Bayesian Analysis of Mass Spectrometry Data Using Wavelet-Based Functional Mixed Models

Jeffrey S. Morris

UT MD Anderson Cancer Center

joint work with Philip J. Brown, Kevin R. Coombes, Keith A. Baggerly

Functional Data Analysis of Mass Spectrometry Data

- **Model as "functional data"**
  - **Idea: Model entire spectrum as single entity, not a collection of data points.**
- **Wavelet-based Functional Mixed Models**
  - **Peak detection**
  - **Identify differentially expressed peaks while controlling Bayesian FDR**
  - **Automatically account for block effects**
  - **Classify samples based on spectra, without having to search high dimensional model spaces**

Outline

- **Introduction**
  - **Examples**
  - **Mixed Models/Functional Mixed Models**
  - **Wavelets**
- **Wavelet-Based Functional Mixed Models**
- **Bayesian Inference for Mass Spectrometry**
- **Apply to Examples**
- **Discussion**

Example: Pancreatic Cancer Study

- **Koomen, et al. (2004)**
- **256 blood serum samples – 141 pancreatic cancer, 115 normal controls**
- **4 MALDI spectra/sample**
  - **Fractions: MYO25, MYO70, BSA25, BSA70**
- **Samples (all fractions) run in 4 blocks on 4 different dates**
- **Goals:**
  - **Identify differentially expressed protein peaks.**
  - **Classify samples as C/N based on spectra.**
- **Must adjust for block effects on spectra**
- **This talk: Focus on MYO25 fraction, 4kD-10kD**

Example:Organ-Cell Line Expt

- 16 nude mice had 1 of 2 cancer cell lines injected into 1 of 2 organs (lung or brain)
- Cell lines:
  - A375P: human melanoma, low metastatic potential
  - PC3MM2: human prostate, highly metastatic
- Blood Serum extracted from each mouse – placed on 2 SELDI chips
- Samples run at 2 different laser intensities (low/ high)
- Total of 32 spectra (observed functions), 2 per mouse

Example: Organ-Cell Line Expt

- **Goal:**

  Find proteins differentially expressed by:
  - **Host organ site (lung/brain)**
  - **Donor cell line (A375P/PC3MM2)**
  - **Organ-by-cell line interaction**
- **Combine information across laser intensities:**

  Requires us to include in modeling:
  - **Functional laser intensity effect**
  - **Random effect functions to account for correlation between spectra from same mouse**

Linear Mixed Models

Linear Mixed Model (Laird and Ware, 1982):

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

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

$$e \sim N(0, \underbrace{R}_{N\times N})$$

- **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 (FMM)

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

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

- $B_j(t)$ = fixed effect functions
- $U_k(t)$ = random effect functions
- $E_i(t)$ = residual error processes

Pancreatic Cancer Example

Let $Y_i(t)$ be MALDI spectrum from sample *i*

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

- X_{i1}=*1* if **cancer**, *-1* if **normal**
  X_{ij} =*1* if **block** *j*, *-1* if block *1* for *j=2,3,4*
- $B_0(t)$ = **overall mean** spectrum
  $B_1(t)$ = **cancer effect** function
  $B_j(t)$ = **block effect** function for *j=2,3,4*
- **No random effects necessary**

Organ-by-Cell Line Example

Let $Y_i(t)$ be the SELDI 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 int function

  $B_4(t)$ = laser intensity effect function

  - **Z_{ik}=*1* if spectrum *i* is from mouse *k* (*k=1, …, 16)***
  - **$U_k(t)$ is random effect function for mouse *k*.**

Functional Mixed Models

- **Key feature of FMM: Does not require specification of parametric form for curves**
- **Methods based on kernels/fixed knot splines not well suited to spiky functional data**
- **Wavelet Regression: nonparametric regression technique that better preserves local features present in the curves.**

Functional Mixed Model
(Discrete version)

Y= *N*-by-*T* matrix containing the observed spectra on sampling grid of size *T*

$$\underset{N\times \mathrm{T}}{Y} = \overset{N\times p}{\overbrace{X}}\ \underset{p\times \mathrm{T}}{B} + \overset{N\times m}{\overbrace{Z}}\ \underset{m\times \mathrm{T}}{U} + \underset{N\times \mathrm{T}}{E}$$

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

- B_{ij} is the effect of covariate i at location t_j
- Q and S are covariance matrices (T x T)
- Note: Some structure must be assumed on form of Q and S (discussed later)

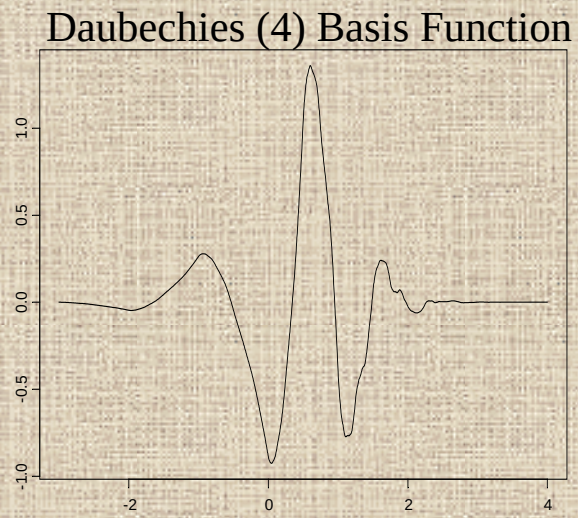
Introduction to Wavelets

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

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

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

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

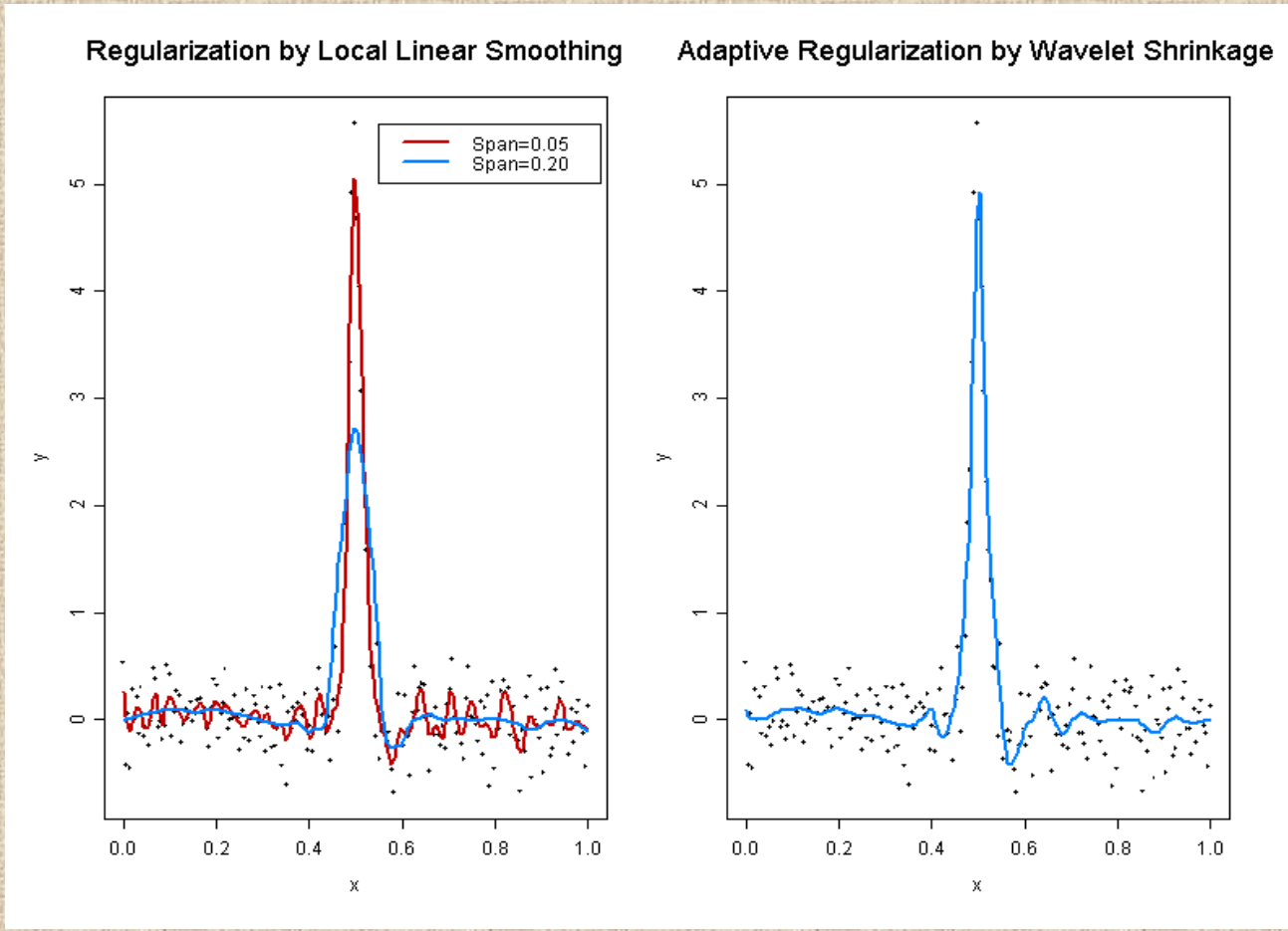


- **Discrete Wavelet Transform (DWT): fast algorithm {O(*T*)} for obtaining *T* empirical wavelet coefficients for curves sampled on equally-spaced grid of length *T*.**
- **Linear Representation:** $d = y\,W'$
  - ***W'* =*T*-by-*T* orthogonal projection matrix**
- **Inverse DWT (IDWT):** $y = d\,W$

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* (plot) nonparametric estimates of function**
- **Morris, et al. (2003) extended to nested functional model (Bayesian)**
- **Morris and Carroll (2004) extended to general functional mixed model framework (Wavelet-based FMM)**

Adaptive Regularization



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

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

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

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

Projecting FMM to Wavelet Space

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

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

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

Projecting FMM to Wavelet Space

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

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

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

Projecting FMM to Wavelet Space

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

$$U_i\mathbf{W'} \sim MVN(0, \mathbf{WQW'})$$

$$E_i\mathbf{W'} \sim MVN(0, \mathbf{WSW'})$$

Projecting FMM to Wavelet Space

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

$$\mathbf{U}_i^{*} \sim MVN(0, \mathbf{Q}^{*})$$

$$\mathbf{E}_i^{*} \sim MVN(0, \mathbf{S}^{*})$$

Model Each Column Separately

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

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

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

Prior Assumptions

Mixture prior on B_{ijk}^*:

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

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

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

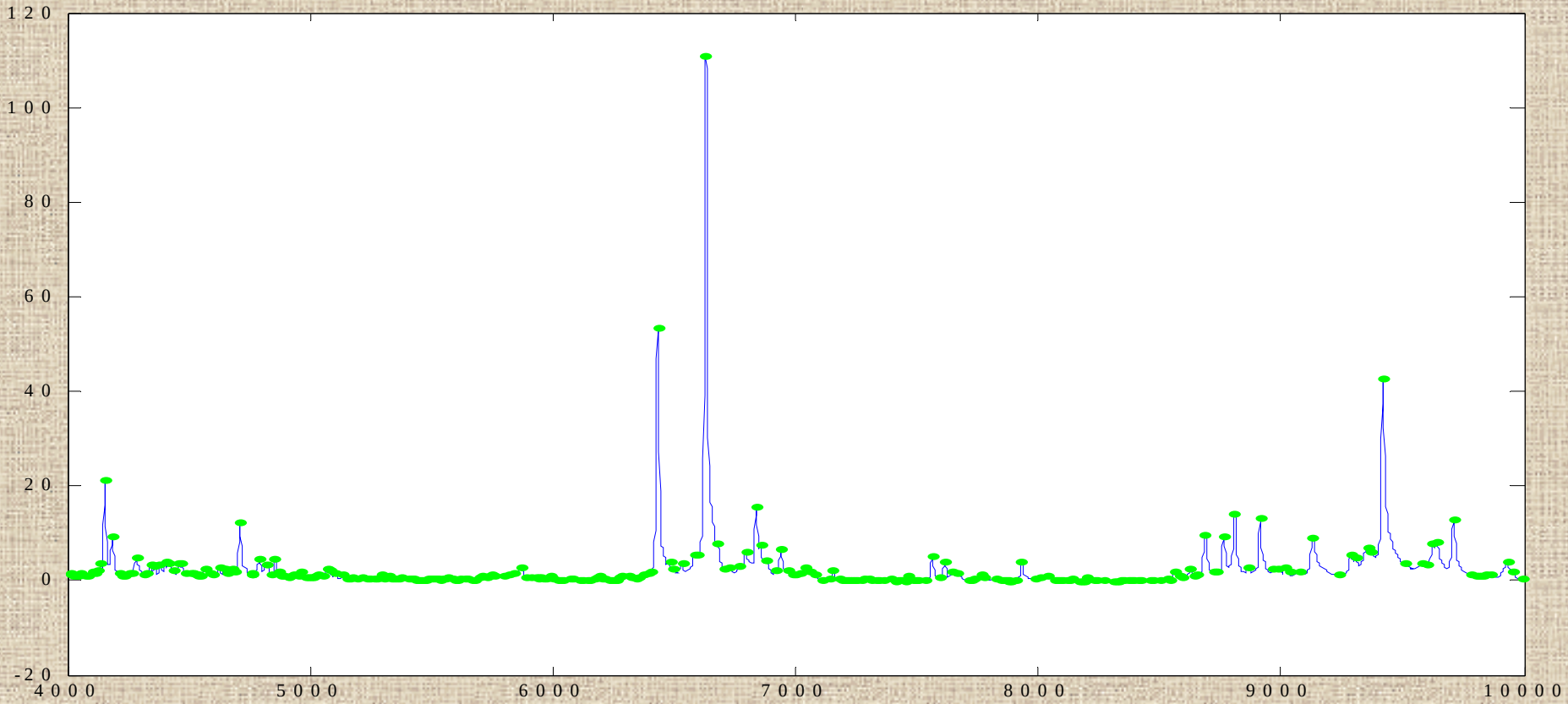
3. Project wavelet-space estimates (posterior samples) back to data space.

- **Apply IDWT to posterior samples of B^* to get posterior samples of fixed effect functions $B_j(t)$ for j=$1,\ldots, p$, on grid t.**
  - **B=B*W**
- **Posterior samples of $U_k(t)$, Q, and S are also available, if desired.**
- **Can be used for Bayesian inference/prediction**

Bayesian Inference:
Peak Detection

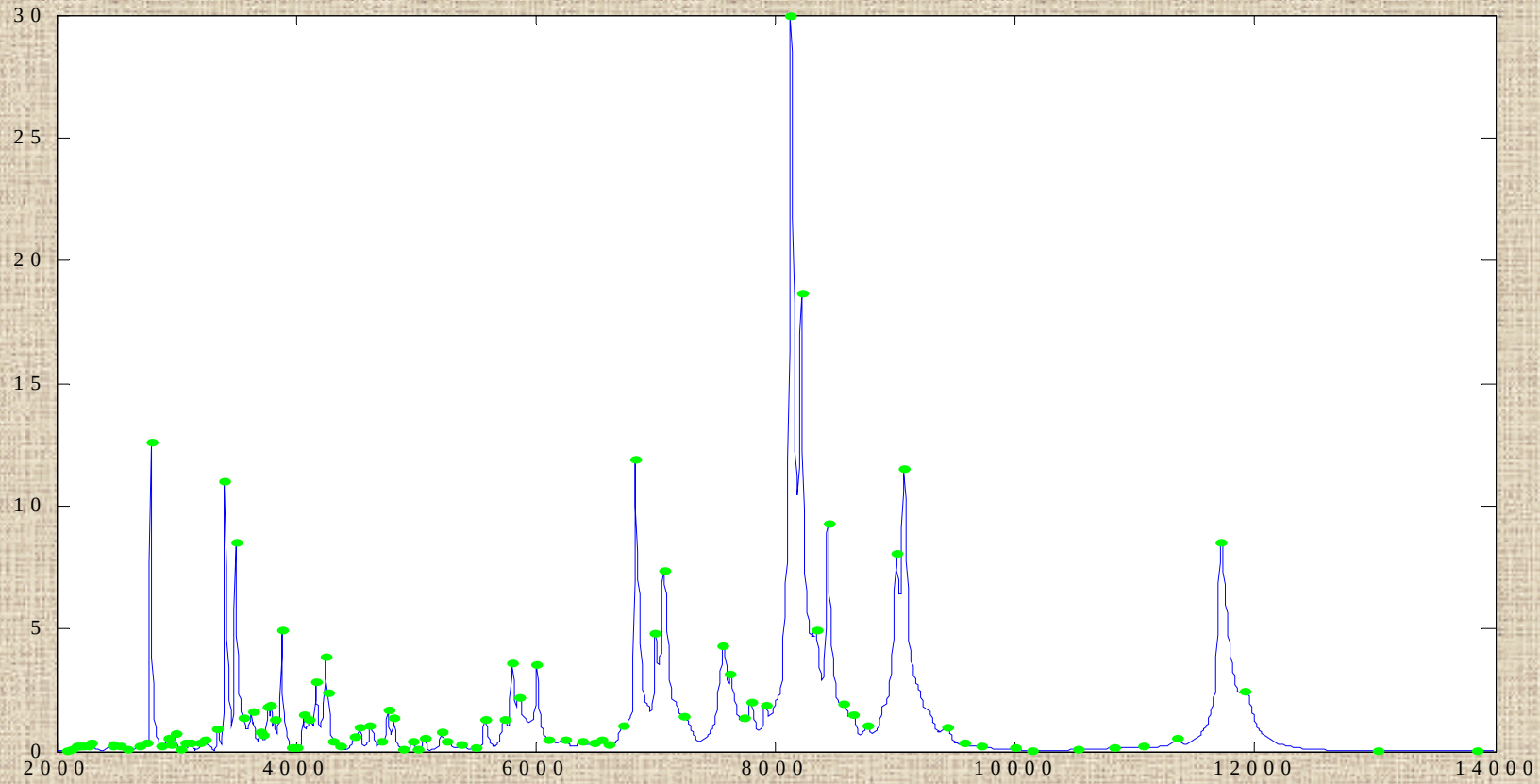
- **Focus specifically on peaks – locations in spectra likely to correspond to proteins/peptides**
- **Can use posterior mean estimate of overall mean spectrum for peak detection (Morris et al. 2005)**
- **All local maxima in (denoised) overall mean spectrum considered peaks, possibly subject to some threshold on Signal-to-Noise ratio ($S/N>\delta$)**
- **Let *K*=# of peaks found**

Pancreatic Cancer:
Peak Detection



- ***K=370* peaks detected**

Organ-by-Cell Line:
Peak Detection



- **Found *K=102* peaks (*58* with *S/N*>2)**

Bayesian Inference:

Identifying Differentially Expressed Peaks

- **Identify which peaks are related to clinical factors of interest (cancer/normal, organ, cell line, interaction)**

Procedure:

1. **Compute posterior probability of differential expression for each peak using posterior samples for suitable fixed effect function (2-sided)**

$$p_{ij}=\min[\Pr\{B_j(t_i)>0\},\Pr\{B_j(t_i)<0\}]$$

$$i=1,\ldots,K \qquad j=1,\ldots,p$$

2. **Rank peaks based on p_{ij}**

Bayesian Inference:

Identifying Differentially Expressed Peaks

Procedure:

1. **Rank peaks in ascending order of their 2-sided posterior probabilities of differential expression.**

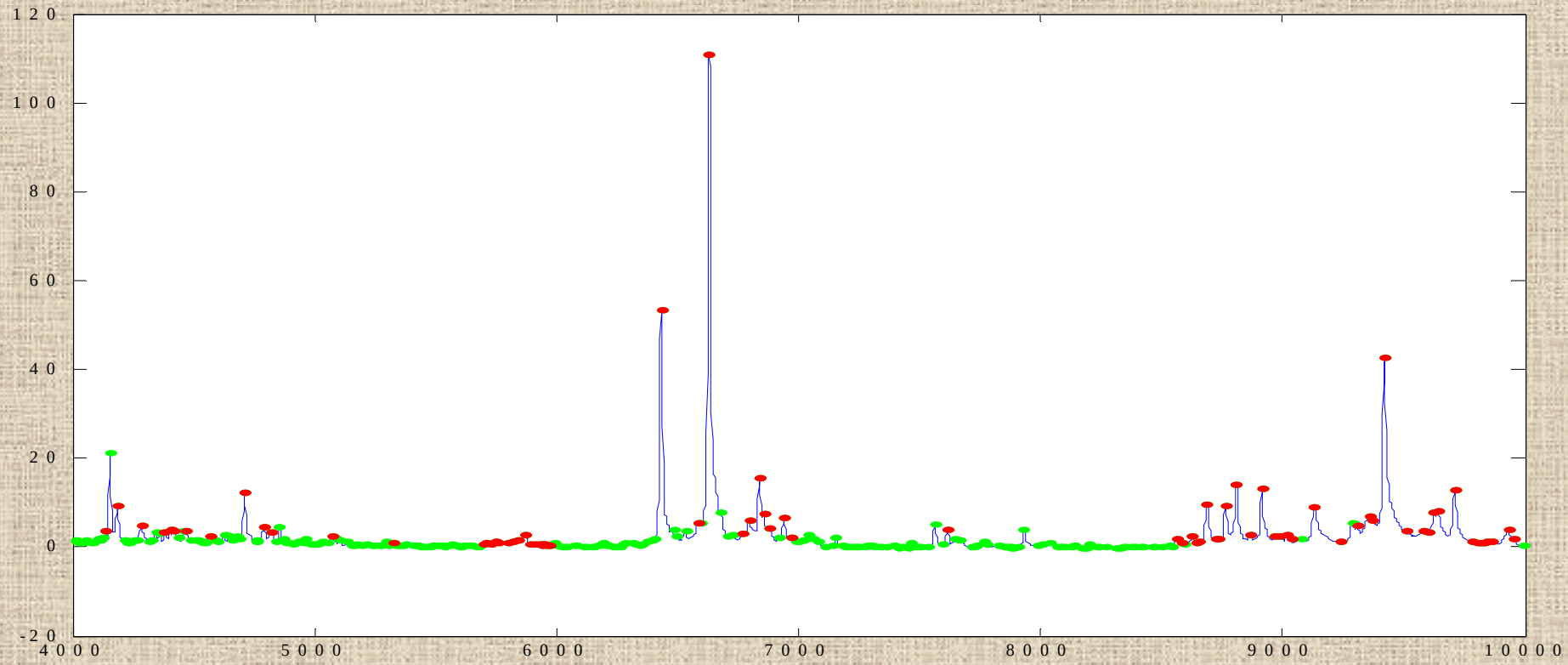
   $p_{(1)}, p_{(2)}, \cdots, p_{(pK)}$

2. **Find K^* such that:**

$$(K^*)^{-1} \sum_{k=1}^{K^*} p_{(k)} < \alpha / 2$$

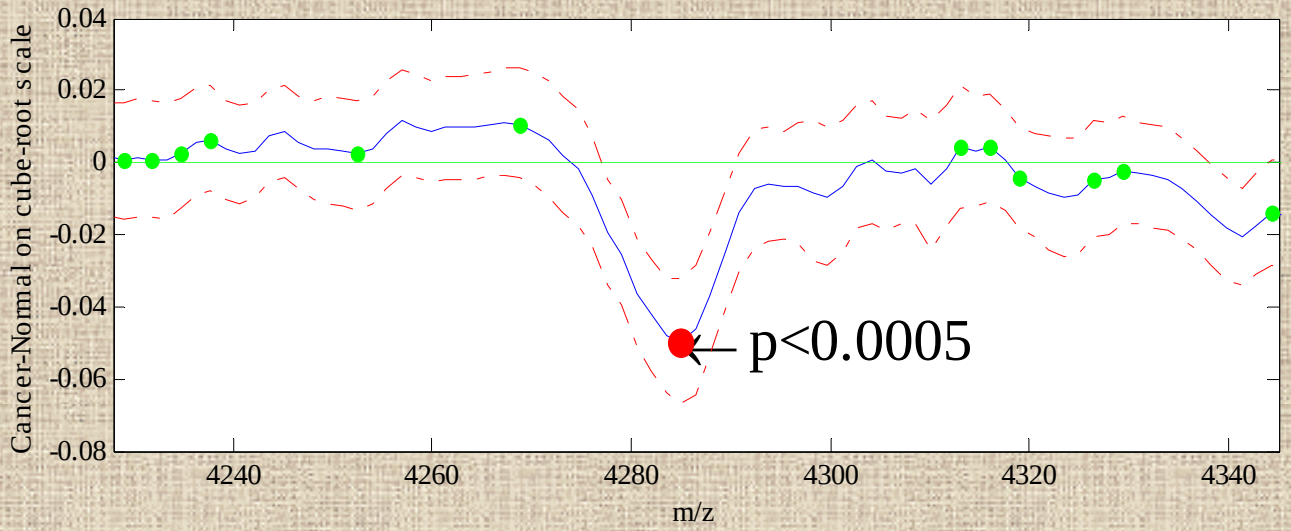
3. **Let $\psi = p_{(K^*)}$. Any peak i with $p_{ij} < \psi$ is called "differentially expressed" for outcome j**

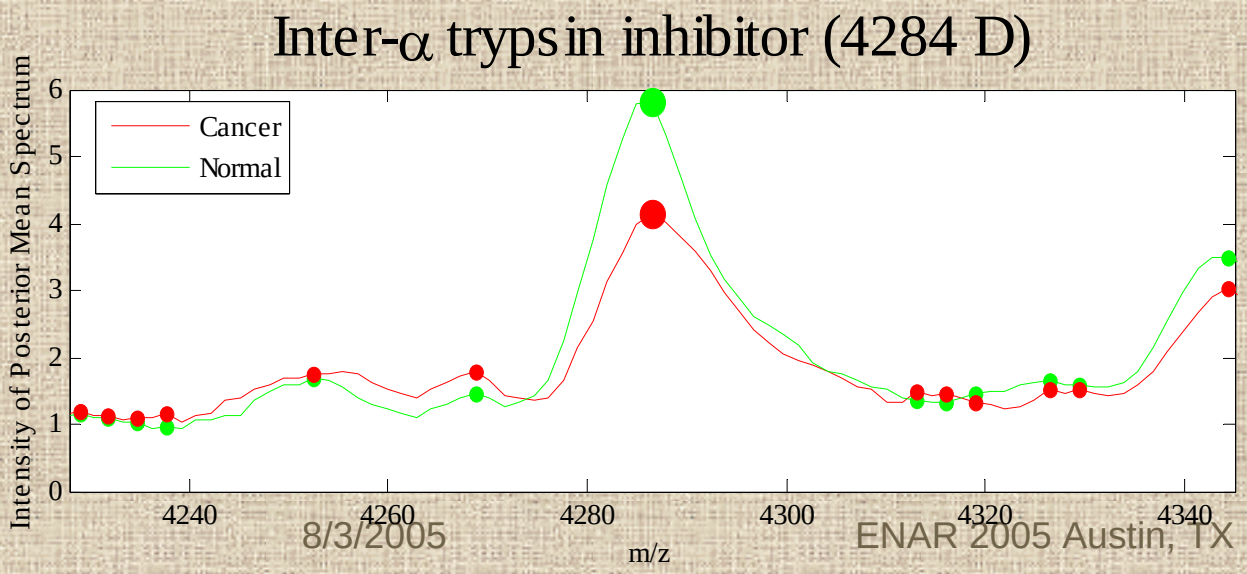
Pancreatic Cancer:
Differentially Expressed Peaks



- **370 peaks detected**
- **83 differentially expressed using α=0.01**

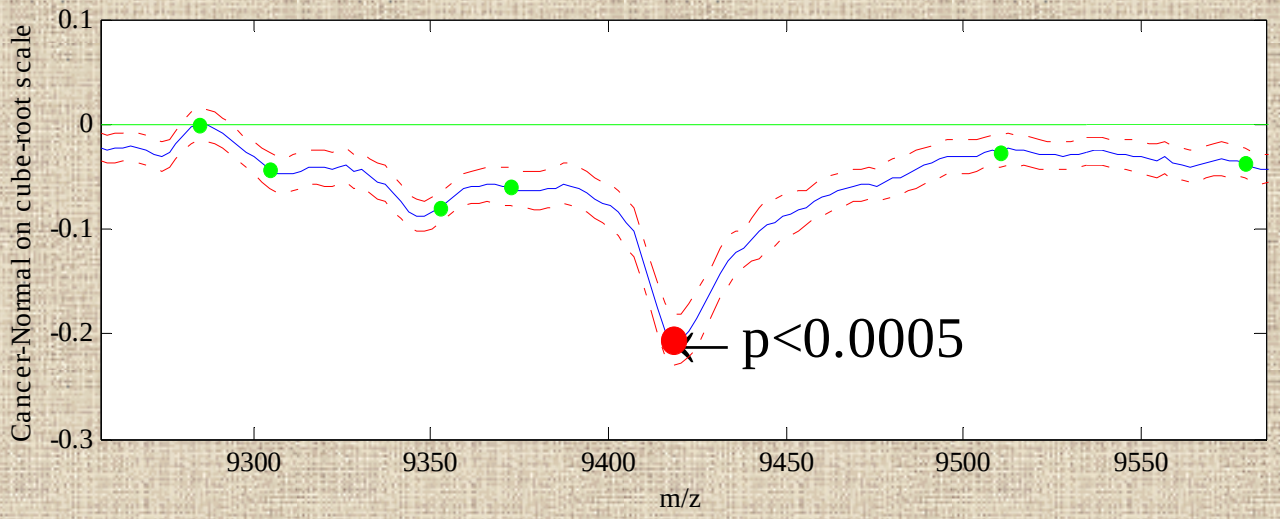
Pancreatic Cancer: Results



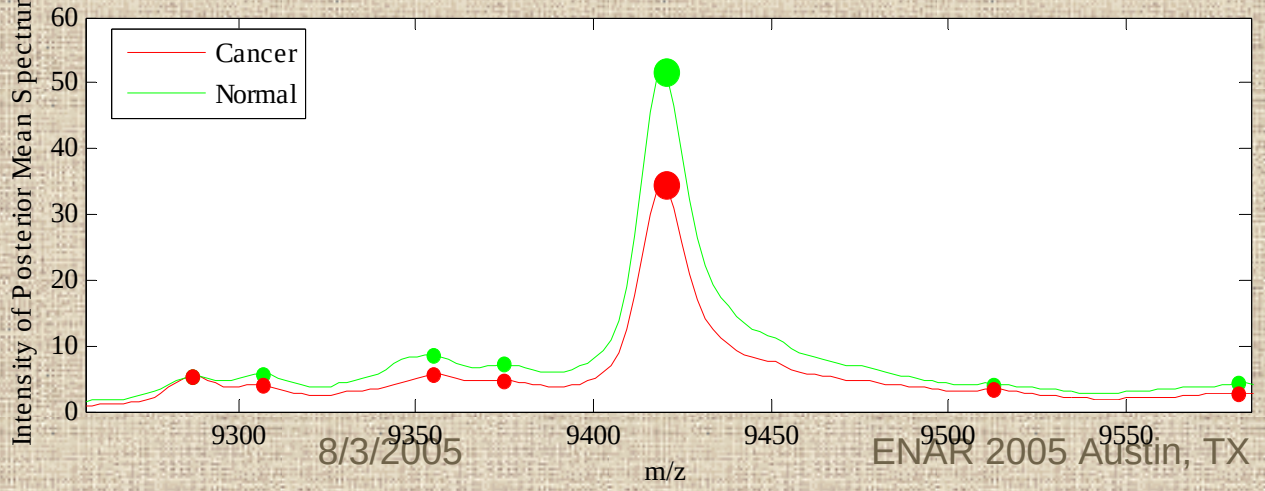


- Known to be related to pancreatic cancer
- Under-expressed in serum of cancerous patients
- May not be specific to pancreatic cancer

Pancreatic Cancer: Results

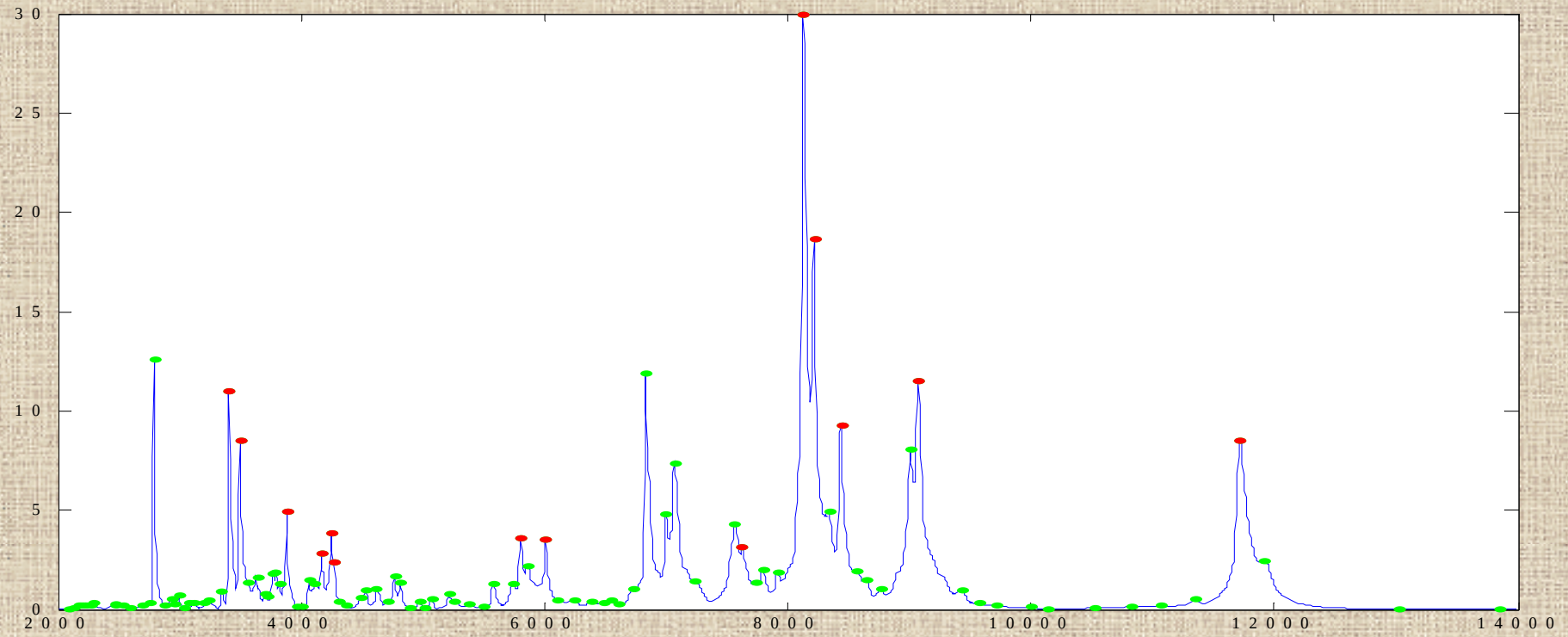


Small inducible cytokine B14 precursor (9419 D)



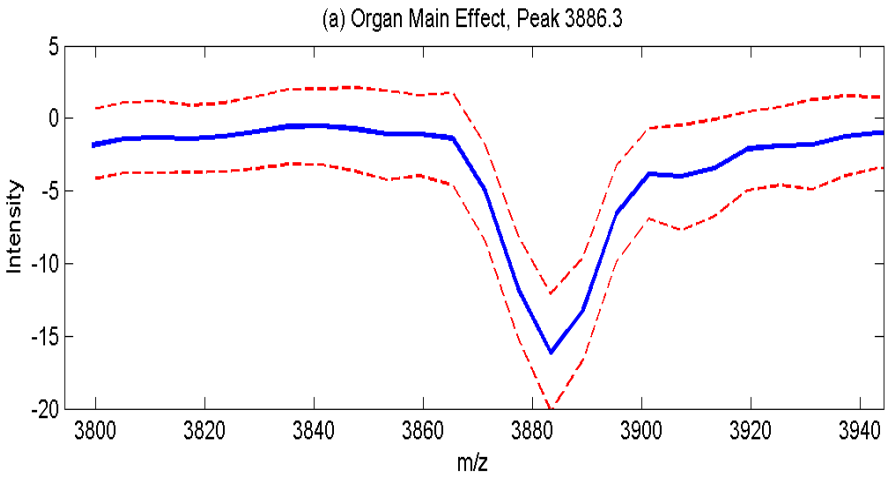
- Secreted from various organs, including pancreas
- Highly expressed in normal tissue with no inflammatory response
- Low expression in cancer cell lines

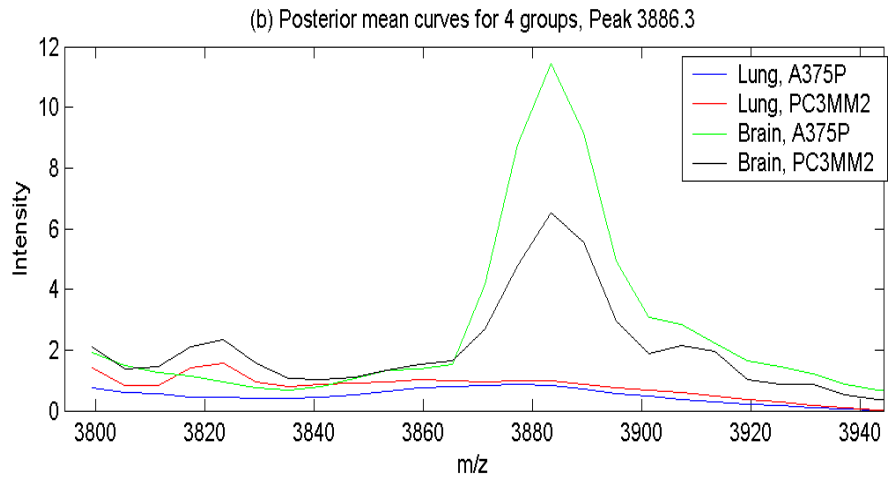
Organ-by-Cell Line:
Differentially Expressed Peaks



- **14/102 differentially expressed using α =0.01**
- **5 interaction, 2 organ, 3 cell line, 4 organ+cell line**

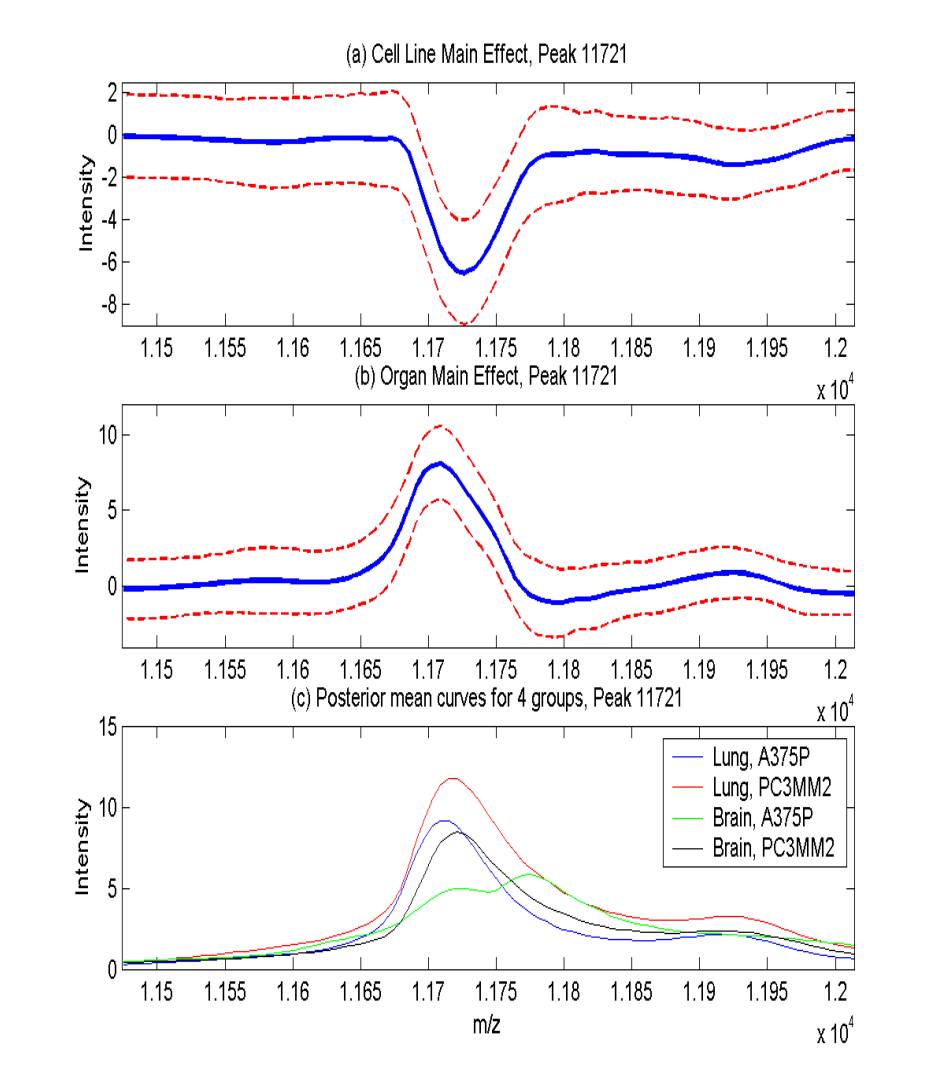
Organ-by-Cell Line: Results





- Specific to brain-injected mice
- May be CGRP-II (3882.34 Dal), peptide in mouse proteome that dilates blood vessels in brain
- Host response to tumor implanted in brain?

Organ-by-Cell Line: Results

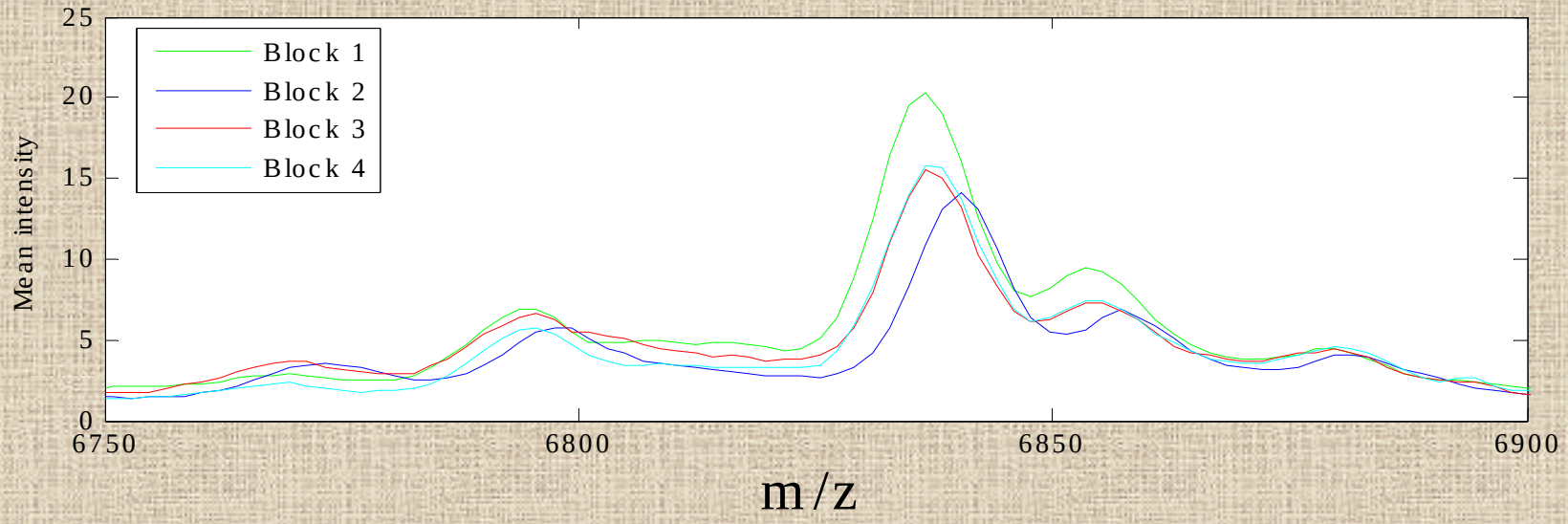


- **Higher in mice injected with metastatic (PC3-MM2) cell line**
- **May be MTS1 (11721.43 Dalt), metastatic cell protein in mouse proteome.**
- **Also higher in lung-injected mice than brain-injected mice**

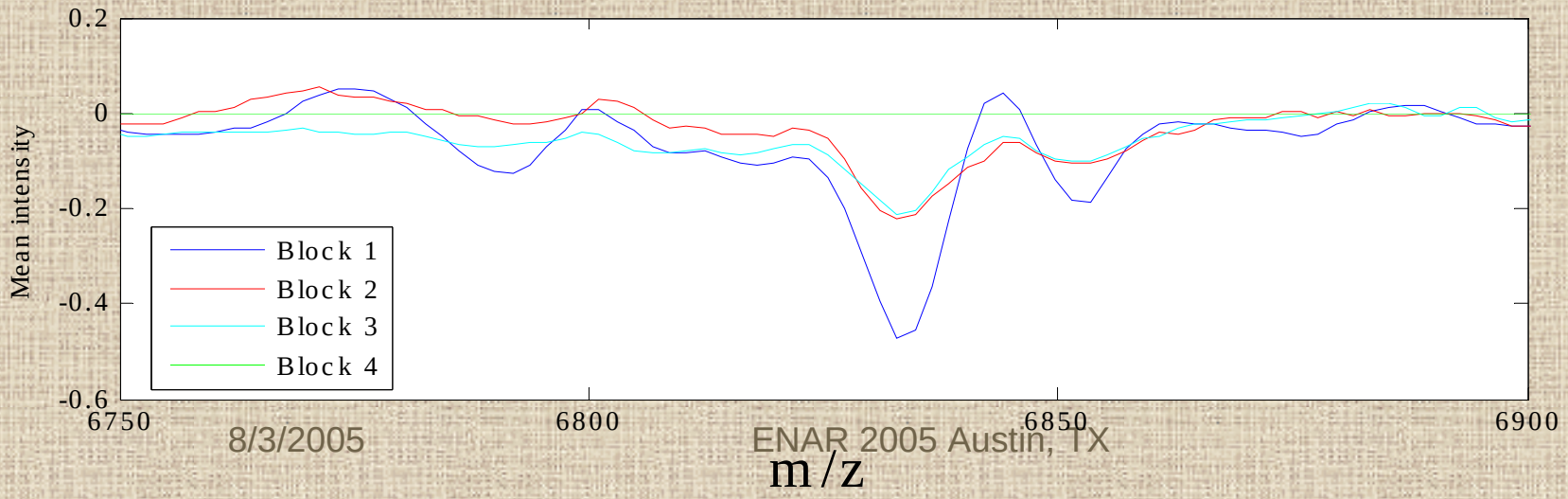
Bayesian Inference:
Investigating Block Effects

- By including fixed effect for blocks, we can adjust for systematic differences in spectra from different blocks (time blocks, laser intensity)
  - Systematic shifts in spectral intensities (y)
  - Systematic shifts in peak locations (x)
- These adjustments are done automatically by the model-fitting.
  - Flexibility of nonparametric fixed effects allows us to adjust for arbitrarily nonlinear misalignments

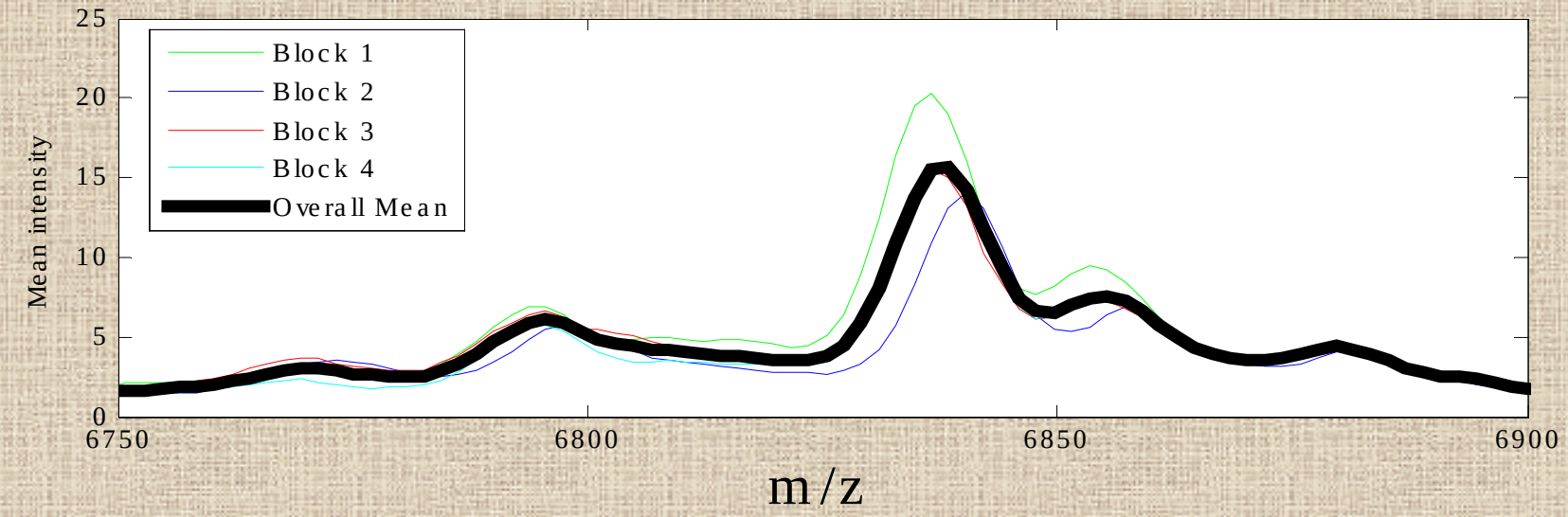
Pancreatic Cancer: Block Effects



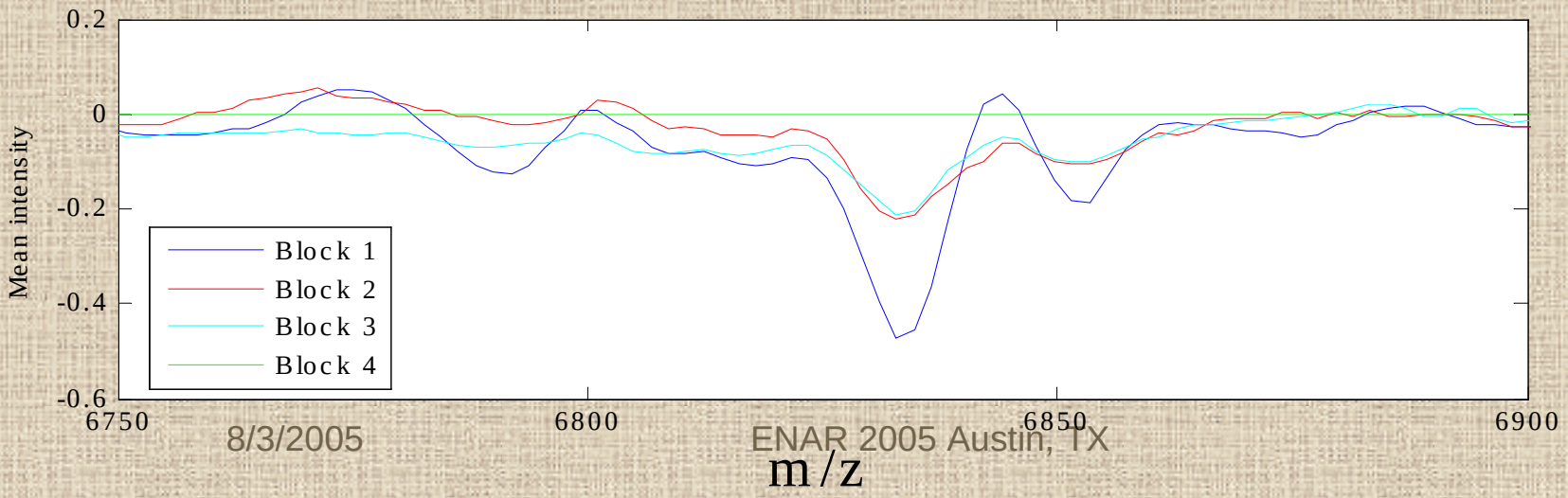
Block Effects



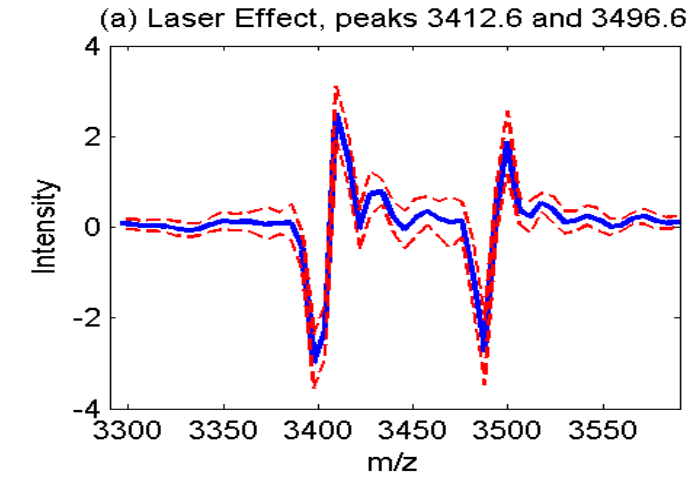
Pancreatic Cancer: Block Effects

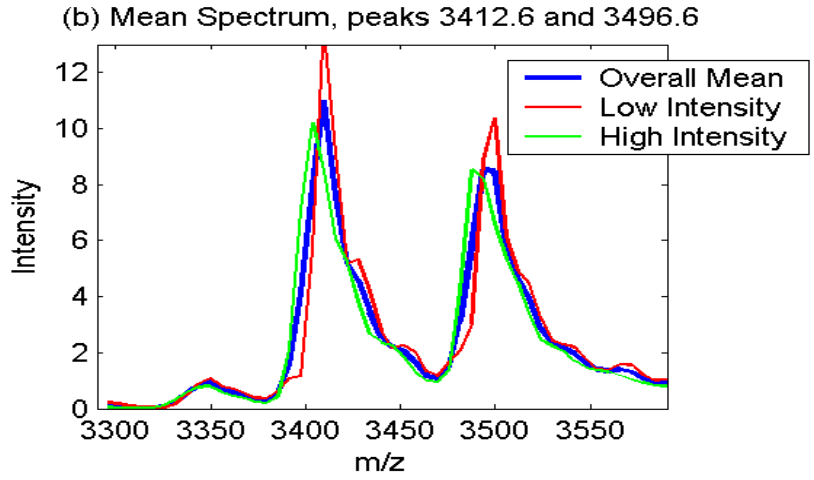


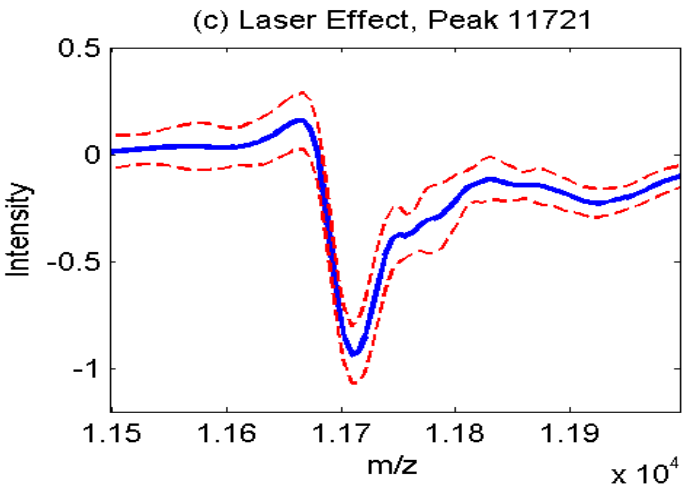
Block Effects

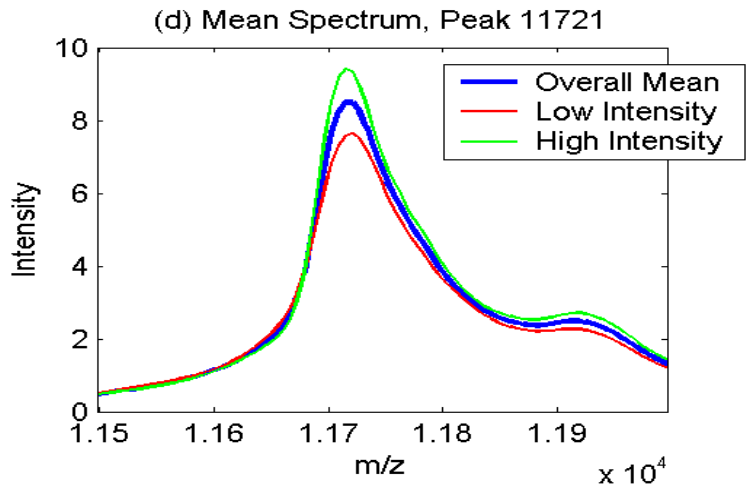


Organ-by-Cell Line: Block Effects









Bayesian Inference: Discrimination/Classification

- New samples can be classified as Cancer/Normal based on their spectra 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
- Can be used to perform classification without having to first do feature selection

Bayesian Inference: Discrimination/Classification

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

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

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

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

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

Bayesian Inference: Discrimination/Classification

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

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

Pancreatic Cancer: Classification Accuracy

	Accuracy	Sensitivity	Specificity
Training Data	81%	78%	83%
Test Data (8-fold CV)	70%	73%	66%

- Koomen, et al. 2004: 90% sensitivity, 77% specificity
- Used entire spectrum and all 4 fractions
- We only used small region of 1 fraction – doing others

Pancreatic Cancer: Classification Accuracy

- Performance improved by not using all wavelet coeffs
  - Leave out those likely to be unrelated to peaks
  - Lowest frequencies removed (j=1,2,3,4): baseline
  - Highest frequency removed (j=16): noise

	Accuracy	Sensitivity	Specificity
Training Data	83%	78%	89%
Test Data (8-fold CV)	74%	75%	73%

Discussion

- **Flexible method for modeling mass spectrometry data**
  - **Multiple fixed effects**
  - **Block effects**
  - **Random effects**
- **Various types of inference possible**
  - **Peak detection, differentially expressed peaks, control FDR, classification without feature selection**
- **Easy-to-use code being developed**
  - **Only necessary inputs: Y, X, Z matrices**
  - **Available by end of Summer 2005.**
- **Method also applies to other types of functional data.**

Acknowledgements

- **Co-authors: Philip J. Brown, Kevin R. Coombes, Keith A. Baggerly**
- **Collaborators on other WFMM projects: Raymond J. Carroll, Marina Vannucci, Louise Ryan, Brent Coull, Naisyin Wang, Betty Malloy**
- **SELDI/MALDI Data: John Koomen, Nancy Shih, Josh Fidler, Stan Hamilton, Donghui Li, Jim Abbruzzesse, and Ryuji Kobayashi**
- **Thanks to Dick Herrick for assistance in optimizing the code for the method, and for converting the Matlab code to C++.**

Wavelet-Based Hierarchical Functional Models

- **Most existing wavelet regression methods are for single function case**
- **Morris, Vannucci, Brown, and Carroll (2003)**
  - **Bayesian wavelet-based method for estimating mean function for functional data from nested design.**
  - **Extended wavelet regression to hierarchical functional context.**
- **Morris and Carroll (2004)**
  - **Extended to functional mixed model framework**
  - **Allowed nonstationary covariance structures**

Example: Model Fitting

- Daubechies 8 wavelet basis, J=11 levels
- Empirical Bayes procedure used to estimate regularization parameters π_{ij} and τ_{ij} from data.
- Burn-in 1000; 20,000 MCMC samples; thin=10
- Took 7hr 53min on Win2000 P-IV 2.8GHz 2GB RAM
  - That is Matlab code; C++ code takes ~2 hours.
- Trace plots indicated good convergence properties
- Metropolis Hastings acceptance probabilities good:
  - Range of (0.04, 0.53)
  - (10^{th},50^{th},90^{th}) percentiles of (0.20, 0.29, 0.50)

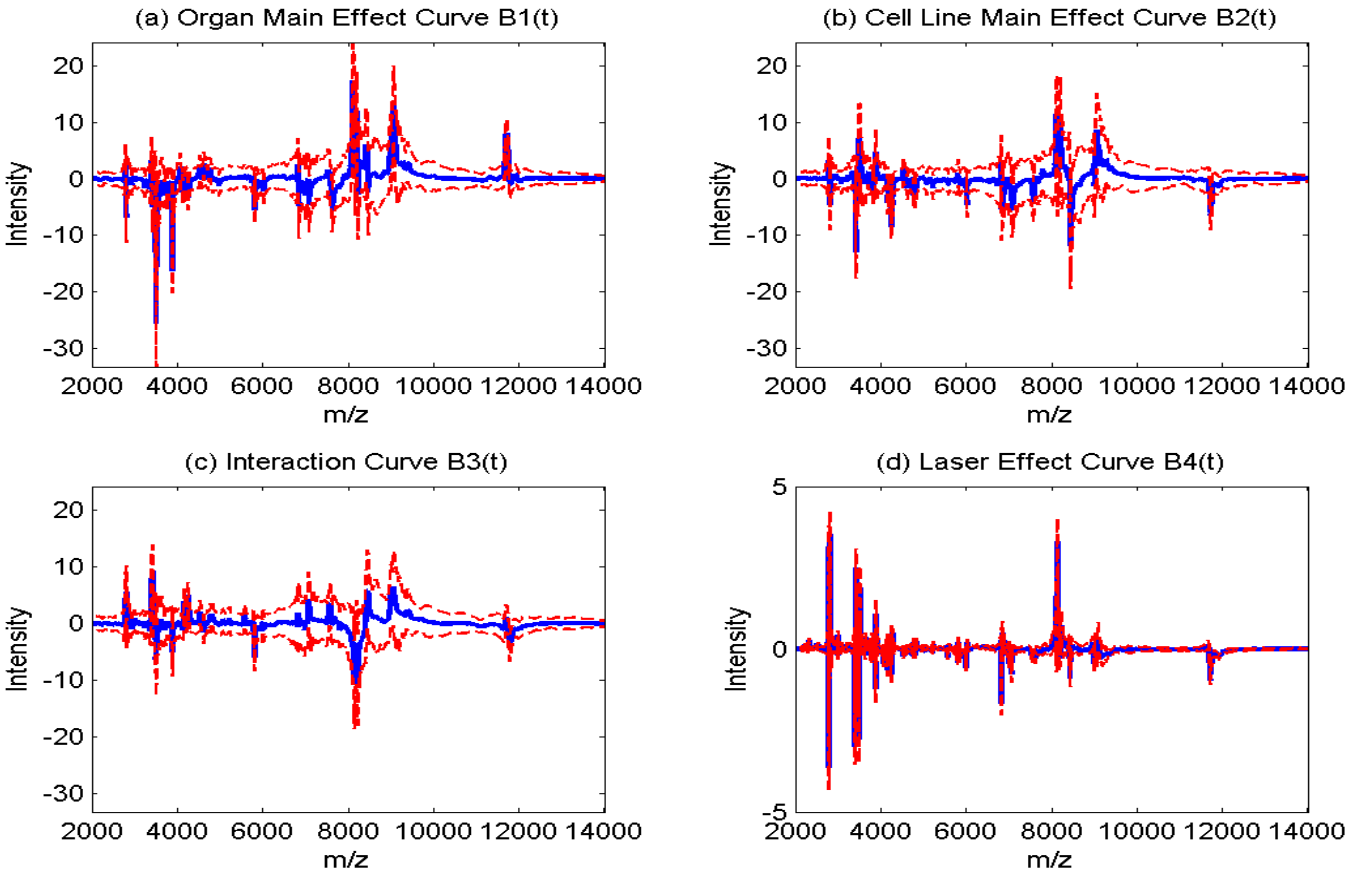
Discussion

- **Introduced unified modeling approach for FDA**
  - **Applied here to MALDI-TOF, but method is general.**
- **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**
  - **Posterior credible intervals; pointwise or joint**
  - **Predictive distributions for future sampled curves**
  - **Predictive probabilities for group membership of new curves**
  - **Bayesian functional inference can be done via Bayes Factors**
- **Since a unified modeling approach is used, all sources of variability in the model propagated throughout inference.**

Discussion

- **Since functions adaptively regularized using wavelet shrinkage, the method is appropriate for spatially heterogeneous functional data.**
- **Approach is Bayesian. The only informative priors to elicit are regularization parameters, which can be estimated from data using empirical Bayes.**
- **Method generalizes to higher dimensional functions, e.g. image data, space/time (fixed domain) data.**
- **We used wavelet bases, but approach can be generalized to other orthogonal basis functions.**
- **Major challenges in developing unified statistical modeling approach for replicated functional data, but worth the effort.**

Organ-by-Cell Line: Results



Organ-by-Cell Line: Flagged peaks

Detecting 'significant' peaks: Top 9 peaks

m/z	Effect	*p*	Comment
3412.6	int.	<0.0005	PC3MM2>A375P for brain-injected only
3496.6	organ	<0.0005	Only expressed in brain-injected mice
3886.3	organ	<0.0005	Only expressed in brain-injected mice
4168.2	int.	0.0005	PC3MM2>A375P in brain-injected only
4252.1	int.	<0.0005	PC3MM2>A375P in brain-injected only
4270.1	cell line	<0.0005	PC3MM2>A375P
5805.3	int.	<0.0005	brain>lung only for mice given A375P cell-line
6015.2	cell line	<0.0005	PC3MM2>A375P
11721	cell line	<0.0005	PC3MM2>A375P
11721	organ	<0.0005	lung>brain