BAYESIAN METHODS FOR VARIABLE SELECTION WITH APPLICATIONS TO HIGH-DIMENSIONAL DATA

Part 3: Functional Data & Wavelets

Marina Vannucci

Rice University, USA

PASI-CIMAT
04/28-30/2010

Part 3: Functional Data & Wavelets

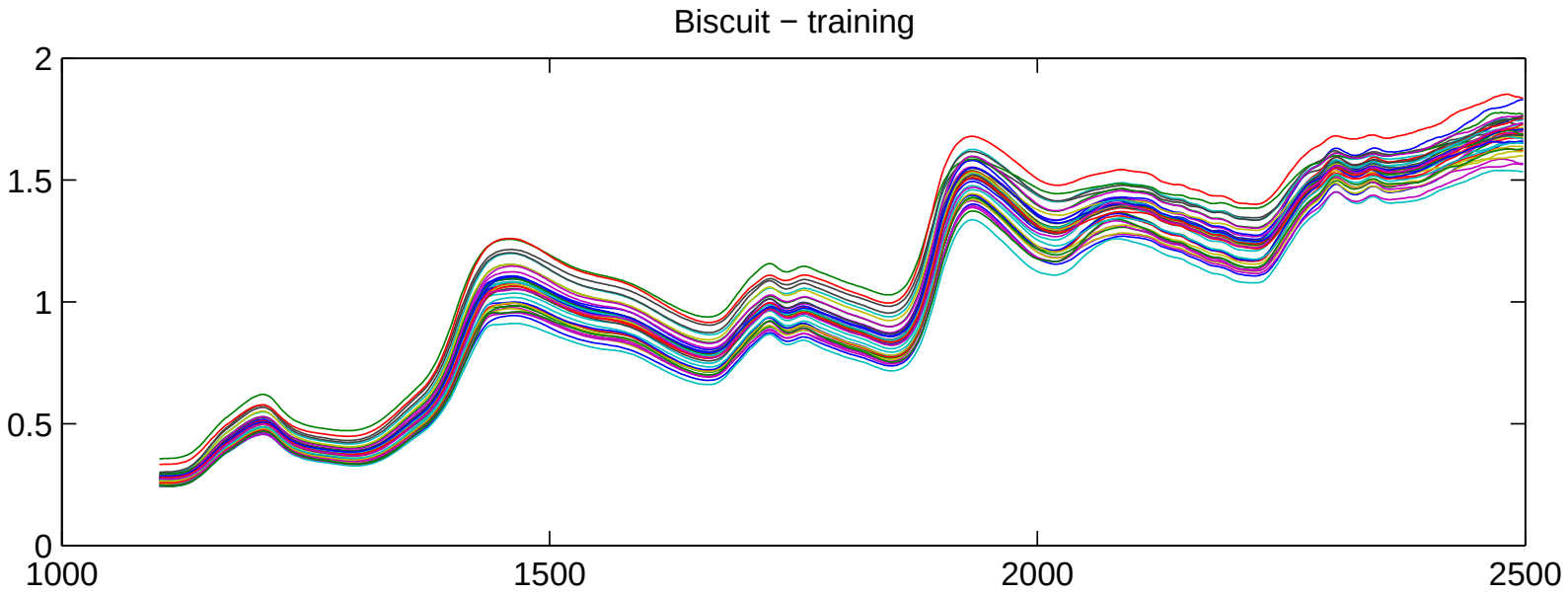
- Functional data
- Brief intro to wavelets
- Curve regression models
- Curve classification
- Applications to Near Infrared spectral data from Chemometrics

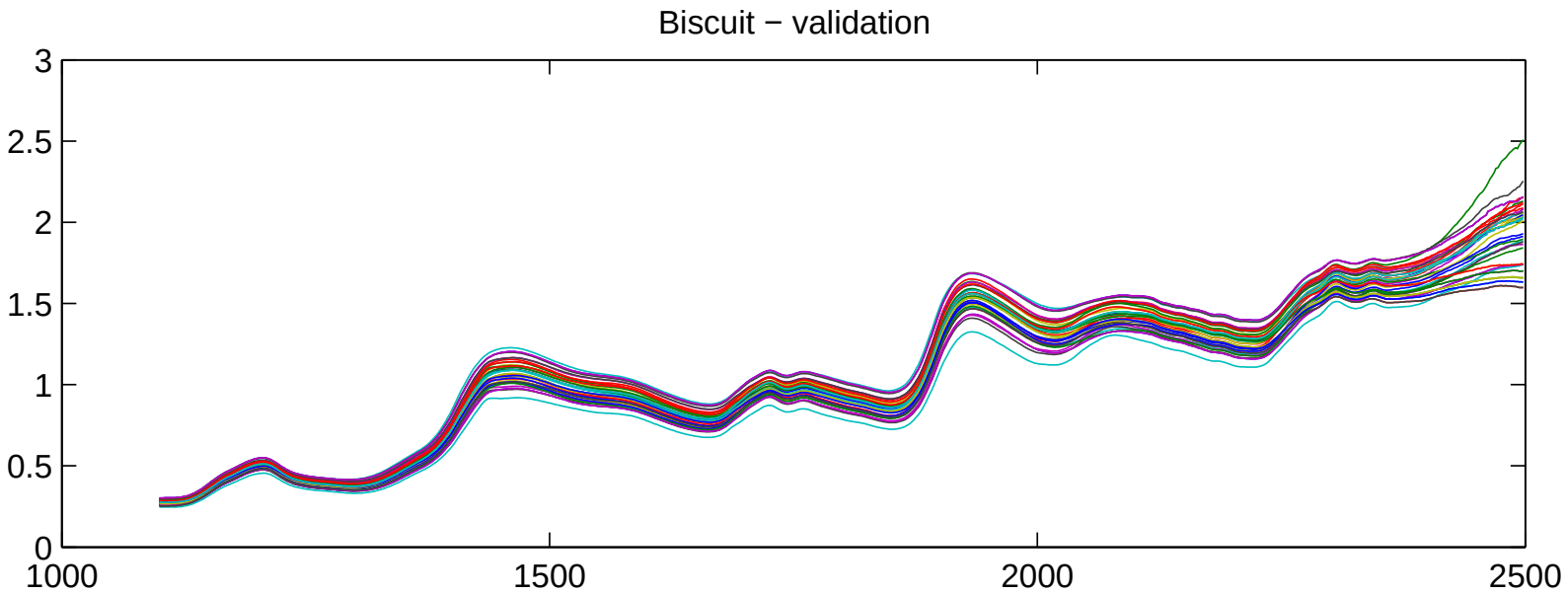
Overall objective: Methods for prediction or classification or clustering based on functional data (multiple curves).

- **Regression:** a continuos response is observed.
- **Classification:** class membership observed in a training set.
- **Clustering:** group membership needs to be uncovered.
- Data are curves.

Motivating Examples from NIR Calibration

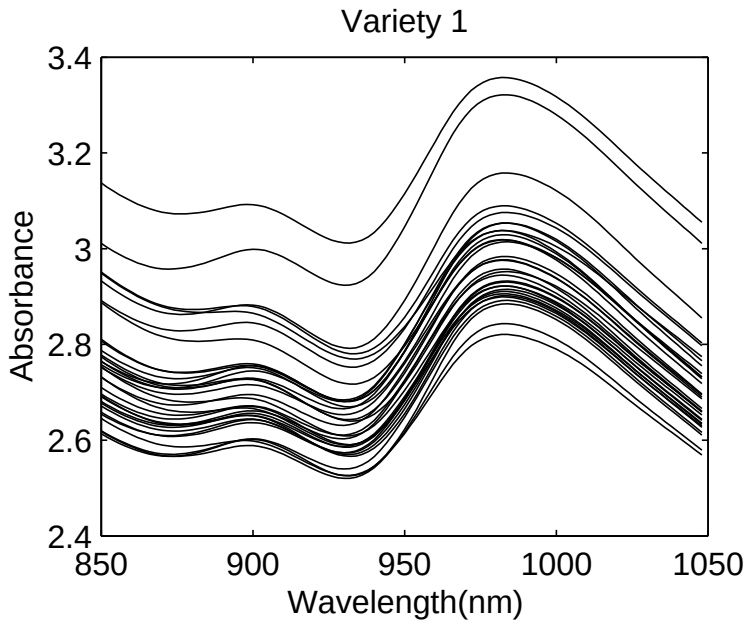
- Experiment: 40 biscuit doughs made with variations in quantities of fat, flour, sugar and water in recipe.
  Fat 15-21%, Sugar 10-23%, Flour 44-54%, Water 11-17%
- Data: Composition (by weight) of the 4 constituents and spectral data at 700 wavelengths (from 1100nm to 2498nm in steps of 2nm) for each dough piece.
- Aim: Use the spectral data to predict the composition.
  Wavelets as a dimension reduction tool that preserves local features.

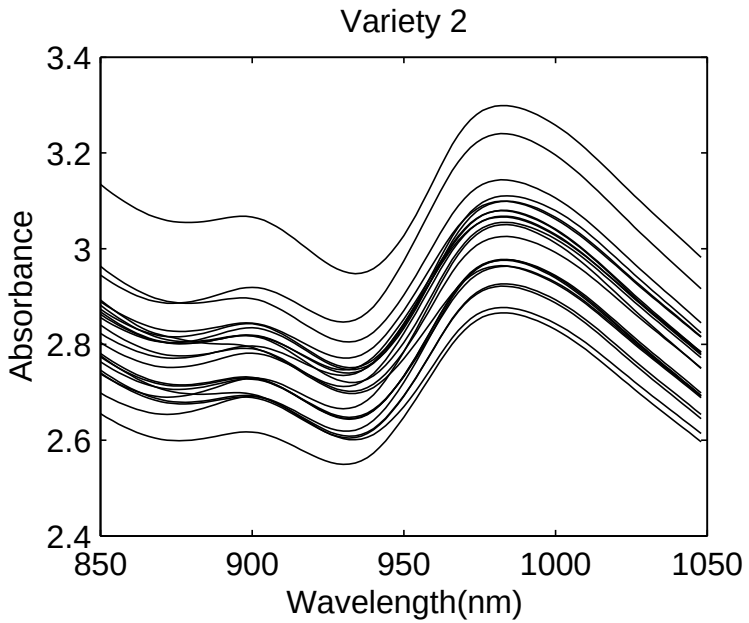

Biscuit − training
2
1.5
1
0.5
0
1000
1500
2000
2500

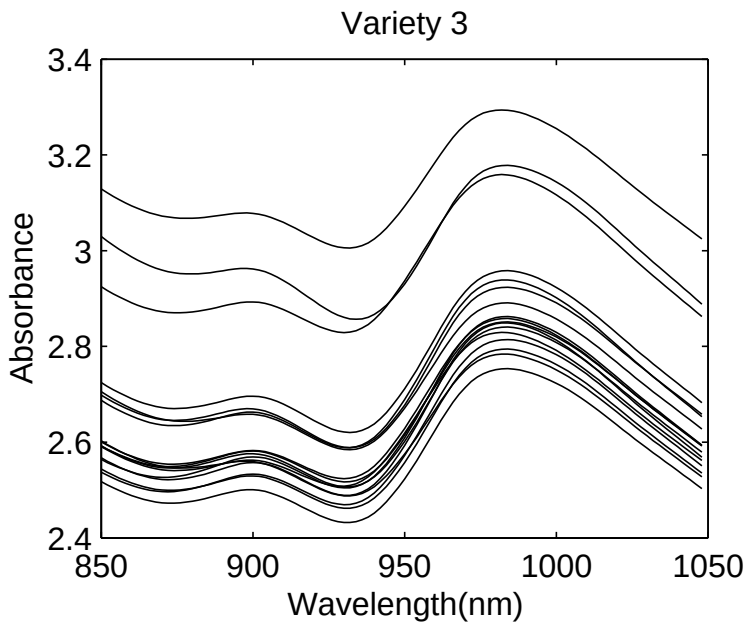

Biscuit − validation
3
2.5
2
1.5
1
0.5
0
1000
1500
2000
2500

Another Example

- NIR spectra used in analysis of food and drink and pharmaceutical products, measured at hundreds of wavelengths.
- 3 varieties of wheat, 94 observations. NIR spectra with 100 wavelengths from 850 to 1048nm in steps of 2nm.
- Aim: Use the spectral data to predict the wheat variety.


Variety 1
Absorbance
3.4
3.2
3
2.8
2.6
2.4
850
900
950
1000
1050
Wavelength(nm)


Variety 2
Absorbance
3.4
3.2
3
2.8
2.6
2.4
850
900
950
1000
1050
Wavelength(nm)


Variety 3
Absorbance
3.4
3.2
3
2.8
2.6
2.4
850
900
950
1000
1050
Wavelength(nm)

Our Approach

- Use wavelets as dimension reduction tool that preserves local features. Apply DWT to curves. Wavelet coefficients summarise curve features in an efficient way, they are localized (unlike Fourier coefficients) and can capture noise in the data.
- Develop Bayesian methods in wavelet domain for simultaneous feature selection and prediction or classification. We use mixture priors and Bayes methods to look for wavelet component that well describe the variation in the responses.
- Wavelet coeffs selection is done by assigning a probability to every possible subset and then searching for subsets with high probability. Small coefficients can be important.

Mayor Milestones on Wavelets

The basic idea behind wavelets is to represent a generic function with simpler functions (building blocks) at different scales and locations.

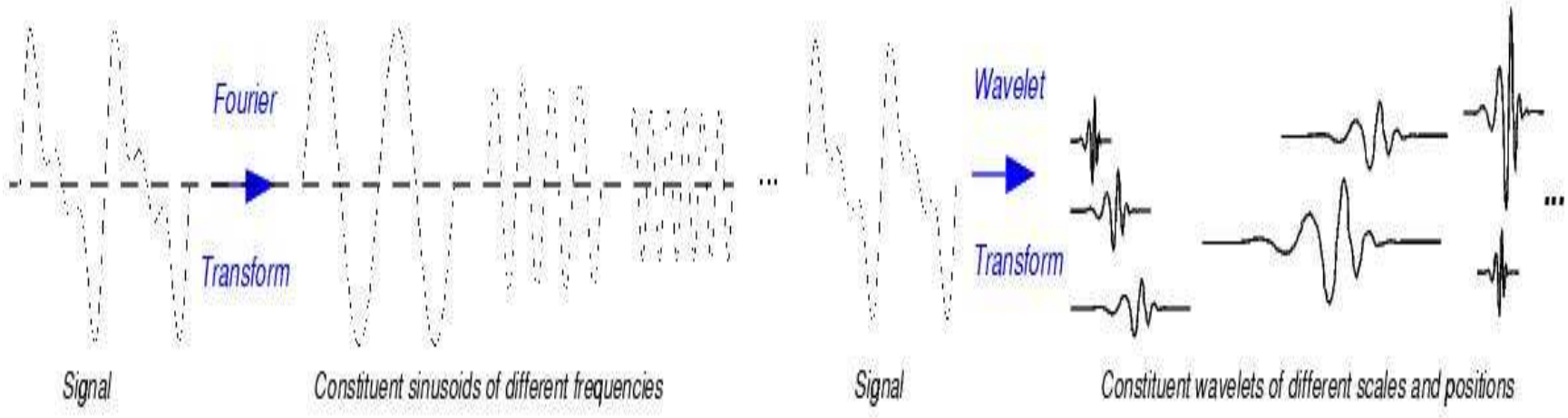
- **1807:** Fourier orthogonal decomposition of periodic signals
- **1946:** Gabor windowed Fourier transform (STFT)
- **1984:** A. Grossmann and J. Morlet introduce the **continuous** wavelet transform for the analysis of seismic signals.
- **1985:** Y. Meyer defines **discrete** orthonormal wavelet bases.
- **1989:** S. Mallat links wavelets to the theory of "multiresolution analysis" (MRA) a framework that allows to construct orthonormal bases. A **discrete wavelet transform** is defined as a simple recursive algorithm to compute the wavelet decomposition of a signal from its approximation at a given scale.
- **1989:** I. Daubechies constructs wavelets with compact support and a varying degree of smoothness.
- **1992:** D. Donoho and I. Johnstone use wavelets to remove noise from data.

Wavelets vs Fourier Representations

Given f and a basis $\{f_1, \ldots, f_n\} \rightarrow$ series expansion

$$f(x) = \sum_i a_i f_i(x) dx$$

$$a_i =< f, f_i >= \int f(x) f_i(x) dx$$

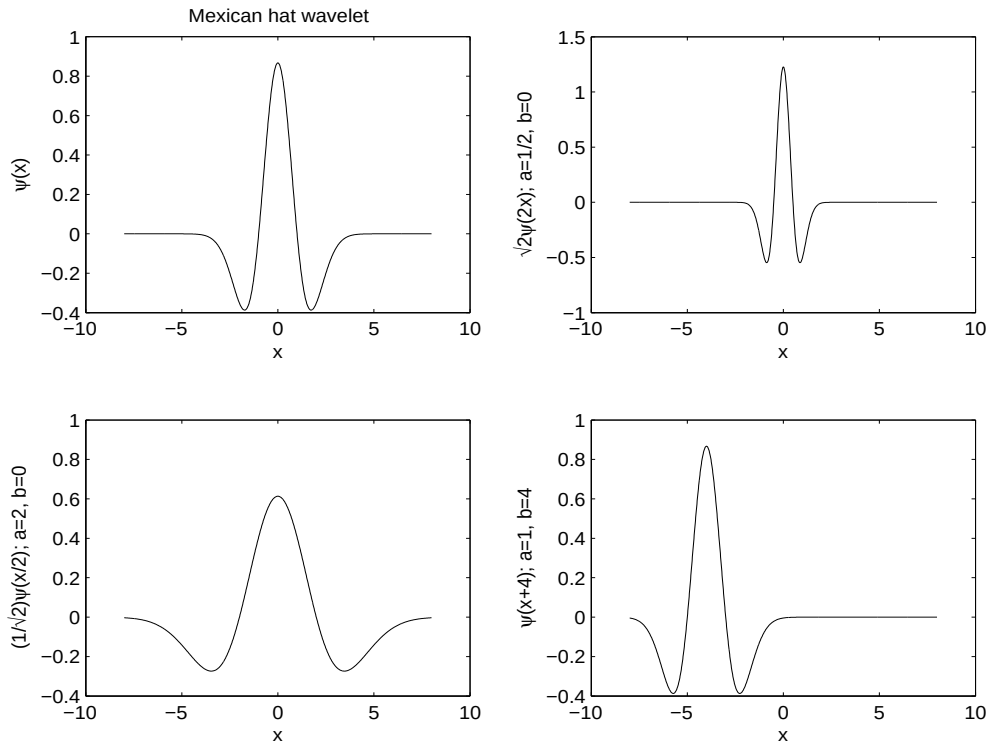


Fourier transforms measure the frequency content of a signal.
Wavelet transforms provide a *time-scale* analysis.

Wavelets as "Small Waves"

Mother wavelet ψ as an oscillatory function with zero mean

$$\psi(x) \quad : \quad \int_{\mathbf{R}} \psi(x)dx = 0$$



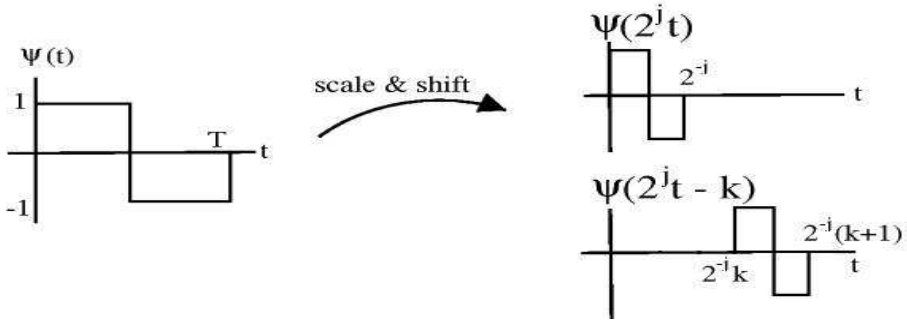
Wavelets as orthonormal basis

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$$

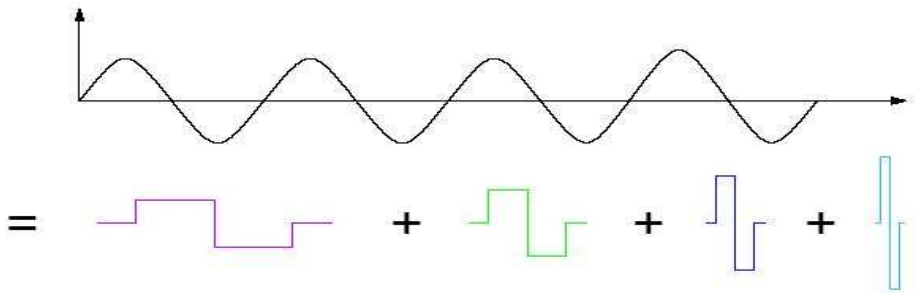
with j, k scale and translation parameters.

Examples of Wavelet Families

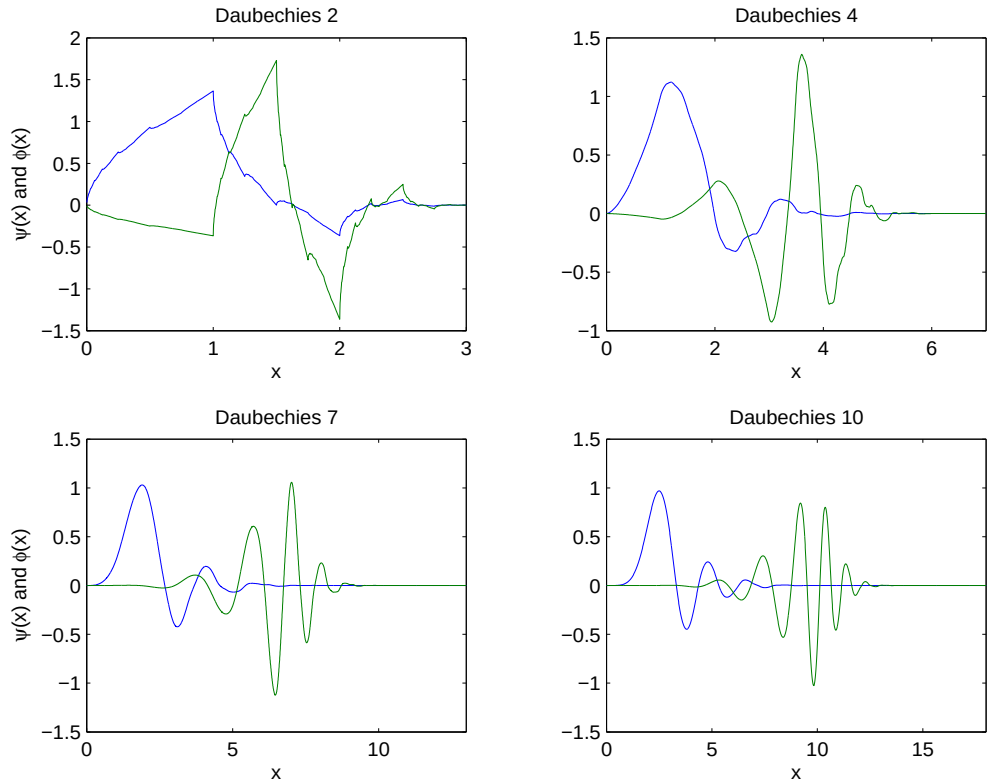
Haar wavelets. The simplest family of wavelets, already known before the formulation of the wavelet theory (Haar, 1909).
$\psi(x) = 1$ for $0 \leq x < 1/2$; -1 for $1/2 \leq x < 1$ and 0 otherwise.
Haar wavelets constructed from ψ via dilations and translations



Given a generic function *f*, Haar wavelets approximate *f* with piecewise constant functions (not continuous)



Daubechies Wavelets



- Compact support (good time localization)
- Vanishing moments $\int x^l \psi(x) dx = 0,\ l = 0, 1, \ldots, N$ ensure decay as $< f, \psi_{j,k} > \leq C 2^{-jN}$, (good for compression)
- Various degrees of smoothness. For large N, $\phi \in C^{\mu N}$, $\mu \approx 0.2$.

Properties of Wavelets

$$\text{Wavelet series: } f(x) = \sum_{j,k} < f, \psi_{j,k} > \psi_{j,k}(x)$$

$$\{W_f(j,k) = \langle f, \psi_{j,k} \rangle = \int f(x)\psi_{j,k}(x)dx\}_{j,k \in Z}$$

describing features of f at different locations and scales.

Properties:

- Small waves with zero mean
- Time-frequency localization
- Good at describing non-stationarity and discontinuities
- Multi-scale decomposition of functions (MRA) - sparsity, shrinkage
- Recursive relationships among coefficients $\rightarrow$ DWT

Wavelets in Practice (DWT)

Let's consider a vector $\mathbf{Y}$ of observations of f at n equispaced points

$$y_i = f(t_i), \quad i = 1, \ldots, n, \text{ with } n = 2^m$$

Discrete Wavelet Transforms operate via recursive filters applied to $\mathbf{Y}$

$$\text{data} \equiv \mathbf{c}^{(m)} \xrightarrow{\mathcal{H}} \mathbf{c}^{(m-1)} \xrightarrow{\mathcal{H}} \ldots \xrightarrow{\mathcal{H}} \mathbf{c}^{(m-r)}$$
$$\mathcal{G} \searrow \qquad \mathcal{G} \searrow \qquad \mathcal{G} \searrow$$
$$\mathbf{d}^{(m-1)} \qquad \ldots \qquad \mathbf{d}^{(m-r)}$$

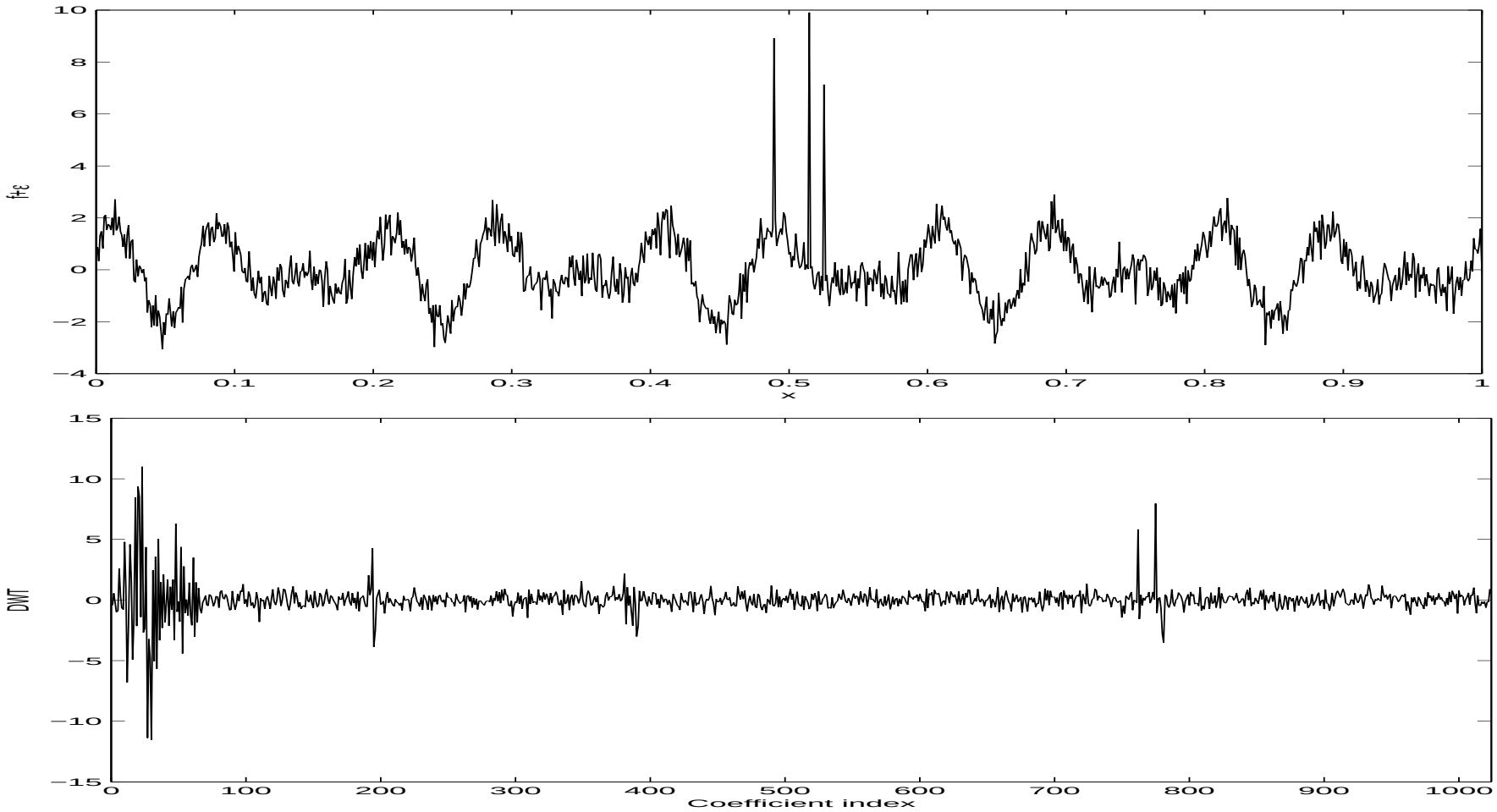
$$\mathcal{H} : c_{m-1,k} = \sum_l h_{l-2k} c_{m,l}, \quad \mathcal{G} : d_{m-1,k} = \sum_l g_{l-2k} c_{m,l}, \ldots$$

Then, in practice,

$\mathbf{Y} \xrightarrow{DWT} \mathbf{d}^{(\mathbf{Y})} = (\mathbf{c}^{(m-r)}, \mathbf{d}^{(m-r)}, \mathbf{d}^{(m-r-1)}, \ldots, \mathbf{d}^{(m-1)})$

discrete approx of $< f, \psi_{j,k} >$ at scales $m-1, \ldots, m-r$

Example



$$\text{data} \equiv \mathbf{Y} \xrightarrow{DWT} \mathbf{d}^{(\mathbf{Y})} = (\mathbf{c}^{(m-r)}, \mathbf{d}^{(m-r)}, \mathbf{d}^{(m-r-1)}, \ldots, \mathbf{d}^{(m-1)})$$

Matrix Notation of the DWT

$$y_i = f(t_i), \quad i = 1, \ldots, n, \text{ with } n = 2^m$$

- DWT: $\mathbf{Y} \stackrel{DWT}{\longrightarrow} \mathbf{d}^{(\mathbf{Y})} = \mathbf{WY}$, $\mathbf{W}$ determined by ψ, $\mathbf{W}'\mathbf{W} = \mathbf{I}$
- Variances and covariances of DWT coefficients:

$$\mathbf{\Sigma}_{\mathbf{d}^{(\mathbf{Y})}} = \mathbf{W}\mathbf{\Sigma}_{\mathbf{Y}}\mathbf{W}'$$

for given $\Sigma_{\mathbf{Y}}(i,j) = [\gamma(|i-j|)]$
$\gamma(\tau)$ autocovariance function of the process generating the data

VANNUCCI & CORRADI (*JRSS, Series B*, 1999).

Fields of Application

- Applied Mathematics: Partial/ordinary differential equations solutions; representation of basic operators ...
- Engineering: Signal and image processing (compression, smoothing) ...
- Statistics: Smoothing of noisy data; nonparametric density and regression estimation; stochastic processes representation, time series, functional data ...
- Physics, Biology, Genetics: Turbolence, DNA sequence analysis, magnetic resonance imaging ...
- Music, Human Vision, Computer Graphics ...

Curve Regression

The basic setup is a multivariate linear regression model, with n observations on a q-variate response and p explanatory variables

$$\mathbf{Y} = \mathbf{1}_n \boldsymbol{\alpha}' + \mathbf{XB} + \mathbf{E}$$

$$\mathbf{Y}(n \times q) - \text{ responses, } \mathbf{X}(n \times p) - \text{ predictors}$$

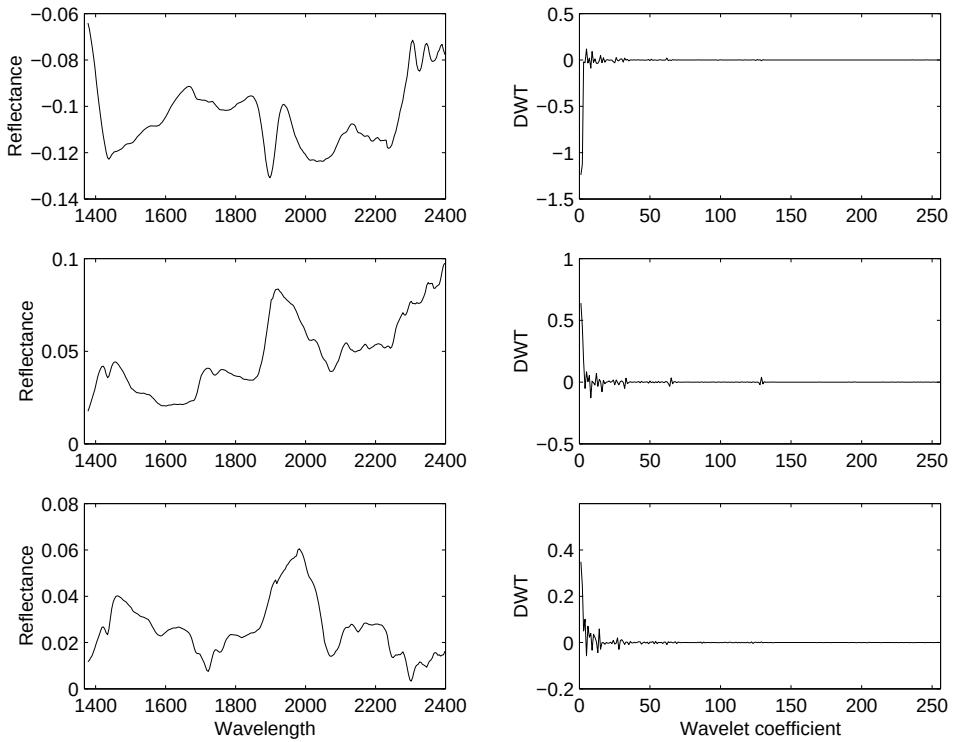
Concern is with functional predictor data, ie the situation where each row of $\mathbf{X}$ is a vector of observations of a curve $x(t)$ at p equally spaced points.

Interest is in situations where p is large$\longrightarrow$ variable selection and/or dimension reduction methods are needed

Wavelet Transformation

A wavelet transform is applied to each row of $\mathbf{X}$

$$\mathbf{Y} = \mathbf{1}_n \alpha' + \mathbf{X}\mathbf{W}'\mathbf{W}\mathbf{B} + \mathbf{E}, \quad \mathbf{W}'\mathbf{W} = \mathbf{I}$$



$$\mathbf{Y} = \mathbf{1}_n \alpha' + \mathbf{D}\tilde{\mathbf{B}} + \mathbf{E}$$

with $\tilde{\mathbf{B}} = \mathbf{W}\mathbf{B}$ and $\mathbf{D} = \mathbf{X}\mathbf{W}'$ a matrix of wavelet coefficients.

Prior Model

If $\mathbf{H}$ is the var-cov matrix of the prior on $\mathbf{B}$ then

$$\tilde{\mathbf{H}} = [\mathbf{WHW}']$$

We compute transformed covariance priors on $\tilde{\mathbf{B}}$ as $\mathbf{WHW}'$ using results of Vannucci and Corradi, JRSSB (1999).

Priors on α and Σ are unchanged
$(\tilde{\mathbf{B}} \rightarrow \mathbf{B}, \tilde{\mathbf{H}} \rightarrow \mathbf{H})$

Selection Prior

Mixture priors to represent negligible coefficients.
A latent binary vector $\boldsymbol{\gamma}$ identifies different models

$$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p), \quad \gamma_j = 0, 1, \quad \gamma_j = 1 \leftrightarrow \mathbf{D}_j \text{ in}$$

Coefficients are drawn from a mixture distribution

$$\mathbf{B} - \mathbf{B}_0 \sim \mathcal{N}(\mathbf{H}\boldsymbol{\gamma}, \boldsymbol{\Sigma})$$

$$\mathbf{B}_{:,j} \sim (1 - \gamma_j)\mathbf{I}_0 + \gamma_j N(0, h_{jj}\boldsymbol{\Sigma})$$

$\pi(\boldsymbol{\gamma})$ as single Bernoulli's or Beta-Binomial or related predictors

$$Prob(\gamma_j = 1) = w_j, \quad w_j = w, \; j = 1, ..., p$$

Metropolis Search

MCMC used to "search" the posterior $g(\gamma) \sim \pi(\gamma|\mathbf{Y}, \mathbf{D})$ looking for good models.

- At each step the algorithm generates γ^{new} from γ^{old} by one of two possible moves:
  Add or Delete a component by choosing at random one component in γ^{old} and changing its value.
  Swap two components by choosing independently at random a 0 and a 1 in γ^{old} and changing both of them.
- The new candidate γ^{new} is accepted with probability

$$\min\{\frac{g(\gamma^{new})}{g(\gamma^{old})}, 1\}$$

Prior Specification

Priors on $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ vague and largely uninformative

$$\boldsymbol{\alpha}' - \boldsymbol{\alpha}_0' \sim \mathcal{N}(h, \boldsymbol{\Sigma}), \quad h \to \infty, \quad \boldsymbol{\Sigma} \sim \mathcal{IW}(\delta, \mathbf{Q})$$

Choices for $\mathbf{H}_{\boldsymbol{\gamma}}$:

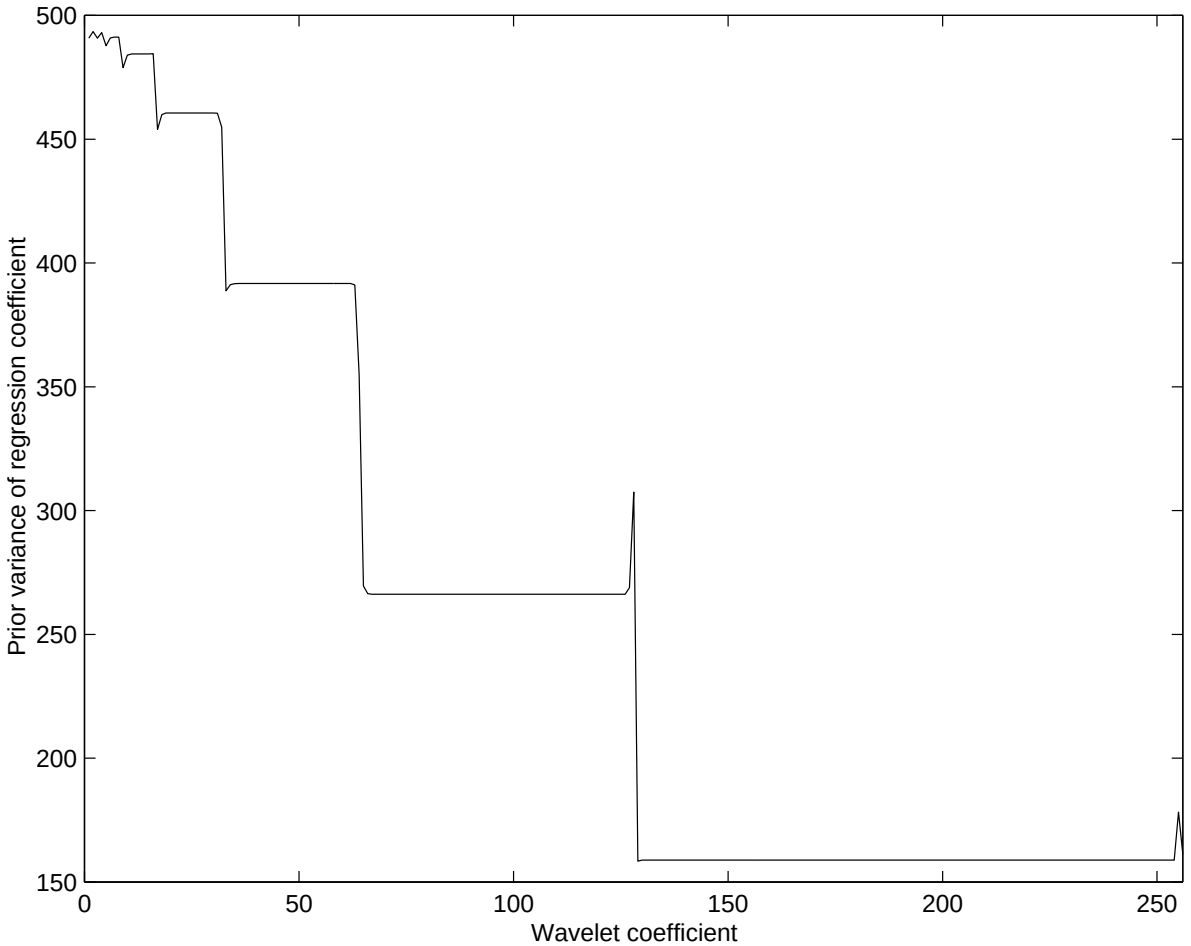
- $\mathbf{H}_{\boldsymbol{\gamma}} = C * [(\mathbf{D}'\mathbf{D})^{-1}]_{\boldsymbol{\gamma}}$
- $\mathbf{H}_{\boldsymbol{\gamma}} = C * [diag(\mathbf{D}'\mathbf{D})^{-1}]_{\boldsymbol{\gamma}}$
- $\mathbf{H} = C\mathbf{I}$
- $\mathbf{H} = AR(\sigma, \rho)$ with EB estimates of hyperparameters,

$$L(\cdot; \mathbf{D}, \mathbf{Y}) = |\mathbf{K}|^{-q/2} |\mathbf{Q}|^{-n/2} |\mathbf{I}_n + \mathbf{K}^{-1}\mathbf{Y}\mathbf{Q}^{-1}\mathbf{Y}'|^{-(\delta+q+n-1)/2}$$

$$\mathbf{K} = \mathbf{I} + \mathbf{D}\mathbf{H}\mathbf{D}'$$

Choice of w:

Diagonal Elements of $\tilde{H} = WHW'$



Posterior Inference

The stochastic search results in a list of visited models $(\gamma^{(0)}, \gamma^{(1)}, \ldots)$ and their corresponding relative posterior probabilities

$$p(\gamma^{(0)}|\mathbf{D}, \mathbf{Y}), p(\gamma^{(1)}|\mathbf{D}, \mathbf{Y}) \ldots$$

- Select variables:
  - in the **"best"** models, i.e. the γ's with highest $p(\gamma|\mathbf{D}, \mathbf{Y})$ or
  - with largest marginal posterior probabilities
- Prediction on future Y^f:
  - via *Bayesian model averaging (BMA)* as posterior weighted average of model predictions
  - via *single model predictions* as LS/Bayes predictions on single best models or LS/Bayes predictions with "threshold" models (eg, "median" model) obtained from estimated marginal probabilities of inclusion.

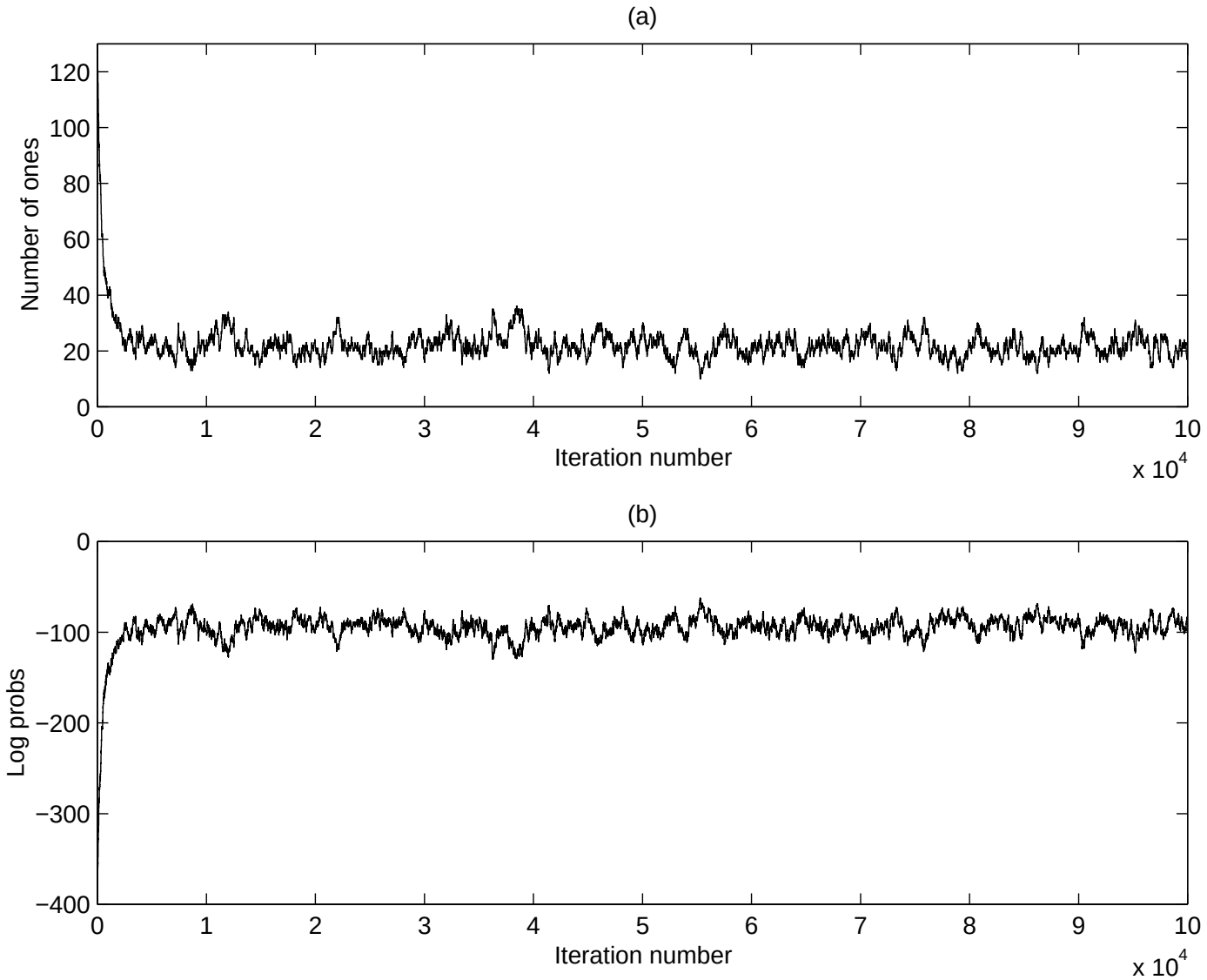
Biscuit Doughs

4 parallel MCMC with 100,000 iterations and about 40,000 successful moves per chain.

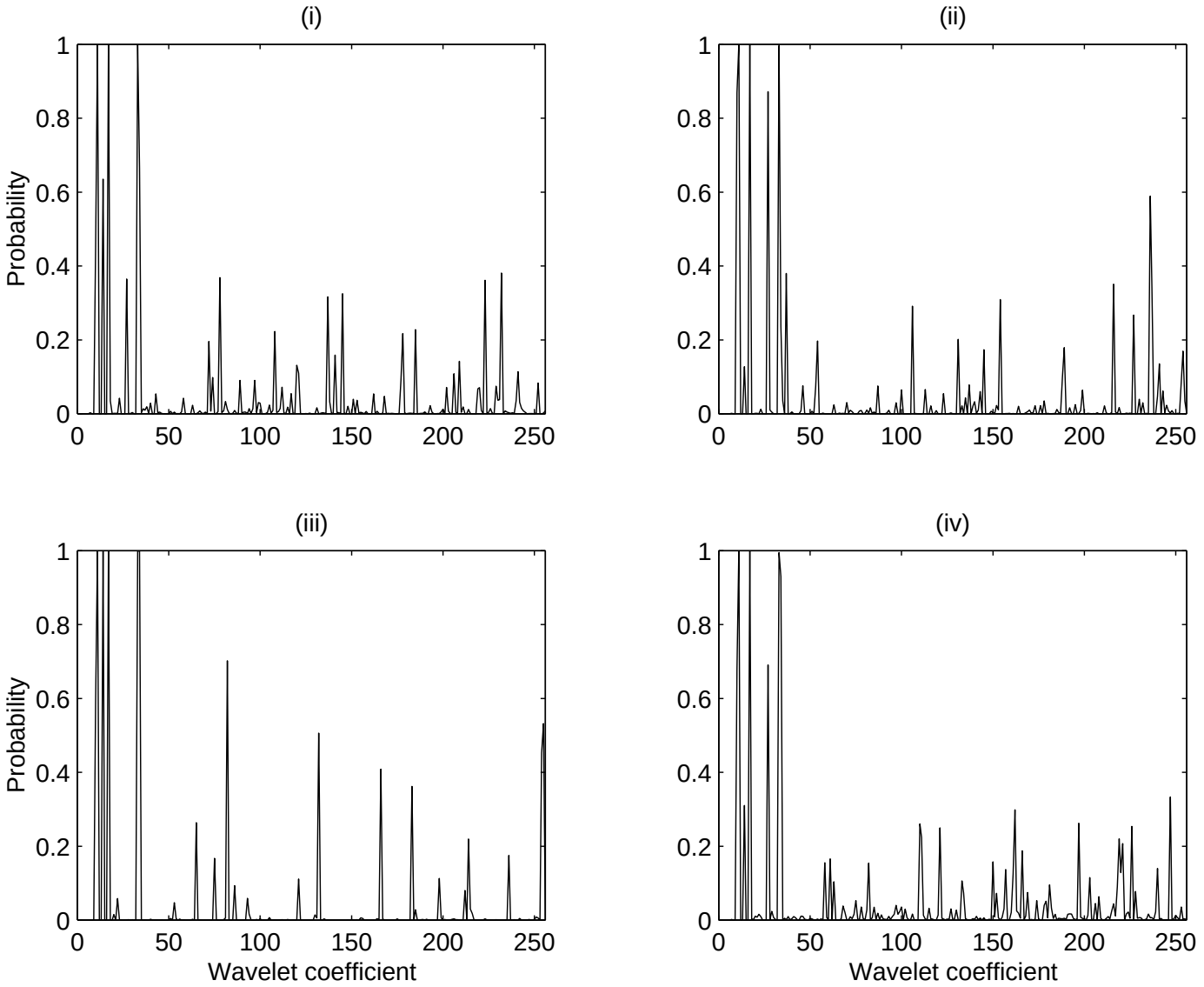
$$H(i,j) = \sigma^2 \rho^{|i-j|}, \sigma^2 = 254, \rho = .32$$

PLS(MSE): .151, .583, .375, .105
PCR(MSE): .160, .614, .388, .106
Stepwise MLR: .044, 1.188, .722, .221
BMA: 500 models and 219 coeffs
.063, .449, .348, .050
Modal (MSE): 10 coeffs ((0%, 0%, 0%, 37%, 6%, 6%, 1%, 2%) from coarsest to finest level)
.059, .466, .351, .047

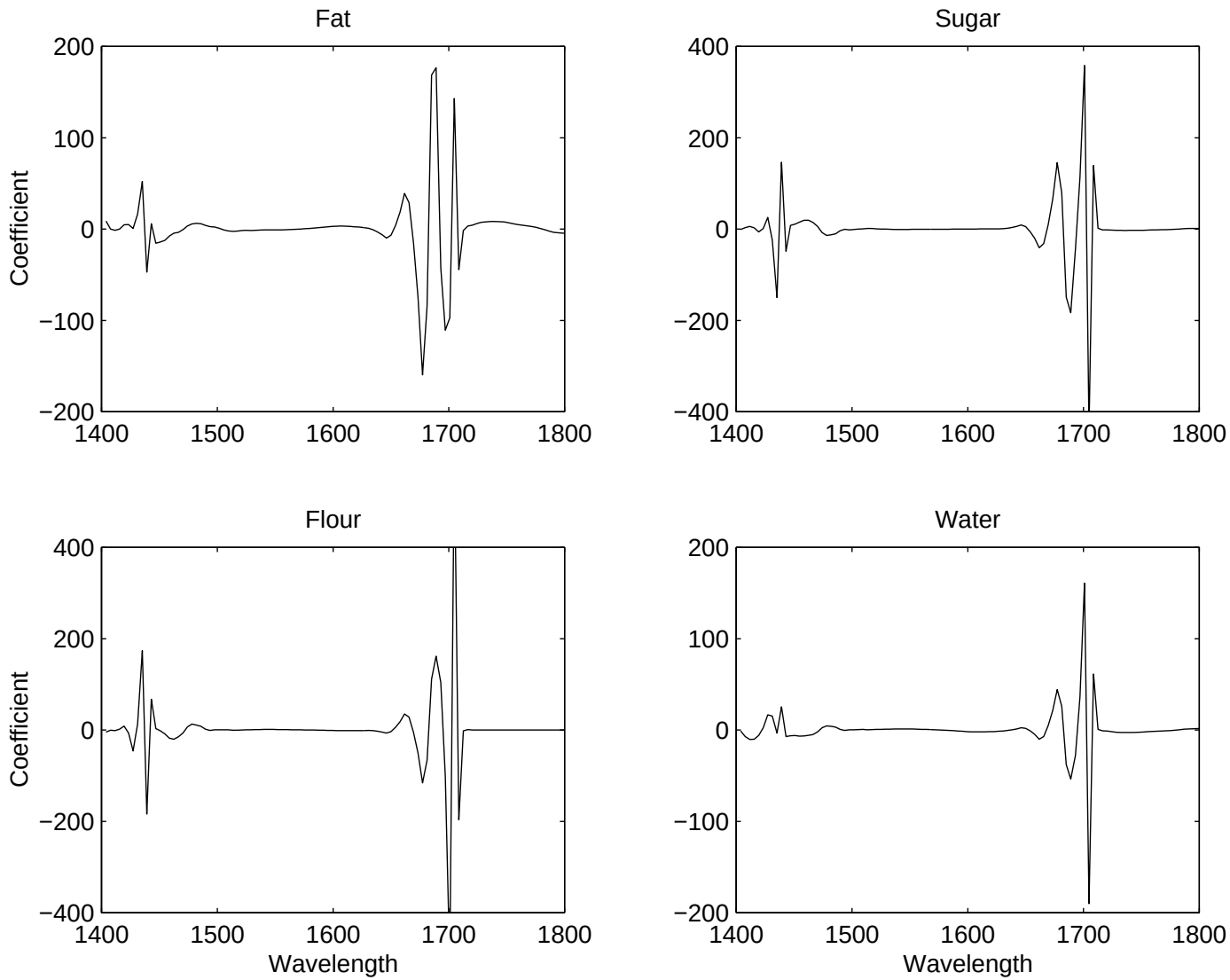
Diagnostic Plots



Marginal Plots



IDWT to Columns of LS Estimate of $\tilde{B}$



Probit Models for Classification

$\mathbf{Z} \rightarrow n \times 1$ categorical response vector (J categories).
$\mathbf{X} \rightarrow n \times p$ predictor matrix. We have functional predictor data.
Probit models use data augmentation and *latent data* to write

$$\mathbf{Y}_i = \boldsymbol{\alpha}' + \mathbf{X}_i'\mathbf{B} + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, n$$

Relationship between the realization z_i and the unobserved $\mathbf{Y}_i$

$$z_i = \begin{cases} 0 & \text{if } y_{i,j} < 0 \text{ for every } j \\ j & \text{if } y_{i,j} = \max\limits_{1 \le k \le J-1} \{y_{i,k}\} \end{cases}$$

... now transform to wavelets, use selection prior on wavelet coefficients and MCMC for inference in a probit model ...

Results for Wheat Data

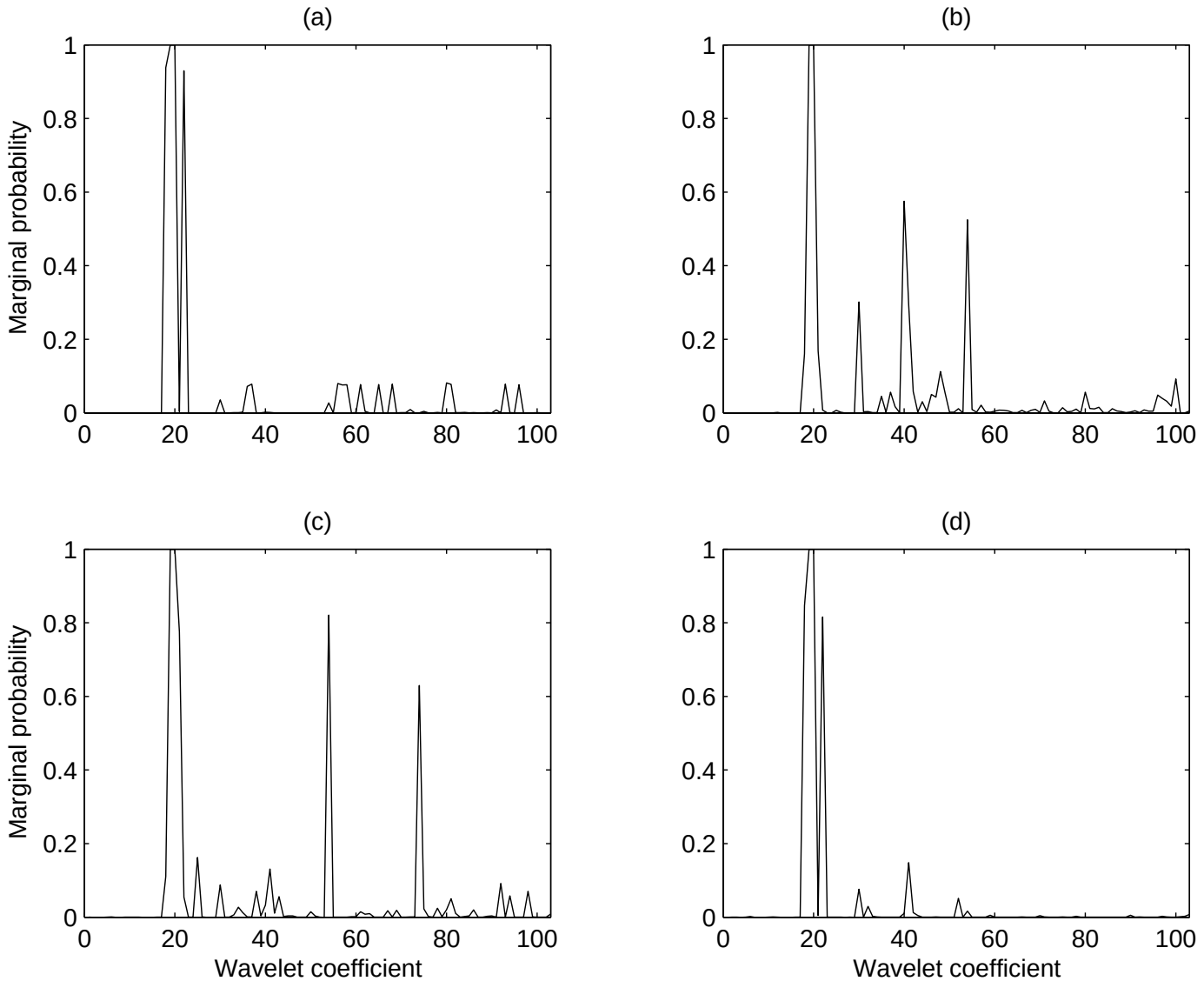
- 3 classes, 94 samples, NIR transmission spectra at p=100 wavelengths from 850 to 1048nm.

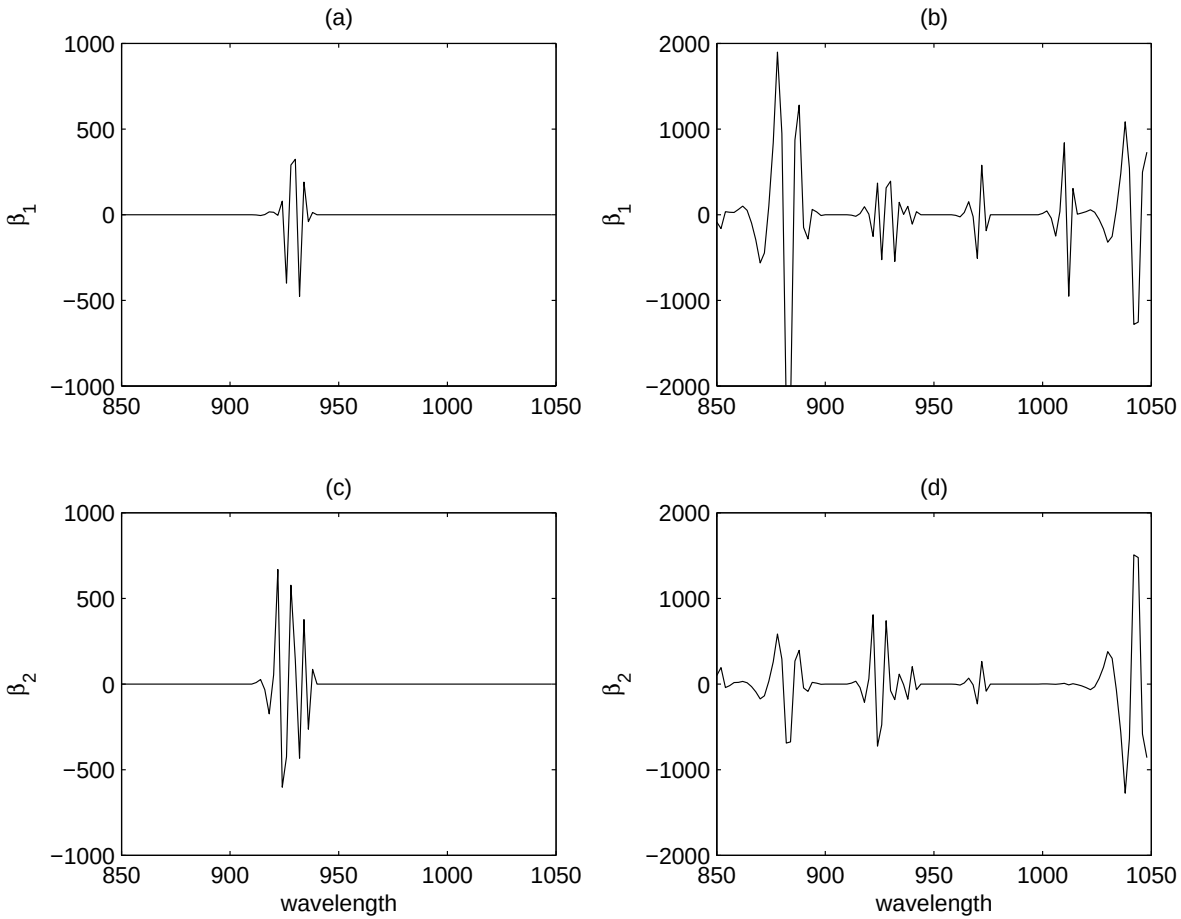
Variety	**1**	**2**	**3**	Total
Training	32	22	17	71
Validation	10	7	6	23

- Fearn *et al.*(2001)
  Bayesian decision approach that balances costs for variables against the loss of misclassification.

Mis-classification Errors

Prediction	Var. Sel./# PC.	Error
"Best" Model	5	3/23
Marginal model	9	3/23
BD(Fearn *et al.*)	6	5/23
BD(Fearn *et al.*)	12	3/23
LDA with PCA	14–18 PCs	4/23
QDA with PCA	8–10 PCs	7/23


(a)
(b)
(c)
(d)
Marginal probability
Wavelet coefficient
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100



Plots (a) and (c) are obtained using 4 wavelet coefficients and plots (b) and (d) using 9 coefficients.

Main References

- Mallat, S.G. (1989). Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Transactions of the American Mathematical Society*, **315(1)**, 69-87.
- Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM.
- Brown, P.J., Fearn, T. and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association,* **96**, 398-408.
- Vannucci, M., Sha, N. and Brown, P.J. (2005). NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection. *Chemometrics and Intelligent Laboratory Systems*, **77**, 139–148.

Code from my Website

- bvsme_wav: Bayesian Variable Selection applied to wavelet coefficients
- Metropolis search
- non-diagonal selection prior
- Bernoulli priors or Beta-Binomial prior
- Predictions by LS and BMA

http://stat.rice.edu/∼marina