Functional Data Analysis for Accelerometer Data

Jeffrey S. Morris

University of Texas M.D. Anderson Cancer Center, Houston, TX

Joint work with: **Louise Ryan, Cassandra Arroyo, Brent Coull**, and **Steve Gortmaker**

Planet Health Study

- **Planet Health:** school-based intervention study to reduce obesity in middle school youth by changing key physical activity and dietary risk factors.
- **Design:**
  - **10 schools** paired up, randomized to intervention/control
  - **1295 children:** nutritional, behavioral, and health-related outcomes measured at baseline (F95) and follow-up (S97)
- **Challenge:** How to measure physical activity levels?
  - Questionnaires
  - **Accelerometers:** objectively quantify activity levels
- **256 children:** monitored using TriTrac-R3D activity monitor for one or two 4-day sessions in S97

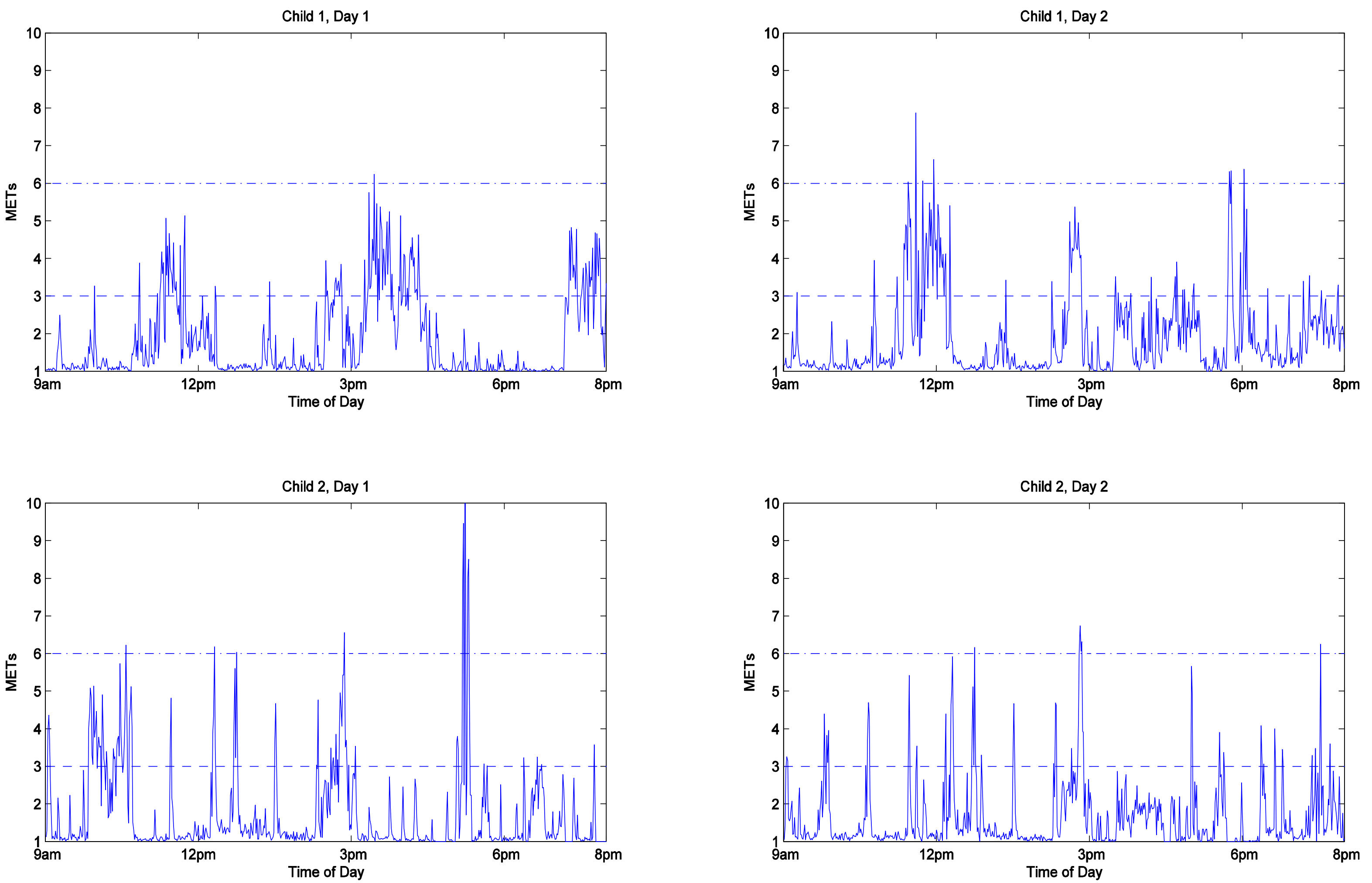
Accelerometers

- **Accelerometers:** small motion sensors that digitally record minute-by-minute activity levels
  - Increasingly used in large surveillance studies and intervention trials to objectively monitor activity
    - Monitoring children's activity levels and correlating with obesity (Abbott & Davies 2004; Going, et al. 2003; Goldfield, et al. 2000; Rowlands, et al. 2000; Rowlands, Eston, and Ingledew 1999)
    - Monitoring physical activity of older adults (Talbot, et al. 2003)
    - Index wandering behavior in Alzheimer patients (Algase, … 2003)
    - Have potential for expanded use, e.g. in monitoring side effects of drugs, sleep studies, …
- Numerous validation studies done
  - 3D accelerometers do better than 1D
  - Do reasonable job of measuring energy expenditure, when correlated with gold standard, for many types of activities
  - Underestimate energy expenditure for certain types of activity, e.g. sedentary activities, climbing, biking, weightlifting

TriTrac-R3D Accelerometer

- TriTrac-R3D: pocket-sized sensor worn on hip
  - Minute-by-minute record of motion in 3 planes
  - Condensed into single activity level measurement/minute, on either acceleration or "METs" scale (multiple of resting metabolic rate)
- METs for different activities (Rowlands, et al. 2004)
  - 1 MET = minimum activity level
  - 3-6 MET = moderately intense activity
  - >6 MET = vigorous activity

Some Accelerometer Profiles



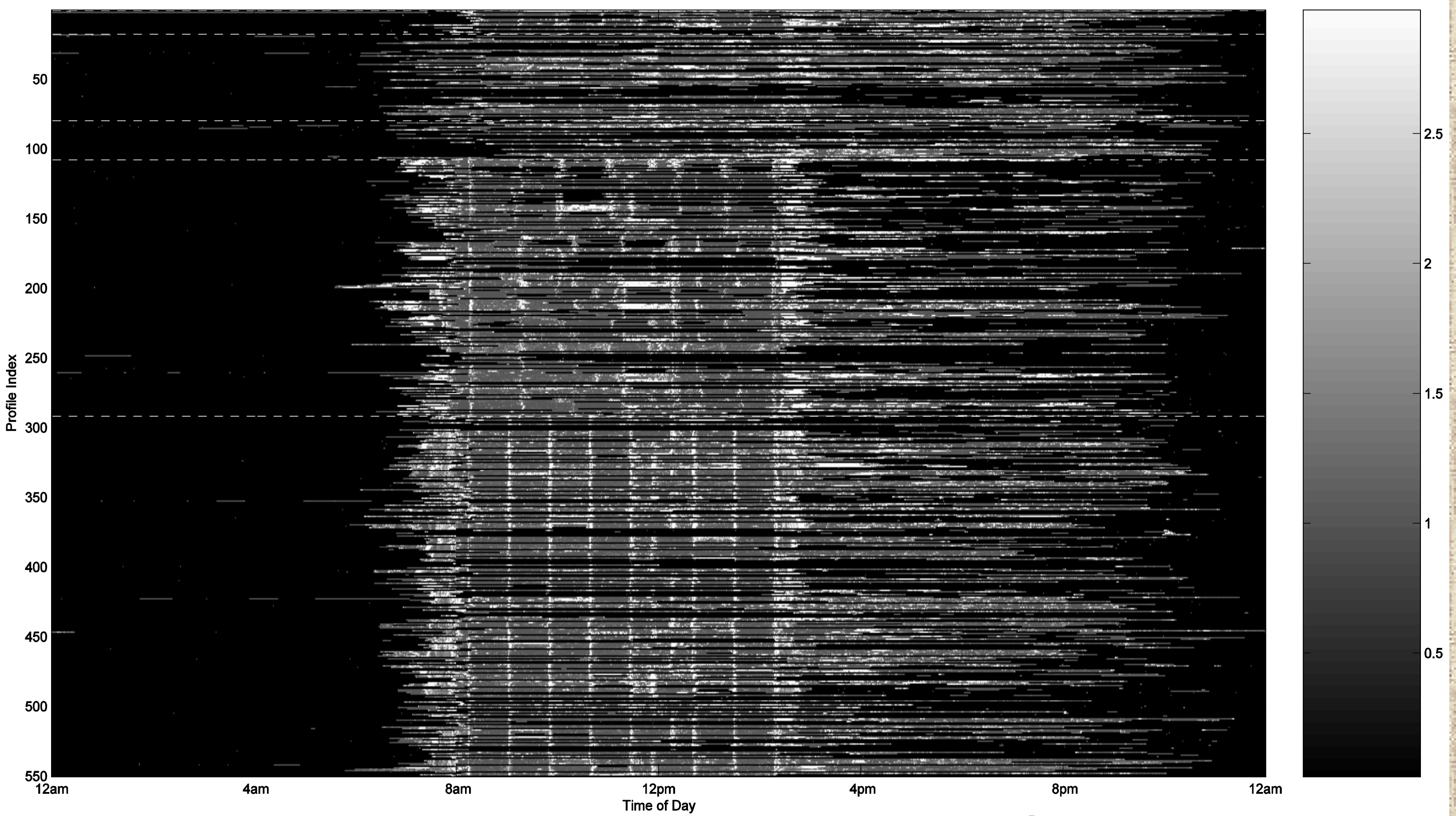
Analyzing Accelerometer Data

- **Major challenge:** How to analyze these data?
- **Standard strategy:** Extract simple **summary measures** from the profiles, then limit the analysis to these.
  1. Average daily activity levels
  2. Proportion of time with activity levels corresponding to sedentary, moderately vigorous, and vigorous activities
  3. 30-minute averages
- **Benefit:** simplifies data into recognized form.
- **Limitations** of these approaches:
  – Do not make full use of information in functional data
  – Cannot effectively deal with incomplete profiles
- We have analyzed these data using a new method that can model the functions themselves and deal with incomplete profiles.

Accelerometer Data

- **Here, we focus only on weekday profiles from children at control schools**
  - **550 profiles from 112 children, 1440 meas./profile**
  - **Enormous amount of data; can be displayed using Heatmap**
- **Lots of *missing data***
  - **Measurement coded as *missing* for time periods for which there was virtually no activity (vector magnitude <10) for at least 30 consecutive minutes**
  - **We focus on profiles >50% complete from 9am-8pm**
    - **292 profiles from 106 children**
  - **We will return to this missingness issue later.**

Heatmap



Accelerometer Data

Accelerometer Data

- **Other measurements:**
  - **Child-level covariates**: school, race, gender, age, weight, height, BMI, % body fat, avg hrs of TV/week
  - **Day-level covariates:** day of week, calendar date
- **Goals:**
  1. Assess how activity levels vary throughout day, across schools, across different days of the week, over time from early to late Spring, and across various child-level covariates.
  2. Assess relative variability in activity levels from day-to-day and child-to-child, in order to guide future study design.

– If we had just 1 measurement/profile, e.g. average daily activity level, we could answer these questions using linear mixed models.

Linear Mixed Models

Linear Mixed Model (Laird and Ware, 1982):

$$\underbrace{Y}_{N\times 1} = \underbrace{X}_{N\times p}\overbrace{\beta}^{p\times 1} + \underbrace{Z}_{N\times m}\overbrace{u}^{m\times 1} + \underbrace{e}_{N\times 1}$$

$$u \sim N(0, \overbrace{P}^{m\times m})$$
$$e \sim N(0, \underbrace{R}_{N\times N})$$

- **Fixed effects part, $X\beta$, accommodate a broad class of mean structures, including main effects, interactions, and linear coefficients.**
- **Random effects part, Zu, provide a convenient mechanism for modeling correlation among the N observations.**

Example: Mixed Model

Suppose we have the average daily activity level computed from 292 daily profiles from 106 children.

$$Y_i = B_0 + X_{i1}B_1 + X_{i2}B_2 + \sum_{k=1}^{106} Z_{ik}U_k + E_i$$

- **Y_i = log(daily activity level) for profile i**
- **X_{i1} = 1 if male, -1 if female; X_{i2}=1 before DST, -1 after**

  B_1 (B_2) : Effects of gender (DST) on activity levels
- **Z_{ik}= 1 if profile i is from child k, 0 otherwise**

  U_k = random effect for child k; $U_k \sim N(0,\sigma_c^2)$
- **E_i = residual error for profile i; $E_i \sim N(0,\sigma_d^2)$**

Functional Data Analysis

- Functional Data Analysis (Ramsay&Silverman 1997)
  - Statistical methods to analyze functional data, where "units of observation" are actually continuous curves.
  - Approach: Model entire functions instead of arbitrarily computed summary measures.
- Potential to extract more information from data
  - Identify what times of day children are most active.
  - Allow possibility that covariates may have different effects at different times of the day.
  - Also allows us to potentially incorporate partial information contained in incomplete profiles.

Functional Mixed Model (FMM)

Suppose we observe a sample of N curves, $Y_i(t)$, i=1, …, N

$$U_k(t) \sim GP\{0, Q(t_1, t_2)\}$$
$$E_i(t) \sim GP\{0, S(t_1, t_2)\}$$

$$Y_i(t) = B_0(t) + \sum_{j=1}^{p} X_{ij} B_j(t) + \sum_{k=1}^{q} Z_{ik} U_k(t) + E_i(t)$$

- **$B_j(t)$: effect of covariate j on activity levels at time t**
- **$U_k(t)$: random effect for child k at time t**
- **Q, S: account for child-to-child and day-to-day var.**
- **No restrictions (e.g. linear) placed on form of fixed/random effect functions and covariances.**
  - **Provides flexibility necessary to model irregular functions**

Example: Model

Let $Y_i(t)$ be a (log) accelerometer profile (from 9am-8pm). Recall we have 292 profiles from 106 children

$$Y_i(t) = B_0(t) + \sum_{j=1}^{13} X_{ij} B_j(t) + \sum_{k=1}^{106} Z_{ik} U_k(t) + E_i(t)$$

- *Covariates modeled*: School (4), gender, % body fat, BMI, day-of-week (4), DST, avg tv hrs/wk
- Examples:
  - For Gender: X_{i5}=1 if boy, -1 if girl
    $B_5(t)$: add to mean log profile if boy, subtract if girl
  - For TV hours: X_{i13}= {$TVHrs_i$ – mean(TVHrs)}/sd(TVHrs)
    $B_{13}(t)$: add (subtract) to mean log profile for each standard deviation above (below) the mean TVHrs.
- Note: $\exp\{B_j(t)\}$ is multiplicative effect on activity levels

Wavelet-Based Functional Mixed Models

- **Morris and Carroll (2004): Describe method for fitting functional mixed models that is appropriate for irregular (spiky) functional data.**
  - **"Wavelet-based functional mixed models" (WFMM)**
  - **Uses mathematical constructs called "wavelets" to efficiently represent the functions.**
  - **Yields estimates and Bayesian inference for model quantities**
- **Problem: WFMM requires *complete profiles***
  - **Cannot incorporate info. from incomplete profiles**

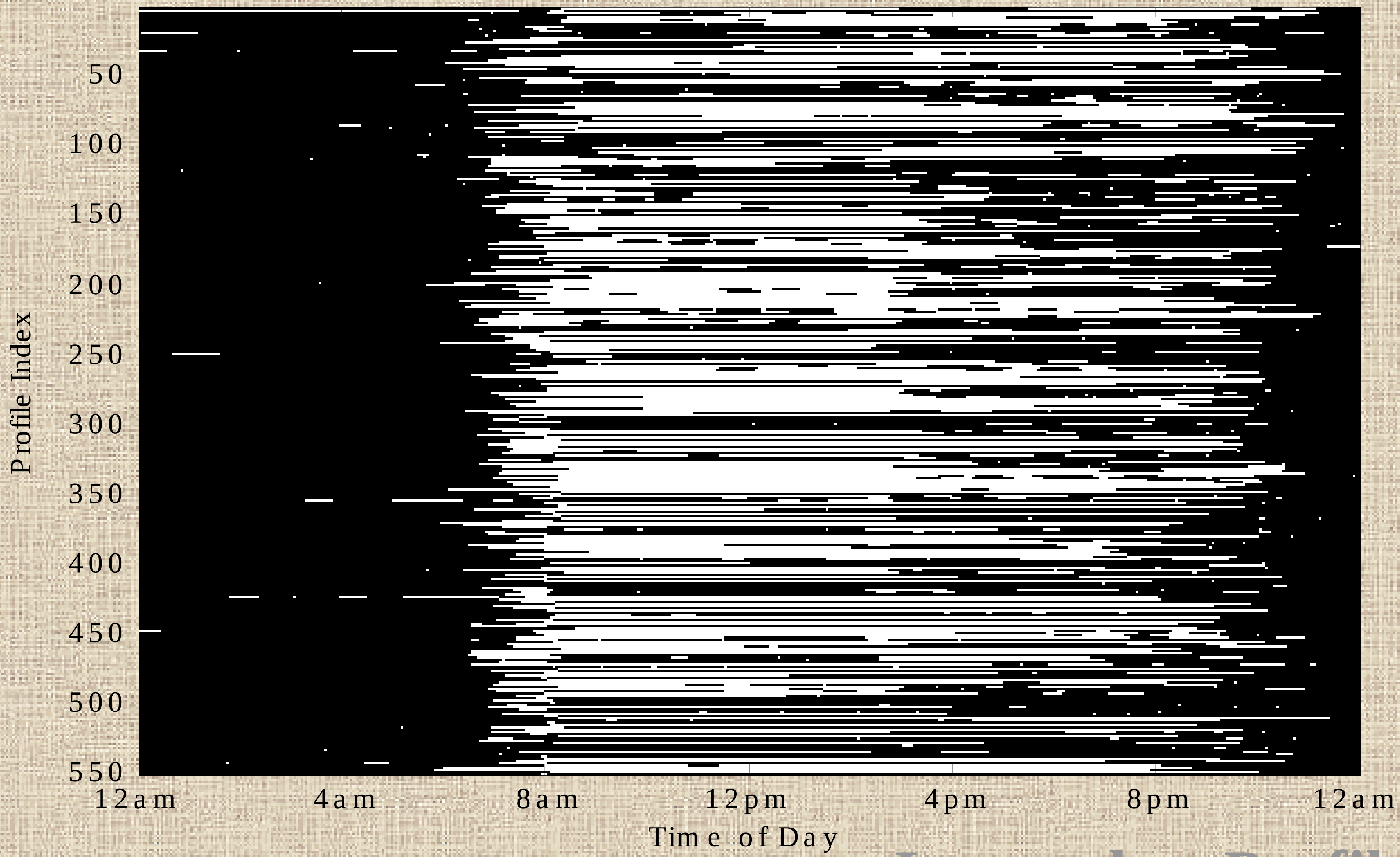
Incomplete Profiles

- **Lots of missing data (Missing Data)**

  Example of *incomplete profile*

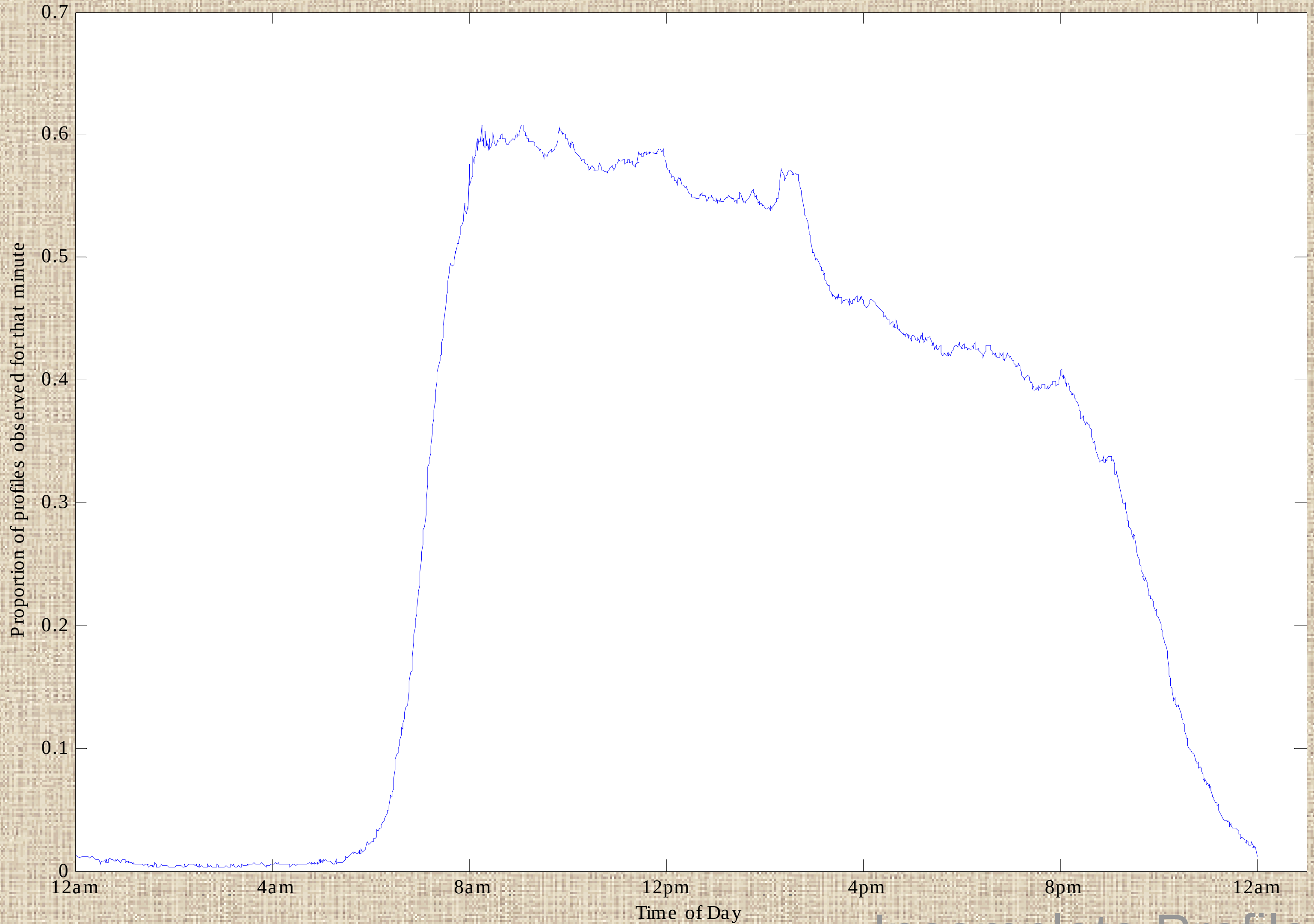
- **We focus on 9am-8pm (Missingness vs. Time-of-Day)**
  - **Most profiles are missing outside this region.**
- **WFMM can only be applied to *complete profiles* (with no missing regions)**
  - **Only 95 of the 292 profiles are *complete* from 9am to 8pm**
- **How do we incorporate information from other 197 *incomplete profiles* ?**

Missing Data



Incomplete Profiles

Missingness vs. Time-of-Day

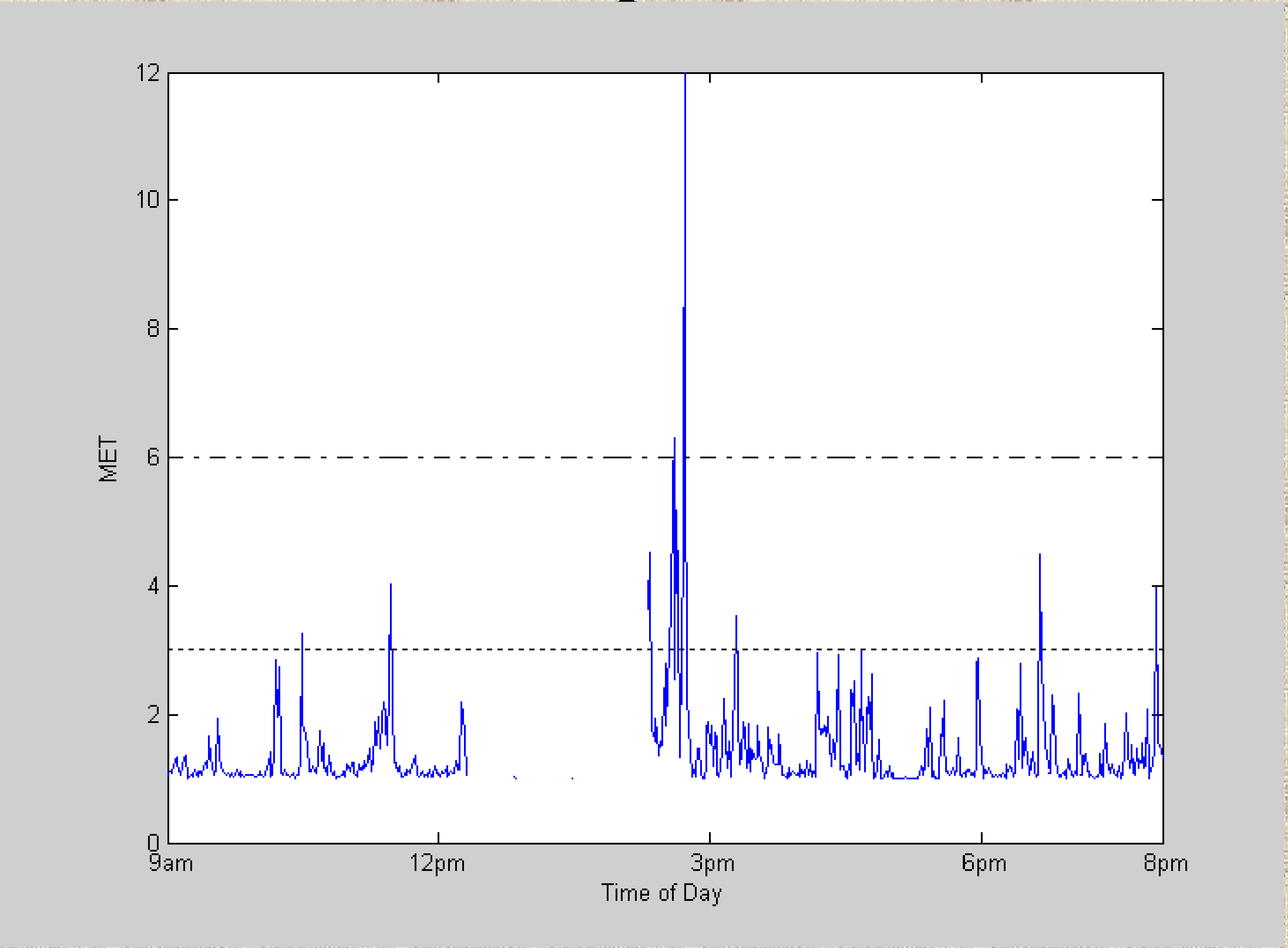


Incomplete Profiles

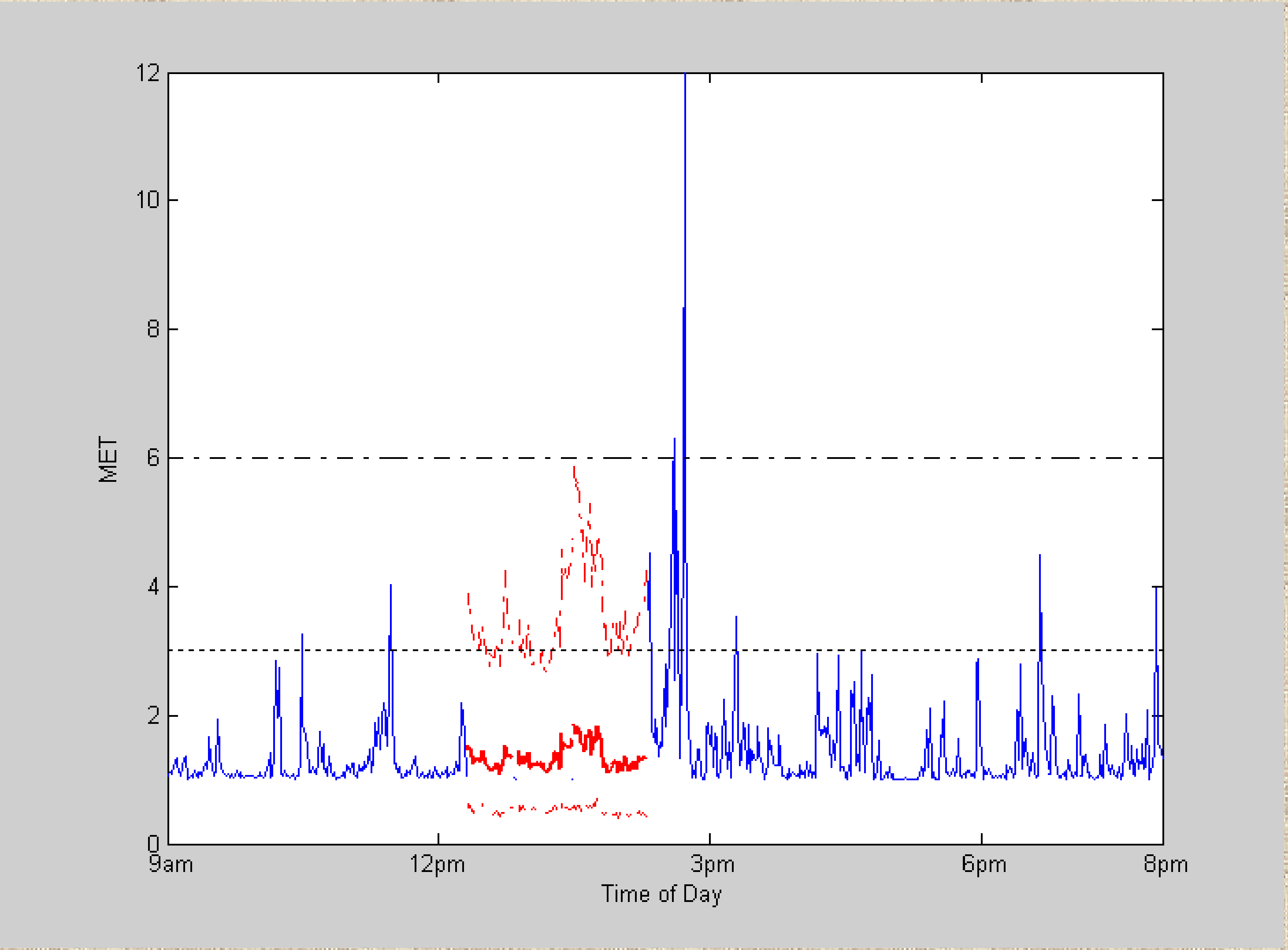
Approach: Incomplete Profiles

1. First fit model to *complete profiles*, get estimates and standard errors for model parameters.
2. Use these to estimate *predictive distributions* for the the incomplete profiles (fig)
   - Borrow information about what the curves in these regions look like.
   - Account for child-specific and day-specific covariates.
3. Regress missing data on the observed data to obtain *imputation distribution* for missing regions (fig)
   - Borrow information from nearby times in incomplete profiles.
   - Makes predictions for missing regions "connected" with observed.
4. Supplement WFMM with step to stochastically impute values for missing data.
   - Inference appropriately accounts for uncertainty in imputation

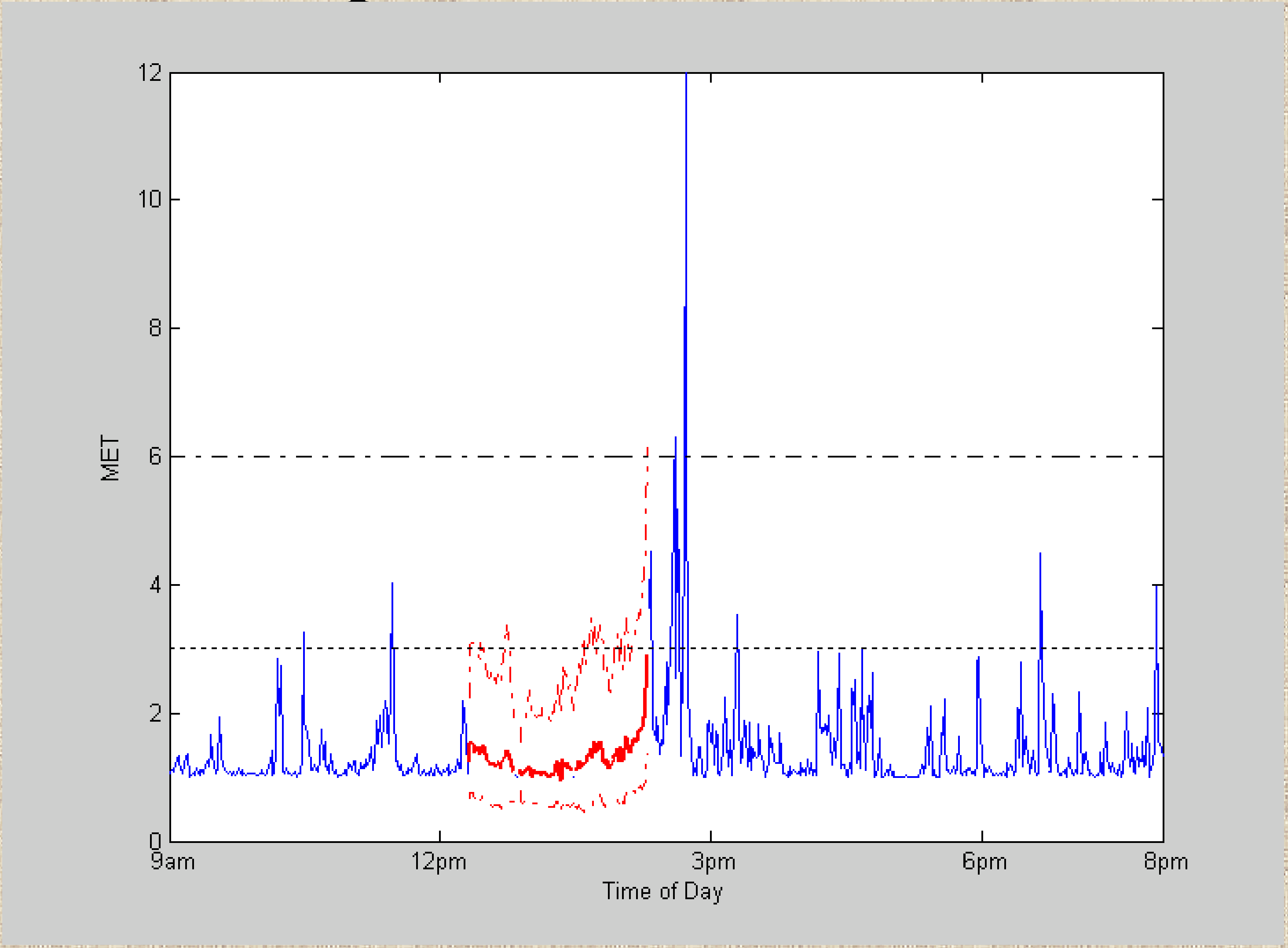
Incomplete Profile



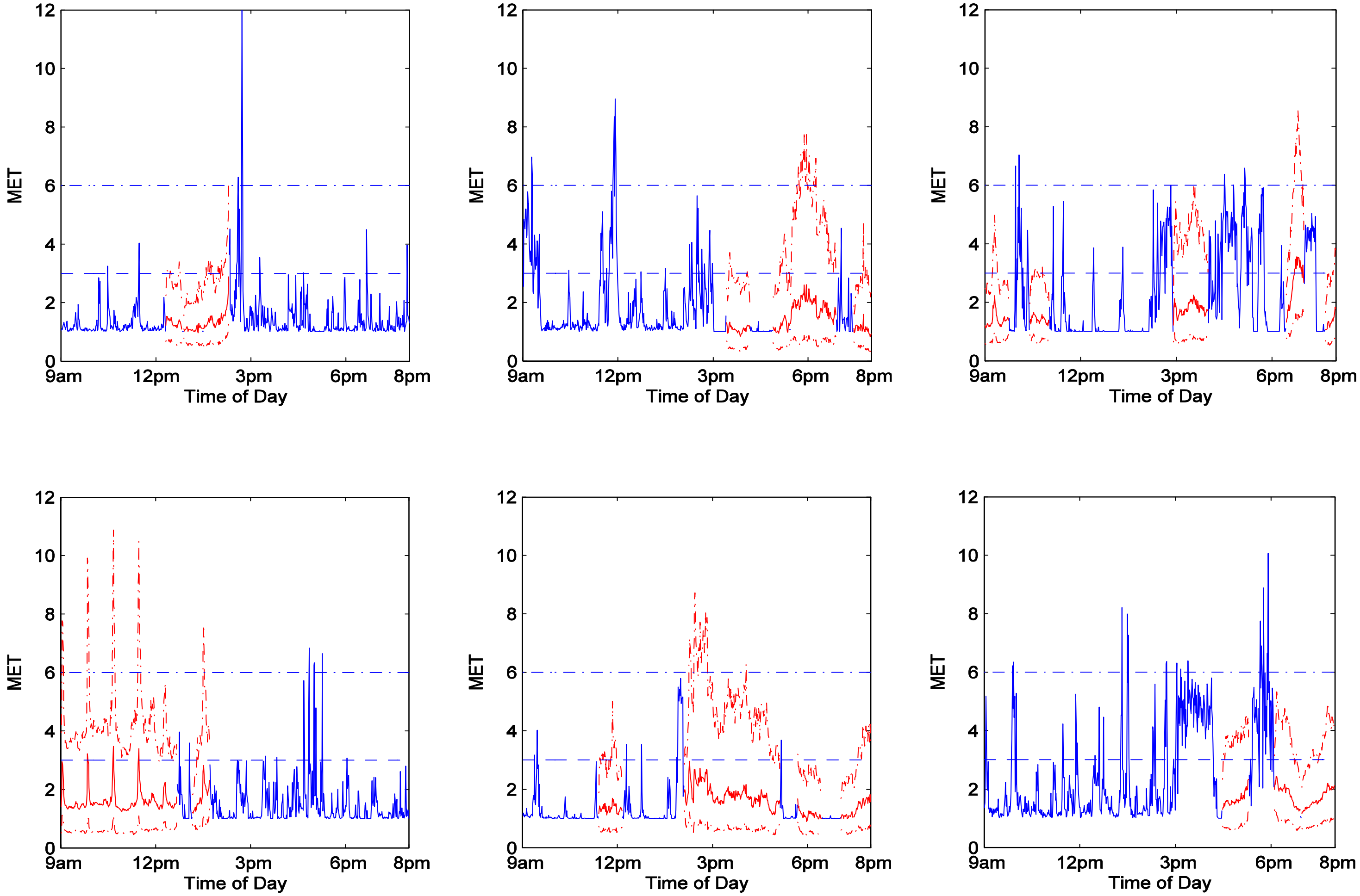
Predictive Distribution



Imputation distribution



Incomplete Profiles



Classical Statistical Inference

Y=data, Θ=parameters (mean functions/covariances)

Likelihood: $f(Y|\Theta)$

Frequentist (classical) approach

- Parameters are *fixed, unknown* quantities
  - Cannot have "distributions"
- Estimate Θ by maximizing the likelihood
- Perform asymptotic inference
  - Assumes "large" sample sizes
- Yields *p-values* and *confidence intervals*, which must be interpreted in terms of "repeated sampling"

Bayesian Statistical Inference

Y=data, Θ=parameters (mean functions/covariances)

Likelihood: $f(Y|\Theta)$

Bayesian approach

- **Parameters are *random* quantities with distribution**

  Prior Distribution: $f(\Theta)$

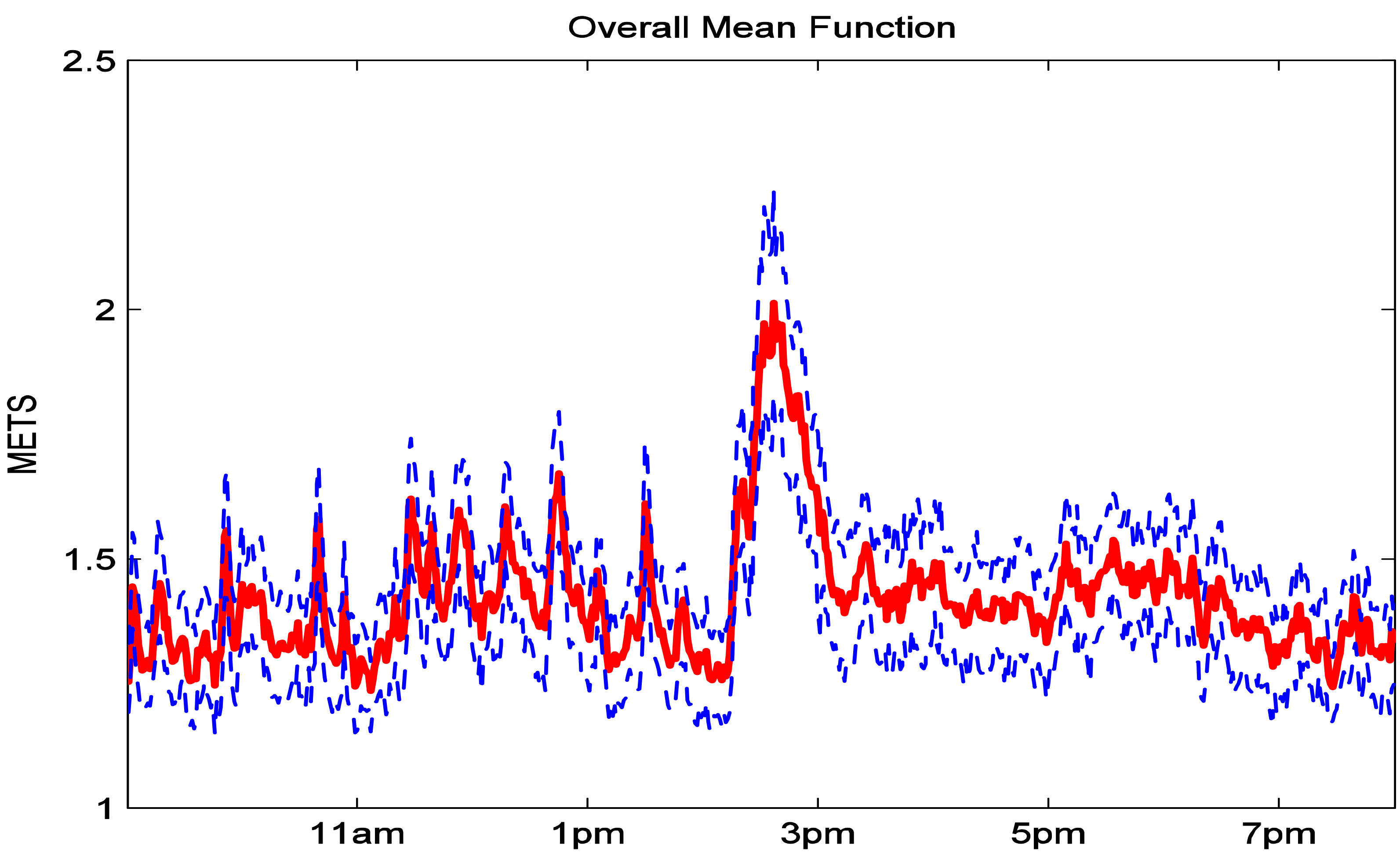
- **Inference based on *posterior distribution* of Θ**
  - **Computed using *Bayes' Rule:*** $f(\Theta|Y) = f(Y|\Theta) \times f(\Theta)/f(Y)$
  - **Yields *posterior probabilities* and *credible intervals*, which can be interpreted as probabilities of parameters**
- **When $f(\Theta|Y)$ not directly computable, MCMC used**
  - **Simulation technique yielding random samples from posterior distribution**

Statistical Inference in FMM

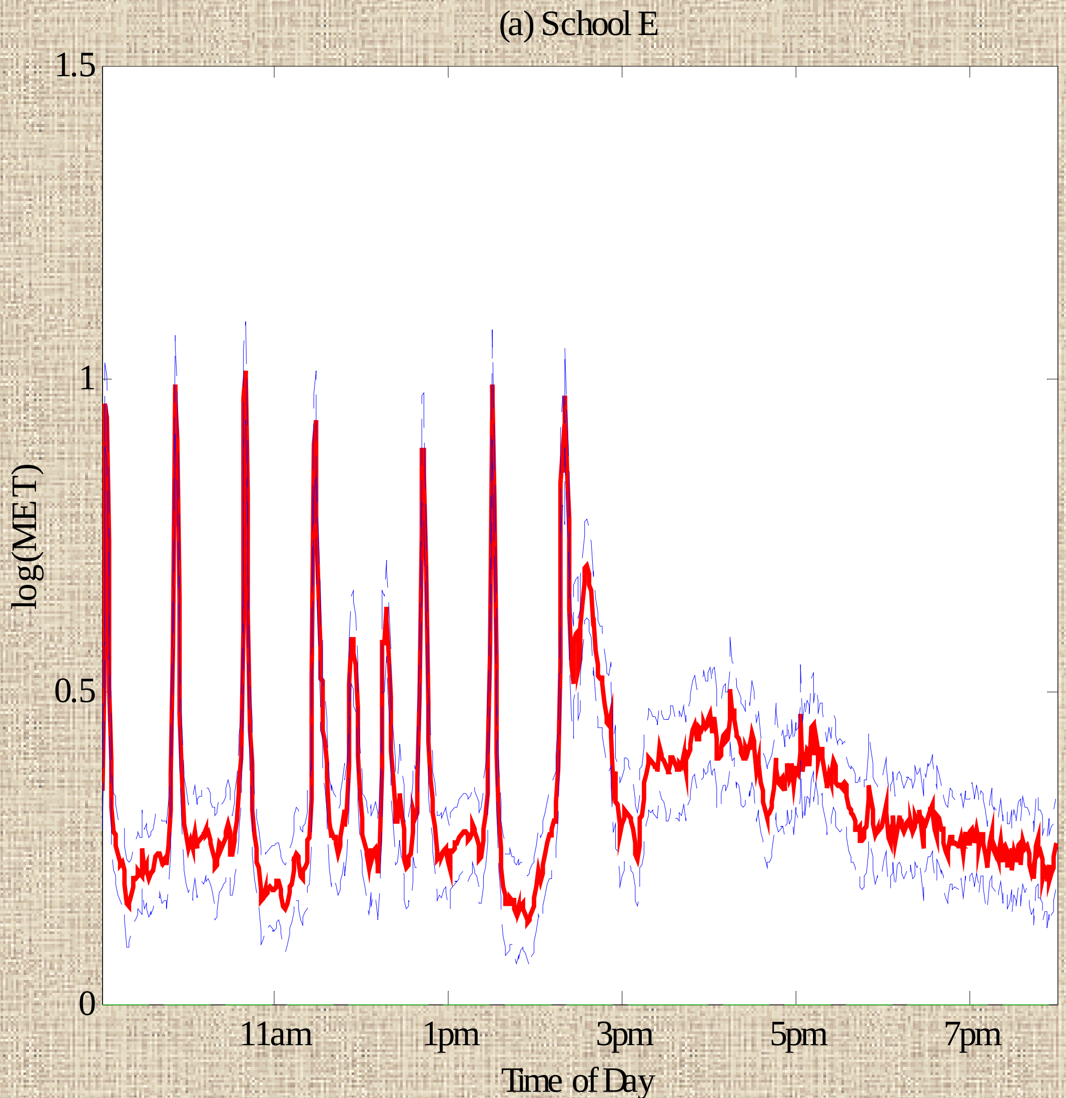
- Prior distributions made "uninformative"
  - Exception: *smoothing parameters*, which can be estimated from the data using an "empirical Bayes" approach
- We use MCMC to draw samples from posterior distributions of $B_i(t)$ functions and covariances.
- For each covariate effect, we compute 90% pointwise credible intervals at each value of t
- We also compute posterior prob. of nonzero effects, aggregating data within following time intervals:

1. All day (9am-8pm)
2. Morning (9am-11:30am)
3. Lunch (11:30am-12:30pm)
4. Afternoon (12:30-2:15pm)
5. Going home (2:15pm-3pm)
6. After school (3pm-5:30pm)
7. Early evening (5:30pm-7pm)
8. Late evening (7pm-8pm)

Results: Overall mean function

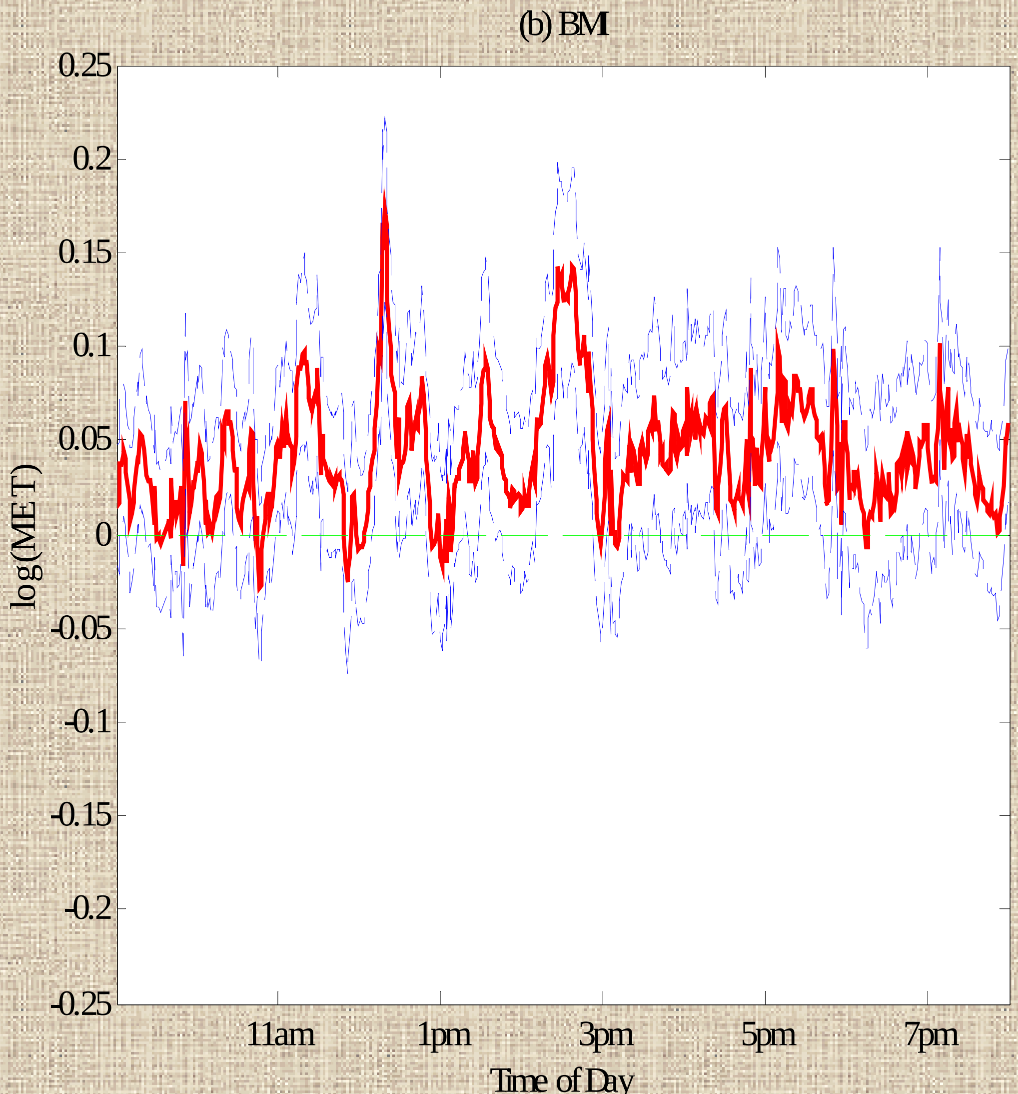


Selected Results: School Effects



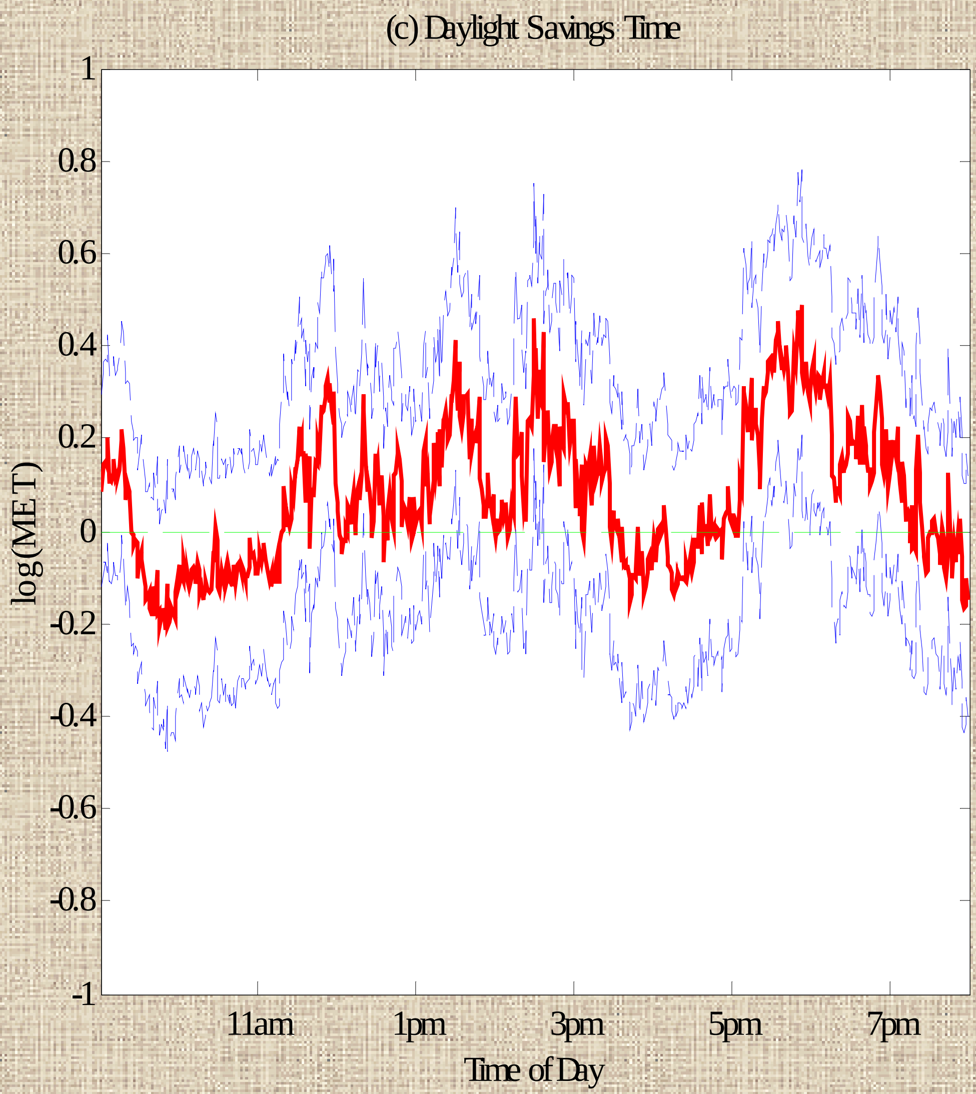
- **School schedules evident in effects**
  - Spikes every 48min (changing classes)
  - 3 lunch periods
  - School out at 2:15pm
- **Not so evident in individual curves**

Selected Results: BMI Effect



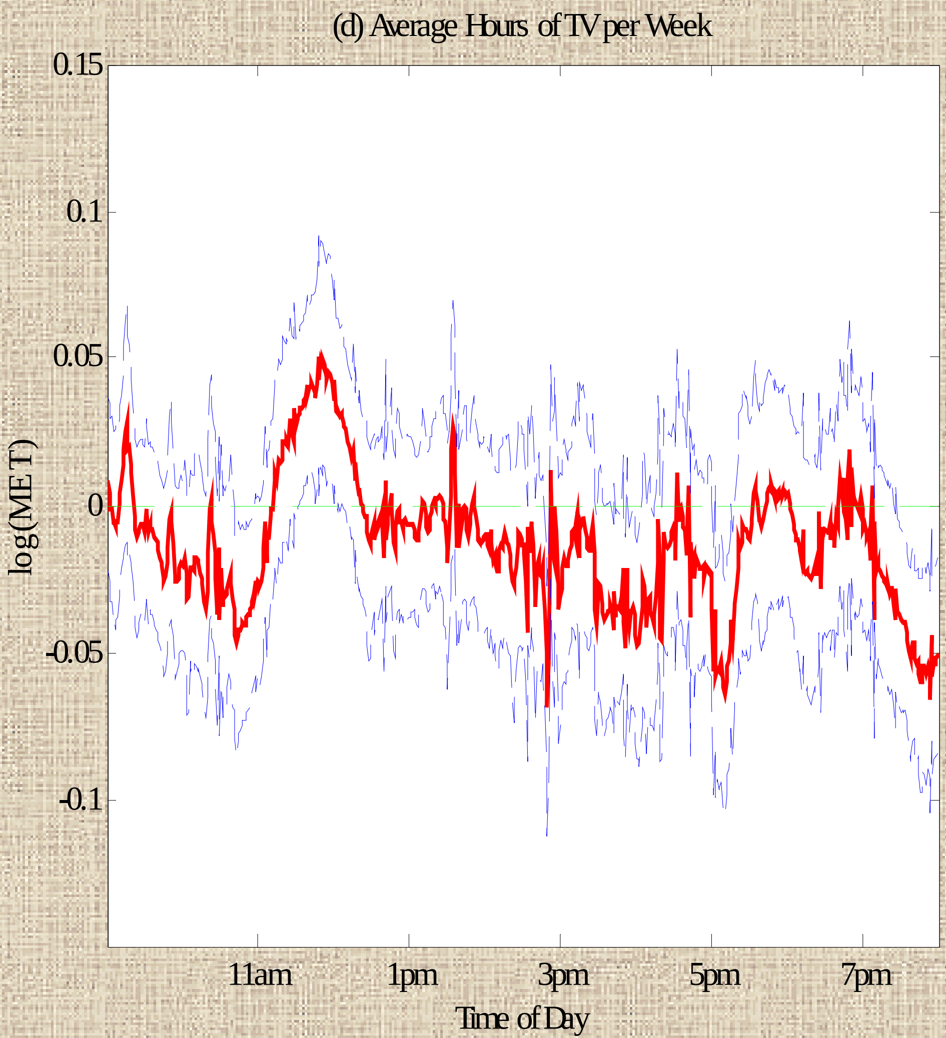
- BMI coded as continuous factor (mean-centered)
- BMI effect positive (*post prob<0.0005*)
  - Higher BMI, more active
  - Preprocessing artifact?
- Should raw activity levels be monitored instead of METs?

Selected Results: DST Effect



- **DST – April 6th**
- **More active after DST (overall 8%, *p=0.062*)**
- **Especially strong:**
  - **As school is letting out (2:15-3:00, 25%, *p=0.03*)**
  - **In early evening (5:30-7:00, 30%, *p=0.01*)**
- **Note: Sunset was**
  - **5:10-6:15 before DST**
  - **7:15-8:10 after DST**

Selected Results: TV hours/wk

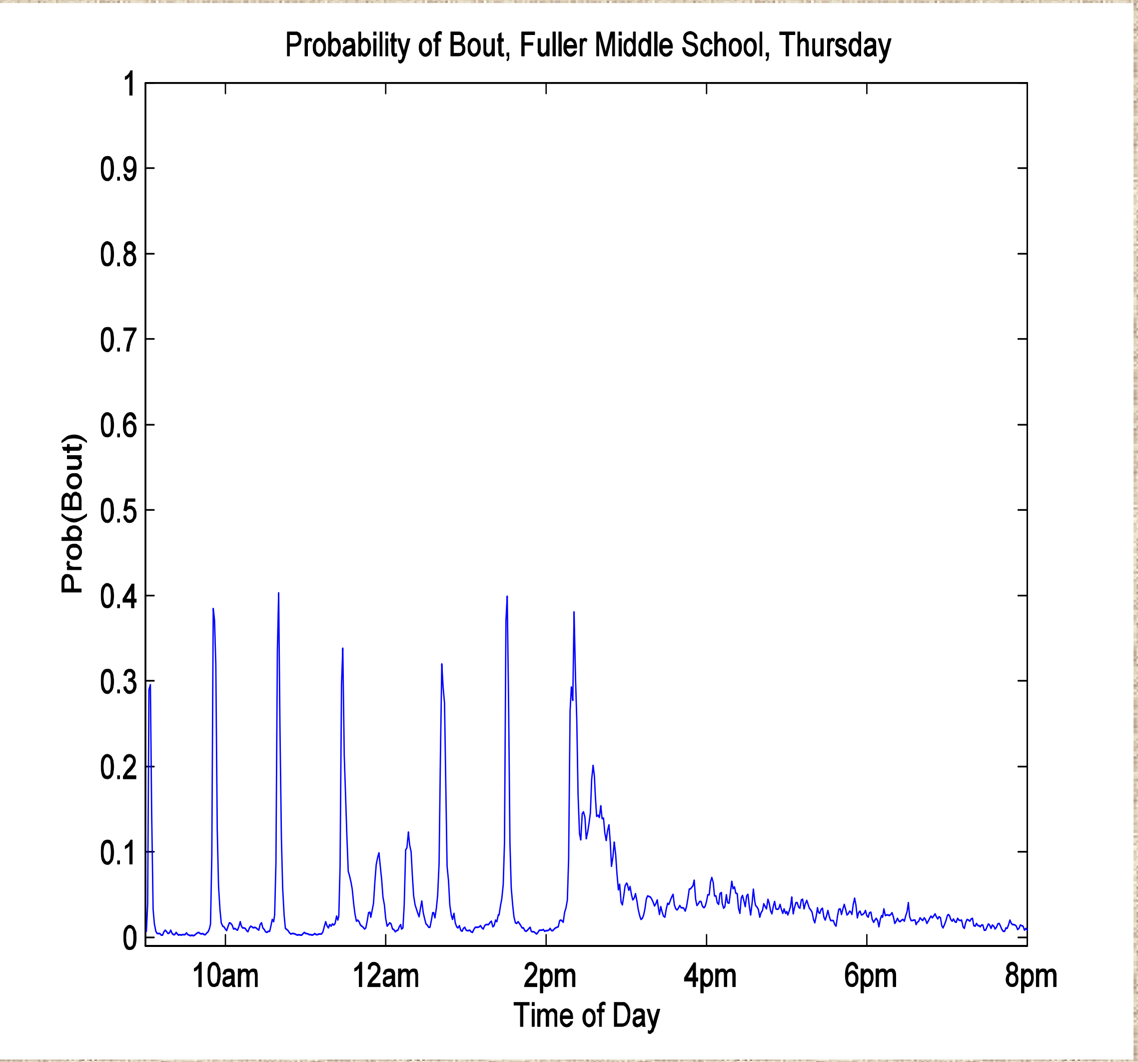


- **TVhrs coded as continuous factor (standardized)**
- **TVhrs effect negative (-1.3% per sd, *p=0.03*)**
  - **More TV, less active**
  - **3:00-5:30, -2.6%, *p=0.02***
  - **7:00-8:00, -3.6%, *p=0.008***
- **Positive effect over lunch**
  - **+2.7%, *p=0.03***
  - **More TV, on average more active over lunch**

Results: Covariance Analysis

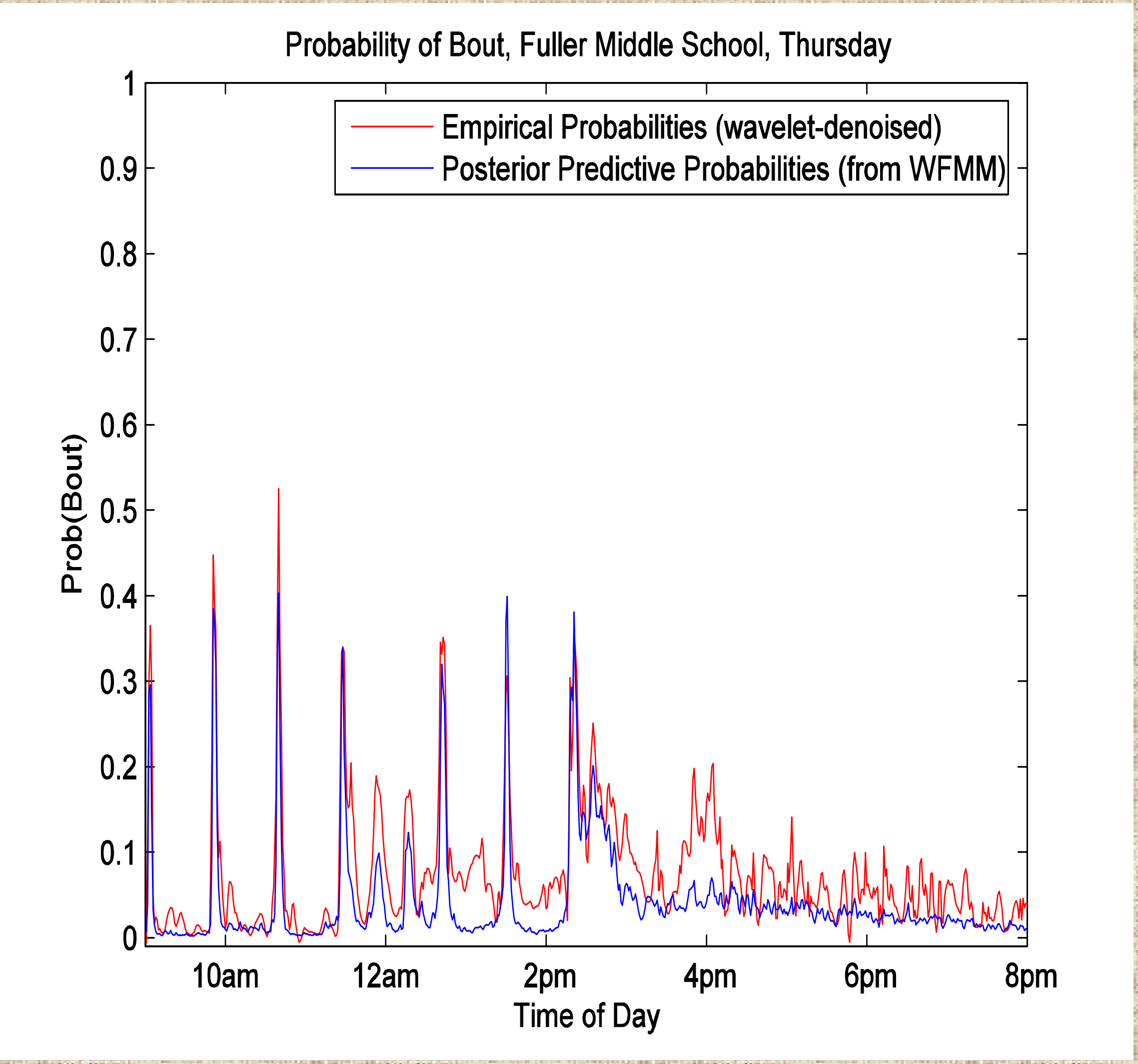
- **Variability: 91% day-to-day, 9% child-to-child**
  - **Important to have many days per child**
- **Study variability as function of *t***
  - **Child-to-child variability: school day > after school**
  - **Day-to-day variability: after school > school day**
  - **Relative day-to-day variability after school: 95%-99%**
- **Equivalent designs:**
  - **108 children, 4 days/child**
  - **72 children, 8 days/child**
  - **54 children, 16 days/child**
- **Less children, more days, save $$$?**

Results: Bouts



- **Can compute posterior predictive probabilities of bouts for children**

Results: Bouts



- **Can compute posterior predictive probabilities of bouts for children**
- **Model-based predictive probabilities not far from empirically-estimated probabilities**

Conclusions

- Found some interesting results in Planet Health
  - School schedules, DST, Day of Week, TV hours
  - Day-to-day variability large – needs lots of days/child
  - Raw vector magnitudes instead of METs?
- Functional mixed models promising method for analyzing accelerometer data
  - Makes use of all of the data
  - Can detect effects that vary over time
  - Can accommodate incomplete profiles
  - After the fact, one can perform inference that averages within different time periods of interest
  - Can still look at probabilities bouts, but as function of time

Acknowledgements

- **Statistical Collaborators on work related to WFMM: Raymond J. Carroll, Louise Ryan, Phil Brown, Marina Vannucci, Kevin Coombes, Keith Baggerly, Naisyin Wang, Brent Coull, and Cassandra Arroyo.**
- **Other Scientific Collaborators: Steve Gortmaker, Joanne Lupton, Josh Fidler, Lihong Wang, Victor Prieto, Stan Hamilton, Ryuji Kobayashi, David Hawke, John Kooman, Nancy Turner, and Meeyoung Hong.**
- **Thanks to Dick Herrick for assistance in optimizing the code for the method, and for converting the Matlab code to C++.**