Applications of Wavelet-Based Functional Mixed Models to Proteomics and Genomics Data

Jeffrey S. Morris
Department of Biostatistics
The University of Texas MD Anderson Cancer Center
Houston, Texas
jefmorris@mdanderson.org
Genomic/Proteomic Data as Functions

- Genomic and proteomic tools used to find biomarkers: genes/proteins related to factors of interest, to use in diagnosis/prognosis of disease

- **Genes:** arrayCGH/SNP chips
  - $t =$ chromosomal location, $Y(t) = \log_2(\text{copy number change})$

- **mRNA:** tiling microarrays
  - $t =$ chromosomal location, $Y(t) =$ mRNA abundance

- **Proteins:** MALDI-MS/2d Gel Electrophoresis
  - $t =$ molecular mass (per unit charge), $Y(t) =$ intensity
  - $t_1 =$ molecular mass, $t_2 =$ pH, $Y(t_1, t_2) =$ intensity

- **Common Characteristics of Data:**
  - Very high dimensional (1000’s to 10,000’s to 1,000,000’s)
  - Functions very irregular, containing various types of nonstationarities, discontinuities and local features.
Preprocessing: Necessary to align, background correct, and normalize data (technology specific)

After preprocessing, usual approach involves 2 steps
1. Extract meaningful features (peaks/spots/segments)
2. Identify which are biomarkers (control for FDR)

Alternative: Model as functions using FDA approach
- Requires very flexible modeling techniques to capture complex local features in data.
- Methods must be computationally efficient enough to handle extremely high dimensions of these data.
- Must find way to adjust for multiple comparisons in functional inference.

Wavelet-Based Functional Mixed Models (Morris and Carroll, 2006 JRSS-B)
Wavelet-Based Functional Mixed Models

- **Goal:** Develop automated method that can be used to model and perform inference on complex, irregular functional and image data.

- **Complexities:**
  - Very irregular signals - not smooth
  - Functions may be correlated (e.g. replicates)
  - We may need to factor out effect of nuisance factors, i.e. covariates
  - We would like to be able to flag certain regions of function/image as related to factors of interest, while giving assessment of uncertainty and controlling for multiple testing (FDR).

- **Generalize linear mixed model to functional setting**
Linear Mixed Models

Linear Mixed Model (Laird and Ware, 1982):

\[
Y_{N \times 1} = X_{N \times p} \beta_{p \times 1} + Z_{N \times m} u_{m \times 1} + e_{N \times 1}
\]

\[
u \sim N(0, D_{m \times m})\\
e \sim N(0, R_{N \times N})
\]

- **Fixed effects** part, \(X\beta\), accommodate a broad class of mean structures, including main effects, interactions, and linear coefficients.
- **Random effects** part, \(Zu\), provides a convenient mechanism for modeling correlation among the \(N\) observations.
Functional Mixed Model (FMM)

- **Idea:** Relate *functional response* to set of scalar predictors through *functional coefficients*, while adjusting for possible *correlation between functions* induced by design.
- Suppose we observe a sample of \( N \) curves, \( Y_i(t), i=1, \ldots, N \), on a closed interval \( T \).

\[
Y_i(t) = \sum_{j=1}^{p} X_{ij} B_j(t) + \sum_{k=1}^{m} Z_{ik} U_k(t) + E_i(t)
\]

- \( B_j(t) \) summarizes partial effect of \( X_j \) on \( Y(t) \)
- \( Q(t_1, t_2) \) and \( S(t_1, t_2) \) are covariance surfaces on \( T \times T \) describing the form of the function-function deviations

\[
U_k(t) \sim GP(0, Q) \\
E_i(t) \sim GP(0, S)
\]
Suppose each observed curve is sampled on a common equally-spaced grid of length $T$.

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

- Rows of $B$ contain fixed effect functions on grid
- $Q$ and $S$ are within-curve covariance matrices $(T \times T)$ approximating surfaces on the grid
  - For irregular functional data, $Q$ and $S$ typically contain many nonstationarities, yet their dimension is too high to leave unstructured

$U_k \sim MVN(0, Q)$
$E_i \sim MVN(0, S)$
Wavelet Space Representation

\[ y(t) = \sum_{j,k \in \mathbb{Z}} d_{jk} \psi_{jk}(t) \]

\[ d_{jk} = \int y(t) \psi_{jk}(t) dt \]

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

Linear Representation:

\[ \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} d_1 \\ \vdots \\ d_T \end{bmatrix} \begin{bmatrix} \psi_{11}(t) & \psi_{12}(t) & \cdots & \psi_{JK}(t) \end{bmatrix} \]

DWT Design Matrix \( W = [\psi_{11}(t) \psi_{12}(t) \cdots \psi_{JK}(t)] \)

Given \( T \)-vector \( y \) consisting of function sampled on equally-spaced grid, a pyramid-based algorithm for DWT (Mallat) can be used to obtain \( d \), \( T \)-vector of wavelet coefficients, in \( O(T) \) operations (converse also true)
Functional Mixed Models

- **Key feature of FMM:** Does not require specification of parametric form for functions (response, fixed, or random)
- **Basis function approach:** \[ Y_i(t) = \sum d_{ijk} \psi_{jk}(t) \]
- **Benefits of Using Wavelet Bases**
  1. **Compact support** allows efficient representations of local features and discontinuities
  2. **Whitening property** allows parsimonious yet flexible representations of Q and S
  3. Decomposes function in both **frequency** \((j)\) and **time** \((k)\) domains
     - Key for **adaptive regularization** of functional estimates
  4. Orthonormal transformation has **linear representation** and special structure allows **fast calculation** of coefficients.
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

- Wavelet basis representation written in matrix form

$$\begin{align*}
D_{N \times T} &= Y_{N \times T} W'_{T \times T} \\
\end{align*}$$

- Orthonormality:

$$WW' = W'W = I_T$$

- Matrix multiplication unnecessary; fast algorithm \{DWT, $O(T)$\} can be applied to each row of $Y$ to get corresponding wavelet coefficients ($D$)

- **Projects** observed functions into space spanned by wavelet coefficients

- **Full rank projection**: can run inverse algorithm (IDWT) on wavelet coefficients and completely recover original observed data.
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 Representations

\[ Y = X + B + Z + E \]

- \( Y = \text{DW} \)
- \( E = E \times W \)
- \( U = U \times W \)
- \( B = B \times W \)

Wavelet Space FMM
Wavelet Space FMM

\[
\begin{align*}
DW_{N \times T} &= X_{p \times T} B^{*W}_{p \times m} + Z_{m \times T} U^{*W}_{m \times N} + E^{*W}_{N \times T} \\
\end{align*}
\]

Wavelet Representations

Y = DW
B = B*W
U = U*W
E = E*W
Wavelet Space FMM

\[
\begin{align*}
\begin{array}{c}
\frac{D}{N \times T} \frac{W}{N \times p} \frac{W'}{N \times m} \frac{W}{N \times T} \\
\frac{X}{p \times T} \frac{B \ast W}{p \times T} \frac{W'}{p \times T} + \\
\frac{Z}{m \times T} \frac{U \ast W}{m \times T} \frac{W'}{m \times T} + \\
\frac{E \ast W}{N \times T} \frac{W'}{N \times T}
\end{array}
\end{align*}
\]

Wavelet Representations

\[
Y = DW
\]

\[
B = B \ast W
\]

\[
U = U \ast W
\]

\[
E = E \ast W
\]

\[
WW' = I
\]
Wavelet Space FMM

\[
D \begin{array}{c}
\scriptstyle N \times m \\
\scriptstyle N \times T
\end{array}
= \begin{array}{c}
\scriptstyle N \times p \\
X
\end{array}
B^* \begin{array}{c}
\scriptstyle p \times T \\
Z
\end{array}
U^* \begin{array}{c}
\scriptstyle m \times T \\
E^*
\end{array}

\text{Wavelet Representations}

YW^\prime = D

BW^\prime = B^*

UW^\prime = U^*

EW^\prime = E^*

WW^\prime = I
Wavelet Space FMM

- \( D \): empirical wavelet coefficients for observed curves
  - Row \( i \) contains wavelet coefficients for observed curve \( i \)
  - Each column double-indexed by wavelet scale \( j \) and location \( k \)

\[
D_{N \times T} = X_{N \times p} B^*_{p \times T} + Z_{N \times m} U^*_{m \times T} + E^*_{N \times T}
\]

- \( B^* = BW' \& U^* = UW' \): Rows contain wavelet coefficients for the fixed and random effect functions, respectively
- \( E^* = EW' \) is the matrix of wavelet-space residuals
- \( Q^* = WQW' \) and \( S^* = WSW' \) model the covariance structure between wavelet coefficients for a given function.
- \( Q^* \) and \( S^* \) are too large for unstructured representation.
  - Our approach: model as diagonal matrices \( Q^* = \text{diag}(q_{jk}) \) (independent but heteroscedastic in wavelet space)
  - Parsimonious, yet accommodates nonstationary \( Q \) and \( S \)
Independent Mixed Models per Column

\[
d_{jk} = \begin{bmatrix} X \end{bmatrix} B_{jk}^* + \begin{bmatrix} Z \end{bmatrix} \begin{bmatrix} u_{jk}^* \end{bmatrix} + \begin{bmatrix} e_{jk}^* \end{bmatrix}
\]

\[
\begin{align*}
    u_{jk}^* & \sim N(0, q_{jk}^*) \\
e_{jk}^* & \sim N(0, s_{jk}^*)
\end{align*}
\]
Prior Assumptions

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 $B(t)$.
- $\tau_{ij}$ & $\pi_{ij}$ are regularization parameters that mitigate the trade-off between bias/variance in function estimation.
- Estimated from data using empirical Bayes approach.
Model Fitting

- MCMC to obtain posterior samples
- Use marginal likelihood: U* integ. out;

MCMC Steps

1. Sample from \( f(B_{ijk}^*|D,q,s) \)
   *Gibbs* step: Spike/Gaussian slab mixture

2. Sample from \( f(q_{jk},s_{jk}|D,B^*) \)
   *Metropolis-Hastings* step: random walk

3. If desired, sample from \( f(U_k^*|D,B^*,\Omega) \)
   *Gibbs* step: 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.**
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 $B_j(t)$ for $i=1,...,p$, on grid $t$.

$$B = B^* W$$

- These posterior samples can be used to perform Bayesian inference, e.g. to figure out for what $t$ the fixed effect functions $B_j(t)$ are significant.
FDR-Based Bayesian Functional Inference

• Given specified effect size $\delta$, compute

$$p_j(t) = 1 - \text{Prob}\{ |B_j(t)| > \delta \mid Y \} \text{ for each } t$$

• $p_j(t) =$ \textit{local FDR estimate} for declaring location $t$ “significant” (region of function with difference $\geq \delta$)

• Global Criterion: Specify $\alpha$, can find cutpoint on $p_j(t)$ for which average FDR controlled to be $\leq \alpha$.

  $\|\text{false positive regions}\| / \|\text{flagged regions}\| \leq \alpha$

• Extends FDR ideas to functional setting, and provides principled solution to multiple testing problem inherent in pointwise inference.
Example: Organ-Cell Line Expt

- 16 mice had 1 of 2 cancer cell lines (A375P or PC3MM2) injected into 1 of 2 organs (lung or brain)
- Blood Serum extracted from each mouse, run on MALDI at 2 laser intensities (low/high)
- Total: 32 spectra (2/mouse), each on grid of 7985

**Goal**: Find proteins differentially expressed by:
- Host organ site (lung/brain)
- Donor cell line (A375P/PC3MM2)
- Organ-by-cell line interaction
Model: Organ-by-Cell Line Experiment

Let $Y_i(t)$ be the (log$_2$) MALDI-TOF spectrum $i$.

$$Y_i(t) = B_0(t) + \sum_{j=1}^{4} X_{ij} B_j(t) + \sum_{k=1}^{16} Z_{ik} U_k(t) + E_i(t)$$

- $X_{i1} = 1$ for lung, -1 brain. $X_{i2} = 1$ for A375P, -1 for PC3MM2
- $X_{i3} = X_1 \cdot X_2$  $X_{i4} = 1$ for low laser intensity, -1 high.
- $B_0(t) =$ overall mean spectrum $B_1(t) =$ organ main effect function
- $B_2(t) =$ cell-line main effect $B_3(t) =$ org x cell-line interaction function
- $B_4(t) =$ laser intensity effect function
- $U_k(t)$ is random effect function for mouse $k$.
- $Z_{ik} = 1$ if spectrum $i$ is from mouse $k$ ($k=1, ..., 16$)
Demonstration of Flexibility of WFMM

- We obtain adaptively regularized estimates of both fixed effect functions and random effect functions
  - Not just estimates, but posterior samples
- We are able to model nonstationarities in between-curve covariances, including heteroscedasticity and spatially-varying autocorrelation (smoothness)
- Model captures complex features: Model-generated posterior predictive spectra look like real spectra.
- Can model out block effects: Inclusion of nonparametric laser intensity effect models systematic differences in location and intensity of peaks, effectively calibrating for common analysis
- Can be applied to very large data sets (1000’s of functions on grid of size in the 10,000’s)
Results: MALDI Example

- Using $\alpha=0.05$, $\delta=1$ (2-fold expression on log$_2$ scale), we flag a number of spectral regions.
Results: MALDI Example

- 3900 D (~100-fold) (CGRP-II): dilates blood vessels in brain
- 7620 D (~5-fold) (neurogranin): active in synaptic modeling in brain (Not detected as peak)
Extension to Higher Dimensions (Images)

- Method can be extended to higher dimensional functions
  - Fixed effect and random effect surfaces
- How? Use 2d (or higher) wavelet transforms
  - Accounts for spatial correlations in both horizontal and vertical directions
- Key: image can be represented as vector, and higher dimensional wavelet transforms can be written as orthonormal linear transformation of this vector.
- Computational considerations:
  - Memory issues: keep subset of wavelet coeffs.
Bayesian Inference: Discrimination/Classification

- Can classify new function $Y_i(t)$ (e.g. cancer/normal) using posterior predictive probabilities
  - $X =$ cancer status of test sample (1 = cancer, -1 = not)
  - $Y =$ test spectrum, $Y^t =$ training spectra
  - Classify as cancer if $Pr(X = 1 | y, Y^t) > 0.50$

- Straightforward to compute given posterior samples of model parameters
- Does not require high dimensional feature selection step
- Can account/adjust for other covariates in the model, clinical and technical
- Straightforward to hierarchically combine together several types of data, functional or clinical, to predict class
Bayesian Inference: Discrimination/Classification

\[ \Pr(X = 1 \mid y, Y^t) = O \div (O + 1) \]

\[ O = \frac{\Pr(X = 1)}{1 - \Pr(X = 1)} \times \frac{\text{Bayes Factor}}{BF} \]

\[ BF = \frac{f(y \mid X = 1, Y^t)}{f(y \mid X = -1, Y^t)} \]

\[ f(y \mid X = 1, Y^t) = \int f(y \mid X = 1, \Theta) f(\Theta \mid Y^t) d\Theta \]

\[ \approx B^{-1} \sum_{b=1}^{B} f(y \mid X = 1, \Theta^{(b)}) \]
Bayesian Inference: Discrimination/Classification

\[
f(y \mid X = 1, \Theta^{(b)}) = f(d \mid X = 1, \Theta^{*(b)}) = \prod_{j,k} f(d_{jk} \mid X = 1, \Theta^{*(b)}_{jk})
\]

\[
BF = \prod_{j,k} BF_{jk}
\]
Discussion

• Presented unified modeling approach for FDA
  – Adaptive enough to handle irregularities in both mean structures and random effects (covariances)

• 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
  – Flag significant regions, while controlling FDR
  – Classify subjects based on genomic/proteomic profile

• Since a unified modeling approach is used, all sources of variability in the model propagated throughout inference.
Discussion

• Approach is Bayesian. The only informative priors to elicit are \textit{regularization parameters}, which can be estimated from data using empirical Bayes.

• Developed \textit{general-use code} (freely available on website) – reasonably fast and straightforward to use \(\rightarrow\) minimum information to specify is Y, X, Z matrices.

• Method can be generalized to model \textit{higher dimensional functions} (e.g. image mixed models, under development)

• The Gaussian/independence assumptions can be relaxed to yield \textit{robust} and even more flexible modeling
Acknowledgements

• Some of the work presented here is from 2 papers


• Supported by NIH Grant R01 CA107304

• Computer code/ papers on web at http://biostatistics.mdanderson.org/Morris/papers.html