STATISTICAL SIGNAL PROCESSING

PARAMETER ESTIMATION

P.R. WHITE

INTRODUCTION

GENERAL COMMENTS

LEAST SQUARES: LINEAR PARAMETERS

LEAST SQUARES: NON-LINEAR PARAMETERS

METHOD OF MAXIMUM LIKELIHOOD

SUMMARY

 © ISVR 1999

INTRODUCTION

The problem addressed here is that of fitting some model to the data.

Assume that there are N measurements of some signal, $x(n)$, collected in the column vector $\underline{x}$ and that the goal is to fit the model $M(\underline{\theta})$[1] to them.

The model, $M(\underline{\theta})$, depends upon K unknown parameters collected in the column vector $\underline{\theta}$.

The object is to determine the values of the parameters $\underline{\theta}$ that result in the model most closely resembling the data $\underline{x}$.

The value of K may in itself be a parameter to be determined (!) but usually $K \ll N$, the role of the model is normally to reduce the dimension of the problem in some sense.

There is a wide range of methods for achieving this and in this lecture we shall explore some of the more widely used.

[1] The notation $M(\underline{\theta})$ is deliberately vague at this point, allowing a general discussion of the rather nebulous concept of "a model".

 © ISVR 1999

GENERAL COMMENTS

If we denote a parameter as θ, then one may naturally ask "What is meant by an estimator of θ?"

The estimator of θ will be denoted $\hat{\theta}$, and any function of the data, $\underline{x}$, can be regarded as an estimator of θ, *i.e.* $\hat{\theta} = f(\underline{x})$.

The question we should ask is "What is a good estimator of θ?"

Since the data, $\underline{x}$, is regarded as random, then so $\hat{\theta}$ is also random.

A "good" estimator is one that on average gives you the correct answer, *i.e.* $E[\hat{\theta}] = \theta$.

One can define the *bias* of an estimator as $\theta - E[\hat{\theta}]$. An unbiased estimator is one for which $E[\hat{\theta}] = \theta$, *i.e.* the bias is zero.

A good estimator should also fluctuate as little as possible, *i.e.* have a small variance.

 © ISVR 1999

LEAST SQUARES

Let us define the vector $y_{\underline{\theta}}(n)$ as the output of the model, when the parameters $\underline{\theta}$ are chosen.

This implies that the model is deterministic, in the sense that once the parameters are fixed then the model output is completely specified.

One, natural, way of selecting the parameters is to make the model output, $y_{\underline{\theta}}(n)$, match the data, $x(n)$, as closely as possible. For example by minimising:

$$\Psi = \sum_{n=1}^{N} \left(x(n) - y_{\underline{\theta}}(n) \right)^2$$

with respect to $\underline{\theta}$.

The cost function can be written in matrix/vector notation as:

$$\Psi = \left(\underline{x} - \underline{y}_{\underline{\theta}} \right)^t \left(\underline{x} - \underline{y}_{\underline{\theta}} \right) = \underline{e}^t \, \underline{e}$$

where $\underline{y}_{\underline{\theta}}$ is the column vector of the model outputs and $\underline{e}$ is the column vector of the errors.

 © ISVR 1999

LINEAR PARAMETERS (AMPLITUDES)

If the model parameters to be fitted to the data are linear, then the solution to the least squares criterion takes a simple form.

For a model with linear parameters

$$\underline{y}_{\underline{\theta}} = \sum_{k=1}^{K} a_k \underline{v}_k = V\underline{a}$$

where a_k are the unknown parameters and $\underline{v}_k$ are vectors of the components to be fitted. The vector $\underline{a}$ is a column vector containing the a_k's and the kth column of the matrix V is $\underline{v}_k$.

$$\Psi = \left(\underline{x} - V\underline{a}\right)^t \left(\underline{x} - V\underline{a}\right) = \underline{x}^t\underline{x} + \underline{a}^t V^t V\underline{a} - 2\underline{x}^t V\underline{a}$$

To minimise Ψ with respect to $\underline{a}$ one seeks a solution to

$$\frac{\partial\Psi}{\partial\underline{a}} = 2V^tV\underline{a} - 2V^t\underline{x} = 0$$

$$\Rightarrow \underline{a} = \left(V^tV\right)^{-1}V^t\underline{x}$$

Substituting this into the expression for Ψ gives:

$$\Psi = \underline{x}^t\underline{x} + \underline{x}^tV\left(V^tV\right)V^t\underline{x} - 2\underline{x}^tV\left(V^tV\right)V^t\underline{x}$$

$$= \underline{x}^t\underline{x} - \underline{x}^tV\left(V^tV\right)V^t\underline{x} = \underline{x}^t\left\{I - V\left(V^tV\right)V^t\right\}\underline{x} = \underline{x}^tP^{\perp}\underline{x} \qquad (1)$$

Note that $\underline{e} = P^{\perp}\underline{x}$

 © ISVR 1999

COMMENTS ON LS ESTIMATES OF LINEAR PARAMETERS

The solution

$$\underline{a} = \left(V^t V\right)^{-1} V^t \underline{x}$$

can be obtained by trying to solve the system of equations $V\underline{a}=\underline{x}$ and using the concept of a pseudo-inverse:

$$\underline{a} = V^{\#}\underline{x} = \left(V^t V\right)^{-1} V^t \underline{x}$$

The model output is given by

$$\underline{y}_{\underline{\theta}} = V\underline{a} = V\left(V^t V\right)^{-1} V^t \underline{x} = P\underline{x}$$

The matrix P is referred to as *orthogonal projection matrix* or just the *projection matrix*.

The matrix $P^{\perp}=I-P$ is the *perpendicular* or *complimentary projection matrix*.

They are both symmetric and idempotent, so that

$$P^2 = PP = V\left(V^t V\right)^{-1} V^t V\left(V^t V\right)^{-1} V^t = V\left(V^t V\right)^{-1} V^t = P$$

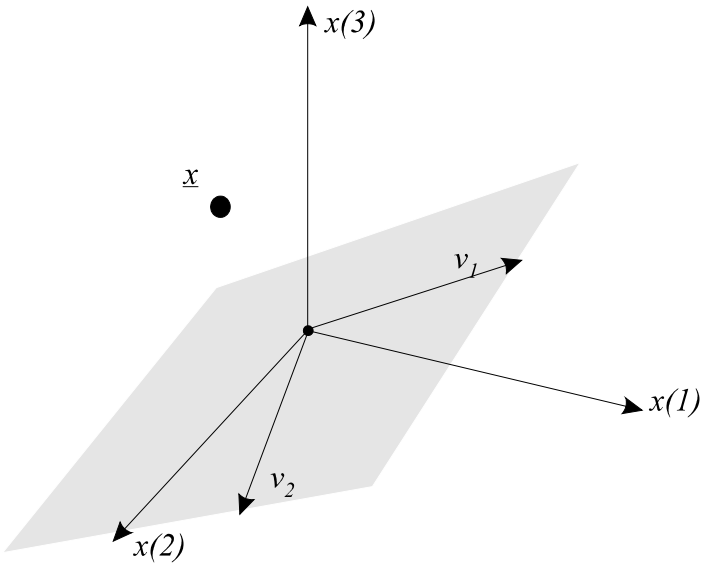
and $P^{\perp}P^{\perp}=P^{\perp}$

 © ISVR 1999

GEOMETRICAL INTERPRETATION

The LS solution can be viewed in a geometric sense.

Consider a problem in which N=3 and K=2.



The shaded plane represents the set of points that can be reached by linearly combining the vectors $\underline{v}_1$ and $\underline{v}_2$.

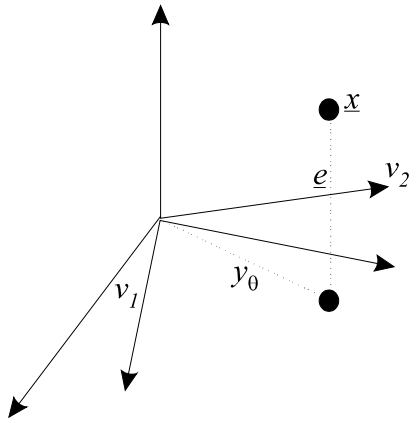
Since the model output $\underline{y}_\theta$ is formed as $a_1\underline{v}_1 + a_2\underline{v}_2$ then it lies some where on the shaded plane.

The LS solution corresponds to the point where the distance between $\underline{x}$ and $\underline{y}_\theta$ is minimised, *i.e.* the point where the plane is closest to $\underline{x}$.

 © ISVR 1999

GEOMETRICAL INTERPRETATION (CONT'D)

Below the previous figure is redrawn so that the shaded plane is horizontal.



Notes:

- The vectors $\underline{v}_1$ and $\underline{v}_2$ are in general not at 90º to each other.
- Adding the vectors $\underline{e}$ and $\underline{y}_\theta$ gives $\underline{x}$.
- $\underline{e}=P^{\perp}\underline{x}$ and $\underline{y}_\theta=P\underline{x}$ are at 90º to each other, *i.e.* they are orthogonal.
- The point $\underline{y}_\theta$ is the image of $\underline{x}$ in the plane of $\underline{v}_1$, $\underline{v}_2$.
- The point $\underline{e}$ is the image of $\underline{x}$ in the plane that is orthogonal to both $\underline{v}_1$ and $\underline{v}_2$.

The optimal least squares linear parameter estimates are obtained by projecting the data into the space (plane) of the vectors $\underline{v}_k$, *i.e.* the column space of V.

The error in this estimate is the projection of the data in the space orthogonal to the vectors $\underline{v}_k$, *i.e.* the null space of V.

 © ISVR 1999

EXAMPLE 1

Consider the problem of fitting a polynomial to data *N* samples of data.

The signal model is:

$$y(n) = \sum_{k=0}^{K-1} a_k n^k$$

Define the *K* vectors so that:

$$\underline{v}_0 = \left[1 1 \cdots 1\right]^t, \underline{v}_1 = \left[1\,2\,3 \cdots N\ \right]^t, \underline{v}_2 = \left[1^2\ 2^2\ 3^2 \cdots N^2\ \right]^t, \cdots$$

Allowing one to write:

$$\underline{y} = \sum_{k=0}^{K-1} a_k \underline{v}_k = V\underline{a}$$

The LS estimate of the amplitudes are given by

$$\underline{a} = \left(V^t V\right)^{-1} V^t \underline{x}$$

and

$$\underline{y} = V\underline{a} = V\left(V^t V\right)^{-1} V^t \underline{x}$$

A special case of this occurs when *K*=1. The resulting algorithm is the familiar linear regression routine.

 © ISVR 1999

NON-LINEAR LEAST SQUARES

For many examples the parameters of interest do not appear in the model in a linear fashion. The least squares criterion:

$$\Psi = \sum_{n=1}^{N} \left(x(n) - y_{\underline{\theta}}(n) \right)^2$$

can still be applied but one can not write down an analytic solution.

The derivative of the cost function can be written:

$$\frac{\partial \Psi}{\partial \underline{\theta}} = -2 \sum_{n=1}^{N} e(n) \frac{\partial y_{\underline{\theta}}(n)}{\partial \underline{\theta}} = 0$$

In general, this system of equations does not simplify and one has to seek the minimum of Ψ using a multi-dimensional minimisation routine, with all the attendant problems that may introduce.

If the model order is small then the problems may be manageable, but for large values of K the computational cost increases dramatically.

Several approaches for such problems have been proposed, including alternating projection and the iterative multi-paramater (IMP) algorithm.

 © ISVR 1999

MIXED PROBLEMS

Many parameter estimation problems consist of fitting a model in which some of the parameters are linear and some are non-linear.

In such circumstances one usually fits the linear parameters first and only then moves on to fit the non-linear parameters.

In this manner the multi-dimensional minimisation is only conducted with respect to the non-linear parameters, so is more efficient.

For example consider a model of the form:

$$y(n) = \sum_{k=1}^{K/2} a_k \underline{v}(\theta_k) \quad \Rightarrow \quad \underline{y} = V\left(\underline{\theta}\right)\underline{a}$$

This model requires the fitting of K parameters, with $K/2$ linear parameters (a_k's) and $K/2$ non-linear parameters (θ_k's).

For the above model one can still write:

$$\underline{a} = V\left(\underline{\theta}\right)^{\#} \underline{x} = \left(V\left(\underline{\theta}\right)^t V\left(\underline{\theta}\right)\right)^{-1} V\left(\underline{\theta}\right)^t \underline{x}$$

and the cost function, Ψ, becomes

$$\Psi = \underline{x}^t \left\{ I - V\left(\underline{\theta}\right)\left(V\left(\underline{\theta}\right)^t V\left(\underline{\theta}\right)\right)^{-1} V\left(\underline{\theta}\right)^t \right\} \underline{x}$$

This is only a function of the non-linear parameters, (θ_k's).

 © ISVR 1999

EXAMPLE 2: SINGLE SINUSOID MODEL

Consider the problem of fitting a model of the form:

$$y(n) = A\cos\left(2\pi fn\Delta + \phi\right) = B\cos\left(2\pi fn\Delta\right) + C\sin\left(2\pi fn\Delta\right)$$

The 2 formulations above represent the same model. Each formulation has 3 unknown parameters; in the first case, A,f and ϕ, in the second case, B,C and f.

Both formulations are mixed (containing some linear and some non-linear parameters), however the second formulation only has 1 non-linear parameter (f).

$$\underline{y} = B\underline{v}_{\cos}(f) + C\underline{v}_{\sin}(f) = V(f)\underline{a}$$

where $\underline{v}_{\cos}(f) = \left[1, \cos\left(2\pi f\Delta\right), \cos\left(2\pi f\,2\Delta\right), \cdots, \cos\left(2\pi f(N-1)\Delta\right)\right]^t$,

$\underline{v}_{\sin}(f) = \left[0, \sin\left(2\pi f\Delta\right), \sin\left(2\pi f\,2\Delta\right), \cdots, \sin\left(2\pi f(N-1)\Delta\right)\right]^t$,

$V(f) = \left[\underline{v}_{\cos}(f) \quad \underline{v}_{\sin}(f)\right]$ and $\underline{a}=[B,C]^t$.

The LS estimate for the amplitudes, $\underline{a}$, is given by

$$\underline{a} = \left(V(f)^t V(f)\right)^{-1} V(f)^t \underline{x}$$

leading to a cost function:

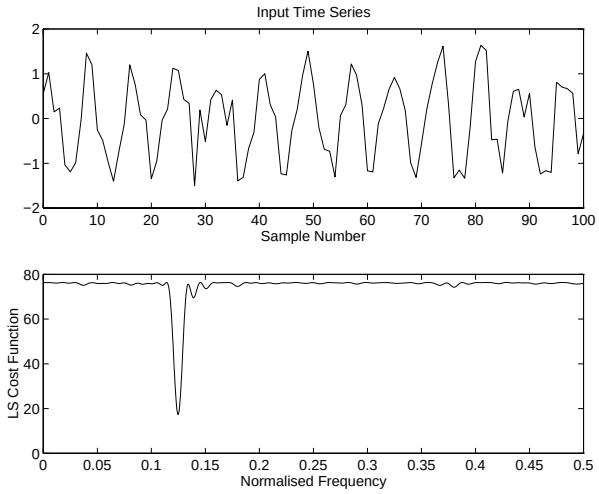
$$\Psi = \underline{x}^t \left\{ I - V(f)\left(V(f)^t V(f)\right)^{-1} V(f)^t \right\} \underline{x}$$

The LS optimal estimate for f is obtained by locating the peak of Ψ.

 © ISVR 1999

EXAMPLE 2: SINGLE SINUSOID MODEL (CONT'D)

Below is an example of the cost function computed for a sine wave of frequency 0.125 at an SNR of 5 dB.



Note the marked minimum at the frequency 0.125.

The cost function, Ψ, can be written:

$$\Psi = \underline{x}^t \underline{x} - \underline{x}^t V(f)\left(V(f)^t V(f)\right)^{-1} V(f)^t \underline{x}$$

The term $\underline{x}^t\underline{x}$ is independent of frequency and so finding the minimum of Ψ is equivalent to finding the **maximum** of Ψ_1, where

$$\Psi_1 = \underline{x}^t V(f)\left(V(f)^t V(f)\right)^{-1} V(f)^t \underline{x}$$

The matrix $V(f)^tV(f)$ is (2×2) and can be expressed as

$$V(f)^t V(f) = \begin{bmatrix} \underline{v}_{\cos}(f)^t \underline{v}_{\cos}(f) & \underline{v}_{\cos}(f)^t \underline{v}_{\sin}(f) \\ \underline{v}_{\sin}(f)^t \underline{v}_{\cos}(f) & \underline{v}_{\sin}(f)^t \underline{v}_{\sin}(f) \end{bmatrix}$$

 © ISVR 1999

EXAMPLE 2: SINGLE SINUSOID MODEL (CONT'D)

$$\underline{v}_{\cos}(f)^t \underline{v}_{\cos}(f) = \sum_{n=0}^{N-1} \cos(2\pi fn\Delta)^2 = \sum_{n=0}^{N-1} \frac{1+\cos(4\pi fn\Delta)}{2} \approx \frac{N}{2}$$

$$\underline{v}_{\sin}(f)^t \underline{v}_{\sin}(f) = \sum_{n=0}^{N-1} \sin(2\pi fn\Delta)^2 = \sum_{n=0}^{N-1} \frac{1-\cos(4\pi fn\Delta)}{2} \approx \frac{N}{2}$$

$$\underline{v}_{\sin}(f)^t \underline{v}_{\cos}(f) = \underline{v}_{\cos}(f)^t \underline{v}_{\sin}(f) = \sum_{n=0}^{N-1} \cos(2\pi fn\Delta)\sin(2\pi fn\Delta) \approx 0$$

Using these approximations one obtains:

$$V(f)^t V(f) \approx \frac{N}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

So that

$$\Psi_1 \approx \frac{N}{2}\underline{x}^t V(f) V(f)^t \underline{x} = \frac{N}{2}\underline{X}^t \underline{X}$$

where $\underline{X}=V(f)^t\underline{x}$, such that $\underline{X}$ is a *2* element column vector, as follows

$$\underline{X} = \begin{bmatrix} \underline{v}_{\cos}(f)^t \underline{x} \\ \underline{v}_{\sin}(f)^t \underline{x} \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} x(n)\cos(2\pi fn\Delta) \\ \sum_{n=1}^{N} x(n)\sin(2\pi fn\Delta) \end{bmatrix}$$

$$\underline{X}^t \underline{X} = \left(\sum_{n=1}^{N} x(n)\sin(2\pi fn\Delta)\right)^2 + \left(\sum_{n=1}^{N} x(n)\cos(2\pi fn\Delta)\right)^2$$

$$= \left|\sum_{n=1}^{N} x(n)e^{-2\pi ifn\Delta}\right|^2 = \left|X_s(f)\right|^2$$

 © ISVR 1999

DISCUSSION OF EXAMPLE 2

We have seen that the cost function for fitting the sinusoidal model using least squares is approximately equal to the magnitude of the Fourier transform of $x(n)$.

The quality of this result depends upon the accuracy of the approximations found at the top of the preceding page.

These approximations are only good as long as the frequency is not close to 0 or $1/(2\Delta)$.

For most frequencies the approximation is very reasonable.

One can conclude that the LS estimate of the frequency of a sinusoid can be computed by locating the maximum of the Fourier transform of $x(n)$.

 © ISVR 1999

EXAMPLE 3: TWO SINUSOID MODEL

Consider a signal that consists of *2* sine waves, so

$$y(n)=a_1\cos(2\pi f_1 n\Delta)+a_2\sin(2\pi f_1 n\Delta)+a_3\cos(2\pi f_2 n\Delta)+a_4\sin(2\pi f_2 n\Delta)$$

There are a total of 6 parameters in this model, 4 amplitude (linear parameters) and 2 frequencies (non-linear parameters).

In a similar manner to example *2* one can write

$$\underline{y} = V(f_1, f_2)\underline{a}$$

where $\underline{a}$ is the vector of amplitudes and $V(f_1,f_2)$ is defined as

$$V(f_1, f_2) = \left[\underline{v}_{\cos}(f_1) \quad \underline{v}_{\sin}(f_1) \quad \underline{v}_{\cos}(f_2) \quad \underline{v}_{\sin}(f_2)\right]$$

Given this change to the matrix *V*, the derivation follows that of example 2 and one can show that one needs to maximise

$$\Psi_1 = \underline{x}^t V(f_1, f_2)\left(V(f_1, f_2)^t V(f_1, f_2)\right)^{-1} V(f_1, f_2)^t \underline{x}$$

which is a function of the 2 frequencies.

This can also be approximated by the Fourier transform of $x(n)$, however the approximation fails if $|f_1-f_2|$ is not large.

For resolving 2 closely spaced sinusoidal components, the LS method has a distinct advantages over the Fourier transform.

 © ISVR 1999

THE METHOD OF MAXIMUM LIKELIHOOD

The LS method, whilst being conceptually simple, lacks a sound theoretical basis.

Estimators generated using the method of Maximum Likelihood (ML) can be shown to have several theoretically attractive properties.

The ML principle is that given a model $M(\underline{\theta})$, dependant upon some parameters $\underline{\theta}$, then one can construct the conditional pdf $p(\underline{x}|\underline{\theta})$ representing the pdf of the data given a set of parameters (and some model[2]).

This conditional probability is referred to as the *likelihood* (also see detection notes).

When using ML one regards $p(\underline{x}|\underline{\theta})$ not as a function of the data, but as a function of the parameters, it is treated as $p(\underline{\theta}|\underline{x})$.

The ML principle is to find the set of parameters for which $p(\underline{x}|\underline{\theta})$ is maximised.

Commonly one seeks to maximise $\log\{p(\underline{x}|\underline{\theta})\}$ since this is equivalent to maximising $p(\underline{x}|\underline{\theta})$ and is often more tractable.

[2] It is normal to suppress the dependence upon the model in the notation, one could more accurately write $p(\underline{x}|M(\underline{\theta}))$, but this is not standard practice and will not be adopted here.

 © ISVR 1999

PROPERTIES OF ML ESTIMATORS

1) The ML estimator is asymptotically (?) unbiased.

2) It is asymptotically efficient

   For sufficiently large data sets the ML estimates achieve the smallest possible variance (see later notes on the Cramér-Rao lower bound).

3) If an efficient estimator exists then it is the ML estimator.

4) It is asymptotically distributed according to a Gaussian distribution.

   For large data sets the ML estimators have a Gaussian pdf, the mean of which is the true parameter (since they are unbiased) and the variance of which corresponds to the Cramér-Rao lower bound.

5) Consider new parameter formed as the result of a one-to-one transformation of θ, say $\theta_{new}=g(\theta)$. Then the ML estimate of θ_{new} is $g(\hat{\theta})$, where $\hat{\theta}$ is the ML estimate of θ.

Properties 2) and 3) make the ML attractive. Since they claim some form of global optimality.

There may be other estimators that will perform better than ML for short data lengths or in problems where no efficient estimator exists.

 © ISVR 1999

EXAMPLE 4

Consider a general model of the form:

$$x(n)=y_{\underline{\theta}}(n)+w(n)$$

where $y_{\underline{\theta}}(n)$ is the signal component, defined via a model with parameters $\underline{\theta}$, the noise $w(n)$ is zero mean, Gaussian and independent.

The noise $w(n)$ can be expressed as:

$$w(n)=x(n)-y_{\underline{\theta}}(n)$$

The noise vector, being independent and Gaussian, is distributed according to:

$$p(\underline{w}) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\sum_{n=1}^{N} w(n)^2/(2\sigma^2)}$$

so that

$$p(\underline{x}\,|\,\underline{\theta}) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\sum_{n=1}^{N} \left(x(n)-y_{\underline{\theta}}(n)\right)^2/(2\sigma^2)}$$

Within $p(\underline{x}|\underline{\theta})$ the only model dependent term occurs in the exponent. Maximising $p(\underline{x}|\underline{\theta})$ is equivalent to minimising

$$\sum_{n=1}^{N} \left(x(n)-y_{\underline{\theta}}(n)\right)^2$$

This is exactly the Least Squares criterion.

 © ISVR 1999

EXAMPLE 5

Example 4 is extended so that the noise is now considered as being correlated (non-white).

The signal model can be expressed in vector form:

$$\underline{x}=\underline{y}(\underline{\theta})+\underline{w}, \quad \Rightarrow \quad \underline{w}=\underline{x}-\underline{y}(\underline{\theta})$$

Then

$$p(\underline{w})=\frac{1}{(2\pi)^{N/2}\,|R|^{1/2}}e^{-\underline{w}^t R^{-1}\underline{w}/2}$$

Such that

$$p(\underline{x}\,|\,\underline{\theta})=\frac{1}{(2\pi)^{N/2}\,|R|^{1/2}}e^{-\left(\underline{x}-\underline{y}(\underline{\theta})\right)^t R^{-1}\left(\underline{x}-\underline{y}(\underline{\theta})\right)/2}$$

where R is the auto-correlation matrix, $E[\underline{x}\underline{x}^t]$.

In this example maximising $p(\underline{x}|\underline{\theta})$ is equivalent to minimising

$$\Psi_1=\left(\underline{x}-\underline{y}(\underline{\theta})\right)^t R^{-1}\left(\underline{x}-\underline{y}(\underline{\theta})\right)=\underline{e}^t R^{-1}\underline{e}$$

Assuming R is positive definite then we can write:

$$R^{-1}=C^t C$$

where the matrix C is one of several possible factorisations of R^{-1}, for example the eigen based decomposition: $C=\Lambda^{-1/2}U^t$.

 © ISVR 1999

EXAMPLE 5 (CONT'D)

In which case we aim to minimise

$$\Psi_1 = \left(\underline{x} - \underline{y}(\underline{\theta})\right)^t C^t C \left(\underline{x} - \underline{y}(\underline{\theta})\right) = \underline{\varepsilon}^t \underline{\varepsilon} = \sum_{n=1}^{N} \varepsilon(n)^2$$

where $\underline{\varepsilon}=C(\underline{x}-\underline{y}(\underline{\theta}))$. The errors $\underline{\varepsilon}$ are weighted versions of the original errors $\underline{e}$.

The above method is a generalisation of Least Squares called *Weighted Least Squares.*

One consequence of this weighting is to render the vector $\underline{x}$ white. If one defines $\underline{\tilde{x}} = C\underline{x}$ then

$$E[\underline{\tilde{x}}\underline{\tilde{x}}^t] = E[C\underline{x}\underline{x}^t C^t] = CE[\underline{x}\,\underline{x}^t]C^t = CRC^t$$

Assuming $C=\Lambda^{-1/2}U^t$ and recalling that $R=U\Lambda U^t$ then this becomes

$$CRC^t = \Lambda^{-1/2}U^t U \Lambda U^t U \Lambda^{-1/2} = I$$

where the last equality arises because U is orthonormal.

Thus $\underline{\tilde{x}}$ is white noise sequence, with unit variance, formed by linearly combinations of $x(n)$, it is referred to as a *whitened* sequence.

 © ISVR 1999

SUMMARY

Least Squares is conceptually simple method for fitting (deterministic) models to data.

Unfortunately least squares, generally, produces results which are sub-optimal.

The method of Maximum Likelihood, is a parameter estimation method that asymptotically gives optimal results.

For deterministic models in independent (white) Gaussian noise the ML and LS are equivalent.

For deterministic models in dependent (coloured) Gaussian noise the ML is equivalent to Weighted LS.

 © ISVR 1999