2008 IIB 4F7 DIGITAL FILTERS AND SPECTRUM ESTIMATION PROJ SJ GODSILL

April 2008

Module 4F7 - worked solutions

DIGITAL FILTERS AND SPECTRUM ESTIMATION

STATIONERY REQUIREMENTS SPECIAL REQUIREMENTS

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

Version: 1

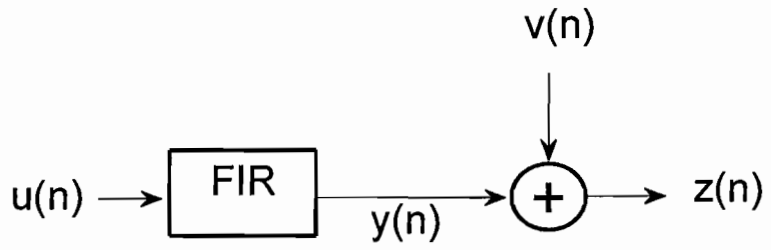

Fig. 1

1 (a) Describe the Recursive Least Square (RLS) method for adaptive filtering. Your answer should include the following points: a definition of the relevant signals and the cost function being minimised; the influence of the forgetting factor on tracking performance and misadjustment; whether the RLS method converges to the Wiener filter for a particular value of forgetting factor and why. [35%]

(b) Fig. 2 depicts an impulse response identification problem. The coefficients of the FIR filter are to be identified using a known input sequence $\{u(n)\}_{n\geq 0}$. The output of the filter $y(n)$ is measured by an imperfect sensor, which can be modelled as the filter output with additive noise $v(n)$. Assume that $E\{v(n)\} = 0$, $E\{v(n)^2\} = \sigma_v^2$ and that the impulse response of the filter to be identified is $[\beta_0, \ldots, \beta_{L-1}]^{\mathrm{T}}$.

Explain how to solve the identification problem using RLS assuming knowledge of L. For the case $L = 2$ and given that $\beta_0 = 1$, compute the RLS solution at time n. [35%]

(c) Explain how to solve the same general identification problem using the Steepest Descent and the Least Mean Square (LMS) algorithm, once again assuming you are given the value of L only. Assuming that$u(n) = \sum_{i=0}^{M-1} \alpha_i w(n-i)$ where $E\{w(n)\} = 0$, $E\{w(n)^2\} = \sigma_w^2$, $E\{w(k)w(l)\} = 0$ when $l \neq k$, what is the stability condition for the LMS algorithm? (Hint: use an appropriate estimate of $\lambda_{\max}$.) [30%]

Solution:

Part (a) The **input** signal $\{u(n)\}$, **reference** signal $\{d(n)\}$, filter $\mathbf{h} = [h_0, h_1, \ldots, h_{M-1}]^{\mathrm{T}} \in R^M$. The **output** signal $\{y(n)\}$ is

$$y(n) = \sum_{k=0}^{M-1} h_k u(n-k) = \mathbf{u}^{\mathrm{T}}(n)\mathbf{h}$$

The **error** signal $\{e(n)\}$, $e(n) = d(n) - \mathbf{u}^{\mathrm{T}}(n)\mathbf{h}$

At time n RLS aims to minimise the cost function

$$J(\mathbf{h},n) = \sum_{k=0}^{n} \lambda^{n-k} e^2(k)$$

where $0 < \lambda \leq 1$. λ is a called the **forgetting factor**

For $\lambda < 1$, $J(\mathbf{h},n)$ regards the past errors as less important since they are weighted by λ^{n-k}. The smaller λ is, the quicker the RLS will respond if the Wiener filter is time varying (better tracking)

For $\lambda = 1$, RLS solution converges to the Wiener filter because

$$\frac{1}{n+1} J(\mathbf{h},n)|_{\lambda=1} = \frac{1}{n+1} \sum_{k=0}^{n} e^2(k)$$

This is a sample average and should converge to $E\left\{e^2(k)\right\}$. So, we can now argue that the RLS solution (asymptotically) and the Wiener filter coincide when $\lambda = 1$ since dividing $J(\mathbf{h},n)$ by $n+1$ does not change the minimizer.

Misadjustment is $\frac{E\left[e^2(n)\right] - J_{\min}}{J_{\min}} \approx \frac{1-\lambda}{1+\lambda}$

Part (b)

Following on from the solution to part (a) the error signal for solving the identification problem should be:

the error signal at time n is $e(n) = z(n) - \mathbf{h}^{\mathrm{T}}\mathbf{u}(n)$ where $\mathbf{h} = [h_0, h_1, \ldots, h_{L-1}]^{\mathrm{T}}$, $\mathbf{u}(n) = [u(n), u(n-1), \ldots, u(n-L+1)]^{\mathrm{T}}$.

Note the length of the filter is L.

For the case when $L = 2$, $z(n) = u(n) + \beta_1 u(n-1) + v(n)$.

Since we are told that $\beta_0 = 1$, we can set $\mathbf{h} = [1, h_1]^{\mathrm{T}}$ so the only parameter to be adapted is h_1. This gives $e(n) = (\beta_1 - h_1)u(n-1) + v(n)$

The cost function at time n is $\sum_{k=0}^{n} e(k)^2 \lambda^{n-k}$

Differentiating w.r.to h_1 gives $\sum_{k=0}^{n} 2e(k)\lambda^{n-k}u(k-1)$

Setting the derivative to zero yields the solution to the RLS problem at time n, which is,

$$\sum_{k=0}^{n} \lambda^{n-k}[z(k) - u(k)]u(k-1) = h_1 \sum_{k=0}^{n} \lambda^{n-k}u(k-1)u(k-1)$$

Note the solution is in terms of the sensor output and known input signal.

Part (c)

The Steepest Descent cost function is $J(\mathbf{h}) = E\{e(n)^2\}$ where $e(n) = z(n) - \mathbf{h}^{\mathrm{T}}\mathbf{u}(n)$ where $\mathbf{h} = [h_0, h_1, \ldots, h_{L-1}]^{\mathrm{T}}$, $\mathbf{u}(n) = [u(n), u(n-1), \ldots, u(n-L+1)]^{\mathrm{T}}$ as in part (b). Note length of filter is L.

The SD recursion is

$$\begin{aligned} \mathbf{h}(n+1) &= \mathbf{h}(n) - \frac{\mu}{2}\nabla J(\mathbf{h})|_{\mathbf{h}=\mathbf{h}(\mathbf{n})} \\ &= \mathbf{h}(n) + \mu(\mathbf{p} - \mathbf{R}\mathbf{h}(n)) \end{aligned}$$

where $\mathbf{p} = E\{z(n)\mathbf{u}(n)\}$ and $\mathbf{R} = E\{\mathbf{u}(n)\mathbf{u}(n)^{\mathrm{T}}\}$

The LMS recursion is

$$\mathbf{h}(n+1) = \mathbf{h}(n) + \mu e(n)\mathbf{u}(n)$$

where $e(n) = z(n) - \mathbf{h}(n)^{\mathrm{T}}\mathbf{u}(n)$

The stability condition for the SD and LMS algorithm is $0 < \mu < 2/\lambda_{\max}$

Use the following estimate for $\lambda_{\max} < \sum_{k=1}^{L} \lambda_k = \mathrm{trace}(\mathbf{R}) = LE\{u(n)^2\}$

For the given input signal $E\{u(n)^2\} = \sum_{i=0}^{M-1} \alpha_i^2 E\{w(n-i)^2\} = \sigma_w^2 \sum_{i=0}^{M-1} \alpha_i^2$

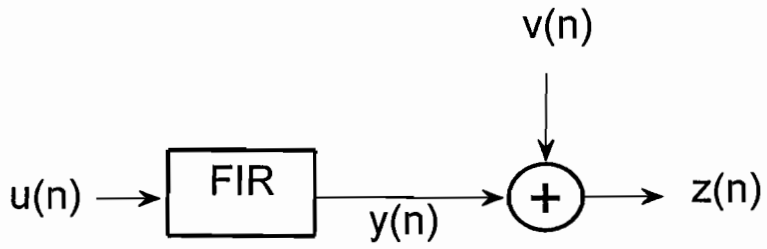

Fig. 2

2 (a) Describe the Recursive Least Square (RLS) method for adaptive filtering. Your answer should include the following points: a definition of the relevant signals and the cost function being minimised; the influence of the forgetting factor on tracking performance and misadjustment; whether the RLS method converges to the Wiener filter for a particular value of forgetting factor and why. [35%]

(b) Fig. 2 depicts an impulse response identification problem. The coefficients of the FIR filter are to be identified using a known input sequence $\{u(n)\}_{n\geq 0}$. The output of the filter $y(n)$ is measured by an imperfect sensor, which can be modelled as the filter output with additive noise $v(n)$. Assume that $E\{v(n)\} = 0$, $E\{v(n)^2\} = \sigma_v^2$ and that the impulse response of the filter to be identified is $[\beta_0, \ldots, \beta_{L-1}]^{\mathrm{T}}$.

Explain how to solve the identification problem using RLS assuming knowledge of L. For the case $L = 2$ and given that $\beta_0 = 1$, compute the RLS solution at time n. [35%]

(c) Explain how to solve the same general identification problem using the Steepest Descent and the Least Mean Square (LMS) algorithm, once again, assuming you are given the value of L only. Assuming that$u(n) = \sum_{i=0}^{M-1} \alpha_i w(n-i)$ where $E\{w(n)\} = 0$, $E\{w(n)^2\} = \sigma_w^2$, $E\{w(k)w(l)\} = 0$ when $l \neq k$, what is the stability condition for the LMS algorithm? (Hint: use an appropriate estimate of $\lambda_{\max}$.) [30%]

3 (a) Consider an infinite collection $\{y_t\}_{t\geq 0}$ of random variables where

$$y_t = x + v_t.$$

Describe the Gram-Schmidt orthogonalization procedure for forming the minimum variance linear estimator of x using the first T random variables from this collection.

For the case $T = 2$, you are given $E(xy_1) = 0.1$, $E(xy_2) = 0.3$, $E(y_1y_2) = -0.2$, $E(y_1^2) = 1$, $E(y_2^2) = 1.5$. Compute the coefficient of y_1 and y_2 of the estimator given by the Gram-Schmidt procedure. (Hint: write $\hat{x}$ as a function of y_1 and y_2 explicitly and compute the coefficients.) [35%]

(b) A constant temperature θ is measured with the use of noisy sensors. The measurement made by sensor i is $y_i = \theta + v_i$ where v_i is a Gaussian random variable with mean 0 and variance σ_i^2. Assume random variables $\{v_i\}_{i=1,2,\ldots,T}$ are independent. Compute the minimum variance linear unbiased estimator for $T = 2$ and compute the minimum variance [35%]

(c) Let the ith sensor variance be $\sigma_i^2 = i$. Using the answer derived in the previous part, estimate the value of T needed to reduce the variance of the estimator to less than 6/11. [30%]

Solution:

Part a)

We must first make $\{y_1, \ldots, y_T\}$ orthogonal. Set $\varepsilon_1 = y_1$. For $T \geq j > 1$,

$$\varepsilon_j = y_j - \sum_{i=1}^{j-1} \frac{E\left(y_j\varepsilon_i\right)}{E\left(\varepsilon_i^2\right)}\varepsilon_i$$

Now project x onto $\{\varepsilon_1, \ldots, \varepsilon_T\}$,

$$\widehat{x} = \sum_{i=1}^{T} \frac{E\left(x\varepsilon_i\right)}{E\left(\varepsilon_i^2\right)}\varepsilon_i.$$

Part b) For $T = 2$, $\widehat{x} = \frac{E(x\varepsilon_1)}{E\left(\varepsilon_1^2\right)}\varepsilon_1 + \frac{E(x\varepsilon_2)}{E\left(\varepsilon_2^2\right)}\varepsilon_2$ where $\varepsilon_2 = y_2 - \frac{E(y_2y_1)}{E(y_1^2)}y_1$

$E\left(x\varepsilon_1\right) = 0.1, E\left(\varepsilon_1^2\right) = 1$

$E\left(x\varepsilon_2\right) = E\left(xy_2\right) - \frac{E(y_2\varepsilon_1)}{E\left(\varepsilon_1^2\right)}E\left(x\varepsilon_1\right) = 0.3 - \frac{-0.2}{1}0.1 = 0.32$

$E\left(\varepsilon_2^2\right) = E(y_2^2) - \frac{E(y_2 y_1)^2}{E(y_1^2)} = 1.5 - 0.04 = 1.46$

So $\widehat{x} = 0.1y_1 + \frac{0.32}{1.46}(y_2 + 0.2y_1) = 0.1438y_1 + 0.2192y_2$

Part c)

Using (y_1, y_2), propose an estimate of θ of the form

$$\widehat{\theta} = a_1 y_1 + a_2 y_2$$

and determine (a_1, a_2) for this estimate to be unbiased and to admit a variance as low as possible. Since

$$\begin{aligned} E\left\{\widehat{\theta}\right\} &= a_1 E\{y_1\} + a_2 E\{y_2\} \\ &= (a_1 + a_2)\,\theta \end{aligned}$$

we require

$$a_1 + a_2 = 1$$

for the estimator to be unbiased. Variance of the estimate:

$$\begin{aligned} var\left\{\widehat{\theta}\right\} &= E\left\{\left(\widehat{\theta} - E(\widehat{\theta})\right)^2\right\} \\ &= E\{(a_1 v_1 + a_2 v_2)^2\} \\ &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 \end{aligned}$$

Now substitute $a_2 = 1 - a_1$ in to get

$$\begin{aligned} var\left\{\widehat{\theta}\right\} &= a_1^2\sigma_1^2 + (1 + a_1^2 - 2a_1)\sigma_2^2 \\ &= a_1^2(\sigma_1^2 + \sigma_2^2) - 2a_1\sigma_2^2 + \sigma_2^2 \end{aligned}$$

Taking the derivative with respect to a_1 and setting it to zero gives

$$\begin{aligned} a_1 &= \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \\ a_2 &= \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}. \end{aligned}$$

The minimum variance is $a_1^2\sigma_1^2 + a_2^2\sigma_2^2$ with above values of a_1 and a_2 which yields

$$\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

Part d)

For $T = 2$, the minimum variance

$$\frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

This answer is to be interpreted as the product of variance divided by the sum of variance.

Substituting numerical values gives 2/3.

For $T = 3$, treat the optimal solution for the $T = 2$ case as a single sensor which measures the temperature with additive noise which has mean zero and variance 2/3.

$$\frac{2/3 * 3}{2/3 + 9/3} = \frac{6}{11}$$

4 (a) Describe the *parametric* approach to power spectrum estimation, including a brief discussion of the ARMA model, the AR model, and their power spectra. [30%]

Answer:

Bookwork as follows from lecture notes (more detailed than required):

- Periodogram-based methods can lead to biased estimators with large variance
- If the physical process which generated the data is known or can be well approximated, then a parametric model can be constructed
- Careful estimation of the parameters in the model can lead to power spectrum estimates with improved bias/variance.
- We will consider spectrum estimation for LTI systems driven by a white noise input sequence.
- If a random process $\{X_n\}$ can be modelled as white noise exciting a filter with frequency response $H(e^{j\omega T})$ then the spectral density of the data can be expressed as:

$$\boxed{S_X(e^{j\omega T}) = \sigma_w^2 |H(e^{j\omega T})|^2}$$

where σ_w^2 is the variance of the white noise process. [It is usually assumed that $\sigma_w^2 = 1$ and the scaling is incorporated as gain in the frequency response]

- We will study models in which the frequency response $H(e^{j\omega T})$ can be represented by a finite number of parameters which are estimated from the data.
- Parametric models need to be chosen carefully - an inappropriate model for the data can give misleading results

ARMA Models

A quite general representation is the autoregressive moving-average (ARMA) model:

- The ARMA(P,Q) model difference equation representation is:

$$\boxed{x_n = -\sum_{p=1}^{P} a_p x_{n-p} + \sum_{q=0}^{Q} b_q w_{n-q}} \quad (1)$$

where:

a_p are the AR parameters,

b_q are the MA parameters

and $\{W_n\}$ is a zero-mean stationary white noise process with unit variance, $\sigma_w^2 = 1$.

•Clearly the ARMA model is a pole-zero IIR filter-based model with transfer function

$$H(z) = \frac{B(z)}{A(z)}$$

where:

$$A(z) = 1 + \sum_{p=1}^{P} a_p z^{-p}, \quad B(z) = \sum_{q=0}^{Q} b_q z^{-q}$$

With $Q = 0$ we have the AR model and with $P = 0$ the MA model.

•Unless otherwise stated we will always assume that the filter is stable, i.e. the poles (solutions of $A(z) = 0$) all lie *within* the unit circle (we say in this case that $A(z)$ is *minimum phase*). Otherwise the autocorrelation function is undefined and the process is technically *non-stationary*.

•Hence the power spectrum of the ARMA process is:

$$S_X(e^{j\omega T}) = \frac{|B(e^{j\omega T})|^2}{|A(e^{j\omega T})|^2}$$

The ARMA model is quite a flexible and general way to model a stationary random process:

•The poles model well the *peaks* in the spectrum (sharper peaks implies poles closer to the unit circle)

•The zeros model troughs in the spectrum

•Complex spectra can be approximated well by large model orders P and Q

Note however, that model order determination is critical for ARMA modelling and an ARMA model may not be appropriate for certain datasets.

(b) Write down an expression for the prediction error at time index n for an autoregressive model of order P.

By considering minimisation of an appropriate function of this prediction error corresponding to a finite length of data $x_0, \ldots, x_{N-1}$, show that the vector of autoregressive parameters **a** may be estimated as

$$\mathbf{a} = -(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{x}$$

where **x** and **X**, which are a vector and matrix containing observed data values, should be carefully defined.

Explain how the covariance and autocorrelation methods can be obtained from this method and briefly summarise the properties of each.

[40%]

Solution:

The AR model can be written equivalently as:

$$x_n = -\sum_{p=1}^{P} a_p x_{n-p} + e_n \tag{2}$$

where e_n is a white noise sequence having variance $\sigma_e^2 = b_0^2$.

An alternative interpretation of this equation is that:

$$x_n = \hat{x}_n + e_n$$

where:

$$\hat{x}_n = -\sum_{p=1}^{P} a_p x_{n-p}$$

is a prediction of x_n from previous data and the term e_n is the *prediction error*. In terms of e_n, equation 2 becomes:

$$e_n = x_n + \sum_{p=1}^{P} a_p x_{n-p}$$

Suppose we write this equation for all values of n such that

$$n_I \leq n \leq n_F$$

All of these equations may be expressed in matrix notation as:

$$\boxed{\mathbf{e} = \mathbf{x} + \mathbf{X}\mathbf{a}}$$

 (TURN OVER for continuation of Question 4

where:

$$\mathbf{e} = \begin{bmatrix} e_{n_I} \\ e_{n_I+1} \\ \vdots \\ e_{n_F} \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} x_{n_I} \\ x_{n_I+1} \\ \vdots \\ x_{n_F} \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} x_{n_I-1} & x_{n_I-2} & \cdots & x_{n_I-P} \\ x_{n_I} & x_{n_I-1} & \cdots & x_{n_I-P+1} \\ \vdots & \vdots & & \vdots \\ x_{n_F-1} & x_{n_F-2} & \cdots & x_{n_F-P} \end{bmatrix} \tag{3}$$

and

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix}$$

Two particular cases for n_I and n_F will be considered shortly.

If we wish to find the AR parameters which fit the observed data 'best', then it would seem reasonable to minimize the prediction error terms (i.e. an 'ideal' model for the data would have zero prediction errors).

A convenient way to achieve this by choosing the parameter vector $\mathbf{a}$ which minimizes the total squared prediction error, E:

$$\mathscr{E} = \sum_{n=n_I}^{n_F} e_n^2 = \mathbf{e}^{\mathrm{T}}\mathbf{e}$$

where $\mathbf{e}^{\mathrm{T}}$ denotes the transpose of $\mathbf{e}$.

We recognise this as a standard least squares estimation problem, as studied in 1B Linear Algebra, so we obtain the solution immediately:

$$\boxed{\mathbf{a} = -(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{x}}$$

An alternative derivation seeks to minimize the function directly. In order to find the minimum of $\mathbf{e}^{\mathrm{T}}\mathbf{e}$ with respect to all of the elements of $\mathbf{a}$ we must solve the P simultaneous equations:

$$\frac{\partial(\mathbf{e}^{\mathrm{T}}\mathbf{e})}{\partial a_i} = 0, \; i = 1,2,...,P$$

We can express the same thing in vector notation as:

$$\frac{\partial(\mathbf{e}^{\mathrm{T}}\mathbf{e})}{\partial \mathbf{a}} = \mathbf{0}_P$$

where

$$\frac{\partial(\mathbf{e}^{\mathrm{T}}\mathbf{e})}{\partial \mathbf{a}} = \begin{bmatrix} \frac{\partial(\mathbf{e}^{\mathrm{T}}\mathbf{e})}{\partial a_1} \\ \frac{\partial(\mathbf{e}^{\mathrm{T}}\mathbf{e})}{\partial a_2} \\ \vdots \\ \frac{\partial(\mathbf{e}^{\mathrm{T}}\mathbf{e})}{\partial a_P} \end{bmatrix} \quad \text{and} \quad \mathbf{0}_P = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

[so $\frac{\partial(\mathbf{e}^{\mathrm{T}}\mathbf{e})}{\partial \mathbf{a}}$ is just the *gradient vector* from vector calculus].

Now, expand $\mathbf{e}^{\mathrm{T}}\mathbf{e}$ and differentiate:

$$\mathbf{e} = \mathbf{x} + \mathbf{X}\mathbf{a}$$

$$\mathbf{e}^{\mathrm{T}}\mathbf{e} = (\mathbf{x} + \mathbf{X}\mathbf{a})^{\mathrm{T}}(\mathbf{x} + \mathbf{X}\mathbf{a})$$

(TURN OVER for continuation of Question 4

$$= \mathbf{x}^T\mathbf{x} + 2\mathbf{x}^T\mathbf{X}\mathbf{a} + \mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a}$$

$$\frac{\partial(\mathbf{e}^T\mathbf{e})}{\partial\mathbf{a}} = 2\frac{\partial(\mathbf{x}^T\mathbf{X}\mathbf{a})}{\partial\mathbf{a}} + \frac{\partial(\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a})}{\partial\mathbf{a}}$$

$$= 2\mathbf{X}^T\mathbf{x} + 2\mathbf{X}^T\mathbf{X}\mathbf{a}$$

Here we have used two standard results from matrix/vector calculus:

$$\frac{\partial(\mathbf{b}^T\mathbf{a})}{\partial\mathbf{a}} = \frac{\partial(\mathbf{a}^T\mathbf{b})}{\partial\mathbf{a}} = \mathbf{b} \text{ and } \frac{\partial(\mathbf{a}^T\mathbf{B}\mathbf{a})}{\partial\mathbf{a}} = 2\mathbf{B}\mathbf{a}$$

for constant vector $\mathbf{b}$ and symmetric matrix $\mathbf{B}$.

[You can verify these by differention term by term. See, for example Therrien, Appendix A.]

For a maximum or minimum of $\mathbf{e}^T\mathbf{e}$:

$$\frac{\partial(\mathbf{e}^T\mathbf{e})}{\partial\mathbf{a}} = \mathbf{0}$$

Therefore,

$$2\mathbf{X}^T\mathbf{x} + 2\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{0} \tag{4}$$

and finally, provided $\mathbf{X}$ is full rank,

$$\boxed{\mathbf{a} = -(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{x}}$$

(cont.

It is of interest to consider the form of the terms in this expression for the AR coefficient vector **a**. First notice that **X** (see equation 3) may be expressed as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{-1} & \mathbf{x}_{-2} & \cdots & \mathbf{x}_{-P} \end{bmatrix}$$

where

$$\mathbf{x}_{-p} = \begin{bmatrix} x_{n_I-p} & x_{n_I-p+1} & \cdots & x_{n_F-p} \end{bmatrix}^{\mathrm{T}}$$

Hence we can write

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} \mathbf{x}_{-1}^{\mathrm{T}} \\ \mathbf{x}_{-2}^{\mathrm{T}} \\ \cdots \\ \mathbf{x}_{-P}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{-1} & \mathbf{x}_{-2} & \cdots & \mathbf{x}_{-P} \end{bmatrix}$$

whose (i, j)th element is

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})_{(i,j)} = \mathbf{x}_{-i}^{\mathrm{T}}\mathbf{x}_{-j} = \sum_{n=n_I}^{n_F} x_{n-i}x_{n-j} \tag{5}$$

Similarly, $\mathbf{X}^{\mathrm{T}}\mathbf{x}$ can be expressed as:

$$\mathbf{X}^{\mathrm{T}}\mathbf{x} = \begin{bmatrix} \mathbf{x}_{-1}^{\mathrm{T}} \\ \mathbf{x}_{-2}^{\mathrm{T}} \\ \cdots \\ \mathbf{x}_{-P}^{\mathrm{T}} \end{bmatrix} \mathbf{x}$$

whose ith element is

$$(\mathbf{X}^{\mathrm{T}}\mathbf{x})_{(i)} = \mathbf{x}_{-i}^{\mathrm{T}}\mathbf{x} = \sum_{n=n_I}^{n_F} x_{n-i}x_n$$

Notice also that equation 4 can be rewritten as:

$$\mathbf{X}^{\mathrm{T}}\mathbf{x} + \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a} = \mathbf{X}^{\mathrm{T}}(\mathbf{x} + \mathbf{X}\mathbf{a}) = \mathbf{X}^{\mathrm{T}}\mathbf{e} = \mathbf{0}$$

Hence we have that $\mathbf{x}_{-i}^{\mathrm{T}}\mathbf{e} = 0$ for $i = 1, 2, ..., P$, i.e. at the least squares solution the error is orthogonal to all the columns of **X**. This is a well-known property of least squares.

We now consider two commonly used methods, based on different values of n_I and n_F. In both cases it is assumed that exactly N data points are available, $x_0, x_1, ..., x_{N-1}$.

Covariance method

 (TURN OVER for continuation of Question 4

•The covariance method minimizes only those error terms which can be fully calculated from the data

•Examine the error equation:

$$e_n = x_n + \sum_{p=1}^{P} a_p x_{n-p}$$

•The first error term that can be fully calculated is e_P and the last is e_{N-1}.

•Hence $n_I = P$ and $n_F = N-1$ in the squared error equation:

$$\mathscr{E}^{\mathrm{C}} = \sum_{n=P}^{N-1} e_n^2$$

•The resulting matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is *not* Toeplitz. Although fast algorithms exist to solve for **a**, they are much more complex than for the autocorrelation method.

•The AR parameter estimate is not guaranteed to be stable

•The method is intuitively appealing as it does not attempt to make guesses about data that aren't observed.

•The covariance method is a good approximation for moderately large N to the true maximum likelihood estimate [see Module 4F6 - Detection and Estimation].

Autocorrelation method

•In the autocorrelation method $n_I = 0$ and $n_F = N+P-1$

•Hence the squared error minimized is:

$$\mathscr{E}^{\mathrm{A}} = \sum_{n=0}^{N+P-1} e_n^2$$

•To calculate these error terms requires data before $n = 0$ and after $n = N-1$. These data points are assumed to be zero.

•Consider the elements $(\mathbf{X}^{\mathrm{T}}\mathbf{X})_{(i,j)}$ (equation 5), when $i \geq j$:

$$\begin{aligned}(\mathbf{X}^{\mathrm{T}}\mathbf{X})_{(i,j)} &= \sum_{n=0}^{N+P-1} x_{n-i}x_{n-j} \\ &= \sum_{n=i}^{N+j-1} x_{n-i}x_{n-j} \\ &= \sum_{n'=0}^{N+j-i-1} x_{n'}x_{n'+(i-j)}\end{aligned}$$

and since $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is symmetrical, for $j \geq i$:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})_{(i,j)} = \sum_{n'=0}^{N+i-j-1} x_{n'} x_{n'+(j-i)}$$

Now, letting $k = |i - j|$ we have overall:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}_{(i,j)} = \sum_{n=0}^{N-k-1} x_n x_{n+k}$$

Hence $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is *Toeplitz*, which means that the efficient Levinson recursion ($O(P^2)$) can be used to solve for **a**.

- Note that the autocorrelation method is equivalent to estimating the autocorrelation function using the *biased* estimate and then solving the matrix Yule-Walker equations directly
- The parameter estimate is guaranteed to be stable
- However, the assumption of zeros before the start and after the end of the data are likely to make the estimate less accurate than the covariance method for small N

A discrete time function is defined as:

$$x_0 = 1, x_1 = -0.9, x_2 = 0.81, \ldots$$

i.e. the general term is $x_n = (-0.9)^n$. Write down an autoregressive model with order $P = 1$ which fits this function perfectly (i.e. with zero prediction error for any $n > 0$.

Now compute estimates of order $P = 1$ autoregressive models from data points $x_0, x_1, \ldots x_{N-1}$, using both the autocorrelation method and the covariance method.

Comment on the similarity between these estimates and the model which fits the data perfectly. What happens to this similarity as N tends to infinity?

[40%]

Solution: The solution to part b) with $P = 1$ has:

$$a_1 = \frac{\sum_{n_I}^{n_F} x_k x_{k-1}}{\sum_{n_I}^{n_F} x_{k-1}^2}$$

For the covariance method, set $n_I = 1$ and $n_F = N - 1$. Then

$$a_1^{\mathrm{cov}} = \frac{\sum_{n_I}^{n_F} (-0.9)^k (-0.9)^{k-1}}{\sum_{n_I}^{n_F} x_{k-1}^2} = \frac{\sum_{n_I}^{n_F} (-0.9)^{2k-1}}{\sum_{n_I}^{n_F} (-0.9)^{2k-2}} = -0.9$$

 (TURN OVER for continuation of Question 4

i.e. the covariance method always matches the perfect fitting model, since it attempts to minimise the prediction error within the data (i.e. not making assumptions of data outside the measured block).

For the autocorrelation method, set $n_I = 0$ and $n_F = N$, but note that some terms in the summations are zero, since we assume $x_{-1} = 0$ and $x_N = 0$. Allowing for these zero terms, we get:

$$\begin{aligned} a_1^{\text{auto}} &= \frac{\sum_1^{N-1}(-0.9)^{2k-1}}{\sum_1^{N}(-0.9)^{2k-2}} \\ &= \frac{\frac{(-0.9)(1-(-0.9)^{N-1})}{1-(-0.9)}}{\frac{(1-(-0.9)^{N})}{1-(-0.9)}} \\ &= -0.9\frac{1-(-0.9)^{N-1}}{1-(-0.9)^{N}} \neq -0.9 \end{aligned}$$

Hence the autocorrelation method is in error. However, as $N \to \infty$ it clearly tends to the correct answer, since the terms $(-0.9)^N$, etc. go to zero in this case.

5 (a) Define the bias and variance of an estimator for a random quantity, explaining how they can be used to evaluate the estimator's performance. [20%]

Solution:

Definition: Unbiased Estimators

- An estimator $\hat{\theta}$ of a random quantity θ is *unbiased* if the expected value of the estimate equals the true value, i.e.

$$E[\hat{\theta}] = \theta$$

Otherwise the estimator is termed *biased.*

- The *variance* of an estimator measures how much variability an estimator has around its mean (expected) value. It is defined as:

$$\text{var}(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

- Hence we expect that a 'good' estimator will make some suitable trade-off between low bias and low variance.

(b) In a power spectrum estimation method for a wide-sense stationary and ergodic random process, the data is first windowed using a window function w_n having length N, i.e. $w_n = 0$ for $n < 0$ and $n > N-1$.

The autocorrelation function is then estimated as

$$\hat{R}_{XX}[k] = \frac{1}{N} \sum_{n=0}^{N-1-k} (w_n x_n)(w_{n+k} x_{n+k}), \text{ for } k = 0, 1, ..., N-1$$

and

$$\hat{R}_{XX}[k] = \hat{R}_{XX}[-k], \text{ for } k = -1, -2, ..., -N+1$$

where the x_n are measured values drawn from the random process.

Show that the expected value of the autocorrelation function estimate is given by

$$E[\hat{R}_{XX}[k]] = R_{XX}[k] \frac{1}{N} \sum_{n=0}^{N-1-k} (w_n w_{n+k})$$

where $R_{XX}[k]$ is the true autocorrelation function for the process.

Is this an unbiased estimate? [20%]

Solution:

$$E[\hat{R}_{XX}[k]] = E[\frac{1}{N}\sum_{n=0}^{N-1-k}(w_n x_n)(w_{n+k}x_{n+k})$$
$$= \frac{1}{N}\sum_{n=0}^{N-1-k} w_n w_{n+k} E[x_n)(x_{n+k})]$$
$$= \frac{1}{N}\sum_{n=0}^{N-1-k} w_n w_{n+k} R_{XX}[k]$$
$$= R_{XX}[k]\frac{1}{N}\sum_{n=0}^{N-1-k}(w_n w_{n+k})$$

i.e. it is biased in general.

A. B. Show that the expected value of the corresponding power spectrum estimate is:

$$E[\hat{S}_X(e^{j\omega T})] = \frac{1}{2\pi N} S_X(e^{j\omega T}) * |W(e^{j\omega T})|^2$$

where $S_X(e^{j\omega T})$ is the true power spectrum of the random process, $W(e^{j\omega T})$ is the DTFT of the window function w_n, and $*$ denotes the convolution operator. [40%]

Solution:

We have from lectures on the periodogram that

$$E[\hat{S}_X(e^{j\omega T})] = E[DTFT\{\hat{R}_{XX}[k]\}] = DTFT\{E[\hat{R}_{XX}[k]]\}$$

Then, note that

$$\sum_{n=0}^{N-1-k}(w_n w_{n+k}) = \{w_n\} * \{w_{-n}\}$$

whose DTFT is:

$$W(e^{j\omega T})W^*(e^{j\omega T}) = |W(e^{j\omega T})|^2$$

But, this term is multiplied (in time) with $R_{XX}[k]$. Hence overall the DTFT is:

$$E[\hat{S}_X(e^{j\omega T})] = \frac{1}{2\pi N} S_X(e^{j\omega T}) * |W(e^{j\omega T})|^2,$$

as required

C. How might this modified autocorrelation method be used improve the performance and properties of the standard periodogram estimator? [20%]

Solution: First, note that the estimate will always be positive, which is an improvement in properties. Second, note that the convolution of a suitable window spectrum $|W(e^{j\omega T})|^2$ will smooth out the spectrum and hence give a trade-off between noise-like randomness against frequency resolution.

END OF PAPER