FEATURE EXTRACTION FOR DATA INPUT TO NEURO-CLASSIFIERS

Gennady Ososkov

Dmitry Baranov

Laboratory of Information Technologies
Joint Intstitute for Nuclear Research,
141980 Dubna, Russia
email: ososkov@jinr.ru

Why ANN for contemporary HEP experiments

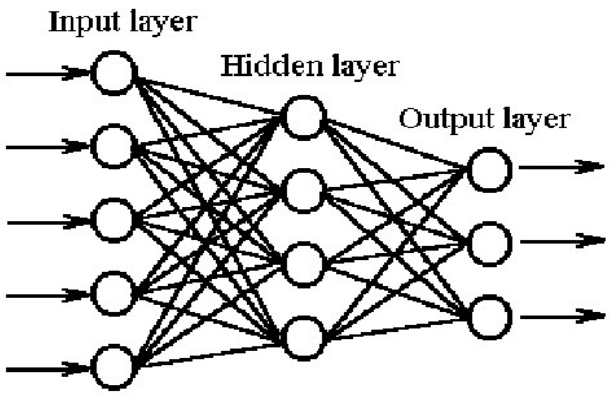
Artificial neural networks (ANN) are widely and successfully used for solving problems of classification, forecasting, and recognition in many scientific applications, in particular and especially in high energy physics (HEP).

Moreover, namely physicists wrote in late 80-ties one of the first NN programing packages – Jetnet. They were also among first neuro-chip users.

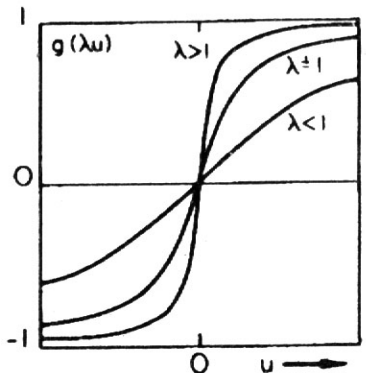
Main reasons were:

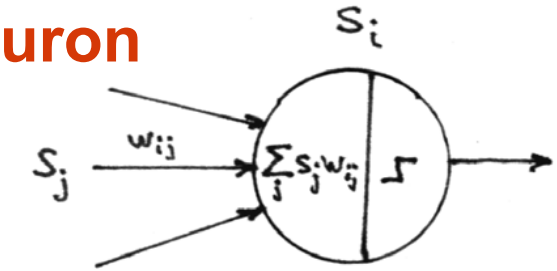
- **The possibility to generate training samples of any arbitrary needed length by Monte Carlo on the basis of some new physical model**
- **neuro-chip appearance on the market at that time which make feasible implementing a trained NN, as a hardware for the very fast triggering and other NN application.**

Brief reminder of ANN basic concepts



Artificial neuron



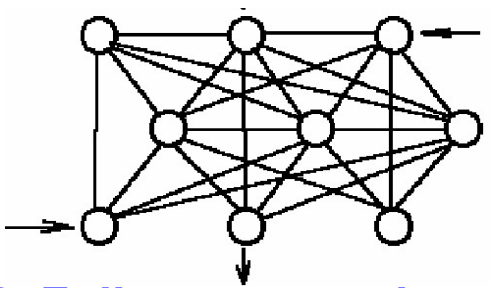


the *i*-th neuron output signal

$h_i = g(\Sigma_j w_{ij}\, s_j)$

Activation function $g(x)$

1. Feed-forward ANN (MLP or RBF- networks)



2. Fully-connected or recurrent ANN

Output-layer neurons transform hidden neurons h_j as

$$y_j = g(\sum_k w_{kj} h_k)$$

Then training sample $(\{x_i\}^{(m)}, \{z_i\}^{(m)})$ is needed to train MLP to obtain weights by the error backpropagation algorithm

$$E = \Sigma_m \Sigma_{ij} (y_i^{(m)} - z_i^{(m)})^2 \rightarrow min_{\{wik\}}$$

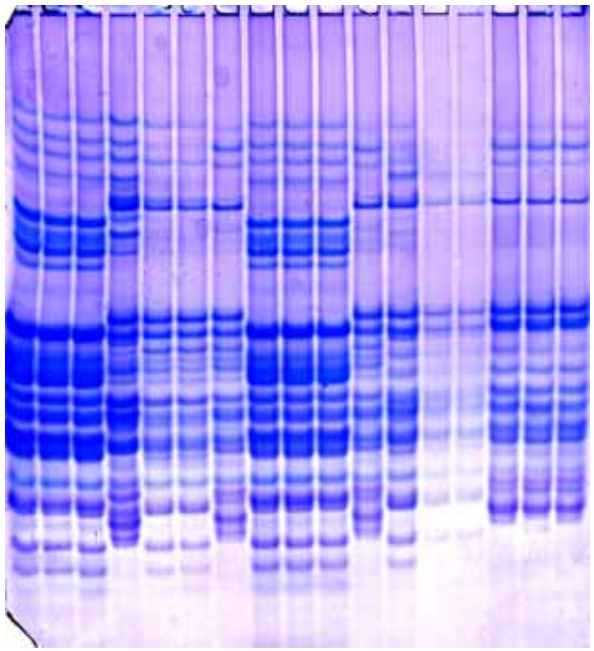
which is based on the steepest descent method.

The "curse of dimensionality" problem arises in many non-physics applications due to

- very high dimensionality of experimental data to be classified
- the amount of these data is scarce for training and quality verification of a trained network

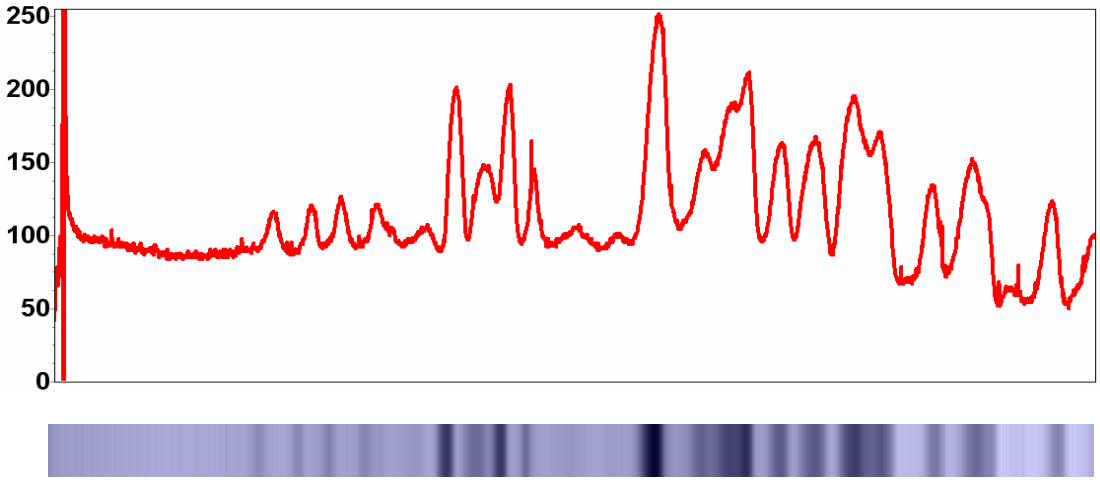
Example: Genetics of proteins

Radiobiology study



Electrophoregram example with 17 wheat cultivar

The genetics of gliadin (alcohol soluble protein) have been studied in detail by a special gel electrophoresis. Each electrophoretogram strip after its digitalization became a densitogram spectrum consisting of about 4000 pixels.



It can be considered as a simple genetic formula, which allows for a qualified expert to classify any spectrum to its corresponding protein.

Such classifying procedure is of great importance in radiobiology and, especially, in agriculture

Therefore it is to be automatized.

Genetics of protein (continue)

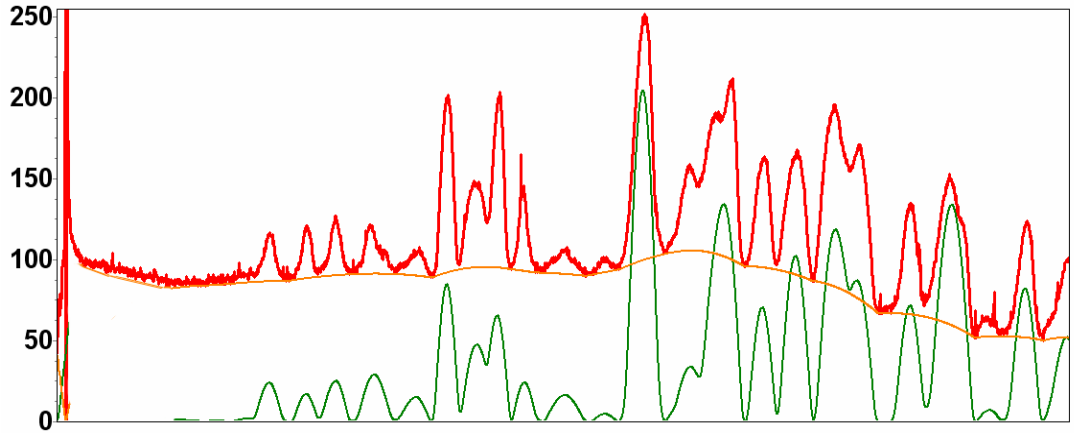
The problem: realize ANN based expert system to identify the wheat cultivar by its spectrum.

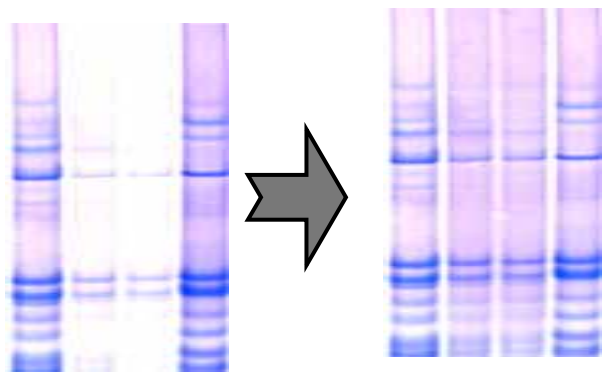
Note: the electrophoregram information must be considerably preprocessed to suit as input to an expert system

Preprocessing stages:

1. digitization and standardization of densitometry data

- smoothing, denoising and eliminating background pedestal;





- density normalization to the range 0-255;

- aligning all strip to fix the beginning and the end of information on each gel (was fulfilled by Hamming neural net)

2. extracting most informative features

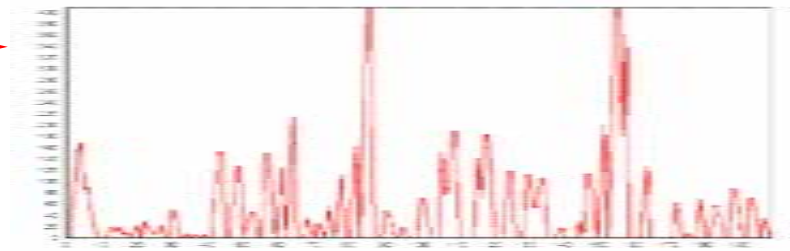
Curse of dimensionality problem

Input: 4000 pixels

Output: 5 ÷ 20 sorts to be classified

MLP dimension D=4000*1000+1000*20=

4.2 *10^6, **i.e millions of weights**



or equations to solve by the error back propagation method!

A cardinal reduction of input data was needed

Feature extraction approaches

1st approach: spectrum coarsing from 4000 points into 200 zones with averaged density (**D**=16400 weights)

The real size of the training sample is 120 etalons preliminarily classified by experts for each of 20 wheat sorts, i.e. for 5 different sorts we have 600 etalons for training.

Result for 5 sorts: after training ANN (200/80/5) the efficiency was 85%.

2d Fast Fourier transform (FFT). Real part of direct FFT was used to transform input data to the frequency domain, where the highest frequencies were cut up to 256 (16 times of reduction)

$$H_n \equiv \sum_{k=0}^{N-1} h_k \exp\left(\frac{kn}{N} 2\pi i\right) \qquad n \in \left[-\frac{N}{2}, \frac{N}{2}\right]$$

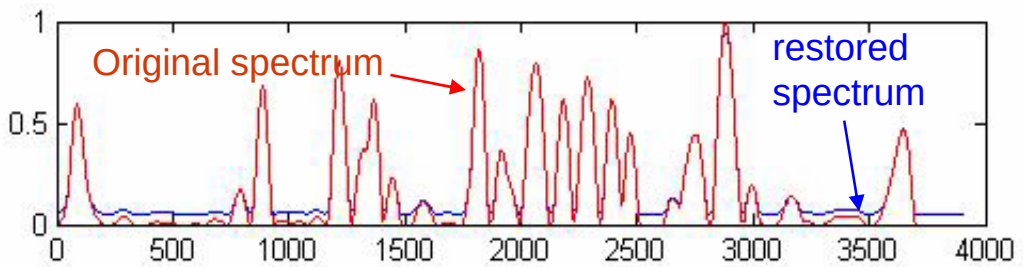
After transforming all training samples to Fourier space NN-classifier (256/40/5) was trained on them and tested again on transformed sample.

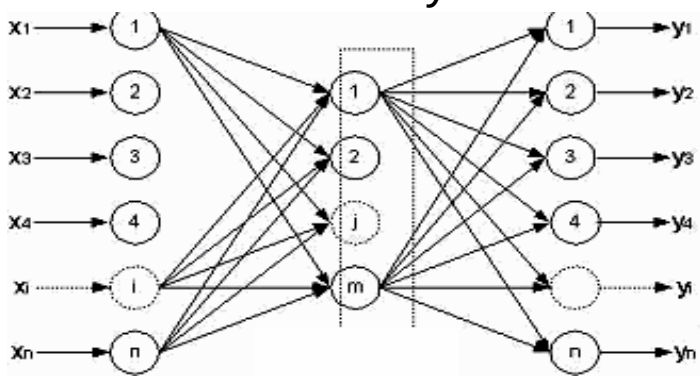
Result for 5 sorts: 100% of efficiency

Feature extraction approaches (contin)

3d. Principal component analysis (PCA) transforms d-dimensional data to a m-dimensional subspace by using the information of their covariance matrix S_X =cov($X_i X_k$). The orthogonal Karhunen–Loeve transform gives the diagonal form of S_X with eigenvalues l_i numbered in decreasing order. Thus, we can retain only *m* most significant eigenvalues $l_1, l_2, \ldots, l_m$ ($m \ll p$) and express the input data in terms of these principal components as $X_i \cong l_{1i}Y_1 + l_{2i}Y_2 + \ldots + l_{mi}Y_m$.

PCA has its neural network implementations what allows to **avoid cumbersome calculations of covariance matrices and their eigenvectors**. It is done by so-called **recircular (autoassociative) NN.** Such NN uses 4000 input neurons, as etalon, output ones. The number of hidden neurons N_{hid} should correspond to the number of principal components. The best efficiency of the next classifying network was obtained for N_{hid} **=150,** i.e. data compression in more than 20 times.





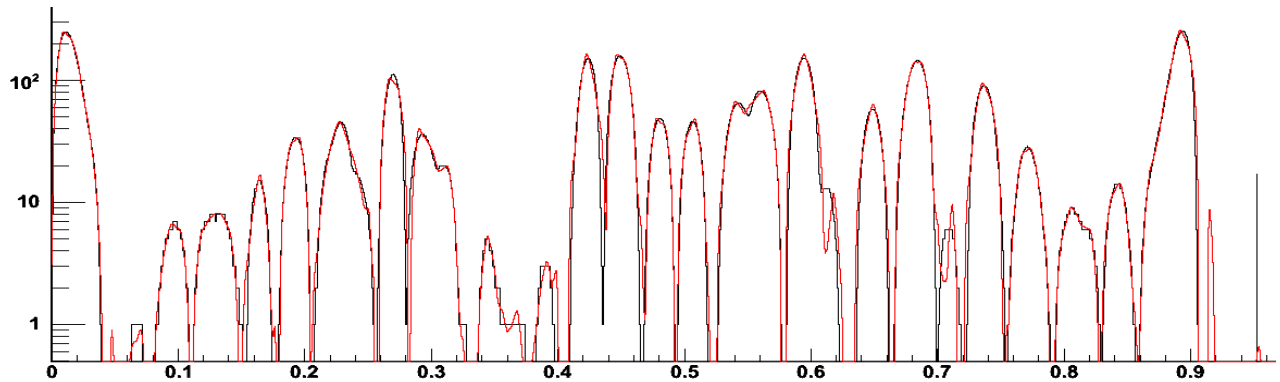
(PCA in its NN form was also successfully used for the face recognition)

Result for 5 sorts: after extracting 150 PC from all data of training and testing samples the PCA classifying efficiency was 99,54%. This efficiency keeps stable while the increase of sorts number till 8 then drops down for 17-25%

Feature extraction approaches (contin)

4th. Discrete wavelet transform (DWT). Coiflet DWT of the 6th order were applied to transform all training and testing samples into wavelet space. Then NN-classifier (256/40/5) was trained and tested on them.

Result for 5 sorts: 100% of efficiency



An example of the quality of DWT transform
——— original spectrum
——— restored spectrum
(ordinata is in logarithmic scale).

5th. Taking into account the heuristics of experts. Experts pay attention mainly to the order in which higher and lower peaks are alternating. So it was proposed to recognize all well-pronounced peaks, fit each of them by some of bell-shaped function, like a Gaussian, in order to evaluate 3 basic parameters of each peak: **position, integral** (square under this peak) and **its rank** according to its integral.

The maximum number of peaks on every of all densitogramms given to us was equal to 37, so there were 37*3=111 input neurons, 5 output and 40 hidden neurons.

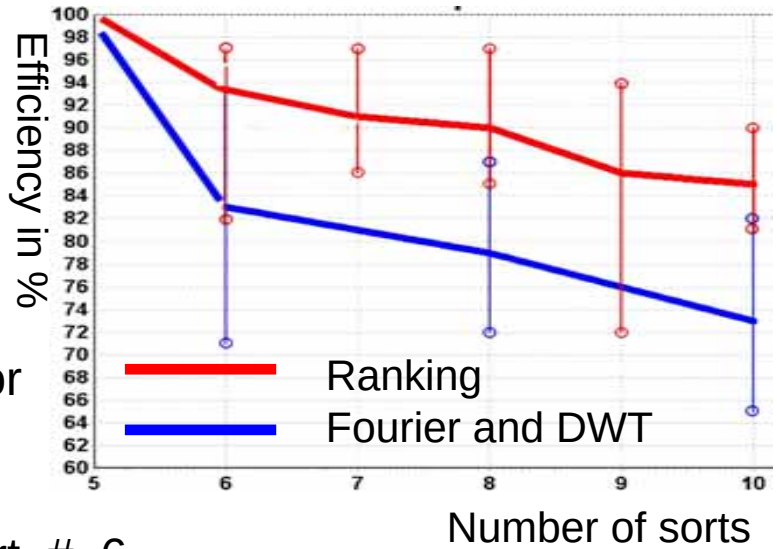
Result for 5 sorts: after training ANN (111/40/5, **D**=4640) the efficiency was **100%!**

Status of the wheat protein classification

On the basis of this study software system was elaborated in the collaboration with the Vavilov Institute of General Genetics RAS (VIGG RAS) for the full chain of genetic analysis of electrophoregrams including the preprocessing stage.
The system is now in use for testing and further developing.

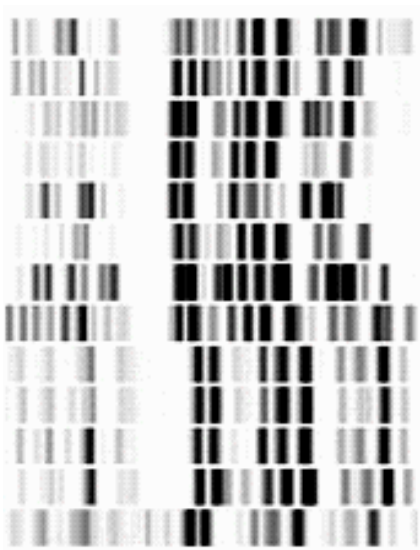
Results for higher number of sorts shows:

- Notable decrease of the classifying efficiency;
- a preference of the ranking method

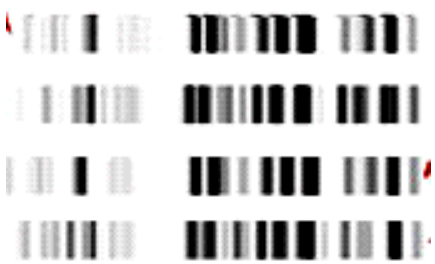


The main reasons are

1. Spectrum variability for the same sort



2. Spectrum similarity for genetically close sorts



Sort # 6
13
#19
26

Spectra for the wheat cultivar #19

Therefore, the further development is thought to apply Kohonen SOM ANN and **to** formalize expert classifying approaches in order to elaborate a hierarchy of NN for the wheat protein classification.

Summary

1. A comparative study of the feature extraction methods was fulfilled on the example of wheat protein genetic classification
2. 5 different methods were tested
   - coarsing data to 200 zones with mean density;
   - PCA by recircular network;
   - FFT;
   - DWT;
   - ranking data by peak integral.
3. Last four methods show satisfactory results on classifying 5 sorts
4. Further increasing sort numbers causes drop of classifying efficiency, ranking method showed its advantage.
5. Software system was elaborated in collaboration with the VIGG RAS Institute for the full chain of electrophoretic data genetic analysis.
6. Further system development is in progress.

Authors thank Dr.A.Kudryavtsev (VIGGS) for the problem formulation and providing all experimental data and also S.Lebedev, S.Dmitrievsky, and E.Demidenko (JINR) for the essential help in performing calculations.

Thanks for your attention!

Formula to estimate the total weight number ***Nw***:
where ***Nx*** number of input neurons,
Ny — number of output neurons,
Q — length of training sample

$$\frac{N_y Q}{1+\log_2 Q} \le N_w \le N_y \left(\frac{Q}{N_x} + 1 \right) \left(N_x + N_y + 1 \right) + N_y$$

Brief introduction to wavelets

One-dimensional wavelet transform (WT) of the signal f(x) has 2D form

$$W_\psi(a,b)f = \frac{1}{\sqrt{C_\Psi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{|a|}} \Psi\left(\frac{b-x}{a}\right) f(x)dx,$$

where the function Ψ is the wavelet, b is a displacement (shift), and a is a scale. Condition $C_\psi < \infty$ guarantees the existence of Ψ and the wavelet inverse transform. Due to freedom in Ψ choice, many different wavelets were invented.

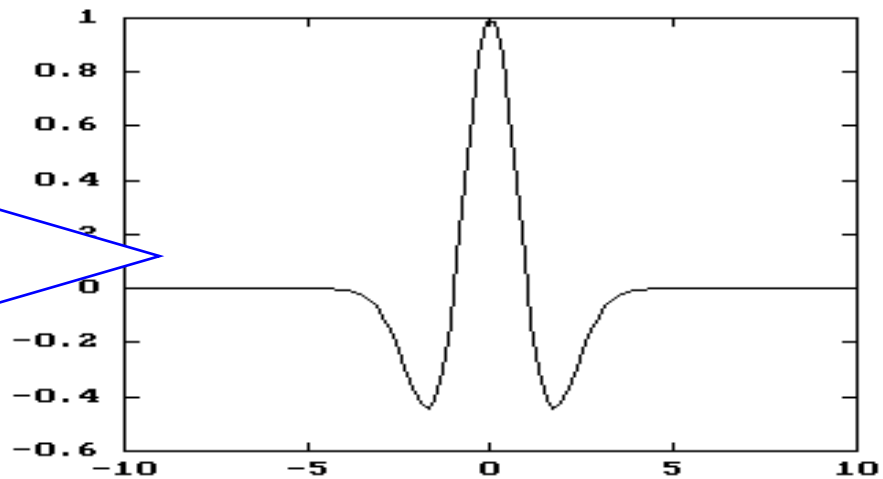
The family of **continuous wavelets** is presented here by **Gaussian wavelets**, which are generated by **derivatives of Gaussian function**

$$g_n(x) = (-1)^{n+1} \frac{d^n}{dx^n} e^{-x^2/2},$$

Two of them, we use, are $g_2(x) = (1-x^2)e^{-\frac{x^2}{2}}$

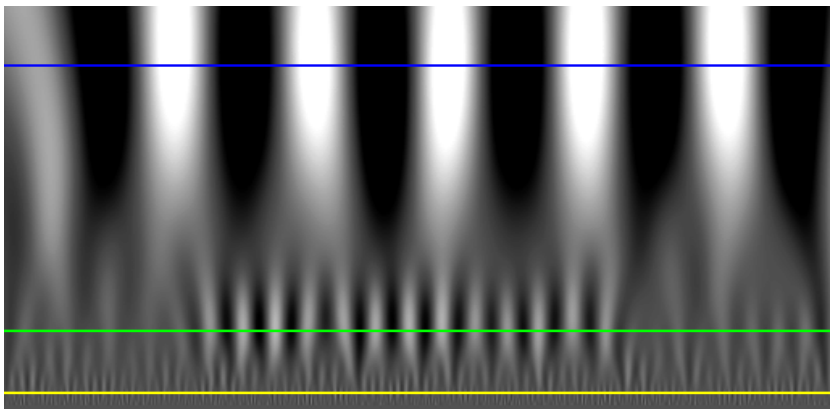
and $g_4(x) = (6x^2 - x^4 - 3)e^{-\frac{x^2}{2}}$.

Most known wavelet **G_2** is named **"the Mexican hat"**

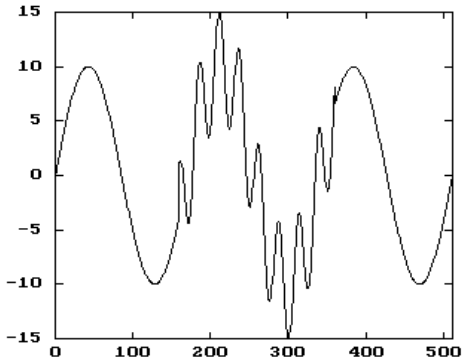


Wavelets can be applied for extracting very special features of mixed and contaminated signal

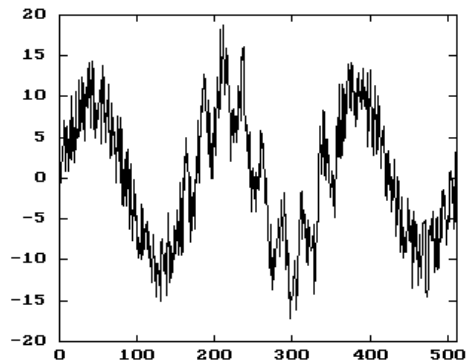
An example of the signal with a localized high frequency part and considerable contamination



G_2 wavelet spectrum of this signal

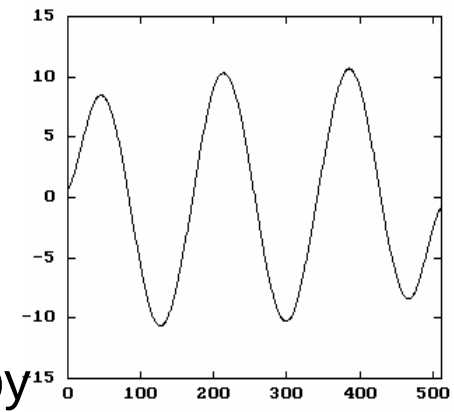


Source sample

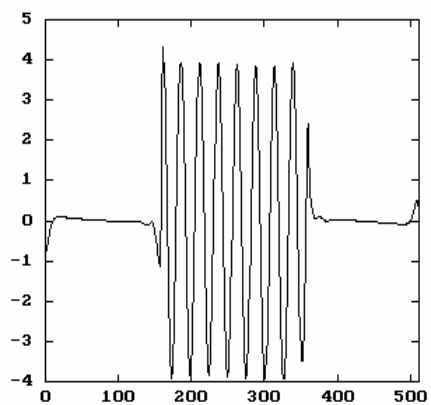


Noise added

then wavelet **filtering** is applied



Low frequency



High frequency

Filtering works in the wavelet domain by thresholding of scales, to be eliminated or extracted, and then by making the inverse transform

Filtering results. Noise is removed and high frequency part perfectly localized

Continuous wavelets: pro and contra

PRO: **- Using wavelets we overcome background estimation**

- Wavelets are resistant to noise (robust)

CONTRA: **- redundancy** $\longrightarrow$ slow speed of calculations

- nonorthogonality (signal distotrs after inverse transform)

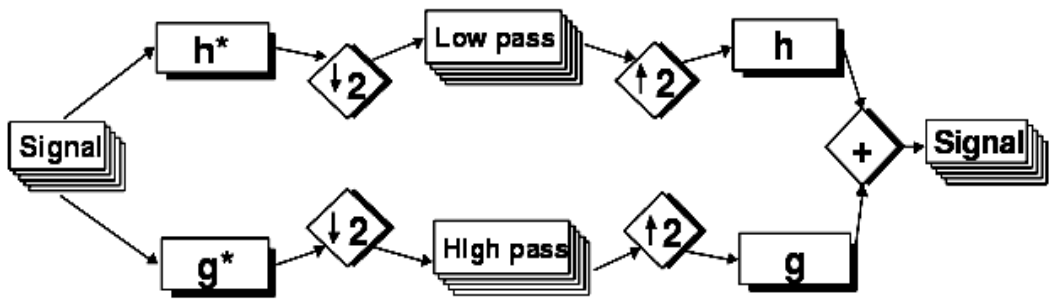
Besides, **real signals** to be analised by computer are **discrete,**

So **orthogonal discrete wavelets** should be preferable.

The discrete wavelet transform (**DWT**) was built by Mallat as **multi-resolution analysis**. It consists in representing a given data as a signal decomposition into basis functions ***φ*** and ***ψ.*** Both these functions must be **compact** in time/space and frequency domains.

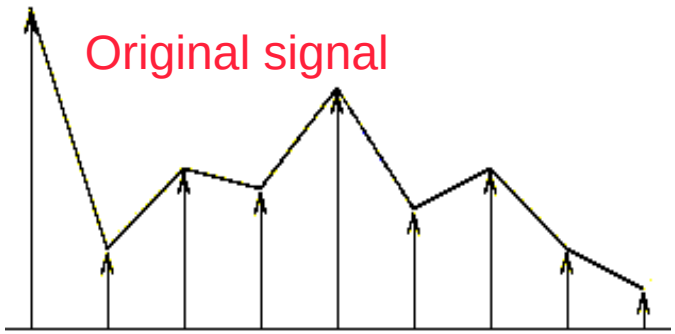
$$\psi_{j,k} = 2^{-j/2}\psi(2^{-j}t - k)$$
$$\phi_{j,k} = 2^{-j/2}\phi(2^{-j}t - k)$$

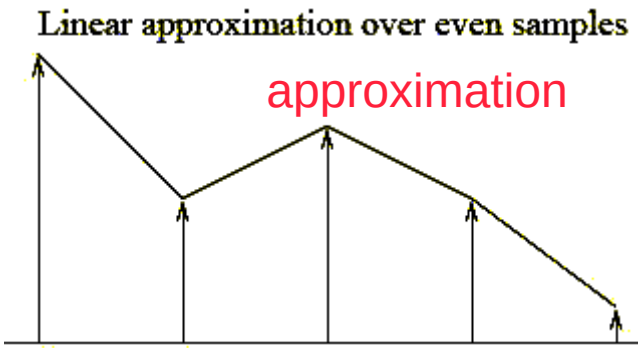
$$f(x_i) = \sum_{k=1}^{N} s_k \phi_{Lk}(x_i) + \sum_{j=1}^{L}\sum_{k=1}^{N} d_{jk}\psi_{jk}(x_i)$$

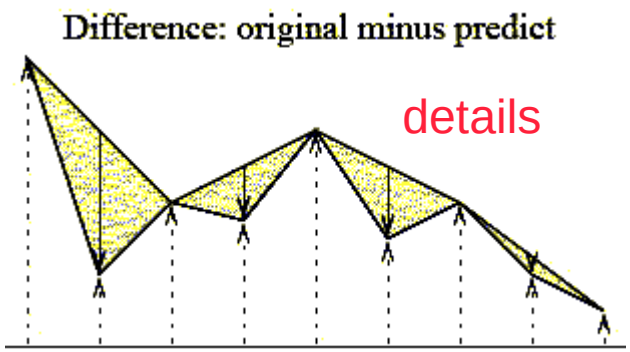


Scheme of one step of the wavelet decomposition and reconstruction

Lifting scheme as an example of discrete wavelets



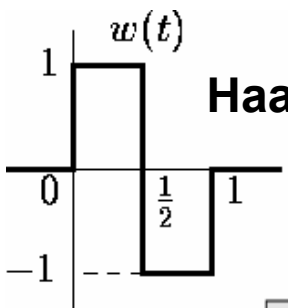




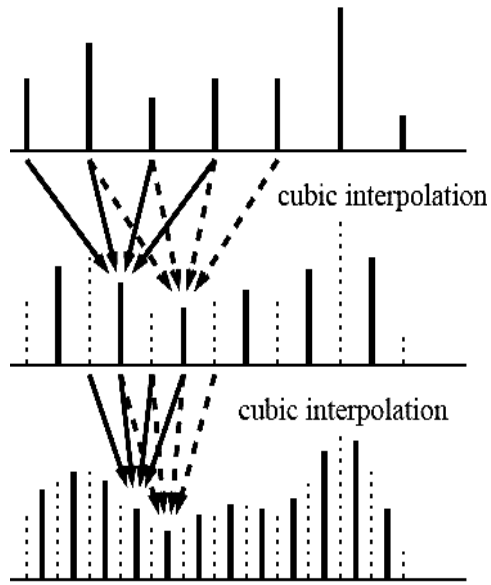
Algorithm:

- **Decimate** into odd - even
- **Predict** and obtain details
- Store s_k and d_k "in place"
- continue recursively

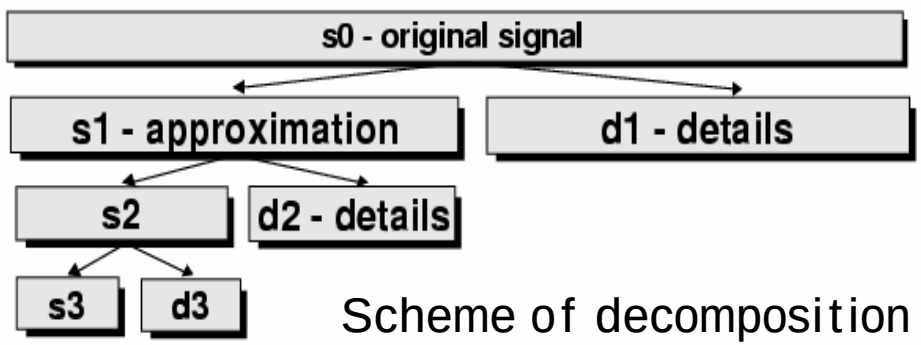
requirement: sample size must be a power of 2 (2^n)



Haar wavelet

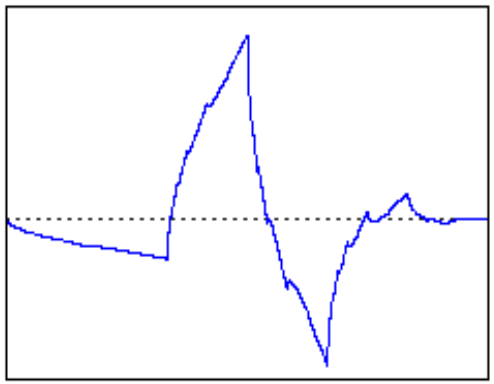


Prediction can be non-linear

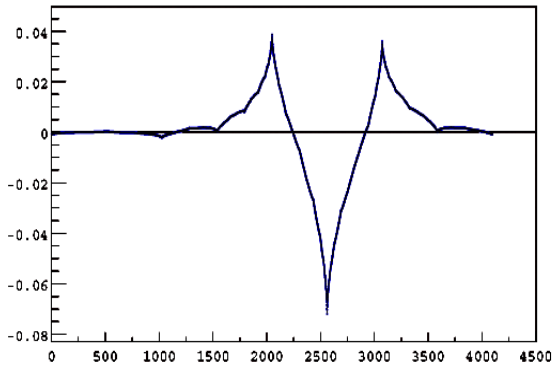


Scheme of decomposition algorithm

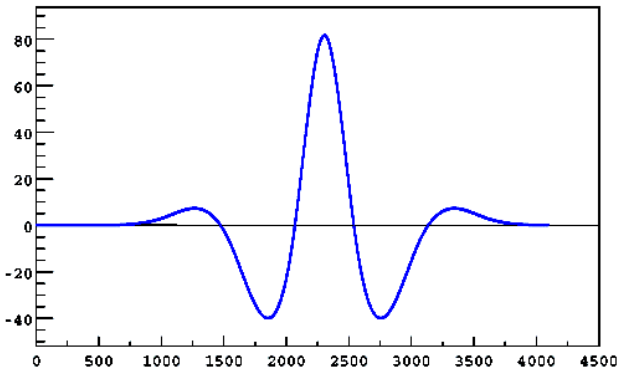
Various types of discrete wavelets



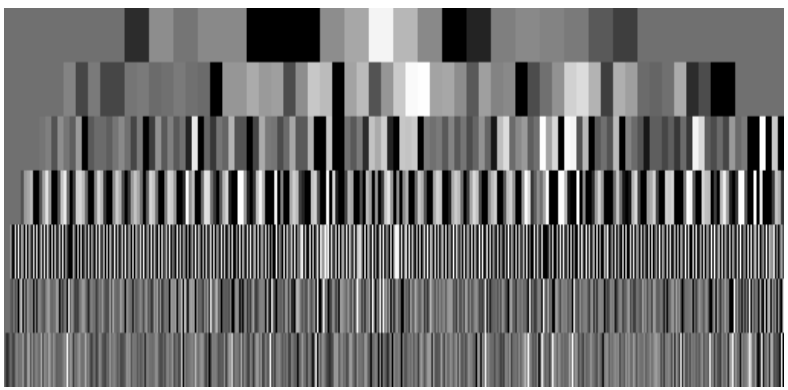
Daubechie's wavelet
with 2 vanishing momenta



Bi-orthogonal
CDF44 wavelet



Coiflet – most symmetric



An example of **Daub2** spectrum

Denoising by DWT shrinking

wavelet shrinkage means, certain wavelet coefficients are reduced to zero:

$$W_\psi = 0, if |W_\psi| < \lambda$$

Our innovation is

the adaptive shrinkage,

i.e. $\lambda_k = 3\sigma_k$ where *k* is decomposition level (*k*=*scale*$_1$,...,*scale*$_n$), σ_k is RMS of *Wψ* for this level (recall: sample size is 2^n)