NSF-CBMS Regional Research Conference

Wavelet Analysis as a Tool for

Computational and Harmonic Analysis

University of Central Florida

• May 4 – 8, 1998 •

Wavelet Based Statistical Procedures. Three Case Studies

Brani Vidakovic • *Duke University*

May 4, 1998

Plan

- **Wavelet Landmarks**

- **Waveletizing Siguira's Procedure**

- **Are the Standard Deviations of Daubechies' Random Variables Linear in Vanishing Moments?**

- **Some Other "Non-standard" Applications**

Figure 1: Infratec whole grain analyzer from Tecator is the Near Infrared instrument that has been approved by the US Dept. of Commerce NIST. The Infratec is now a standard for the international grain industry.

Wavelet Landmarks

- The amounts of four chemical constituents (**y's**) in grain are of interest to researchers: **fat, sucrose, flour, and water.**

- For each run in the experiment measurements are made by the standard laboratory method (Kjeldahl nitrogen on a ground subsample)

- Near Infrared transmission spectra are obtained [Tecator Infratec] and sampled at 256 points.

- Data [**y's** and associated spectra] are randomly divided on the calibration sample (size 39) and prediction sample (size 32).

- The aim is to predict the content four chemical constituents (expensive and time-consuming measurements) from the spectrum (fast and inexpensive measurements).

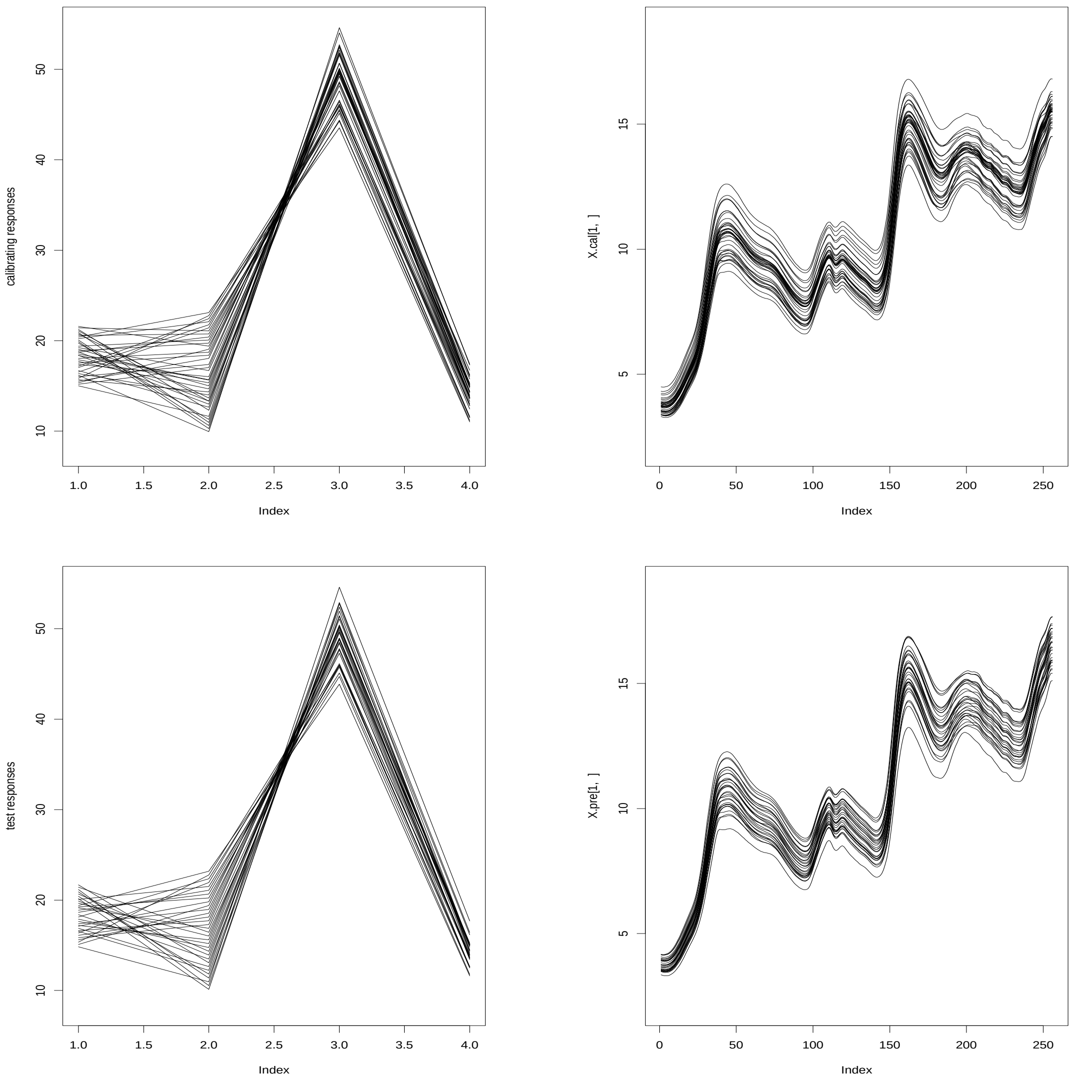

Figure 2: Data

Panels (a) and (c) depict measurements of **fat, sucrose, flour, and water,** in that order (connected with the lines, each signifying one experimental run), for the calibration and prediction data sets. Panels (b) and (d) depict the associated spectral curves.

• We selected one of the $\mathbf{y}$'s [fat content], $\mathbf{y}_1 = (\mathbf{y}_{1,1}, \ldots, \mathbf{y}_{1,39})$ and found the coefficient of corelation between $\mathbf{y}_1$ and vectors $\mathbf{d}_{jk}$ at each **detail** position in the matrix (size 39×256) of wavelet coefficients of the 39 calibration curves.

• The *best* wavelet coefficient vector $\mathbf{d}^*$ (one maximizing the correlation) is selected.

• A linear regression between vectors $\mathbf{y}_1$ and $\mathbf{d}^*$ is found and applied in the prediction data set. The predicted vector was $\hat{\mathbf{y}}_1$.

• Two findings are significant:

(i) The coefficien of correlation between $\mathbf{y}_1$ and $\hat{\mathbf{y}}_1$ was 0.956, and $\Sigma_i |\hat{\mathbf{y}}_{1,i} - \mathbf{y}_{1,i}|^2 / \Sigma \hat{\mathbf{y}}_i^2 = 0.00215$.

(ii) the average of $\mathbf{d}^*$ was moderate to small in magnitude (approximately 2.5% of an average wavelet coefficient).

This is a sharp contrast in what is percieved to be an important wavelet coefficient from an "energy preserving" point of view.

Waveletizing Siguira's Procedure

• Wavelets can sucessfully replace standard orthonormal bases and enhance many of the existing methods.

• Siguira [Osaka Math. J. 14, 253-263, 1962] developed a method for approximate calculation of moments of order stattistics using orthogonal series.
• Legandre polynomials;
• only several coefficients;
• errors of the order 10^{-4}.

• Joshi [JASA, 64, 1617-1624, 1969] improved and generalized Siguira's method.

• Let $\{\psi_i,\ i \in \mathbb{N}\}$ be an orthonormal basis of $\mathbb{L}_2$. Let $f, g \in \mathbb{L}_2$.

$$f(x) = \Sigma_i d_i \psi_i(x) \quad \text{and} \quad g(x) = \Sigma_i e_i \psi_i(x).$$

$$\begin{aligned} \langle f, g \rangle_{\mathbb{L}_2} &= \langle \underset{\sim}{d}, \underset{\sim}{e} \rangle_{\ell_2} \\ |\langle f, g \rangle_{\mathbb{L}_2} - \langle \underset{\sim}{d}, \underset{\sim}{e} \star \mathbf{1}_K \rangle_{\ell_2}| \ &\leq \ ||\underset{\sim}{d} \star \mathbf{1}_R|| \, ||\underset{\sim}{e} \star \mathbf{1}_R||. \end{aligned}$$

• $\underset{\sim}{a} \star \underset{\sim}{b} = \{a_1 b_1, a_2 b_2, \ldots\}$ is the Hadamard product of vectors $\underset{\sim}{a} = \{a_1, a_2, \ldots\}$ and $\underset{\sim}{b} = \{b_1, b_2, \ldots\}$.
• $\mathbf{1}_A$ is an indicator of a set $A \subset \mathbb{N}$, $(\mathbf{1}_A)_i = \mathbf{1}(i \in A),\ i \in \mathbb{N}$.
• K ("keep") and R ("reject") - an arbitrary partition of integers $\mathbb{N}$ such that $K \cup R = \mathbb{N}$, and $K \cap R = \emptyset$.

• Let $X_1, X_2, \ldots, X_n$ be a sample from a population with an absolutely continuous distribution F, with finite second moment. Let $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ be the ordered sample (order statistic).

• The distribution of $X_{r:n}$ is given by

$$F_{r:n}(x) = \mathbf{I}_{F(x)}(r, n-r+1),$$

where $\mathbf{I}_p(a,b) = \frac{1}{B(a,b)} \int_0^p t^{a-1}(1-t)^{b-1} dt, \;\; a > 0, b > 0,$ $0 \leq p \leq 1$, is the *incomplete Beta* function.

Let $u = F(x)$ and $x(u) = F^{-1}(u)$.

• If F has a finite second moment,

$$x(u) \in \mathbb{L}_2([0,1]).$$

$\int_0^1 x^2(u) du = \int_R x^2 \frac{du}{dx} dx = \int_R x^2 dF(x) < \infty.$

• Let $\{\phi_{00}, \psi_{jk}, j \geq 0; \; 0 \leq k \leq 2^j - 1\}$ be periodized wavelet orthonormal basis on $[0,1]$.

• Let

$$\begin{aligned} d_{jk} &= \langle x(u), \psi_{jk}(u) \rangle \\ e_{jk} &= \langle \frac{u^{r-1}(1-u)^{n-r}}{B(r, n-r+1)}, \psi_{jk}(u) \rangle \end{aligned}$$

be the wavelet coefficients of $x(u)$ and $\frac{u^{r-1}(1-u)^{n-r}}{B(r,n-r+1)}$.

• For any 'keep" subset K of indices,

$$|EX_{r:n} - \langle \underset{\sim}{d}, \underset{\sim}{e} \star \mathbf{1}_K \rangle| \leq ||\underset{\sim}{d} \star \mathbf{1}_R|| \; ||\underset{\sim}{e} \star \mathbf{1}_R||,$$

where $R = \mathbb{N} \backslash K$.

Example: Expectations of order statistics from the standard normal distribution.

- (< 20) Teichroew [Ann. Math. Stat. 27, 410-426, 1956]
- (2 – 50) Tietjen, Kahaner and Beckman [Select. Tables Math. Stat, Vol 5, 1977]
- Exact calculations are possible for $n \leq 5$. For example,

$$EX_{2:5} = \frac{30}{\pi\sqrt{\pi}} \arctan(\sqrt{2}) - \frac{10}{\sqrt{\pi}}.$$

- The wavelet based approximation of $EX_{2:5}$ is

```
> a_ExpX(n=5, r=2, N=2^19)
[1] -0.4950189704756
```

with an error of order 10^{-11}.

For example,

```
>  a <- ExpX(n=100, r=40, N=2^19)
[1] -0.2657229858
>  a <- ExpX(n=10000, r=4000,
+       N=2^19)
[1] -0.2534710049
>  a <- ExpX(n=1000000000,
+      r=400000000,  N=2^19)
[1] -0.2533471032
```

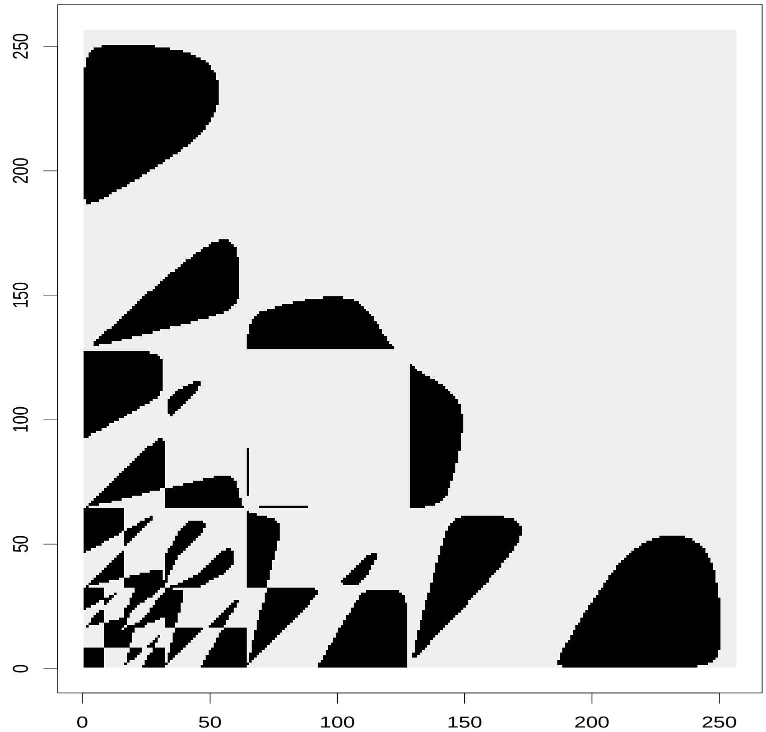

Figure 3: Black are “keep”-coefficients in the product matrix

$$EX_{1:4}X_{4:4} = -\frac{3}{\pi}.$$

$$N = 256, \quad \text{Error} \approx 10^{-5}. \quad \text{“Keep”} > 10^{-8}.$$

- $|EX_{r:n}X_{s:n} - \langle \underset{\sim}{d}, \underset{\sim}{e} \star \mathbf{1}_K \rangle| \leq ||\underset{\sim}{d} \star \mathbf{1}_R|| \; ||\underset{\sim}{e} \star \mathbf{1}_R||,$
- $\underset{\sim}{d} = \{d_{j;k,l} = d_{jk} \cdot d_{jl};\ d_{jk} = \langle x(u), \psi_{jk}(u) \rangle \}$
- $\underset{\sim}{e} = \{e_{j;k,l} = \int_0^1 \int_0^1 g(u,v) \psi_{j;k,l}(u,v) du dv \},$
- $g(u,v) = \frac{1}{2B(r,s-r,n-s+1)} (\min(u,v))^{r-1} |u-v|^{n-r+1} (1 - \max(u,v))^{n-s};\ 0 \leq u \neq v \leq 1,$

where $B(a_1, a_2, \ldots, a_n) = \dfrac{\prod_{i=1}^n \Gamma(a_i)}{\Gamma(\sum_{i=1}^n a_i)}$.

Are the standard deviations of Daubechies' random variables linear in vanishing moments?

• Scaling function as a probability density?

Except in the Haar case, a scaling function can not be a density, even though $\int \phi(x)dx = 1$.

• [Janssen, IEEE Trans. on Inf. Th, 38, 884-885, 1993]

Let ϕ be a scaling function generating an orthogonal multiresolution analysis. Except for the Haar wavelet, ϕ can not be a probability density.

Assume that ϕ is a probability density. Then all h_n in $\underset{\sim}{h}$ are non-negative, because

$$h_n = \int \sqrt{2}\phi(2x-n)\,\phi(x)\,dx = \sqrt{2}E\phi(2X-n) \geq 0.$$

$$\begin{aligned} 1 &= m_0(0) = \frac{1}{\sqrt{2}}\Sigma h_n \text{ and} \\ 0 &= m_0(\pi) = \frac{1}{\sqrt{2}}\Sigma(-1)^n h_n, \end{aligned}$$

$$\Longrightarrow \qquad \Sigma_{n \text{ even}} h_n = \Sigma_{n \text{ odd}} h_n = \frac{1}{\sqrt{2}}.$$

Since $\Sigma_n h_n^2 = 1$ and $h_n \geq 0$, the sums $\Sigma_{n \text{ even}}$ and $\Sigma_{n \text{ odd}}$ must contain a single element each.

Walter and Shen's Construction

Non-negative Scaling Functions [Manuscript, November 1997]
$\rho_r(x) = \Sigma_n r^{|n|}\phi(x-n)$.
$\tilde{\rho}_r(x) = \frac{(1+r^2)\phi(x) - r(\phi(x+1)+\phi(x-1))}{2\pi(1-r^2)}$
• $\{2^{j/2}\rho_r(2^j x - k),\ k \in \mathbb{Z}\}, \{2^{j/2}\tilde{\rho}_r(2^j x - k),\ k \in \mathbb{Z}\}$
are primary and dual bases of V_j and $\tilde{V}_j = V_j$.

$$\frac{1-r}{1+r}\rho_r(x)$$

is a probability density, for $r \geq r_0, \quad 0 < r_0 < 1$. (Such r_0 is always possible to find since $\Sigma_k \phi(x-k) = 1$)

Example: DAUB #2, $r_0 \approx 0.2579$.
$X \sim 0.3/1.7\rho_{0.7}(x)$,
$E(X) = 0.629642$ and Var$(X) = 15.13675$.

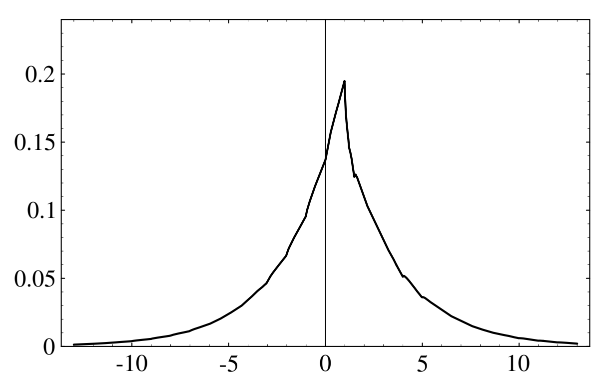

Figure 4: Densitiy $0.3/1.7 \cdot \rho_{0.7}(x)$.

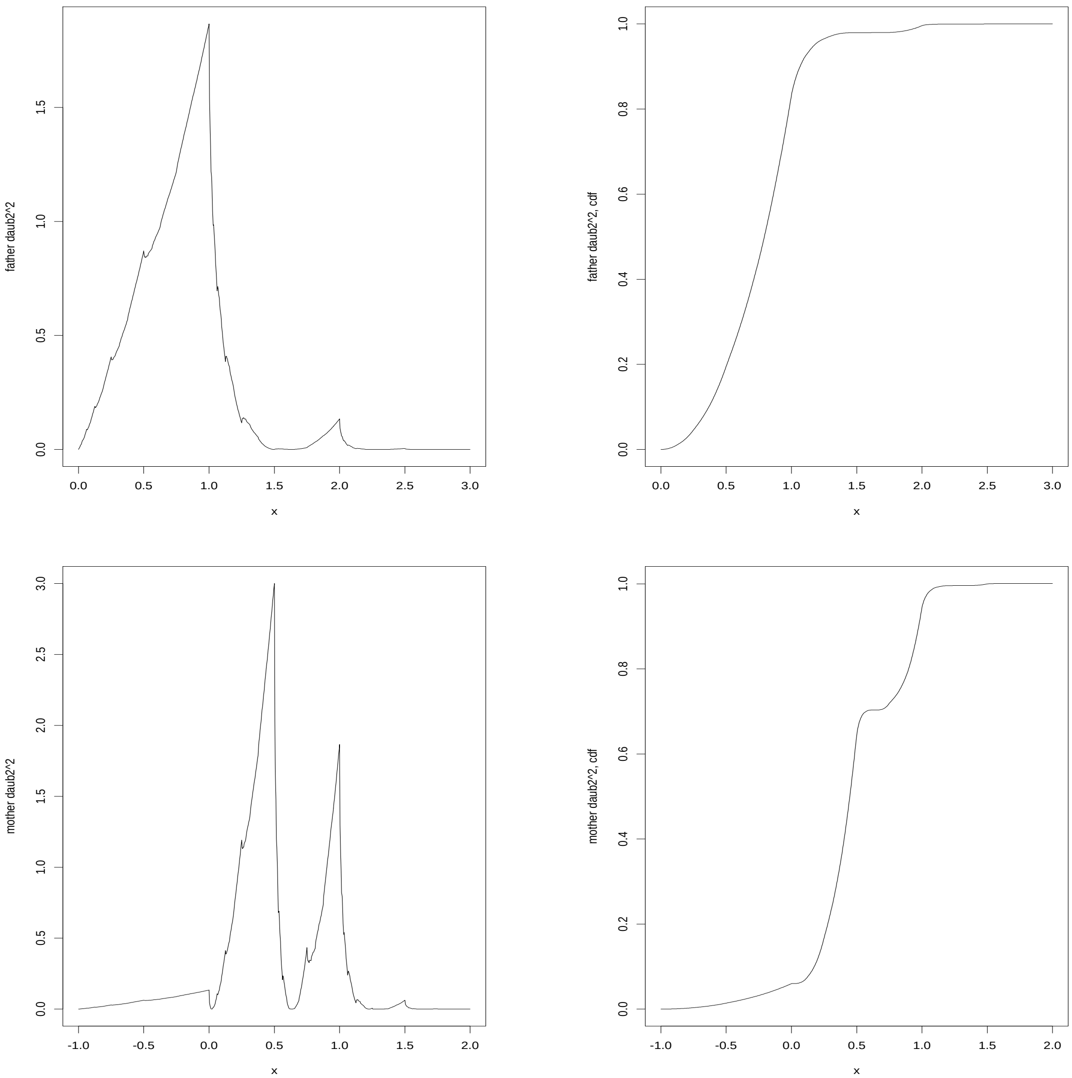

Figure 5: Densities and corresponding cumulative distribution functions for squares of DAUB2 scaling and wavelet functions.

$$\bullet \int \phi^2(x)\,dx = \int \psi^2(x)\,dx = 1.$$

• The mean $\xi_{1,0} = \int x\psi^2(x)dx$ is at the center of support of $\psi(x)$. ($\xi_{1,0} = 0.5$, if the support is $[1-N, N]$.) [L. Villemoes; Ma, Strang and Vidakovic; Shann and Yan]

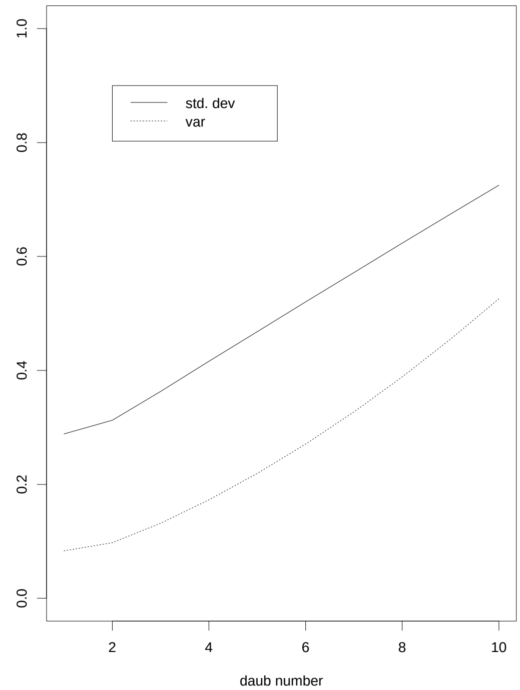

Figure 6: Variances and standard deviations of Daub#n random variables

	MEAN $\mu_{1,0}$	VARIANCE
DAUB1	0.5	0.083333
DAUB2	0.770948	0.097718
DAUB3	1.022422	0.132056
DAUB4	1.266408	0.172921
DAUB5	1.506244	0.219200
DAUB6	1.743334	0.270688
DAUB7	1.978412	0.327253
DAUB8	2.211921	0.388751
DAUB9	2.444157	0.455041
DAUB10	2.675332	0.525993

- [?] $X_n \sim \phi^2$, n vanishing moments $\Longrightarrow \sigma_{X_n} = \alpha n + \beta$.
- Residuals $< 10^{-5}$.

- Statistical Shape Descriptors.

- Wavelets as a Calculational Tool.

- Wavelets as Building Blocks for Statistical Mixture Modeling.

Some Other Non-standard Applications in Statistics

- Bayesian Models in the Wavelet Domain and Induced Wavelet Shrinkage.

- Random Functions and Densities via Wavelets.

- Wavelet Priors. Polya Trees and Wavelets.

- Wavelet Tools in Time Series Analysis.

- Wavelet Based Discrimination Rules.