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, $Y(t)$ = $\log_2$(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, $Y(t_t, t_2)$ = 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.**

Statistical Modeling

- 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):

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

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

- **Fixed effects** part, *Xβ,* accommodate a broad class of mean structures, including main effects, interactions, and linear coefficients.
- **Random effects** part, Z*u,* 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 $\mathcal{T}$

$$U_k(t) \sim GP(0,Q)$$
$$E_i(t) \sim GP(0,S)$$

$$\underbrace{Y_i(t)}_{\text{response functions}} = \sum_{j=1}^{p} X_{ij} \underbrace{B_j(t)}_{\text{fixed effect function}} + \sum_{k=1}^{m} Z_{ik} \underbrace{U_k(t)}_{\text{random effect functions}} + \underbrace{E_i(t)}_{\text{residual error functions}}$$

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

Discrete Version of FMM

Suppose each observed curve is sampled on a common equally-spaced grid of length *T*.

$$U_k \sim MVN(0,Q)$$

$$E_i \sim MVN(0,S)$$

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

- Rows of ***B*** contain fixed effect functions on grid
- ***Q*** and ***S*** are within-curve covariance matrices (*T* × *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

Wavelet Space Representation

$$y(t) = \sum_{j,k \in \mathfrak{I}} d_{jk} \psi_{jk}(t)$$

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

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

Linear Representation:

$$\overbrace{\mathbf{y}}^{1\times T} = \underbrace{\mathbf{d}}_{1\times T} \overbrace{\mathbf{W}}^{T\times T}$$

$$\overbrace{\mathbf{d}}^{1\times T} = \underbrace{\mathbf{y}}_{1\times T} \overbrace{\mathbf{W}}^{T\times T}{}'$$

DWT Design Matrix **W**=[$\psi_{11}(\mathbf{t})$ $\psi_{12}(\mathbf{t})$ … $\psi_{JK}(\mathbf{t})$]

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:** $Y_i(t)=\Sigma d_{ijk}\psi_{jk}(t)$
- **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

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

$$\boxed{\begin{array}{l}\text{Orthonormality:}\\ WW'=W'W=I_T\end{array}}$$

- 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.

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 Space FMM

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

Wavelet Representations

Y=DW

B=B*W

U=U*W

E=E*W

Wavelet Space FMM

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

Wavelet Representations

Y=DW

B=B*W

U=U*W

E=E*W

Wavelet Space FMM

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

Wavelet Representations

Y=DW

B=B*W

U=U*W

E=E*W

WW’=I

Wavelet Space FMM

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

Wavelet Representations

WW’=I

YW’=D

BW’=B*

UW’=U*

EW’=E*

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*

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

$$U_k^* \sim MVN(0, Q^*)$$
$$E_i^* \sim MVN(0, S^*)$$

- *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**=diag(q_{jk}) (independent but heteroscedastic in wavelet space)
  - Parsimonious, yet accommodates nonstationary ***Q*** and ***S***

Independent Mixed Models per Column

$$\underbrace{d_{jk}}_{N\times 1} = \overbrace{X}^{N\times p}\underbrace{B^{*}_{jk}}_{p\times 1} + \overbrace{Z}^{N\times m}\underbrace{u^{*}_{jk}}_{m\times 1} + \underbrace{e^{*}_{jk}}_{N\times 1}$$

$$u^{*}_{jk} \sim N(0, q^{*}_{jk})$$

$$e^{*}_{jk} \sim N(0, s^{*}_{jk})$$

Prior Assumptions

Mixture prior on B_{ijk}^{*}:

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

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

- Nonlinearly shrinks B_{ijk}^{*} towards 0, leading to ***adaptively regularized*** estimates of $B_i(t)$.
- τ_{ij} & π_{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.**

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 $\boldsymbol{B_j(t)}$ for $i=1,\ldots, p,$ on grid **t**.

$$\mathbf{B}=\mathbf{B}^{*}\mathrm{W}$$

- These posterior samples can be used to perform Bayesian inference, e.g. to figure out for what t the fixed effect functions $\boldsymbol{B_j(t)}$ are significant

FDR-Based Bayesian Functional Inference

- Given specified effect size δ, compute

  $p_j(t) = 1$ - Prob{ $|B_j(t)| > \delta$ | Y } for each t
- $p_j(t)$ = ***local FDR estimate*** for declaring location t "significant" (region of function with difference $\geq\delta$)
- Global Criterion: Specify α, can find cutpoint on $p_j(t)$ for which average FDR controlled to be $\leq\alpha$.

  || *false positive regions* || / || *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

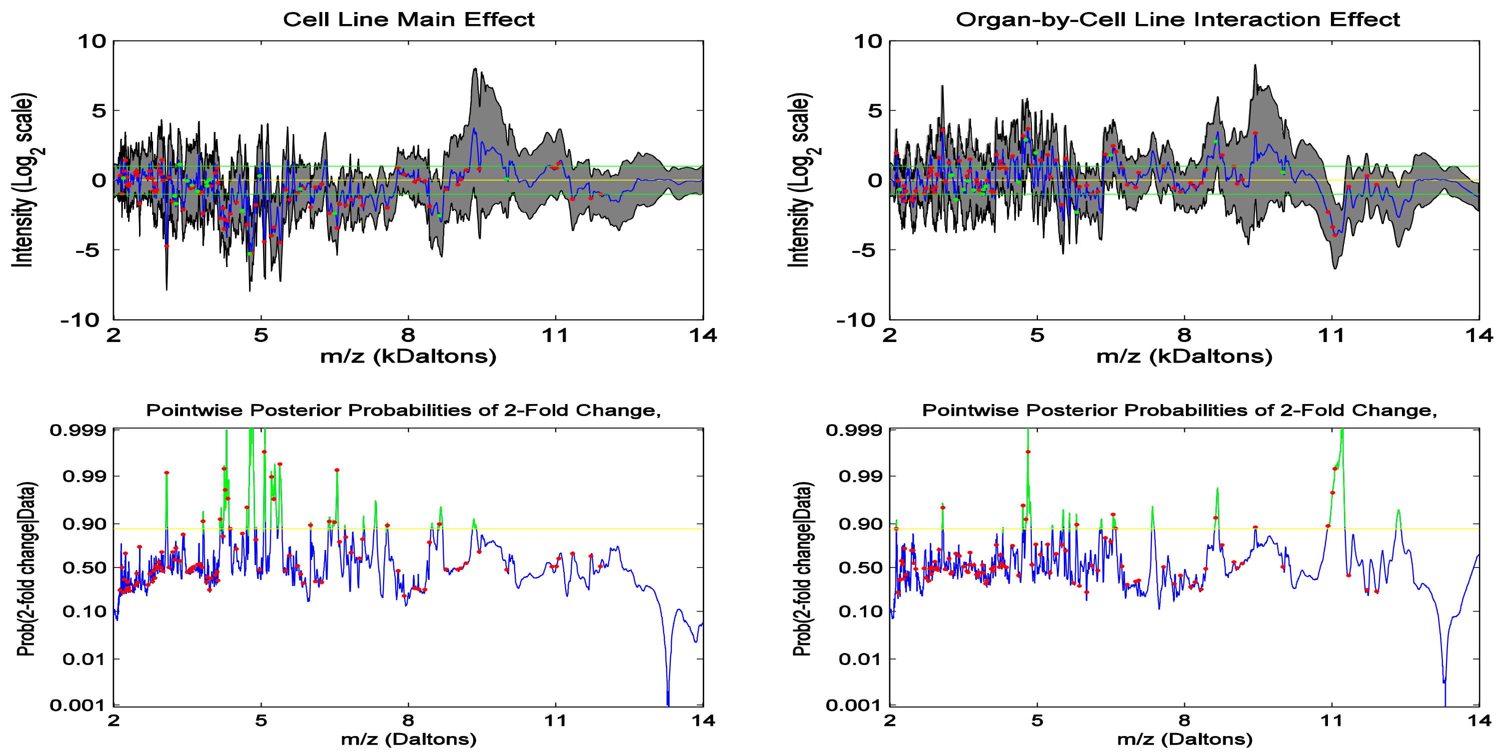
$$Y_i(t) = B_0(t) + \sum_{j=1}^{4} X_{ij} B_j(t) + \sum_{k=1}^{16} Z_{ik} U_k(t) + E_i(t)$$

- $X_{i1} = 1$ for lung, -1 brain. $X_{i2} = 1$ for A375P, -1 for PC3MM2
  $X_{i3} = X_1 * X_2$ $X_{i4} = 1$ 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.
- $Zik=1$ if spectrum i is from mouse k ($k=1, \ldots, 16$)

Demonstration of Flexibility of WFMM

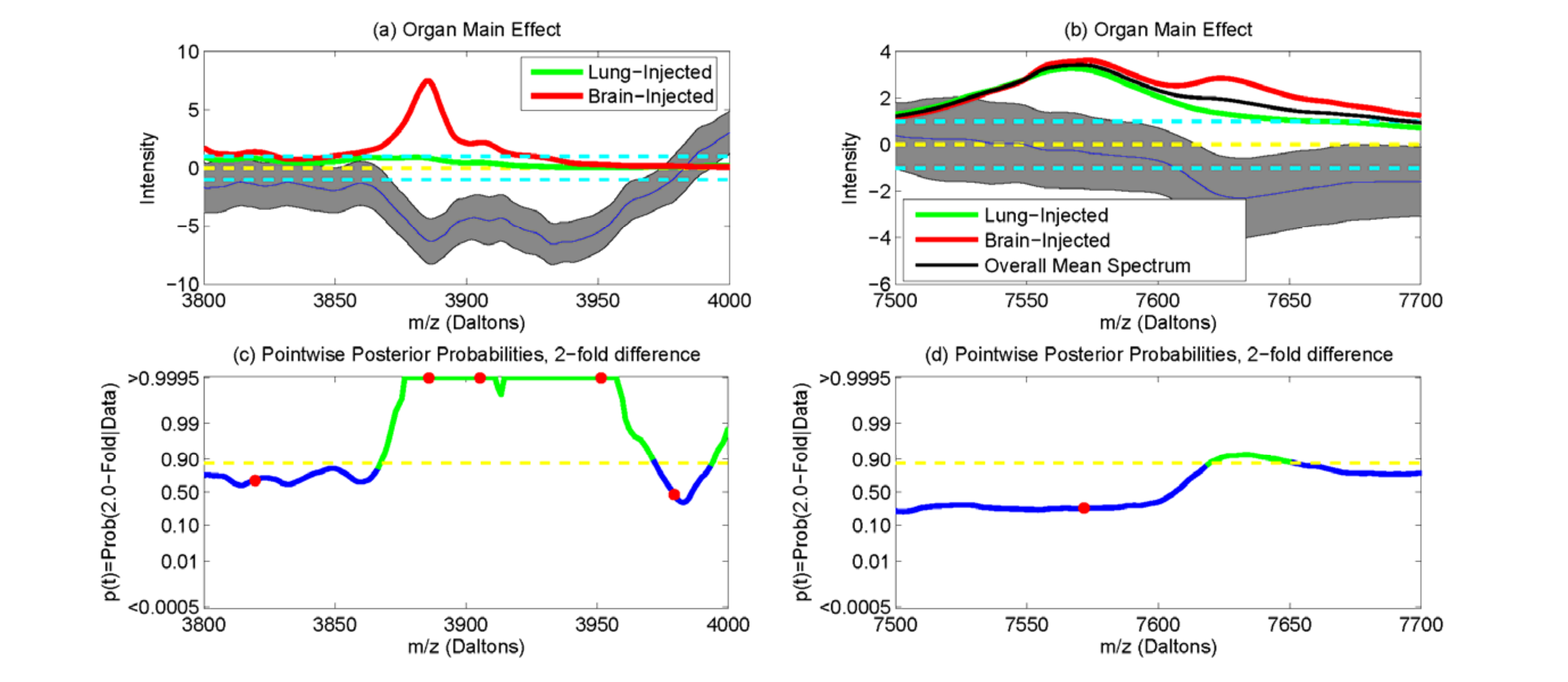
- We obtain adaptively regularized estimates of both fixed effect functions and random effect functions
  - Not just estimates, but posterior samples
- We are able to model nonstationarities in between-curve covariances, including heteroscedasticity and spatially-varying autocorrelation (smoothness)
- 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 α=0.05, δ=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

Details

Bayesian Inference: Discrimination/Classification

$$\Pr(X=1 \mid y, Y^t) = O/(O+1)$$

$$O = \overbrace{\frac{\Pr(X=1)}{1-\Pr(X=1)}}^{\text{prior odds}} \times \overbrace{BF}^{\text{Bayes Factor}}$$

$$BF = \frac{f(y|X=1,Y^t)}{f(y|X=-1,Y^t)}$$

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

$$\approx B^{-1} \sum_{b=1}^{B} f(y \mid X=1, \Theta^{(b)})$$

More Details

Bayesian Inference: Discrimination/Classification

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

$$BF = \prod_{j,k} BF_{jk}$$

Return

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 *regularization parameters*, which can be estimated from data using empirical Bayes.
- Developed general-use code (freely available on website) – reasonably fast and straightforward to use → minimum information to specify is Y, X, Z matrices.
- Method can be generalized to model higher dimensional functions (e.g. image mixed models, under development)
- The Gaussian/independence assumptions can be relaxed to yield robust and even more flexible modeling

Acknowledgements

- **Some of the work presented here is from 2 papers**

1. "*Wavelet-Based Functional Mixed Models*" (2006) Jeffrey S. Morris and Raymond J. Carroll, *JRSS-B*, 68(2): 179-199.
2. "*Bayesian Analysis of Mass Spectrometry Proteomics Data using Wavelet Based Functional Mixed Models*" (2007) Jeffrey S. Morris, Philip J. Brown, Richard Herrick, Keith A. Baggerly, and Kevin R. Coombes, *Biometrics*, doi:10.1111/j.1541-0420.2007.00895.x (online)

- **Supported by NIH Grant R01 CA107304**
- **Computer code/papers on web at** `http://biostatistics.mdanderson.org/Morris/papers.html`