortality in US adults: The NHANES study. bioRxiv: 182337.

Predicting 5-year mortality in NHANES

- See vignette "5-year mortality prediction in NHANES with Lab Measurements" in the *rnhanesdata* package

```
browseVignettes(package="rnhanesdata")
```

- Assesses the predictive value of scalar accelerometry features compared to standard predictors of mortality

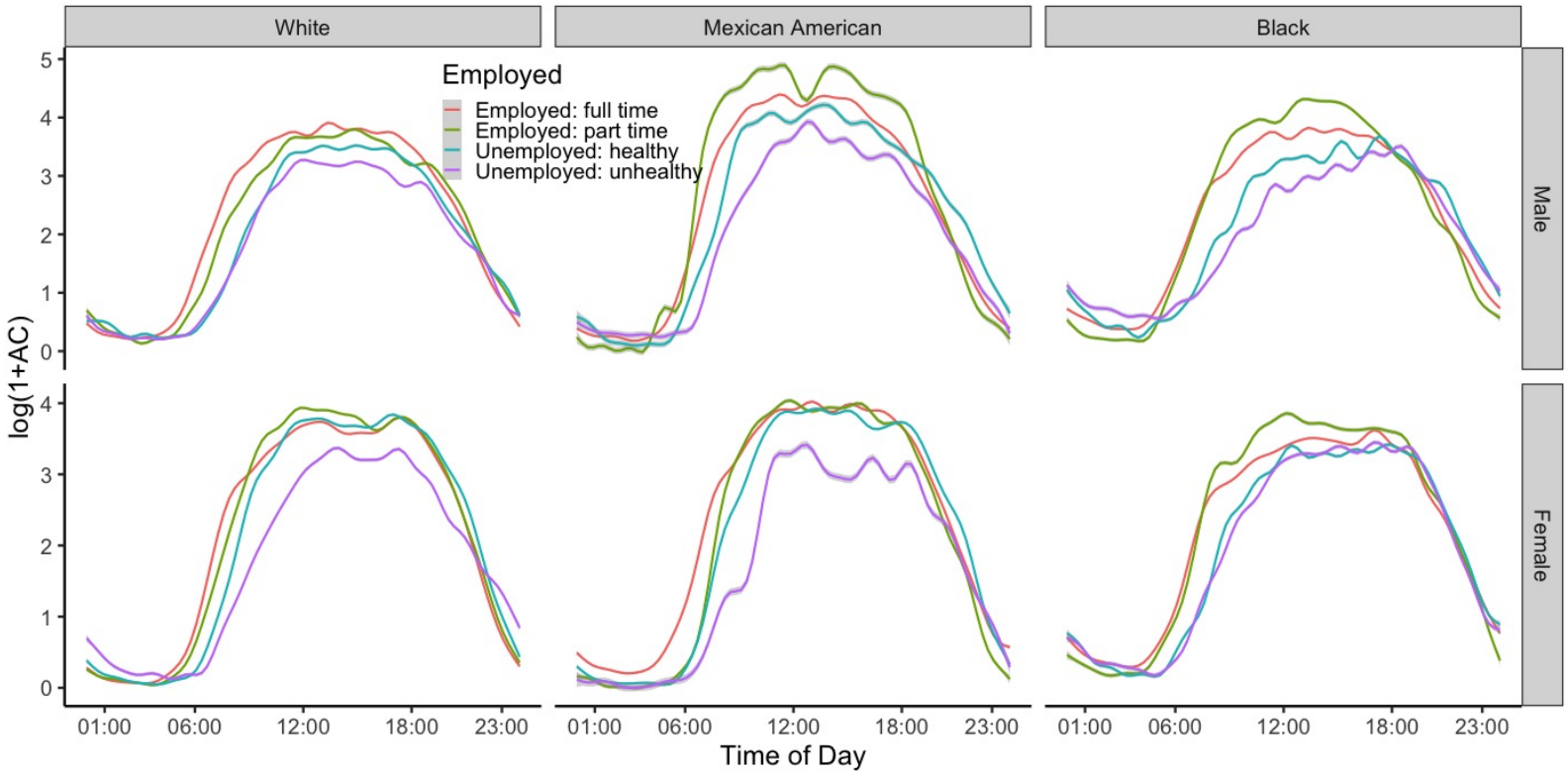
Accelerometry as Functional Data

- Fundamentaly, we think of data as "functional" data when there is some (underlying) smooth process which we measure
  - child growth (height and weight)
  - heart rate + other biosignals
  - neuroimaging
- Physical activity is inherently a “continuous" process
- In regression analyses, functional data can either be the outcome of interest or a predictor
- In this course we'll discuss 2 methods
  - Function-on-scalar regression (FoSR)
  - Scalar-on-function regression (SoFR)

Physical Activity and Employment

- Scientific question: How do patterns of low/light activity vary by employment status among 30-55 year olds? Do these patterns differ by Race? By gender?
- Non-model based approach: group people into employment, race, and gender categories, take average at each time of the day

Physical Activity and Employment



Physical Activity and Aging

- Scientific question: How do patterns of low/light activity change with age? Do these patterns differ between weekends and weekdays? By gender?

Physical Activity and Aging

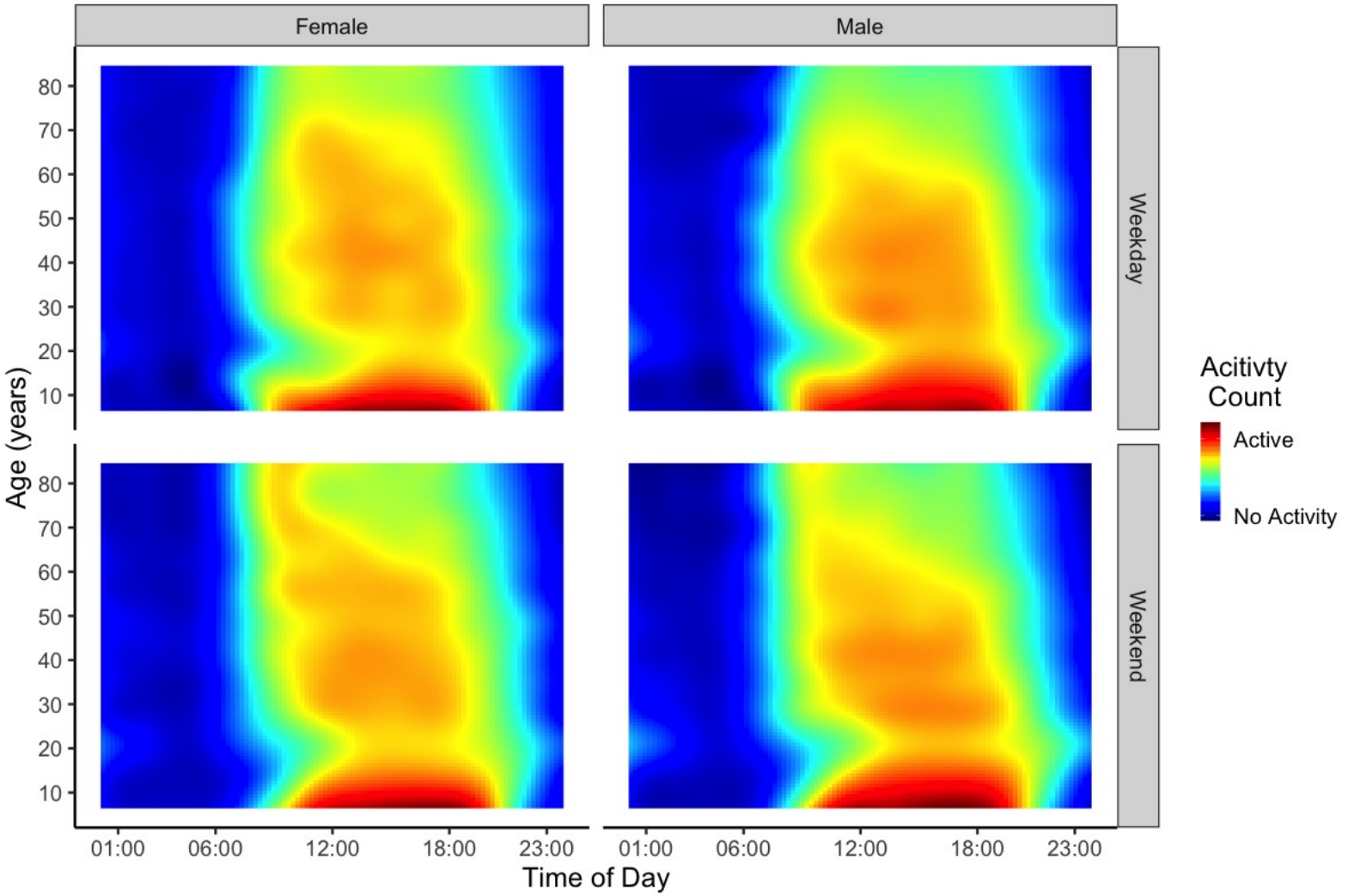
- Scientific question: How do patterns of low/light activity change with age? Do these patterns differ between weekends and weekdays? By gender?
- $i = 1, \ldots, N$ subject, $j = 1, \ldots, J_i$ day, $t = 1, \ldots, 1440$ minute of the day
- Let M_i be the indicator that subject i is male, and D_{ij} be the indicator that day j for subject i is a weekend

$$\log(1 + \text{AC}_{ij}(t)) = f_{\mathbf{0}}(t) + f_{\mathbf{1}}(t, \text{Age}_i) M_i D_{ij} + f_{\mathbf{2}}(t, \text{Age}_i) M_i (1 - D_{ij}) + f_{\mathbf{3}}(t, \text{Age}_i)(1 - M_i) D_{ij} + f_{\mathbf{4}}(t, \text{Age}_i)(1 - M_i)(1 - D_{ij}) + \epsilon_{ij}(t)$$

$$\epsilon_i(t) \sim N(0, \sigma^{\mathbf{2}})$$

- PA modelled as smooth function of age and time of day separately for each gender and weekday vs weekend
- Ignores within subject correlation

Physical Activity and aging



Functional regression in *R*

- In *R*, the main software package designed specifically for fitting functional regression models is the *refund* package[4]
- Basically, *refund* contains wrapper functions for the *gam()/bam()* functions in the *mgcv* package[5][6]
- Reduces user burden: data transformations, numeric integration, extracting estimated functional coefficients, and dealing with identifiability constraints.

---

[6]Goldsmith J, Scheipl F, Huang L, et al. (2018). refund: Regression with Functional Data. R package version 0.1-17. `https://CRAN.R-project.org/package=refund`

[6]Wood, SN (2017). Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC.

[6]Wood, SN (2018). mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation and GAMMs by REML/PQL. R package version 1.8-25. `https://CRAN.R-project.org/package=mgcv`

Class Project Ideas

- Missing data
  - (Medium/Easy) Look in depth at missing data patterns (within a day, across days, by age, etc.)
  - (Hard) Impute missing activity data at the minute level
- Functional Regression [I have ideas for each of these]
  - (Hard) Propose and estimate a functional transition model
  - (Hard/Medium) Model time dependent fragmentation
  - (Hard/Medium) Model PA profiles using models for zero inflated count data
- (Medium) Try to beat TAC as a predictor of 5-year mortality
- (Medium) Develop a Shiny application for impressive visualization of the data
- (Medium) Assess the weekend vs. weekday effect of PA on an outcome
- (Medium/Easy) Associate "adjusted" fPCA scores with mortality (patterns unaccounted for by age-, sex- specfic average activity)