JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

WIENER FILTERING

During the war, Wiener worked on aircraft fire control. (1 plane for 10,000 rounds of ammunition). Wiener worked on a electronic gunsight.

Where to aim the gun? We must predict the place where the plane will be in the future!!!

The contribution was to think of the plane position in STATISTICAL terms. From all possible paths choose the one that is more probable.

This was a break with deterministic thinking. Helped establish the conceptual framework for information theory.

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

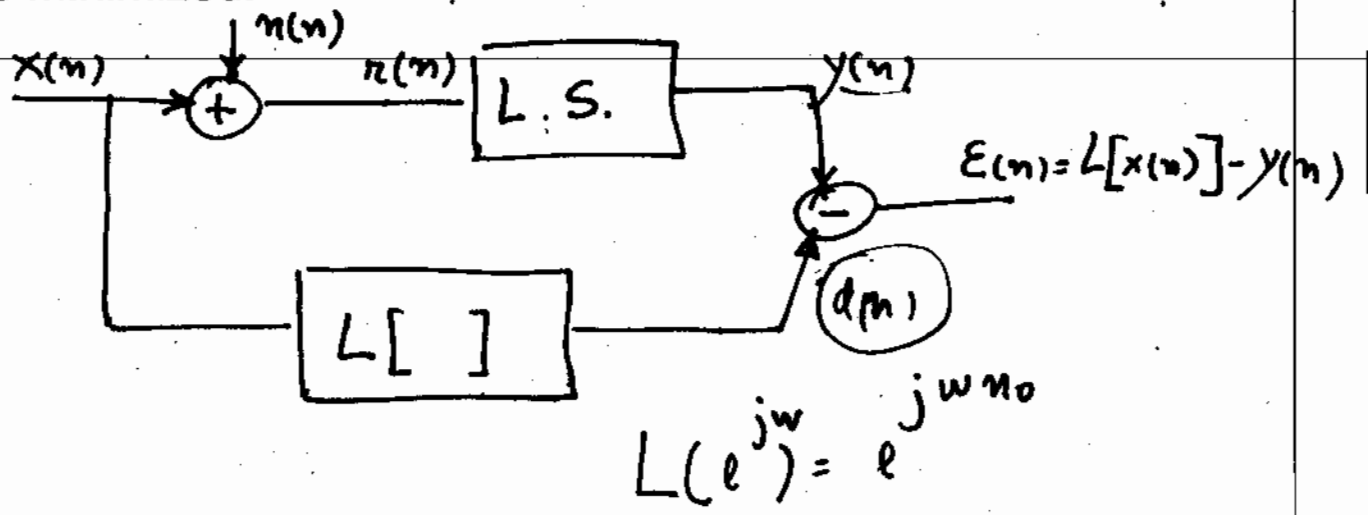
We will only formulate the problem of pure prediction (it can be extended also to estimation).

A random signal x(n) in an additive random noise n(n), is going to be sent throught some linear system, such that the output y(n) is an approximation to some linear operator L[x(n)] on the input signal.

The filter is to be designed such that the least méan square error

$$J = E\{\varepsilon^2(n)\} = E\{(L[x(n)] - y(n))^2\}$$

is to be minimized.



JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

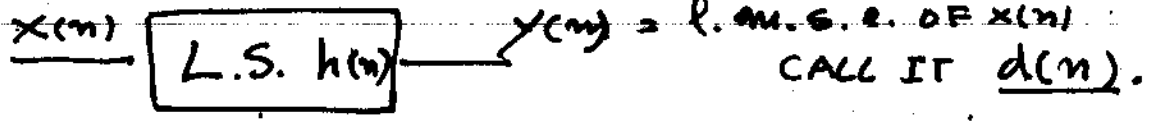
Three cases

⇒ 1) $n_0 > 0$; $n(n) = 0$ PURE PREDICTION OF $x(n)$

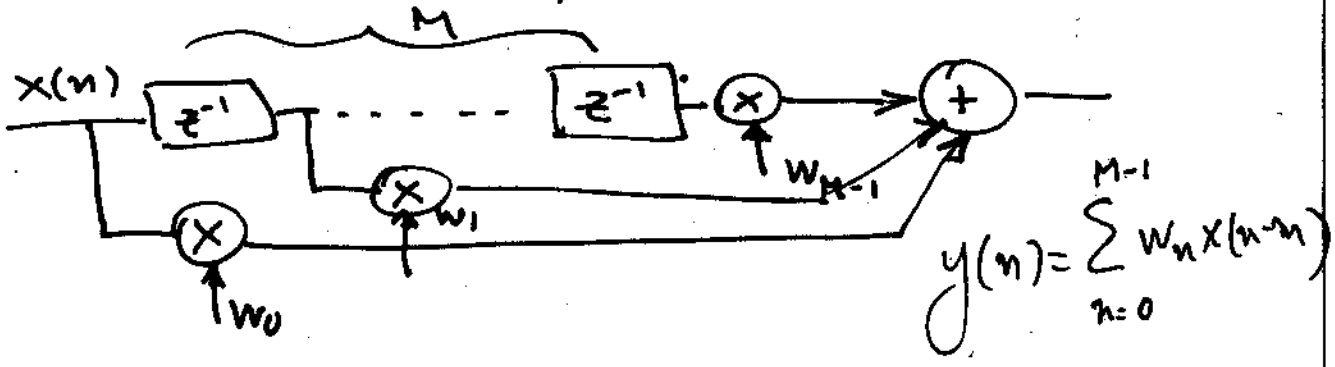
2) $n_0 > 0$; $n(n) \neq 0$ PREDICTION + ESTIMATION

3) $n_0 < 0$; $n(n) \neq 0$ ESTIMATION OF $x(n)$

We will only treat the pure prediction case. We will predict the input at time n using M previous samples.



with x(n) widesense stationary, and h(n) physically realizable. Let us assume an FIR structure, i.e. that



JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

INNER PRODUCT (DOT PRODUCT)

$$y(n) = \sum_{k=0}^{M-1} x(n-k)\, w(k)$$

$$\begin{bmatrix} x^T(n) \end{bmatrix} \begin{bmatrix} w \end{bmatrix}$$

Using matrix notation (COLUMN VECTORS)

$$\vec{w}^T = [w_0, \ldots . w_{M-1}]$$

$$\vec{x}^T_{(n)} = [x(n), \ldots . , x(n-M-1)]$$

$$\Longrightarrow \quad y(n) = \vec{x}^T(n)\vec{w} = \vec{w}^T\vec{x}(n)$$

Now the mean square error is

$$\min_w J = E\{\varepsilon^2(n)\} = E\left\{\left| d(n) - \sum_{k=0}^{M-1} x(n-k)\, w(k)\right|^2\right\}$$

This is a set of M equation in M unknowns, the filter coefficients.

To minimize the error, we equate to zero the derivatives of the error with respect to the unknowns.

$$\frac{\partial J}{\partial w_i} = 0 \qquad i = 0, \ldots . M-1$$

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

$$\frac{\partial J}{\partial w_i} = E\left\{ 2 \frac{\partial}{\partial w_i} \left(d(n) - \sum_{k=0}^{M-1} x(n-k)\, w_k \right) \left(d(n) - \sum_{k=0}^{M-1} x(n-k) w(k) \right) = \right.$$

$$\Rightarrow E\left\{ -2\, x(n-i) \left(d(n) - \sum_{k=0}^{M-1} x(n-k)\, w(k) \right) \right\} =$$

$$\Rightarrow E\left\{ -x(n-i)\, d(n) \right\} + E\left\{ \sum_{k=0}^{M-1} x(n-i)\, x(n-k)\, w(k) \right\} = 0$$

$$E\left\{ x(n-i)\, d(n) \right\} = \sum_{k=0}^{M-1} E\left\{ x(n-i)\, x(n-k) \right\} \cdot \underline{w}(k)$$

$$i = 0, \ldots, M-1$$

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

Calling

$$E\{x(n-i)\,d(n)\} = P(i)$$

the crosscorrelation function between d(n) and input, and

$$E\{x(n-i)\,x(n-k)\} = R(i-k)$$

the autocorrelation function of the input,.we get

$$P(i) = \sum_{k=0}^{M-1} R(i-k)\,w(k) \qquad i=0,\ldots M-1$$

This set of M equations in M unknowns is called the NORMAL EQUATIONS, or the Wiener- Hopf equations.

In our matrix notation if we call P the colum vector

and W^T the optimal weight vector, we can write $P = R\,W_{opt}$

$$\begin{bmatrix} P(0) \\ \vdots \\ P(M-1) \end{bmatrix} = \begin{bmatrix} R(0) & R(1) & \cdots & R(M-1) \\ R(1) & R(0) & \cdots & R(M-2) \\ \vdots & & & \\ R(M-1) & \cdots & & R(0) \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_{M-1} \end{bmatrix}$$

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

If we use the vector (matrix) notation we can write the error as

$$J = E\{\varepsilon^2(n)\} = E\{(d(n) - \vec{W}^T\vec{X}(n))^2\}$$

So that

$$\frac{\partial J}{\partial W} = -2E\{d(n) - \vec{W}^T\vec{X}_n\}\vec{X}_n^T \equiv 0$$

$$\Rightarrow E\{d(n)\vec{X}_n^T\} = E\{\vec{W}^T\vec{X}_n\vec{X}_n^T\}$$

$$[\vec{P} = \vec{W}_{OPT}^T\vec{R}] \quad P = RW_{OPT} \quad W_{OPT} = W^*$$

To obtain the coefficients we have to invert the equation

$$\boxed{\vec{W}_{OPT} = \vec{R}^{-1}\vec{P}}$$

This is the set of coefficients that minimize the mean square error between the predicted value y(n) and x(n).

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

VECTOR SPACE INTERPRETATION

Each signal x(n) with M components can be regarded as a vector with a length $\sqrt{E\{x_n\}^2}$

The inner product between d(n) and say x(n-2) is

$$E\{d(n)\,x(n-2)\} = P(2)$$

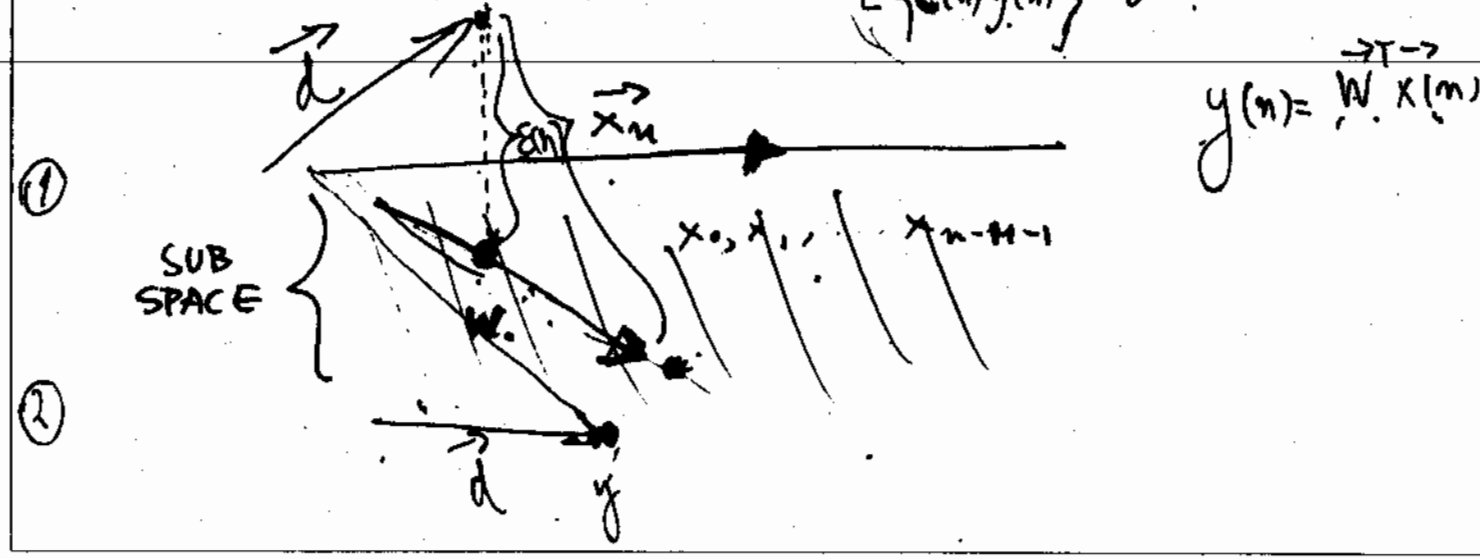
$$E\{x(i)\,x(i-1)\} = R(1)$$

Two vectors are orthogonal if the inner product is zero.

The output of the linear system is a linear combination of the input with the system coefficients, so it must lie in the hiperplane defined by the input.

$E\{\vec{e}(n)\,\vec{y}(n)\} = 0$

$y(n) = \vec{W}^T \vec{X}(n)$



JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

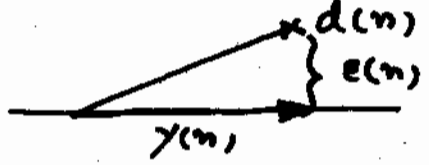
$\varepsilon_n = d(n) - y(n)$

The desired output can be written as

$$d(n) = y(n) + \varepsilon(n) = \sum a(i)x(n-i) + \varepsilon(n)$$

Now if d(n) exists in the subspace defined by x(0), x(n-M-1) we can solve the problem with no error. Just find y(n) that coincides with y(n).

But in general d(n) will not exist in this subspace due to the componenet e(n).



In this case since y(n) must lie in the x(n) subspace, the best we can do to minimize the error is to choose the the orthogonal projection of d(n) in the subspace x(n). This will give the best m.s.q.e. estimate.

It is obvious to see that e(n) is orthogonal to y(n), so to every x(i).

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

$$E\{x(n-i)\varepsilon(n)\} = 0 \Longrightarrow$$

$$E\{x(n-i)[y(n)-d(n)]\} = E\{x(n-i)(\sum_k x(n-k)w_k - d(n))\}$$

$$= E\{x(n-i)d(n)\} = E\{\sum_k x(n-k)\,x(n-i)\}w_k$$

$$P(i) = \sum_{k=0}^{M-1} R(k-i)\,w(k)$$

This is the same equation as before. So minimizing the the m.s.q.e is to choose the error orthogonal to the input samples!!!

(NOTE: there is a fund amental theorem called the WOLD decomposition that says that an unpredicatable signal can be decomposed as a sum of two orthogonal signals: one predictable and the other that is an autoregressive process).

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

Notice that the error is uncorrelated with the input.

$$\varepsilon(k) = d(k) - X^T(k)\, W$$

In fact, multiplying both sides by X(k)

$$\varepsilon(k)\, X(k) = d(k)\, X(k) - X^T(k) X(k)\, W$$

and taking the expected value

$$E\{\varepsilon(k) X(k)\} = P - R\, W$$

For Wopt

$$E\{\varepsilon(k) X(k)\} = P - P = 0$$

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

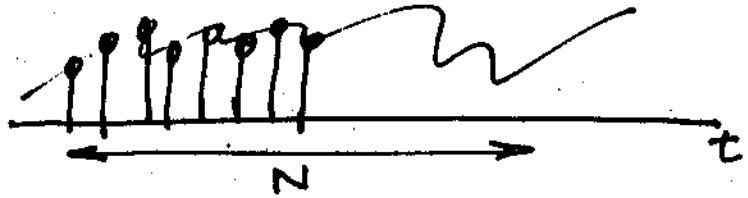
904-335-8444

principe@brain.ee.ufl.edu

TEMPORAL AVERAGES

This is what we normally do!

Consider a time segment of N points, from a realization of a stationary random process assumed ergodic.



Let us formulate the mean square error minimization in this context.

We define the error as

$$J_n = \sum_{m=0}^{N-1} e_n^2(m) = \sum_m \left(d_n(m) - y_n(m)\right)^2$$

$$= \sum_{m=0}^{N-1} \left[d_n(m) - \sum_{k=0}^{M-1} w(k)x(m-k)\right]^2$$

and we proceed as before

M delay

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

$$\frac{\partial J_n}{\partial w_i} = 0 \qquad i = 0, \ldots M-1$$

$$2 \sum_n \left(d(n) - \sum_k w(k) x(n-k) \right) x(n-i) = 0$$

$$\Rightarrow \sum_n d(n) x(n-i) = \sum_n \sum_k w(k) x(n-k) x(n-i)$$

DEFINE:

$$\begin{cases} \sum_{n=0}^{N-1} d(n) x(n-i) = P_n(i) \\ \sum_{n=0}^{N-1} x(n-k) x(n-i) = R_n(i-k) \end{cases}$$

So

$$P_n(i) = \sum_{k=0}^{M-1} R_n(i-k) w(k)$$

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

We get the set of normal equations again, but now the auto/ cross correlation functions are computed in time, and with a finite number of points.

Question: If the segment length is changed or moved in time does R_n stay the same?

SOLUTION OF TH E OPTIMAL VECTOR

If R^{-1} exists, then we can compute the optimal solution directly

$$W_{op} = R^{-1} P$$

If R is NxN, can use Gausian elimination, but it is very inefficient computationally (complexity of algorithm is $O(N^3)$)

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

However R has a special structure as we have seen. It is Toeplitz (symmetric with all elements in the diagonal equal). For Toeplitz matrices there are much more efficient algorithms to solve the equation. One commonly used recursive solution was proposed by Durbin using the Levinson Robinson algorithm. The complexity drops to $O(N^2)$.

If the matrix involving the statistics of the input is not the autocorrelation function, but the crosscorrelation function (as sometimes is the case), we can use the Cholesky decomposition which is still more efficient than the Gaussian elimination.

There are also iterative solutions to W. Consider the estimation of W(l+1) as

$$\bar{W}(l+1) = (I - \mu\bar{R})\,\bar{W}(l) + \mu\bar{P}$$

where $W(0) = 0$

and is such that $\bar{V}' = \bar{V}[I - \mu R]$ $\bar{V}'$ SHALLER IN LENGTH THAN V

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

This method requires $O(N^2)$ multiplication per iteration. Solution converges to $R^{-1}P$. In fact

$$W(1) = (I - \mu R)\cdot 0 + \mu P$$

$$W(2) = (I - \mu R)\mu P + \mu P$$

$$\vdots$$

$$W(k) = \mu \left[\sum_{m=0}^{k-1} (I - \mu R)^m\right] P$$

$$\text{IN THE LIMIT } W = \mu \left[\frac{I}{I-(I-\mu R)}\right] P \longrightarrow R^{-1}P$$

- Direct method is generally more efficient, but for large data segments they are comparable.
- When R is singular, i.e. not full rank (i.e when det(R)=0), direct method does not work. However by iteration we get a solution (that is not unique).
- When R is almost singular, the direct method is very sensitive to roundoff errors. The iterative solution tends to be more robust.

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

RECAP:

Wiener filter is an adaptive filter.

Adaptation requires desired signal and an error criteria. Prediction uses the next sample as the desired signal, and uses the mean square error. This leads to a very elegant and mathematical treatable solution- the normal equations.

R^{-1} must exist. Otherwise the solution is not unique.

In vector spaces, the Wiener solution finds the orthogonal projection of the desired signal vector, onto the subspace spanned by the input. Remmember, the error is orthogonal to all of the input vectors.

Normally we use time averages to solve the problem. Estimates of the autocorrelation function may vary. Assume ergodicity. Can use Durbin algorithm.

This means that we need statistics of the input, estimated by segment, to find the optimal coefficients.

[redacted]

We get the set of normal equations again, but now the auto/cross correlation functions are computed in time, and with a finite number of points.

Question1: If the segment length is changed or moved in time does R stay the same?

Question2: What do we gain with an estimation that changes in time?

ESTIMATIONS OF THE AUTOCORRELATION FUNCTION

There are two basic methods: COVARIANCE and AUTOCORRELATION.

In Covariance the error is windowed. In Autocorrelation the data is windowed. This gives two different estimators.

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

AUTOCORRELATION

Assume that the the signal is windowed

$$x_n(m) = x(m+n)\,f(m) \qquad f(m) = \begin{cases} 1 & 0 \le m \le N-1 \\ 0 & \end{cases}$$

If the signal is nonzero only for $0 \le m \le N-1$, then the error for a M order filter (predictor) will be nonzero for $0 \le m \le N-1+M$, then,

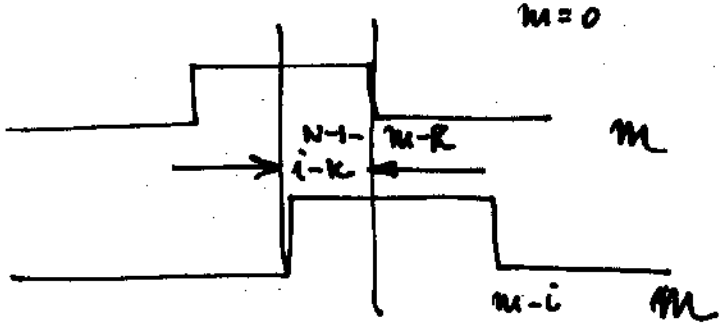
$$E_n = \sum_{m=0}^{N+M-1} e_n^2(m)$$

The limits in the summation of the function

$$\phi_n(i,k) = \sum_{m=0}^{N+M-1} x_n(m-i)\,x_n(m-k) \qquad \begin{matrix} 1 \le i \le M \\ 0 \le k \le M \end{matrix}$$

are the same as for the error. However, since the data is zero outside $0 \le m \le N-1$,

$$\phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} x_n(m)\,x_n(m+i-k) \qquad 1 \le i \le M \quad 0 \le k \le M$$



JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

FORMALLY $\ell = m - i$ $\begin{cases} m=0 \Rightarrow \ell = -i \\ m=0\ N+M-1 \Rightarrow \ell = N+M-1-i \end{cases}$

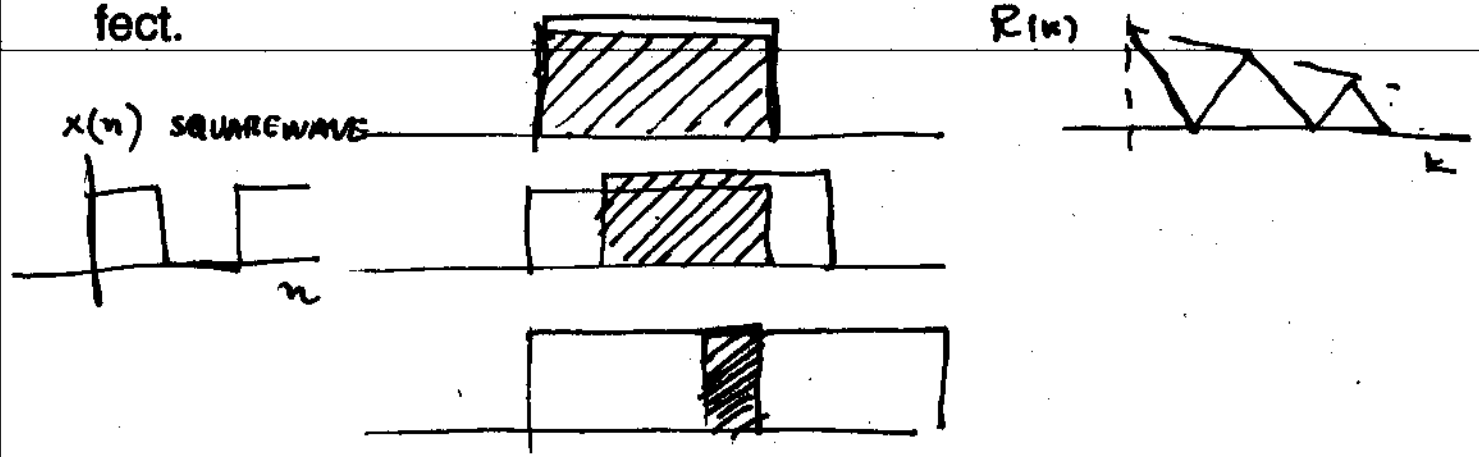
$$\sum_{\ell=-i}^{N+M-1-i} x(\ell)\, x(\ell+i-k) \Rightarrow \sum_{\ell=0}^{N+1-(i-k)} x(\ell)\, x(\ell+i-k)$$

This can be put in the form of the short term autocorrelation function, which is defined as the windowed autcorrelation fucntion.

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)\, w(n-m)\, x(m+k)\, w(n-k-m)$$

$$\Rightarrow \sum_{m=0}^{N-1-k} [x(n+m)\, w'(m)][x(n+m+k)\, w'(k+m)]$$

Notice that the short term autocorrelation function uses less and less data points for larger lags, which creates a fall-off effect.



JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

Notice that this estimation gives very large errors in the beginning and at the end of the window (predict nonzero values from zero and the converse). Therefore a window that tapers the data (such as the Hamming) is recommended.

HAMMING WINDOW

$$w(n) = \begin{cases} 0.54 - 0.46 \cos\left(2\pi n/(L-1)\right) & 0 \le n \le L-1 \\ 0 & \text{otherwise.} \end{cases}$$

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 99

904-335-8444

principe@brain.ee.ufl.edu

COVARIANCE

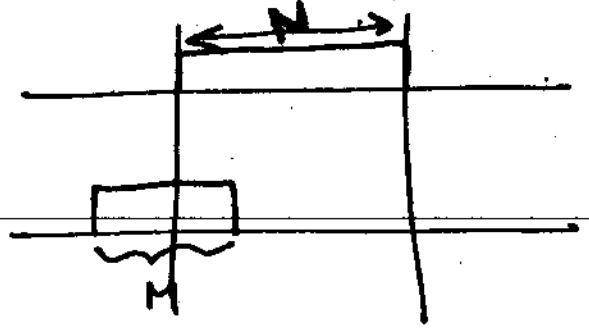
Here we are going to window (limit) the error calculation. We have than to see what are the effective data points that contribute to the error window.

$$E_n = \sum_{m=0}^{N-1} e_n^2(m)$$

Then

$$\phi_n(i,k) = \sum_{m=0}^{N-1} x_n(m-i)\, x_n(m-k) \qquad \begin{matrix} 1 \le i \le M \\ 0 \le k \le M \end{matrix}$$

If we change the index in the summation



$$m - k = \ell \quad \begin{cases} m = 0 \Rightarrow \ell = -k \\ m = N-1 \\ \ell = N-k-1 \end{cases}$$

one obtains

$$\phi(i,k) = \sum_{m=-k}^{N-k-1} x_n(m)\, x_n(m+k-i) \qquad \begin{matrix} 0 \le i \le M \\ 0 \le k \le M \end{matrix}$$

JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

Notice that for the calculation of the error we are using points outside the window (0-N), i.e. in the worst case $-M \leq m \leq N-1$. Now this function can be interpreted as a cross correlation function of two time series that are the same, but of different lengths.

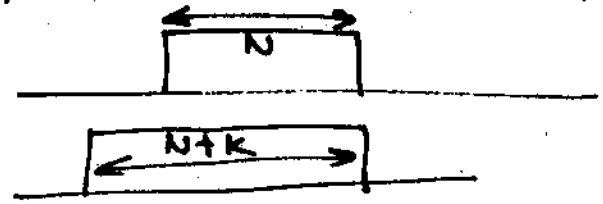
In fact the modified short term autocorrelation function is defined as

$$\hat{R}_M(k) = \sum_{n=-\infty}^{\infty} [x(n+m)\, w_1'(n)]\, [x(n+m+k)\, w_2'(n+k)]$$

where the window w_2' is larger than w_1' by M points. In this case the fall-off intrinsic to the short term autocorrleation function is eliminated.

$$\hat{R}_M(k) = \sum_{n=0}^{N-1} x(n+m)\, x(n+m+k)$$

Effectively we are making two different sequences (different lengths) from the same time series, so the name cross correlation.



JOSE C. PRINCIPE

UNIVERSITY OF FLORIDA

EEL 6935- SPRING 90

904-335-8444

principe@brain.ee.ufl.edu

The implication for the solution is that diagonal elements are different, but the matrix is still symmetric. No windowing is necessary, but the matrix is no longer Toeplitz.

$$\begin{bmatrix} \hat{R}_n(1,1) & \hat{R}_n(1,2) & \cdots & \hat{R}_n(1,M) \\ \hat{R}_n(2,1) & \hat{R}_n(2,2) & & \\ \vdots & & \ddots & \\ & & & \hat{R}_n(M,M) \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_M \end{bmatrix} = \begin{bmatrix} P_n(1) \\ \vdots \\ P_n(M) \end{bmatrix}$$

$$\hat{R}_n(1,2) = \hat{R}_n(2,1)$$