SELF-ORGANIZATION AS A PROCESSING OF HIDDEN CAUSAL INFORMATION

Miloš Milovanović

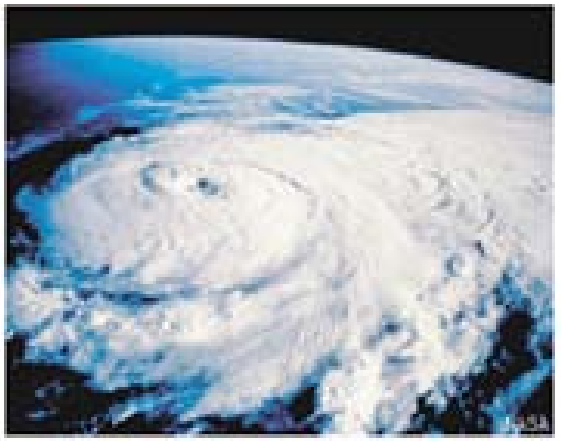
Mathematical Institute SASA

milosm@mi.sanu.ac.rs

SELF-ORGANIZATION

- Emerging structures or organized behavior without the external influence
- Second law of thermodynamics (entropy)
- Ilya Prigogine (Nobel prize in 1950.)
- Living organisms, brain – neural networks, flocks of birds, shoals of fish, swarms of bees, herds, social networks, nations, religious groups

EXAMPLES OF SELF-ORGANIZATION

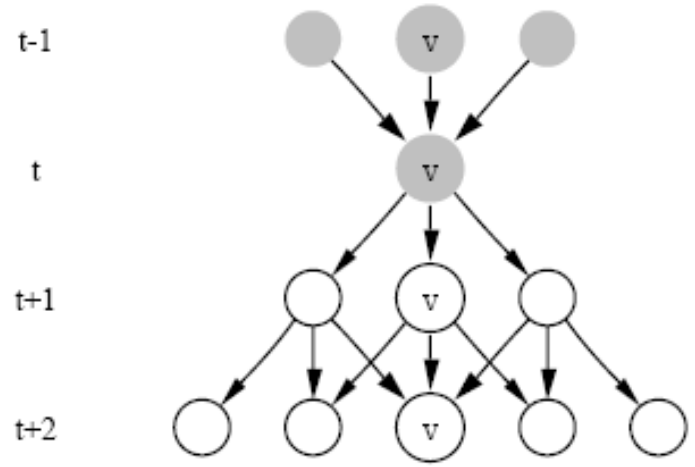


MODEL FOR SELF-ORGANIZATION

- The information flow through fixed graph
- Light cones
- Local causal state
- Local complexity

$$C(v,t) = H(S(v,t))$$

- Markov field of local causal states



STATISTICAL COMPLEXITY

- Periodical and random – complex
- Entropy – rough distinction of elements
- Internal computability of a process
- Statistical complexity – minimal information required for optimal prediction $C = H(S)$
- Self-organization – increase of statistical complexity

PRIMARY OF WAVELETS

- Signal space $L^2(R) = \{ f : R \to C \mid \int_{-\infty}^{+\infty} | f(x) |^2 \, dx < \infty \}$
- Signal energy $\| f \|^2 = \int_{-\infty}^{+\infty} | f(x) |^2 \, dx$
- Scalar product $< f, g >= \int_{-\infty}^{+\infty} f(x) g^*(x) dx$
- Orthonormal wavelets

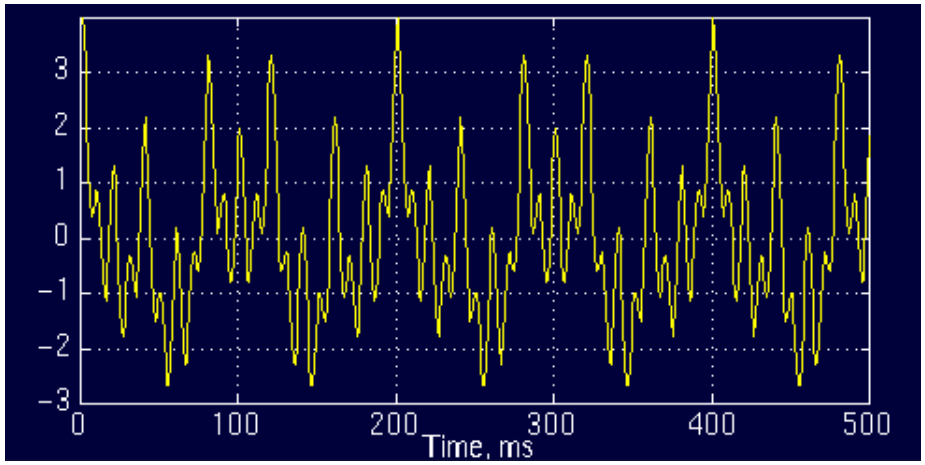
  $\{ \psi_{j,k}(x) = 2^{\frac{j}{2}} \psi(2^j x - k) \mid j,k \in Z \}$ is o.n.b. of $L^2(R)$
- Atomic decomposition of a signal

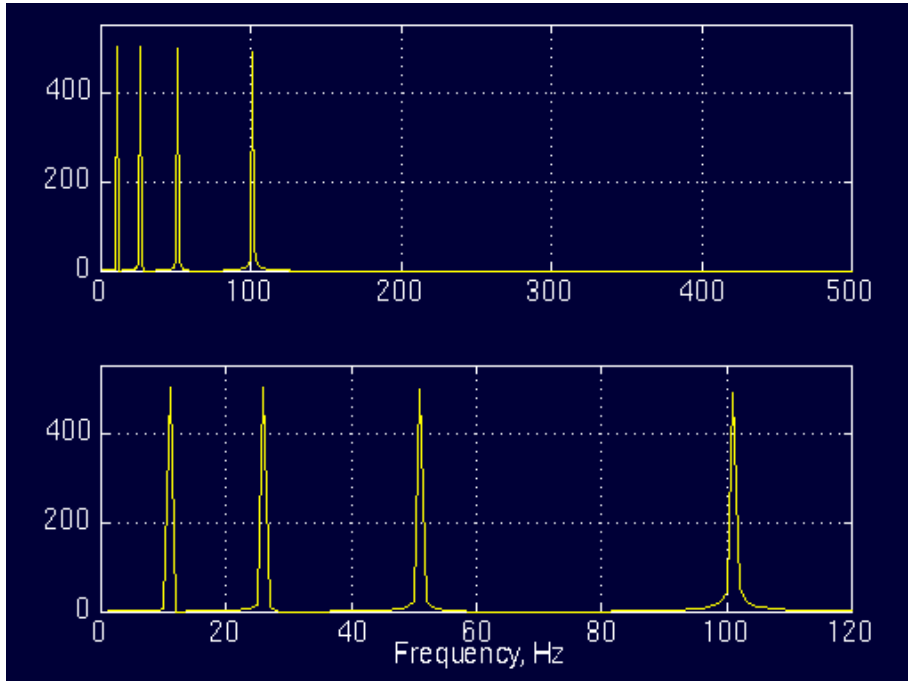
$$f = \sum_{j=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} D_j[k] \psi_{j,k} \qquad D_j[k] =< f, \psi_{j,k} >$$

FOURIER TRANSFORMATION

- F – transform
- Periodic signals

$$\hat{f}(\xi) = \int_{-\infty}^{+\infty} f(x) e^{-ix\xi} dx$$

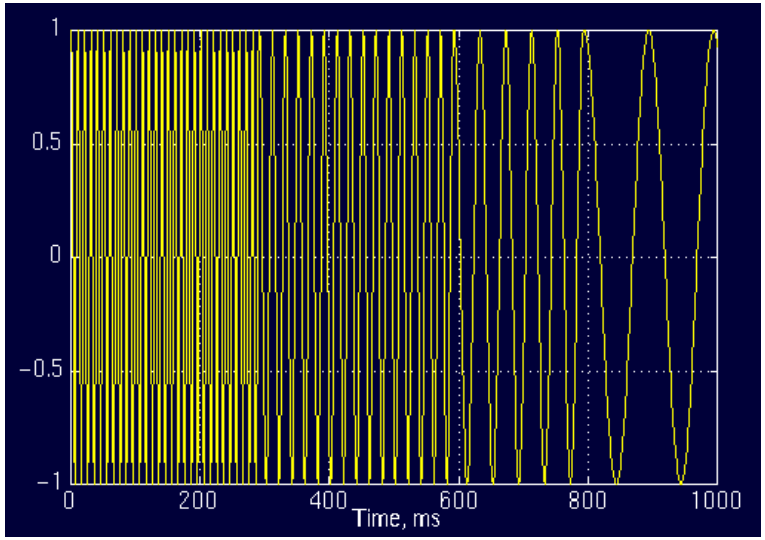


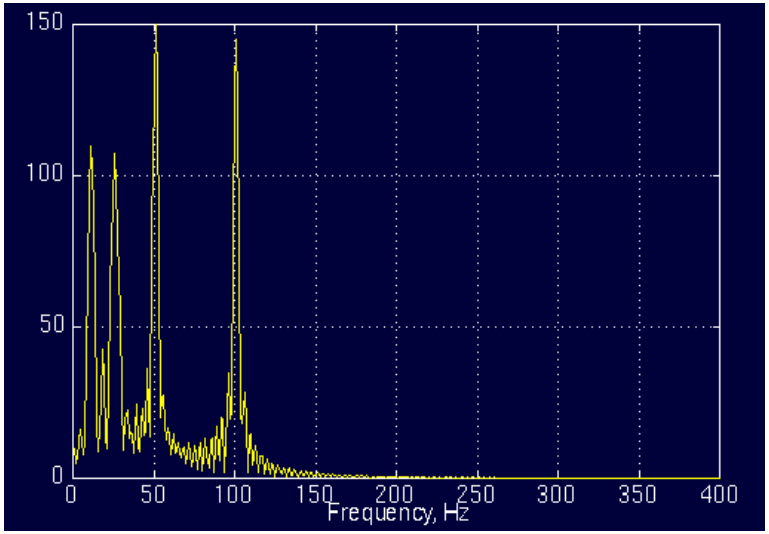


DEFECTS OF F-TRANSFORM

- Non-periodic signals
- Energy limitation
- Uncertainty principle

$$\Delta x \Delta \xi \geq \frac{1}{4\pi}$$





WAVELET TRANSFORMATION

- W-transform

$$D_j[k] =< f, \psi_{j,k} >= \int_{-\infty}^{+\infty} f(x)\psi_{j,k}(x)dx$$

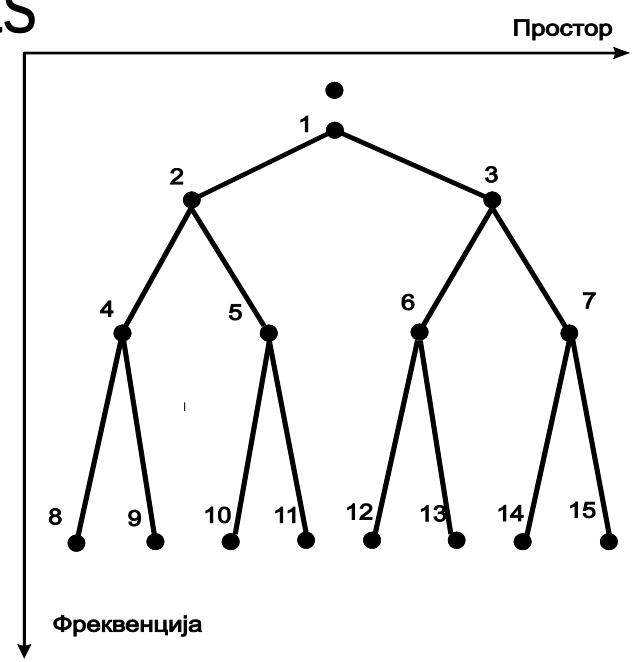
$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$$

- Atomic decomposition of a signal

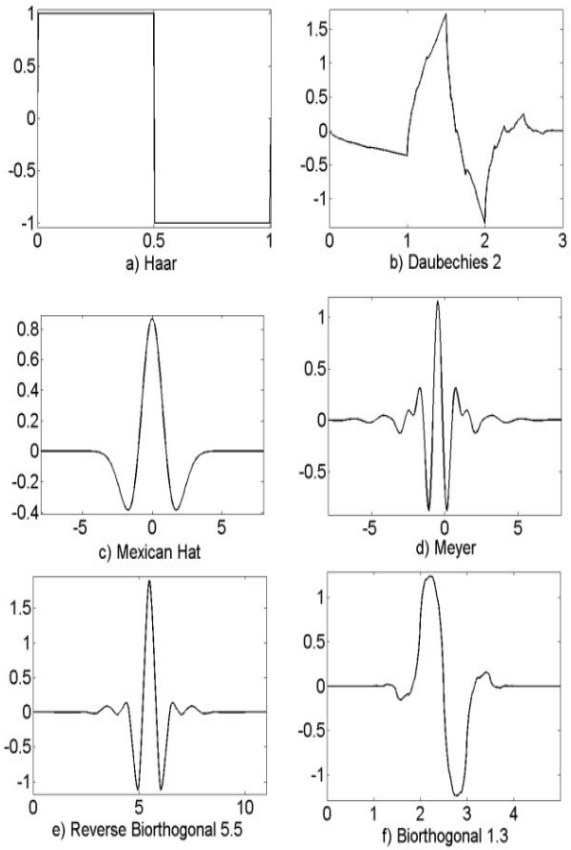
$$f = \sum_{j=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} D_j[k]\psi_{j,k}$$

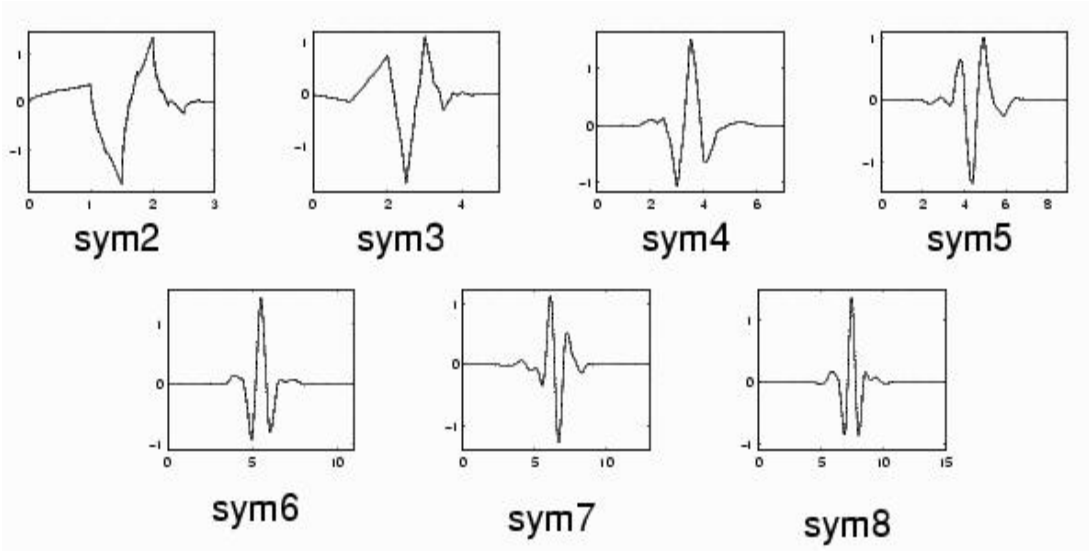
- Pyramid of detail coefficients

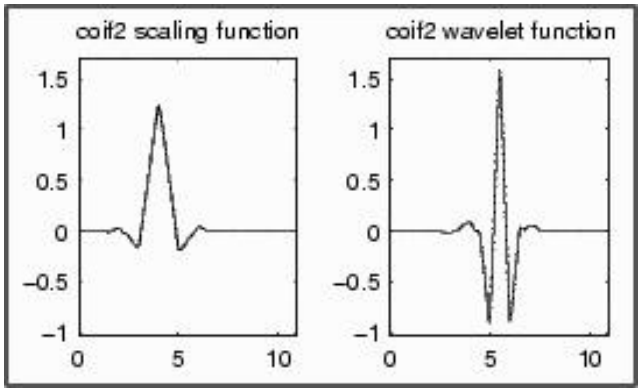
(pseudo-numeration)



SOME SIGNIFICANT WAVELETS







OPTIMAL WAVELET

- Representation of a signal in the basis

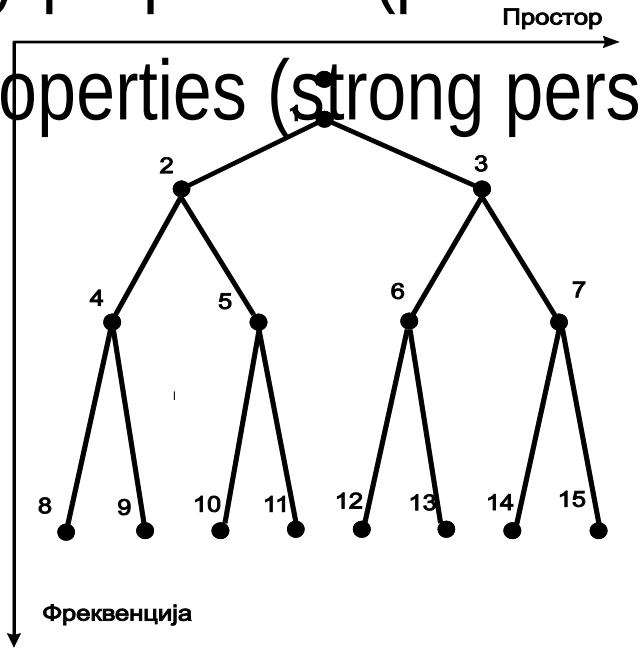
$$1,0,0,0,0\ldots$$

$$3,17,1,-5,33\ldots$$

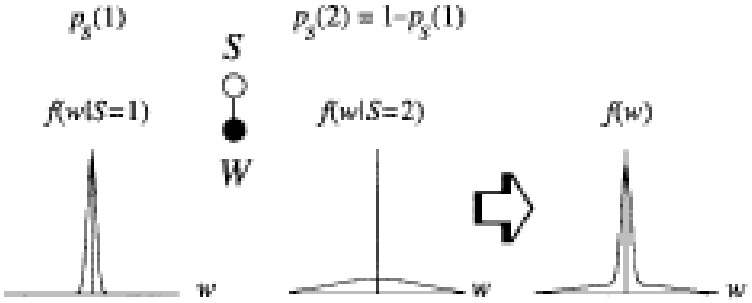
- Complexity of the system – minimal information required for optimal prediction
- Optimal representation – maximal complexity

PROPERTIES OF W-TRANSFORM

- Primary properties (multiresolution, locality, singularity detection, approximative decorrelation, energy compactition)
- Secondary properties (persistence, clustering)
- Tertiary properties (strong persistence, exp. decay)



COEFFICIENT DISTRIBUTION



- Mixed Gaussian model

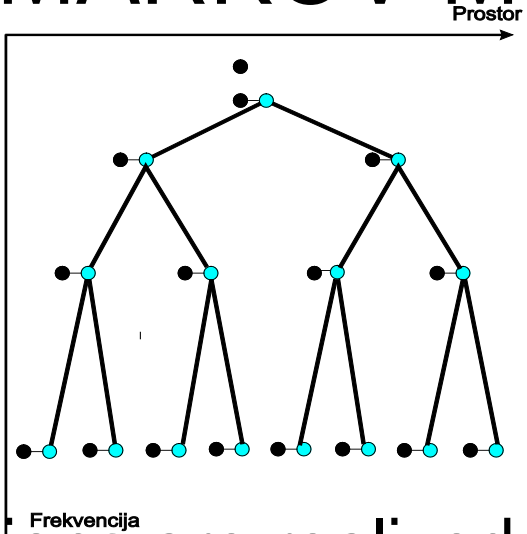
$$P(D[i]=d)=\sum_m P(S[i]=m)g(d,\mu_i^m,\sigma_i^m)$$

- Hidden variable $S[i]:\begin{pmatrix}\alpha & \omega\\ p_i^\alpha & p_i^\omega\end{pmatrix}$

$$P(D[i]\,|\,S[i]=m)=g(\cdot,\mu_i^m,\sigma_i^m)$$

$$p_i^\alpha << p_i^\omega$$
$$\sigma_i^\omega << \sigma_i^\alpha$$

HIDDEN MARKOV MODEL



- Coefficient correlations are realized locally through hidden states
- Model parameters $\varepsilon_i^{mn} = P(S[i] = m \mid S\rho[i] = n)$

$$\vartheta = (p_1^m, \varepsilon_i^{mn}, \mu_i^m, \sigma_i^m \mid i = 1...I; m, n = \alpha...\omega)$$

EM-ALGORITHM

- Tying – coefficients on a common scale are equally distributed
- Iterative procedure estimating both model parameters and hidden state probabilities
- Initialization : $\vartheta^0, l = 0$
- E-step : $P(S \mid d, \vartheta^l)$
- M-step : $\vartheta^{l+1} = \arg\max < \log P(D, S \mid \vartheta) \mid D, \vartheta^l >_S$
- Repeat E-step for $l+ = 1$ until convergence

BAUM-WELCH ALGORITHM

the forward-backward algorithm in the HMM literature [18] and as the upward–downward or inward–outward algorithm in the artificial intelligence literature [20], [27], [32].) We will then develop the EM steps for multiple trees. We will finish by incorporating into the EM steps the notion of tying within trees from Section IV-C.

We first focus on processing a single size-P wavelet tree containing observed wavelet coefficients $\mathbf{w} = [w_1 w_2 \cdots w_P]$ having hidden states $\mathbf{S} = [S_1 S_2 \cdots S_P]$ that take on values $m = 1, \cdots, M$. The primary task of the E step is to calculate the hidden state probabilities $p(S_i = m|\mathbf{w}, \boldsymbol{\theta})$ and $p(S_i = m, S_{\rho(i)} = n|\mathbf{w}, \boldsymbol{\theta})$. To obtain these probabilities, we introduce a number of intermediate variables.

A. Setup

We now introduce some notation for trees of observed wavelet coefficients. Similar in structure to the trees of Fig. 5, these trees are formed by linking the wavelet coefficients rather than the hidden states. We define $\mathcal{T}_i$ to be the subtree of observed wavelet coefficients with root at node i so that the subtree $\mathcal{T}_i$ contains coefficient w_i and all of its descendants. Now, if $\mathcal{T}_j$ is a subtree of $\mathcal{T}_i$ (i.e., W_j and all its descendants are members of $\mathcal{T}_i$), then we define $\mathcal{T}_{i \setminus j}$ to be the set of wavelet coefficients obtained by removing the subtree $\mathcal{T}_j$ from $\mathcal{T}_i$. Without loss of generality, we order $\mathbf{W}$ so that w_1 is at the root of the entire tree. Thus, $\mathcal{T}_1$ is the entire tree of observed wavelet coefficients (a tree-structured version of the vector $\mathbf{W}$). In our probability expressions, we will interchange $\mathcal{T}_1$ and $\mathbf{W}$ when convenient.

For each subtree $\mathcal{T}_i$, we define the conditional likelihoods

$$\beta_i(m) \equiv f(\mathcal{T}_i|S_i = m, \boldsymbol{\theta}) \tag{10}$$

$$\beta_{i,\rho(i)}(m) \equiv f(\mathcal{T}_i|S_{\rho(i)} = m, \boldsymbol{\theta}) \tag{11}$$

$$\beta_{\rho(i)\setminus i}(m) \equiv f(\mathcal{T}_{\rho(i)\setminus i}|S_{\rho(i)} = m, \boldsymbol{\theta}) \tag{12}$$

and the joint probability functions

$$\alpha_i(m) \equiv p(S_i = m, \mathcal{T}_{1\setminus i}|\boldsymbol{\theta}) \tag{13}$$

with S_i taking discrete values and the coefficients in $\mathcal{T}_{1\setminus i}$ taking continuous values.

Based on the HMT properties from Section III-B, the trees $\mathcal{T}_i$ and $\mathcal{T}_{1\setminus i}$ are independent given the state variable S_i. This fact, along with the chain rule of probability calculus, leads to the desired state probabilities in terms of the α's and β's. First, we obtain

$$p(S_i = m, \mathcal{T}_1|\boldsymbol{\theta}) = \alpha_i(m)\beta_i(m) \tag{14}$$

and

$$p(S_i = m, S_{\rho(i)} = n, \mathcal{T}_1|\boldsymbol{\theta}) = \beta_i(m)\epsilon^{mn}_{i,\rho(i)}\alpha_{\rho(i)}(n)\beta_{\rho(i)\setminus i}(n). \tag{15}$$

The likelihood of $\mathbf{W}$ is then

$$f(\mathbf{w}|\boldsymbol{\theta}) = f(\mathcal{T}_1|\boldsymbol{\theta}) = \sum_{m=1}^{M} p(S_i = m, \mathcal{T}_1|\boldsymbol{\theta}) = \sum_{m=1}^{M} \beta_i(m)\alpha_i(m). \tag{16}$$

Bayes rule applied to (14)–(16) leads to the desired conditional probabilities

$$p(S_i = m|\mathbf{w}, \boldsymbol{\theta}) = \frac{\alpha_i(m)\beta_i(m)}{\sum_{n=1}^{M}\alpha_i(n)\beta_i(n)} \tag{17}$$

and

$$p(S_i = m, S_{\rho(i)} = n|\mathbf{w}, \boldsymbol{\theta}) = \frac{\beta_i(m)\epsilon^{mn}_{i,\rho(i)}\alpha_{\rho(i)}(n)\beta_{\rho(i)\setminus i}(n)}{\sum_{n=1}^{M}\alpha_i(n)\beta_i(n)}. \tag{18}$$

B. E Step for a Single Wavelet Tree (Upward–Downward Algorithm)

All state variables within our HMT model are interdependent; in determining probabilities for the state variables, we must propagate state information throughout the tree. The upward–downward algorithm is an efficient method for propagating this information. The up step calculates the β's by transmitting information about the fine-scale wavelet coefficients to the states of the coarse-scale wavelet coefficients; the down step calculates the α's by propagating information about the coarse-scale wavelet coefficients down to the states of the fine-scale wavelet coefficients. Combining the information from the α's and β's via (17) and (18), we obtain conditional pmfs for the state of each wavelet coefficient in the tree.

For our derivation, we will focus on models with mixing components that are Gaussian with density

$$g(w; \mu, \sigma^2) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(w-\mu)^2}{2\sigma^2}\right]. \tag{19}$$

More general densities can also be treated. Recall that we assign to each node i in the tree a scale $j(i) \in \{1, \cdots, L\}$ with $J = 1$ the finest scale and $J = L$ the coarsest scale. In addition, recall that $\rho(i)$ is the parent of node i and $\mathcal{C}(i)$ the set of children to node i.

Up Step:

Initialize: For all state variables S_i at the finest scale $J = 1$, calculate for $m = 1, \cdots, M$:

$$\beta_i(m) = g(w_i; \mu_{i,m}, \sigma^2_{i,m}). \tag{20}$$

1) For all state variables S_i at scale J, compute for $m = 1, \cdots, M$

$$\beta_{i,\rho(i)}(m) = \sum_{n=1}^{M} \epsilon^{mn}_{i,\rho(i)}\beta_i(n) \tag{21}$$

$$\beta_{\rho(i)}(m) = g\left(w_{\rho(i)}; \mu_{\rho(i),m}, \sigma^2_{\rho(i),m}\right) \times \prod_{i \in \mathcal{C}(\rho(i))} \beta_{i,\rho(i)}(m) \tag{22}$$

$$\beta_{\rho(i)\setminus i}(m) = \frac{\beta_{\rho(i)}(m)}{\beta_{i,\rho(i)}(m)}. \tag{23}$$

2) Set $J = J + 1$ (move up the tree one scale).
3) If $J = L$, then stop; else return to step 1.

Down Step:

Initialize: For state variable S_1 at the coarsest scale $J = L$, set for $m = 1, \cdots, M$

$$\alpha_1(m) = p_{S_1}(m). \tag{24}$$

1) Set $J = J - 1$ (move down the tree one scale).
2) For all state variables S_i at scale J, compute for $m = 1, \cdots, M$

$$\alpha_i(m) = \sum_{n=1}^{M} \epsilon^{mn}_{i,\rho(i)}\alpha_{\rho(i)}(n)\beta_{\rho(i)\setminus i}(n). \tag{25}$$

3) If $J = 1$, then stop; else return to step 1.

C. E Step for Multiple Wavelet Trees

To handle $K > 1$ wavelet trees, we add a superscript k to denote the tree number. We denote the observed wavelet coefficients as $\mathbf{w} = [\mathbf{w}^1\ \mathbf{w}^2 \cdots \mathbf{w}^K]$ and the hidden states as $\mathbf{S} = [\mathbf{S}^1\ \mathbf{S}^2 \cdots \mathbf{S}^K]$. The vectors $\mathbf{w}^k = [w_1^k\ w_2^k \cdots w_P^k]$ and $\mathbf{S}^k = [S_1^k\ S_2^k \cdots S_P^k]$ contain the wavelet coefficients and states of the kth tree, respectively.

To implement the E step at iteration l of the EM algorithm, we apply the upward–downward algorithm independently to each of the K wavelet trees. Using the parameter estimates $\boldsymbol{\theta} = \boldsymbol{\theta}^l$, we calculate the probabilities $p(S_i^k = m|\mathbf{w}^k, \boldsymbol{\theta}^l)$ and $p(S_i^k = m, S_{\rho(i)}^k = n|\mathbf{w}^k, \boldsymbol{\theta}^l)$ for each tree via (17) and (18).

D. M Step

Once the probabilities for the hidden states are known, the M step is straightforward. We update the entries of $\boldsymbol{\theta}^{l+1}$ as

$$p_{S_i}(m) = \frac{1}{K}\sum_{k=1}^{K} p(S_i^k = m|\mathbf{w}^k, \boldsymbol{\theta}^l) \tag{26}$$

$$\epsilon^{mn}_{i,\rho(i)} = \frac{\sum_{k=1}^{K} p(S_i^k = m, S_{\rho(i)}^k = n|\mathbf{w}^k, \boldsymbol{\theta}^l]}{K p_{S_{\rho(i)}}(n)} \tag{27}$$

$$\mu_{i,m} = \frac{\sum_{k=1}^{K} w_i^k p(S_i^k = m|\mathbf{w}^k, \boldsymbol{\theta}^l)}{K p_{S_i}(m)} \tag{28}$$

$$\sigma^2_{i,m} = \frac{\sum_{k=1}^{K} (w_i^k - \mu_{i,m})^2 p(S_i^k = m|\mathbf{w}^k, \boldsymbol{\theta}^l)}{K p_{S_i}(m)}. \tag{29}$$

The updates for the state probabilities $p_{S_i}(m)$ and $\epsilon^{mn}_{i,\rho(i)}$ are performed by summing the individual state probabilities and then normalizing so that the probabilities sum to one. Just as for the IM model [26] and the hidden Markov chain model [18], updates for the Gaussian mixture means and variances are performed by a weighted averaging of the empirical means and variances with the weights chosen in proportion to the probabilities of each mixture.

As should be clear from the E and M steps, the per-iteration computational complexity of the EM algorithm is linear in the number of observed wavelet coefficients. The overall complexity may involve a large multiplicative constant, depending on the number of hidden states used and the number of iterations required to converge. However, as shown throughout this paper, even the simplest two-state HMT model can approximate many densities quite well.

E. Tying Within Trees

The M step changes slightly when tying is performed within trees, such as tying wavelet coefficients and their states within a certain subband or scale. (See Section IV-C for the basic idea behind tying.) With tying, we perform extra statistical averaging over coefficients that are tied together within each tree. For the kth tree $\mathbf{w}^k$ with wavelet coefficients w_i^k, we write $i \sim j$ if w_i^k and w_j^k (and their states) are tied, which means that they are modeled with the same underlying density parameters. The set $[i] = \{j | w_j^k \sim w_i^k\}$ denotes the equivalence class of i, with $|[i]|$ the number of elements in the class.

For simplicity, we assume that all trees are tied in the same fashion (that is, the coefficients in the trees $\mathbf{w}^1, \mathbf{w}^2, \cdots, \mathbf{w}^K$ are tied in the same manner) according to the collection of equivalence classes given by the $[i]$'s. In this scenario, the M step becomes

$$p_{S_i}(m) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|[i]|}\sum_{j \in [i]} p(S_j^k = m|\mathbf{w}^k, \boldsymbol{\theta}^l) \tag{30}$$

$$\epsilon^{mn}_{i,\rho(i)} = \frac{1}{K p_{S_{\rho(i)}}(m)}\sum_{k=1}^{K}\frac{1}{|[i]|}\sum_{j \in [i]} \times p(S_j^k = n, S_{\rho(j)}^k = m|\mathbf{w}^k, \boldsymbol{\theta}^l) \tag{31}$$

$$\mu_{i,m} = \frac{1}{K p_{S_i}(m)}\sum_{k=1}^{K}\frac{1}{|[i]|}\sum_{j \in [i]} \times w_j^k p(S_j^k = m|\mathbf{w}^k, \boldsymbol{\theta}^l) \tag{32}$$

$$\sigma^2_{i,m} = \frac{1}{K p_{S_i}(m)}\sum_{k=1}^{K}\frac{1}{|[i]|}\sum_{j \in [i]} \times (w_j^k - \mu_{i,m})^2 p(S_j^k = m|\mathbf{w}^k, \boldsymbol{\theta}^l). \tag{33}$$

Although (30)–(33) appear more computationally intensive than (26)–(29), the computational complexity remains the same since the common parameters for each equivalence class $[i]$ are calculated only once.

REFERENCES

[1] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Stat. Assoc.*, vol. 90, pp. 1200–1224, Dec. 1995.

[2] J.-C. Pesquet, H. Krim, and E. Hamman, "Bayesian approach to best basis selection," in *IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP*, Atlanta, GA, 1996, pp. 2634–2637.

[3] H. Chipman, E. Kolaczyk, and R. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J. Amer. Stat. Assoc.*, vol. 92, 1997.

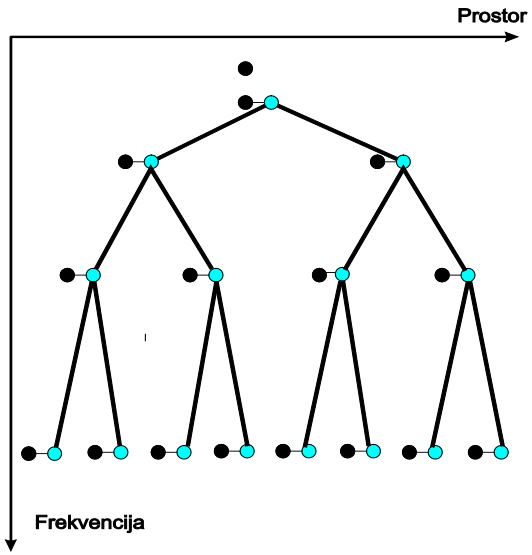
VITERBI ALGORITHM

- Maximum *a posteriori* probability estimation of hidden states S for realized values d and model parameters ϑ
- Minimizing entropy $H_\vartheta(S \mid d)$

$$H_\vartheta(S \mid d) \approx 0 \Rightarrow (S \mid d) = f_\vartheta(d) a.c. \Rightarrow S = f_\vartheta(D) a.c.$$

SELF-ORGANIZATION IN HIDDEN MARKOV MODEL

- Time axis – dyadic frequency axis
- Local causal states – hidden states
- Local complexity $C[i] = H(S[i])$
- Global complexity $C = H(S)$



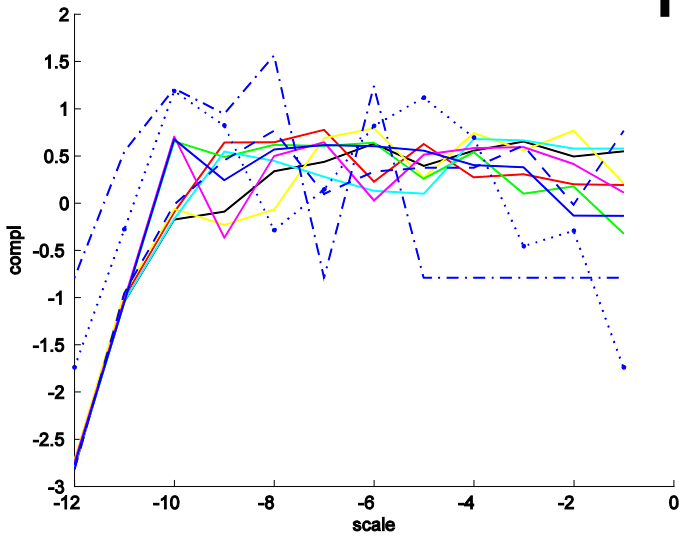
THREE THEOREMS

- Global complexity measures increase of local complexity in temporal domain
- Global complexity measures the accessibility of denoising the signal in WGN
- Global complexity is minimal information required for optimal prediction in spatial domain

DECOMPOSITION OF INFORMATION

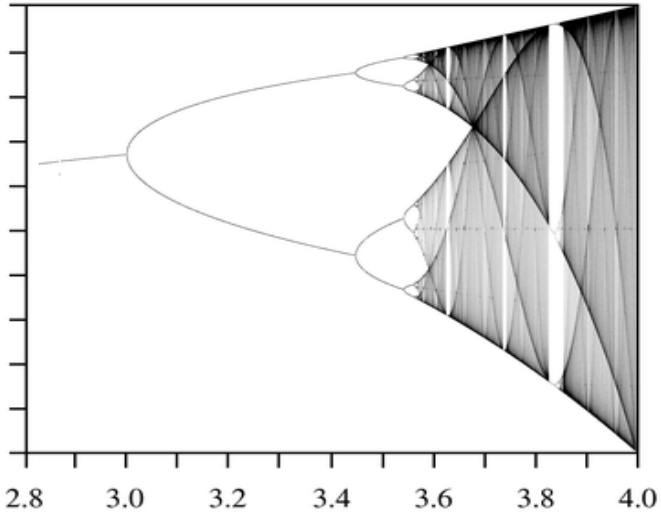
- $H(D) = H(S) + H(D \mid S)$ (noise)
- $H(S)$ informational content of causal variable (acausality – statistical causality)
- Measure of self-organization
- Minimal information for optimal prediction
- Optimal representation of a signal and denoising

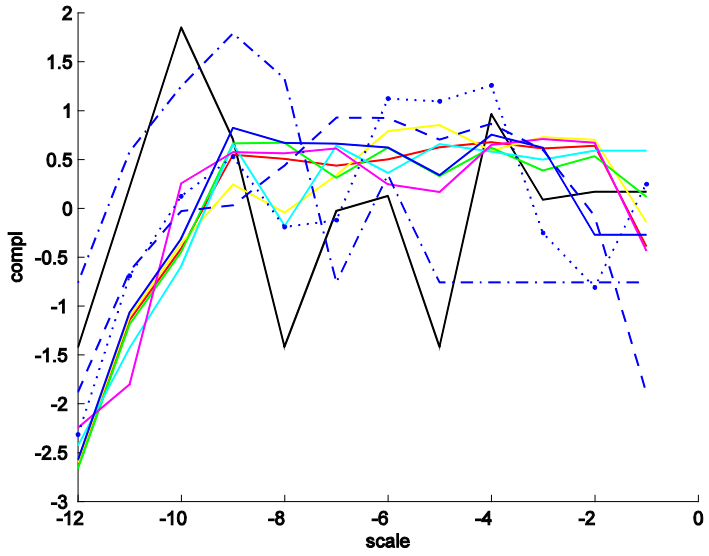
RESULTS



haar	db2	sym3	coif1	bior1.3	rbio1.3	dmey
0.6138	**0.3888**	**0.3234**				
4	3	5				
0.2984	**0.6474**	**0.7300**				

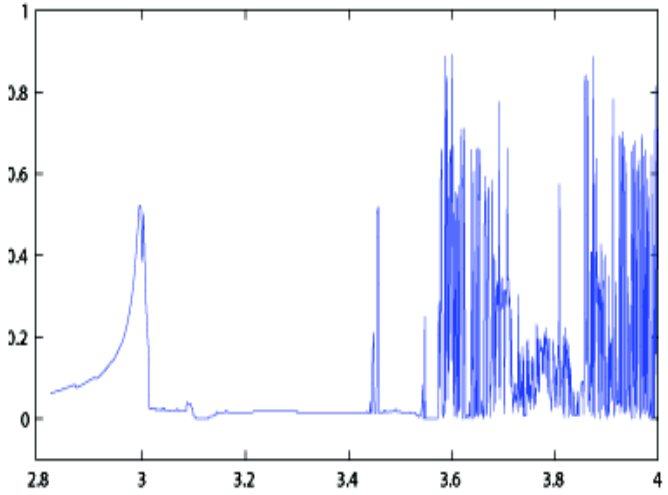
haar	db2	sym3
0.5934	0.3862	**0.3216**
2	2	2
0.2655	0.5518	**0.4171**





	coif1	bior1.3
l	**0.9786**	1.0307
	2	2
l	**0.8867**	0.7108

	coif1	bior1.3
)	**0.9784**	1.0307
	3	3
3	**0.8005**	0.6597



SUMMARY

- Signal – Discrete representation
- Statistical model – Hidden variables
- Self-organization – Complexity
- Mathematico-physical code (relevant information)

SERBIA - SRBIJA

- Not difficult – complex

REPUBLIKA SRPSKA

- 21 year - adulthood

THE END