E4.13
AS2
SO15
ISE4.31

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2007

MSc and EEE/ISE PART IV: MEng and ACGI

Corrected Copy

SPECTRAL ESTIMATION AND ADAPTIVE SIGNAL PROCESSING

Wednesday, 2 May 10:00 am

Time allowed: 3:00 hours

There are FIVE questions on this paper.

Answer ONE of questions 1,2 and TWO of questions 3,4,5.

All questions carry equal marks

Any special instructions for invigilators and information for candidates are on page 1.

Examiners responsible First Marker(s) : D.P. Mandic
Second Marker(s) : M.K. Gurcan

© University of London 2007

1) Consider the problem of periodogram based spectral estimation.

a) Write down the expression for the periodogram based power spectrum estimate. In your own words explain the operation of the periodogram. How is this estimate related to the estimate of the autocorrelation function? [4]

b) The ways to improve the properties of periodogram based spectrum estimation include: averaging over a set of periodograms, applying window functions to the data, and overlapping windowed data segments.

i) Explain how the averaging over a set of periodogams influences the bias, variance and resolution of the periodogram. [2]

ii) Explain how applying different windows to the data influences the bias, variance and resolution of the periodogram. [2]

iii) Explain how overlapping of windowed data segments influences the bias, variance and resolution of the periodogram. [2]

c) Consider the problem of estimating the power spectrum of two sinusoids in white Gaussian noise $w[n] \sim \mathcal{N}(0,1)$, given by

$$x[n] = sin\,(0.2\pi n + \Phi_1) + 2sin\,(0.32\pi n + \Phi_2) + w[n], \qquad n = 0, 1, \ldots, N-1$$

The total number of data points is $N = 512$. Assuming that the successive sequences are offset by D points and that each sequence is $L = 128$ points long, and if K sequences cover the entire N points, then

$$N = L + D(K-1)$$

i) If the sequences are allowed to overlap by 50% ($D = L/2$), explain how we can maintain the same resolution as Bartlett's method while reducing the variance. [4]

ii) If the sequences are allowed to overlap by 50% ($D = L/2$), explain how we can increase the resolution while reducing the variance. [4]

iii) Sketch a general shape of the estimated power spectra for cases *i*) and *ii*). [2]

2) One class of spectrum estimation methods assumes a harmonic model of the process (Pisarenko, MUSIC, Principal Components spectrum estimation).

a) Explain the mechanism behind these techniques. [4]

b) Principal components spectrum estimation is based on the eigendecomposition of the autocorrelation matrix $\mathbf{R}_{xx}$, given by

$$\mathbf{R}_{xx} = \sum_{i=1}^{M} \lambda_i \mathbf{v}_i \mathbf{v}_i^H = \underbrace{\sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^H}_{signal} + \underbrace{\sum_{i=p+1}^{M} \lambda_i \mathbf{v}_i \mathbf{v}_i^H}_{noise}$$

where $\lambda_i, i = 1, \ldots, M$ are the eigenvalues and $\mathbf{v}_i$ the eigenvectors of $\mathbf{R}_{xx}$.

i) Write down the expression for power spectrum estimate based on the signal subspace within the above decomposition. [2]

ii) Explain in your own words the difference between this method and methods based on the noise subspace. [4]

iii) Explain how we can use principal component analysis (PCA) in conjunction with other spectrum estimation methods. [2]

iv) Sketch a simple example combining PCA with the maximum entropy and autoregressive spectrum estimation methods. [2]

c) We would like to estimate the power spectrum of the autoregressive (AR) process of order two ($AR(2)$), given by

$$x[n] = a_1 x[n-1] + a_2 x[n-2] + w[n]$$

where $w[n]$ is a unit variance white noise. However, the measurements of $x[n]$ are noisy, and we can only observe the process

$$y[n] = x[n] + v[n]$$

where $v[n]$ is uncorrelated with $x[n]$ and can be modelled as a moving average (MA) process. The autocorrelation functions for $y[n]$ and $v[n]$ are given respectively by

$$r_{yy}[0] = 5 \quad r_{yy}[1] = 2 \quad r_{yy}[2] = 0 \quad r_{yy}[3] = -1 \quad r_{yy}[4] = 0.5$$
$$r_{vv}[0] = 3 \quad r_{vv}[1] = 1$$

Explain how you would estimate the power spectrum of $x[n]$.

(Hint: for uncorrelated processes $r_{yy}[n] = r_{xx}[n] + r_{vv}[n]$)

[6]

3) Consider the class of linear finite impulse response (FIR) adaptive filters.

a) Figure 3.1 shows the block diagram of a general adaptive system. Write down the equations explaining the input–output relationships of this structure. Explain in your own words the operation of such a structure. [6]

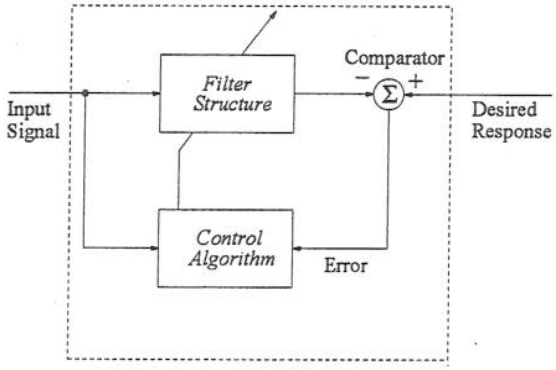
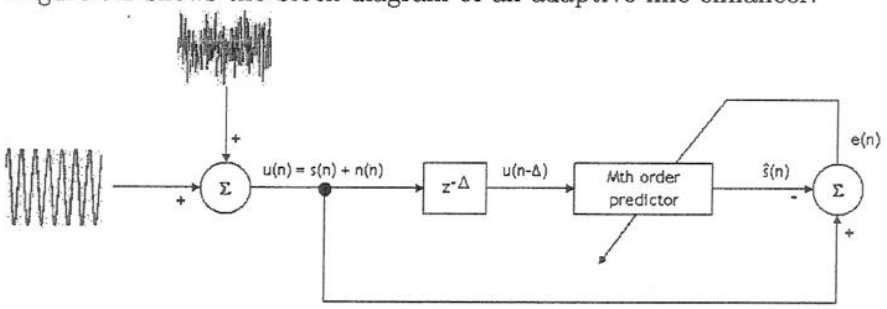


Figure 3.1: General adaptive system

b) Figure 3.2 shows the block diagram of an adaptive line enhancer.



i) State and explain the class of applications for this adaptive filtering scheme. [3]

ii) Write down the equations describing the operation of this structure. Comment on the role of the delay operator $z^{-\Delta}$. [3]

c) The predictor within this structure is trained using the Normalised Least Mean Square (NLMS) algorithm.

i) Derive the NLMS algorithm by minimising the a posteriori prediction error. (Hint: Expand the a posteriori error $e(n+1)$ using the Taylor series expansion around $e(n)$) [4]

ii) Explain the role of the learning rate (step size) within NLMS and provide the stability bounds for step size parameter μ. [4]

4) One way to derive the steepest descent algorithm for solving the normal equations $\mathbf{R}_{xx}\mathbf{w} = \mathbf{r}_{dx}$ is to use a power series expansion for the inverse of $\mathbf{R}_{xx}$. This expansion is

$$\mathbf{R}_{xx}^{-1} = \mu \sum_{k=0}^{\infty} (\mathbf{I} - \mu \mathbf{R}_{xx})^k$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{r}_{dx}$ is the crosscorrelation between the desired input d and the input signal x, and μ is a positive constant. In order for this expression to converge, $\mathbf{R}_{xx}$ must be positive definite and the constant μ must lie in the range

$$0 < \mu < \frac{2}{\lambda_{max}}$$

where λ_{max} is the largest eigenvalue of $\mathbf{R}_{xx}$.

a) Let

$$\mathbf{R}_{xx}^{-1}(n) = \mu \sum_{k=0}^{n} (\mathbf{I} - \mu \mathbf{R}_{xx})^k$$

be the n–th order approximation to $\mathbf{R}_{xx}^{-1}$, and let

$$\mathbf{w}_n = \mathbf{R}_{xx}^{-1}(n)\mathbf{r}_{dx}$$

be the n–th order approximation to the desired solution $\mathbf{w} = \mathbf{R}_{xx}^{-1}\mathbf{r}_{dx}$. Express $\mathbf{R}_{xx}^{-1}(n+1)$ in terms of $\mathbf{R}_{xx}^{-1}(n)$, and show how this may be used to derive the steepest descent algorithm, given by

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \mu \left[\mathbf{R}_{xx}\mathbf{w}_n - \mathbf{r}_{dx}\right] \qquad [8]$$

(Hint: Multiply both sides of the ACF inverse by $\mathbf{r}_{dx}$)

b) If the statistics of $x(n)$ are unknown, then $\mathbf{R}_{xx}$ is unknown and the expansion for $\mathbf{R}_{xx}^{-1}$ in part a) cannot be evaluated. However, suppose we approximate $\mathbf{R}_{xx} = E\{\mathbf{x}(n)\mathbf{x}^T(n)\}$ at time n as follows

$$\hat{\mathbf{R}}_{xx}(n) = \mathbf{x}(n)\mathbf{x}^T(n)$$

and use, as the n–th order approximation to $\mathbf{R}_{xx}^{-1}$

$$\hat{\mathbf{R}}_{xx}^{-1}(n) = \mu \sum_{k=0}^{n} \left[\mathbf{I} - \mu \mathbf{x}(k)\mathbf{x}^T(k)\right]^k$$

Express $\hat{\mathbf{R}}_{xx}^{-1}(n+1)$ in terms of $\hat{\mathbf{R}}_{xx}^{-1}(n)$ and use this expression to derive a recursion for the weight vector $\mathbf{w}_{n+1}$. [8]

c) Compare your recursion derived in part b) to the LMS algorithm. [4]

5) The Least Mean Square (LMS) adaptive filter minimises the instantaneous squared error

$$J(n) = \frac{1}{2} e^2(n)$$

Consider the modified cost function

$$J'(n) = \frac{1}{2} e^2(n) + \frac{\beta}{2} \mathbf{w}^T(n)\mathbf{w}(n)$$

where $\beta > 0$, $\mathbf{w}(n)$ is the filter weight vector, and $(\cdot)^T$ the vector transpose operator.

a) Derive the LMS coefficient update equation for $\mathbf{w}(n)$ that mimimises $J'(n)$. [6]

b) The cost function $J'(n)$ has two terms, one minimising the mean squared error and the second penalising for large values of the weight vector. Explain in your own words the principle behind such a cost function. [3]

c) Determine the condition on the step size that will ensure that $\mathbf{w}(n)$ converges in the mean.

(Hint: Apply the expectation operator to $\mathbf{w}(n+1)$ to yield $E\{\mathbf{w}(n+1)\}$, and set the norm of the homogeneous term to be < 1) [3]

d) Hybrid filters consist of a convex combination of two standard LMS type adaptive filters with different learning rates, which are updated separately, based on their own instantaneous output errors $e_1(n)$ and $e_2(n)$. The outputs of the consitutive filters $y_1(n)$ and $y_2(n)$ are combined to give the output of the hybrid filter

$$y_H(n) = \lambda\, y_1(n) + (1 - \lambda)\, y_2(n)$$

where $0 \leq \lambda \leq 1$ is the convex mixing parameter. If λ is made adaptive, show that

$$\lambda(n+1) = \lambda(n) + \mu_\lambda\, e(n)\ (y_1(n) - y_2(n))$$

Explain the similarities and differences between the hybrid filter and the adaptive filter based on cost function $J'(n)$ from a). [8]

E4.13 / Ex 4.21 / 2015

SPECTRAL ESTIMATION and ADAPTIVE SIGNAL PROCESSING

2007

1/5

Solutions:

1) a) [**bookwork**]

$$\hat{\mathbf{P}}_{per}(f) = \frac{1}{N}\left|\sum_{k=0}^{N-1} x[k]e^{-\jmath 2\pi fk}\right|^2$$

Clearly the periodogram is related to the estimation of the ACF, te better this estimate the better the spectrum estimate. From the functional expression, the standard periodogram uses a rectangular window to process the data, which introduces problems due to the convolution in the frequency domain between the true power spectrum and the *sinc* function. The significant sidelobe of the sinc in the frequency domain is the main problem with this approach.

b) [**bookwork and intuitive reasoning**]

i) The periodogram is asymptotically unbiased. By averaging a number of periodograms, very much like with any other estimation problem, the variance reduces up to the factor given by the number of averages (best for uncorrelated data). The resolution is proportional to the number of data points, hence it reduces appropriately ($0.89 \times K\frac{2\pi}{N}$)

ii) The windowed periodogram remains asymptotically unbiased. The windowing offers a trade off between spectral resolution (main lobe width) and spectral masking (sidelobe amplitude). The variance estimate using this method, however is not consistent.

iii) By overlapping segments of (possibly windowed) data, we combine the properties of the above two modifications. By choosing appropriately the size and number of windows and the degree of overlapping, we can vritually span the whole range of possible combinations of periodogram modifications.

c) [**Analysis of new example**]

i) The variance is reduced by means of averaging, ideally (for uncorrelated data) by the number of averaged segments. By allowing the segments to overlap, we generate a larger number of segments, which in turn help with the number of estimates to be averaged. If there is no overlap ($D = L$), we have $K = N/L$ sections of lenth L (Bartlett's method). If the sequences are overlapping by 50%, then $D = L/2$ and we may form $K = 2N/L - 1$ sections of length L this maintaining the same resolution as Bartlett's method while doubling the number of modified periodograms that are averaged, thereby reducing the variance.

ii) with 50% overlap we can also form $K = N/L - 1$ sequences of length $2L$, thus increasing the resolution, whle maintaining the same variance as Bartlett's method. Therefore, by allowing the segments to overlap, we can trade a reduction in variance for a reduction in resolution.

iii) Straightforward from the corresponding examples from course notes.

2) a) [**bookwork**]
These methods assume that power spectum at a discrete set of frequencies has physical meaning (information bearing such is in radar, sonar, speech). The idea is to use eigendecomposition to decompose the autocorrelation matrix of the (noisy) data into the signal related part and noise related part. Based on the orthogonality between the useful signal and noise we may use the noise subspace or signal subspace for power spectrum estimation at the desired set of frequencies of interest. The so produced power spectrum estimate need not be accurate outside the discrete set of frequencies of interest.

b) [**bookwork and intuitive reasoning**]
i)

$$\hat{R}_s = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^H$$

ii) The methods based on the noise subspace estimation produce a peak in the spectrum for a discrete set of frequencies of interest. The PCA based spectrum estimate on the other hand imposes a rank constraint on the signal subspace and provides an estimate of the ACF of the signal.

iii) Since this method produces an estimate of ACF, it can be used in conjunction with other standard methods which rely on an estimate of ACF.

iv) $\mathbf{R}_{ss}$ from above can be used directly within autoregressive spectrum estimation since it provides an estimate of ACF. It can also be used within MEM, due to the duality between the autoregressive and MEM spectrum estimation.

c) [**new example**]
Due to the orthogonality between signal and noise and the MA noise model we have

$$\begin{aligned} r_{xx}[n] &= r_{yy}[n] - r_{vv}[n] \quad \text{which gives} \\ r_{xx}[0] = 2 \qquad & r_{xx}[1] = 1 \quad r_{xx}[2] = 0 \quad r_{xx}[3] = -1 \quad r_{xx}[4] = 0.5 \end{aligned}$$

We therefore know the dimension of the signal subspace and by performing since we know the ACF of the data, we can employ any ACF based spectrum estimation method.

3) a) [**bookwork**]

- Filter architecture (FIR, IIR, linear, nonlinear)
- Input $\{x\}$, output $\{y\}$, and desired $\{d\}$ signal
- Filter function:- prediction, system identification, inverse system modelling, noise cancellation
- Adaptation:- Based on the error $e(n) = d(n) - y(n)$

b) i) and ii)[**bookwork, worked example**]

- ☐ Adaptive line enhancement (ALE) refers to the case where a noisy signal, $u(n) = 'sin(n)' + 'wn(n)'$
- ☐ ALE consists of a de-correlation stage, symbolised by $z^{-\Delta}$ and an adaptive predictor
- ☐ The de-correlation stage attempts to remove any correlation that may exist between the samples of noise, by shifting them Δ samples apart
- ☐ A phase shift introduced (input Δ steps behind)
- ☐ This way, if the decorrelation is performed in a satisfactory way, denoising boils down to adaptive prediction.

c) i) [**bookwork, coursework, and worked example**]

- Start from the **independence** assumptions, that is that $\mathbf{w}$, $\mathbf{x}$, μ, and e are independent and mutually Gaussian
- **Normalisation $\Leftrightarrow$ minimisation** of the *a posteriori* error $e(k+1)$

$$e(k+1) = \mathbf{x}^T(k)\mathbf{w}(k+1)$$

- Perform Taylor Series Expansion around $e(k)$ to obtain

$$\begin{aligned} e(k+1) &= e(k) + \sum_{i=1}^{N} \frac{\partial e(k)}{\partial w_i(k)} \Delta w_i(k) + \sum_{i=1}^{N} \frac{\partial e(k)}{\partial x_i(k)} \Delta x_i(k) \\ &\quad + \frac{\partial e(k)}{\partial d(k)} \Delta d(k) + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\partial^2 e(k)}{\partial w_i(k) \partial x_j(k)} \Delta w_i(k) \Delta x_j(k) + \cdots \end{aligned}$$

- Now the partial derivatives from TSE expansion become

$$\begin{aligned} \frac{\partial e(k)}{\partial w_i(k)} &= -x(k-i+1) = -x_i(k), \quad i = 1, 2, \ldots, N \\ \Delta w_j(k) &= \mu e(k) x_j(k) \end{aligned}$$

- And finally

$$e(k+1) = e(k) - \sum_{i=1}^{N} \mu e(k) x_i^2 = e(k) \left[1 - \mu \parallel \mathbf{x}(k) \parallel_2^2 \right]$$

- to give

$$\eta(k) = \frac{\mu}{\parallel \mathbf{x}(k) \parallel_2^2}$$

ii) [**bookwork and intuitive reasoning**] The learning rate is normalised by the power of the input vector in the memory of the filter. This is a very simple estimate of the autocorrelation function, and as such, the normalisation aims at decorrelating the tap input in order to provide faster convergence. For stability $0 < \mu \leq 2$.

4) [**New example and bookwork**]

a) Using the n–th order approximation to $\mathbf{R}_{xx}^{-1}$,

$$\mathbf{R}_{xx}^{-1}(n) = \mu \sum_{k=0}^{n} (\mathbf{I} - \mu \mathbf{R}_{xx})^k$$

we have

$$\mathbf{R}_{xx}^{-1}(n+1) = \mu \sum_{k=0}^{n+1} (\mathbf{I} - \mu \mathbf{R}_{xx})^k = \mu (\mathbf{I} - \mu \mathbf{R}_{xx}) \sum_{k=0}^{n} (\mathbf{I} - \mu \mathbf{R}_{xx})^k + \mu \mathbf{I}$$

Therefore

$$\mathbf{R}_{xx}^{-1}(n+1) = (\mathbf{I} - \mu \mathbf{R}_{xx})\, \mathbf{R}_{xx}^{-1}(n) + \mu \mathbf{I}$$

Multiplying both sides of the equation ty $\mathbf{r}_{dx}$ on the right, we have

$$\mathbf{w}_{n+1} = (\mathbf{I} - \mu \mathbf{R}_{xx})\, \mathbf{w}_n + \mu \mathbf{r}_{dx}$$

which is the steepest descent algorithm.

b) Using the approximation $\hat{\mathbf{R}}_{xx} = \mathbf{x}(n)\mathbf{x}^T(n)$, we have

$$\hat{\mathbf{R}}_{xx}^{-1}(n+1) = \left(\mathbf{I} - \mu \mathbf{x}(n)\mathbf{x}^T(n)\right) \hat{\mathbf{R}}_{xx}^{-1}(n) + \mu \mathbf{I}$$

Multiplying both sides of the equation by $\mathbf{r}_{xd}$ on the right we have

$$\mathbf{w}_{n+1} = \left[\mathbf{I} - \mu \mathbf{x}(n)\mathbf{x}^T(n)\right] \mathbf{w}_n + \mu \mathbf{r}_{dx}$$

c) If we use the approximation $\mathbf{r}_{dx} = d(n)x(n)$ then the above recursion becomes equivalent to the LMS algorithm.

5/5

5) [**New example**]
We first need to evaluate the gradient of $J'(n)$, that is

$$\nabla J'(n) = \nabla[1/2e^2(n)] + \beta/2\nabla[\mathbf{w}^T(n)\mathbf{w}(n)]$$

Using the standard LMS type of approach we have

$$\begin{aligned} \nabla[1/2e^2(n)] &= -e(n)\mathbf{x}(n) \\ \nabla[1/2\mathbf{w}^T(n)\mathbf{w}(n)] &= \mathbf{w}(n) \end{aligned}$$

which gives the final update based on $J'(n)$

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu e(n)\mathbf{x}(n) - \mu\beta\mathbf{w}(n) = (1-\mu\beta)\,\mathbf{w}(n) + \mu e(n)\mathbf{x}(n)$$

b) The additional term in $J'(n)$ represents the regularisation parameter. Cost function $J'(n)$ is convex and has a unique minimum, which provides balance between the minimisation of MSE and the size of the weights. The principle behind this cost function is similar to that in methods for determining the order of an AR model (MDL, AIC).

c) Using the standard LMS type approach we have

$$E\{\mathbf{w}(n+1)\} = [(1-\mu\beta)\mathbf{I} - \mu\mathbf{R}_{xx}]\,E\{\mathbf{w}(n)\} + \mu\mathbf{r}_{dx}$$

Therefore for stability (convergence in the mean), we require

$$|(1-\beta\mu) - \mu\lambda_k| < 1, \quad k = 0, 1, 2, \ldots, p \quad \Rightarrow \quad 0 < \mu < \frac{2}{\beta + \lambda_{max}}$$

d) From

$$\lambda(n+1) = \lambda(n) - \nabla_\lambda \frac{1}{2} e^2(n)$$

and knowing that the inputs to the λ update are the outputs of the consitutive filters $y_1(n)$ and $y_2(n)$ we obtain the update equation straightforwardly.

Clearly, cost functions with two terms can be thought of as two adaptive filters running in parallel. If those terms were combined in a convex manner with an adaptive λ, the cost function $J'(n)$ could be seen as a cost function of a hybrid filter.