Prediction of Stationary Processes  
Tidsserieanalys SF2943  
Jan Grandell & Timo Koski

- LINEAR MINIMAL MEAN SQUARE ESTIMATION (LMMSE).
- LINEAR PREDICTION OF STATIONARY PROCESSES & TOEPLITZ MATRICES.
- DURBIN-LEVINSON & INNOVATIONS ALGORITHMS.
1 Introduction

This document covers linear prediction of stationary processes. The main theorem below can, however, be applied to other problems (e.g., filtering, smoothing, i.e., estimation of missing values) of linear estimation of random variables than merely to prediction (forecasting). We refer to the general estimation problem as LMMSE.

2 LMMSE

We are dealing with a sequence of random variables $Y_1, \ldots, Y_N$ and an additional random variable $X$. We assume everywhere in the sequel that $\text{Var}(X) < \infty$, with $\text{Var}(Y_i) < \infty$, for $i = 1, 2, \ldots, N$.

We assume for simplicity of writing that all these variables have expectations equal to zero. This restriction is lifted in subsection 4 below.

The problem is to estimate $X$ with a linear combination $a_1Y_1 + \ldots + a_NY_N$. This means that we must determine the parameters $a_1, \ldots, a_N$ in some manner. If we choose $a_1, \ldots, a_N$ so that

$$E \left[ (X - (a_1Y_1 + \ldots + a_NY_N))^2 \right]$$

(2.1)

is minimized, we talk about linear minimal mean square estimation of $X$ based on the zero mean sequence $Y_1, \ldots, Y_N$. We abridge linear minimal mean square estimation with LMMSE. We should clearly write

$$a_{1,N}Y_1 + \ldots + a_{N,N}Y_N$$

because the coefficients may change with $N$, too, but we refrain from this for the moment. We can also call

$$Z = a_1Y_1 + \ldots + a_NY_N$$

(2.2)

the linear mean square estimator. $\hat{X}$ is the minimal mean square estimator.
3 Main theorems for LMMSE

3.1 Orthogonality and the main theorem

Theorem 3.1 [THE ORTHOGONALITY PRINCIPLE] Assume that \( Y_1, \ldots, Y_N \) and \( X \) are random variables with zero means such that

\[
\gamma_{mk} = \text{Cov}(Y_m, Y_k) = E[Y_mY_k], \quad m = 1, \ldots, N; \quad k = 1, \ldots, N
\]

(3.1)

\[
\gamma_{om} = \text{Cov}(Y_m, X) = E[Y_mX], \quad m = 1, \ldots, N.
\]

If \( a_1, \ldots, a_N \) are solutions to the linear system of equations

\[
\sum_{k=1}^{N} a_k \gamma_{mk} = \gamma_{om}; \quad m = 1, \ldots, N,
\]

(3.2)

then it holds that

\[
E\left[Y_m (X - (a_1 Y_1 + \ldots + a_N Y_N))\right] = 0; \quad m = 1, 2, \ldots, N.
\]

(3.3)

Proof: We expand the left hand side of (3.3) by well known rules for manipulating expectation and get

\[
E\left[Y_m (X - (a_1 Y_1 + \ldots + a_N Y_N))\right] = E[Y_mX] - a_1 E[Y_mY_1] - \ldots - a_N E[Y_mY_N] =
\]

\[
= \gamma_{om} - a_1 \gamma_{m1} - \ldots - a_N \gamma_{mN} = \gamma_{om} - \sum_{k=1}^{N} a_k \gamma_{mk} = 0
\]

in view of (3.1).

The orthogonality principle in (3.3) is formulated in words as the statement that 'the estimation error should be orthogonal to data' in LMMSE and is illustrated in figure 3.1.
Here we may think of a linear space of random vectors, where the correlation defines a scalar product $\langle X, Y \rangle$ and the associated norm $\|X\|$ by
\[
\langle X, Y \rangle = E[X \cdot Y], \quad \|X\| = \sqrt{\langle X, X \rangle} = \sqrt{E[X^2]}.
\]
This leads to the main theorem.

**Theorem 3.2** [MAIN THEOREM] Assume that $Y_1, \ldots, Y_N$ and $X$ are random variables with zero means and covariances as in the preceding theorem. Then the expression
\[
E \left[ (X - (a_1 Y_1 + \ldots + a_N Y_N))^2 \right]
\]
(3.4)
is minimized if the coefficients $a_1, \ldots, a_N$ satisfy the system of equations
\[
\sum_{k=1}^{N} a_k \gamma_{mk} = \gamma_{om}; \quad m = 1, \ldots, N.
\]
(3.5)

**Proof:** Let $b_1 Y_1 + \ldots + b_N Y_N$ be a linear combination of $Y_1, \ldots, Y_N$ with arbitrary coefficients $b_1, \ldots, b_N$. We write also
\[
\hat{X} = \sum_{k=1}^{N} a_k Y_k.
\]
Then we can write the estimation error using $b_1 Y_1 + \ldots + b_N Y_N$ as

$$
\varepsilon = X - (b_1 Y_1 + \ldots + b_N Y_N) = \left( X - \hat{X} \right) + \sum_{k=1}^{N} (a_k - b_k) Y_k
$$

We square this identity and compute expectations and get

$$
E[\varepsilon^2] = E\left[ \left( X - \hat{X} \right)^2 \right] + E\left[ \left( \sum_{k=1}^{N} (a_k - b_k) Y_k \right)^2 \right] + 2 \sum_{k=1}^{N} (a_k - b_k) E\left[ Y_k \cdot \left( X - \hat{X} \right) \right].
$$

But the orthogonality principle (3.3), which holds by assumption (3.5), gives that $E\left[ Y_k \cdot \left( X - \hat{X} \right) \right] = 0$ for every $k$. Hence we get

$$
E[\varepsilon^2] = E\left[ \left( X - \hat{X} \right)^2 \right] + E\left[ \left( \sum_{k=1}^{N} (a_k - b_k) Y_k \right)^2 \right] \geq E\left[ \left( X - \hat{X} \right)^2 \right],
$$

since the omitted term $E\left[ \left( \sum_{k=1}^{N} (a_k - b_k) Y_k \right)^2 \right]$ is non-negative. Thus

$$
E\left[ \left( X - \sum_{k=1}^{N} b_k Y_k \right)^2 \right] \geq E\left[ \left( X - \hat{X} \right)^2 \right].
$$

which proves the assertion, since $b_1, \ldots, b_N$ are arbitrary.

**3.2 Estimation Error**

**Theorem 3.3** [OPTIMAL ESTIMATION ERROR VARIANCE] Let

$$
\hat{X} = \sum_{k=1}^{N} a_k Y_k
$$

be the minimal mean square estimator. Then minimal mean square error variance is

$$
\sigma^2_{\text{min}} = E\left[ \left( X - \hat{X} \right)^2 \right] = E[X^2] - \sum_{k=1}^{N} a_k \gamma_{ok}. \quad (3.6)
$$
Proof:

\[
E \left[ (X - \hat{X})^2 \right] = E \left[ (X - \hat{X}) \left( X - \sum_{k=1}^{N} a_k Y_k \right) \right] = \\
E \left[ (X - \hat{X}) X \right] - E \left[ (X - \hat{X}) \sum_{k=1}^{N} a_k Y_k \right] = \\
E \left[ (X - \hat{X}) X \right] - \sum_{k=1}^{N} a_k E \left[ (X - \hat{X}) Y_k \right].
\]

In the last expression \( E \left[ (X - \hat{X}) Y_k \right] = 0 \) for every \( k \) by the orthogonality principle. Thus

\[
E \left[ (X - \hat{X})^2 \right] = E \left[ (X - \hat{X}) X \right] = E \left[ X^2 \right] - E \left[ \hat{X} X \right] = \\
E \left[ X^2 \right] - \sum_{k=1}^{N} a_k E \left[ Y_k X \right] = E \left[ X^2 \right] - \sum_{k=1}^{N} a_k \gamma_{ok}.
\]
as was to be proved. \( \blacksquare \)

3.3 A matrix solution for LMMSE

From the above we see that LMMSE reduces to two operations:

1. Computation of the covariances in (3.1)

2. Solving the equations (3.2) with respect to \( a_1, \ldots, a_N \)

We can formulate the equations in (3.2) with a piece of linear algebra. Let

\[
a = (a_1, \ldots, a_N)^T
\]

be a column vector (\(^T\) is the transpose) and

\[
\gamma_{XY} = \left( \gamma_{o1}, \ldots, \gamma_{oN} \right)^T.
\]
The \( N \times N \) covariance matrix for \( Y_1, \ldots, Y_N \) is

\[
\Gamma_{YY} := \begin{pmatrix}
\gamma_{11} & \gamma_{12} & \cdots & \gamma_{1N} \\
\gamma_{21} & \gamma_{22} & \cdots & \gamma_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
\gamma_{N1} & \gamma_{N2} & \cdots & \gamma_{NN}
\end{pmatrix}
\] (3.7)
Note that this a symmetric matrix. Then we can write the system of equations in (3.2) as
\[ \Gamma_{YY}a = \gamma_{XY}. \] (3.8)
If \( \Gamma_{YY} \) is invertible (has full rank = \( N \)), then
\[ a = \Gamma_{YY}^{-1}\gamma_{XY}. \] (3.9)
As is well-known, and easily seen, a covariance matrix \( \Gamma_{YY} \) is non-negative definite. If \( \Gamma_{YY} \) is positive definite, then \( \Gamma_{YY}^{-1} \) exists.

4 LMMSE with non-zero means

Let us now return to the general case where the means are non-zero. We consider the case where
\[ E[X] = E[Y_1] = \cdots = E[Y_N] = \mu. \]
We have in the preceding found \( a_1, \ldots, a_N \) to estimate \( (X - \mu) \) so that
\[ E \left[ \left( (X - \mu) - \sum_{i=1}^{N} a_i (Y_i - \mu) \right)^2 \right] \] (4.1)
is minimized, i.e., \( a_1, \ldots, a_N \) satisfy (3.2). We can write this as
\[ E \left[ \left( X - \left( \mu + \sum_{i=1}^{N} a_i (Y_i - \mu) \right) \right)^2 \right] \] (4.2)
But then we can write the estimator of \( X \) as
\[ \hat{X} = \mu + \sum_{i=1}^{N} a_i (Y_i - \mu). \] (4.3)
Thus the best linear estimator of \( X \) can be found by adding \( \mu \) to the best linear predictor of \( (X - \mu) \) based on linear combinations of \( Y_1 - \mu, \ldots, Y_N - \mu \). Next we can write
\[ \hat{X} = \mu - \mu \cdot \sum_{i=1}^{N} a_i + \sum_{i=1}^{N} a_i Y_i. \]
Let us set

\[ a_0 = \mu \cdot \left( 1 - \sum_{i=1}^{N} a_i \right). \] (4.4)

Then we can write the optimal linear estimator of \( X \) as

\[ \hat{X} = a_0 + \sum_{i=1}^{N} a_i Y_i, \] (4.5)

where \( a_1, \ldots, a_N \) satisfy (3.2) and \( a_0 \) satisfies (4.4).

We can also regard (4.5) as the optimal mean square estimator of \( X \) based on linear combinations of \( 1, Y_1, \ldots, Y_N \). We set \( P_N X = \hat{X} \)

\[ P_N X = a_0 + \sum_{i=1}^{N} a_i Y_i \] (4.6)

to denote the best linear mean square estimator of \( X \) based on linear combinations of \( 1, Y_1, \ldots, Y_N \). We summarize this as

**Theorem 4.1** Assume that \( Y_1, \ldots, Y_N \) and \( X \) are random variables with means

\[ E[X] = E[Y_1] = \cdots = E[Y_N] = \mu \]

and \( \text{Var}(X) < \infty, \text{Var}(Y_i) < \infty, \) for \( i = 1, 2, \ldots, N \). Then the best linear mean square estimator of \( X \) can be found by adding \( \mu \) to the best linear predictor of \( (X - \mu) \) based on linear combinations of \( Y_1 - \mu, \ldots, Y_N - \mu \).

Equivalently, the optimal linear estimator of \( X \) is

\[ \hat{X} = a_0 + \sum_{i=1}^{N} a_i Y_i, \] (4.7)

where \( a_1, \ldots, a_N \) satisfy (3.2) and \( a_0 \) satisfies (4.4). This is also the best linear mean square estimator of \( X \) based on linear combinations of \( 1, Y_1, \ldots, Y_N \).

Hence we can without restriction assume that the mean of a stationary times series is zero.
5 Some Simple Examples

Example 5.1 Let us consider LMMSE on basis of one variable \((N = 1)\), where all means are zero or minimization of

\[
E \left[ (X - aY)^2 \right].
\]

We determine \(a\). Then the formula (3.9) boils down to

\[
a = \frac{Cov(X, Y)}{Var(Y)}.
\]

This we can rewrite using the correlation coefficient \(\rho_{X,Y}\) between \(X\) and \(Y\) as

\[
a = \rho_{X,Y} \cdot \frac{\sqrt{Var(X)}}{\sqrt{Var(Y)}}.
\]

Hence the LMMSE is

\[
\hat{X} = \rho_{X,Y} \cdot \frac{\sqrt{Var(X)}}{\sqrt{Var(Y)}} \cdot Y.
\]

The error variance is from (3.6)

\[
\sigma^2_{min} = E \left[ X^2 \right] - a \cdot \gamma_{o1} = E \left[ X^2 \right] - \rho_{X,Y} \cdot \frac{\sqrt{Var(X)}}{\sqrt{Var(Y)}} \cdot Cov(X, Y) =
\]

\[
= E \left[ X^2 \right] - \rho_{X,Y} \cdot \frac{Cov(X, Y)}{\sqrt{Var(X)} \sqrt{Var(Y)}} Var(X) =
\]

\[
= E \left[ X^2 \right] - \rho^2_{X,Y} \cdot Var(X) = Var(X) \left( 1 - \rho^2_{X,Y} \right).
\]

since here \(E \left[ X^2 \right] = Var(X)\) by assumption of zero means so that

\[
\sigma^2_{min} = Var(X) \left( 1 - \rho^2_{X,Y} \right).
\]

Example 5.2 Let us note now that if \((X, Y)\) have a two-dimensional normal distribution, then the conditional distribution of for \(X\) given \(Y = y\) is the normal distribution

\[
N \left( \mu_1 + \rho \cdot \frac{\sigma_1}{\sigma_2} (y - \mu_2), \sigma_1^2 (1 - \rho^2) \right).
\]
where \( \mu_1 = E(X) \), \( \mu_2 = E(Y) \), \( \sigma_1 = \sqrt{\text{Var}(X)} \), \( \sigma_2 = \sqrt{\text{Var}(Y)} \) och \( \rho = \rho_{X,Y} \).

By comparison with example 5.1 and (5.3) we see, using (4.3), that the conditional expectation \( E[X|Y = y] = \mu_1 + \rho \cdot \frac{\sigma_1}{\sigma_2} (y - \mu_2) \) is the LMMSE of \( X \) given \( Y = y \). The conditional variance \( \sigma_1^2 (1 - \rho^2) \) equals in view of (5.4) the variance of the minimal estimation error.

\[ \]

6 Predicting of a Stationary Time Series

6.1 The linear minimal mean square \( h \)-step predictor

Let \( \{X_n \mid n = 0, \pm 1, \pm 2, \ldots, \} \) be a stationary process with mean function equal to zero and with autocovariance function \( \gamma_X(k) \). As the autocovariance function of a stationary process this satisfies

\[ \gamma_X(k) = \gamma_X(-k). \] (6.1)

Our conventions are that

\[ E[X_m \cdot X_{m+k}] = \gamma_X(k) = \gamma_X(-k) = E[X_{m+k} \cdot X_m]. \]

We consider the problem of estimating, or predicting, the value \( X_{n+h}, h > 0 \), in terms of the \( n \) most recent values of that process at time \( n \). Our predictor is of the form

\[ P_n X_{n+h} = a_1 X_n + \ldots + a_n X_1 \] (6.2)

We want to find \( a_1, \ldots, a_n \) so that the mean square prediction error

\[ E[(X_{n+h} - P_n X_{n+h})^2] \]

is minimized.

In our generic formulation (2.2) above we take \( N = n, X = X_{n+h}, Y_1 = X_n, \ldots, Y_N = X_1 \) or in one formula

\[ Y_i = X_{n+1-i}, i = 1, \ldots, n, \] (6.3)

so that

\[ P_n X_{n+h} = a_1 X_n + \ldots + a_n X_1 = \sum_{i=1}^{n} a_i X_{n+1-i}. \]
We use (3.1) and (6.3)

\[ \gamma_{mk} = E[Y_m Y_k] = E[X_{n+1-m} X_{n+1-k}] = \gamma_X(m - k), \]

\[ m = 1, \ldots, n; k = 1, \ldots, n \]

(6.4)

\[ \gamma_{om} = E[Y_m X] = E[X_{n+1-m} X_{n+h}] = \gamma_X(h + m - 1), m = 1, \ldots, n. \]

Thus, if \( a_1, \ldots, a_n \) are solutions to the linear system of equations (3.2), now expressed as

\[ \sum_{k=1}^{n} a_k \gamma_X(m - k) = \gamma_X(h + m - 1); m = 1, \ldots, n, \]

(6.5)

then we have found the best \( h \)-step predictor in minimal mean square sense.

We want to solve these equations in a useful matrix formulation. Since the process is stationary we have \( \gamma_X(m - k) = \gamma_X(k - m) \). Then the covariance matrix (3.7) corresponding to (6.5) is by (6.4) equal to

\[
\begin{pmatrix}
\gamma_X(0) & \gamma_X(1) & \cdots & \gamma_X(n-2) & \gamma_X(n-1) \\
\gamma_X(1) & \gamma_X(0) & \cdots & \gamma_X(n-2) & \gamma_X(n-1) \\
\gamma_X(2) & \gamma_X(1) & \cdots & \gamma_X(n-3) & \\
\vdots & \vdots & \ddots & \vdots & \\
\gamma_X(n-2) & \gamma_X(n-3) & \cdots & \cdots & \gamma_X(1) \\
\gamma_X(n-1) & \gamma_X(n-2) & \cdots & \cdots & \gamma_X(0)
\end{pmatrix}
\]  

(6.6)

In addition we set

\[ \gamma_n(h) = (\gamma_X(h), \gamma_X(h + 1), \ldots, \gamma_X(h + n - 1))^T. \]

Then (6.5) becomes

\[ \Gamma_n a = \gamma_n(h). \]

(6.7)

If \( \Gamma_n \) is invertible (has full rank = \( n \)), then

\[ a = \Gamma_n^{-1} \gamma_n(h). \]

(6.8)

We introduce also the process \( \{ \varepsilon_n \mid n = 0, \pm 1, \pm 2, \ldots \} \) of \( h \)-step prediction errors

\[ \varepsilon_n = X_{n+h} - P_n X_{n+h} = X_{n+h} - (a_1 X_n + \ldots + a_n X_1). \]
We introduce also
\[ \sigma^2_\epsilon = E[\epsilon_n^2]. \]

Then the minimal error variance is by (3.6)
\[ \sigma^2_\epsilon = E[X_{n+h}^2] - E[X_{n+h} \cdot P_nX_{n+h}], \]
(6.9)
\[ = \gamma_X(0) - \sum_{i=1}^{n} a_i \gamma_X(h + i - 1). \]
(6.10)

6.2 Toeplitz matrices

The covariance matrix for a finite \( p \)-segment of a stationary process \( \{X_n \mid n = 0, \pm 1, \pm 2, \ldots \} \) with \( \text{Var}(X_n) < \infty \), is in general given by

\[
\begin{pmatrix}
\gamma_X(0) & \gamma_X(1) & \ldots & \gamma_X(p-2) & \gamma_X(p-1) \\
\gamma_X(1) & \gamma_X(0) & \ldots & \gamma_X(p-3) & \gamma_X(p-2) \\
\gamma_X(2) & \gamma_X(1) & \gamma_X(0) & \ldots & \gamma_X(p-4) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\gamma_X(p-2) & \gamma_X(p-3) & \ldots & \gamma_X(1) \\
\gamma_X(p-1) & \gamma_X(p-2) & \ldots & \gamma_X(1) & \gamma_X(0)
\end{pmatrix}.
\]

This is a symmetric matrix with a special property: the elements along every diagonal are identical. Matrices with this property are known as Toeplitz matrices. A useful function exploiting this is implemented in MATLAB\textsuperscript{TM}.

The following is an edited excerpt of a MATLAB\textsuperscript{TM} -session.

\[
\text{>> help toeplitz}
\]
\[
\text{>> TOEPLTZX Toeplitz matrix. TOEPLTZX(\Gamma) is a symmetric (or Hermitian) Toeplitz matrix.}
\]
\[
\text{>> r = [ 0.5 0.2 0.5 1 ] .}
\]
\[
\text{>> toeplitz(r)}
\]
\[
\text{ans = }
\begin{pmatrix}
0.5000 & 0.2000 & 0.5000 & 1.0000 \\
0.2000 & 0.5000 & 0.2000 & 0.5000 \\
0.5000 & 0.2000 & 0.5000 & 0.2000 \\
1.0000 & 0.5000 & 0.2000 & 0.5000
\end{pmatrix}.
\]

There are a number of algorithms for inverting \( \Gamma_p \) that take advantage of the Toeplitz property, e.g. Levinson - Durbin.
6.3 Innovations

Let us next consider the one-step predictor of a stationary process \((h = 1\) in \((6.2)\)). Then we refer to the random variables \(\varepsilon_n\) in

\[
\varepsilon_n = X_{n+1} - p_n X_{n+1} = X_{n+h} - (a_1 X_n + \ldots + a_n X_1).
\]

\[
(6.11)
\]
as innovations. We can write

\[
X_{n+1} = a_1 X_n + \ldots + a_n X_1 + \varepsilon_n.
\]

The meaning of innovation is that the new information in the process at \(n\) is not \(X_n\) but \(\varepsilon_n\).

7 Some Simple Examples of Prediction and estimation of Stationary Processes

**Example 7.1** [LINEAR PREDICTION] The process \(\{X_n \mid n = 0, \pm 1, \pm 2, \ldots\}\) is stationary with mean function \(E[X_n] = 0\) and autocovariance function

\[
\gamma_X(h) = \phi^{|h|}, h = 0, \pm 1, \pm 2, \ldots
\]

\[
(7.1)
\]

where \(|\phi| < 1\). We wish to use the current value of the process to predict (to estimate) future values. We formulate this in terms of the LMMSE theory above. We set \(X = X_{n+k}\) and \(Y = X_n\) and \(N = 1\). Thus we wish to predict the process \(k > 0\) steps ahead using the value of the process at time \(n\) using the linear minimal mean square criterion. We denote the predictor by \(\hat{X}_{n+k}\).

We have actually already once solved the problem in example 5.1, and get from \((5.3)\) the result

\[
\hat{X}_{n+k} = \rho_{X_{n+k},X_n} \cdot \frac{\sqrt{Var(X_{n+k})}}{\sqrt{Var(X_n)}} \cdot X_n = \phi^k X_n,
\]

\[
(7.2)
\]

\[\text{We can represent } \{X_n \mid n = 0, \pm 1, \pm 2, \ldots\} \text{ as}
\]

\[
X_{n+1} = \phi X_n + Z_n,
\]

where \(\{Z_n \mid n = 0, \pm 1, \pm 2, \ldots\}\) is WN\((0, \sigma^2_Z)\), and, since \(|\phi| < 1\),

\[
\sigma^2_Z = 1 - \phi^2.
\]
since $\text{Var}(X_{n+k}) = \text{Var}(X_n) = \gamma_X(0) = 1$ and $\text{Cov}(X_{n+k}, X_n) = \gamma_X(-k) = \phi^k$ for $k > 0$. The prediction error variance is from (5.4)

$$\sigma^2_{\text{min}} = 1 - \phi^{2k}.$$ (7.3)

From (7.2) and (7.3) we see that if $k$ is large, then the predictor $c^k X_n$ should often be small, close to $E[X_{n+k}] = 0$ and the error variance close to its maximum value. This expresses the reasonable idea that the reliability of the prediction should decrease for prediction large steps ahead. If $\phi$ is close to zero, then again the prediction is close to $E[X_{n+k}] = 0$, the a priori prediction. If $|\phi|$ is close to 1, then the predictor is highly correlated with the predicted variable.


Example 7.2 [LINEAR ONE-STEP PREDICTION OF AR(1)] The process $\{X_n \mid n = 0, \pm 1, \pm 2, \ldots, \}$ is stationary AR(1) with mean function $E[X_n] = 0$

$$X_{n+1} = \phi X_n + Z_n,$$ (7.4)

where $|\phi| < 1$ and $\{Z_n \mid n = 0, \pm 1, \pm 2, \ldots, \}$ is WN(0, $\sigma^2_Z$). Hence the autocovariance function is

$$\gamma_X(h) = \frac{\sigma^2_Z}{1 - \phi^2} \phi^{|h|}, h = 0, \pm 1, \pm 2, \ldots$$ (7.5)

We want to predict $X_{n+1}$ by

$$P_n X_{n+1} = a_1 X_n + \ldots + a_n X_1,$$ (7.6)

where we use $a_1, \ldots, a_n$ so that the mean square prediction error

$$E[(X_{n+1} - P_n X_{n+1})^2]$$

is minimized. This differs from example 7.1 above in that we do not from the outset restrict the predictor to consist of just one variable. However, we shall now show that for an AR(1) process the optimal values of the coefficients are in fact $a_1 = \phi$, $a_2 = \ldots = a_n = 0$. We need to solve

$$\Gamma_n a = \gamma_n(1).$$ (7.7)
The pertinent Toeplitz matrix is now

$$
\Gamma_n = \frac{\sigma_Z^2}{1 - \phi^2} \begin{pmatrix}
1 & \phi & \ldots & \phi^{n-2} & \phi^{n-1} \\
\phi & 1 & \phi & \ldots & \phi^{n-2} \\
\phi^2 & \phi & 1 & \ldots & \phi^{n-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\phi^{n-2} & \phi^{n-3} & \ldots & \ldots & \phi \\
\phi^{n-1} & \phi^{n-2} & \ldots & \phi & 1
\end{pmatrix}
$$

(7.8)

and

$$
\gamma_n(1) = \frac{\sigma_Z^2}{1 - \phi^2} (\phi, \phi^2, \ldots, \phi^n)^T.
$$

The first equation in the system of equations (7.7) is

$$
a_1 + \phi a_2 + \ldots + \phi^{n-2} a_{n-1} + \phi^{n-1} a_n = \phi
$$

and $a_1 = \phi, a_2 = \ldots = a_{n-1} = a_n = 0$ satisfies this equation. It is straightforward to check that actually

$$
a_1 = \phi, a_2 = \ldots = a_{n-1} = a_n = 0
$$

is a solution of all of (7.7). Hence an optimal one-step predictor of a stationary AR(1)-process is

$$
P_n X_{n+1} = \phi X_n.
$$

There are easier ways to find this than the one given above. We have thus that the innovations are

$$
\varepsilon_n = X_{n+1} - P_n X_{n+1} = X_{n+1} - \phi X_n = Z_n.
$$

The minimal one-step prediction error variance (variance of the innovations) is from (6.9)

$$
\sigma_\varepsilon^2 = \frac{\sigma_Z^2}{1 - \phi^2} - \phi \frac{\sigma_Z^2}{1 - \phi^2} \phi
$$

$$
= \sigma_Z^2 \frac{1 - \phi^2}{1 - \phi^2} = \sigma_Z^2.
$$
Example 7.3 \textit{[LINEAR FILTERING]} The process \( \{X_n \mid n = 0, \pm 1, \pm 2, \ldots, \} \) is stationary with mean \( E[X_n] = 0 \) and autocovariance function

\[
\gamma_X(h) = \phi^{|h|} h = 0, \pm 1, \pm 2, \ldots
\]  

(7.9)

where \( |\phi| < 1 \). The random sequence \( \{Y_n \mid n = 0, \pm 1, \pm 2, \ldots, \} \) is a noisy measurement of (a signal) \( \{X_n \mid n = 0, \pm 1, \pm 2, \ldots, \} \) or

\[ Y_n = X_n + Z_n, \]

where \( \{Z_n \mid n = 0, \pm 1, \pm 2, \ldots, \} \) is WN(0, \( \sigma_Z^2 \)) measurement noise and is independent of \( \{X_n \mid n = 0, \pm 1, \pm 2, \ldots, \} \). We wish to estimate \( X_n \) using \( Y_n \). This is known as a filtering estimate of a signal in noise. The generic LMMSE formulation is \( X = X_n \) and \( Y = Y_n \). We have in principle done the problem twice already, and the result is

\[
\hat{X}_n = \rho_{X_n,Y_n} \cdot \frac{\sqrt{\text{Var}(X_n)}}{\sqrt{\text{Var}(Y_n)}} \cdot Y_n = \frac{1}{1 + \sigma_Z^2} Y_n, \quad (7.10)
\]

since \( \text{Var}(X_n) = 1, \text{Var}(Y_n) = \text{Var}(X_n) + \text{Var}(Z_n) = 1 + \sigma_Z^2 \) and

\[
\text{Cov}(X_n, Y_n) = E[X_n \cdot Y_n] = E[X_n \cdot X_n] + E[X_n \cdot Z_n] = 1 + 0 = 1.
\]

The optimal error variance is

\[
\sigma_{\min}^2 = 1 - \frac{1}{1 + \sigma_Z^2}.
\]

\[ \blacksquare \]

Example 7.4 \textit{[LINEAR PREDICTION OF AN MA(1) PROCESS]} Let \( \{Z_n \mid n = 0, \pm 1, \pm 2, \ldots, \} \) be WN(0, 1) and set

\[
X_n = Z_n + \frac{1}{2} Z_{n-1}. \quad (7.11)
\]

This determines an MA(1) - process. Then we have (check this) the autocovariance function

\[
\gamma_X(k) = \begin{cases} 
\frac{3}{2} & k = 0 \\
\frac{1}{2} & k = \pm 1 \\
0 & k \text{ otherwise}.
\end{cases}
\]
We wish to predict $X_n$ using

$$P_2 X_n = a_1 X_{n-1} + a_2 X_{n-2}$$

and LMMSE to determine $a_1$ and $a_2$. This can be done by (6.8). From $\gamma_X(k)$ we obtain

$$\Gamma_2 = \begin{pmatrix} \frac{5}{2} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

and

$$\gamma_2(1) = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}.$$

(Check these !) Then

$$a = \Gamma_2^{-1} \gamma_2 = \begin{pmatrix} 0.9524 & -0.3810 \\ -0.3810 & 0.9524 \end{pmatrix} \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix} = \begin{pmatrix} 0.4762 \\ -0.1905 \end{pmatrix}.$$  \hspace{1cm} (7.12)

8  The Durbin-Levinson & Innovations Algorithm

8.1 Operator Formalism and Linear Span

For any given $Y_1, \ldots, Y_N$ and $X$ with finite second moments we have in (4.3) and (3.2) shown how to compute the minimal mean square linear estimate of $X$ in terms of $1, Y_1, \ldots, Y_N$. We assume now that the means are zero.

The linear span of $Y_1, \ldots, Y_N$ is defined as the set of all linear combinations of $Y_1, \ldots, Y_N$. We denote the linear span by

$$\overline{sp} \{ Y \}.$$

Then the minimal mean square estimate

$$\hat{X} = \sum_{i=1}^{N} a_i Y_i.$$
is given a new formal symbol by

\[ P \left( X \mid \mathbf{sp} \{ Y \} \right) = \hat{X}. \]

We regard \( P \left( X \mid \mathbf{sp} \{ Y \} \right) \) as a **prediction operator**

\[ X \mapsto P \left( X \mid \mathbf{sp} \{ Y \} \right) \in \mathbf{sp} \{ Y \}. \]

Then we can also consider, for example,

\[ \mathcal{K}_1 = \mathbf{sp} \{ Y_2, \ldots, Y_N \} \quad (8.13) \]

and

\[ \mathcal{K}_2 = \mathbf{sp} \{ Y_1 - P (Y_1 \mid \mathcal{K}_1) \}. \quad (8.14) \]

Then \( P (X \mid \mathcal{K}_1) \in \mathcal{K}_1 \) and \( P (X \mid \mathcal{K}_2) \in \mathcal{K}_2 \), it holds by the orthogonality principle in Theorem 3.1, i.e., (3.3), that

\[ E \left[ P (X \mid \mathcal{K}_1) \cdot P (X \mid \mathcal{K}_2) \right] = 0. \quad (8.15) \]

To check this, let us note that

\[ P (X \mid \mathcal{K}_2) = a \left( Y_1 - P (Y_1 \mid \mathcal{K}_1) \right) \]

for some (optimal) \( a \), and

\[ P (X \mid \mathcal{K}_1) = \sum_{i=2}^{N} a_i Y_i. \]

Here by the orthogonality principle, i.e., (3.3)

\[ E [Y_i (Y_1 - P (Y_1 \mid \mathcal{K}_1))] = 0, i = 2, 3, \ldots, N, \]

and we have verified (8.15).

Then we can write for any \( X \), with \( E [X^2] < \infty \), as

\[ P (X \mid \mathbf{sp} \{ Y \}) = P (X \mid \mathcal{K}_1) + P (X \mid \mathcal{K}_2). \quad (8.16) \]
8.2 The Durbin-Levinson Recursion

If \( \{X_n \mid n = 0, \pm 1, \pm 2, \ldots \} \) is a zero-mean stationary time series. We express the best linear predictor \( \hat{X}_{n+1} \) of \( X_{n+1} \) in terms of \( X_1, X_2, \ldots, X_n \) is

\[
\hat{X}_{n+1} = P(X \mid s_p\{X_1, X_2, \ldots, X_n\}) = \sum_{i=1}^{n} a_{n,i} X_{n+1-i}, \quad n = 1, 2, \ldots,
\]

where

\[
a_n = \begin{pmatrix} a_{n,1} \\ \vdots \\ a_{n,n} \end{pmatrix} = \Gamma^{-1}_n \gamma_n, \quad \gamma_n = \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(n) \end{pmatrix}
\]

If \( \gamma(0) > 0 \) and \( \gamma(h) \to 0 \) as \( h \to \infty \), then all \( \Gamma_n \) are invertible (the proof is omitted). The mean-squared error is

\[
v_n = E|X_{n+1} - \hat{X}_{n+1}|^2 = \gamma(0) - a_n^T \gamma_n = \gamma(0) - \gamma_n^T \Gamma_n^{-1} \gamma_n.
\]

If \( n \) is small, this works well. If \( n \) is very large, one may consider approximations. If \( n \) is of moderate size, computations of \( \Gamma_n^{-1} \) may be rather difficult.

We shall now consider recursive methods, i.e. at time \( n - 1 \) we know \( X_1, X_2, \ldots, X_{n-1} \) and have computed \( \hat{X}_n \). When we then get information of \( X_n \) we want to compute \( \hat{X}_{n+1} \) based on \( X_1, X_2, \ldots, X_n \). The real new information is the innovation \( X_n - \hat{X}_n \) rather than \( X_n \).

Assume now that we have computed \( \hat{X}_n \), which really means that we know \( a_{n-1} \) and \( v_{n-1} \). We have the following algorithm.

**Theorem 8.1 (The Durbin–Levinson Algorithm)** If \( \{X_n \mid n = 0, \pm 1, \pm 2, \ldots \} \) is a zero-mean stationary time series such that \( \gamma(0) > 0 \) and \( \gamma(h) \to 0 \) as \( h \to \infty \), then \( a_{1,1} = \gamma(1)/\gamma(0) \), \( v_0 = \gamma(0) \),

\[
a_{n,n} = \left[ \gamma(n) - \sum_{j=1}^{n-1} a_{n-1,j} \gamma(n-j) \right] v_{n-1}^{-1}
\]

\[
\begin{pmatrix} a_{n,1} \\ \vdots \\ a_{n,n-1} \end{pmatrix} = \begin{pmatrix} a_{n-1,1} \\ \vdots \\ a_{n-1,n-1} \end{pmatrix} - a_{n,n} \begin{pmatrix} a_{n-1,n-1} \\ \vdots \\ a_{n-1,1} \end{pmatrix}
\]

and

\[
v_n = v_{n-1} [1 - a_{n,n}^2].
\]
**Proof:** We set
\[ \mathbb{P} \{ Y \} = \mathbb{P} \{ X_1, X_2, \ldots, X_n \}, \]
\[ K_1 = \mathbb{P} \{ X_2, \ldots, X_n \} \]
and
\[ K_2 = \mathbb{P} \{ X_1 - P(X_1 | K_1) \}. \]

Then, as argued above, we can write for \( X_{n+1} \)
\[ P(X_{n+1} | \mathbb{P} \{ Y \}) = P(X_{n+1} | K_1) + P(X_{n+1} | K_2). \] (8.17)

For simplicity of writing we set
\[ \hat{X}_{n+1} = P(X_{n+1} | \mathbb{P} \{ Y \}). \] (8.18)

Then we have
\[ \hat{X}_{n+1} = P(X_{n+1} | K_1) + a(X_1 - P(X_1 | K_1)). \] (8.19)

Here we have (c.f., (5.2))
\[ a = \frac{\text{Cov}(X_{n+1}, X_1 - P(X_1 | K_1))}{\text{Var}(X_1 - P(X_1 | K_1))}. \] (8.20)

Now, by stationarity, \((X_1, X_2, \ldots, X_n)^T\) has the same covariance matrix as \((X_n, X_{n-1}, \ldots, X_1)^T\) and \((X_2, X_3, \ldots, X_{n+1})^T\) so that
\[ P(X_1 | K_1) = \sum_{i=1}^{n-1} a_{n-1,i} X_{i+1} \] (8.21)
and
\[ P(X_{n+1} | K_1) = \sum_{i=1}^{n-1} a_{n-1,i} X_{n+1-i} \] (8.22)

Then
\[ \text{Var}[(X_1 - P(X_1 | K_1))] = E[(X_1 - P(X_1 | K_1))^2] = E[(X_{n+1} - P(X_{n+1} | K_1))^2] = E\left[\left(X_n - \hat{X}_n\right)^2\right] = v_{n-1}. \] (8.23)

From the equations (8.19), (8.21) and (8.22) we get
\[ \hat{X}_{n+1} = aX_1 + \sum_{i=1}^{n-1} (a_{n-1,i} - aa_{n-1,n-i}) X_{n+1-i}. \] (8.24)
Here from (8.20) and (8.21) by uniqueness (implied by our assumptions) we can compare the coefficients in (8.24) with
\[ \hat{X}_{n+1} = \sum_{i=1}^{n} a_{n,i} X_{n+1-i}. \]

Therefore we deduce that
\[ a_{n,n} = a \]
\[ a_{n,i} = \begin{cases} 
  a_{n-1,i} - a_{n,n} a_{n-1,n-i} & \text{if } i = 1, \ldots, n-1, \\
  a & \text{if } i = n,
\end{cases} \]
as was claimed. We have also
\[ a_{n,n} = E \left[ X_{n+1} \right] - \sum_{i=1}^{n-1} a_{n-1,i} E \left[ X_{n+1} X_{i+1} \right]. \]
\[ = \gamma X(n) - \sum_{i=1}^{n-1} a_{n-1,i} \gamma X(n-i) \]
\[ v_{n-1} \]

It remains to be established that
\[ v_n = v_{n-1} [1 - a_{n,n}^2]. \]
In order to see this we use (8.19) so that
\[ v_n = E \left[ X_{n+1} - \hat{X}_{n+1} \right]^2 \]
\[ = E \left[ X_{n+1} - P \left( X_{n+1} \mid K_1 \right) - P \left( X_{n+1} \mid K_2 \right) \right]^2 = \]
\[ = E \left[ X_{n+1} - P \left( X_{n+1} \mid K_1 \right) \right]^2 + E \left[ P \left( X_{n+1} \mid K_2 \right) \right]^2 \]
\[ - 2E \left[ \left( X_{n+1} - P \left( X_{n+1} \mid K_1 \right) \right) \left( P \left( X_{n+1} \mid K_2 \right) \right) \right] \]

We have here that
\[ E \left[ X_{n+1} - P \left( X_{n+1} \mid K_1 \right) \right]^2 = v_{n-1} \]
by (8.23) and
\[ E \left[ P \left( X_{n+1} \mid K_2 \right) \right]^2 = E \left[ a \left( X_1 - P \left( X_1 \mid K_1 \right) \right) \right]^2 = a^2 v_{n-1} \]
also by (8.23), and
\[ E \left[ \left( X_{n+1} - P \left( X_{n+1} \mid K_1 \right) \right) a \left( X_1 - P \left( X_1 \mid K_1 \right) \right) \right] \]
\[ = aE \left[ X_{n+1} (X_1 - P (X_1 \mid K_1)) \right] - aE \left[ P (X_{n+1} \mid K_1) (X_1 - P (X_1 \mid K_1)) \right]. \]

It can be shown in the same way as shown in (8.15) that

\[ E \left[ P (X_{n+1} \mid K_1) (X_1 - P (X_1 \mid K_1)) \right] = 0. \]

From (8.20) and (8.23) we get

\[ E \left[ X_{n+1} (X_1 - P (X_1 \mid K_1)) \right] = av_{n-1}. \]

Hence we have finally

\[ = v_{n-1} + a^2v_{n-1} - 2a^2v_{n-1} = v_{n-1}[1 - a^2_{n,n}]. \]

\[ \blacksquare \]

### 8.3 Innovations Algorithm

Since the new information at time \( n \) is \( X_n - \hat{X}_n \) rather than \( X_n \) it might be natural to consider predictors which are linear combinations of the innovations \( X_1 - \hat{X}_1, \ldots, X_n - \hat{X}_n \). Actually this makes no difference, since

\[ \mathbb{P}\{X_1, X_2, \ldots, X_n\} = \mathbb{P}\{X_1 - \hat{X}_1, \ldots, X_n - \hat{X}_n\}. \]

In this case we do not need to assume stationarity, and we have the following algorithm:

**Theorem 8.2 (The Innovations Algorithm)** If \( \{X_n \mid n = 0, \pm 1, \pm 2, \ldots\} \) has zero-mean and \( E(X_iX_j) = \kappa(i, j) \), where the matrix

\[
\begin{pmatrix}
\kappa(1,1) & \cdots & \kappa(1,n) \\
\vdots & \ddots & \vdots \\
\kappa(n,1) & \cdots & \kappa(n,n)
\end{pmatrix}
\]

is non-singular, we have

\[
\hat{X}_{n+1} = \begin{cases} 
0 & \text{if } n = 0, \\
\sum_{j=1}^{n} \theta_{n,j}(X_{n+1-j} - \hat{X}_{n+1-j}) & \text{if } n \geq 1,
\end{cases} \tag{8.25}
\]

and

\[
v_0 = \kappa(1, 1),
\]

\[
\theta_{n,n-k} = v_k^{-1}\left(\kappa(n+1, k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right), \quad k = 0, \ldots, n-1,
\]

\[
v_n = \kappa(n+1, n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2v_j.
\]
**Proof:** From (8.25) with \( X_{k+1} - \tilde{X}_{k+1} \) and using orthogonality we get

\[
E \left[ \tilde{X}_{n+1} \cdot \left( X_{k+1} - \tilde{X}_{k+1} \right) \right] = \theta_{n,n-k} v_k,
\]

and thus

\[
\theta_{n,n-k} = v_k^{-1} \cdot E \left[ X_{n+1} \cdot \left( X_{k+1} - \tilde{X}_{k+1} \right) \right].
\]

From (8.25), with \( n \) replaced by \( k \), we obtain

\[
\theta_{n,n-k} = v_k^{-1} \left( \kappa(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j} E \left[ X_{n+1} \cdot \left( X_{j+1} - \tilde{X}_{j+1} \right) \right] \right).
\]

The form of \( v_n \) follows immediately from \( E|X_n - \tilde{X}_n|^2 = E|X_n|^2 - E|\tilde{X}_n|^2 \).

9 One-step prediction using the entire past

Let \( \{X_n \mid n = 0, \pm 1, \pm 2, \ldots\} \) again be a stationary random sequence with mean function equal to zero and with autocovariance function \( \gamma_x(k) \). We consider the problem of predicting with one-step the present value \( X_n \) in terms of its entire (or infinite past) so that

\[
\tilde{P}X_n = \lim_{p \to \infty} P_pX_n = a_1X_{n-1} + \ldots + a_pX_{n-p} = \sum_{j=1}^{\infty} a_jX_{n-j}.
\]

We hope that the random sum converges in the mean square sense (\( l.i.m. \)).

Next we take the limit \( p \to \infty \) in (6.5) and get

\[
\sum_{k=1}^{\infty} a_k\gamma_x(m-k) = \gamma_x(m); m = 1, \ldots
\]

which constitutes a system of infinitely many linear equations in infinitely many unknowns. If these equations have a solution such that \( \sum_{j=1}^{\infty} a_jX_{n-j} \) converges in mean square (\( l.i.m. \)), then we have found the one-step predictor in LMMSE sense based on the entire past of the process.
10 Problems

1. **Expectation as MMSE**
   Suppose that we want to estimate the random variable $X$ by a constant $a$ so that the mean square error
   $$E[(X - a)^2]$$
   is minimized. Show *without using differential calculus* that the optimal choice of $a$ is $a = E[X]$. *Hint:* Use the same method that was applied in the proof of the main theorem 3.2: Let $b$ be an arbitrary constant and write
   $$E[(X - b)^2] = E[(X - E[X] + E[X] - b)^2],$$
   expand and bound downwards.

2. **Conditional Expectation as MMSE**
   Suppose that we want to estimate the random variable $X$ by a function $Z(Y)$ of another random variable $Y$ so that the mean square error
   $$E[(X - Z(Y))^2]$$
   is minimized. Show *without using differential calculus* that the optimal choice is $Z(Y) = E[X|Y]$. There are two alternative ways (I) – (II) of doing this.
   
   (I) Use first the iterated expectation
   $$E[(X - Z(Y))^2] = E[E[(X - Z(Y))^2 | Y]]$$
   and after that the first exercise 1 above.
   
   (II) Use the same method that was applied in the proof of the main theorem 3.2: Let $g(Y)$ be an arbitrary function of $Y$ and write
   $$E[(X - g(Y))^2] = E[(X - E[X|Y] + E[X|Y] - g(Y))^2]$$
   and expand.

3. Let $(X_n)_{n=0}^\infty$ be a stationary process with zero expectation and with autocovariance $\gamma_X(k)$. We predict $X_n$ by the one-step predictor
   $$P_2X_n = a_1X_{n-1} + a_2X_{n-2}.$$ 
   Give an explicit formula for $a_1$ and $a_2$ using LMMSE and determine the optimal error variance.
4. Suppose that \( \{X_n \mid n = 0, \pm 1, \pm 2, \ldots, \} \) is a stationary process with zero mean and with the autocovariance function

\[
\gamma_X(k) = \begin{cases} 
1 & k = 0 \\
\frac{\lambda}{1+\lambda^2} & k = \pm 1 \\
0 & \text{otherwise},
\end{cases}
\]

where \( 0 < \lambda < 1 \). Show that the linear predictor based on the entire past

\[
\tilde{P} X_n = \sum_{k=1}^{\infty} a_k X_{n-k}
\]

is given by

\[
a_k = -(-\lambda)^k, k = 1, 2, \ldots
\]

and find the optimal error variance.