Wavelet-Based Functional Mixed Models

Jeffrey S. Morris

University of Texas,

M.D. Anderson Cancer Center

Houston, TX, USA

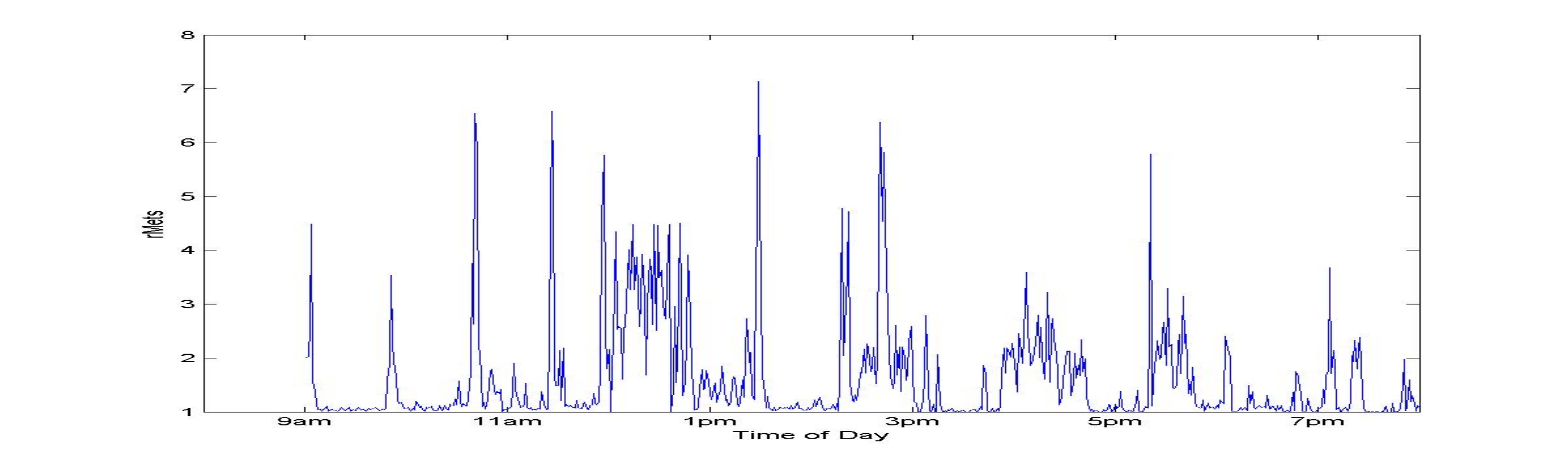
joint work with **Raymond J. Carroll**

Functional Data Analysis

- **Functional Data:**
  - **Ideal units of observation: curves**
  - **Observed data: curves sampled on fine grid**
- **With the development of new automated data collection devices, these type of data are increasingly encountered in scientific research.**

Accelerometer Data

Planet Health Study

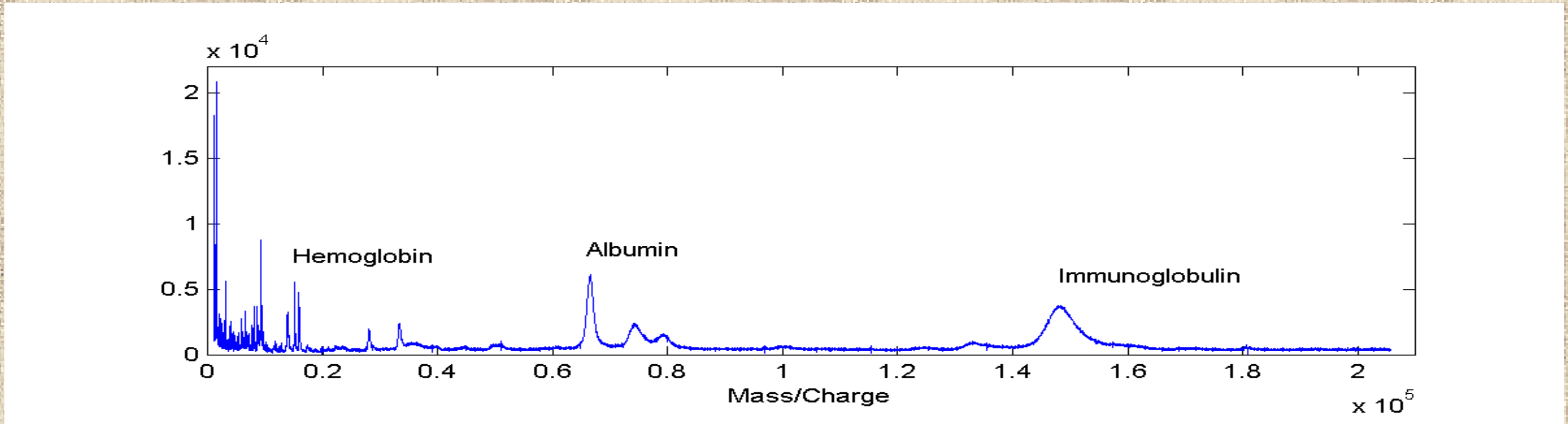


- **Planet Health**: 5 intervention/5 control schools in Boston
- **Accelerometer profile:** *x*=time, *y*=activity level
  - ~400 children, ~4 days/child (~1600 profiles)

Goal: Assess effect of intervention on childrens' activity levels, while controlling for various covariates.

Mass Spectrometry Proteomics

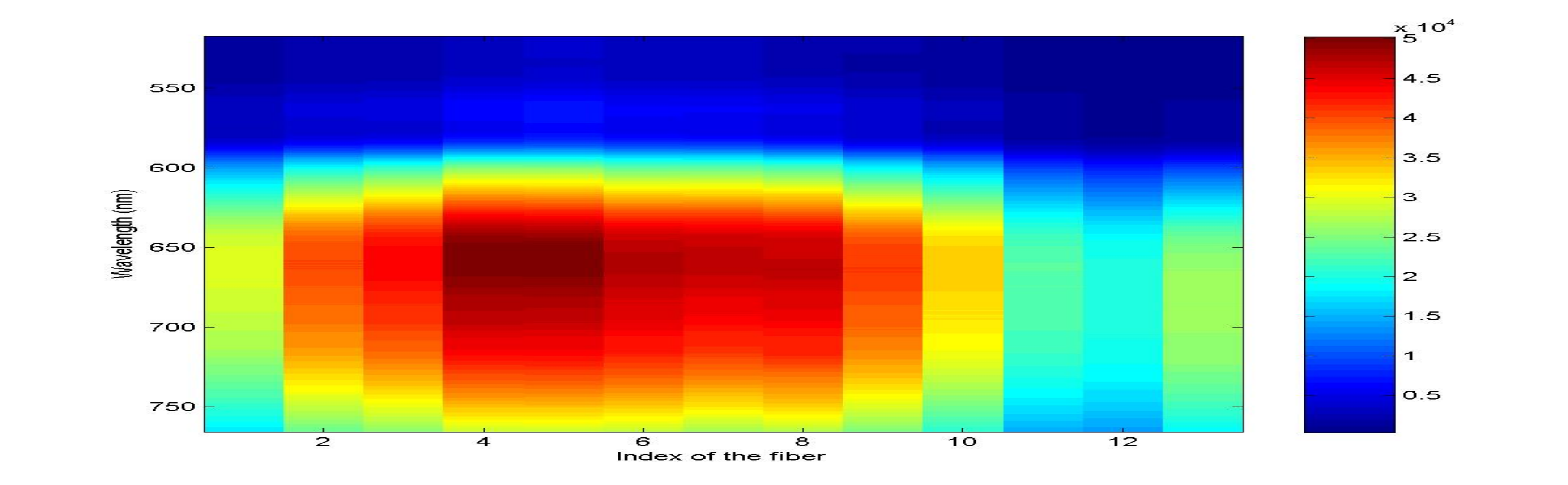
Cancer Cell Line Study



- **Cell line Study:** ~180 rats; 3 cell lines/4 organ sites
- **Proteomic Spectrum:** *x*=molecular mass, *y*=intensity
  - Peaks correspond to proteins present in sample
- **Goal:** Identify proteins in blood serum proteome associated with implanted cell line, host organ site, and their interaction.

SOIR (Spectroscopic Oblique-Indicence Reflectometry)

Melanoma Diagnosis Study



- **Melanoma Study:** Diagnose melanoma using SOIR
- **SOIR:** *x*=optical fiber number, *y*=wavelength , *z*=intensity
  - 42 lesions; histology known (malignant/not)
  - 5 SOIR images from each lesion and adjacent normal tissue
- **Goal:** Develop SOIR-based diagnostic tool for melanoma

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.

Functional Mixed Model (Discrete Version)

Y= ***N*-by-*T* matrix containing the observed curves all sampled on a common equally-spaced grid of length *T*.**

$$\underbrace{Y}_{N\times \mathrm{T}} = \overbrace{X}^{N\times p} \underbrace{B}_{p\times \mathrm{T}} + \overbrace{Z}^{N\times m} \underbrace{U}_{m\times \mathrm{T}} + \underbrace{E}_{N\times \mathrm{T}}$$

$$U \sim MN(P,Q)$$
$$E \sim MN(R,S)$$

- ***U* and *E* follow the Matrix Normal distribution**
  - **–** U**~MN(**P**,**Q**) implies Cov{**U_{ij}**,** $U_{i'j'}$**}=**$P_{ii'}$*****$Q_{jj'}$
- ***P* and *R* are between-curve covariance matrices**
- ***Q* and *S* are within-curve covariance matrices (*T* × *T*)**

Functional Mixed Models

- **Key feature of FMM**: Does not require specification of parametric form for curves
- Most existing literature for nonparametrically modeling functional data is based on kernels or splines.
- Kernels/fixed-knot splines may not work well for spatially heterogeneous functional data
  - e.g. curves have lots of local features, like peaks
- **Wavelet Regression**: nonparametric regression technique that better preserves local features present in the curves.

Introduction to Wavelets

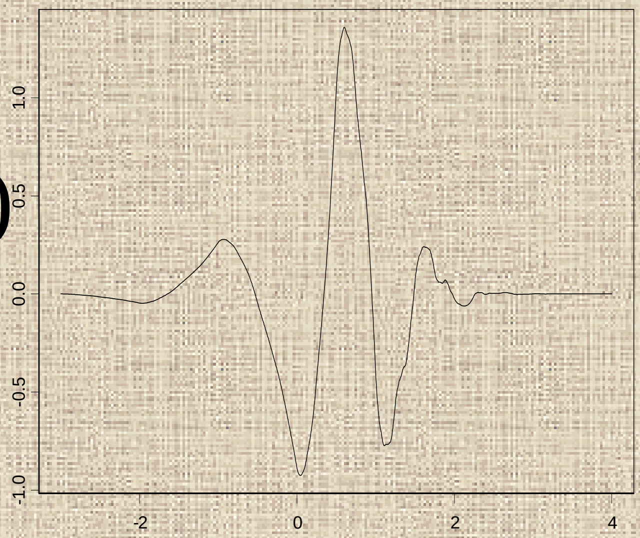
- Wavelets: **families of orthonormal basis functions**

$$g(t) = \sum_{j,k \in \Im} d_{jk} \psi_{jk}(t)$$

$$\psi_{jk}(t) = 2^{-j/2} \psi(2^{-j/2} t - k)$$

$$d_{jk} = \int g(t) \psi_{jk}(t) dt$$

Daubechies-4 Basis Function



- **Discrete Wavelet Transform (DWT): fast algorithm {O(*T*)} for obtaining *T* empirical wavelet coefficients for curves sampled on equally-spaced grid of length *T*.**

Wavelet Regression

- **Wavelet Regression – 3 step process**
  1. **Project data into wavelet space**
  2. **Threshold/shrink coefficients**
  3. **Project back to data space**
- **Yields *adaptively regularized* nonparametric estimates of function**
- **Most work limited to single function case**

Wavelet-Based FMM:

General Approach

1. **Project observed functions** Y **into wavelet space.**
2. **Fit FMM in wavelet space.**

   (Use MCMC to get posterior samples)
3. **Project wavelet-space estimates (posterior samples) back to data space.**

Wavelet-Based FMM:

General Approach

1. **Project observed functions Y into wavelet space.**
2. **Fit FMM in wavelet space**

   (Use MCMC to get posterior samples)
3. **Project wavelet-space estimates (posterior samples) back to data space.**

Wavelet-Based FMM

1. **Project** **observed functions** Y **to wavelet space**

- **Apply DWT to rows of** Y **to get wavelet coefficients corresponding to each observed function**

$$\underbrace{D}_{N\times T} = \underbrace{Y}_{N\times T} \underbrace{W'}_{T\times T}$$

- **Projects the observed curves into the space spanned by the wavelet bases.**

Wavelet-Based FMM:

General Approach

1. **Project** **observed functions Y** **into wavelet space.**
2. **Fit FMM** **in wavelet space**

   (Use MCMC to get posterior samples)

3. **Project** **wavelet-space estimates (posterior samples)** **back to data space.**

Projecting FMM to Wavelet Space

$$\underbrace{Y}_{N\times \mathrm{T}} = \overbrace{X}^{N\times p} \underbrace{B}_{p\times \mathrm{T}} + \overbrace{Z}^{N\times m} \underbrace{U}_{m\times \mathrm{T}} + \underbrace{E}_{N\times \mathrm{T}}$$

$$U \sim MN(P,Q)$$
$$E \sim MN(R,S)$$

Projecting FMM to Wavelet Space

$$\underbrace{Y}_{N\times \mathrm{T}}\overbrace{\mathbf{W'}}^{T\times T} = \overbrace{X}^{N\times p}\underbrace{B}_{p\times \mathrm{T}} + \overbrace{Z}^{N\times m}\underbrace{U}_{m\times \mathrm{T}} + \underbrace{E}_{N\times \mathrm{T}}$$

$$\begin{gathered} U \sim MN(P,Q) \\ E \sim MN(R,S) \end{gathered}$$

Projecting FMM to Wavelet Space

$$\underbrace{Y}_{N\times \mathrm{T}} \overbrace{\mathbf{W}'}^{T\times T} = \overbrace{X}^{N\times p} \underbrace{B}_{p\times \mathrm{T}} \overbrace{\mathbf{W}'}^{T\times T} + \overbrace{Z}^{N\times m} \underbrace{U}_{m\times \mathrm{T}} \overbrace{\mathbf{W}'}^{T\times T} + \underbrace{E}_{N\times \mathrm{T}} \overbrace{\mathbf{W}'}^{T\times T}$$

$$U \sim MN(P,Q)$$
$$E \sim MN(R,S)$$

Projecting FMM to Wavelet Space

$$\underbrace{\mathbf{D}}_{N \times \mathrm{T}} = \overbrace{X}^{N \times p} \underbrace{\mathbf{B}^*}_{p \times \mathrm{T}} + \overbrace{Z}^{N \times m} \underbrace{\mathbf{U}^*}_{m \times \mathrm{T}} + \underbrace{\mathbf{E}^*}_{N \times \mathrm{T}}$$

$$U \sim MN(P, Q)$$
$$E \sim MN(R, S)$$

Projecting FMM to Wavelet Space

$$\underbrace{\mathbf{D}}_{N\times \mathrm{T}} = \overbrace{X}^{N\times p}\underbrace{\mathbf{B}^{*}}_{p\times \mathrm{T}} + \overbrace{Z}^{N\times m}\underbrace{\mathbf{U}^{*}}_{m\times \mathrm{T}} + \underbrace{\mathbf{E}^{*}}_{N\times \mathrm{T}}$$

$$U\mathbf{W}' \sim MN(P, \mathbf{W}Q\mathbf{W}')$$
$$E\mathbf{W}' \sim MN(R, \mathbf{W}S\mathbf{W}')$$

Projecting FMM to Wavelet Space

$$\underbrace{\mathbf{D}}_{N\times \mathrm{T}} = \overbrace{X}^{N\times p}\underbrace{\mathbf{B}^{*}}_{p\times \mathrm{T}} + \overbrace{Z}^{N\times m}\underbrace{\mathbf{U}^{*}}_{m\times \mathrm{T}} + \underbrace{\mathbf{E}^{*}}_{N\times \mathrm{T}}$$

$$\mathbf{U}^{*} \sim MN(P,\mathbf{Q}^{*})$$
$$\mathbf{E}^{*} \sim MN(R,\mathbf{S}^{*})$$

Covariance Assumptions

- **We choose parametric structures for *P* and *R* to model the covariance structure between the curves.**
  - **Based on the experimental design**
  - **As in linear mixed models.**
- **We assume the between-wavelet covariance matrices *Q** and *S** are diagonal.**
  - **Assume wavelet coefficients within given random effect function or residual error process are independent**
  - **Heuristically justified by whitening property of DWT**
  - **Common assumption in wavelet regression**
  - **Is parsimonious in wavelet space (*T* parameters), yet leads to flexible class of covariance structures in data space**

Adaptive Regularization via Shrinkage Prior

Mixture prior on B_{ijk}^*:

$$B_{ijk}^* = \gamma_{ijk}^* N(0, \tau_{ij}) + (1 - \gamma_{ijk}^*)\delta_0$$

$$\gamma_{ijk}^* = \text{Bernoulli}(\pi_{ij})$$

- **Nonlinearly shrinks B_{ijk}^* towards 0, leading to adaptively regularized estimates of B_i.**
- **τ_{ij} & π_{ij} are regularization parameters**
  - **Can be estimated from the data using empirical Bayes**
  - **Extend Clyde&George (1999) to functional mixed model**

Model Fitting

- **MCMC to obtain posterior samples of model quantities**
  - **Work with marginal likelihood; U* integrated out;**
- Let Ω be a vector containing ALL covariance parameters (i.e. for P, Q^*, R, and S^*).

MCMC Steps

1. **Sample from $f(B^*|D,\Omega)$:**

   Mixture of normals and point masses at 0 for each *i,j,k*.

2. **Sample from $f(\Omega|D,B^*)$:**

   Metropolis-Hastings steps for each *j,k*

3. **If desired, sample from $f(U^*|D,B^*,\Omega)$:**

   Multivariate normals

Wavelet-Based FMM:

General Approach

1. **Project** observed functions Y **into wavelet space.**
2. **Fit FMM** in wavelet space
   (Use MCMC to get posterior samples)
3. **Project** wavelet-space estimates (posterior samples) **back to data space**.

Wavelet-Based FMM

3. **Project wavelet-space estimates (posterior samples) back to data space.**

- **Apply IDWT to posterior samples of *B** to get posterior samples of fixed effect functions B_i for $i=1,\ldots,p$, on grid t.**
  - **B=B*W**
- **Posterior samples of *U*, *P*, *Q*, *R*, and *S* are also available, if desired.**

Wavelet-Based FMM: Inference

- **Posterior samples of model quantities enable various types of Bayesian inference & prediction:**
  - **Posterior credible intervals (pointwise or joint)**
  - **Predictive distributions for future sampled curves**
    - **Used in missing data methodology for incomplete profiles**
  - **Predictive probabilities for group membership of new curves**
  - **Bayes Factors for functional inference/model selection**
- **Structure of model makes computation feasible in spite of enormous size of data set.**

Model: Planet Health Study

Y = 503-by-660 matrix, N=503 weekday profiles (200 children)
Profile=minute-by-minute activity levels from 9am-8pm.

$$\boxed{Y = XB + ZU + E}$$

- $X=\{1, X_{interv}, X_{cov}, X_{inter*cov}\}$, where
  X_{interv} = 1 if profile from child in intervention school, -1 control.
  X_{cov} = covariates month, day-of-week, gender, obesity, race
  $X_{inter*cov}$ = various intervention-by-covariate interactions
- $B=\{B_0^T, B_{int}^T, B_{cov}^T, B_{int*cov}^T\}^T$, where
  B_0 = overall mean profile ($1 \times T$), B_{int} = intervention effect ($1 \times T$)
  B_{cov} = covariate effects ($p \times T$), $B_{int*cov}$ = interaction effects ($p \times T$)
- $Z=\{Z_{school}, Z_{child}\}$ $U=\{U_{school}, U_{child}\}$

Discussion

- **Introduced unified modeling approach for FDA**
- **Method based on mixed models; is** FLEXIBLE
  - **Accommodates a wide range of experimental designs**
  - **Addresses large number of research questions**
- **Posterior samples allow Bayesian inference and prediction**
- **Since a unified modeling approach is used, all sources of variability in the model propagated throughout inference.**
- **Since functions adaptively regularized using wavelet shrinkage, the methodology can be used for spatially heterogeneous functional data.**

Discussion

- Approach is Bayesian. The only informative priors to elicit are regularization parameters, which can be estimated from data using empirical Bayes.
- Method complex, but code straightforward to implement
- Normality assumption can be relaxed through use of scale mixtures of normals (or Dirichlet processes?)
- Method generalizes to higher dimensional functions, e.g. image data, space/time (fixed domain) data.
- We used wavelet bases, but approach generalizable to use other orthonormal basis functions.
- Major challenges in developing unified statistical modeling approach for replicated functional data, but worth the effort.