Linear least squares estimation

A short primer

P. Dewilde - N&S
Stochastic variables

We look at stochastic variables on a given probability space! Assumptions on the variables (say \(A, B, C\) or \(X[n], X[n-1], \ldots\))

- Zero means: \(\mathbb{E}A = \mathbb{E}B = \cdots = 0\)
- Covariance data known: \(\mathbb{E}|a|^2, \mathbb{E}(ab^*)\), etc...

*where '\(\mathbb{E}\)' is the 'means' or expectation operator; \(\mathbb{E}\) is linear in its arguments:*

\[
\mathbb{E}(k_1A + k_2B) = k_1\mathbb{E}(a) + k_2\mathbb{E}(b)
\]
Two variable case

Given two stochastic variables $A$ and $B$

Try to estimate $B$ from knowledge (samples) of $A$!

Estimate: $B^\wedge$. We take it proportional to $A$: $B^\wedge=kA$, and wonder what the best coefficient $k$ could be (linear estimation)!

We use the Wiener principle:

The error (innovations) $B-B^\wedge$ will be minimal in the mean least square sense when it is uncorrelated with the known information:

$$E(|B-B^\wedge|^2) \text{ will be minimal if } E(B-B^\wedge)A^* = 0$$

hence: $k=E(BA^*)/E(|A|^2)$
Geometric interpretation

Stochastic variables can be represented as vectors in an (abstract) space that spans them - in fact linear combinations of stochastic variables are nothing but linear combinations of functions on the probabilistic 'space'. Distance is measured by their variance:

$$\|A - B\| = \sqrt{\text{E}(|A - B|^2)}$$

orthogonality means

$$\text{E}[(B - B^\ast)A^\ast] = 0$$
Ilse in a stochastic process

Estimate $X[n]$ from $X[n-1] \cdots X[n-p]$

Linear estimation:

$$X^\wedge [n] = -a_1 X[n-1] - a_2 X[n-2] - \cdots - a_p X[n-p]$$

Gives for the error or innovations:

$$X[n] - X^\wedge [n] = X[n] + a_1 X[n-1] + \cdots + a_p X[n-p]$$

which should be uncorrelated to the known data

(Wiener principle)
Wiener principle (2)

Hence:
\[ \mathbb{E}[(X[n] - X^*[n])X^*[n-1]] = 0; \]
\[ \mathbb{E}[(X[n] - X^*[n])X^*[n-2]] = 0; \]
\[ \vdots \]
\[ \mathbb{E}[(X[n] - X^*[n])X^*[n-p]] = 0 \]

For example (first equation):
\[ \mathbb{E}(X[n]X[n-1]^*) + a_1 \mathbb{E}(X[n-1]X[n-1]^*) + \cdots + a_p \mathbb{E}(X[n-p]X[n-1]^*) = 0 \]

Let's us now assume the process to be stationary in the sense:
\[ \mathbb{E}(X[n]X[n]^*) = \mathbb{E}(X[n-1]X[n-1]^*) = \cdots = \mathbb{E}(X[n-p]X[n-p]^*) = \kappa_X[0] \]
\[ \mathbb{E}(X[n]X[n-1]^*) = \mathbb{E}(X[n-1]X[n-2]^*) = \cdots = \kappa_X[1] \]
\[ \mathbb{E}(X[n-1]X[n]^*) = \mathbb{E}(X[n-2]X[n-1]^*) = \cdots = \kappa_X[-1] \]
\[ \text{etc...} \]
Wiener principle (3)

then the equations become:

\[
\begin{align*}
\kappa_x[1] + a_1\kappa_x[0] + a_2\kappa_x[-1] + \cdots + a_p\kappa_x[-p + 1] &= 0 \\
\kappa_x[2] + a_1\kappa_x[1] + a_2\kappa_x[0] + \cdots + a_p\kappa_x[-p + 2] &= 0 \\
&\quad \cdots \\
\kappa_x[p] + a_1\kappa_x[p - 1] + a_2\kappa_x[p - 2] + \cdots + a_p\kappa_x[0] &= 0
\end{align*}
\]

To this we add (as first equation) the equation describing the variance of the error:

\[
\mathbb{E}((X[n] - X^*[n])^2) = \mathbb{E}((X[n] - X^*[n])X[n]^*) = \kappa_x[0] + a_1\kappa_x[-1] + \cdots + a_p\kappa_x[-p]
\]

(because $X[n]-X^*[n]$ is orthogonal to $X[n-1] \cdots X[n-p]$)
Matrix form

Putting all this in matrix form, we obtain:

\[
\begin{bmatrix}
\kappa_X[0] & \kappa_X[-1] & \cdots & \cdots & \kappa_X[-p] \\
\kappa_X[1] & \kappa_X[0] & \kappa_X[-1] & \kappa_X[-p+1] \\
\vdots & \kappa_X[1] & \kappa_X[0] & \vdots & \vdots \\
\vdots & \vdots & \vdots & \kappa_X[-1] & \kappa_X[-1] \\
\kappa_X[p] & \kappa_X[p-1] & \cdots & \kappa_X[1] & \kappa_X[0]
\end{bmatrix}
\begin{bmatrix}
1 \\
1 \\
\vdots \\
1 \\
1
\end{bmatrix}
= 
\begin{bmatrix}
F_p \\
0 \\
\vdots \\
0 \\
0
\end{bmatrix}
\]

where \( F_p \) is the variance of the \( p \)th order innovation. The matrix in the equation is called the ‘covariance matrix’ \( K_p \).

These are the so called NORMAL or Yule-Walker equations.
Property of the covariance matrix

If the process $X[n]$ is undetermined (i.e. if no linear combination of the variables has variance exactly zero), then the covariance matrix $K_p$ is strictly positive definite (and hence invertible).

*note: we say that a matrix $K$ is positive definite if for all row-vectors $u$ of suitable dimension, $uKu^* \geq 0$, and strictly positive definite if it is positive definite and in addition, $uKu^* = 0 \Rightarrow u = 0$.*

In the following slide we give an indication why this is so! In particular, an undetermined covariance matrix is non singular.
Motivation of the positive definiteness

We have:

\[ K_p = \mathbb{E} \begin{pmatrix} X[n] \\ X[n-1] \\ \vdots \\ X[n-p+1] \\ X[n-p] \end{pmatrix} \begin{pmatrix} X[n] & X[n-1] & \cdots & X[n-p+1] & X[n-p] \end{pmatrix} \]

Now, let \( u \) be a vector of dimension \( p+1 \), then \( uK_p u^* = \mathbb{E}(WW^*) \), for

\[ W = u_0 X[n] + u_1 X[n-1] + \cdots + u_p X[n-p] \]

Hence, for all such \( u \), \( uK_p u^* \geq 0 \), and if this would equal zero, then the corresponding stochastic variable \( W \) would have zero variance, in which case the process would not be undetermined.
Toeplitz matrices

In addition to being positive definite, the covariance matrix of a time invariant stochastic process is Toeplitz meaning that elements on the same principal diagonals are equal: $K_{ij} = \kappa|i-j|$. This property is exploited in the Levinson and Schur algorithms.
Fourier transform and the Kolmogoroff isomorphism

Idea: represent the ‘stochastic function’ $X[n]$ by a function on the unit circle of the complex plane

$$X[n] \approx e^{jn\theta}$$

Also ‘inner products’ must be converted. Let

$$W_X(\theta) = \sum_{k=-\infty}^{\infty} \kappa_X[k]e^{-jk\theta} \quad \text{the Power Spectral Density Function}$$

and let us define the inner product on function of the unit circle:

$$<f(e^{j\theta}), g(e^{j\theta})>_{W_X} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{j\theta})W_X(\theta)g(e^{j\theta})^* d\theta$$

then we get

$$\kappa_X[n-m] = \text{E}(X[n]X[m]^*) = <e^{jn\theta}, e^{jm\theta}>_{W_X} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{jn\theta}W_X(\theta)e^{-jm\theta} d\theta = \kappa_X[n-m]$$
Interpretation

The ‘abstract’ stochastic inner product gets replaced by a concrete inner product of functions on the unit circle, weighted on the Power Density Function $W_x$. This simple device will allow us to study stochastic modeling through filtering. A parametric model will then approximate the power spectral density function. The next sheet shows a direct application of this theory.
Generating a Power Spectral Density Function by filtering

The principle:

\[
\begin{align*}
\text{stochastic process} &\quad X[n] \quad \text{with PSDF} \quad W_X(\theta) \\
\text{Filter with transfer function} &\quad H(e^{j\theta}) \\
\text{stochastic process} &\quad Y[n] \quad \text{with PSDF} \quad |H(e^{j\theta})|^2 W_X(\theta)
\end{align*}
\]

so if we choose \( X[n] \) to be white noise, then \( W_X(\theta) = 1 \), then the output stochastic process has PSDF \( |H(e^{j\theta})|^2 \).

**Proof:** by the Kolmogoroff isomorphism! Let \( h[k] \) be the impulse response of the filter, corresponding to \( H(e^{j\theta}) \). Then

\[
Y[n] = \sum_{k=0}^{\infty} h[k] X[n - k] \leftrightarrow \sum_{k=0}^{\infty} h[k] e^{j(n-k)\theta} = e^{jn\theta} H(e^{j\theta})
\]
Proof (continued)

and we find for the covariance:

\[ \kappa_Y[n-m] = \mathbb{E}(Y[n]Y[m]^*) = \frac{1}{2\pi} \int_{-\pi}^{\pi} [e^{jn\theta} H(e^{j\theta})] W_X(\theta)[e^{jm\theta} H(e^{j\theta})]^* d\theta \]

\[ = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\theta})|^2 W_X(\theta)e^{j(n-m)\theta} d\theta \]

hence, by reverse Fourier transformation:

\[ W_Y(\theta) = |H(e^{j\theta})|^2 W_X(\theta) \]

qed

(a direct proof is possible as well, but is complicated!)
The autoregressive model

Let us take a $p^{th}$ order ‘autoregressive’ model for a process. Suppose that $X[n]$ satisfies

$$X[n] = -A_{p,1}X[n-1] - \cdots - A_{p,p}X[n-p] + \sigma N[n]$$

in which $N[n]$ is unit variance white noise ($\sigma > 0$). The filtering picture is, with

$$A_p(z) = 1 + A_{p,1}z^{-1} + \cdots + A_{p,p}z^{-p}$$

and by the previous filtering theory, we find for the PSDF:

$$W_X(\theta) = \frac{\sigma^2}{|A_p(e^{j\theta})|^2}$$
Normal equations

One may wonder which equation the \( p \)th order coefficients satisfy, if indeed this is a good model for the process. With little effort one finds (by the white noise property):

\[
\begin{bmatrix}
\kappa_X[0] & \kappa_X[-1] & \ldots & \kappa_X[-p] & \ldots \\
\kappa_X[1] & \kappa_X[0] & \ddots & \ddots & \ddots \\
\vdots & \vdots & \ddots & \ddots & \ddots \\
\kappa_X[p] & \kappa_X[0] & \cdots & \kappa_X[0] & A_{p,1} \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\
\end{bmatrix}
\begin{bmatrix}
1 \\
A_{p,1} \\
A_{p,2} \\
\vdots \\
A_{p,p} \\
0
\end{bmatrix} =
\begin{bmatrix}
\sigma^2 \\
0 \\
0 \\
\vdots
\end{bmatrix}
\]

In other words, the \( p \)th order least squares predictor of a process that satisfies a \( p \)th order autoregressive model produces an innovation process that is white.